Translating Fieldwork into Datasets:

the development of a corpus for the quantitative investigation of grammatical phenomena in Eibela

Grant Aiton
Grant.Aiton@anu.edu.au
Centre of Excellence for the Dynamics of Language
Australian National University
9 Fellows Road, Canberra, Australia

Abstract

This extended abstract details the process of constructing an annotated XML corpus suitable for quantitative analysis of morphosyntactic and phonetic phenomena in the Eibela language of Papua New Guinea. Preliminary results will also be included, which investigate the semantic, phonetic, and discourse correlates of argument realization. The goal of this paper is to illustrate how legacy materials can be enriched and investigated using computational methodologies including forced alignment of phonetic segments using bulk processing of data in Python and R, the Montreal Forced Aligner (MFA), and morphosyntactic annotation developed as part of the Multilingual Corpus of Annotated Spoken Texts (Multi-CAST).

1 Introduction

This extended abstract details the process of constructing an annotated XML corpus suitable for quantitative analysis of morphosyntactic and phonetic phenomena in the Eibela language of Papua New Guinea. Preliminary results will also be included, which investigate the semantic, phonetic, and discourse correlates of argument realization. The goal of this paper is to illustrate how legacy materials can be enriched and investigated using computational methodologies including forced alignment of phonetic segments using bulk processing of data in Python and R, the Montreal Forced Aligner (MFA; McAuliffe et al. 2017), and morphosyntactic annotation developed as part of the Multilingual Corpus of Annotated Spoken Texts (Multi-CAST) (Haig and Schnell 2020).

2 Data Source and Language Information

Eibela (ISO 639-3 AIL, also appearing as Aimele or Kware in some sources) is a member of the Bosavi language family spanning the border between Western Province and Southern Highlands Province. There are currently approximately 250 speakers of the language, but there appears to be a linguistic shift to speaking the neighboring Kamula language and the national language Tok Pisin.

Data for this analysis is a subset of data gathered from 2012 to 2017 in Lake Campbell, Western Province, and includes approximately two hours of recorded narratives. This project is intended as a preliminary attempt at developing a workflow for processing legacy materials into an extensive annotated corpus suitable for extensive quantitative analysis of morphosyntactic and phonetic phenomena. To this end, recordings with multiple speakers were excluded and the recordings with the highest possible recording quality were prioritized for annotation and analysis. Conversational language or recordings with considerable background noise were excluded, but it is hoped that the corpus will be expanded to include these materials once these methodologies prove successful on more easily processed data. The data chosen for this corpus initially had narrow phonetic transcriptions and free translation of clauses aligned to an audio file using ELAN (2020), and were then processed according to the workflow described in section 3.

Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages: Vol. 2 Resource Papers and Extended Abstracts, pages 1–3,
Online, March 2–3, 2021.

3 Corpus Construction Methodology

The annotation of this data into a research corpus follows three stages of processing. The initial stages of data annotation follow an ELAN-FLEX-ELAN workflow first introduced to this researcher during training by the Endangered Languages Documentation Programme using training materials developed by Tim Gaved, Sophie Salffner, and Katia Chikova (Gaved et al 2015). Following this workflow, free translations and phonetic transcriptions were first exported from ELAN to a Fieldworks Language Explorer database (FIELDWORKS 2020), where the phonetic transcription underwent morphosyntactic parsing and interlinearization. Once this was complete, the texts were reimported into ELAN in order to align morphosyntactic annotations and morpheme segmentations with the original audio recording. This resulted in a searchable corpus of time-aligned ELAN files with part of speech tagging, glossing data, free translations and morphophonological alternations.

3.1 Multi-CAST Annotation

The interlinearized ELAN files generated in the ELAN-FLEX-ELAN workflow above were then used to create a corpus of texts for the Multilingual Corpus of Annotated Spoken Texts (Multi-CAST). This annotation scheme focuses on the syntactic and semantic properties of referents in discourse and is split into two primary annotation schemes. First, the GRAID (Grammatical Relations and Animacy in Discourse) annotation scheme labels the animacy, person reference, form, and syntactic function of every referent (Haig and Schnell 2014). Animacy of arguments was classified as human, anthropomorphized non-human, animal, and non-animate. Person marking was divided into first, second, or third person, with no explicit labeling of number distinctions. The form of the referent refers to whether the referent is realized as a noun phrase, a reduced form such as a pronoun, or is not overtly realized. Zero arguments are labeled and included in the annotation in cases where the referent is clearly identifiable in discourse and an overt argument could grammatically be present in the construction fulfilling that syntactic role. Finally, the syntactic role of referents were labeled as S, A, or P according to the transitivity of the clause and syntactic role of the argument. Predicates, clause boundaries, and non-referential elements of a clause are labeled as well, but in far less detail. For the purposes of this project, additional annotations were added beyond the standard GRAID annotations, including case marking, topicality, affectedness on arguments, and switch reference marking in clauses. The second annotation scheme is RefIND (Referent Indexing in Natural-language Discourse; Schiborr, Schnell, Thiele 2018), which assigns every referential entity of a text a unique identifier in order to track the number of references and proximity of reference for specific discourse entities. This allows concepts related to newness, topicality, and contrasting reference to be approximated quantitatively.

3.2 Phonetic Annotation with Montreal Forced Aligner and PRAAT

The final stage of annotation required the phonetic parsing and forced alignment of segments using the Montreal Forced Aligner (MFA) (McAuliffe et al 2017). This required the automated export of PRAAT TEXTGRID files of the phonetic transcription tiers from the existing ELAN files using the Python pympi-ling module (Lubbers and Torreira 2018), which were then processed by MFA in order to segment phrase level transcriptions into individual phones, which were then aligned with the acoustic data. The resulting segmented TEXTGRID files were then used to generate measurements of frequency, formant values, and amplitude using PRAAT and reimported into the ELAN EAF file for the text. The reliability of the automatic alignment was verified by hand-checking a subset of the data for each speaker, as well as verifying the accuracy of any unusual outliers identified during exploratory data analysis.

4 Applications and Research Plans

These three levels of annotation have all been merged into a single comprehensive XML database which can be used to investigate an enormous variety of analytical questions regarding the cooccurrence of morphological, syntactic, and phonetic phenomena. This database was created using ELAN to add data sources as tiers, and then using the resulting EAF file as an XML database which may then be imported as a data source into R for quantitative analysis. Each level of annotation offers rich avenues of exploration, and the creation of a single database expands the possibilities exponentially. For example, FLEX interlinearization allows the investigation of morphologically annotated concordances, allowing the discovery of phrasal collocations and the extent to which various bound morphemes cooccur in natural speech. The addition of Multi-CAST annotations allows the additional factors of topicality, animacy, and syntactic role to be investigated as a conditioning factor in morphological realization and phrasal collocation. Similarly, the Multi-CAST annotations alone allow a researcher to track whether the constituent order, prior mentions, or semantic properties of a particular referent affect the formal realization of that referent in terms of case marking, clause position, or reduced form as a pronoun or zero argument. The addition of phonetic data further allows a researcher to ask whether the number of mentions affects the pitch or duration of a referential NP or pronoun, or whether dislocated arguments of a clause have unique intonational characteristics.

The first pilot study using this corpus is an investigation of optional case-marking and argument realization in Eibela using mixed effects statistical modelling. Arguments are frequently elided from discourse in Eibela, and the core ergative and absolutive case-marking suffixes may also be omitted. This investigation will therefore take a detailed look at how syntactic, morphological, and discourse factors influence whether an argument of a clause is case-marked, a zero argument, a full noun phrase, or a reduced pronominal form. Among the conditioning factors will be whether the argument is dislocated such that it deviates from Eibela's canonical SV/APV constituent order, it's syntactic role (A, S or P), number of prior references to the referent in the discourse (i.e. "newness"), and semantic factors such as animacy, affectedness, and agency. Phonetic characteristics such as pitch and amplitude will also be investigated in order to determine the intonational attributes of various types of arguments.

Acknowledgments

This research is supported by funding from the Endangered Languages Documentation Programme, the Firebird Foundation for Anthropological Research, the Centre of Excellence for the Dynamics of Language, and the Australian National University.

References

ELAN (Version 5.9) [Computer software]. 2020. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan/.

FIELDWORKS (Version 9.0) [Computer software]. 2020. SIL International. Retrieved from https://software.sil.org/fieldworks/.

Gaved, Tim, Sophie Salffner, and Katia Chirkova. 2015. Working with ELAN and FLEx together. (https://archive.mpi.nl/tla/elan/thirdparty?fbclid=IwAR01wkP0NppCfDJO0fTmGae8DgR79yL85Mc2lJ9eoBpUfdDi0D4Rg5_HQ4A) (Accessed 2020-09-08)

Haig, Geoffrey & Schnell, Stefan (eds.), Multi-CAST: Multilingual corpus of annotated spoken texts. (multicast.aspra.uni-bamberg.de/) (Accessed 2020-09-08).

Haig, Geoffrey & Schnell, Stefan. 2014. Annotations using GRAID (Grammatical Relations and Animacy in Discourse): Introduction and guidelines for annotators (version 7.0). (https://multicast.aspra.uni-bamberg.de/#annotations) (Accessed 2020-09-08).

Lubbers, Mart and Torreira, Francisco. 2018. pympi-ling: a Python module for processing ELAN's EAF and Praat's TextGrid annotation files (Version 1.69). (https://pypi.python.org/pypi/pympi-ling) (Accessed 2020-09-08).

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner [Computer program] (Version 0.9.0), (http://montrealcorpustools.github.io/Montreal-Forced-Aligner/) (accessed 202-09-08).

Schiborr, Nils N., Stefan Schnell, and Hanna Thiele. 2018. RefIND — Referent indexing in natural-language discourse Annotation guidelines (version 1.1). (https://multicast.aspra.uni-bamberg.de/#annotations) (Accessed 2020-09-08).