Diversifying Documentary Corpora

Tyler M. Heston
University of Hawai‘i at Mānoa
“If there is one and only one chance ever to document a language, it makes especially good sense to strive for a result that is as complete as may be possible”

Woodbury (2011:181)
“If there is one and only one chance ever to document a language, it makes especially good sense to strive for a result that is as complete as may be possible”

Woodbury (2011:181)
What are the elements of a maximally useful corpus?
Overview

1. The role of controlled data in a corpus
2. Support from first-hand fieldwork
3. Best practice recommendations
Naturally-occurring data

Recording of any relatively *natural* communicative event

Relatively little researcher intervention

Stories, rituals, conversations, monologues...
Controlled data

Linguistic data organized by the researcher

Wordlists, staged events, perception experiments, grammatical elicitation...

Focus today on phonetic production data
Disciplinary focus on naturally-occurring data

Controlled data recognized as necessary (Lüpke 2010, Himmelmann 2007), but often disparaged

I argue controlled data forms a crucial component of a well-rounded corpus
Natural data

Broad usefulness (e.g. Lüpke 2010)
  › Syntax, morphology, discourse, conversation analysis, naturalistic prosody...

Both naturalistic and multipurpose

Extremely difficult to isolate confounding factors
Goals

If goal is a diverse, multipurpose corpus, (Himmelmann 2006)
... we must accept wide variety of data

While texts are ideal for some research, for other questions, controlled data is essential
Methodology

Naturalistic observation and experimentation complement one another

Neither complete in itself

Important to acknowledge benefits and limitations of each available data type
Benefits

Controlled data helps **answer key questions**

Makes corpus relevant to **broader audience**

Establishes a **baseline** for the analysis of naturally-occurring data
In the Field
East Timor
Goals

Documentation

Phonological analysis

Problematic word prosody
Challenges of texts

Texts are naturalistic, but confound all levels

Is a pitch peak:
- tone?
- stress?
- pitch accent?
- phrase-level intonation?
- segmental aberration?
Analyzing words in isolation offers similar challenge.

In isolation, each word is also an utterance.

No way to differentiate word- and sentence-prosody (Himmelmann 2007).
Strategies – pt. 1

Analyzing sentences of 2-3 words (Jun & Fletcher 2015)

Voiced segments ideal

Manipulate word length and position
Carrier phrases ensure same segmental and prosodic environment (Maddieson 2001)

- “___ is an English word”
- “Say ___ again”
- “I said ___”
Results from Fataluku

Previous work confounded word- and phrase-prosody (e.g. Campagnolo 1973)

“Accent” or “tone” actually phrase-level contour

Future work to address prosody in naturally-occurring speech
Experimental data crucial to understanding Fataluku prosody, and therefore texts

Similar needs for segmental phonetic analysis

Extremely beneficial, very low time commitment
What data is most useful for **phonetic analysis**?
Recommendations: Minimal

10 voiced three-word sentences

Every phoneme illustrated in the same neutral context, in a frame

3 repetitions x 4 speakers (2M 2F)
Recommendations: Better

15 voiced sentences, as declarative, question, command, broad and narrow focus

Every phoneme in every position

3 repetitions x 8-12 speakers (M & F)
Importance of controlled data

Controlled data is crucial for disentangling confounds, especially in phonetics

Equally applicable to other subfields

Establishes baseline for textual data
Conclusion

Linguistic analysis often difficult from texts alone (e.g., prosody)

Need only a small amount of carefully-selected data

If a corpus is to serve as a final, multipurpose record of a language, controlled data crucial
Selected References
