4.4 Advantages of Well-Formed Data in Analysis, Publication, and Presentation

Interaction with field recordings via their aligned transcripts allows the transcripts to improve as your understanding of the language grows. As an example, the first author's research language, South Efate, has two phonologically similar pronouns, *ga* '3SG' and *gaŋ* '2SG.POS', that need to be carefully distinguished, because in preverbal position the latter in fact acts as a benefactive. In the early versions of the South Efate transcripts, the distinction between pronouns was not always made correctly because of the difficulty of hearing the final velar nasal in *gaŋ*. As the author's knowledge of the language—and his awareness of the benefactive forms—grew, he was able to easily return to the primary recordings and look for discrepancies in the transcript. By having well-structured data, he was able to improve the transcription and confirm the analysis of benefactives in South Efate.

p. 116 A well-formed corpus allows us to seek answers to linguistic questions that are difficult to ask when data is limited to what can be expressed on the printed page. Much of language structure that is embodied in phenomena like intonation, tempo, and stress, is observable only acoustically; certain aspects of spontaneous language use like disfluencies, gesture, and speaker overlap are difficult to represent on paper. Even the best transcription systems—for example, ToBI (Beckman and Hirschberg 1994), Discourse Transcription (Du Bois et al. 1992; 1993), Conversation Analysis (Schegloff 2007)—are only substitutes for direct observation. Being able to cite archived, time-aligned examples of the phenomenon you describe in your publications by their permanent handles, down to the level of the clause, word, or even segment, allows others to confirm, challenge, and build on your claims. Well-formed data is a powerful tool for research and publication, allowing linguistics as a discipline to participate in the practice of open data that has long been an ethos of scientific inquiry.[36]

The possibilities for multimedia presentation of language material are enticing, and linguists have long recognized the value of dynamic media in capturing and representing the dynamism of language. When multimedia software was first developed sufficiently, a number of projects like CD-based talking dictionaries immediately took advantage of the ability to link text to audio or video as a pedagogical and presentational tool. Unfortunately, many of these early projects had very limited lives. As the software they were built in aged, they became unplayable. Even worse, in some cases, producing the multimedia CDs was the primary goal of the language project, so the data has become effectively lost.

A better objective is to build well-formed, archival data from fieldwork, and then to derive delivery forms of the media from that base. These delivery forms can be on-line or local or paper-based, but they all benefit from being derived from well-structured underlying data. Time-aligned transcripts, with their links to specific points in their associated media, will allow you to install an iTunes jukebox in a local school (see also Barwick §7.3.4.1 below), or develop a simple Flash presentation online or on CD. Well-structured lexical databases will allow you to deliver targeted topical dictionaries on paper or over the web.