

Figure 4.2.

\id	061:005
\aud	AHT-MP-20100305-Session.wav 02: 19.320-02: 21.780
\tx	Ga ɪdu' ben yii taghiɪ'aa.
\mr	ga ɪdu' ben yii ta- ghi- ɬ- 'aa
\mg	DEM FOC lake in water ASP CLF linear.extend
\fg	'As for that one (river), it flows into the lake.'

Structure of interlinear glossed texts.

4.1.4 Reuse of field data

Reuse of our field recordings and research outputs is central to the methodology advocated here. We need to be able to reuse our own field materials in our analysis of the language, and also ensure that others can access them in the future. It is increasingly common for linguists to deposit primary material with an archive early and then to regularly keep adding to their collection. Once items are archived, any subsequent work—transcribing, annotating, or selecting example sentences for published analyses—can be cited back to an archival file. The archive provides this citability for you, so if you archive your files before you begin your analysis, you allow your research to be embedded within the data (Thieberger 2009) and then to be verified by other researchers with reference to the primary data.

4.1.5 A workflow for well-formed data

Creating well-formed data from linguistic fieldwork entails a workflow like the one in Fig. 4.3. The workflow begins with project planning, which for our purposes includes preparing to use technology in the field and deciding on which file naming and metadata conventions you will use before you make your first recording. (Clearly there is a great deal of non-data-related planning to be done before the first recording can be made, but that is not the focus of this chapter.) After recordings are made and metadata are collected, data must be transcribed and annotated with various software tools, and then used for analysis and in ↴ representation of the language via print or multimedia. Note that depositing materials in an archive is carried out at every phase of the procedure.

Because you will subject your data to a number of processes, there is always a risk that some parts of it may be lost along the way if the outputs of one software tool do not align exactly with the input requirements of the next tool in the workflow. For example, importing a time-aligned transcript from ELAN (see §4.3.5) into Toolbox for interlinearization may result in some extra fields appearing in Toolbox. These fields correspond to the original ELAN data categories, but are not necessary for the Toolbox interlinearization process. You must take care, then, that you do not delete these extra fields so that they will still be present when you re-import the data into ELAN.

Another significant problem that arises from using a variety of tools in a workflow is that the exports of one tool may not be in a format that fits into the import routine of the next tool in the workflow. For example, time-alignment software (e.g. ELAN or Transcriber) typically produces a native XML file which can be exported to a number of interchange formats, but none of these can be opened *directly* in, say, Toolbox or Fieldworks (although more import and export formats are being added to these tools regularly, so this may