The LST Guide to Your Linguistic Career

2009 Fieldwork Workshop

Department of English, National Taiwan Normal University

April 11, 2009

Field Data Management

Hsiu-chuan Liao

National Tsing Hua University

Linguistic Project: Processes

- ***recording***: media (audio, video, image) and text
- ***capture***: the encoding and transfer of an analogue recording (as on a cassette or reel-to-reel tape) or text written on paper to the digital domain as a computer file
- ***analysis***: transcription, translation, annotation, and notation of metadata
- ***archiving***: creating archived objects, and assigning access and usage rights
- ***mobilization***: publication and distribution of the material in various forms

Fieldwork: Things to be Noted-I

- ***Never*** record in a compressed format (e.g. MP3)
- ***Never*** record direct to computer hard-disk.
  -Such techniques risk irrecoverable data loss.
- Make note of ***abbreviations*** and ***symbols*** used in the front of the notebook.
- Start each section of notes with the ***date***, ***time*** and ***place***, and ***speaker***.

Fieldwork: Things to be Noted-II

- Fieldnotes should be written in ***ball-point pen*** (*not* pencil and *not* washable ink!).
  --***Ball-point pens*** seem to be more durable—they are smudge-proof (cf. *Ink* will bleed when the paper gets wet; *pencil* does smudge if you are carrying your notebook around a lot and the pages rub against each other).
- Write on ***one side*** of the page *only*.
  --Use the other side for notes, later corrections, and cross-references.
- ***Do not*** write on ***every line***.
  --If you cross things out or write glosses above words in your notes, you will *not* want to write on every line.

Fieldwork: Things to be Noted-III

- **Make sure that pieces of paper are *not* lost**.
  -- ***Avoid*** the temptation of **writing vocabulary items on the backs of envelopes**, or at the very least **stick the envelope into a notebook** or **copy the information to your fieldnotes *as soon as possible***.
- **Always use *a notebook with a binding***:
  (a) a ***hardback*** *notebook*: durable and stable binding, but quite heavy
  (b) a ***spiral bound*** (A5 100-page) *notebook* with cardboard or plastic covers: easily balanced on one's knee and not too heavy

Fieldwork: Things to be Noted-IV

- ***Go over your notes after the session*** and ***add anything that you remember from the session that isn't in the notes and annotate your fieldnotes*** **for hypotheses about the data**, questions, comments, and notes on cross-referencing.
  -- Use *highlighters* or *post-it notes* and *page markers* (although bear in mind that highlighters aren't very archivally friendly) or *different colored pens* (but that information will be lost when you photocopy your notes)
  -- ***Tag comments in your database electronically*** (this is preferable).

Fieldwork: Things to be Noted-V

- At the very least you will want to **organize your data** so that you can find things in it later on. Therefore, it is worth investing some time and thought in how you organize your fieldnotes.

  -- ***Don't* rely on your memory**; you won't remember what's in which recording in six months, or what a particular question meant. Did it mean that the gloss is suspicious, or that you aren't quite sure of the transcription, or that you want to check that the word is in your main database?

  -- **The media should be *durable***. It would be a disaster if you lost everything because someone spilt a drink on your laptop, or you lose the piece of paper with your coding system.

Make a Backup-I

- **Make backup copies of your recordings *as soon as possible*.**
  --This goes for *audio recordings* and *any other files* (e.g. *your database* where you keep all the information about what is in which recording).
- **Make sure that the backup worked** (i.e. that the copy did ***not*** become corrupted).
- ***Never* directly edit your *original* recordings.**
  --You could accidentally delete part of the recording, or resample it, or a file could become corrupted.

Make a Backup-II

- As soon as the media is recorded, **make sure you *cannot* accidentally write over it.**
- When transferring files, **make sure that you always know which folder is the *original* version and which is the *updated* version**.
- Make sure you have ***audition sheets* for the recordings you have made** and that **your recordings are labeled in an *informative* way**.

Make a Backup-III

- There are several ways of backing up your recordings and files. At least ***two*** different backups (stored in different places) are recommended.

  -- ***Portable hard drives*** are not very expensive.

  -- ***Data DVDs*** are an efficient way of backing up large amounts of sound data.

  -- ***CDs*** are also good.

Make a Backup-IV

- ***How often should you make backups of your data?***

  -- Back up your transcription files ***each day*** in the field, and burn CDs ***every three or four days*** (depending on how much ***transcriptions*** that you have done).

  -- Back up audio files ***as soon as it's transferred to computer***, and post the backups every few weeks to your snail mail address.

Capture-I

- ***Capture***: the encoding and transfer of an analogue recording (as on a cassette or reel-to-reel tape) or text written on paper to *the digital domain* as a computer file.
- When using digital capture software, it is important to make sure you ***use appropriate settings***.
- It is advisable **to transfer fieldnotes from notebooks to computer files, ideally *as soon as possible after recording*** so you do not forget notes, abbreviations, and comments.

Capture-II

- As for ***recording***, it is imperative **to name your computer files *consistently* and *clearly***, making sure that you should *not* rely on directory structure to disambiguate file names.
  -- Different naming schemes can be used, but *clarity* and *transparency* is the goal.

- It is also essential **to record the relevant *metadata* for the data files you create** as you make them, ideally *in a structured way* such as a relational table using standard terminology.

  -- ***metadata***: *data about data*, i.e. structured information about events, recordings, and data files.

Structure of the Project

- **Work out a basic directory structure *before* you leave for the field**. That way, you will not be trying to organize parts and files the same time that you are collecting your first data.

- **Consider *the directory or file structure* that you will use**: where will you keep all the different files that you will be creating?

  (a) Keep all files in a *single directory*.

  (b) ***A set of directories* is preferable:** have different directories for *audio files*, *transcriptions*, *budget forms* and *other reports*, *secondary analysis* such as articles, assignments, or your dissertation; *lexicon files*, or *other categories of this type* that you find useful.

Ways of Organizing Fieldnotes

- The most useful system will depends on *how you organize your field sessions*.

  -- Use ***several notebooks****: one for elicitation, one for transcription of texts, and another for miscellaneous queries*.

  --Use *a* ***single notebook*** *for everything*.

Organization of the Recordings

- **Record a *summary* of what is on the track at the end of the day.**

- **Do *an audition sheet* for your recordings at the end of each day.**

- ***Having some sort of numbering systems*:**

  -- Number each tape or recording through your ***career*** (e.g. 1-1,000).

  -- Number by ***collection***: Each fieldtrip has a number (1.1-10, 2.1-100) etc.

  -- Number sequentially by ***language* worked on** (Tag1-100, Ilk1-100).

  -- **Use *the date of recording*** (20090307a, 20090307b, 20090411).

Label the Recordings

- Whatever system you use, **each recording should be *uniquely identified*.**
  --If you are using reusable media, such as compact flash cards, each session (or ‘episode’) might get its own number. Whatever the system, stick to it and document it.

Other Recordings

- **Decide how *previous linguists’ recordings, or radio recordings*, or *recordings made by speakers themselves* will be incorporated into your cataloguing system (or if they will be).**

  -- **Will they get the *same* numbering system as your own recordings?** If so, how will you keep them separate?

  -- If they are kept separate, will you be able to find things on them?

Software for Data Processing-I

- ***Three*** **most important aspects of fieldwork software:**

  (a) You must be able to ***get your data into the program easily***.

  (b) You need to be able to ***find things in your data easily***.

  (c) You need to be able to ***get data out of the database***: in producing reports for the language community, getting examples out of your database and into the text of your reference grammar, and when converting between programs.

Software for Data Processing-II

- Some software programs are easier to use than others. If you find mastering new programs difficult, ***use one that is on the easier end***.
  -- There is no point in using a program with multiple capabilities if you know you won't be using any of them.
- ***Minimize*** **the time you have to spend retyping or reentering data between programs.**
  -- Ideally you should be able to transcribe your tapes and then move data around, annotate it and include it in a final write-up without having to retype.

Tools for Linguistic Analysis and Processing-I

- ***general purpose software***: the user must design the data structures and can write application programs to manipulate the data and carry out various tasks.

  -- *MS Word* and *Excel*, and *File Maker Pro*.

  -- Such software is powerful and flexible; however, they store data in a *proprietary format* which is *not optimal* for long-term storage and access.

Tools for Linguistic Analysis and Processing-II

- ***specific purpose software***: is designed to be used for particular tasks.

  -- ***Transcriber*** and ***EXMARaLDA*** (EXtensible MARkup Language for Discourse Annotation): for time-aligned audio annotations

  -- ***Shoebox***/***Toolbox***: for text and lexicon annotation

  -- ***Praat***: for speech analysis and annotation

  -- ***ELAN***: for audio and video annotation

  -- ***IMDI Browser***: for cataloguing and administration metadata

Ways to Organize Data-I

- **The simplest way to organize data is *to type everything in a word processor*.** You could have one file for your fieldnotes, another or your lexicon, and a third for analysis. **However, this approach is *not* recommended.**

  -- Formatting a dictionary completely by hand, as a Word document, with correct alphabetization, formatting and so on, **would be very difficult to do.**

  -- Even if you want to do something as simple as displaying all the nouns in the data, you will have to go through the examples by hand.

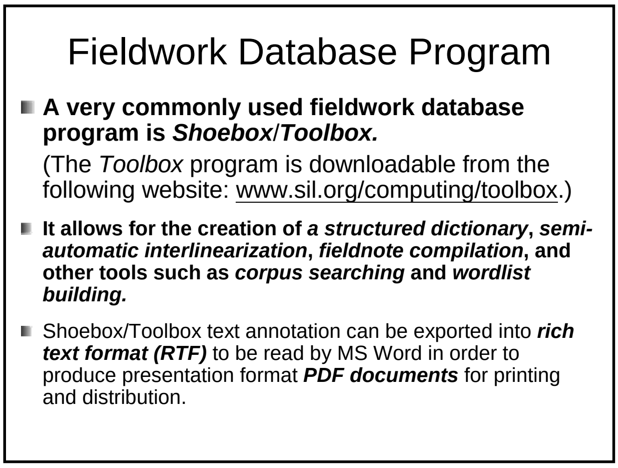
Ways to Organize Data-II

- A much better way **to store your fieldnotes is in a *database***.

  -- There are programs which let you build a dictionary and export the records in a consistent format to another program.

  -- Even in the 'old days' before computerized database software, **linguists used *card files* to organize their data** before a dictionary type-script was produced.

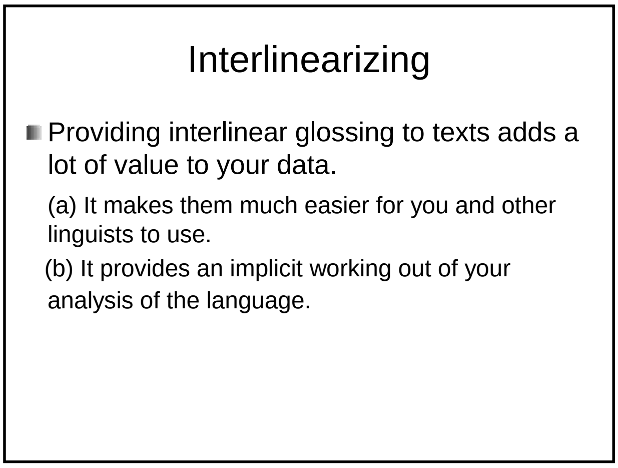
  -- A ***computer database*** allows you to do the same thing as the card files.



Fieldwork Database Program

- **A very commonly used fieldwork database program is *Shoebox*/*Toolbox.***

  (The *Toolbox* program is downloadable from the following website: www.sil.org/computing/toolbox.)
- **It allows for the creation of *a structured dictionary*, *semi-automatic interlinearization*, *fieldnote compilation*, and other tools such as *corpus searching* and *wordlist building.***
- Shoebox/Toolbox text annotation can be exported into ***rich text format (RTF)*** to be read by MS Word in order to produce presentation format ***PDF documents*** for printing and distribution.

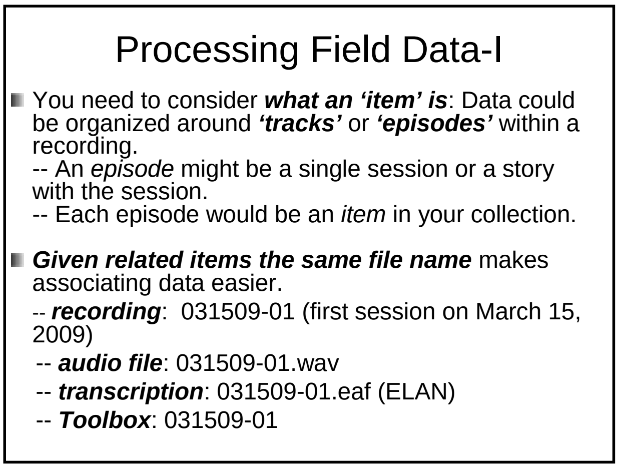


Interlinearizing

- Providing interlinear glossing to texts adds a lot of value to your data.

  (a) It makes them much easier for you and other linguists to use.

  (b) It provides an implicit working out of your analysis of the language.



Processing Field Data-I

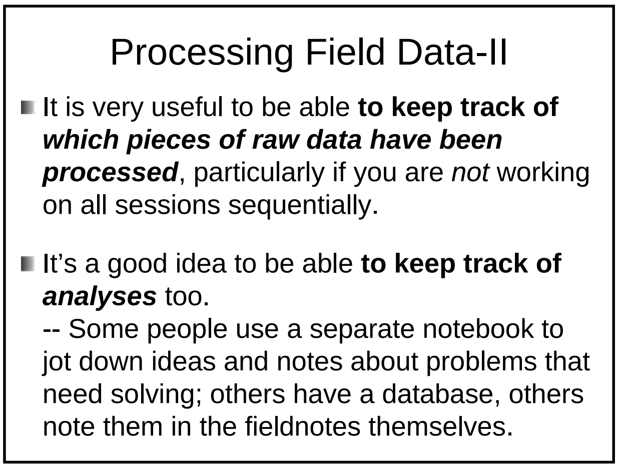
- You need to consider ***what an 'item' is***: Data could be organized around ***'tracks'*** or ***'episodes'*** within a recording.
  -- An *episode* might be a single session or a story with the session.
  -- Each episode would be an *item* in your collection.
- ***Given related items the same file name*** makes associating data easier.

  -- ***recording***: 031509-01 (first session on March 15, 2009)

  -- ***audio file***: 031509-01.wav

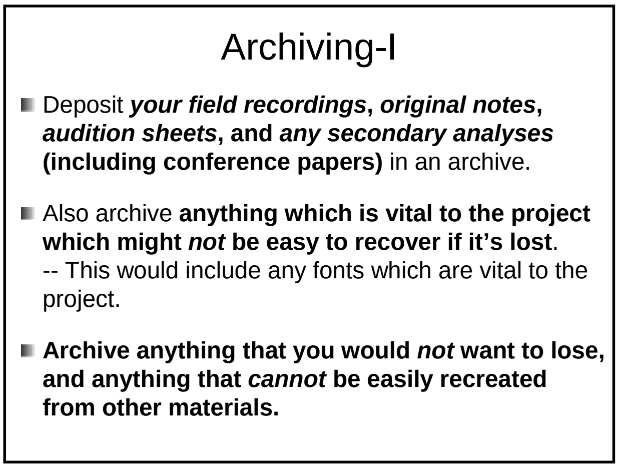
  -- ***transcription***: 031509-01.eaf (ELAN)

  -- ***Toolbox***: 031509-01



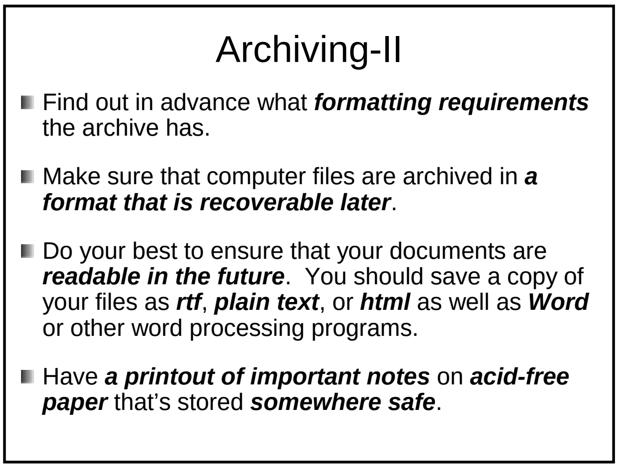
Processing Field Data-II

- It is very useful to be able **to keep track of *which pieces of raw data have been processed***, particularly if you are *not* working on all sessions sequentially.
- It's a good idea to be able **to keep track of *analyses*** too.
  -- Some people use a separate notebook to jot down ideas and notes about problems that need solving; others have a database, others note them in the fieldnotes themselves.



Archiving-I

- Deposit ***your field recordings***, ***original notes***, ***audition sheets***, **and *any secondary analyses* (including conference papers)** in an archive.
- Also archive **anything which is vital to the project which might *not* be easy to recover if it's lost**.
  -- This would include any fonts which are vital to the project.
- **Archive anything that you would *not* want to lose, and anything that *cannot* be easily recreated from other materials.**



Archiving-II

- Find out in advance what ***formatting requirements*** the archive has.
- Make sure that computer files are archived in ***a format that is recoverable later***.
- Do your best to ensure that your documents are ***readable in the future***. You should save a copy of your files as ***rtf***, ***plain text***, or ***html*** as well as ***Word*** or other word processing programs.
- Have ***a printout of important notes*** on ***acid-free paper*** that's stored ***somewhere safe***.

Data Formats in Different Contexts

	working	*archiving*	*presentation*
text	Word, XLS, FMpro, Shoebox/ Toolbox	XML	PDF, HTML
audio	WAV	WAV, BWF	MP3, WMA, RA
video	MPEG2	MPEG2, MPEG4	QuickTime, AVI, WMV

On-line Resources

- **Leipzig Glossing Rules** http://www.eva.mpg.de/lingua/files/morpheme.html
- Linguistics computing resources on the internet (http://www.sil.org/linguistics/computing.html)
- Typological tools for field linguistics (http://www.eva.mpg.de/lingua/tools-at-lingboard/tools.php)
- Praat: doing phonetics by computer (http://fonsg3.hum.uva.nl/praat/)
- WordCorr: A tool for comparative-historical linguists (http://www.wordcorr.org/)
- On-line journal: *Language Documentation & Conservation (LD&C)* (http://www.nflrc.hawaii.edu/ldc/)
- The Hans Rausing Endangered Languages Projects (http://www.hrelp.org/languages/resources/orel/)

References

Bowern, Claire. 2008. *Linguistic fieldwork: A practical guide*. Palgrave MacMillan.

Crowley, Terry. 2007. *Field linguistics: A beginner's guide*. Oxford: Oxford University Press.

Gippert, Jost, Nikolaus P. Himmelmann, and Ulrike Mosel, eds. 2006. *Essentials of language documentation*. Berlin and New York: Mouton de Gruyter.

Ladefoged, Peter. 2003. *Phonetic data analysis: An introduction to fieldwork and instrumental techniques*. Oxford: Blackwell.

Newman, Paul, and Martha Ratliff, eds. 2001. *Linguistic fieldwork*. Cambridge: Cambridge University Press.

Vaux, Bert. 2007. *Linguistic field methods*. Wipf & Stock Publishers.