A Plan for Managing the Data from a Language Documentation Project

Gary F. Simons

24 January 2009

1. Introduction

In a documentation project, the data files and their associated metadata accumulate quickly. Without a method for managing all the files and information, things can become overwhelming in no time at all, and data will start getting lost. This document describes a method for keeping track of that information using an Index spreadsheet that is introduced in section 2. Sections 3 through 6 describe the sheets within that spreadsheet which record, respectively, the items in the corpus, the files that comprise the items, the contributors, and category lists that are used for choices in the other sheets. Finally, section 7 talks about how to save an archival form of this information in the corpus submission.

2. The Index spreadsheet

The file MD_Index.xls in the course package is an empty Microsoft Excel spreadsheet that can be used to manage information about the items, files, and contributors in your project. Make a copy to your own disk and rename it appropriately. Go into the File / Properties template and fill in the metadata for your spreadsheet.

3. The *Items* sheet

The items in the corpus, with their descriptive metadata, are listed in the *Items* sheet. The first column records the unique Item ID for the item. This identifier is foundational to this system of data management; it is used in the File sheet and in the file names. It is recommended that you use a single letter to identify the basic item type according to Himmelmann's three-way distinction: E for a communicative event, L for an elicited list, and T for a discussion topic. This should be followed by a three-digit number (with leading zeroes) that is assigned serially within each category, e.g. E001, E002, E003, and so on. The advantage of using identifiers that include the leading zeroes is that they will sort into numerical order when the numbers reach the tens and hundreds.

The remaining columns of the sheet provide the descriptive metadata for the item. A few deserve special mention:

- *Event Type* and *Gender* are for recording dimensions of the sampling. They are linked to the pick lists on the Categories sheet. If gender is not a sampling dimension, remove it. If you have other sampling dimensions, add them. The point of these columns is to document which cell of your sampling grid an item falls into. They can be used in conjunction with the numbers in the *Event Words* column to compute the distribution of the corpus contents with respect to the various categories of the sampling dimensions.

- *Access* records the access category for the item, e.g. whether it can be openly accessed and used by anyone, or whether access and use are restricted in any way. The access categories come from a choice list in the *Categories* sheet. If you have any restricted items, you will need to add a category to the *Categories* sheet for each kind of restriction that needs to be enforced by the archive (and supply a definition following the category name in the *Categories* sheet).

- *Event Words* and *List Words* are used to record the contribution of the item toward the size of the corpus. For E (event) items, estimate the number of words of running text by multiplying the duration of the original recording by a typical speaking speed in words per minute and enter it into the *Event Words* column. For L (list) items, count the number of lexical items that were elicited and enter it into the *List Words* column.

- For *Speakers*, list the individuals who are performers in the event. Rather than a full name, use the short name that serves as the unique identifier in the *Contributors* sheet.

- For *Location,* name the geographical location, such as the village or town. For *Setting,* give the location from a more cultural standpoint, such as “In X’s home” or “Outside the church” or “Beside the river”. For *Situation*, give any other details that seem relevant about the context for the event or the course of events that caused the event to take place.

4. The *Files* sheet

All of the files that comprise the items of the corpus are listed in the *Files* sheet. The unique identifier for each file is the *File Name* recorded in the third column. The following four part scheme for naming files is recommended:

> ItemID-TypeCode-Other.Extension

The parts of the file name are as follows:

- *ItemID* is the unique one-letter-plus-three-digit item identifier from the *Items* sheet.
- *TypeCode* is a three-letter code for the documentation task represented by the content of the file (e.g. ori = original recording, otc = oral transcription, otl = oral translation, oxd = oral discussion, wtc = written transcription, wtl = written translation, wxd = written discussion). The codes given here are designed to have the property that when sorted, they come out in the order listed above. Thus, sorting the file names will list all the files for an item together and in the above order which corresponds to the temporal sequencing of the tasks.
- *Other* (with its preceding hyphen) is optional. It is used only when there is more than one file for a particular task type. When there is, add something to identify each uniquely.
- *Extension* is the regular file extension used by the operating system to identify the file type and associate it with the software that can open it.

For instance, E001-ori.wav is a single original audio recording, while E002-ori-sp1.wav and E002-ori-sp2.wav name the original recordings of the same event in which the two participants were recorded on separate devices.

During the actual recording events, the recording device will assign arbitrary file names. You will need to write these (and the label of the particular piece of recording medium) on your metadata sheets as you make a record of the recording session. However, these are not necessarily unique (since the device could assign the same names on different SD cards); nor are they very informative. It is recommended that as soon as possible, you transfer your recordings from the recording media to your computer, and in the process you rename the files according to the above scheme and enter them (with all the other metadata) into the *Files* sheet. It is also recommended that you speak the basic session metadata onto the beginning of the recording so that you can still know which file is which even if you lose the notes from the recording session.

Here are some comments on the other columns of the *Files* sheet:

- *Item ID* must match the ID from the *Items* sheet.
- *Task Type* gives a drop-down of the recognized documentation task types (from the *Categories* sheet).
- *Continued by* is used when a file does not go to the end of the event. The name of the file that continues the event is placed in this cell.
- *Date* and *Time* for an original recording are redundant with the date and time of the item on the *Items* sheet.
- For *Recordist* give the short name (from the *Contributors* sheet) of the person who operated the recording device. In the case of a written task type, it identifies the person who operated the software to create the written transcription.
- For *Commenter* give the short name (from the *Contributors* sheet) of the person who speaks the oral transcription or translation or discussion. For an original recording, leave it blank (since the original speakers are identified on the *Items* sheet). For written task types, leave it blank if it is the same as the person who created the file; otherwise, enter the short name of the person who wrote down the material that the “recordist” entered into the file.
- The next set of columns has to do with recording the technical data of the equipment and settings used for the recordings. Leave them blank for written task types.
- In the final *Notes* column, give notes that explain exactly what is in the file if there is more than one file for the same task type of the same item. The column should also be used to describe known problems in the recording.

5. The *Contributors* sheet

The *Contributors* sheet maintains an index of all the people who have contributed as speakers in the original recorders, as transcribers or translators, and as recordists. Each contributor is uniquely identified by a *Short Name.* This is not the person’s nickname, but the short identifier that is used in the other sheets to identify the original speakers of the items and the other contributors to all the files.

It is not necessary to fill in all the columns for people who contribute only by operating equipment, but for any whose voices are recorded or who make written transcriptions it is important to record basic background information like age, gender, level of education, location where they first learned the language, and first language of both parents.

Finally, this is the place to keep track of the informed consent documentation. If individual signed (or recorded) statements are compiled, then identify the file name for this person's consent document. Similarly if there are multiple group statements, use this column to record which group statement this person is covered by.

6. The *Categories* sheet

The final *Categories* sheet records the pick lists that are used in other sheets. You may change the category lists or add entirely new lists by editing this sheet.

After editing a category list, you will need to verify that the named region still encompasses the entire list. Click on the drop-down icon just above the A1 cell; it should drop down a list of the named regions. Select the name for the category list you have changed. If the region that is selected encompasses the complete list, then everything is okay. If the region does not include the whole list, then make a note of the exact region name that shows in the Name Box above A1. Now drag from the first category in the list to the last one in order to make a selection for the whole list. Finally, type the region name in the Name Box to assign the name to the region.

To make a new list altogether, create the list you want on the *Categories* sheet. Then assign it a region name by dragging from the first item to the last item and then typing a name into the Name Box above A1. Now go to the sheet on which you want to use the pick list and select the cells in which you want to use the pick list. Select the Validation option from the Data menu. Select "List" from the "Allow:" drop-down list; then in the Source box, type "=" followed by the name of the region you defined on the *Categories* sheet. When you select a cell in the column, it should now show a drop-down control that shows your choice list.

7. Creating the archival form

The Excel spreadsheet is in a proprietary format. It is suitable for use as a working form as you build the index to your corpus, but it is not suitable as an archival form since users who do not have Microsoft Excel will not be able to use it; nor would the users of the future when Microsoft changes its format.

For your documentary corpus, you should produce the information in a format like PDF or HTML that any user will be able to read. For instance, use the File / Save As Web Page command. If you save the entire workbook, it creates a single base file with a corresponding folder that contains all the individual sheets and other resources. Or, you can save the sheets as individual pages. To ensure that future users will also have access to all the metadata in a form they can manipulate, you can use File / Save As to save the individual sheets in tab-delimited TXT format or comma-delimited CSV format.