

## Transcription Format

### 1. Objective

The purpose of this document is to describe the format to be used for producing and checking transcriptions in this course.

### 2. Conventions

The main transcription system we will use in the course is described in “Outline of Discourse Transcription” (Du Bois et al., 1993) and *Discourse Transcription* (Du Bois et al., 1992), as updated in the Appendix to *Representing Discourse* (2005). Note that, although the theory, discussion, and examples of transcription *categories* as presented in the earlier documents generally remain valid, the transcription *symbols* have undergone a few revisions in the time since the earlier publications appeared. For all the transcribing we do in this course, it is important to use the most recent updated symbol conventions, as indicated in class handouts.

### 3. Font

You may wish to use a Unicode font, such as Gentium, Doulos SIL, or Arial Unicode MS. (The first two are available for free on the Internet from SIL International, while the last is supplied with most Microsoft software.) This will give you the widest range of options for using special symbols, including characters capable of writing any of the languages of the world in their traditional standard orthography, plus the International Phonetic Alphabet, and many other specialized symbols. (The more familiar Times New Roman font is another option; it has some Unicode characters, but not as complete a set.)

### 4. Tabs

Tabs are used in transcribing to help organize the presentation of transcription information, displaying it in a usefully iconic way. For our purposes, tabs are to be used for one purpose only: to distinguish between separate “fields” of data in each line. For a basic transcription, there are two fields. The first field in each line indicates who the speaker is (via a speaker label written in capital letters, followed by a semi-colon). The second field represents the actual utterance, or transcribed speech.

Define your tab location as follows (for the whole file). Define one tab at 2.5 cm (or 1 inch). Insert exactly one tab character in each transcription line:

- if there is a speaker label in the line, the tab is inserted immediately following it
- if there is no speaker label in the line, the tab is the first character of the line

Note that you should NOT use tabs for other purposes, such as inserting blank space in a line in order to align overlaps vertically (just insert spaces instead).

(In a more complex transcription there may be more than two fields, but this issue will be addressed separately.)

## 5. Line spacing

Do not use double-spacing, and do not leave blank lines between turns. (The exception is if you are transcribing a language other than English which needs glossing; see below).

## 6. Timestamp

Following the last word spoken in each weekly transcription increment, insert a time stamp into your transcription, indicating the location of the corresponding audio in your digitized audio file. For example, if the last word of your first week's transcription ends at 60.6 seconds into the computer audio file, insert the following notation at the end of the line: <T= 60.6>. This notation allows you, your teammate(s), and your instructor to quickly find the relevant place in the audio file, in order to keep track of where each week's new transcribed portion begins for checking purposes, and so on. You may wish to add additional time stamps at various points in your transcription, for convenience in locating specific portions of the transcription, for checking, feedback/consulting sessions, presenting segment proposals, and so on. (Or, you may find it useful to time-align the entire transcription with the corresponding audio, using appropriate software. This will amount to time-stamping the start and end of every intonation unit.)

## 7. Index

It is useful to index the transcription to make it easier to refer to specific portions of the data. Each line (or other unit) in the transcription is given its own unique index. There are at two methods which are particularly useful for this: numbering each line, and/or timestamping each line.

For line numbering, it is best not to number lines manually, since any revision of the transcription which adds even one new intonation unit could require you to renumber the entire subsequent transcription. It is better to add (temporary) line numbers automatically, at least while the transcription is in progress and has not been finalized.

***In OpenOffice.org Writer:*** To automatically insert line numbers in the entire document, select "Tools/Line Numbering" from the main menu. Check the box for "Show numbering," set the "Interval" to 1, and then click OK.

***In Microsoft Word:*** To automatically insert line numbers in the entire document, select "File" from the menu; select PageSetup/Layout/LineNumbers; check "Add line numbering" and "Continuous".

Timestamping is more stable, but can be somewhat laborious. If you choose this method, you should do it using software specifically designed to facilitate timestamping (e.g. SoundForge, Transcriber, etc.).

***In SoundForge (etc.):*** To manually insert time stamps for each intonation unit, one at a time, follow the procedures discussed in the relevant manuals or handouts for SoundForge, or other similar program.

## 8. Anonymity

Names of participants used in the transcripts (both in speaker labels, and when one discourse participant utters another participant's name) should not be the speakers' real names, unless you have received written consent from the participant to use their first name. Otherwise, make up an appropriate pseudonym – one which is comparable sociolinguistically and prosodically – and use it consistently. Any pseudonym (or other words you have modified because they could compromise anonymity, such as an utterance of a participant's street address or phone number) should be indicated with the pseudograph

symbol (i.e. tilde ~). (Note, however, that the tilde is used only for uttered words on the recording, not for speaker attribution labels, even if these are pseudonyms.)

For speaker attribution labels, give the complete first name, all in capital letters, not just an initial (e.g. use JILL: rather than J:).

Before disseminating your recording via the web or CD-ROM, you will need to make sure that any anonymity-sensitive words have been “bleeped” or otherwise obscured. You will learn simple techniques for doing this in this course.

## **9. Checking Format**

Indicate your suggestions to your teammate by marking your proposed changes in a copy of your teammate’s computer file. Use the “Record Changes” function in OpenOffice.org Writer (see below), or the “Track Changes” function in Microsoft Word, or the equivalent function in another program. This will allow you, your partner, and the instructor to easily distinguish between your suggested changes and the original transcription. Your suggestions will normally be displayed in underlined or strikeout font, or in color, or some combination of these. In addition, they may include the name of the person who made the change.

Using the “Record Changes” or “Track Changes” function on a word processor file will also allow you to exchange the checked files easily via email attachments. (If you are unable to use the Track Changes or Record Changes function, you may mark up a paper copy of your partner’s transcription, using colored ink or pencil so it will stand out.)

**OpenOffice.org Writer:** To turn on the “Record Changes” function, select Edit/Changes/Record from the main menu. Your insertions and deletions will now be tracked (using a format similar to that used by Microsoft Word, see below). To review and accept or reject your partner’s suggestions, from the main menu select Edit/Changes/Accept\_or\_Reject.

**Microsoft Word:** To turn on the “Track Changes” function, select Tools/Track Changes from the menu. Or, as a shortcut, type CTRL-SHIFT-e. To turn off the Track Changes function, do the same again. To view the changes in Word, make sure the “Reviewing” toolbar is displayed. Normally it will be displayed automatically once you select the “Track Changes” function. (To display the “Reviewing” toolbar, if necessary: From the menu, select View/Toolbars, and then make sure the “Reviewing” option is checked.) The reviewing toolbar has several icons representing functions you may find useful for accepting (or rejecting) your partner’s suggestions. You may wish to experiment with these.

## **10. Glossing**

If you are transcribing a language other than English, you may have to gloss it for your instructor. Normally the best way to do this is with interlinear gloss format (two-line or three-line format). Glossing requires some extra work, but the burden is relatively small in comparison with the work of transcribing. (For extensive information on glossing, see the “Leipzig Glossing Rules” at <http://www.eva.mpg.de/lingua/files/morpheme.html>.)

## **11. File Sharing**

Because of the emphasis on teamwork, this course will involve a lot of sharing of recordings and transcriptions in the form of computer files. We will do as much as possible of this data-sharing via the

Internet (and/or the computers and local area network in the Linguistics Lab). Further details about how to exchange data files with your partners will be given separately. The two most important types of files we will be using are the transcription file and the recording file.

## 12. Transcription File

You should transcribe each recording in a single computer file (e.g. a word processor file). You will add new material to this file each week, incorporating each additional minute of the recording as you transcribe it.

If possible, use **OpenOffice.org Writer 2.0** (or later), which saves your data in the Open Document format (**.odt**). This has the great advantage that your data will be saved in an open standard (using open source software). If this is not feasible for you, you may use another word processor such as Microsoft Word (**.doc**). (If you use a different word processor from either of these, save your file in the “plain text” (**.TXT**) or HTML formats, to facilitate sharing with other course participants.)

## 13. Recording Files

You are responsible for making sure that the appropriate digitized computer files corresponding to your recording (audio and/or video) are available to all relevant participants (e.g. your teammates and the instructor), via course web pages, the Internet, CD-ROM, and/or the Linguistics Lab computers and the Linguistics Department’s Local Area Network. Your computer files should be available in advance of the time they will be needed, e.g. prior to your team meetings, classroom feedback sessions, consultations in office hours, and so on, as well as being ready for your partner to check. Test your file setup before any meeting to make sure it works with the computer you will be using.

To include your files as part of the Class Corpus for Linguistics 212 at UC Santa Barbara for the year 2005 (for example), look in the “Everyone” drive (a.k.a. the “i drive”). There will be a folder with a name like “**Ling212**,” and under this, the “**Class05**” directory. You should make a point of regularly placing your updated files there when you are ready to share them.

## 14. Filenames

To make it possible for course participants (especially your instructor and your partner) to reliably keep track of all the computer data files that we will be exchanging, you should use a consistent practice to assign filenames to all your files. This includes your various recording files (called Transcription Excerpt, Research Segment, and Complete), your transcription file, and your checking file. Note that the checking file will normally be a file that was first originated by your partner (as her/his transcription file), which you have subsequently modified by adding your own corrections, suggestions, and comments to it. Each filename should make it clear who the file belongs to (i.e. you, as the one who worked on it most recently), and what it contains, including the relevant minute of transcription and checking.

Construct the filenames for your various digitized *recording files* as follows. Use your name (first or last, but let’s be consistent), in lower case, as a starting point.

- For your *Transcription Excerpt* recording file (usually 4-8 minutes long): **yourname.wav**
- For your *Research Segment* recording file (usually 20-30 minutes long): , add “**\_segment**”.
- If you made a “*Complete*” digitized version of your recording file (i.e. a digitized file corresponding to the entire original recording), add “**\_complete**”.

Avoid using spaces in the filenames; use an underscore instead. The audio or video editing software should automatically add the appropriate file extension indicating what kind of computer file you are

creating, e.g. an audio file (normally .wav) or a video file, as the case may be. The result should look like the following (assuming you are working with an audio recording):

- your Transcription Excerpt recording file: yourname.wav
- your Research Segment recording file: yourname\_segment.wav
- your Complete recording file (optional): yourname\_complete.wav

The following conventions apply to the filenames for your text-based *transcription files* (i.e. word processor files only). (Note that if you are doing your transcriptions using a sound editor like SoundForge, you should keep your file name the same during the whole course.) The filenames for your various transcription and checking files should be on the following pattern:

- your transcription: yourname.doc
- your checking of your partner's transcription: hername\_checkedby\_yourname.doc

To make a 4-minute version of your wav file, the best way to do it is as follows:

- Open the wav file you have been using to mark regions.
- Highlight the relevant 4 minutes, using the mouse.
- Mark a region that will consist of the whole 4 minutes. Be sure the start and end times for the 4 minutes are included in the region name, and add the word "Source" at the beginning of the region name. Click OK to save the region.
- Copy the excerpt to the Windows clipboard, using the Edit menu or control-C.
- Paste the excerpt from the clipboard into a new file, using control-E.

## **15.Exchange**

See the handout on "Transcribing Procedures" for a detailed discussion of how transcriptions are to be exchanged between course participants.

## **References**

- Du Bois, John W. 2004. *Representing Discourse*. MS, University of California, Santa Barbara.
- Du Bois, John W., Cumming, Susanna, Schuetze-Coburn, Stephan, and Paolino, Danae. 1992. Discourse transcription. *Santa Barbara Papers in Linguistics* 4.
- Du Bois, John W., Schuetze-Coburn, Stephan, Cumming, Susanna, and Paolino, Danae. 1993. Outline of discourse transcription. In *Talking data: Transcription and coding in discourse research*, eds. Jane A. Edwards and Martin D. Lampert, 45-89. Hillsdale, NJ: Erlbaum.

[Rev. 11-Oct-2005]