

## Preparing a Transcription for Computer Analysis

### ***Objective***

The objective of this document is to describe how to prepare a transcription file so that you can work on it using various computer software tools, such as programs for making concordances or aligning transcriptions with sound. This document assumes that you have already made a transcription of some spoken discourse data, and that you now want to analyze it using a computer tool of some kind.

### ***Preparing your transcription***

In order to prepare for working with a transcription using computational tools, it is often necessary to first do some further preparation or checking of the transcription file you have made. This is needed to make sure that your transcription is in a format suitable for analyzing with a concordance program or other similar software (e.g. MonoConc Pro, Wordsmith, etc.). This kind of preparation may require some “cleaning up” of your transcription file, to make sure there is nothing in it that will confuse the concordance software (or other computer tool).

### ***Transcription vs. Coding***

One way to prepare a transcription for computer analysis is to remove any features that are not transcription per se. In general, transcription (as opposed to coding) properly consists primarily of symbols corresponding to actual, audible or visible events which took place in the conversation or other speech event you are transcribing. In other words, the transcription contains mostly words spoken, vocal noises, gestures, actions by participants, salient environmental events, and so on. In contrast, background information such as a description of the setting, a title, or extensive commentary would normally not be considered to be transcription as such. Of course, this information is very important and should accompany the transcription, in some form, such as a linked database, or comment lines which can be included or removed as necessary for different analytical purposes. But it should be distinguished from the transcription per se, lest computer analysis of this background information produce a nonsensical result.

Nor does transcription per se consist of coding or analytical information such as part-of-speech tagging, syntactic bracketing, and so on. (Of course there are other valid approaches to computational analysis of discourse data, in which the insertion of lots of analytical codes into a transcription file is common practice.)

For doing the kind of file preparation we are aiming for in this session (e.g. for concordance-making), you will normally want to have a more or less “straight transcription”, that is, a transcription which is not encumbered by lots of coding symbols or analytical notations. Note that brief comments embedded in the transcription are usually not a problem. For example, a comment that is included along with some words spoken as part of a single intonation unit, is not something that needs to be avoided or edited out.

In sum, transcription means representing what you hear and see on your recording. The preparation we are doing here mainly involves stripping away everything that is not true transcription.

## Procedures

1. If you have not already done so, make sure that all your one-minute increments of transcription are consolidated in a single computer file.
2. It is not so important that this file should include all your minutes of transcription. What is important is that you should do the preparation on whatever transcription data you have ready.
3. Make a copy of your transcription file first. You will be making a few changes to this file in preparation for analyzing it, so it is wise to preserve your original file separately.
4. Remove any lines containing only header information, like the title of the transcription, name of the transcriber, and so on.
5. Remove any lines of text containing nothing but a comment, if possible. Short comments (e.g. those written in double parentheses, etc.) are okay, when they are embedded in (or at the end of) a line with an intonation unit containing words spoken.
6. Remove any lines which consist exclusively of interlinear glosses, or other analytical coding.
7. Remove all blank lines. There should be no blank lines (lines containing no characters) at all in your transcription file. Don't use double spacing. Check the beginning and end of your file to eliminate blank lines there, too.
8. If you have special characters (e.g. Unicode characters) in your transcription, you may need to experiment a bit to find which format best preserves the work you have done. (While the future of corpus and linguistic research may be a Unicode future, the reality is that the future is not quite here yet, at least for some current software. (For example, the current version of the concordance program MonoConc Pro [version 2.2] unfortunately does not handle Unicode characters.) Use the search-and-replace function to replace the IU truncation sign (em dash) with two hyphens (an alternate notation for IU truncation). Other Unicode characters (e.g. mostly the tone and arrow symbols used for intonation transcription) should be replaced with alternate notations, or simply deleted. Note that when you save your file in a plain text format like ANSI or ASCII (see below), any remaining Unicode characters in your file will be automatically replaced by some arbitrary character like a question mark, or a square box. It is better to make a more rational disposition of the Unicode characters yourself, substituting something meaningful for them.
9. When you have finished preparing your file according to the above steps, save it, using your usual format (e.g. if you are using OpenOffice.org Writer, this would be the **.odt** format; if you are using Microsoft Word, this would be the **.doc** format).
10. Now save your file again, this time in a plain text format, such as ANSI or ASCII. This format is required by many concordance programs and other analysis software. This version of the transcription will be the one you will use to make your concordances or other analysis.

**OpenOffice.org.** If you are using OpenOffice.org Writer, save your file as a plain text, as follows: From the main menu select “**File/Save as**”. When the dialog box comes up, look at the very bottom where it says “**Save as type**.” Scroll down in this box until you find “**Text (.txt)**”, and select it. Click **Save**. Your file will be saved with the file extension **.txt**.

**Microsoft Word.** To save your file in plain text format using Microsoft Word, select **File** from the menu, then **Save as**. When the dialog box comes up, look at the very bottom where it says

“**Save as type.**” Scroll down in this box until you find the “**Text with Layout (\*.ans)**” format, and select it. Then click on **Save**. Your file will be saved with the file extension **.ans**. (Alternative file formats that can be used are “MS-DOS Text with Layout (\*.asc)” or “Plain Text (\*.txt)”.)

11. Give your file a name that makes clear what is in it. Use the naming conventions specified in the Appendix on “Transcription Format” under the heading “Filenames.”
12. Now you have a file that you can use with a concordance program, to do research on questions of discourse and grammar.
13. Preparing your file in this way shouldn't take long, normally less than an hour.
14. Copy your prepared file onto the appropriate directory in the Linguistics Lab network. Be sure to do this **BEFORE** the class session in which we will be using the files.
15. Note that any files which are not prepared according to the above guidelines, or which are not copied into the appropriate directory in advance of the class tutorial session, may be unusable or unavailable for use in the concordance tutorial.

*[rev.11-Oct-2005]*