

Toolbox/ELAN conversion exercise

ELDP Training, 14th June 2008

The aim of this exercise is to produce a time-aligned interlinearised transcription of a Cicipu language text using both Toolbox and ELAN, making use of the import and export facility in ELAN.

Preliminary steps

1. Open the **Cicipu.prj** project in the Toolbox directory under **d:\users\ElanToolboxConversion**.
2. Make sure text **saat002.001** is displayed
3. Highlight reference **saat002.001.001** and interlinearise this line *only*. This is important or the import will fail later on.
4. Exit Toolbox.
5. Make a backup copy of **saat002.001.txt** just in case.
6. Using Notepad open the file **Text.typ**. You can find this in **d:\users\ElanToolboxConversion\Toolbox**.
7. Near the top of this file you will see the text **\mkrRecord id**. Change this to **\mkrRecord ref**, save the file and then leave it open in Notepad (this step is needed because each **id** in the Cicipu Toolbox database relates to a different text. In your own text files, if you use a separate **id** for each utterance (rather than for each text) then you can omit this step).

Importing the Toolbox file into ELAN

8. Open ELAN
9. Choose **File->Import->Shoobox File...**
10. Tick the **All markers are Unicode** check box.
11. Select the **Text.typ** file (make sure you use the one *local* to the Cicipu project, as in step 6) in the **Shoobox typ file** box. This tells ELAN how the Toolbox field markers (**\tx**, **\mb** etc...) are related to each other.
12. For the **Shoobox file** select the file **saat002.001.txt**
13. The default block duration is about right for this text so leave it unchanged. For your own texts you should work out the average duration of your Toolbox references in milliseconds (i.e. file duration divided by the number of references in the text). Doing this will carefully will make it quicker to time-align the text later.
14. Press **OK**

Tidying up and linking to the WAV file

15. To make the file look nicer, right-click on **nt@Amos** in the lower pane. Choose **Sort Tiers->Sort by Hierarchy**. If you want to, hide the **nt** and **ph** tiers for each speaker.
16. Link the transcription to the WAV file by selecting **Edit->Linked Files...**
17. Click **Add**, and then select the file **saat002.001.wav**.
18. Press **Apply** and then 'Cancel' to leave the dialog.
19. Note that the wave form now appears above the transcription
20. **VERY IMPORTANT!!! Select Options->Propagate Time Changes->Bulldozer Mode. If you don't do this you will lose data when aligning transcriptions.**
21. Align the transcriptions with the WAV file (if you have not used ELAN yet you will probably need to ask for help here). Just do a few utterances for now as this is time-

consuming, but in real-life you would probably do the whole file before exporting back to Toolbox. P.S. Don't forget the children!

22. Save the ELAN file in the ELANToolboxConversion directory – choose **Save As...**, and call the file **saat002.001.eaf**.

Exporting the time-aligned ELAN file back to Toolbox

23. Choose **File->Export As->Shoebox File**.
24. Make sure you are happy with the order of the tiers (you may want to move the nt and ph tiers below the more standard Toolbox markers tx, mb, etc... if they are not already there)
25. Uncheck **Wrap blocks**.
26. Make sure the **Encode all markers in Unicode** check box is ticked.
27. Check the correct TYP file is selected.
28. Click OK
29. Choose the same file you imported into ELAN (**saat002.001.txt**), and overwrite it.
30. Exit ELAN

Opening the (now) time-aligned Toolbox file in Toolbox

31. Before opening Toolbox, open **saat002.001.txt** in Notepad.
Immediately before the text **\ref saat002.001.001** line add the text
\id saat002.001, on a new line on its own, save, and exit Notepad. This step needs a bit of care. The top of your file should look something like this:

```
\_sh v3.0 400 Text
\_DateStampHasFourDigitYear

\id saat002.001
\ref saat002.001.001
\ELANBegin 9.800
\ELANEnd 10.800
\ELANParticipant Amos
\tx mísòní mísòní
...
```

32. Open Toolbox. You should see your file with the time alignment added.
33. Do a bit more interlinearisation (e.g. three or four lines). There will be a few ambiguity selection boxes – it really doesn't matter which you choose here – this is not a test of your Cicipu!
34. Exit Toolbox
35. Resave the **Text.typ** file that you should have left open from step 7 in Notepad.

Finally, repeat steps 8-12 to reimport the file into ELAN and you should see a time-aligned interlinear transcription. Congratulations!

- Remember this is only one possible 'workflow'. You may do all your interlinear transcriptions in Toolbox first, then convert to ELAN at the end. Conversely you may transcribe directly into ELAN first, and then do the interlinearisation at the end.

Questions to think about:

- Does the \tx field in ELAN look how you would expect?
- For your own documentation project, would you expect to do this process only once or multiple times per text?