ELAN → Toolbox → ELAN WORKFLOW

Simon Devylder & Lucie Laporte-Devylder May 1ˢᵗ 2020

This document walks you through the steps needed when you want to collect your data for several speakers in ELAN, then gloss it with Toolbox, and transfer the glossed data back into ELAN without losing any time-encoded information.
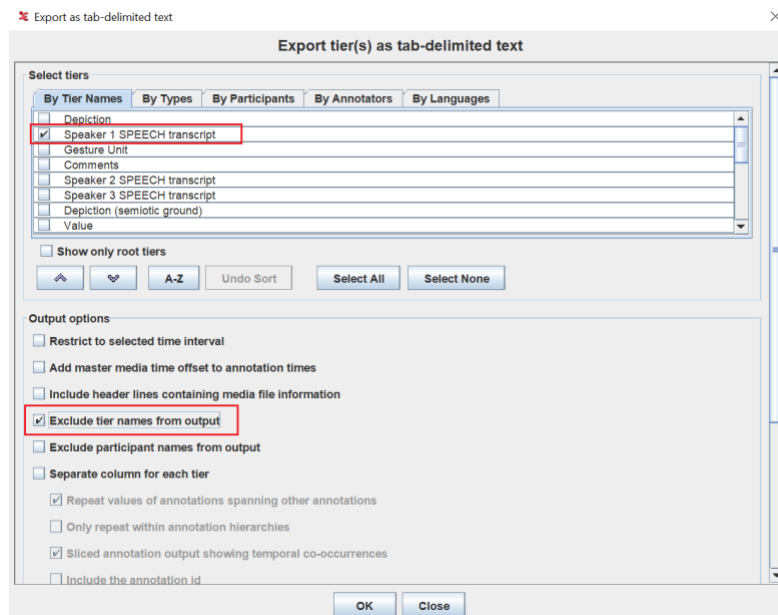
1) Open file in ELAN [e.g. *SAND06.eaf*]

→ click "File" > "Export as" > "Tab-delimited Text"
→ select all tiers where the transcripts are recorded, for every speaker: "SPEAKER 1 SPEECH transcript", "SPEAKER 2 SPEECH transcript", "SPEAKER 3 SPEECH transcript", etc.
→ make sure the box "Exclude tier name from output" is **not** ticked
→ save under the name *SAND06_transcript.txt*

2) Open *SAND06_transcript.txt* in Notepad

→ click "Edit" > "Replace"
- in "Find what", enter "Speaker 1 SPEECH transcript" (i.e the name of the tier for Speaker 1)
- in "Replace by", enter "# " (i.e the # symbol plus a space)
  → click "Replace all"
- repeat the operation, using a different symbol for every speaker
  eg. "Speaker 1 SPEECH transcript" replaced by "# "
       "Speaker 2 SPEECH transcript" replaced by "+ "
       "Speaker 3 SPEECH transcript" replaced by "* "
  etc.
- delete the added symbols for every first line (Speaker 1, line 1 ; Speaker 2, line 1 ; Speaker 3, line 1 ; etc.)

Your document should now look like this:

BEFORE   **AFTER**

3) Open Toolbox

→ click "File" > "New"
       Create a new file and call it SAND06
       Select "Database Type" : "Text"

→ next to \id marker, add "SAND06" (without the quotation marks)
→ copy and paste all the lines from Speaker 1 (the lines that now have a # symbol at their start) from the document SAND06_transcript.txt opened in Notepad in front of the \tx marker

→ click "Tools" > "Break/Number text"
- in "Text-break punctuation", enter "#"
- click "use this name" and enter "SAND06_SP1"

Simon Devylder & Lucie Laporte-Devylder May 1st 2020
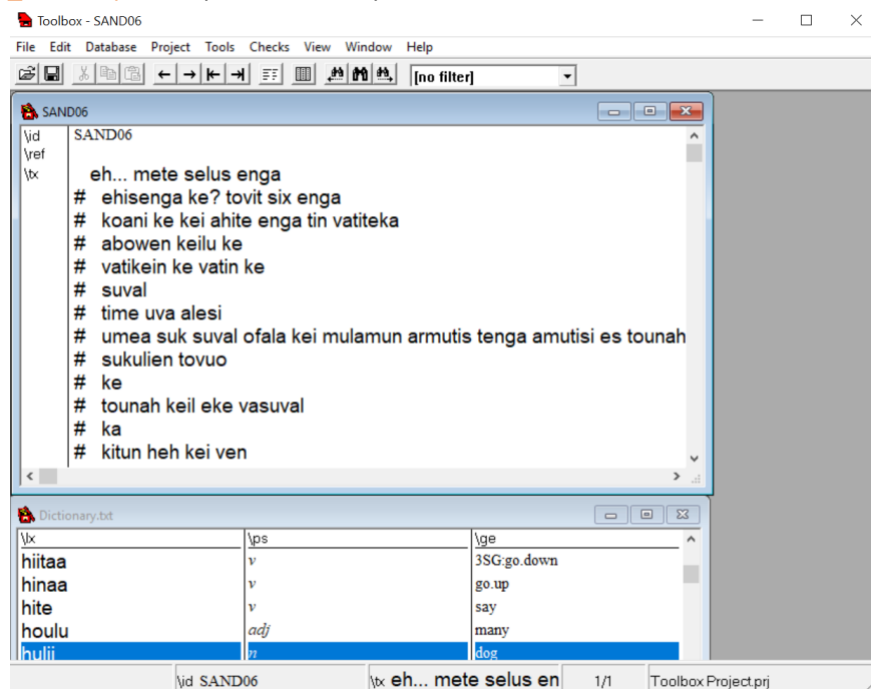
- in "starting number" check that "001" is entered
  → click OK

→ click "Database" > "Insert Record" > name it SAND06 SP2
→ copy and paste all the lines from Speaker 2 (the lines that now have a + symbol at their start) from the document *SAND06_transcript.txt* opened in Notepad in front of the \tx marker
→ click "Tools" > "Break/Number text"

- in "Text-break punctuation", enter "+"
- click "use this name" and enter "SAND06_SP2"
- in "starting number" check that "001" is entered
  → click OK

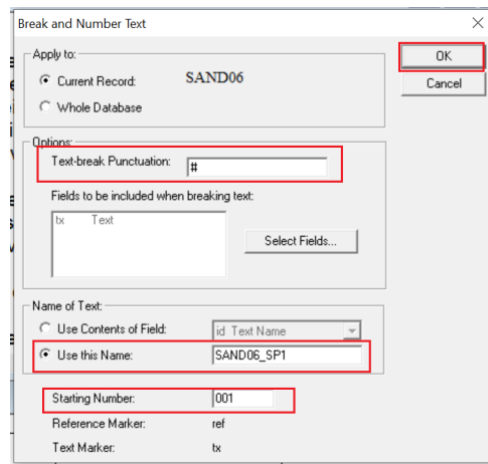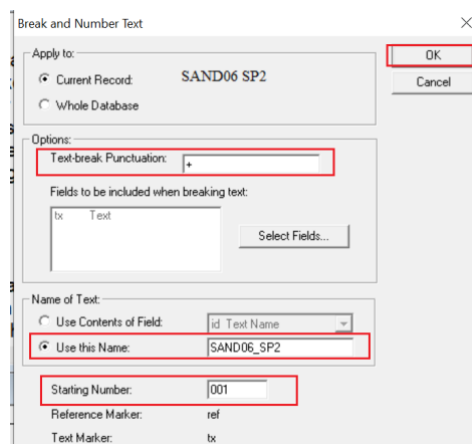→ repeat as many times as necessary, to eventually have 1 sheet per speaker (eg. If your file has 4 speakers you should have 4 sheet: SAND06 ; SAND06 SP2 ; SAND06 SP3 ; SAND06 SP4)
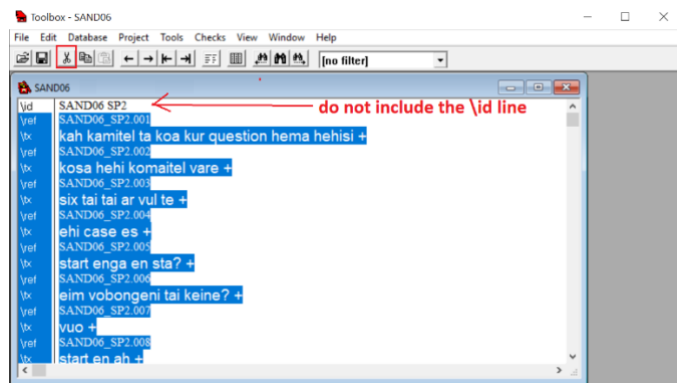Nb. Speaker 1 should have \ref markers that should look like this: SAND06_SP1.001
Nb. Speaker 2 should have \ref markers that should look like this: SAND06_SP2.001
Nb. Speaker 3 should have \ref markers that should look like this: SAND06_SP3.001
etc.

→ cut and paste the text with their \ref form every speaker into the first sheet where speaker 1 is (i.e the sheet with the \id "SAND06")

→ keep only the sheet with the \id "SAND06", delete every other sheet (i.e any sheet with an \id that include a speaker name "SAND06 SP2" , "SAND06 SP3", etc.). To do so, click on "Database" > "Delete record", once you are on the sheet you want to delete

Your Toolbox file should now look something like:
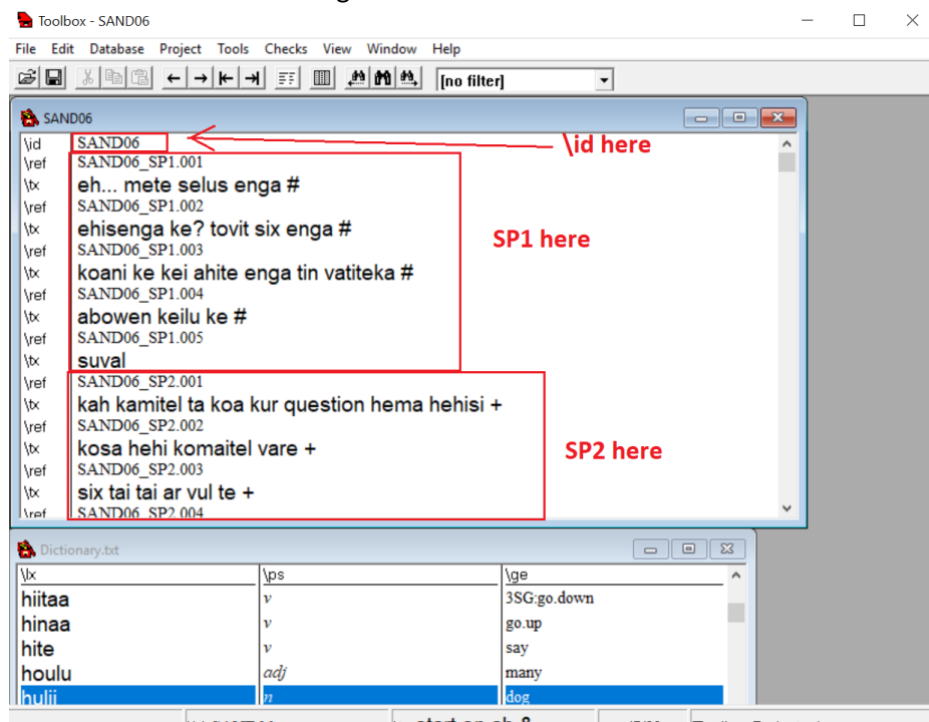
→ click "File" > "Save" (i.e leave Toolbox open with the SAND06 file you just created)

4) Open again file in ELAN [SAND06.eaf]

→ create a copy of this file in your working folder
        → click "Save as" > name it "SAND06_toolexport"

→ in ELAN, under the Annotation mode, right click on "SPEAKER 1 SPEECH transcript"
    > select "change attribute of"
        • in "Tier name", replace "SPEAKER 1 SPEECH transcript" by "ref@SP1"
        • in "Participant", add "SP1"
        → click "Change"

> repeat for each speaker

      Speaker 2   → Tier name: ref@SP2

                                → Participant: SP2

      Speaker 3   → Tier name: ref@SP3

                                → Participant: SP3

      etc.

→ click "Tier" > "Label and Number annotation"

- select the tier now named "ref"
- click "include label part" and enter "SAND06"
- click "insert other delimiter" and enter "."
- Click "min number of integer digit" and enter "3"
- Check that the label will look like "SAND06_SP1.001" in the blue frame at the bottom of the box

    → click "OK" > "Close"

> repeat for every speaker, simply change the number in the case "include label part" : SAND06_SP2

Simon Devylder & Lucie Laporte-Devylder May 1st 2020

Generate label and number for annotations

Generate label and number for annotations

Select tier(s)
- ● Single tier
- ○ Multiple tiers

☐ Gesture Unit
☐ Comments
☑ ref@SP2
☐ ref@SP3
☐ Depiction (semiotic ground)
☐ Value

Options
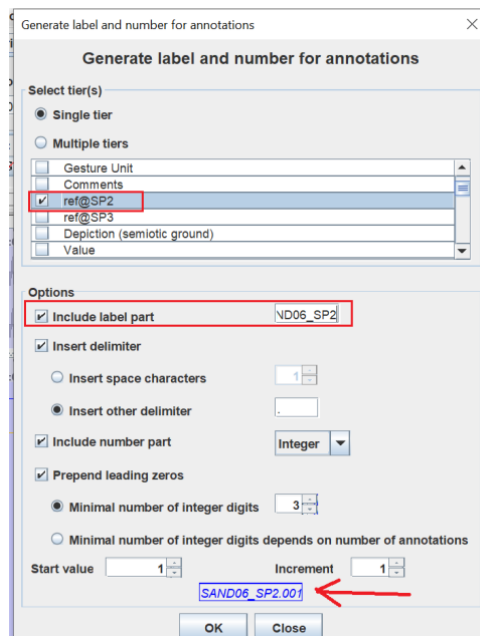☑ Include label part   ND06_SP2
☑ Insert delimiter
  ○ Insert space characters   1
  ● Insert other delimiter   .
☑ Include number part   Integer
☑ Prepend leading zeros
  ● Minimal number of integer digits   3
  ○ Minimal number of integer digits depends on number of annotations
Start value   1   Increment   1
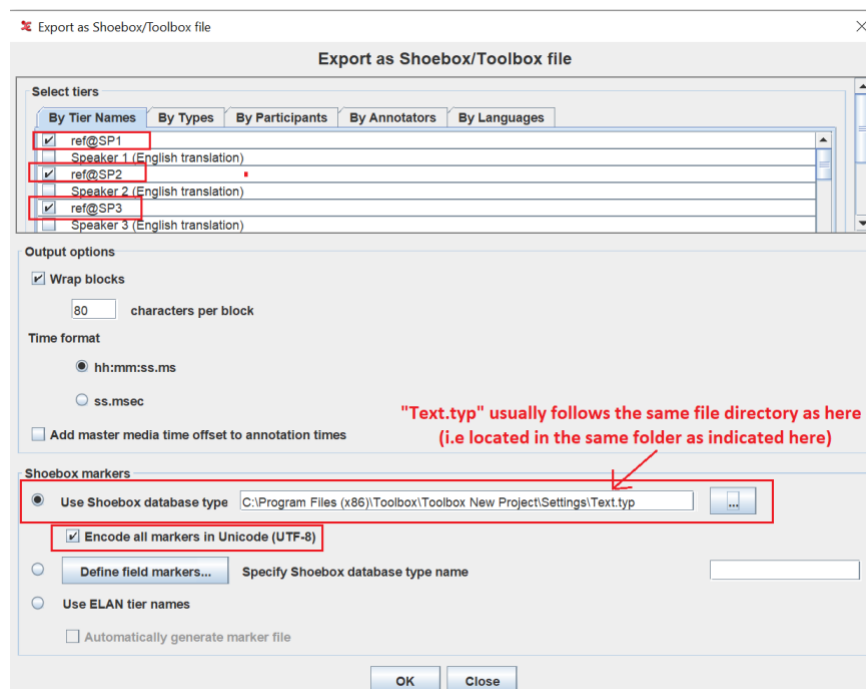
SAND06_SP2.001

OK   Close

→ click "File" > "Export" > "as Shoebox file"
- • in "Select tier" select all speskers: "ref@SP1", "ref@SP2", "ref@SP3", "ref@SP4", etc.
- • in "use Shoebox database type" select the database type "Text.typ"

[n.b. If you are using Windows, this file "Text.typ" is usually located in the folder where Toolbox is installed, in the folder called "Toolbox New Project" > "Setting" > "Text.typ"]

- • make sure the box "Encode all markers in Unicode (UTF-8)" is ticked
- → click "OK"

Export as Shoebox/Toolbox file

Export as Shoebox/Toolbox file

Select tiers

| By Tier Names | By Types | By Participants | By Annotators | By Languages |

☑ ref@SP1
☐ Speaker 1 (English translation)
☑ ref@SP2
☐ Speaker 2 (English translation)
☑ ref@SP3
☐ Speaker 3 (English translation)

Output options
☑ Wrap blocks
    80   characters per block

Time format
  ● hh:mm:ss.ms
  ○ ss.msec
☐ Add master media time offset to annotation times

"Text.typ" usually follows the same file directory as here (i.e located in the same folder as indicated here)

Shoebox markers
● Use Shoebox database type   C:\Program Files (x86)\Toolbox\Toolbox New Project\Settings\Text.typ   ...
  ☑ Encode all markers in Unicode (UTF-8)
○   Define field markers...   Specify Shoebox database type name
○   Use ELAN tier names
    ☐ Automatically generate marker file

OK   Close

→ save under the name *SAND06_elanexport.txt* (/!\ Make sure you select the file format ".txt")

5) Open *SAND06_elanexport.txt* in Notepad

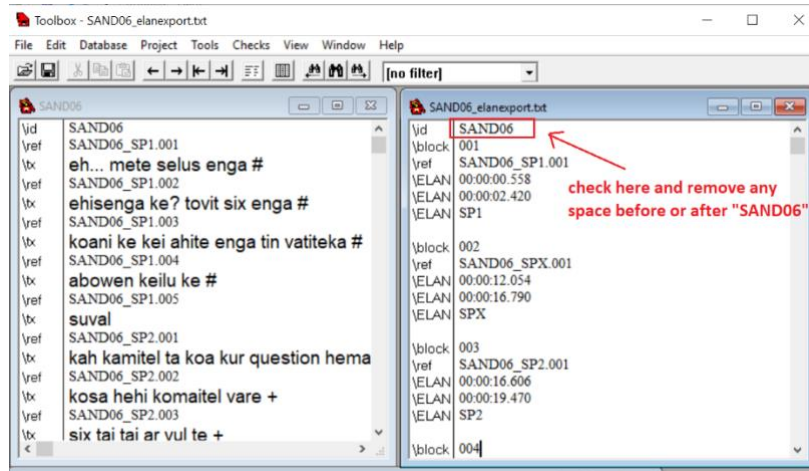→ add manually "\id SAND06" above the line "\block 001"
(/!\ Make sure the is no space after "\id SAND06" (* "\id SAND06 "))
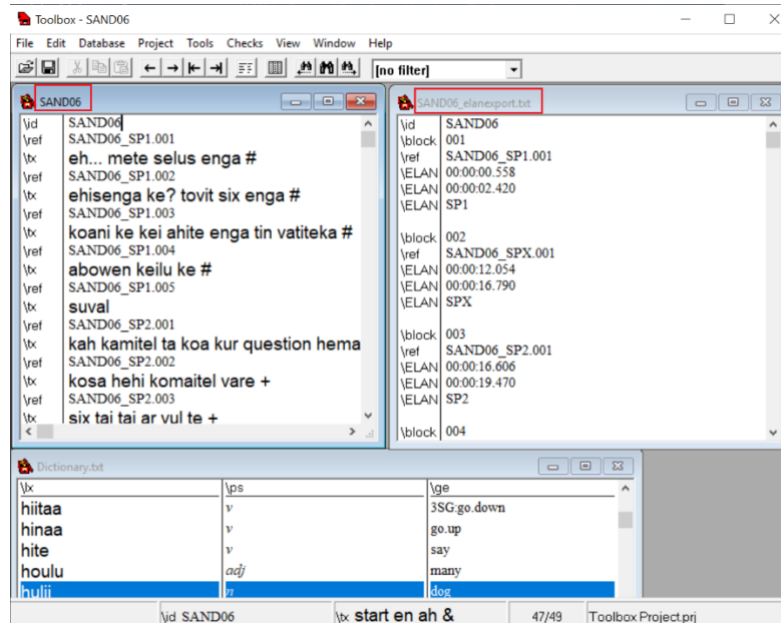
6) Open *SAND06_elanexport.txt* in Toolbox

→ click "File" > "Open" > *SAND06_elanexport.txt*

(/!\ Make sure here again that Toolbox did not add any space after "\id SAND06" (* "\id SAND06 "))

(/!\ Make sure also that there isn't any space in the \ref marker (e.g."\ref SAND06.001" ~~(not* "\ref SAND06.001")~~)

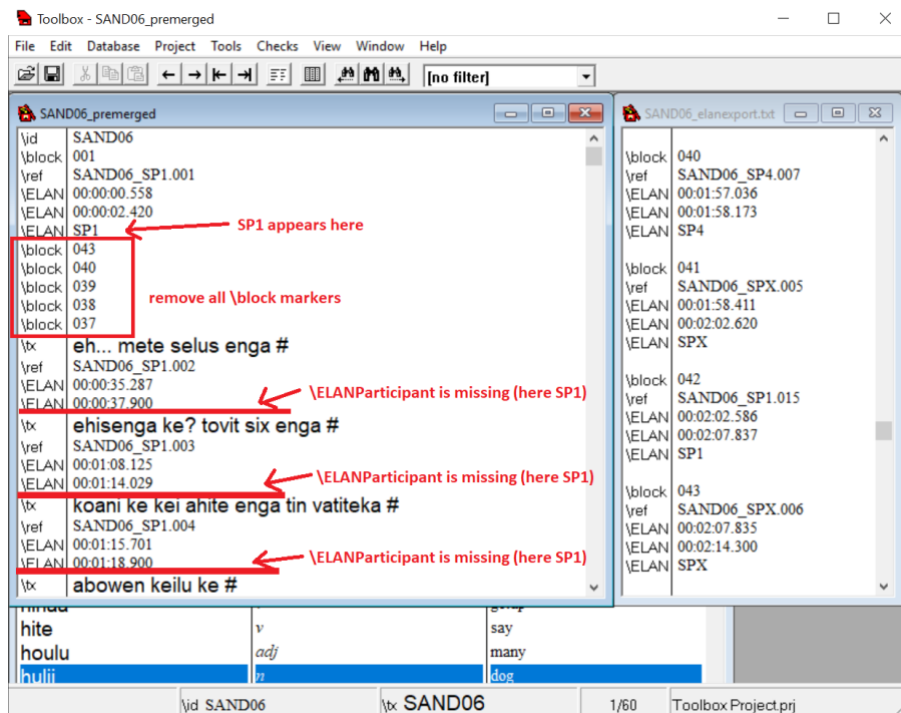You should now have in Toolbox :

→ click on the window of "SAND06" that you created at step 3 (on the left in the picture above)
→ create a copy of it, click "File" > "Save as" > save under the name *SAND06_premerged.txt*

→ click "Database" > "Merge database" > select "SAND06_elanexport" > "OK"

You should now have in Toolbox :

→ close the window of the file "SAND06_elanexport.txt" (on the right in the picture above)

→ in the window of the file "SAND06_premerged"

- delete all "\block" markers (such as those framed in red in the picture above)
- add manually the marker \ELANParticipant with the relevant participant ("SP1", "SP2", "SP3", etc.) everywhere it is missing
- for clarity, move all \ELANMedia markers at the bottom of the file

Your Toolbox file is now ready to be interlinearized, and should look like this:
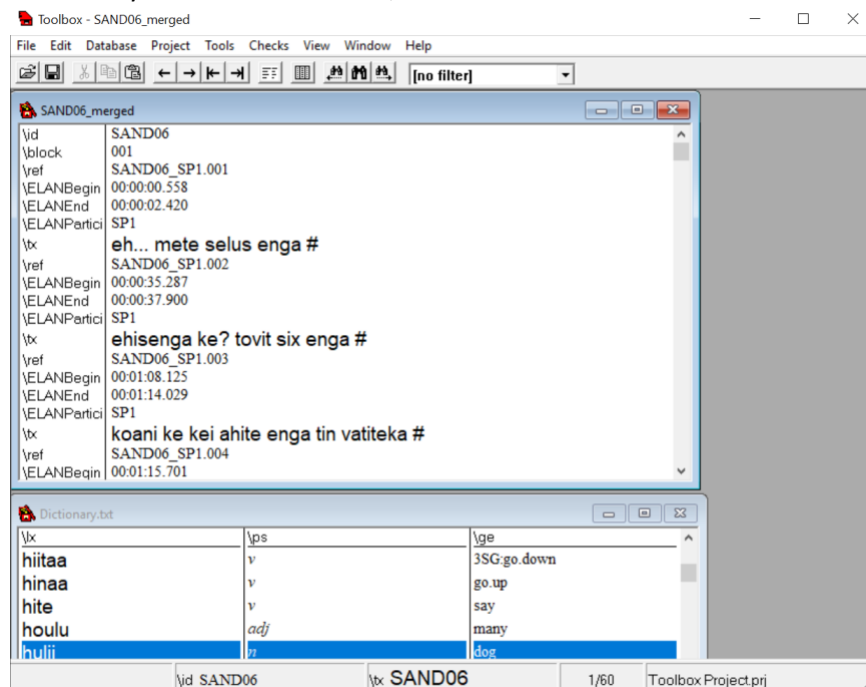
→ click "File" > "Save as" > save under the name *SAND06_merged.txt*

Simon Devylder & Lucie Laporte-Devylder May 1st 2020

<u>7) Gloss your transcript in Toolbox</u>

→ add the "free translation" (marker "\ft") after each transcription line
→ make sure to save regularly

<u>8) Copy and paste into your working folder the file "Text.typ" from the folder Toolbox > Toolbox New Project > Settings</u>

→ open "Text.typ" in Notepad
- replace the line "\mKrRecord id" by "\mKrRecord ref"
→ save under the name "TextELAN"

<u>9) Open ELAN</u>

→ click "Import" > "Shoebox file"
- in "Shoebox file" select the file SAND06_merged.txt
- in "Shoebox database type" select the modified file from step 7: "TextELAN.typ"
- make sure "All marker are Unicode" is ticked
  → OK
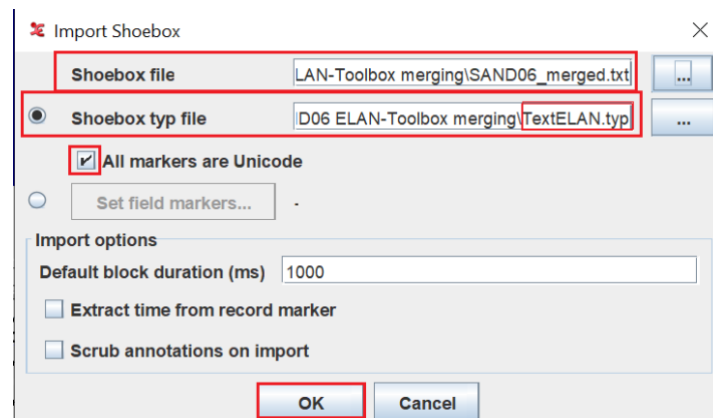
Example of ELAN file after you have imported the glossed transcript from Toolbox:

GOOD LUCK ! ☺

Simon Devylder & Lucie Laporte-Devylder May 1st 2020