Annotation tool Toolbox

how to gloss/annotate in Toolbox

Language Documentation

SEBASTIAN DRUDE

2012-07

Topics

1. Data and Annotation (Theory)
2. Annotation Tools (Overview and Comparison)
3. Intro to Interlinearization (not time-aligned)

   Excurse: Text- vs. sentence-based databases
4. Time-aligned Annotation
   1. ELAN-Generated Annotation
   2. Play chunks of an audio file in Toolbox

      Excurse: Regular Expressions

      Excurse: UNICODE and UTF-8
   3. Transcriber-Generated Annotation; Conversions
   4. Round-trip Configuration ELAN--Toolbox

Topics

1. **Data and Annotation (Theory)**
2. Annotation Tools (Overview and Comparison)
3. Intro to Interlinearization (not time-aligned)
   Excurse: Text- vs. sentence-based databases
4. Time-aligned Annotation
   1. ELAN-Generated Annotation
   2. Play chunks of an audio file in Toolbox
      Excurse: Regular Expressions
      Excurse: UNICODE and UTF-8
   3. Transcriber-Generated Annotation; Conversions
   4. Round-trip Configuration ELAN--Toolbox

Data and Annotation

Data

Data is always data FOR something, or at least OF something – usually it is ***a systematic representation of*** physical states and events

In linguistics, *primary data* is a direct rendering or result of *communicative (speech) events*, for instance a ***written text*** or, in partiuclar, an ***audio/video recording*** of a speech event

Data and Annotation

Annotation

Annotation of data is ***a symbolic representation*** of properties of the state/event represented in the data

In linguistics, the most common and basic types of annotation are a ***transcription*** and a ***translation*** of the linguistic expressions represented in primary data (e.g., in an audio / video recording)

Data and Annotation

Global vs. unit-oriented Annotation

Global or *holistic annotation* represents properties of the event as a whole and is part of the *metadata*

Unit-oriented annotation refers to specific parts of the data, in particular, utterances of individual sentences or words or sounds etc.

We speak of *individual* ***annotations*** (plural)

Data and Annotation

Secondary and derived data

If unit-oriented annotation is directly based on primary data (such as a written text or a audio or video recording), then it is *secondary data*

Annotation based on secondary data would be tertiary data, and so forth recursively

In sum, all unit-o. annotation is *derived data*

There are other types of derived data (lexicon...)

Data and Annotation

Time-aligned annotation

Annotation of a media file is *time-aligned anotation* if each piece of annotation is explicitly associated with the corresponding *chunk* (***time-span, segment***) of the media file

This is usually done by using the time position of the ***start and end points*** of the respective chunk, the *time marks*

Data and Annotation

Linguistic types of annotations

Annotations differ according to the *types of properties* of the speech event that are represented in the annotation

Annotations can be *phonetic, phonological, morphological, syntactic, semantic, pragmatic*, (possibly others), and on each level they can focus on the ***units***, or on ***structures*** of units, or on ***relations*** that hold among units, etc.

Data and Annotation

Coverage of annotation

Basic annotation: only *transcriptions*, *translations* and perhaps *notes*, on a sentence level

Basic glossing: additionally information on *individual morphs*: a *gloss* (indication of meaning or function) and perhaps a part-of-speech tag

Advanced glossing: one or several of *additional levels*, from phonetic to pragmatic (for instance, a *prosodic transcription*, or annotation of the *syntactic structure*, of *grammatical relations*, etc.)

Advanced Glossing:
a syntactic glossing table

I	1	2	3
II	[di.	ʔʊn̩y.bɐ.zɪç.tʰɪç̊n̩	pʁo.blee.mə]
III	L	H_f M_f L H H H	H H_r L_f
IV	/ ˌdii. /	/ ˈʔun.ˌʔyy.bə.r-ˌziX.t-ˌliχ̟ə.n /	/ pro.ˈblee.mə /
V	L	H L L H H H	H **H_r** L
VI	*die*	*unübersichtlichen*	*probleme*
VII	DefArt	Adj	Sub Neut
VIII	Nom Pl Unm_G Str	Unm_C Pl Unm_G Wk	Nom Pl Unm_D
IX	Nom Pl Def	'involved'	'problem'
X	1,3: Nf 2: Nf **1**,2,**3**: NGr		
XI	mod: 2 1,3		
XII	Die unübersichtlichen Probleme.		
XIII	E: The involved problems.		

Advanced Glossing:
a morphological glossing table

I	1	2	3	4	5
II	/ ʔun.	ʔyy.bə.r-	ziX.t-	liX̣	ə.n /
III	H	H,L L	H,L	H,L	L
IV	/ ˈʔun. /	/ ˈʔyy.bə.r /	/ ˈziX.t /	/ ˌliX. /	/ ə.n /
V	H	H,L L	H,L	H,L	L
VI	*un*	*über*	*sicht*	*lich*	*en*
VII	$Pref_i$	$PrepSt_j$	$SubSt_k$	$SubSt_l/AdjSt_m$	$AdjFlex_n$
VIII	–	–	–	–	–
IX	not	‘over’	‘view’	suitable-for	Unm_C Pl Unm_G Wk
X	1: Af 2: Stf	3: Stf 4: Af	5: Af 2,**3**: Stf	**2,3**,4: Stf	1,**2,3,4**: Stf **1,2,3,4**,5: StGr
XI	m-mod: 2 3	m-qual: 4 2,3 3	m-mod: 1 2,3,4	m-qual: 5 1,2,3,4	1,2,3,4
XII	unübersichtlichen				
XIII	‘involved’				

Topics

1. Data and Annotation (Theory)
2. **Annotation Tools (Overview and Comparison)**
3. Intro to Interlinearization (not time-aligned)

   Excurse: Text- vs. sentence-based databases
4. Time-aligned Annotation
   1. ELAN-Generated Annotation
   2. Play chunks of an audio file in Toolbox

      Excurse: Regular Expressions

      Excurse: UNICODE and UTF-8
   3. Transcriber-Generated Annotation; Conversions
   4. Round-trip Configuration ELAN--Toolbox

Annotation Tools

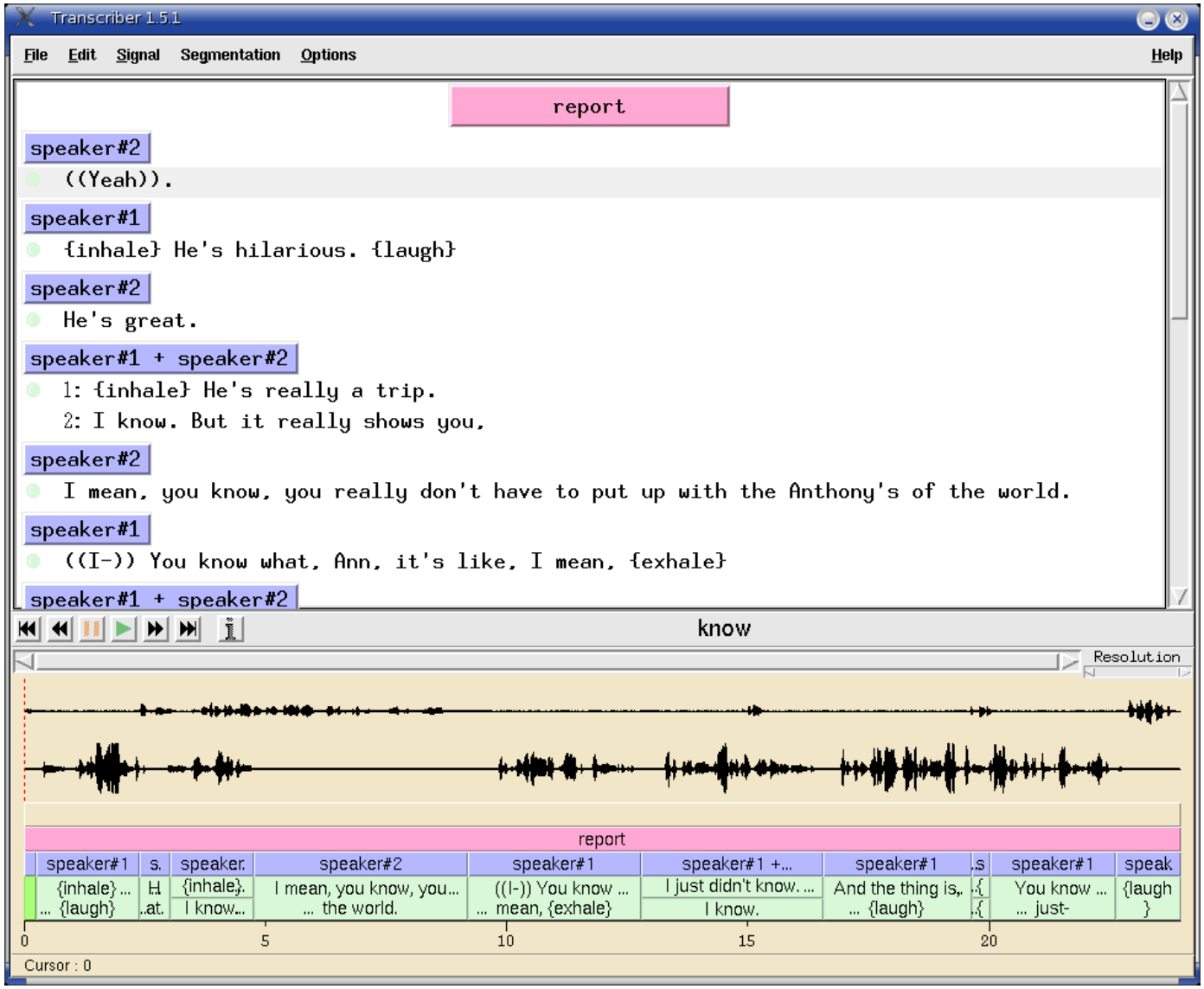
Transcriber

Tool for the segmentation and transcription of audio files

Pros: Compatible with MAC, Windows & Linux; very easy to use; produces simple XML-files

Cons: No Unicode input possible; only one line of annotation; no video; no lexicon (new version not tested)

Transcriber



Annotation Tools

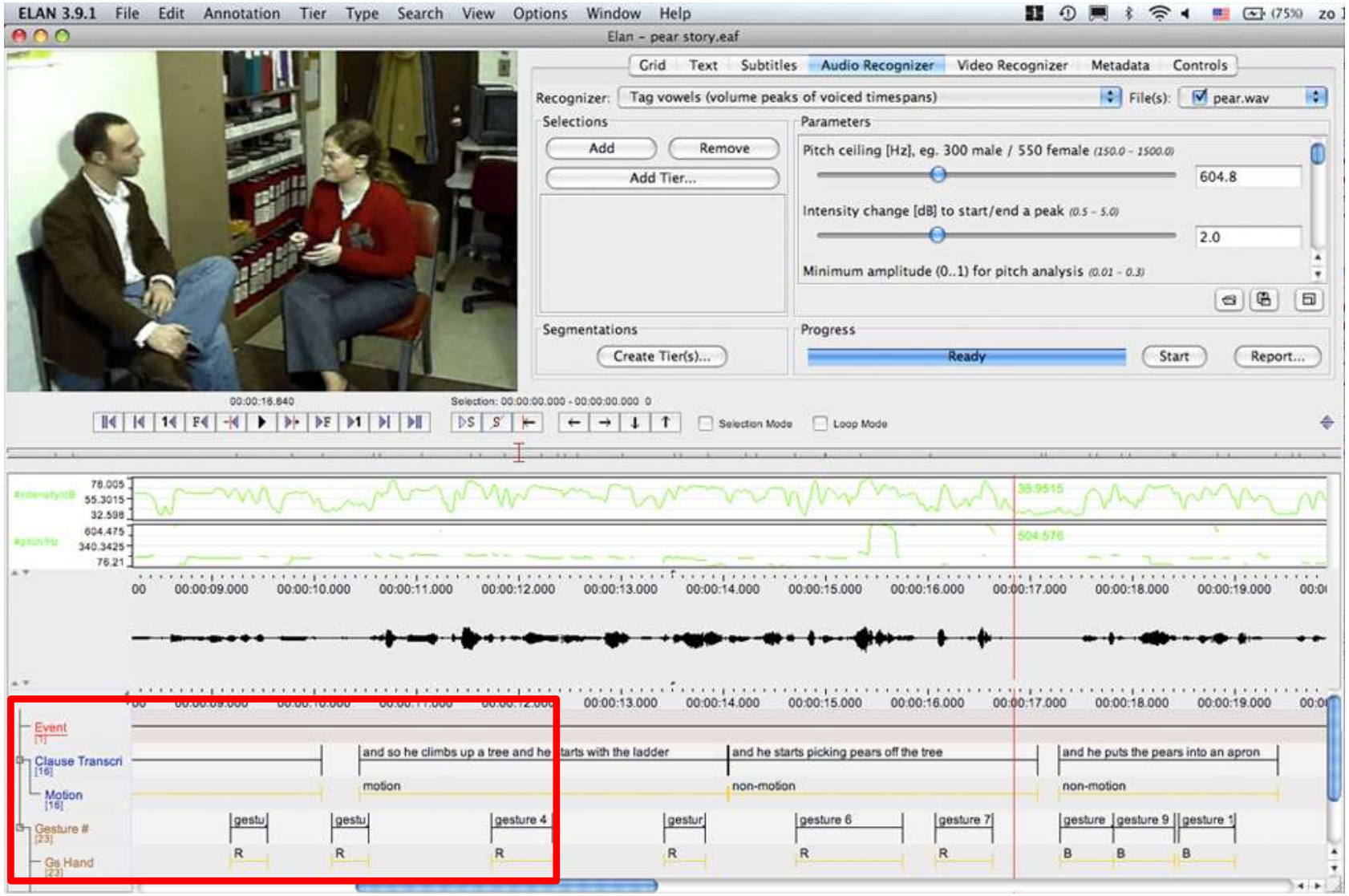
ELAN

Tool for the complex annotation of audio and video files

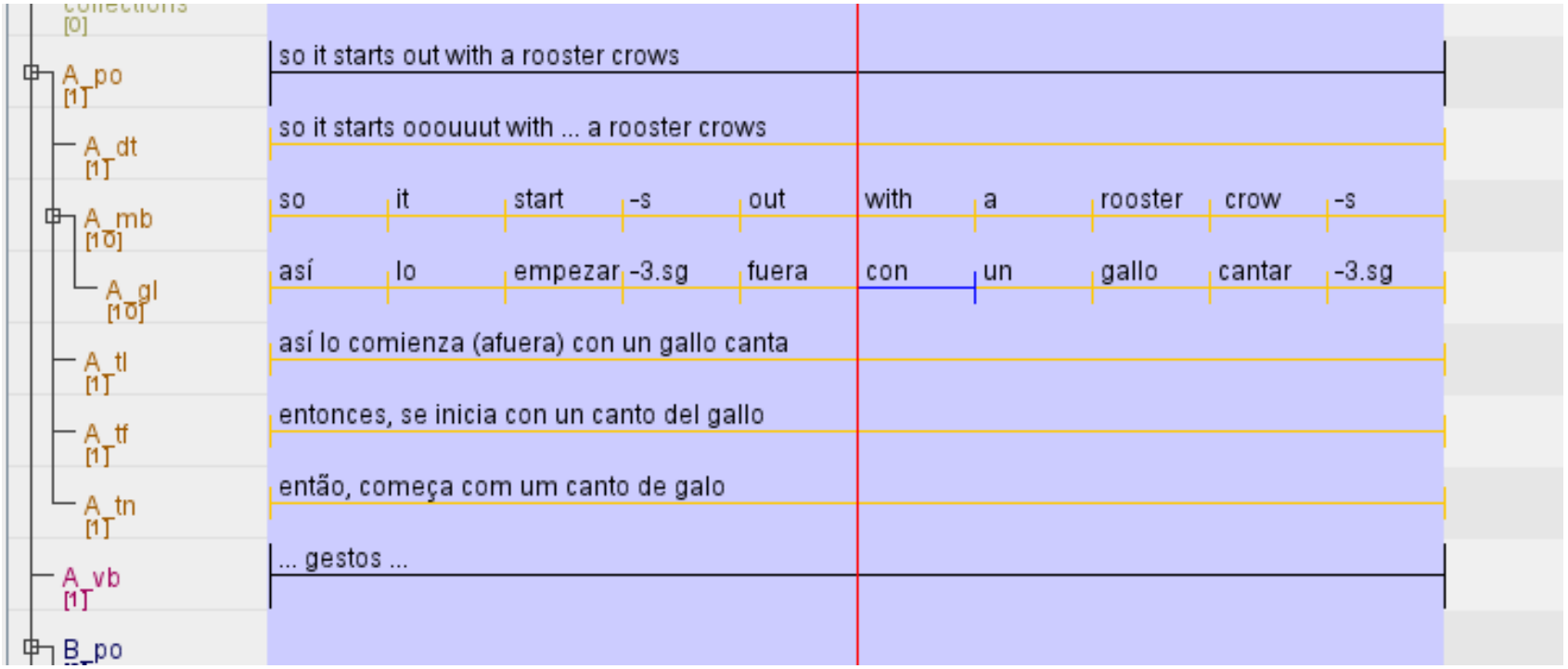
Pros: Compatible with MAC, Windows & Linux; audio and multiple video files; unlimited tiers for different speakers; state-of-the-art; wide user community; XML output (but complex)

Cons: Complex tool for beginners (but now: easier transcription mode); no lexicon (yet)

ELAN


ELAN 3.9.1 File Edit Annotation Tier Type Search View Options Window Help
Elan - pear story.eaf
Grid Text Subtitles Audio Recognizer Video Recognizer Metadata Controls
Recognizer: Tag vowels (volume peaks of voiced timespans)
File(s): pear.wav
Selections
Add Remove
Add Tier...
Parameters
Pitch ceiling [Hz], eg. 300 male / 550 female (150.0 - 1500.0)
604.8
Intensity change [dB] to start/end a peak (0.5 - 5.0)
2.0
Minimum amplitude (0..1) for pitch analysis (0.01 - 0.3)
Segmentations
Create Tier(s)...
Progress
Ready
Start
Report...
00:00:16.840
Selection: 00:00:00.000 - 00:00:00.000
Selection Mode
Loop Mode
78.005
55.3015
32.598
604.475
340.3425
76.21
00:00:09.000
00:00:10.000
00:00:11.000
00:00:12.000
00:00:13.000
00:00:14.000
00:00:15.000
00:00:16.000
00:00:17.000
00:00:18.000
00:00:19.000
Event
Clause Transcri
Motion
Gesture #
Gs Hand
and so he climbs up a tree and he starts with the ladder
and he starts picking pears off the tree
and he puts the pears into an apron
motion
non-motion
non-motion
gestu
gestu
gesture 4
gestur
gesture 6
gesture 7
gesture
gesture 9
gesture 1
R
R
R
R
R
R
B
B
B

ELAN


A_po
[1]
so it starts out with a rooster crows
A_dt
[1]
so it starts ooouuut with ... a rooster crows
A_mb
[10]
so it start -s out with a rooster crow -s
A_gl
[10]
así lo empezar -3.sg fuera con un gallo cantar -3.sg
A_tl
[1]
así lo comienza (afuera) con un gallo canta
A_tf
[1]
entonces, se inicia con un canto del gallo
A_tn
[1]
então, começa com um canto de galo
A_vb
[1]
... gestos ...
B_po

Annotation Tools

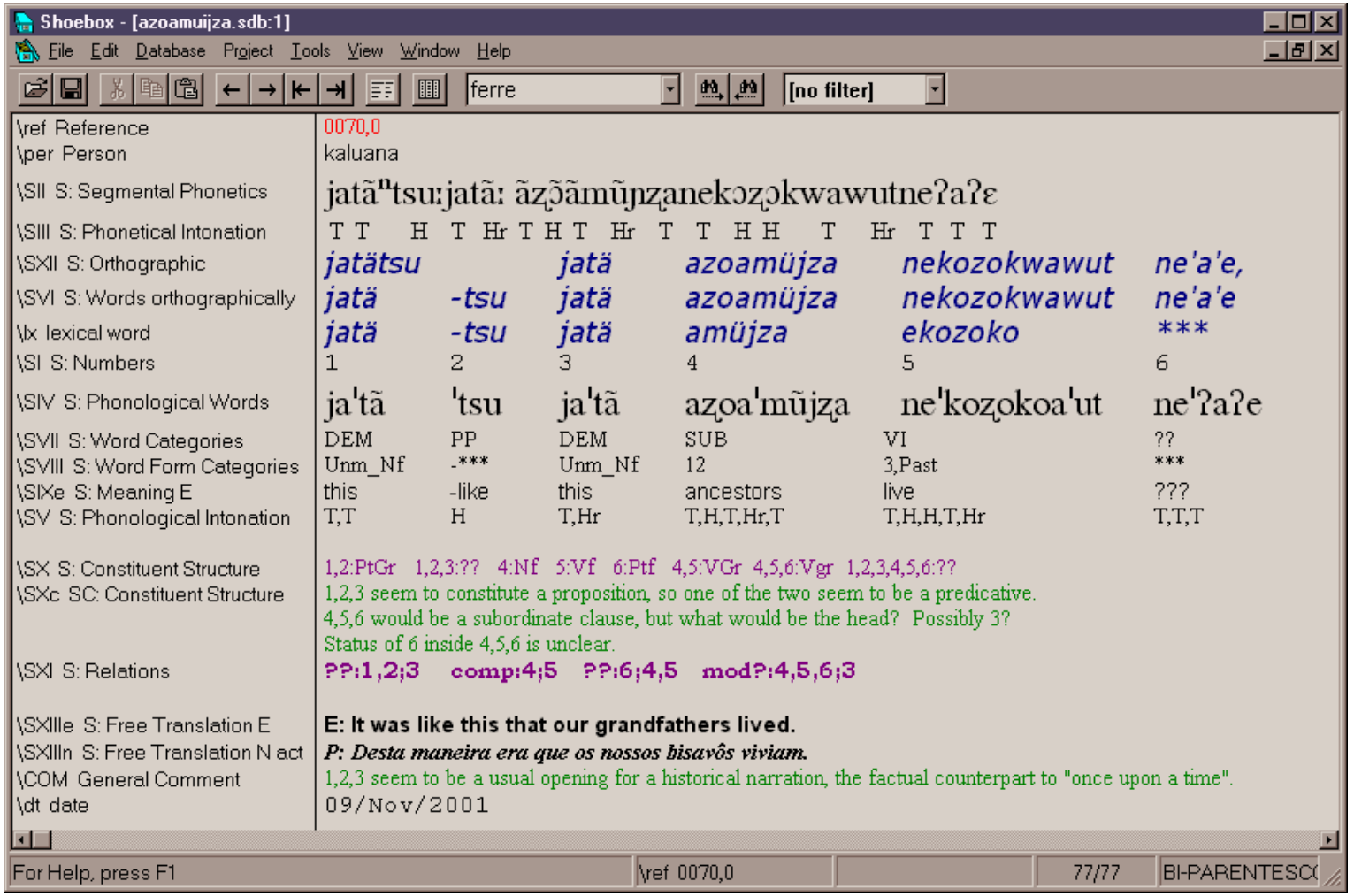
Toolboox

Text-oriented general database tool for linguistic fieldwork with lexicon and texts

Pros: Flexible and powerful; Export to different formats (incl. XML); therefore easy to integrate with other tools; many users

Cons: Too flexible; poor data format “Standard Format”; complex to set up; tricky on MAC/Linux; no video and no time-aligning; at end of life-cycle; produced by SIL

Toolbox


Shoebox - [azoamuijza.sdb:1]
File Edit Database Project Tools View Window Help
ferre
[no filter]
\ref Reference 0070,0
\per Person kaluana
\SII S: Segmental Phonetics jatãⁿtsu:jatã: ãzɔ̃ãmũjnzanekɔzɔkwawutne?a?ɛ
\SIII S: Phonetical Intonation T T H T Hr T H T Hr T T H H T Hr T T T
\SXII S: Orthographic jatätsu jatä azoamüjza nekozokwawut ne'a'e,
\SVI S: Words orthographically jatä -tsu jatä azoamüjza nekozokwawut ne'a'e
\lx lexical word jatä -tsu jatä amüjza ekozoko ***
\SI S: Numbers 1 2 3 4 5 6
\SIV S: Phonological Words ja'tã 'tsu ja'tã azɔa'mũjzɑ ne'kozɔkoa'ut ne'?a?e
\SVII S: Word Categories DEM PP DEM SUB VI ??
\SVIII S: Word Form Categories Unm_Nf -*** Unm_Nf 12 3,Past ***
\SIXe S: Meaning E this -like this ancestors live ???
\SV S: Phonological Intonation T,T H T,Hr T,H,T,Hr,T T,H,H,T,Hr T,T,T
\SX S: Constituent Structure 1,2:PtGr 1,2,3:?? 4:Nf 5:Vf 6:Ptf 4,5:VGr 4,5,6:Vgr 1,2,3,4,5,6:??
\SXc SC: Constituent Structure 1,2,3 seem to constitute a proposition, so one of the two seem to be a predicative.
4,5,6 would be a subordinate clause, but what would be the head? Possibly 3?
Status of 6 inside 4,5,6 is unclear.
\SXI S: Relations PP:1,2;3 comp:4;5 PP:6;4,5 modP:4,5,6;3
\SXIIIe S: Free Translation E E: It was like this that our grandfathers lived.
\SXIIIn S: Free Translation N act P: Desta maneira era que os nossos bisavôs viviam.
\COM General Comment 1,2,3 seem to be a usual opening for a historical narration, the factual counterpart to "once upon a time".
\dt date 09/Nov/2001
For Help, press F1
\ref 0070,0
77/77
BI-PARENTESC

Annotation Tools

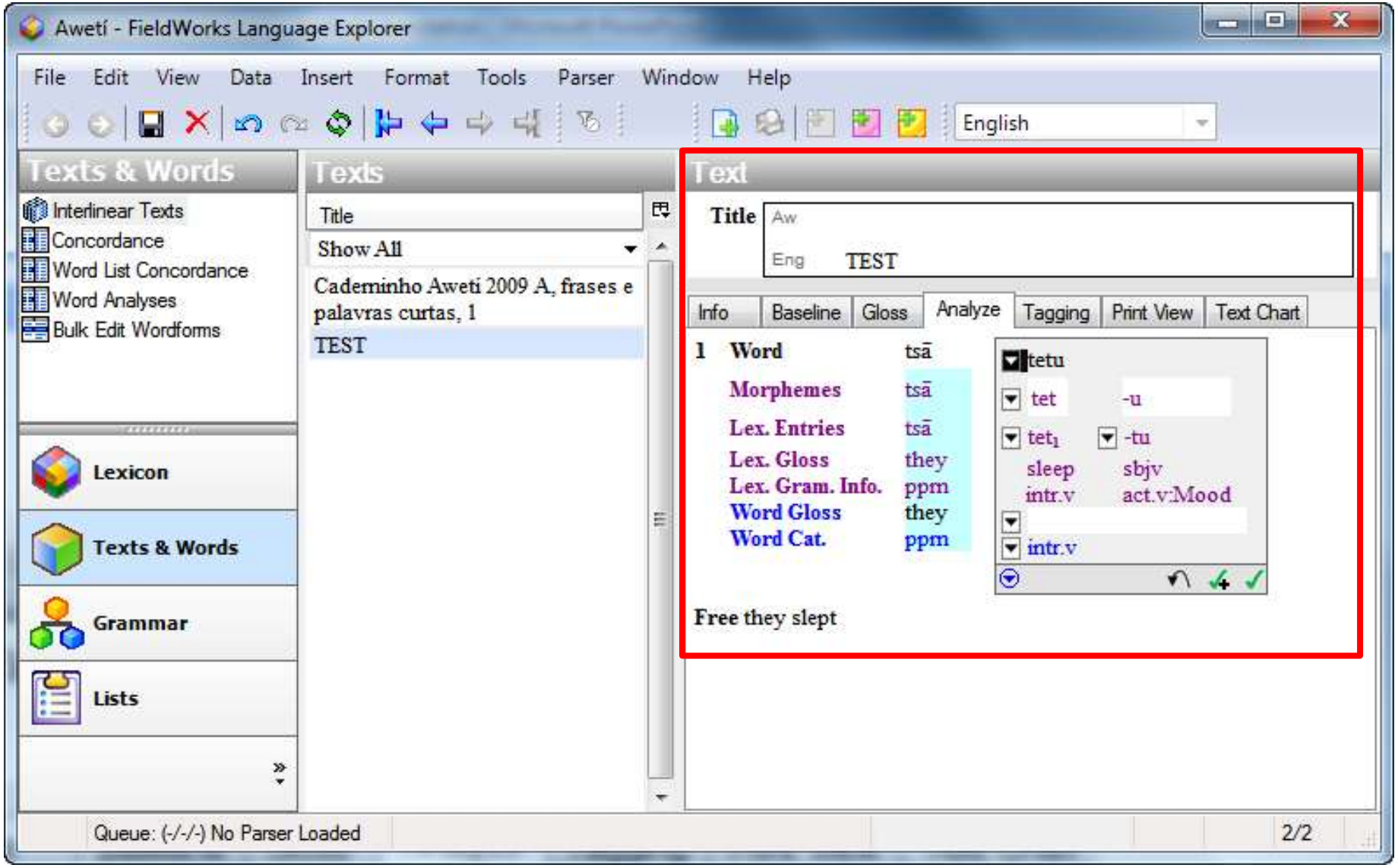
FLEX

Extensive linguistic database tool for linguistic fieldwork with lexicon and texts

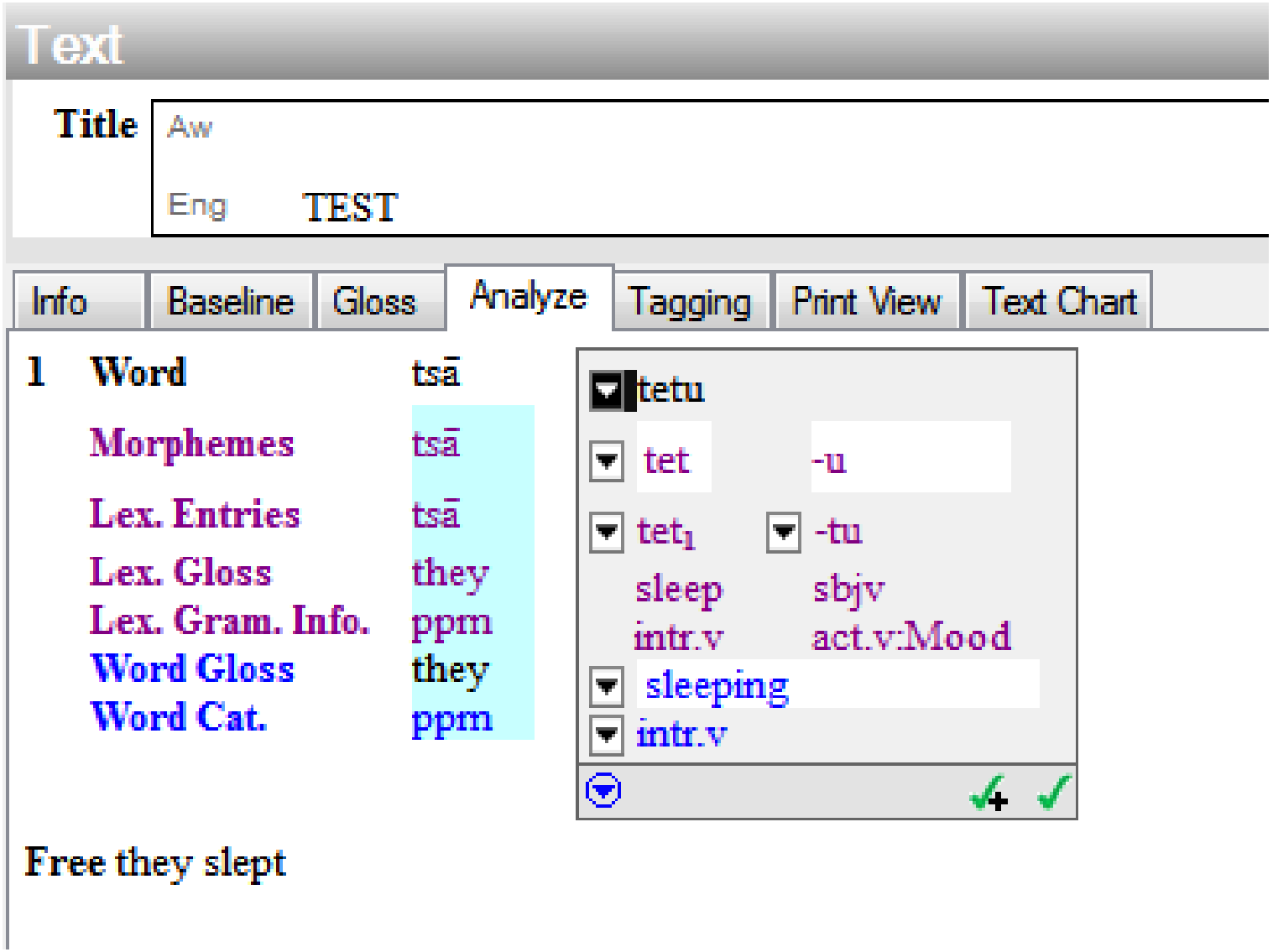
Pros: Powerful and well-designed; inbuilt ontology and analysis tools; growing user community

Cons: Not flexible (8 tiers); one huge XML database with no good import or export function for texts; Windows only; difficult to configure; no audio, no video, no time-alignment; produced by SIL

FLEX


Awetí - FieldWorks Language Explorer
File Edit View Data Insert Format Tools Parser Window Help
English
Texts & Words
Interlinear Texts
Concordance
Word List Concordance
Word Analyses
Bulk Edit Wordforms
Lexicon
Texts & Words
Grammar
Lists
Texts
Title
Show All
Caderninho Aweti 2009 A, frases e palavras curtas, 1
TEST
Text
Title
Aw
Eng TEST
Info Baseline Gloss Analyze Tagging Print View Text Chart
1 Word tsā
Morphemes tsā
Lex. Entries tsā
Lex. Gloss they
Lex. Gram. Info. pprn
Word Gloss they
Word Cat. pprn
tetu
tet -u
tet1 -tu
sleep sbjv
intr.v act.v:Mood
intr.v
Free they slept
Queue: (-/-/-) No Parser Loaded
2/2

FLEX


Text
Title
Aw
Eng
TEST
Info
Baseline
Gloss
Analyze
Tagging
Print View
Text Chart
1
Word
tsā
Morphemes
tsā
Lex. Entries
tsā
Lex. Gloss
they
Lex. Gram. Info.
ppm
Word Gloss
they
Word Cat.
ppm
tetu
tet
-u
tet1
-tu
sleep
sbjv
intr.v
act.v:Mood
sleeping
intr.v
Free they slept

Annotation Tools

Other tools

Praat for segmenting, best for phonetic annotation.

CLAN does audio and video annotation, in the CHAT or CA (Conversation Analysis) formats, for child language data (CHILDES project).

ANVIL seems to be similar to ELAN (not tested).

The *EXMARaLDA Partitur-Editor* (U. Hamburg) is widely used for discourse analysis.

Audiamus and *Eopas* (N. Thieberger) organize (not create) annotation. ***There are several others.***

Annotation Tools

	Transcriber	ELAN	Toolbox	FLEX
Complexity	Easy	Complex, w. easier modes	Complex to configure	Complex
Audio	Yes	Yes	No (can play)	No
Video	No	Yes	No	No
Tiers	1 per speaker	Unlimited	Unlimited	Fixed: 8
Lexicon interop., automatic glossing	No	No (is planned)	Yes	Yes
Unicode	No input	Yes	Yes	Yes
Data format	Simple XML	Compl. XML	Faulty TXT	XML database
Interoperability	Good	Fair	Good	Bad
User community / support	Small?, no support?	Large, good support	Large, fair support	Small, good support
Life cycle	Old (but new version 2011)	Constantly developed	Not officially supported, old	New, being developed

Annotation Tools

	Transcriber	ELAN	Toolbox	FLEX
Complexity	Easy	Complex, w. easier modes	Complex to configure	Complex
Audio	Yes	Yes	No (can play)	No
Video	No	Yes	No	No
Tiers	1 per speaker	Unlimited	Unlimited	Fixed: 8
Lexicon interop., automatic glossing	No	No (is planned)	Yes	Yes
Unicode	No input	Yes	Yes	Yes
Data format	Simple XML	Compl. XML	Faulty TXT	XML database
Interoperability	Good	Fair	Good	Bad
User community / support	Small?, no support?	Large, good support	Large, fair support	Small, good support
Life cycle	Old (but new version 2011)	Constantly developed	Not officially supported, old	New, being developed

Annotation Tools

	Transcriber	ELAN	Toolbox	FLEX
Complexity	Easy	Complex, w. easier modes	Complex to configure	Complex
Audio	Yes	Yes	No (can play)	No
Video	No	Yes	No	No
Tiers	1 per speaker	Unlimited	Unlimited	Fixed: 8
Lexicon interop., automatic glossing	No	No (is planned)	Yes	Yes
Unicode	No input	Yes	Yes	Yes
Data format	Simple XML	Compl. XML	Faulty TXT	XML database
Interoperability	Good	Fair	Good	Bad
User community / support	Small?, no support?	Large, good support	Large, fair support	Small, good support
Life cycle	Old (but new version 2011)	Constantly developed	Not officially supported, old	New, being developed

Annotation Tools

	Transcriber	ELAN	Toolbox	FLEX
Audio	Yes	Yes	No (can play)	No
Video	No	Yes	No	No
Lexicon interop., automatic glossing	No	No (is planned)	Yes	Yes

Topics

1. Data and Annotation (Theory)
2. Annotation Tools (Overview and Comparison)
3. **Intro to Interlinearization (not time-aligned)**

   Excurse: Text- vs. sentence-based databases
4. Time-aligned Annotation
   1. ELAN-Generated Annotation
   2. Play chunks of an audio file in Toolbox

      Excurse: Regular Expressions

      Excurse: UNICODE and UTF-8
   3. Transcriber-Generated Annotation; Conversions
   4. Round-trip Configuration ELAN--Toolbox

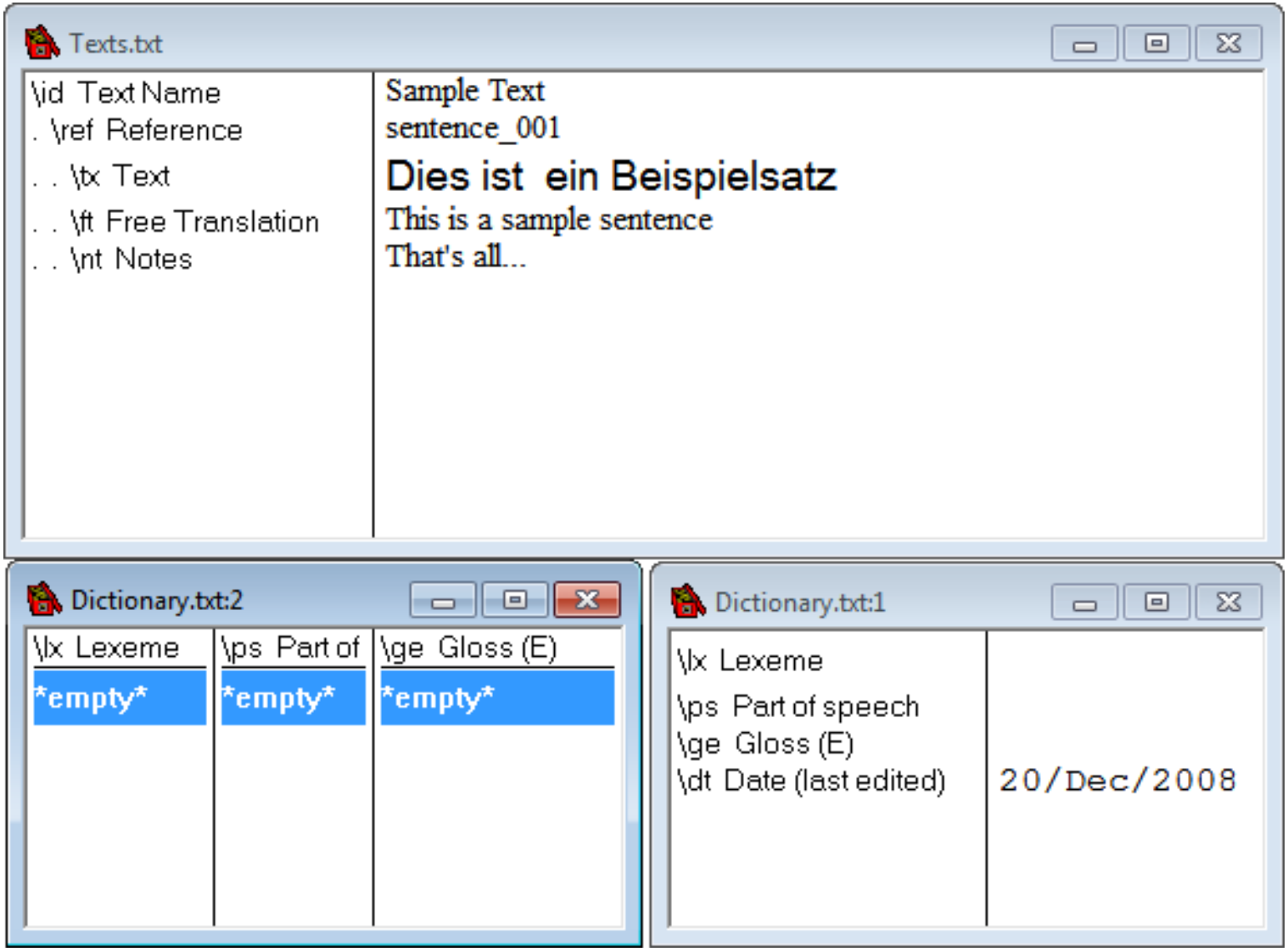
Annotation without time-linking

If you do not have a project yet,
install a new toolbox project.
Use INSTALLTOOLBOXNEWPROJECT###.EXE

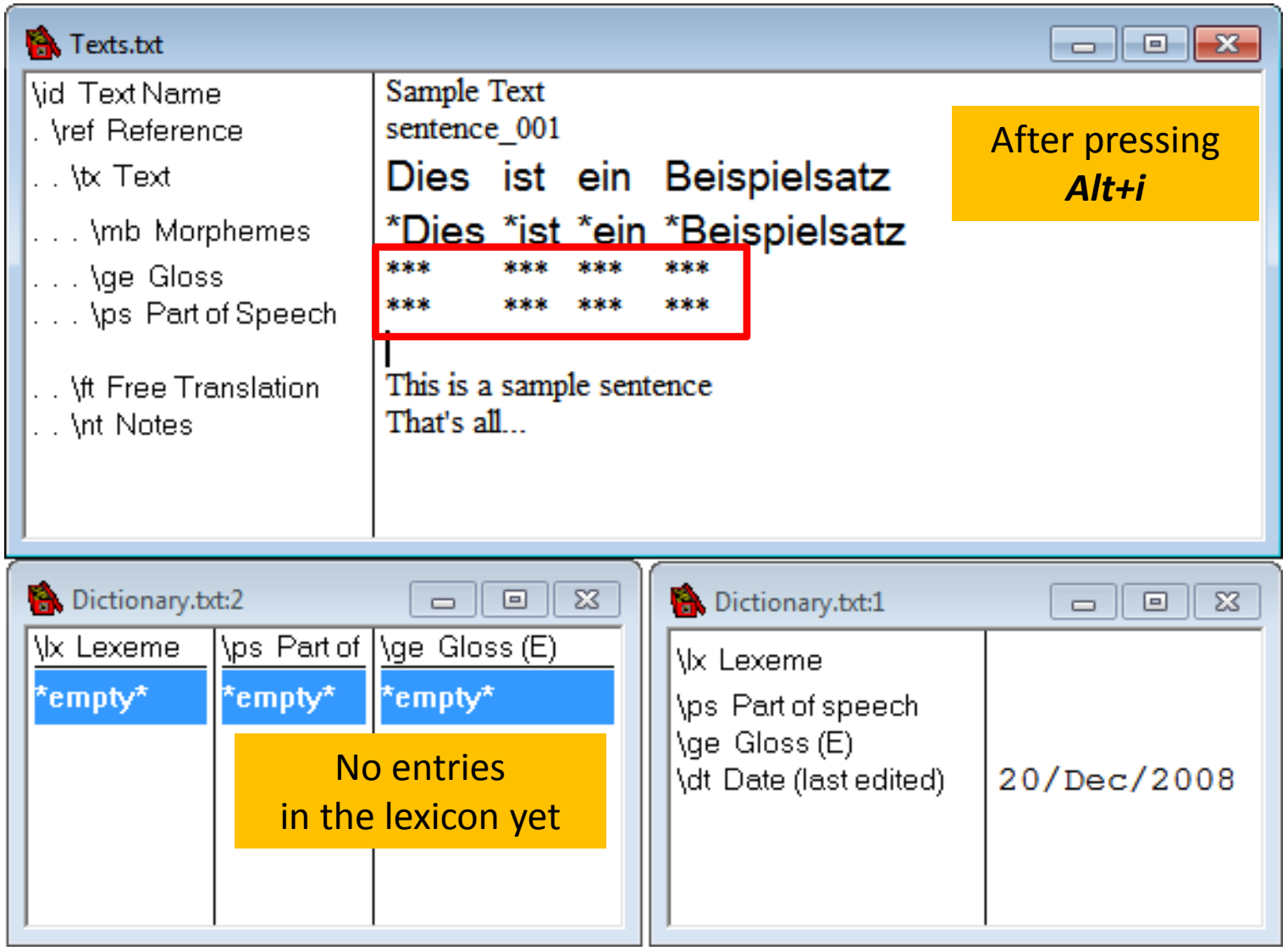
TEXT.TYP provides the set-up for *basic glossing*:

\REF Reference (should be unique)
\TX Text (sentence)
\MB Morphemes (basic form)
\GE Gloss (English)
\PS Part of Speech (on morphological level)
\FT Free translation (English)
\NT Notes

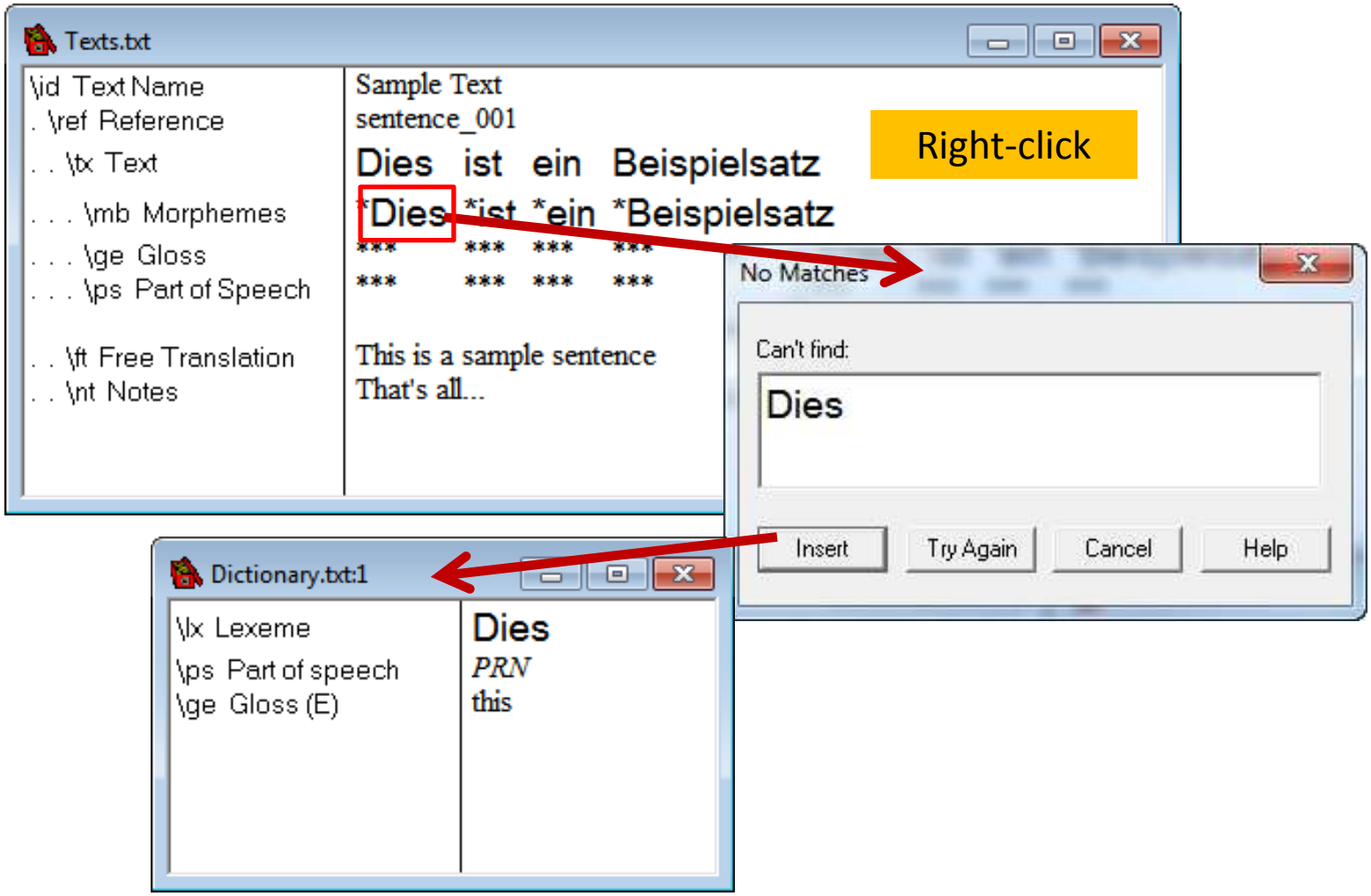
Toolbox default setting



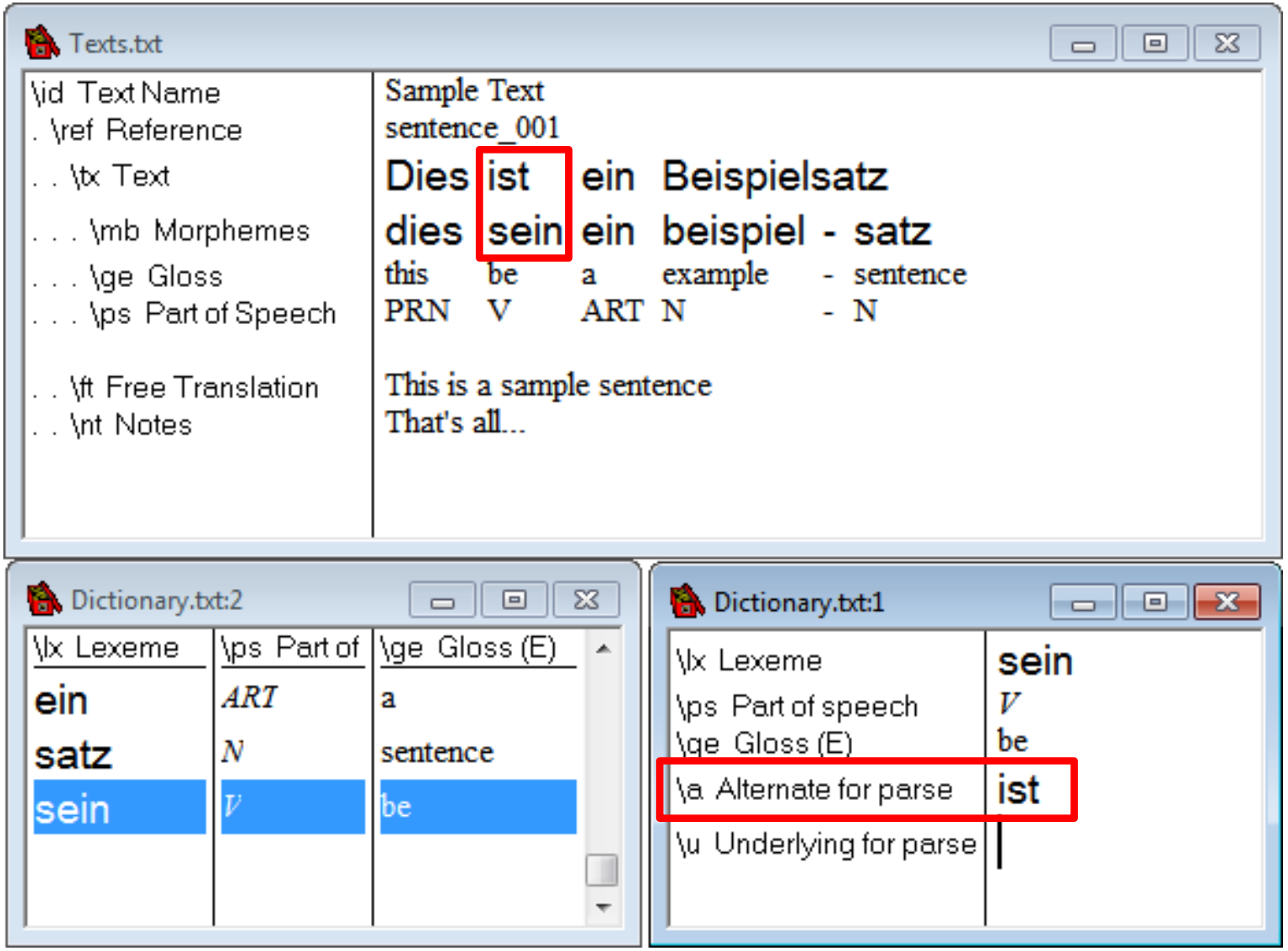
Interlinearizing



Interlinearizing: adding lexical entries



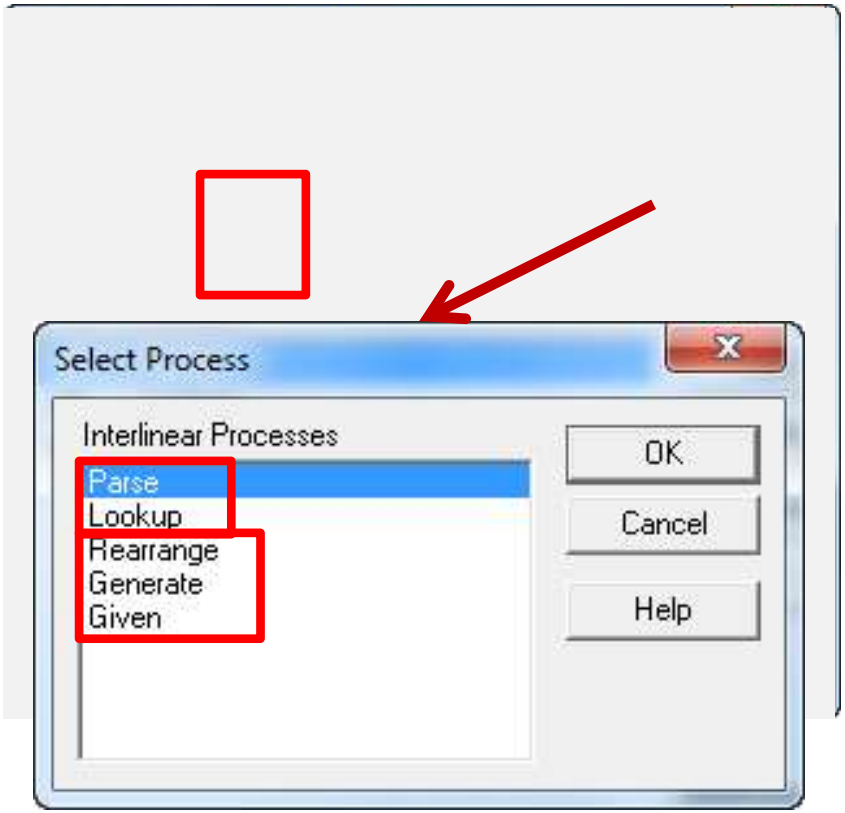
Toolbox default setting: interlinearized



Toolbox: Text and lexicon

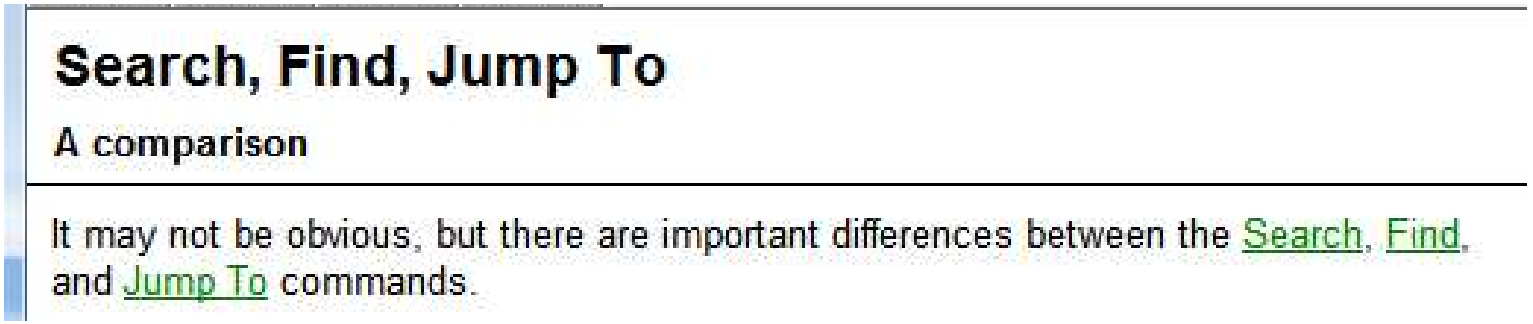
There are three principle ways in which the texts can be connected to the dictionary (or dictionaries):

- **Jump path**
- **Parse** (interlinearization)
- **Lookup** (interlinearization)
- Other interlinearization options are less often used



Toolbox: Jump paths

If a *jump path* for a field is defined, right-clicking in that field searches for identical content in another field in another (or the same) database, and opens the corresponding record in that database -- it is like a ***hypertext link***



Toolbox: Interlinearization processes

Parse
A Parse process breaks words into morphemes by separating affixes from roots and separating compound roots. It uses the affixes and roots in the lexicon to do this. There is normally only one parse process, and it is usually the first process.

Lookup
A Lookup process looks up morphemes in a lexicon and outputs the content of another field. Lookup is used for gloss and part of speech.

Toolbox: Parse details

Toolbox' parser works well with most mainly *isolating* or *agglutinative* languages, less good for *fusional* or (worse) *polysynthetic* languages

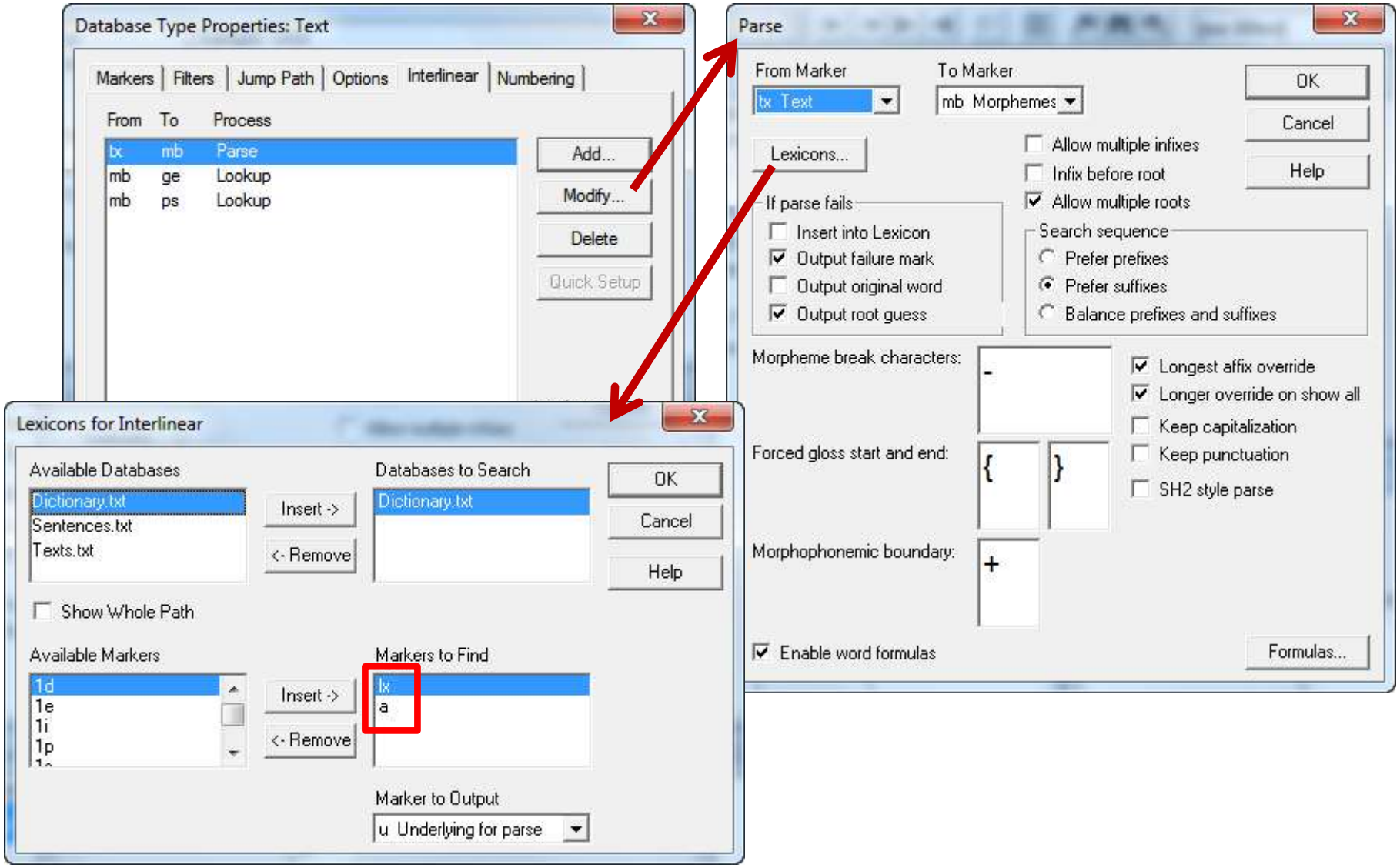
Allomorphy can be covered by using the

\va *variant form* - field and the
\a *alternate form* - field in the lexicon

- *Morpho-phonology*, *sandhi* and *suppletition*:
  \a + **\u *underlying form*** - field, for example:

```
\a went
\u go -ed
```

Interlinearization settings



Shoebox manual

The other field markers are correct and can be left alone. Their meanings will be explained later. The Quick Setup Lexicon Markers dialog box should look like the following:



Fig. 213. Quick Setup Lexicon Markers for interlinearization

10. Click OK



Fig. 214. Interlinear Process List for Axininca

As in the last exercise, you will see the Interlinear Process List now showing the three processes set up for you. The list of To markers shows the interlinear lines that will be generated by

Bookmarks

- Basic interlinearization
- Interlinear setup
- More on interlinearization
  - Interlinear setup
  - Interlinearizing
  - Review
- Affixes
- Morphophonemics
- User's Guide
- Basic information
- Frequently asked questions
- Notes on interlinearization
  - Whole-word parses
  - Longest-match principle
  - Word formulas
  - Forced values
  - Reduplication
    - General reduplication
    - Template reduplication
  - Infixes
  - Compound stems
  - Hyphens in the lexicon
  - Morpheme break characters
  - Multiple parse processes
  - Multiple words
  - Multiple glosses
  - Spell check
  - Related language adaptation
  - Editing interlinear text
  - Copying interlinear text
- Appendixes
- Appendix A: Importing from MS-DOS

Shoebox as it interlinearizes. The generated lines will be **m**, **g**, and **p**.

The names of the processes tell what they do. A Parse process parses words into morphemes. A Lookup process looks up glosses of words or morphemes in a lexicon. Notice that the two Lookup processes both look up morphemes from the **m** line. The first finds the gloss of the morpheme and the second finds the part of speech.

11. Click OK to return to the main window.



Fig. 215. Axininca project ready for interlinearization

Put the text insertion point at the beginning of the first sentence of text (in the \t field).

1. Click to start interlinearization.



Fig. 216. Wildboar text interlinearization

Topics

1. Data and Annotation (Theory)
2. Annotation Tools (Overview and Comparison)
3. Intro to Interlinearization (not time-aligned)

   ***Excurse:* Text- vs. sentence-based databases**
4. Time-aligned Annotation
   1. ELAN-Generated Annotation
   2. Play chunks of an audio file in Toolbox

      Excurse: Regular Expressions

      Excurse: UNICODE and UTF-8
   3. Transcriber-Generated Annotation; Conversions
   4. Round-trip Configuration ELAN--Toolbox

Text- vs. sentence-based databases

The text db. *record marker* in the Toolbox default setup is

\ID *Text name*

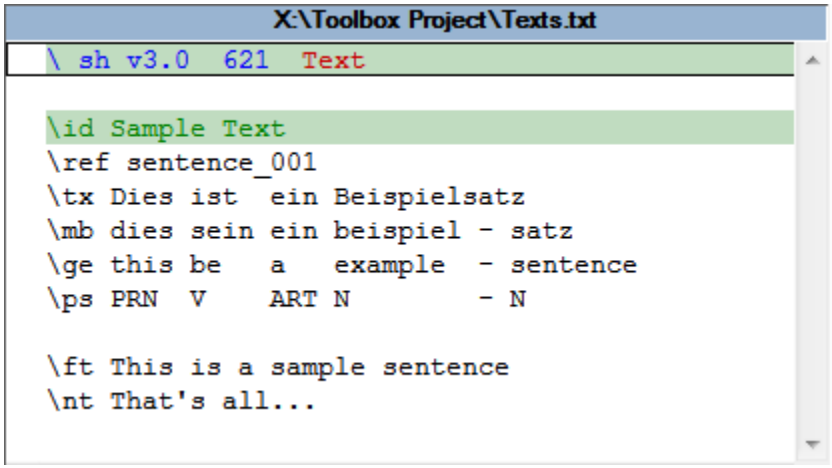
Each record corresponds to one entire text.

This setting is not practical for several reasons, for instance:

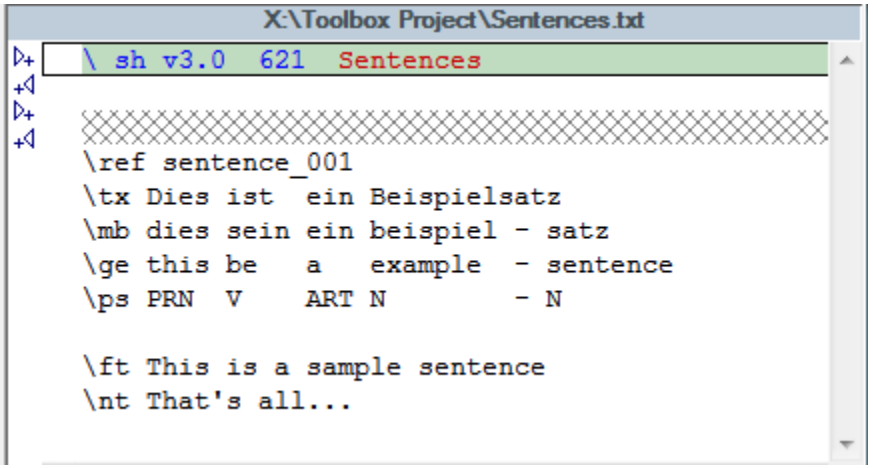
- We need separate files for different stories if we want to export them to ELAN
- If one searches or filters, the hits (results) refer to whole texts
- If one wants to do *advanced glossing*, the screen becomes confusing

Adjust text db. records to sentences

Original **text file** with text-level records



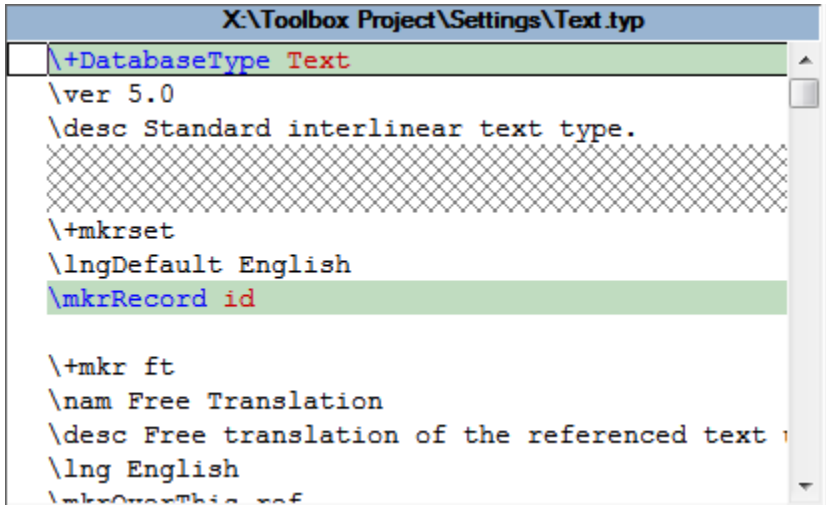
Adjusted **text file** with sentence-level records

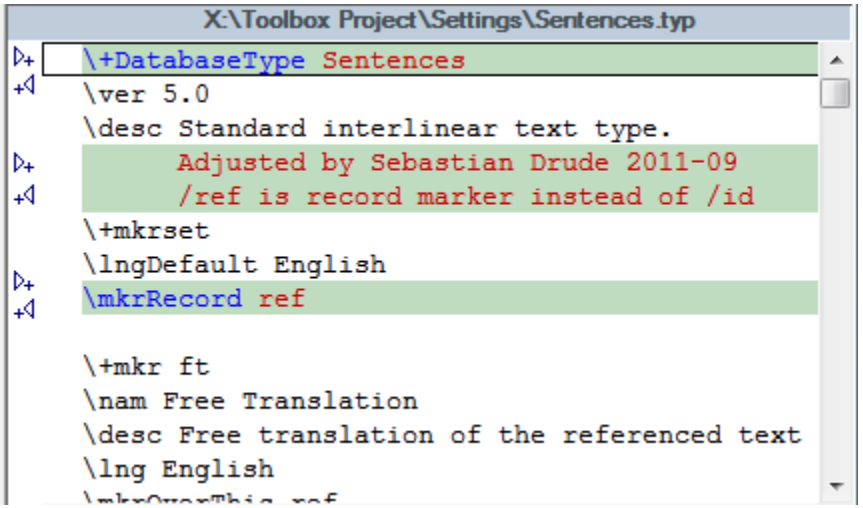


Adjust records to sentences

Original **.typ**-file with text-level records

Adjusted **.typ**-file with sentence-level records

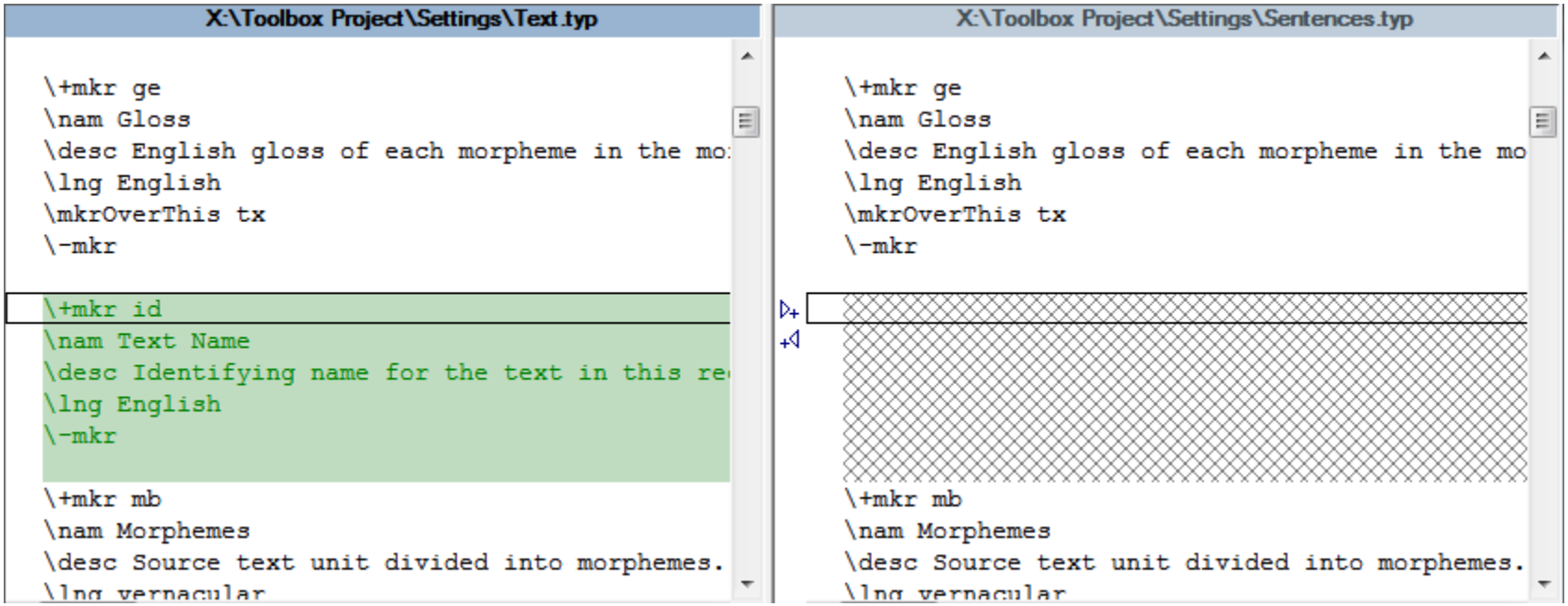




Adjust records to sentences

Original **.typ**-file with text-level records

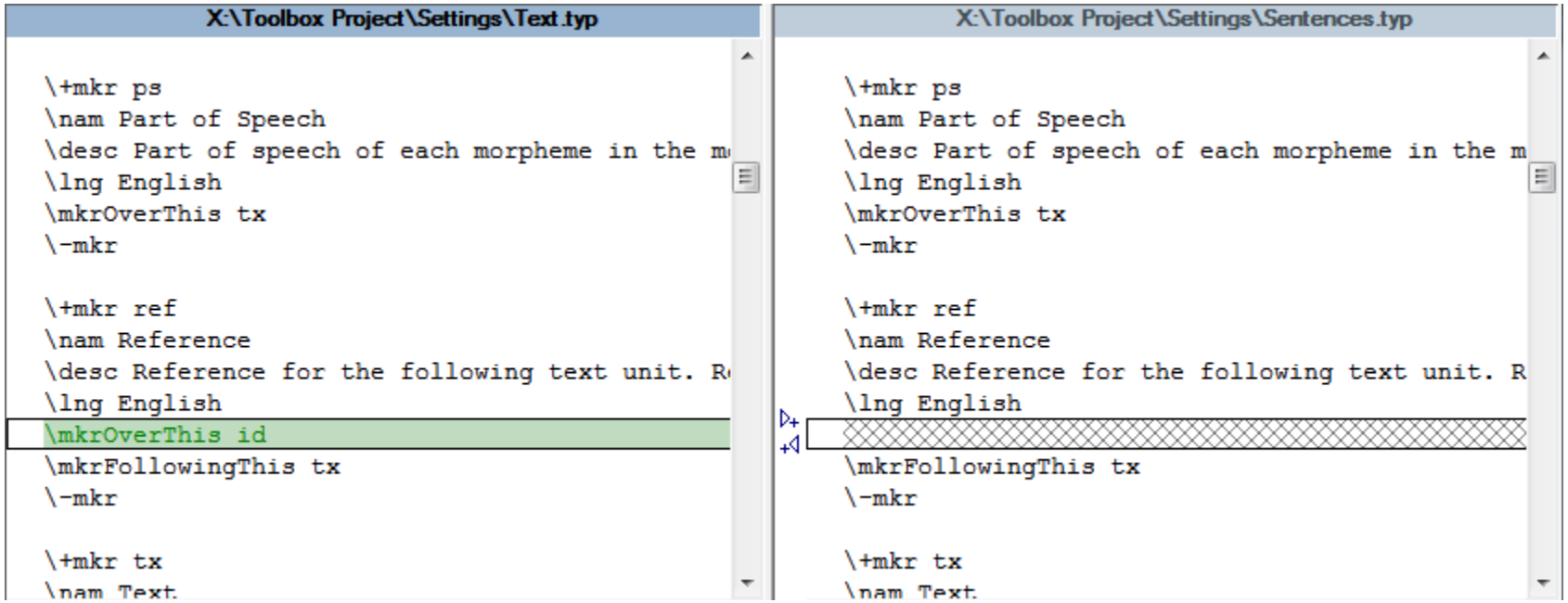
Adjusted **.typ**-file with sentence-level records



Adjust records to sentences

Original **.typ**-file with text-level records

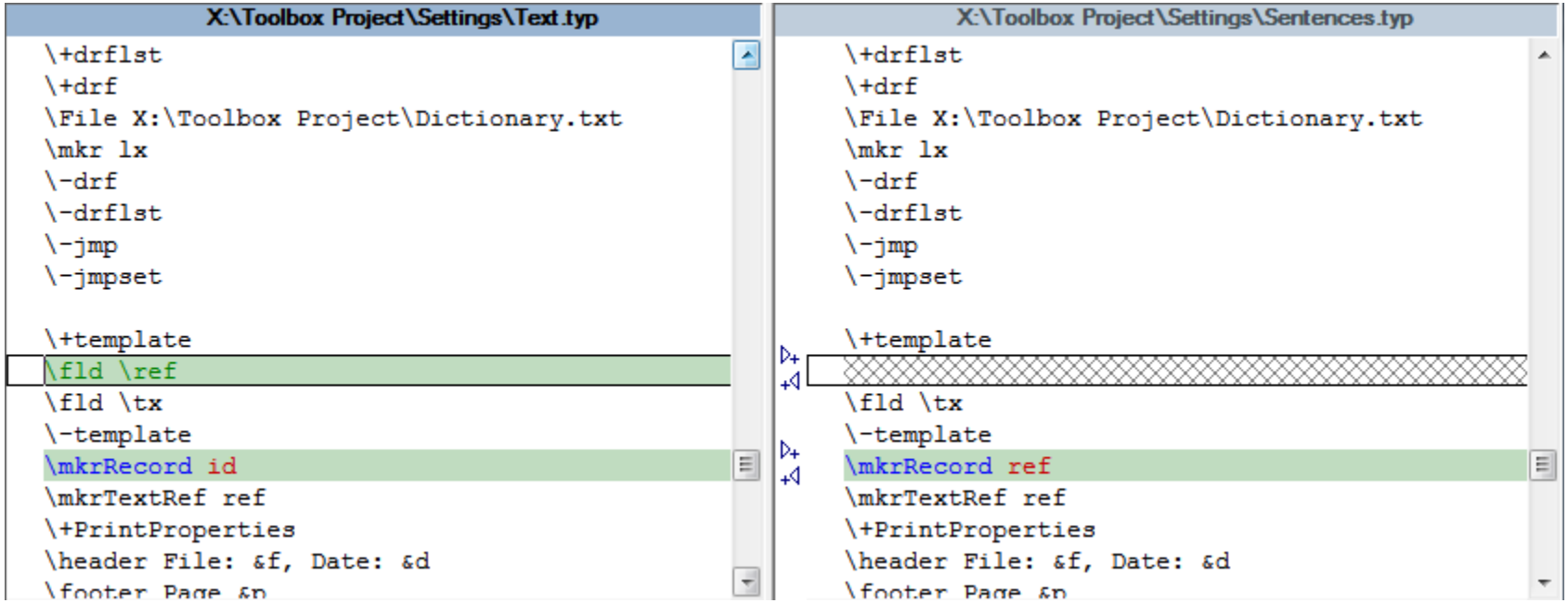
Adjusted **.typ**-file with sentence-level records



Adjust records to sentences

Original **.typ**-file with text-level records

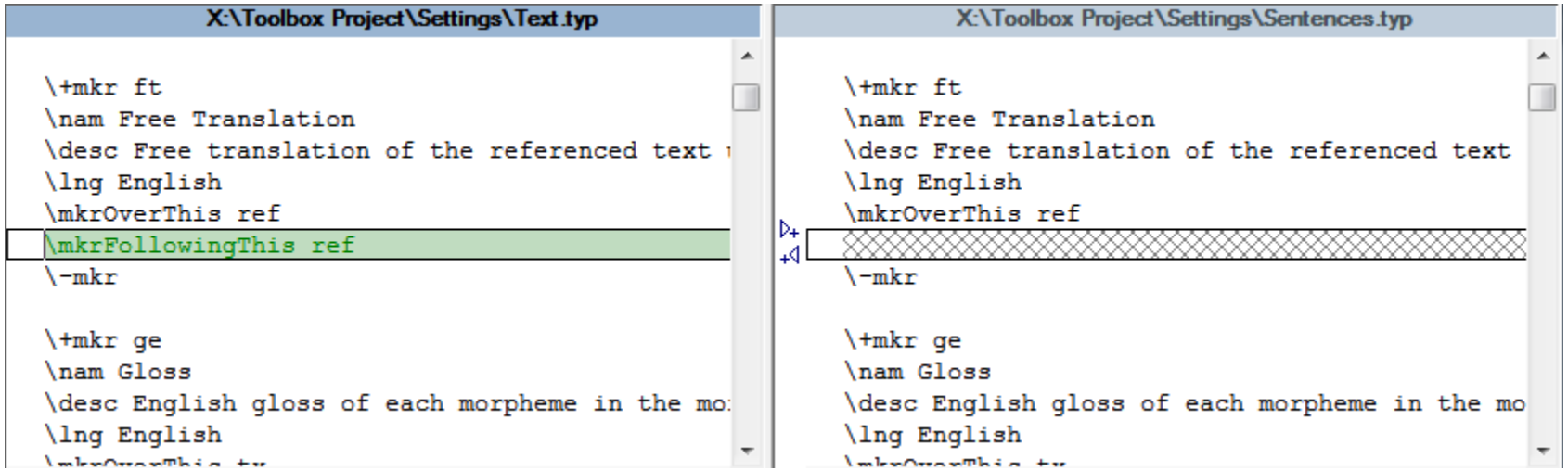
Adjusted **.typ**-file with sentence-level records



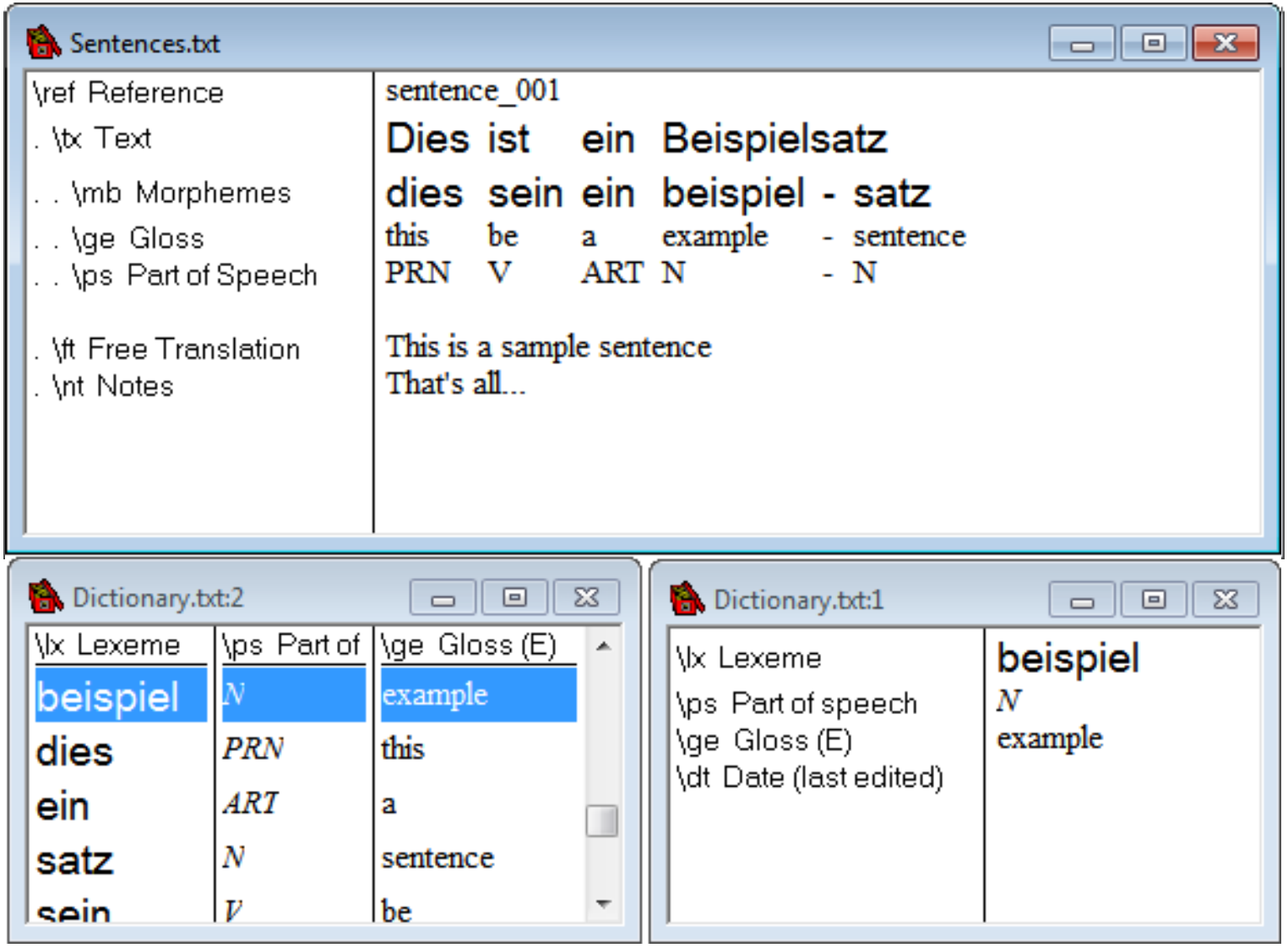
Adjust records to sentences

Original **.typ**-file with text-level records

Adjusted **.typ**-file with sentence-level records



New Toolbox setting



Topics

1. Data and Annotation (Theory)
2. Annotation Tools (Overview and Comparison)
3. Intro to Interlinearization (not time-aligned)

   Excurse: Text- vs. sentence-based databases

4. **Time-aligned Annotation**
   1. ELAN-Generated Annotation
   2. Play chunks of an audio file in Toolbox

      Excurse: Regular Expressions

      Excurse: UNICODE and UTF-8
   3. Transcriber-Generated Annotation; Conversions
   4. Round-trip Configuration ELAN--Toolbox

Annotation with time-alignment

Time-linking is the activity and result of ***specifying the time-alignment*** of each annotation associated with a certain chunk in the media file

Time marks: the start/end times of each chunk

Toolbox can ***play*** chunks of audio files, but cannot practically be used to ***change*** the time marks. In fact, doing so by hand can lead to problems, especially if chunks overlap.

Annotation with time-alignment

The *time-linking* has to be done in some ***other tool***, usually together with the first *transcription* (for identification of each chunk)

We focus on two tools, *ELAN* and *Transcriber*

Both are not topic of this tutorial by themselves, but we here mention on some aspects related to Toolbox

Topics

1. Data and Annotation (Theory)
2. Annotation Tools (Overview and Comparison)
3. Intro to Interlinearization (not time-aligned)
   Excurse: Text- vs. sentence-based databases
4. Time-aligned Annotation
   1. **ELAN-Generated Annotation**
   2. Play chunks of an audio file in Toolbox
      Excurse: Regular Expressions
      Excurse: UNICODE and UTF-8
   3. Transcriber-Generated Annotation; Conversions
   4. Round-trip Configuration ELAN--Toolbox

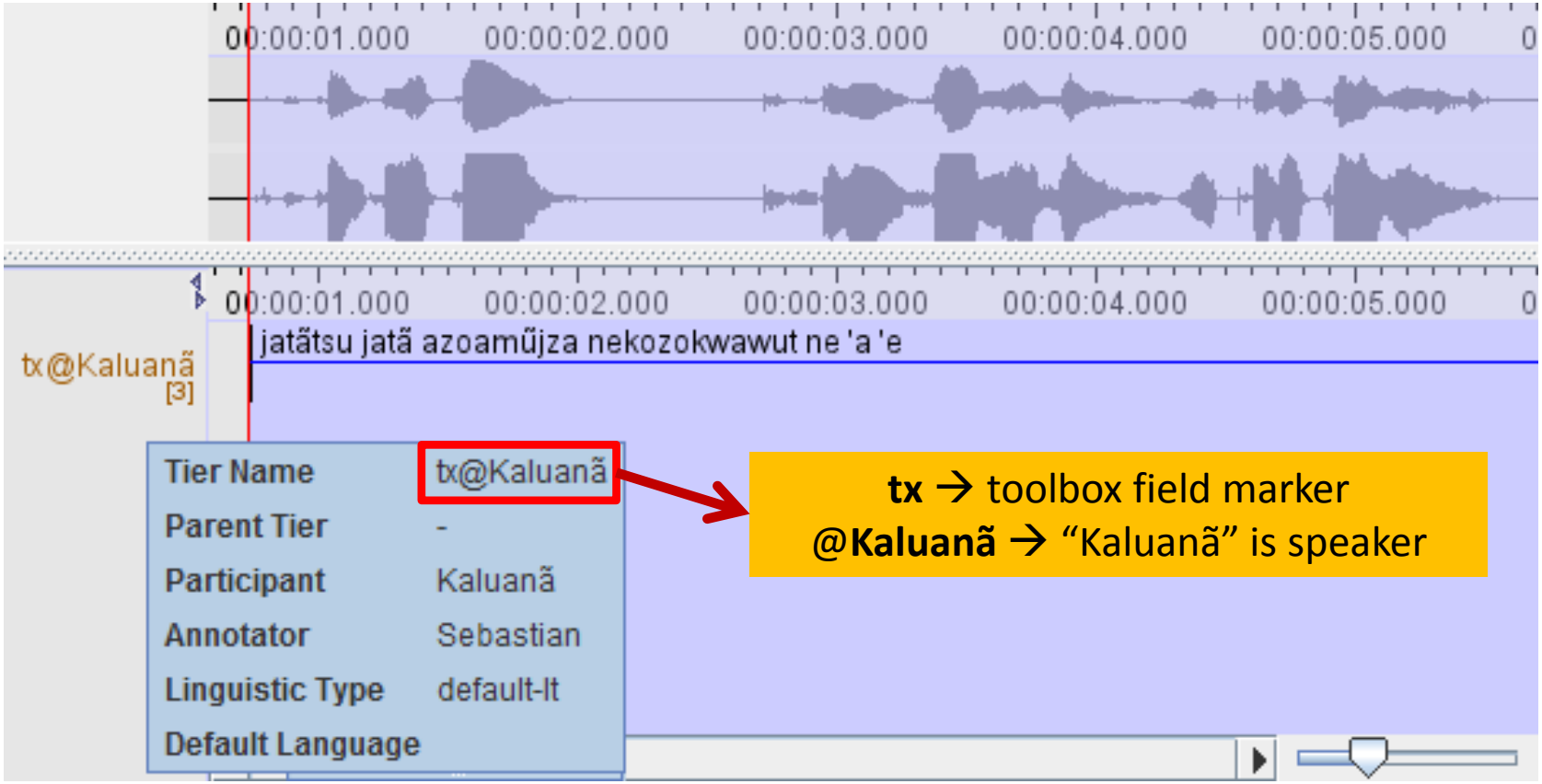
Segmenting and transcribing in ELAN

Segmenting (of a media file): *identification* of relevant chunks and their *time marks*

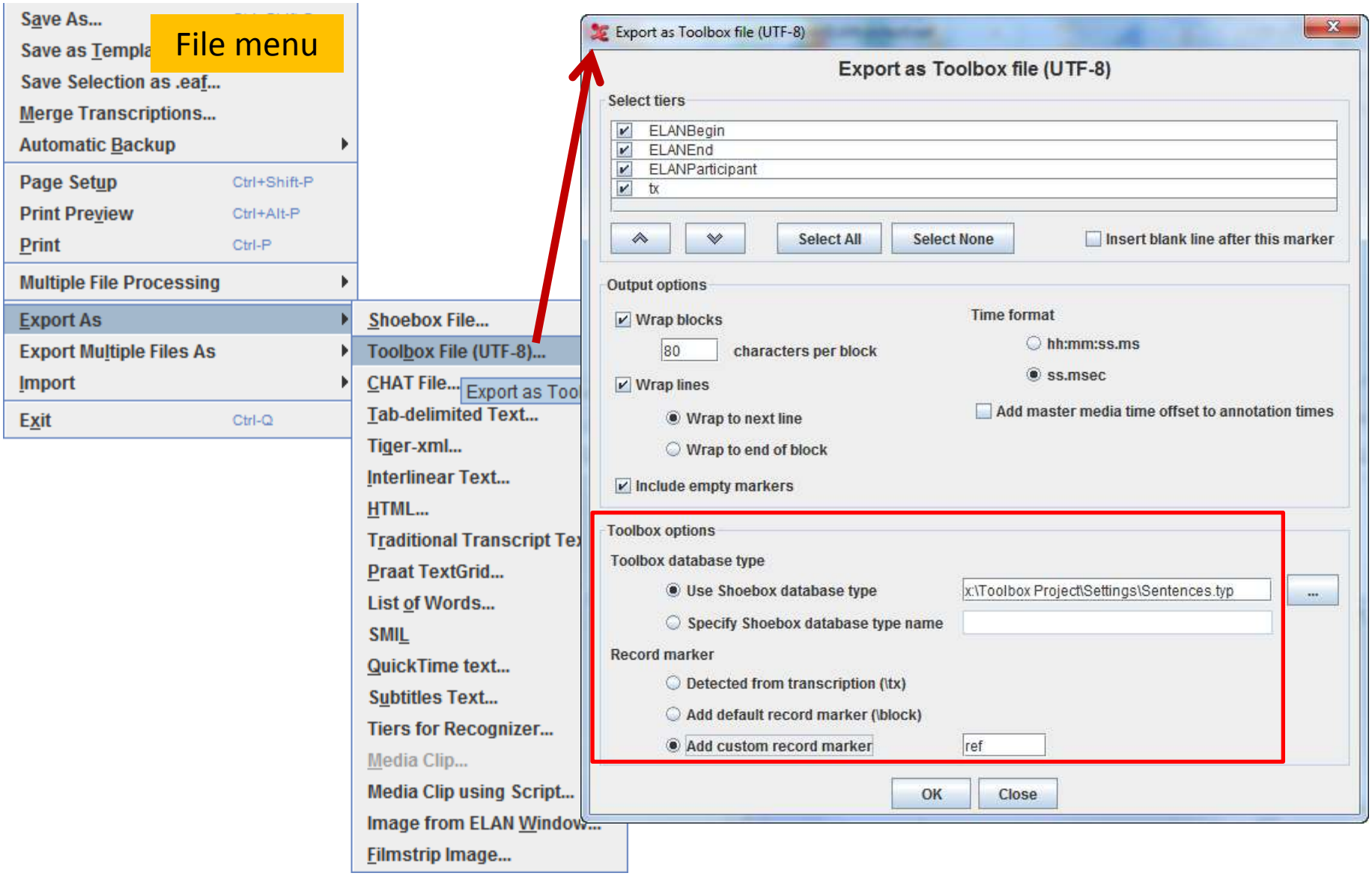
Transcribing : Writing a representation (=annotating) of the expressions in the object language (orthographical, phonemic, or phonetic)

ELAN can be used for both. You can *export* ELAN annotation data to Toolbox format ("Standard Format"), an open it with Toolbox. The results vary depending on the ELAN configuration.

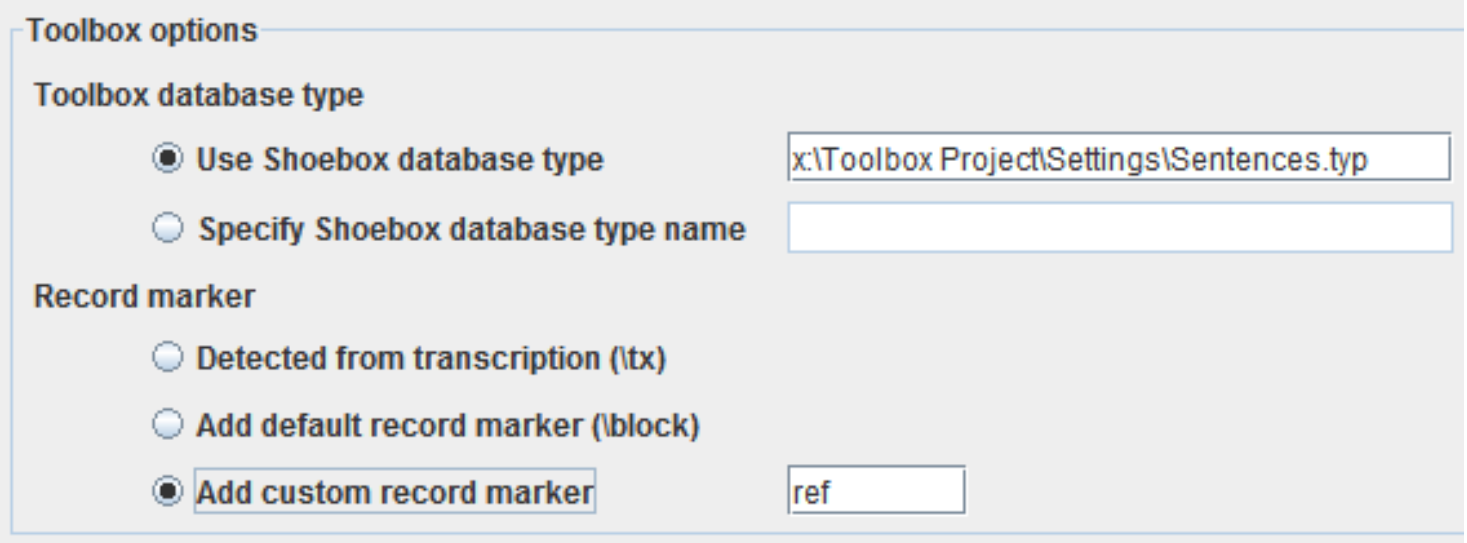
A single ELAN tier



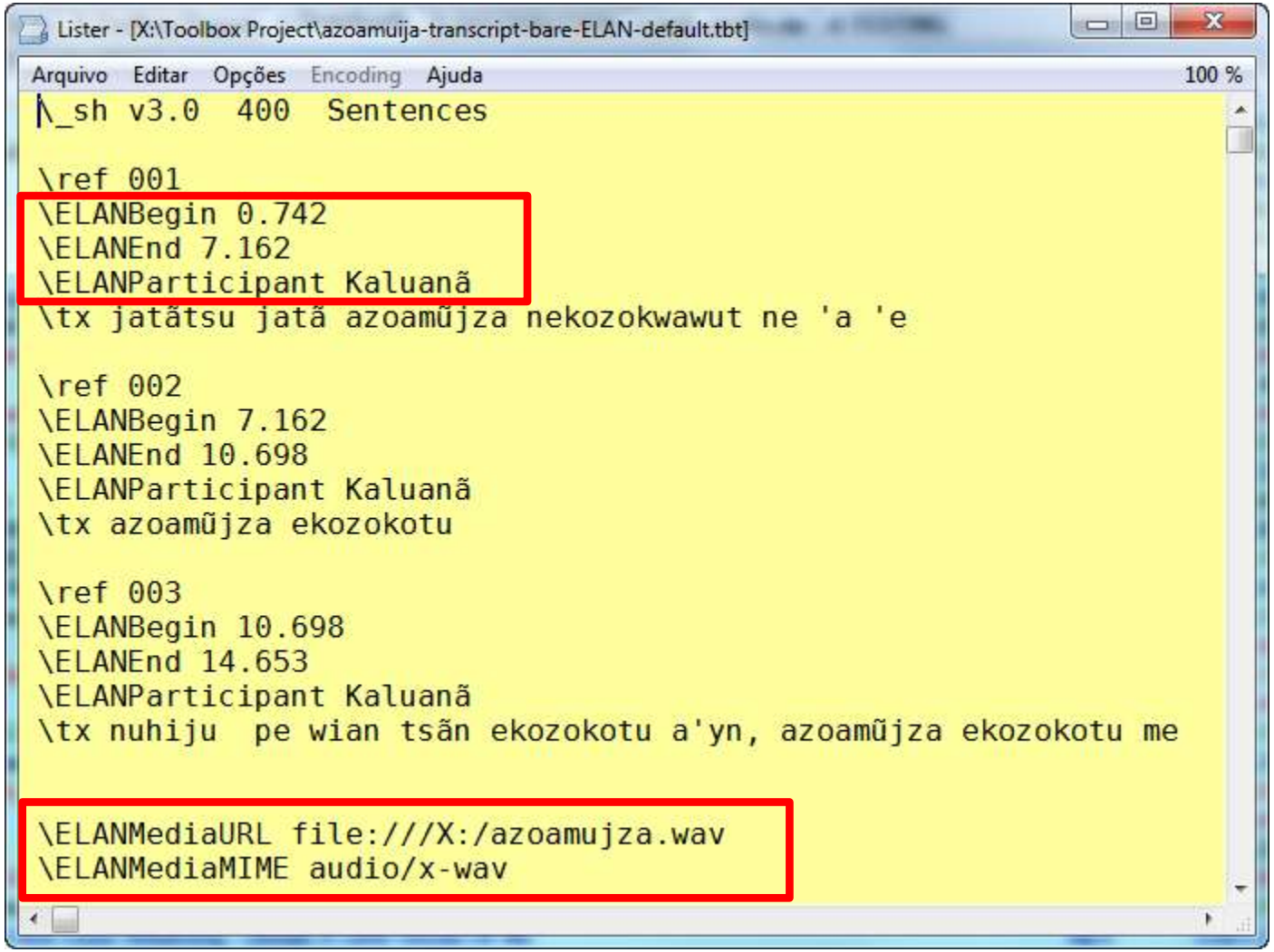
ELAN Toolbox export



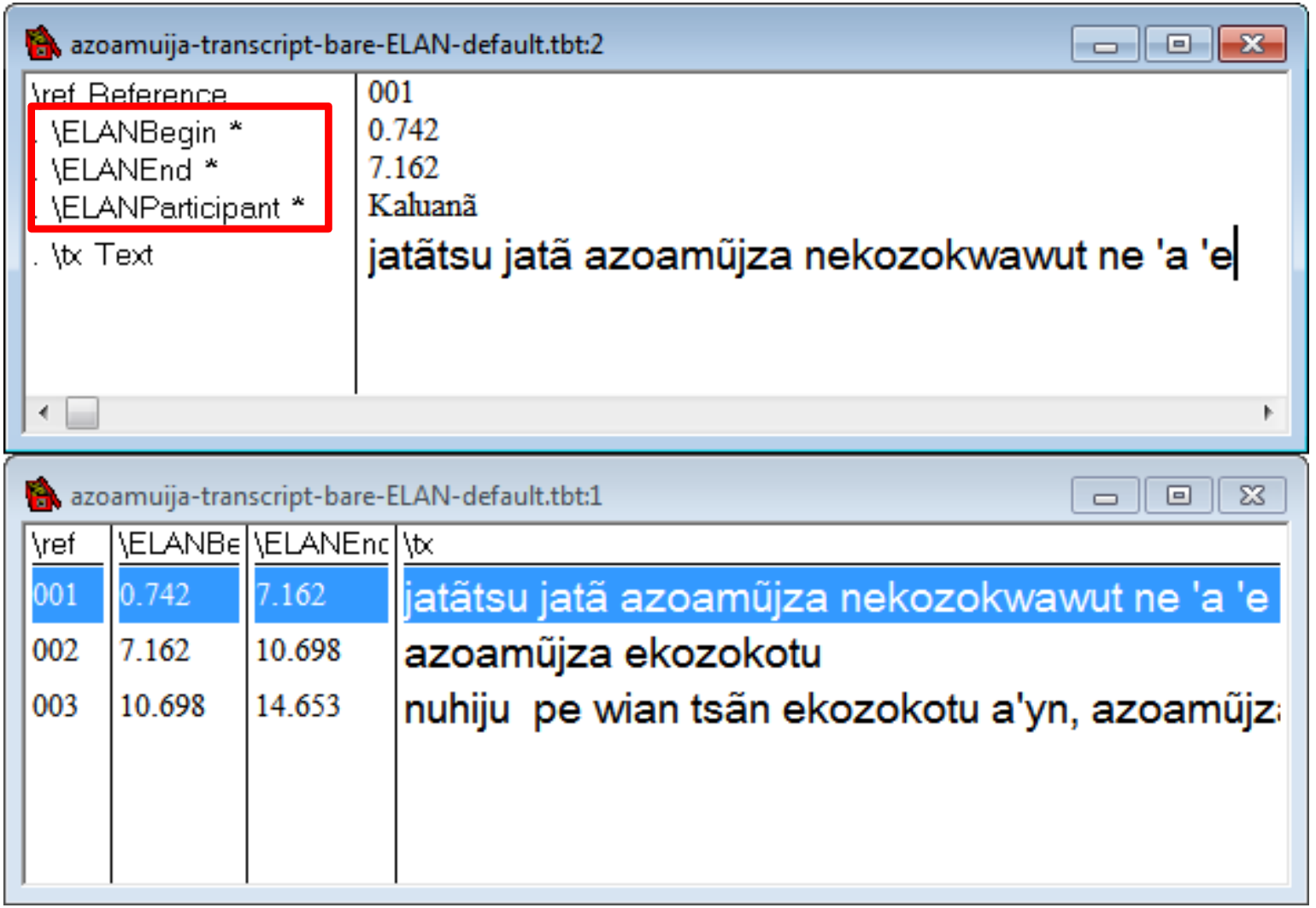
ELAN Toolbox export dialog



ELAN Toolbox export: result



Toolbox import from ELAN



Toolbox import from ELAN: ***.typ*** *file*

X:\Toolbox Project -1\Settings\Sentences.typ

```
\+DatabaseType Sentences
\ver 5.0
\desc Standard interlinear text type.
      Adjusted by Sebastian Drude 2011-09
      /ref is record marker instead of /id
\+mkrset
\lngDefault English
\mkrRecord ref

\+mkr ft
```

X:\Toolbox Project\Settings\Sentences.typ

```
\+DatabaseType Sentences
\ver 5.0
\desc Standard interlinear text type.
      Adjusted by Sebastian Drude 2011-09
      /ref is record marker instead of /id
\+mkrset
\lngDefault English
\mkrRecord ref

\+mkr ELANBegin
\nam *
\lng English
\mkrOverThis ref
\-mkr

\+mkr ELANEnd
\nam *
\lng English
\mkrOverThis ref
\-mkr

\+mkr ELANMediaMIME
\nam *
\lng English
\mkrOverThis ref
\-mkr

\+mkr ELANMediaURL
\nam *
\lng English
\mkrOverThis ref
\-mkr

\+mkr ELANParticipant
\nam *
\lng English
\mkrOverThis ref
\-mkr

\+mkr ft
```

Topics

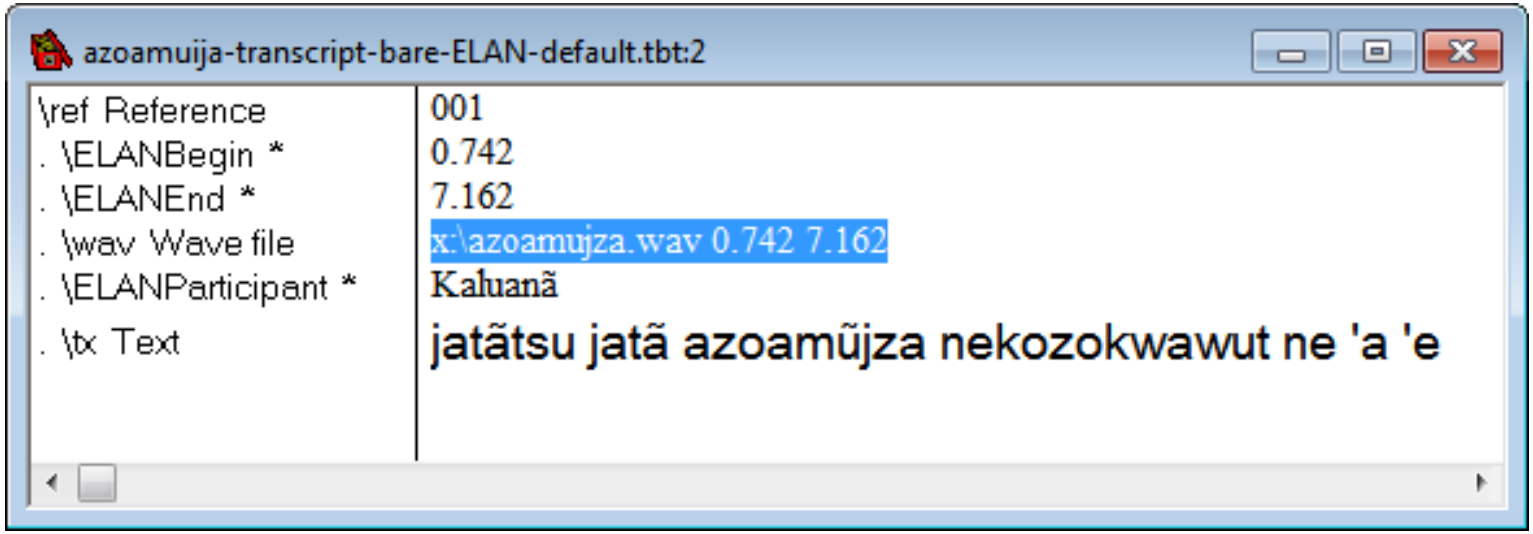
1. Data and Annotation (Theory)
2. Annotation Tools (Overview and Comparison)
3. Intro to Interlinearization (not time-aligned)
   Excurse: Text- vs. sentence-based databases
4. Time-aligned Annotation
   1. ELAN-Generated Annotation
   2. **Play chunks of an audio file in Toolbox**
      Excurse: Regular Expressions
      Excurse: UNICODE and UTF-8
   3. Transcriber-Generated Annotation; Conversions
   4. Round-trip Configuration ELAN--Toolbox

Play chunks of an audio file in Toolbox

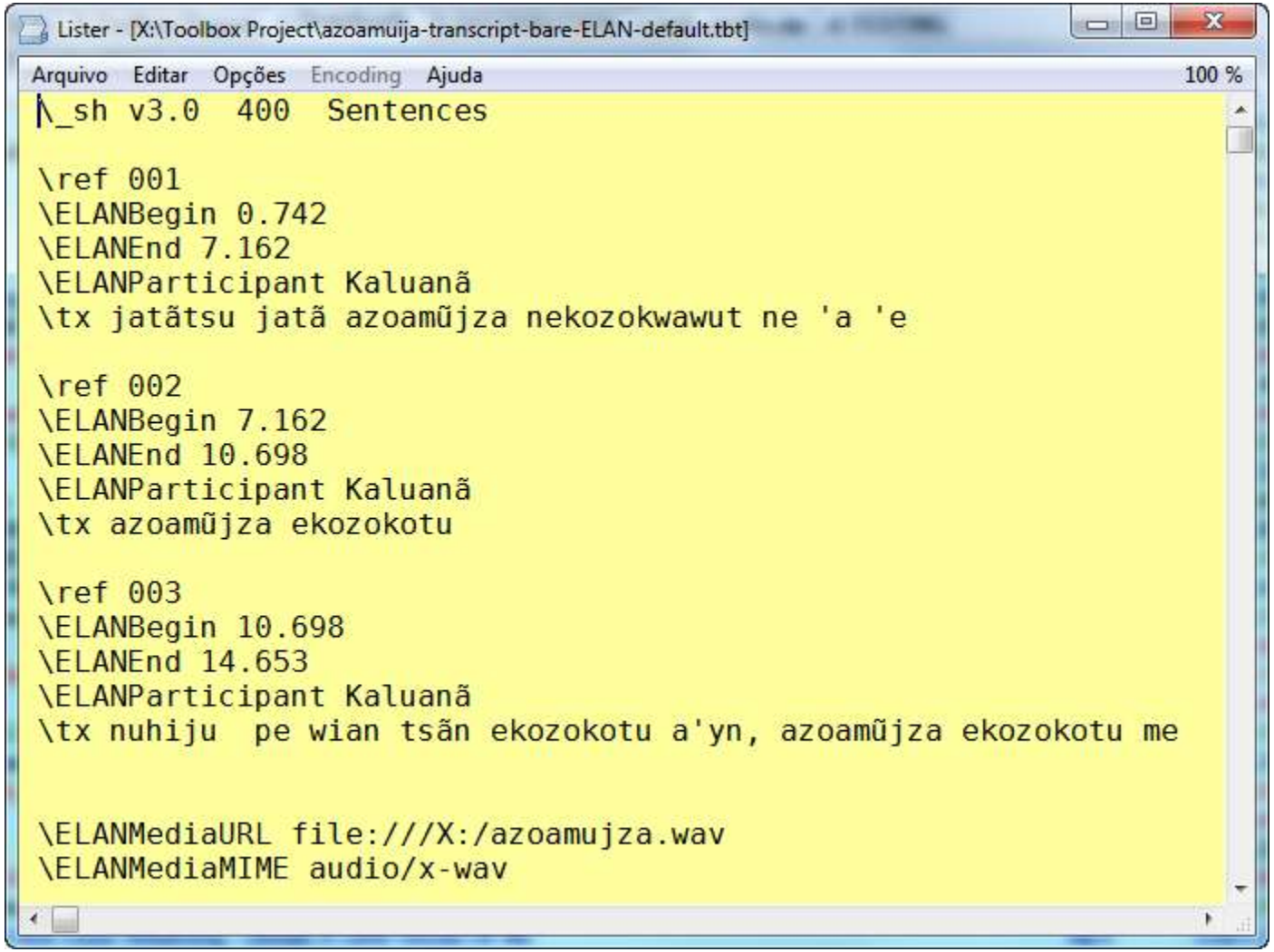
Format: *Path\Filename.wav sss.mmm sss.mmm*

for instance: **X:\azoamujza.wav 0.742 7.162**

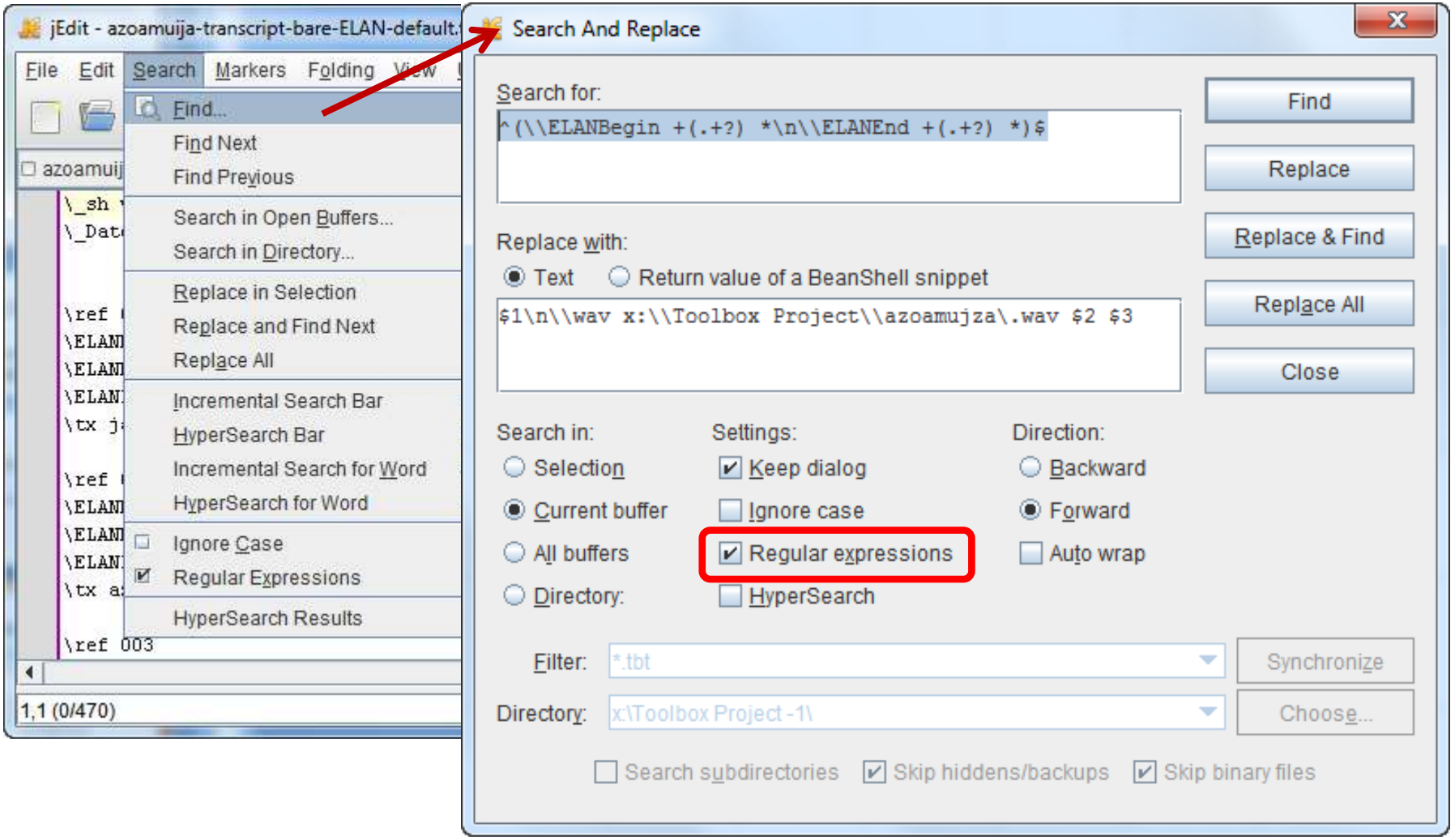
Use **Shift+F4** to play (*Tools* > *Play sound*)



ELAN Toolbox export: result



Creating the audio field



Topics

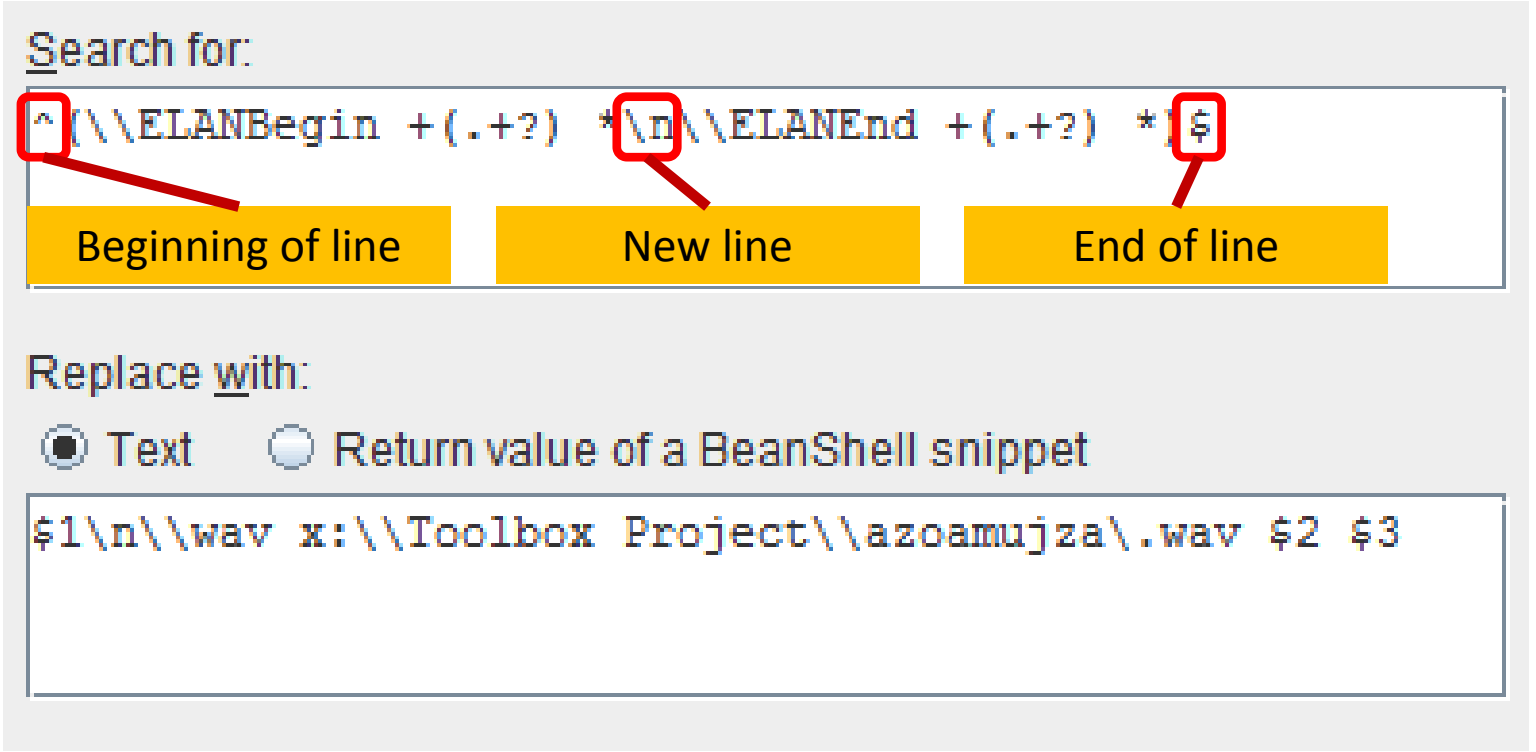
1. Data and Annotation (Theory)
2. Annotation Tools (Overview and Comparison)
3. Intro to Interlinearization (not time-aligned)

   Excurse: Text- vs. sentence-based databases
4. Time-aligned Annotation
   1. ELAN-Generated Annotation
   2. Play chunks of an audio file in Toolbox

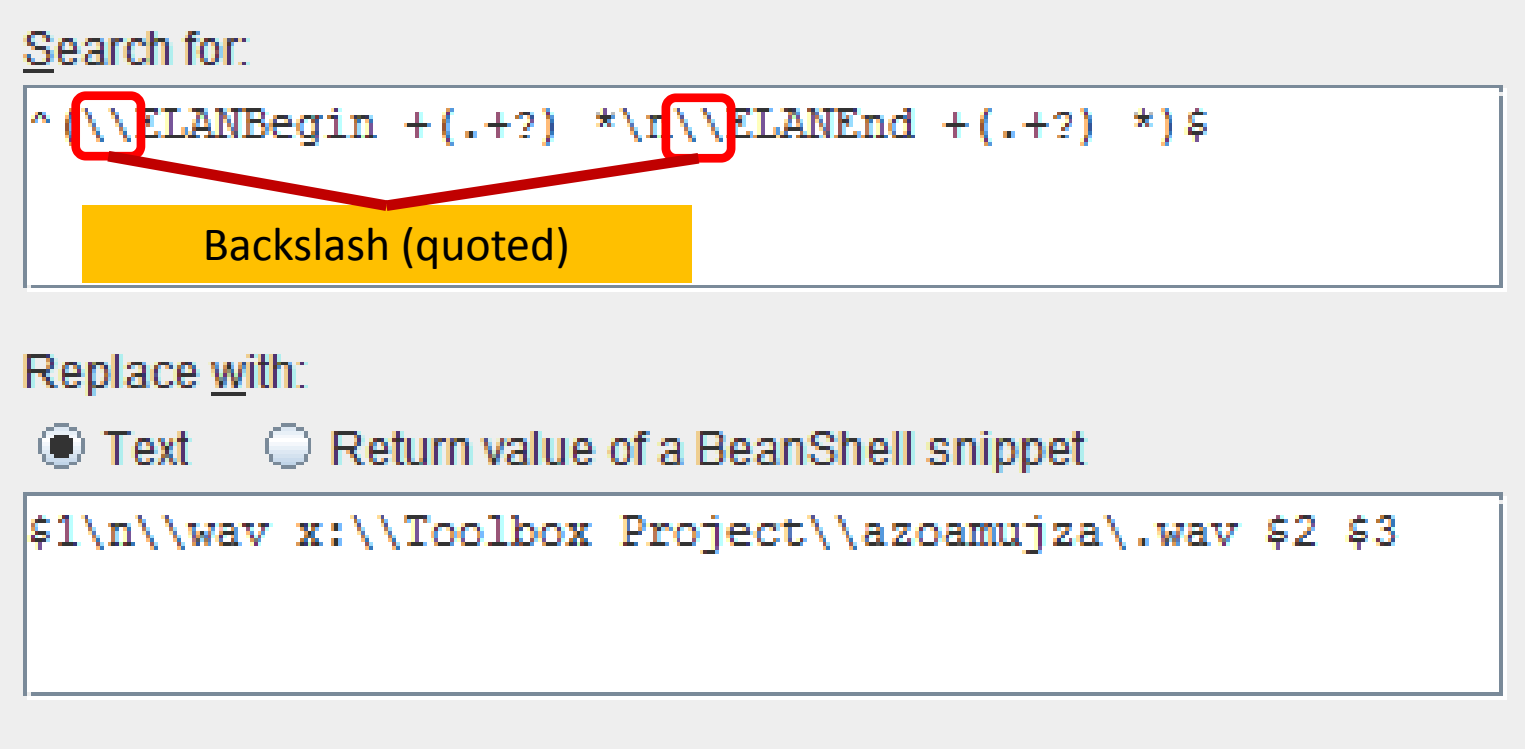
      ***Excurse:* Regular Expressions**

      Excurse: UNICODE and UTF-8
   3. Transcriber-Generated Annotation; Conversions
   4. Round-trip Configuration ELAN--Toolbox

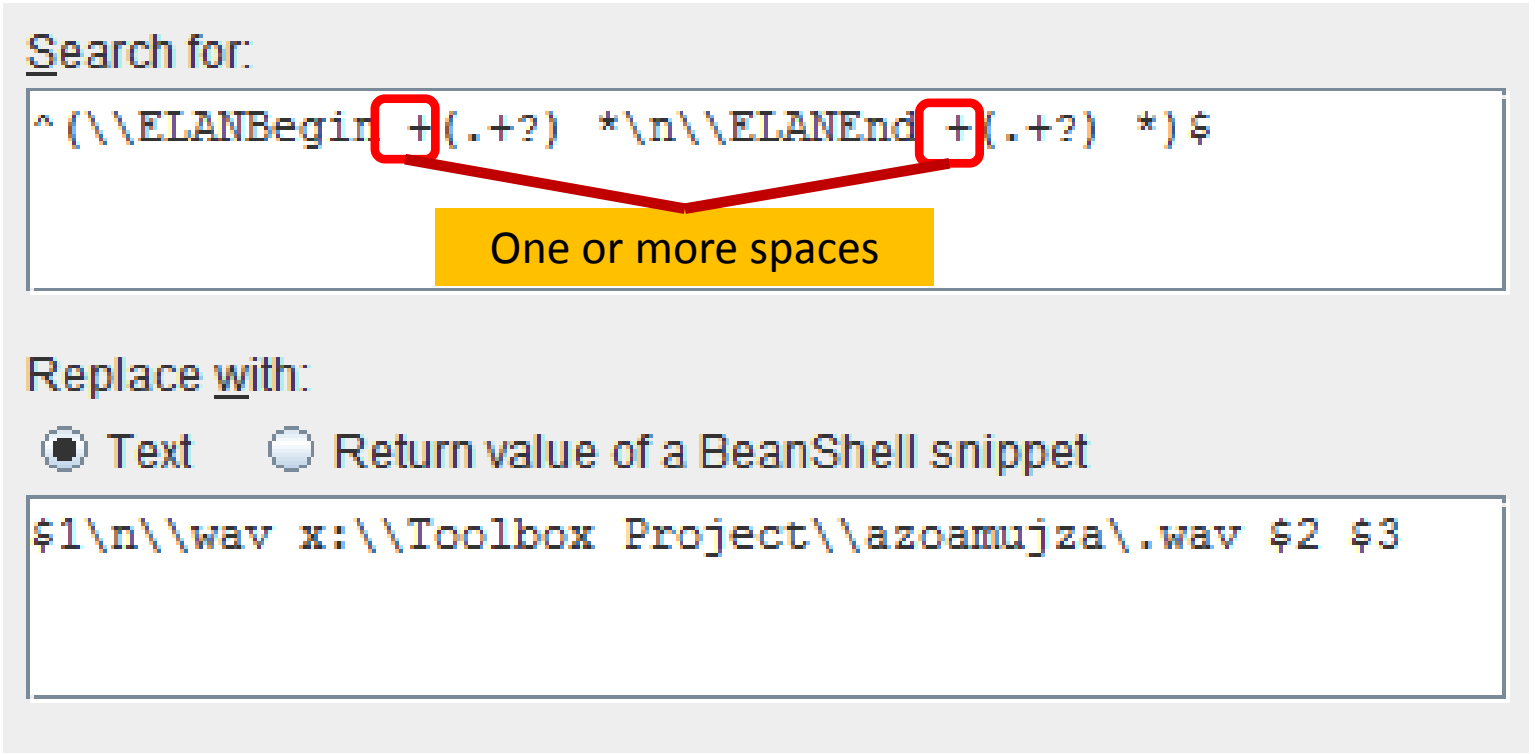
Regular expressions: special characters



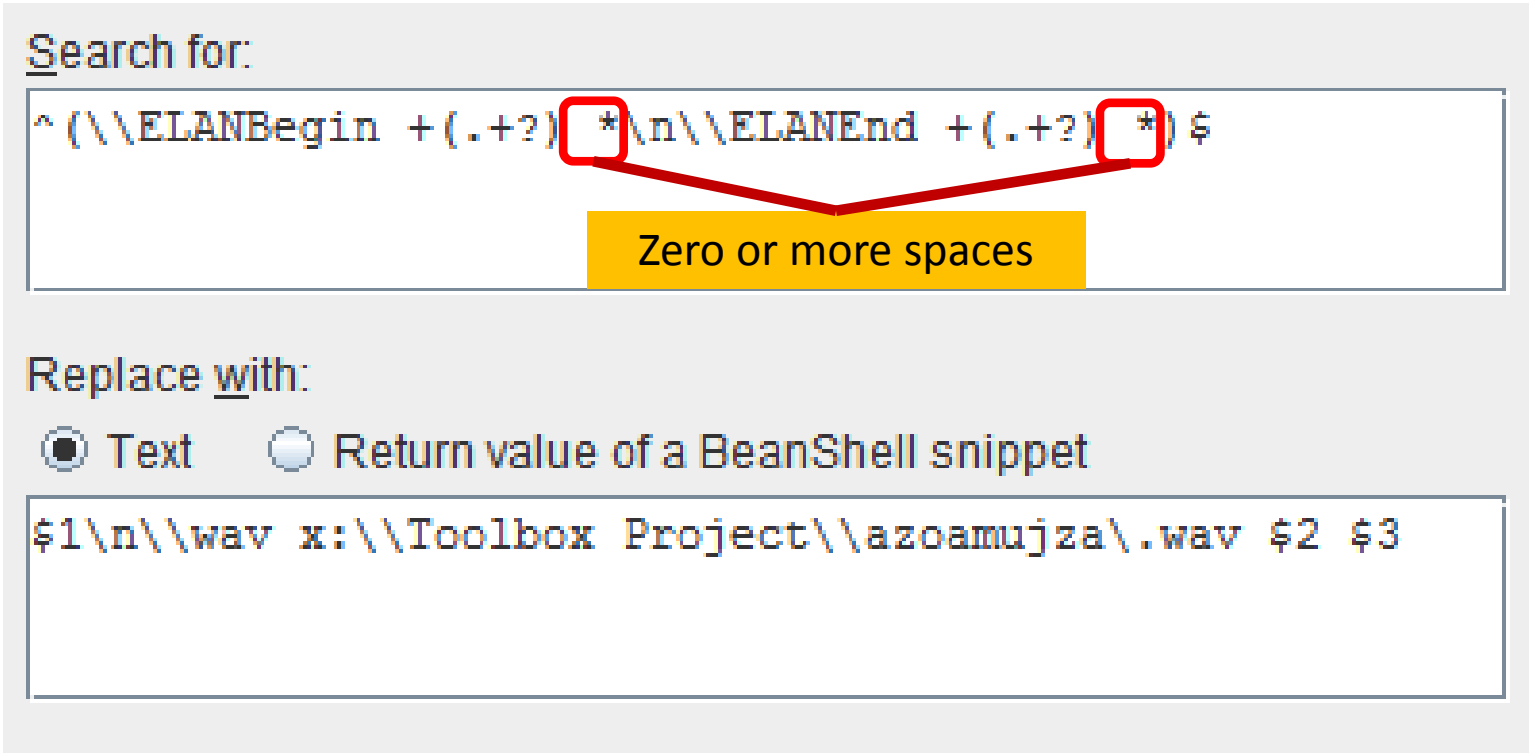
Regular expressions: Quoted characters



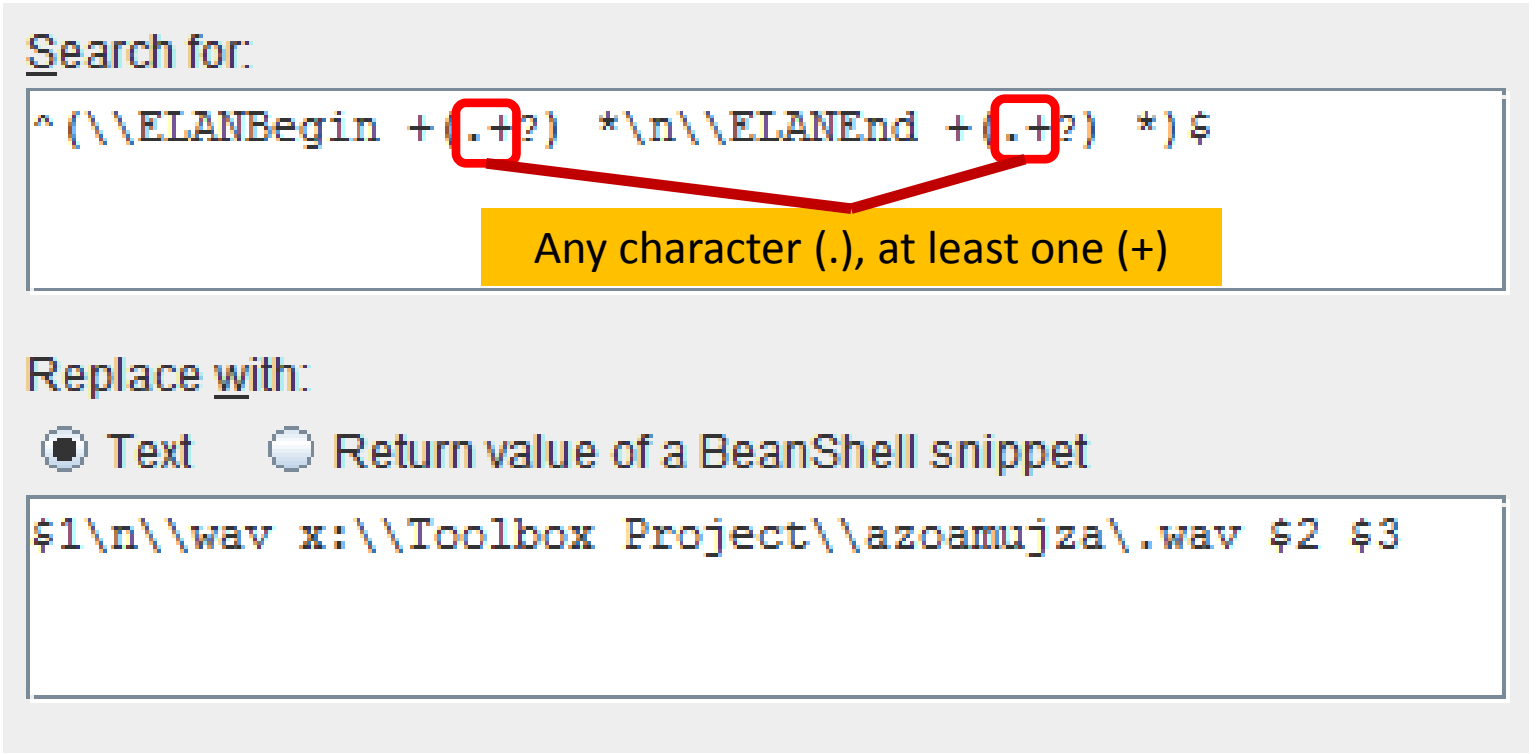
Regular expressions: Modifier characters



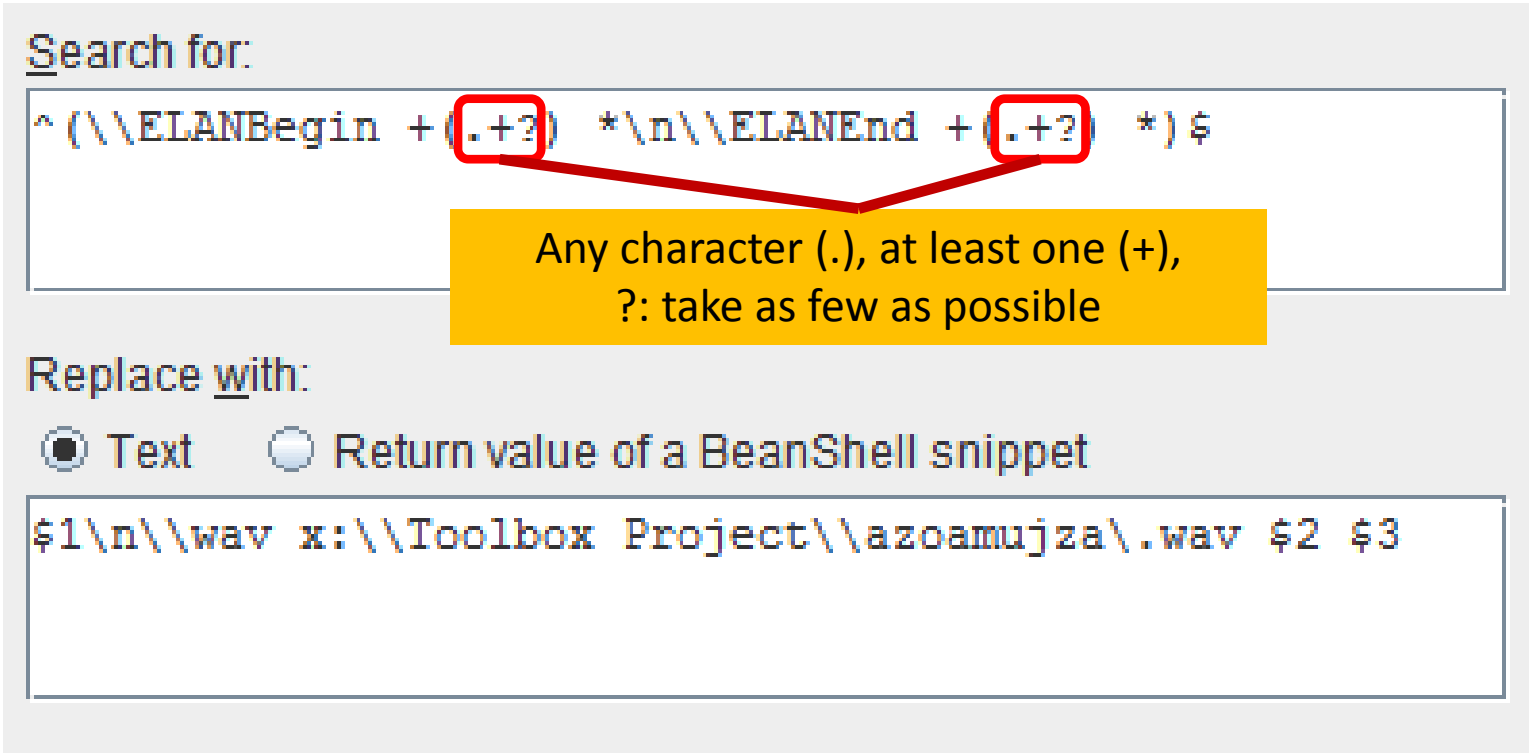
Regular expressions: Modifier characters



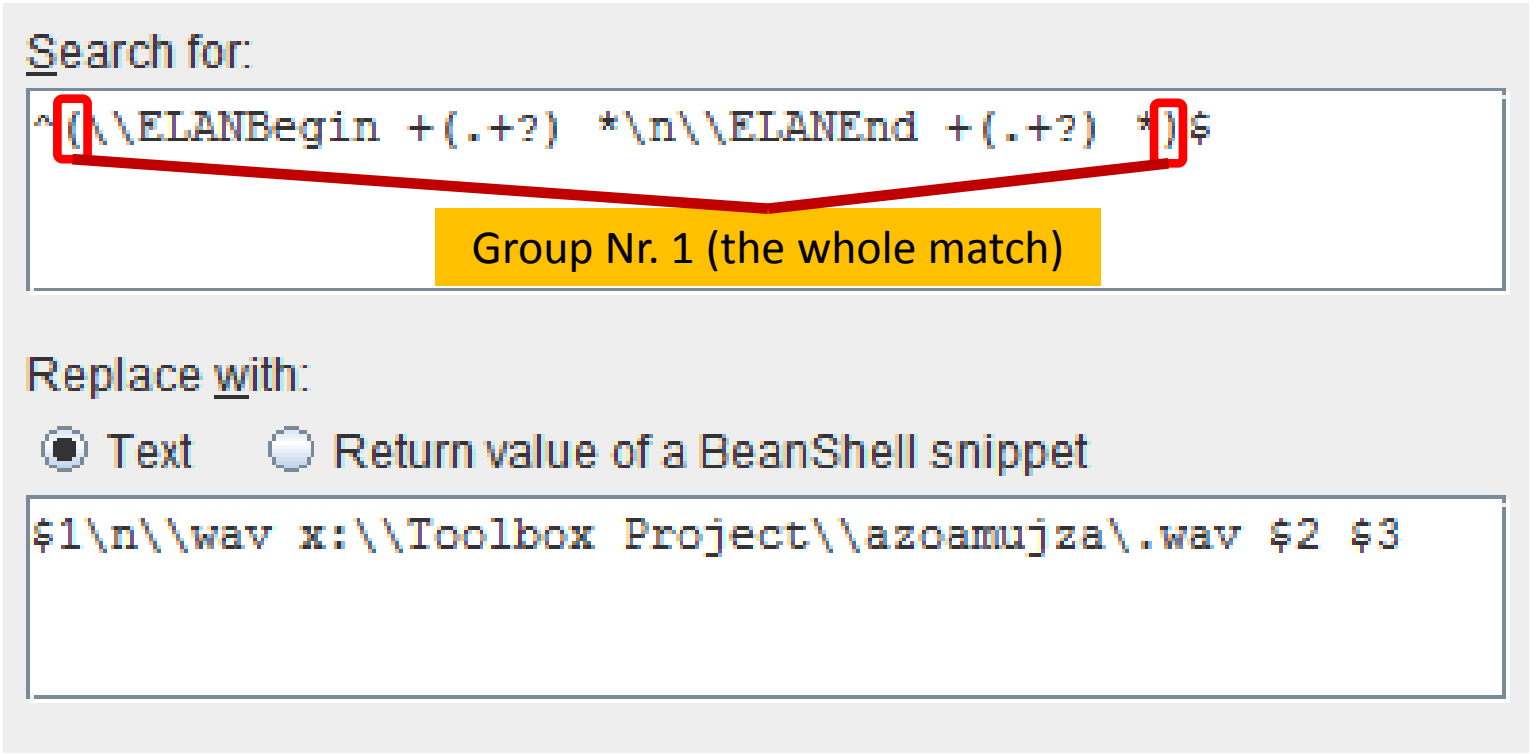
RegExp: Wildcard and modifier characters



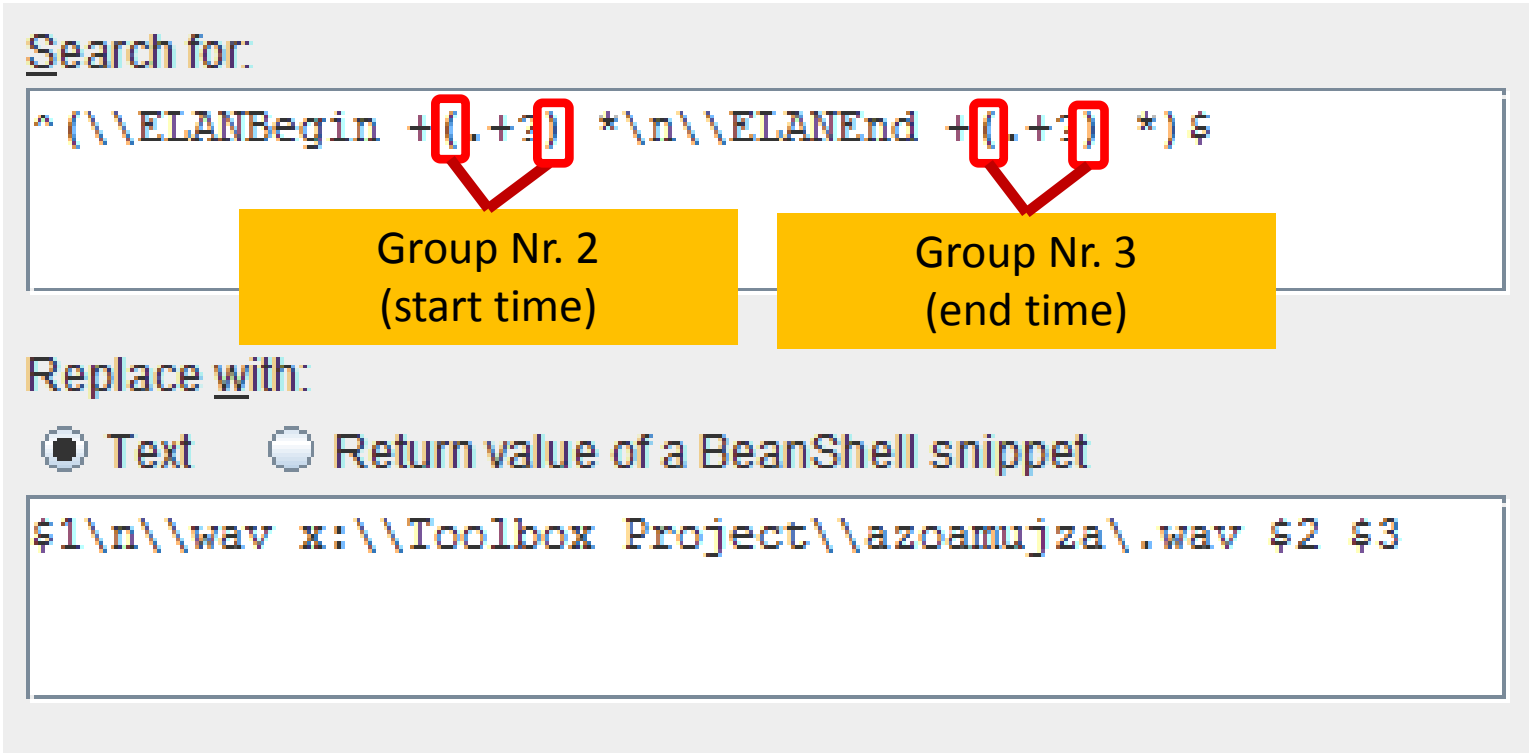
RegExp: "Non-greedy" modifier characters



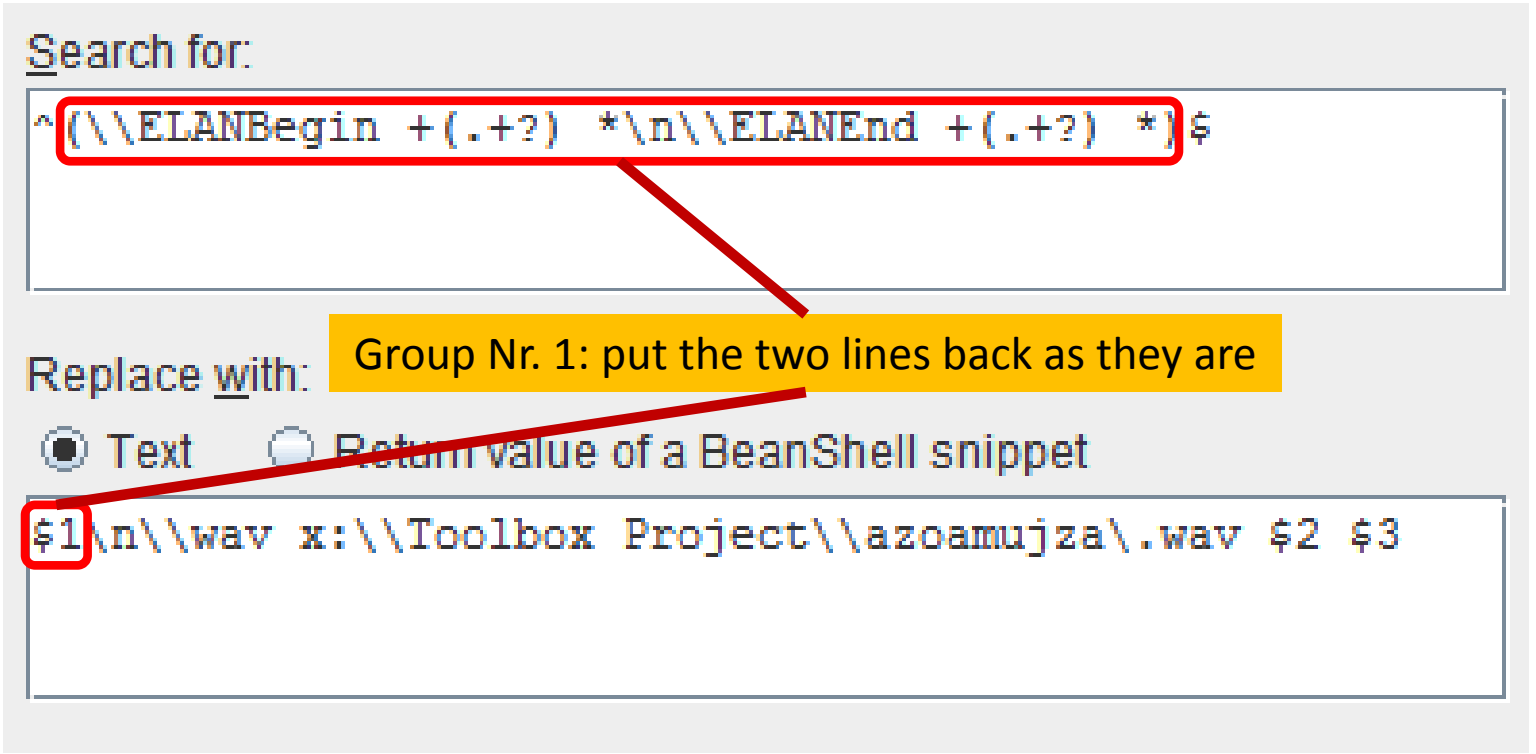
RegExp: Groups in the search expression



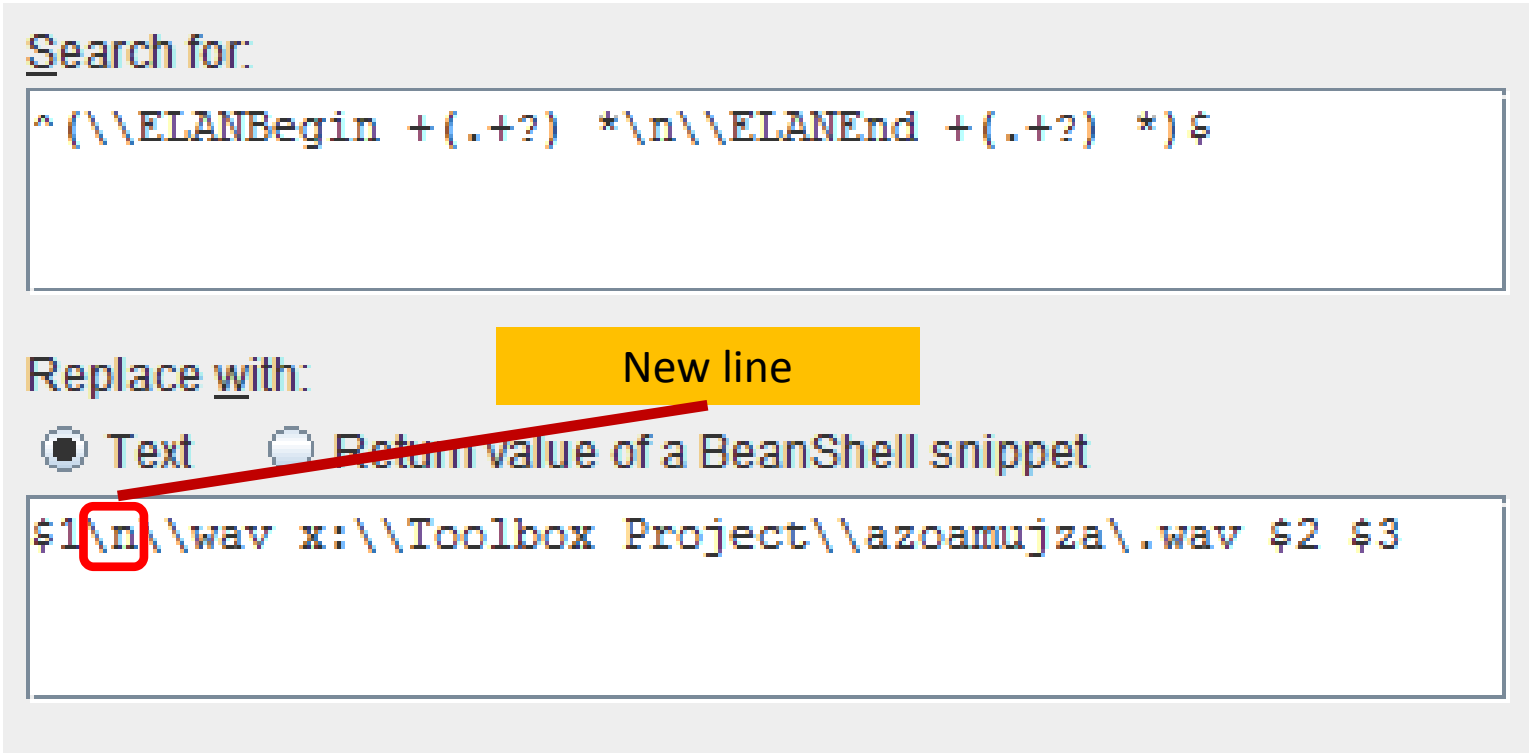
RegExp: Groups in the search expression



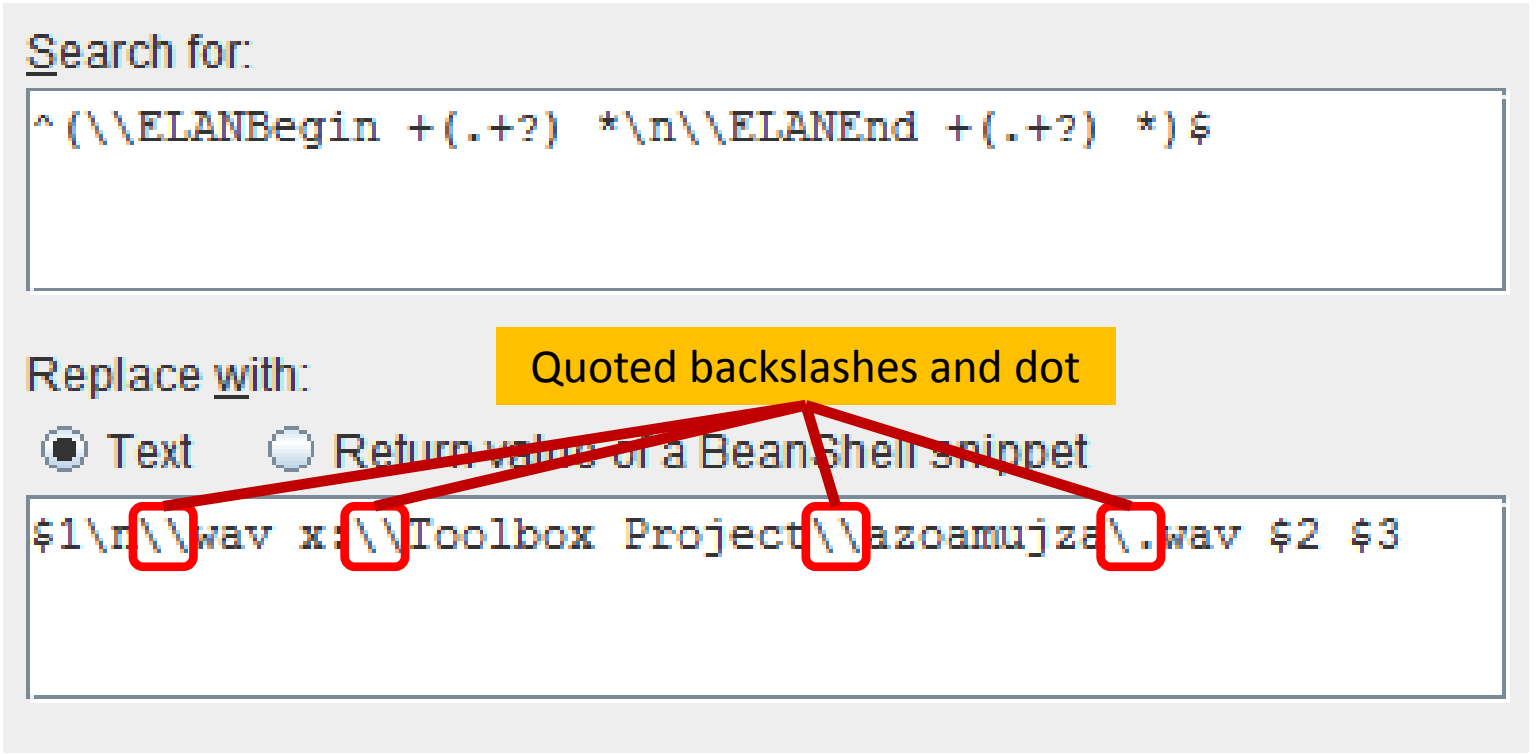
RegExp: Groups in the replace expression



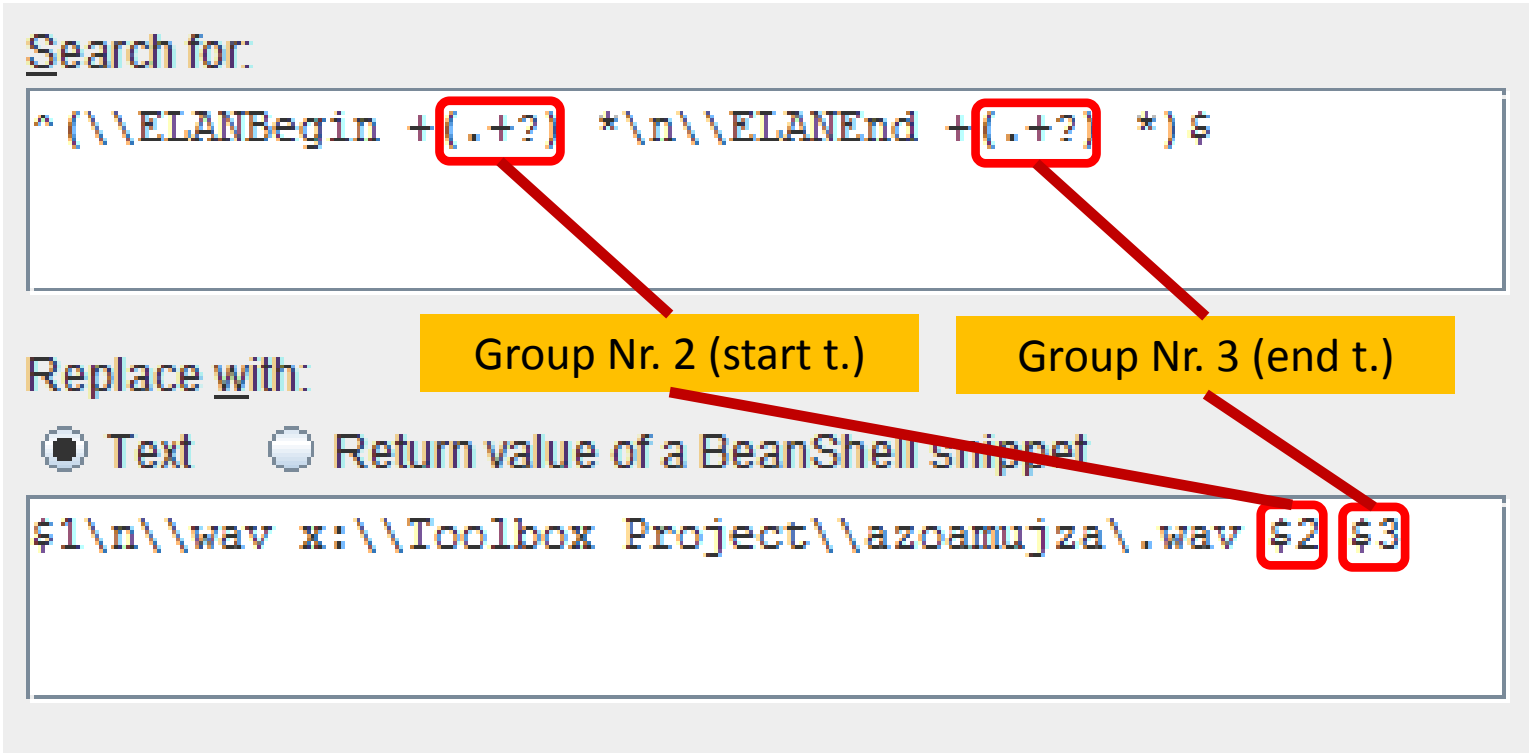
RegExp: Special chars. i. t. replace expr.



RegExp: Quoted chars. i. t. replace expr.



RegExp: Groups in the replace expression



The created the audio fields ***\wav***

X:\...\azoamuija-transcript-bare-ELAN-default.tbt

```
\ref 001
\ELANBegin 0.742
\ELANEnd 7.162
\ELANParticipant Kaluanã
\tx jatãtsu jatã azoamũjza nekozokwawut ne 'a 'e

\ref 002
\ELANBegin 7.162
\ELANEnd 10.698
\ELANParticipant Kaluanã
\tx azoamũjza ekozokotu

\ref 003
\ELANBegin 10.698
\ELANEnd 14.653
\ELANParticipant Kaluanã
\tx nuhiju  pe wian tsãn ekozokotu a'yn, azoamũjza


\ELANMediaURL file:///X:/azoamujza.wav
\ELANMediaMIME audio/x-wav
```

X:\...\azoamuija-transcript-bare-ELAN-default.tbt

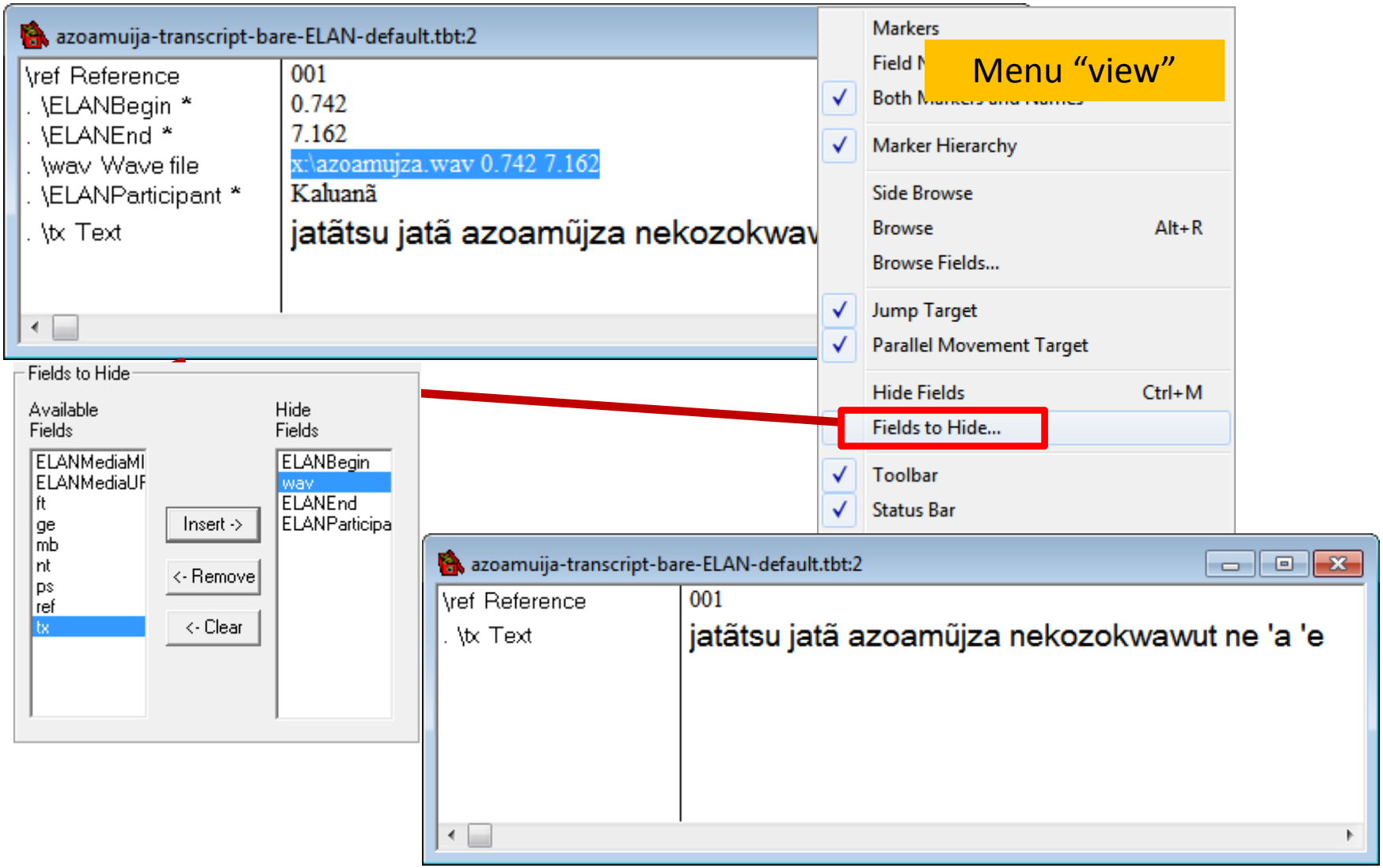
```
\ref 001
\ELANBegin 0.742
\ELANEnd 7.162
\wav x:\azoamujza.wav 0.742 7.162
\ELANParticipant Kaluanã
\tx jatãtsu jatã azoamũjza nekozokwawut ne 'a 'e

\ref 002
\ELANBegin 7.162
\ELANEnd 10.698
\wav x:\azoamujza.wav 7.162 10.698
\ELANParticipant Kaluanã
\tx azoamũjza ekozokotu

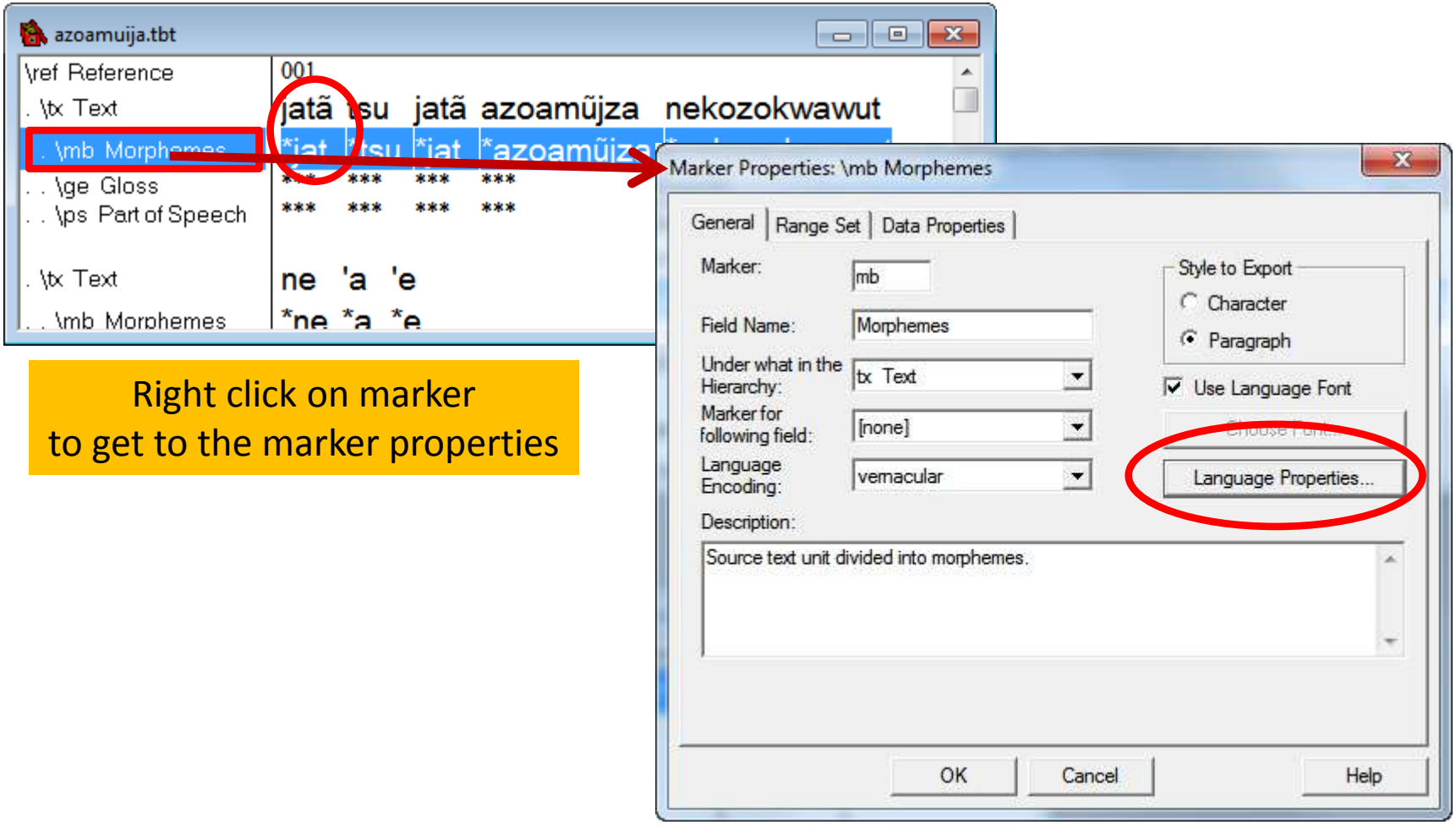
\ref 003
\ELANBegin 10.698
\ELANEnd 14.653
\wav x:\azoamujza.wav 10.698 14.653
\ELANParticipant Kaluanã
\tx nuhiju  pe wian tsãn ekozokotu a'yn, azoamũjza


\ELANMediaURL file:///X:/azoamujza.wav
\ELANMediaMIME audio/x-wav
```

Hiding the fields with “technical data”

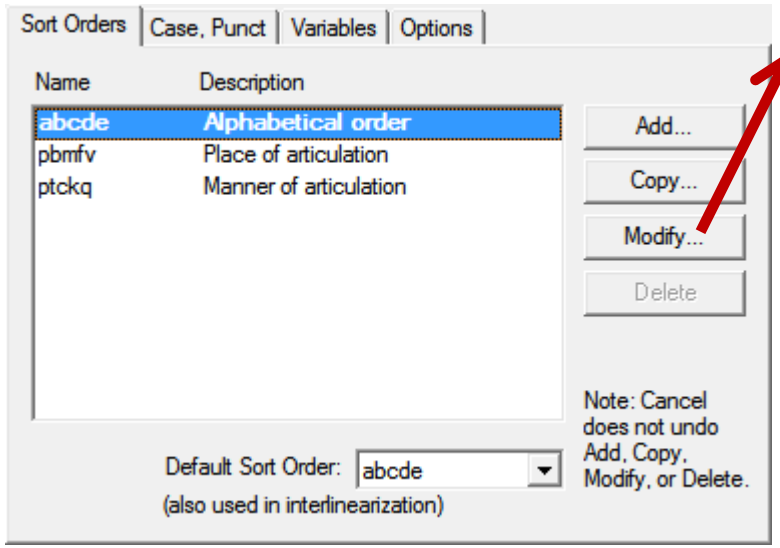


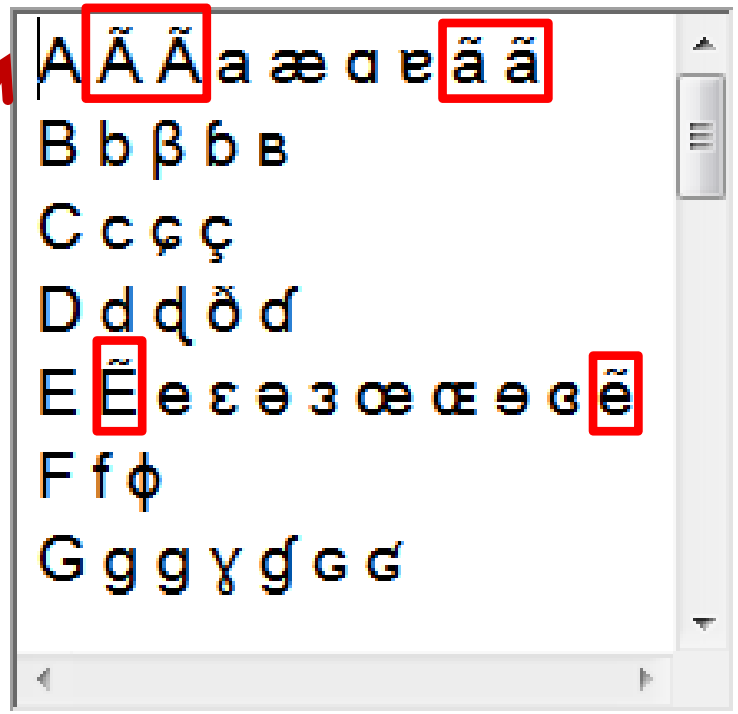
Adjusting the language properties



Right click on marker
to get to the marker properties

Adjusting the language properties





There are two UNICODE representations of ***a + tilde***:

U+00E3 (a+tilde) -- two bytes

U+0061 & **U+0303** (a) & (tilde) -- three bytes

Topics

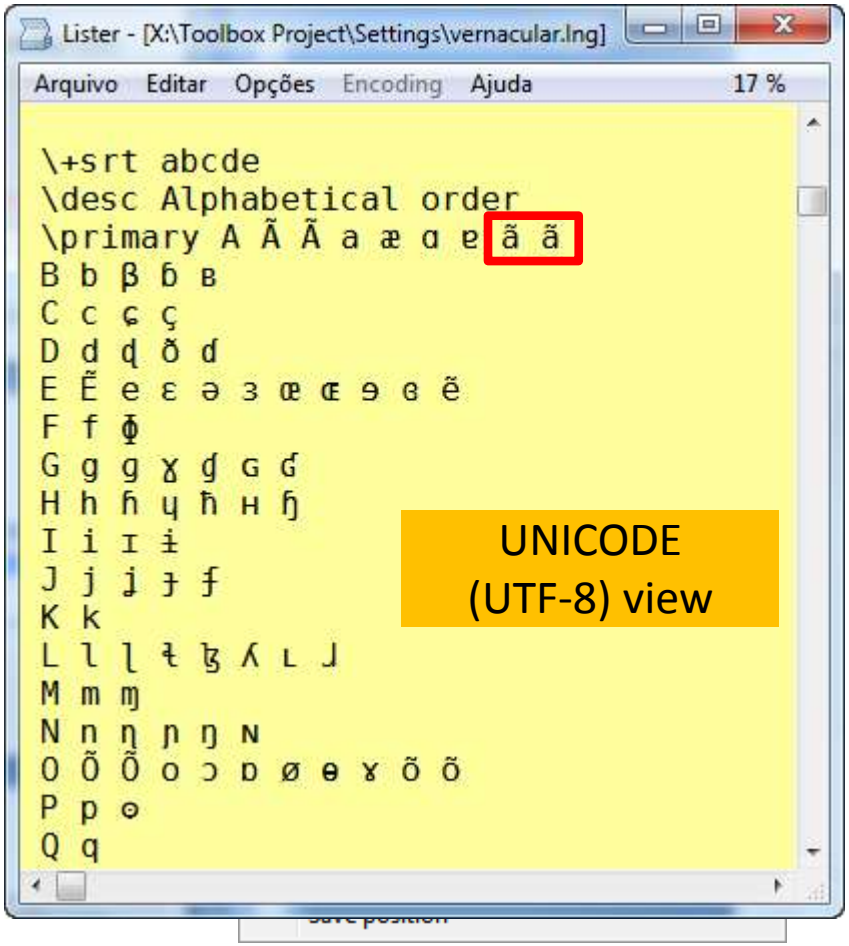
1. Data and Annotation (Theory)
2. Annotation Tools (Overview and Comparison)
3. Intro to Interlinearization (not time-aligned)

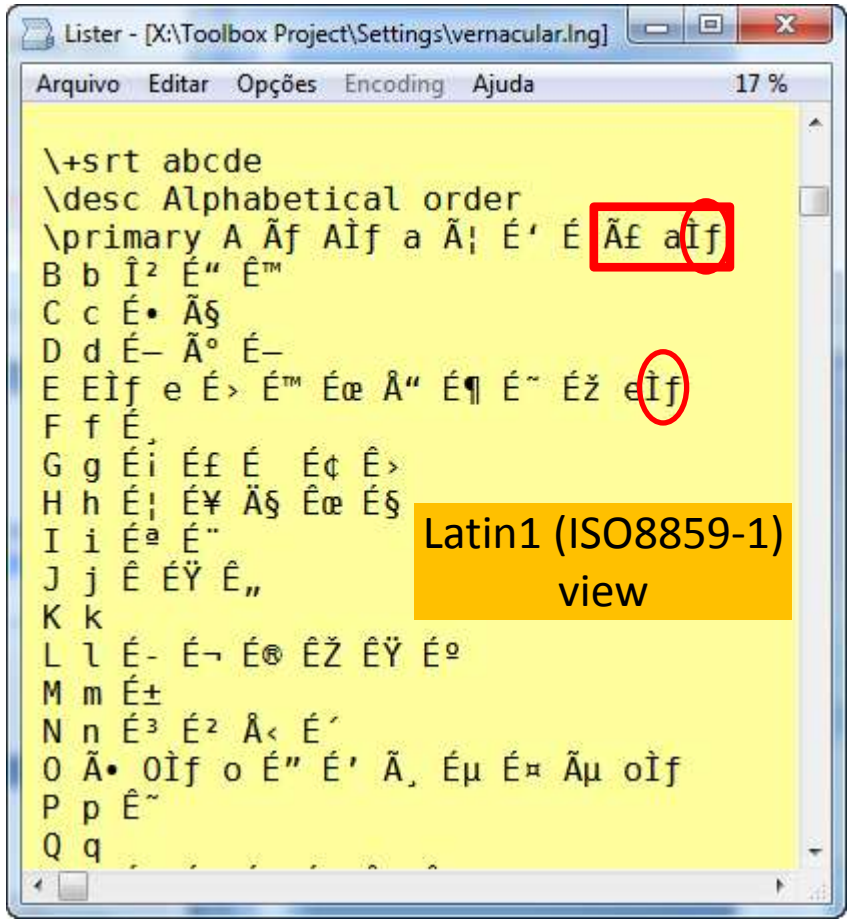
   Excurse: Text- vs. sentence-based databases
4. Time-aligned Annotation
   1. ELAN-Generated Annotation
   2. Play chunks of an audio file in Toolbox

      Excurse: Regular Expressions

      ***Excurse:* UNICODE and UTF-8**
   3. Transcriber-Generated Annotation; Conversions
   4. Round-trip Configuration ELAN--Toolbox

Excurse: UNICODE and UTF-8





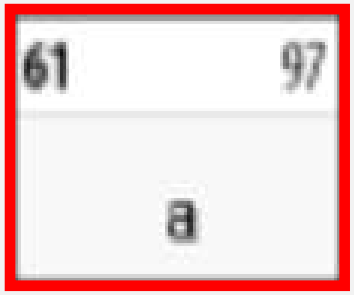
Bits and bytes

- Each letter is, for the computer, a sequence of *bits* – zeros and ones
- The letter “a” is the sequence 01100001, one *byte*, in decimal notation this is the number “97” (= 1*64 + 1*32 + 1)
- Hexadecimal: 0 1 2 3 4 5 6 7 8 9 A B C D E F
- In hexadecimal (basis: 16 instead of 10) this number is “61” (6*16 + 1 = 96 + 1 = 97)

Encodings

- With one byte, one can represent $2^8 = 256$ different letters or other symbols
- *Encoding*: fixed relation of ***number---symbol***
- 256 is enough for upper and lower letters, the numbers, interpunctuation, and a selection of letters with accents, tilde etc.
- The problem is, each language needs different letters, and some need more than 256 -- think of Chinese!

ASCII-encoding: Numbers 0 to 127 (7 bit)



The old Latin1 (ISO8859-1) encoding

Nonstandard (Windows only, so avoid them) character set

128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143
		‚	*ƒ*	„	…	†	‡	ˆ	‰	Š	‹	Œ			
		‚	*ƒ*	„	…	†	‡	ˆ	‰	Š	‹	Œ			
144	145	146	147	148	149	150	151	152	153 trade	154	155	156	157	158	159
	‘	’	“	”	•	–	—	˜	™	š	›	œ			Ÿ
	‘	’	“	”	•	–	—	˜	™	š	›	œ			Ÿ

Extended character set

160 nbsp	161 iexcl ¡ ¡	162 cent ¢ ¢	163 pound £ £	164 curren ¤ ¤	165 yen ¥ ¥	166 brvbar ¦ ¦	167 sect § §	168 uml ¨ ¨	169 copy © ©	170 ordf ª ª	171 laquo « «	172 not ¬ ¬	173 shy	174 reg ® ®	175 macr ¯ ¯
176 deg ° °	177 plusmn ± ±	178 sup2 ² ²	179 sup3 ³ ³	180 acute ´ ´	181 micro µ µ	182 para ¶ ¶	183 middot · ·	184 cedil ¸ ¸	185 sup1 ¹ ¹	186 ordm º º	187 raquo » »	188 frac14 ¼ ¼	189 frac12 ½ ½	190 frac34 ¾ ¾	191 iquest ¿ ¿
192 Agrave À À	193 Aacute Á Á	194 Acirc Â Â	195 Atilde Ã Ã	196 Auml Ä Ä	197 Aring Å Å	198 AElig Æ Æ	199 Ccedil Ç Ç	200 Egrave È È	201 Eacute É É	202 Ecirc Ê Ê	203 Euml Ë Ë	204 Igrave Ì Ì	205 Iacute Í Í	206 Icirc Î Î	207 Iuml Ï Ï
208 ETH	209 Ntilde	210 Ograve	211 Oacute	212 Ocirc	213 Otilde	214 Ouml	215 times	216 Oslash	217 Ugrave	218 Uacute	219 Ucirc	220 Uuml	221 Yacute	222 THORN	223 szlig

UNICODE

- Unicode is not much more than an assignment of one unique name and one unique number to ANY letter or symbol in ANY language
- The number has a "**U+**"-prefix and is hexadecimal
- For example, the phonetic symbol "**ɔ**" is in UNICODE the character ***U+0254*** (=596), and is called ***LATIN LETTER SMALL OPEN O***
- The basic letters (ASCII) are the same as before in Latin1: "**a**" = ***U+0061*** (=97) with the name ***LATIN SMALL LETTER A***

Fonts

Whether and how a character (a number) is graphically rendered / displayed depends on the ***font***

Some have no "*glyph*" (image) at all for a given character

ɔ Calibri

ɔ Arial

ɔ Times new Roman (serif, UNICODE)

□□ Marlett (UNICODE, but has no glyph)

❄ Absalom (not a UNICODE font)

Keyboard

How to enter UNICODE characters to your program? This depends on the program and operation system. Here tips for Windows.

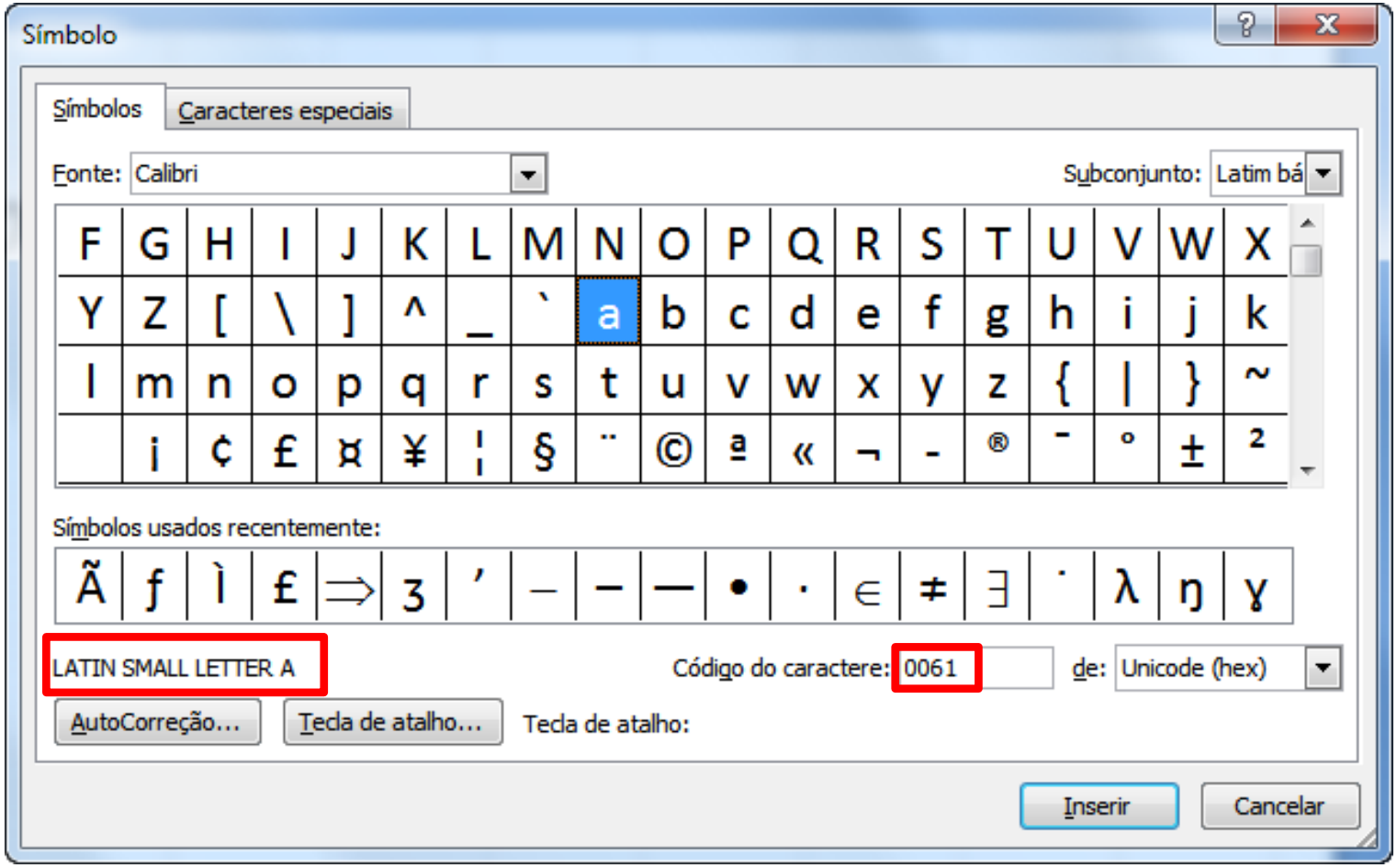
For phonetics I recommend the free **IPA Unicode 5.1 (ver. 1.2) MSK Keyboard** http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=UniIPAKeyboard&_sc=1
Drawback: it presupposes the US Keyboard layout

For sporadic access to arbitrary UNICODE characters, there is a little practical tool at http://www.fileformat.info/tool/unicodeinput/

UTF (Unicode Transformation Format) 8

- In order to represent all the tousands of UNICODE characters, one would need three bytes for each character -- that is not practical
- Different UNICODE-encodings exist
- A very popular and practical one is UTF-8
- UTF-8 is a “compromise” character encoding that can be as compact as ASCII (if the file is just plain English text) but can also contain any UNICODE characters -- some have four bytes

The simple UNICODE character ***a***



The simple UNICODE character ***a***

Unicode Character 'LATIN SMALL LETTER A' (U+0061)

Image
a

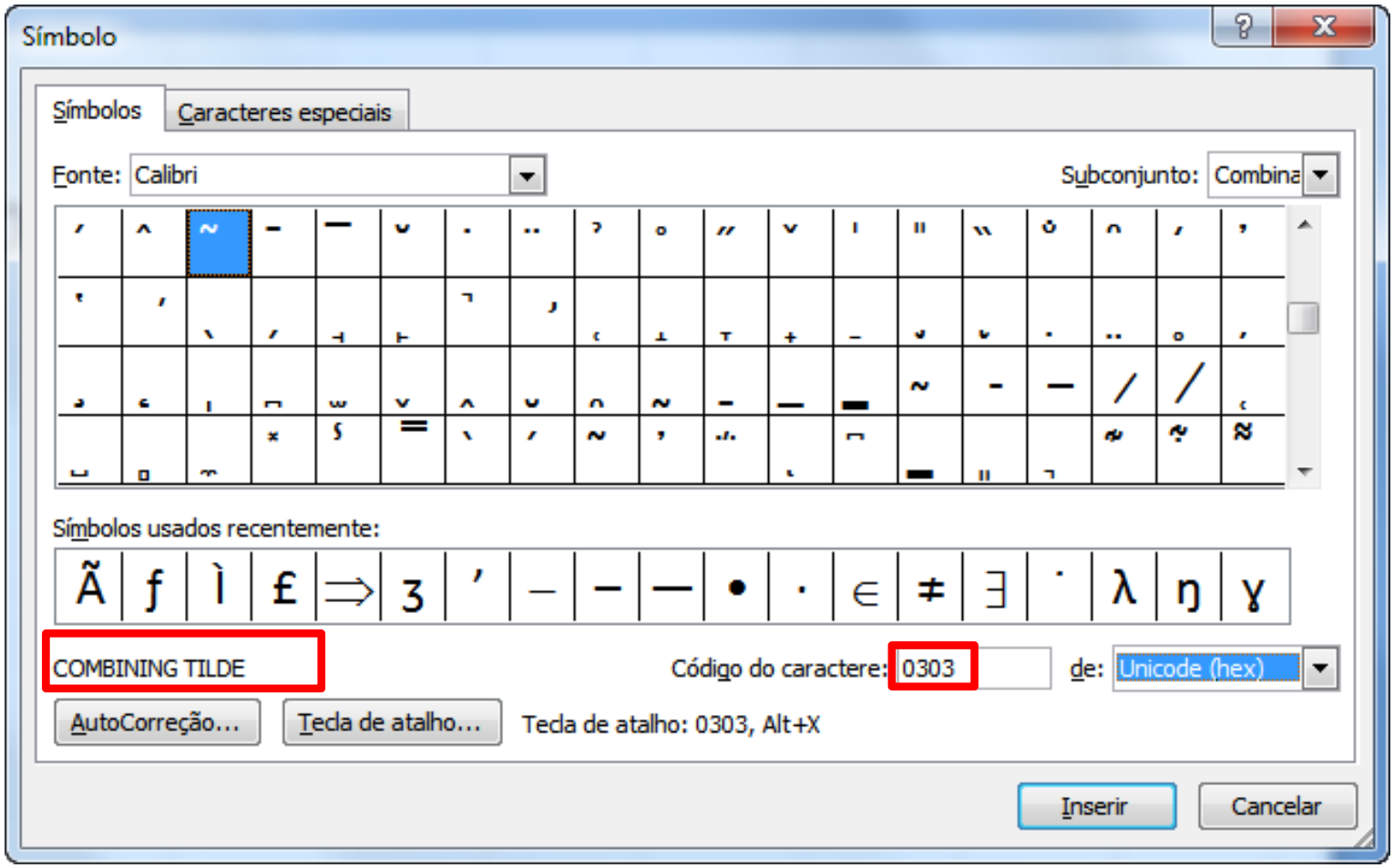
Encodings	
HTML Entity (decimal)	a
HTML Entity (hex)	a
How to type in Microsoft Windows	Alt +61 Alt 097 Alt 97
UTF-8 (hex)	0x61 (61)
UTF-8 (binary)	01100001
UTF-16 (hex)	0x0061 (0061)
UTF-16 (decimal)	97
UTF-32 (hex)	0x00000061 (61)
UTF-32 (decimal)	97
C/C++/Java source code	"\u0061"
Python source code	u"\u0061"
	More...

UTF-8 uses
one byte
to represent
this ***character***:

0x61 = 97 =
01100001

In Latin1, this number is **a**, *too.*

The combining UNICODE character ~



The combining UNICODE character ~

Unicode Character 'COMBINING TILDE' (U+0303)

Image
~

Encodings	
HTML Entity (decimal)	̃
HTML Entity (hex)	̃
How to type in Microsoft Windows	Alt +303
UTF-8 (hex)	0xCC 0x83 (cc83)
UTF-8 (binary)	11001100:10000011
UTF-16 (hex)	0x0303 (0303)
UTF-16 (decimal)	771
UTF-32 (hex)	0x00000303 (303)
UTF-32 (decimal)	771
C/C++/Java source code	"\u0303"
Python source code	u"\u0303"
	More...

UTF-8 uses
two bytes
to represent
this ***character***:

0xCC = 204 =
11001100 > Ì

0x83 = 131 =
10000011 > ƒ

UNICODE UTF-8 ***a & tilde*** *(sequence)*

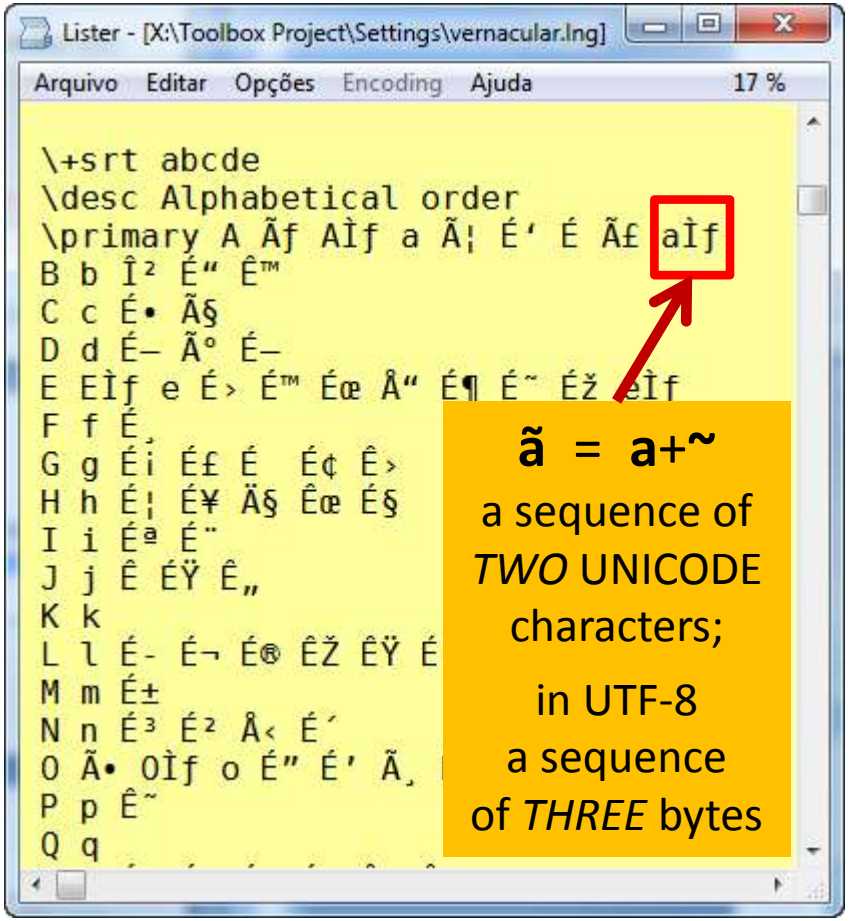
(a) & (tilde):

"LATIN SMALL LETTER A" &

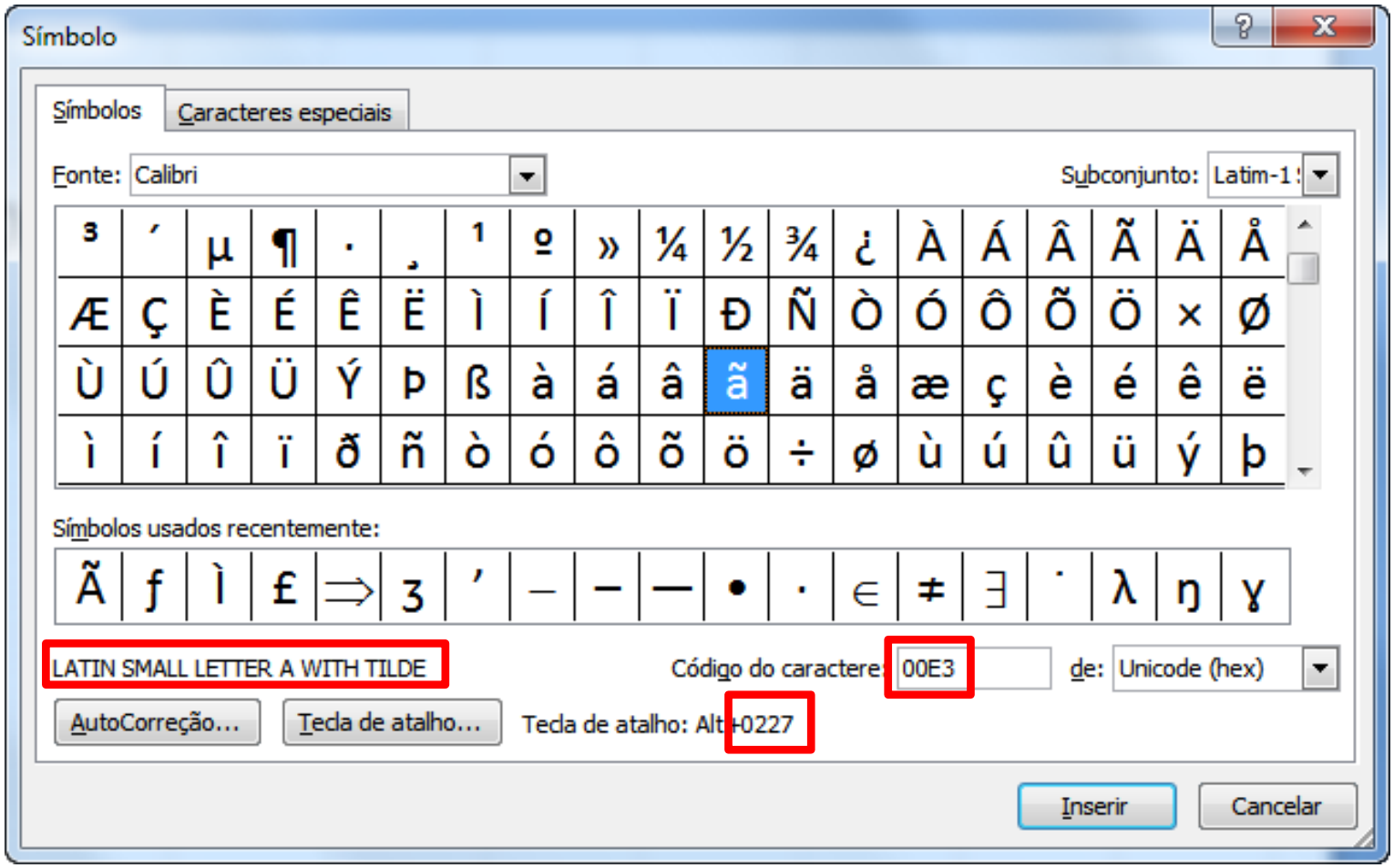
"COMBINING TILDE"

UNICODE: **U+0061** (=97) &
U+0303 (=771)

UTF-8: **0x61** & **0xCC 0x83**
= 97 (Latin1: **a**) &
204 131 (Latin1: **Ì ƒ**)
= 01100001 &
11001100 10000011



The complex UNICODE character ***ã***



*The complex UNICODE character **ã***

Unicode Character 'LATIN SMALL LETTER A WITH TILDE' (U+00E3)

Image
ã

Encodings	
HTML Entity (decimal)	ã
HTML Entity (hex)	ã
HTML Entity (named)	ã
How to type in Microsoft Windows	Alt +E3 Alt 0227
UTF-8 (hex)	0xC3 0xA3 (c3a3)
UTF-8 (binary)	11000011:10100011
UTF-16 (hex)	0x00E3 (00e3)
UTF-16 (decimal)	227
UTF-32 (hex)	0x000000E3 (e3)
UTF-32 (decimal)	227
C/C++/Java source code	"\u00E3"
Python source code	u"\u00E3"
	More...

UTF8 uses
two bytes
to represent
this ***character***:

0xC3 = 195 =
11000011 > **Ã**

0xA3 = 163 =
10100011 > **£**

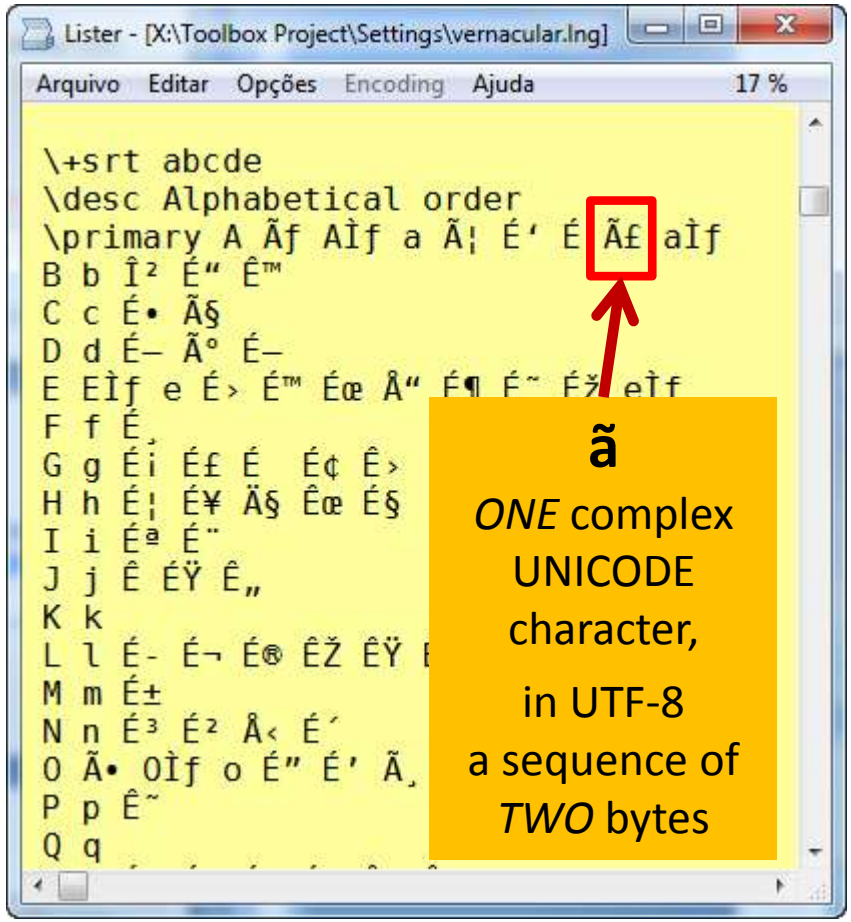
UNICODE UTF-8 ***a+tilde*** *(combined)*

(a+tilde):

"LATIN SMALL LETTER A
WITH TILDE"

UNICODE: **U+00E3** (=227)

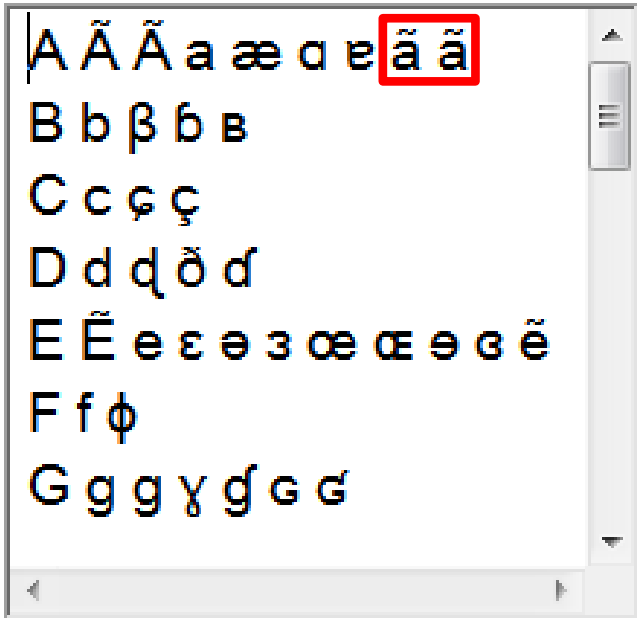
UTF-8: **0xC3 0xA3**
= 195 160 (Latin1: **Ã £**)
= 11000011 10100011

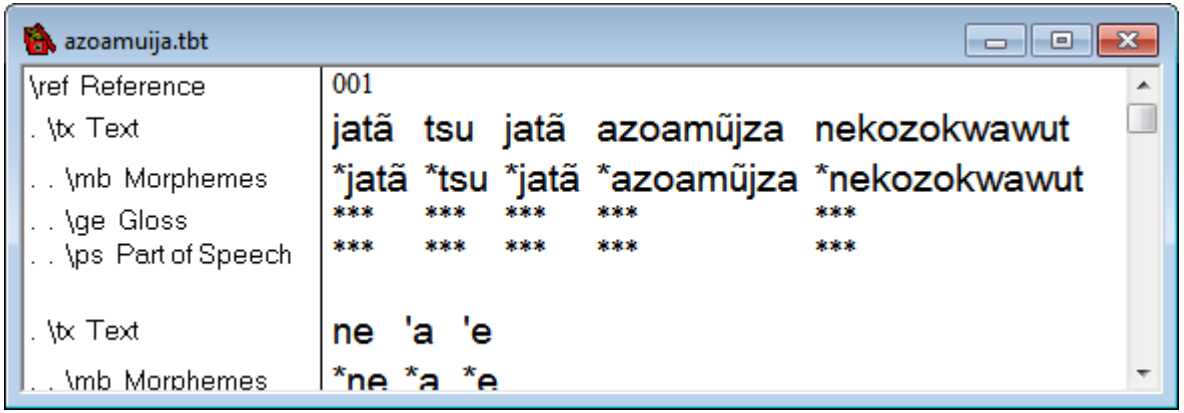


Adjusting the language properties

It is important to enter
ALL possible UNICODE representations
of the letters of the language
for interlinearization to work

But it is also much safer
to use **always the same representation** for any letter





Almost identical looking characters

Be careful with (almost) identical looking characters (depending on the font). For instance, for **ejectives** or the **glottal stop**, use the ***modifier letter apostrophe***, not the *apostrophe* and also not the *right single quotation mark*, although in most fonts they look (almost) the same!

Glyph	Name	UNICODE	Decimal	UTF-8	Bytes in Latin1
'	Apostrophe	U+0027	39	0x27 39	'
ʼ	Modifier letter apostrophe	U+02BC	700	0xCA 0xBC 202 188	Ê ¼
’	Right single quotation mark	U+2019	8217	0xE2 0x80 0x99 226 126 153	Â € ™

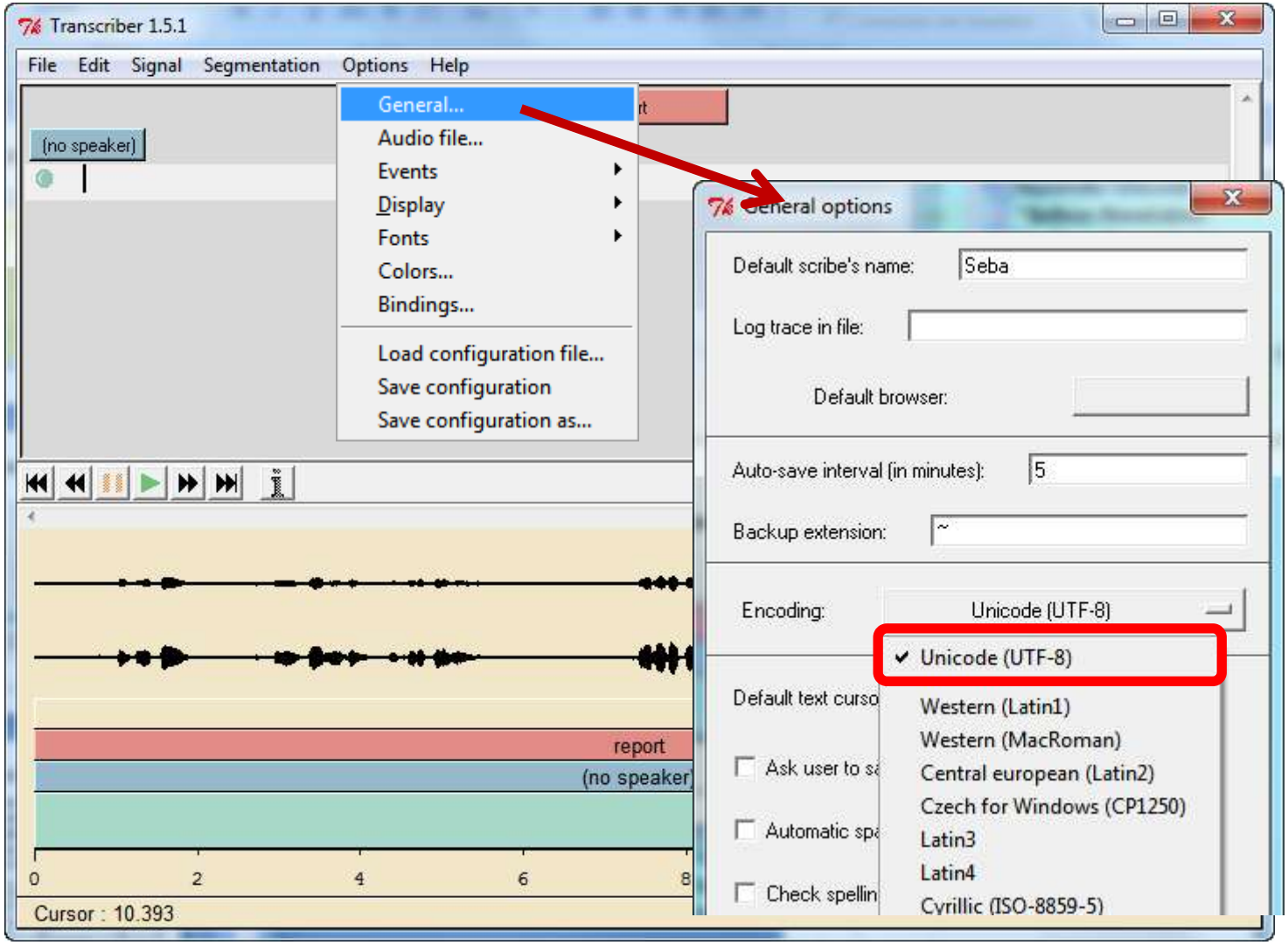
Topics

1. Data and Annotation (Theory)
2. Annotation Tools (Overview and Comparison)
3. Intro to Interlinearization (not time-aligned)
   Excurse: Text- vs. sentence-based databases
4. Time-aligned Annotation
   1. ELAN-Generated Annotation
   2. Play chunks of an audio file in Toolbox
      Excurse: Regular Expressions
      Excurse: UNICODE and UTF-8
   3. **Transcriber-Generated Annotation; Conversions**
   4. Round-trip Configuration ELAN--Toolbox

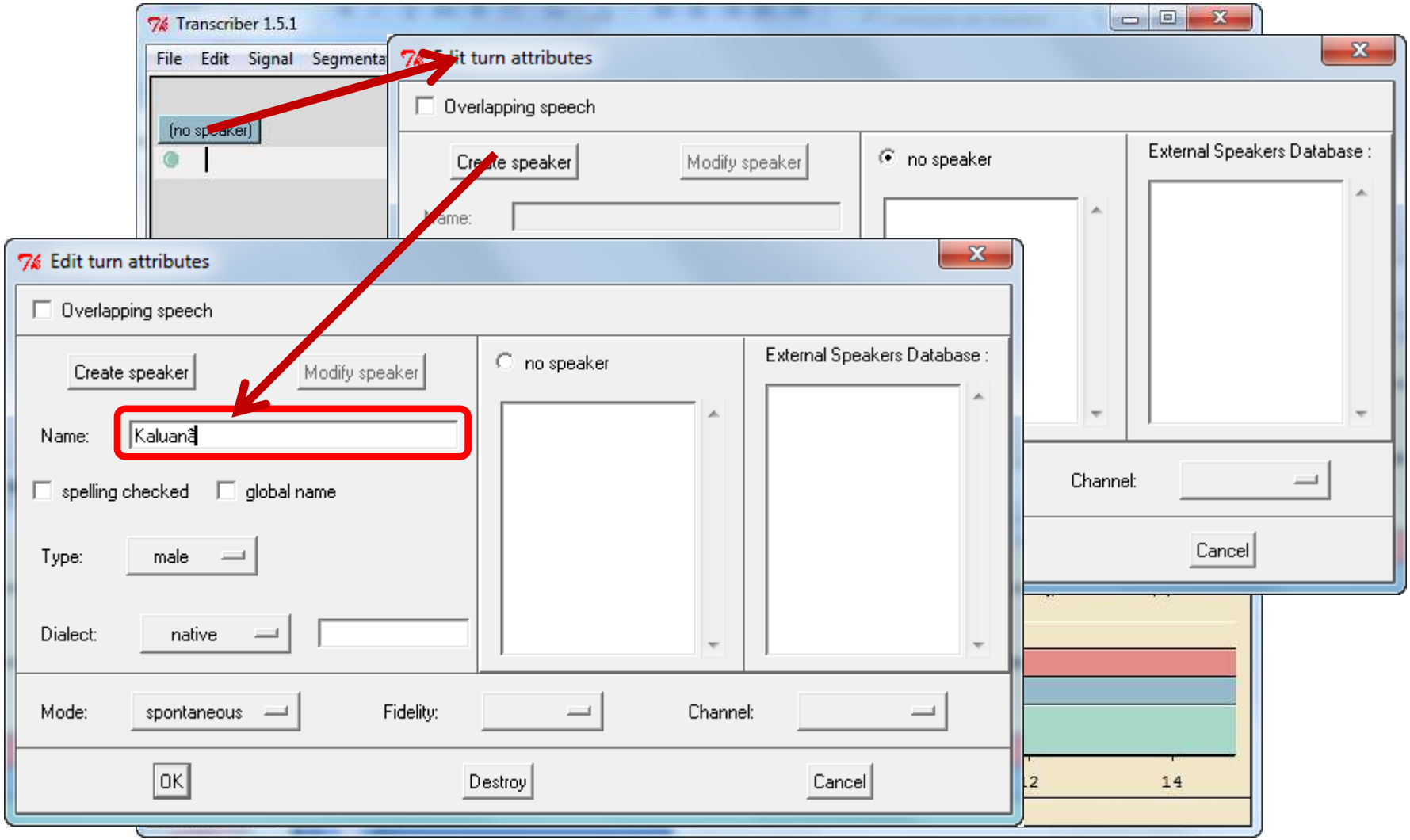
Segmenting and transcribing in Transcriber

- Until recently, the major advantage (ease of use) of Transcriber outweighed its major disadvantage (no UNICODE input).
- Now, ELAN has the new *transcription mode*, and is a viable alternative for efficient segmenting and transcribing even for novice users. Still, Transcriber may be an alternative, and has been used by many documentation projects.

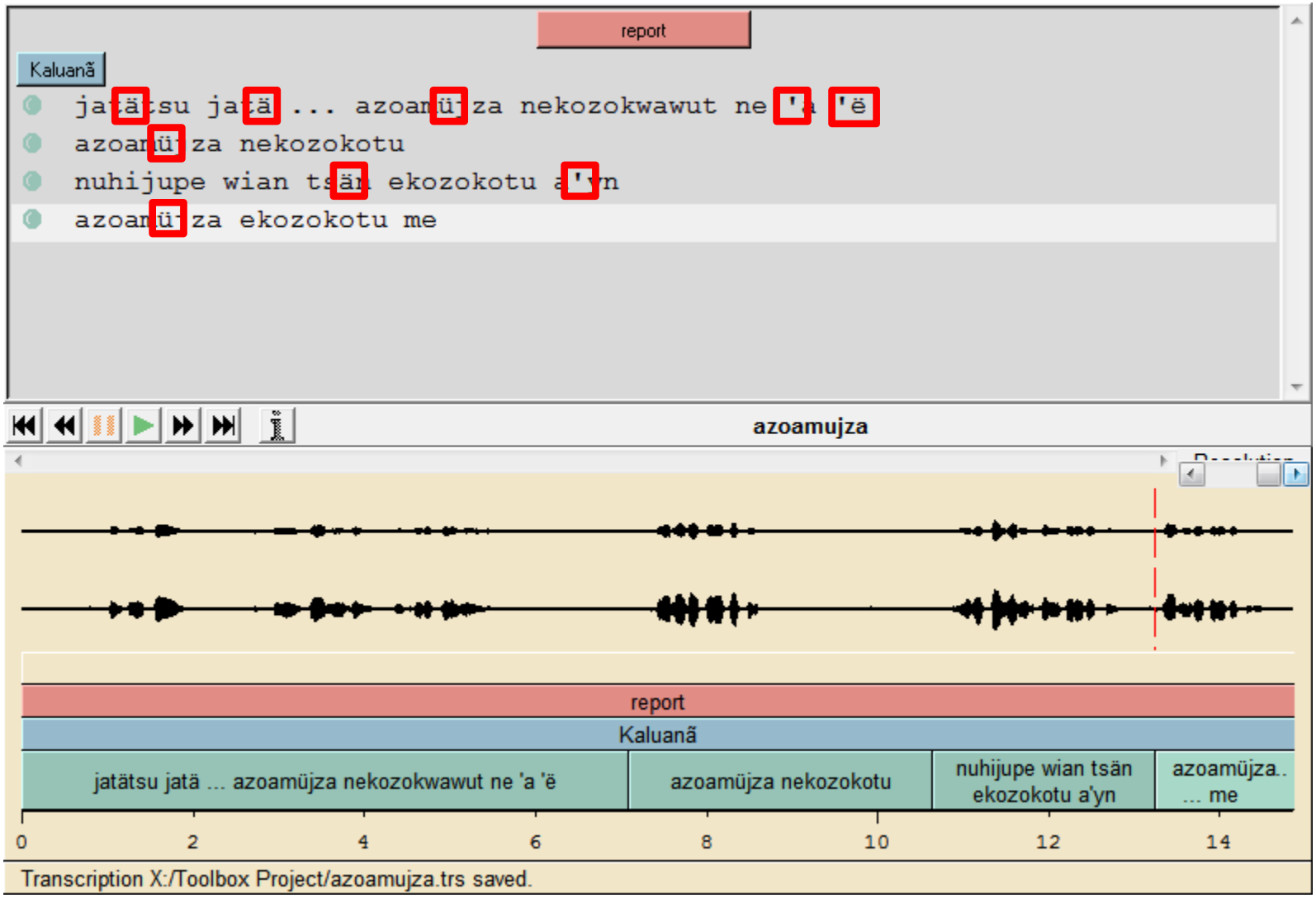
Transcriber: UNICODE encoding



Transcriber: Create speaker



Transcription with Transcriber



*Transcriber generated XML file (**.trs**)*

Lister - [X:\Toolbox Project\azoamujza.trs]

Arquivo Editar Opções Encoding Ajuda 100 %

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE Trans SYSTEM "trans-14.dtd">
<Trans scribe="Seba" audio_filename="azoamujza" version="1" version_date="110903">
<Speakers>
<Speaker id="spk1" name="Kaluanã" check="no" type="male" dialect="native" accent="'
</Speakers>
<Episode>
<Section type="report" startTime="0" endTime="14.8607709751">
<Turn startTime="0" endTime="14.8607709751" speaker="spk1" mode="spontaneous">
<Sync time="0"/>
jatätsu jatä ... azoamüjza nekozokwawut ne 'a 'ë
<Sync time="7.104"/>
azoamüjza nekozokotu
<Sync time="10.649"/>
nuhijupe wian tsän ekozokotu a'yn
<Sync time="13.254"/>
azoamüjza ekozokotu me
</Turn>
</Section>
</Episode>
</Trans>
```

From Transcriber to Toolbox

There are three principle possibilities to import Transcriber files into Toolbox:

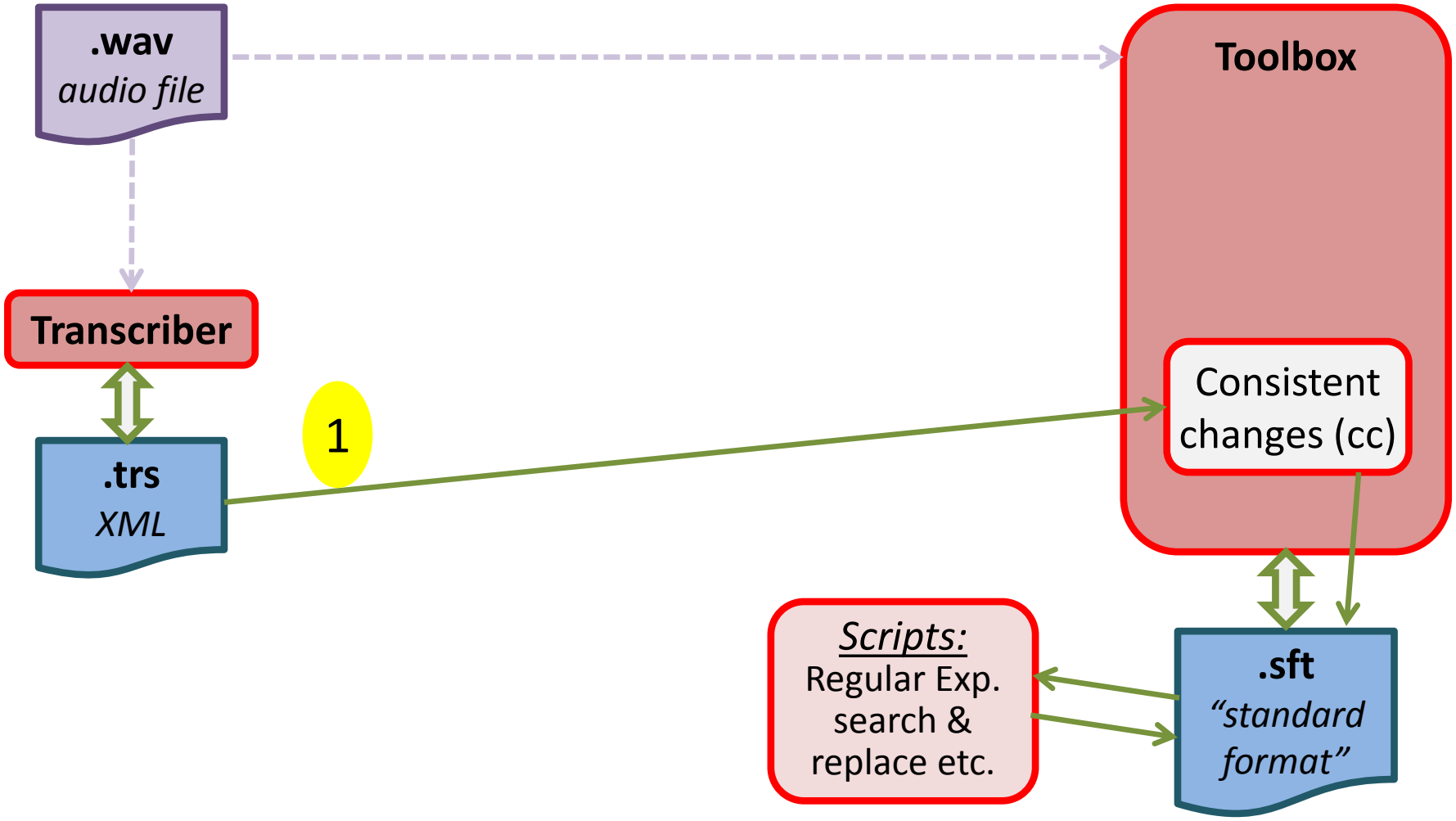
1. “Direct import” of Toolbox (using a CC table)
2. Using a converter (ECONV, Linguistic Software Cv.)
3. Via ELAN

None of these procedures is ideal

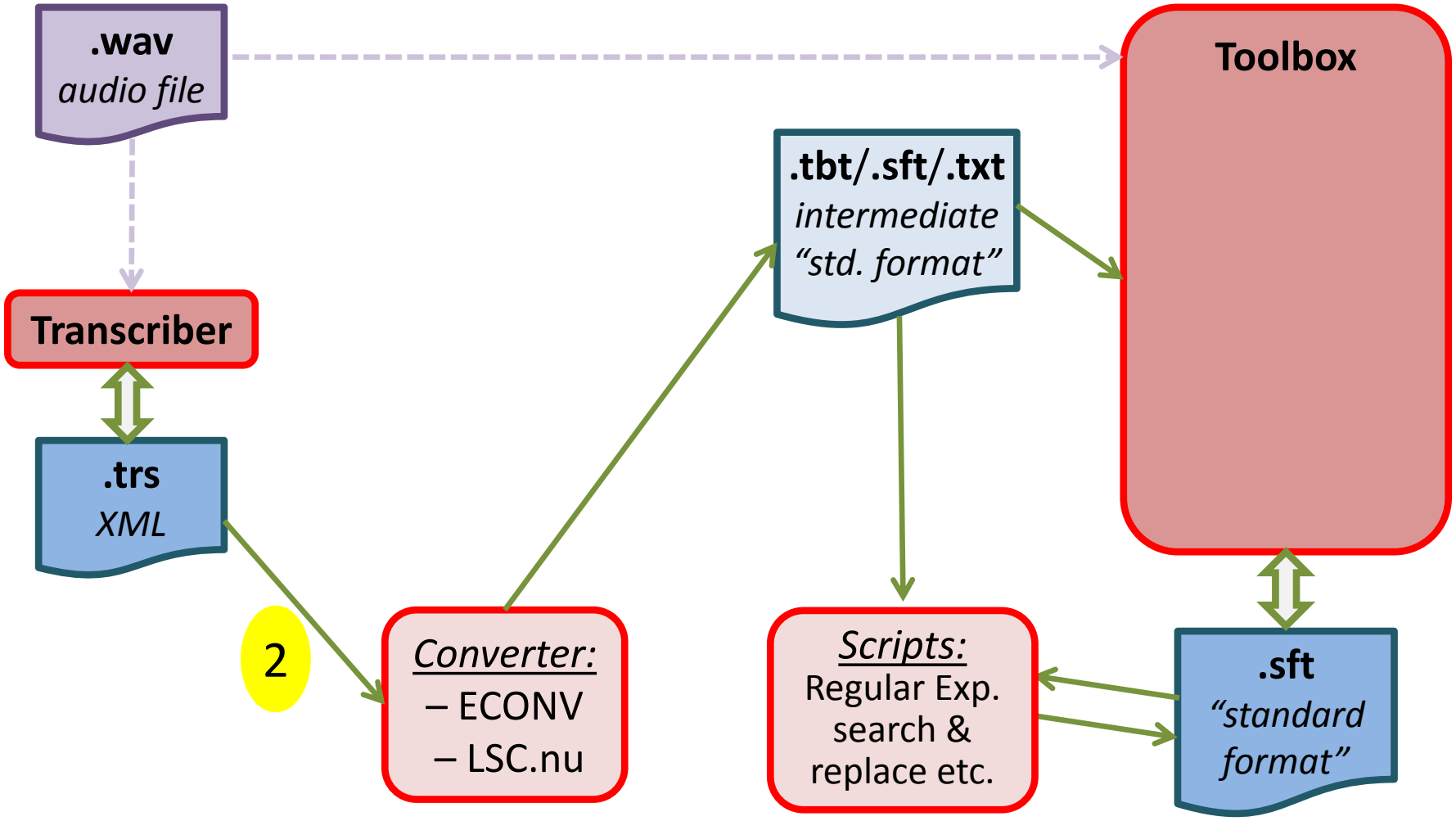
Additional scripts will almost always be needed

In any case, one needs to convert the preliminary makeshift characters to UNICODE characters, either before or after converting to Standard Format

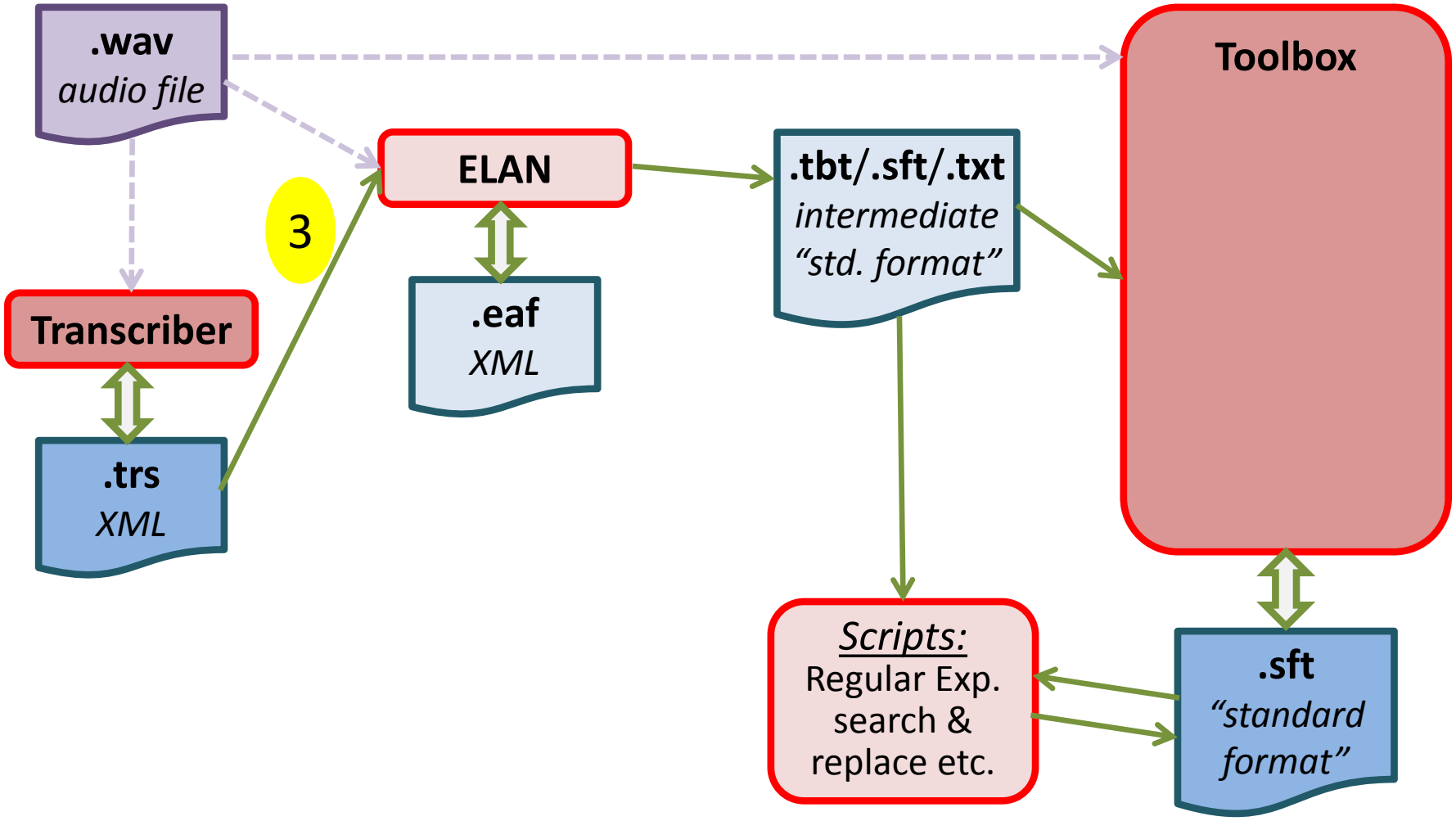
1: Direct import in Toolbox (cc)



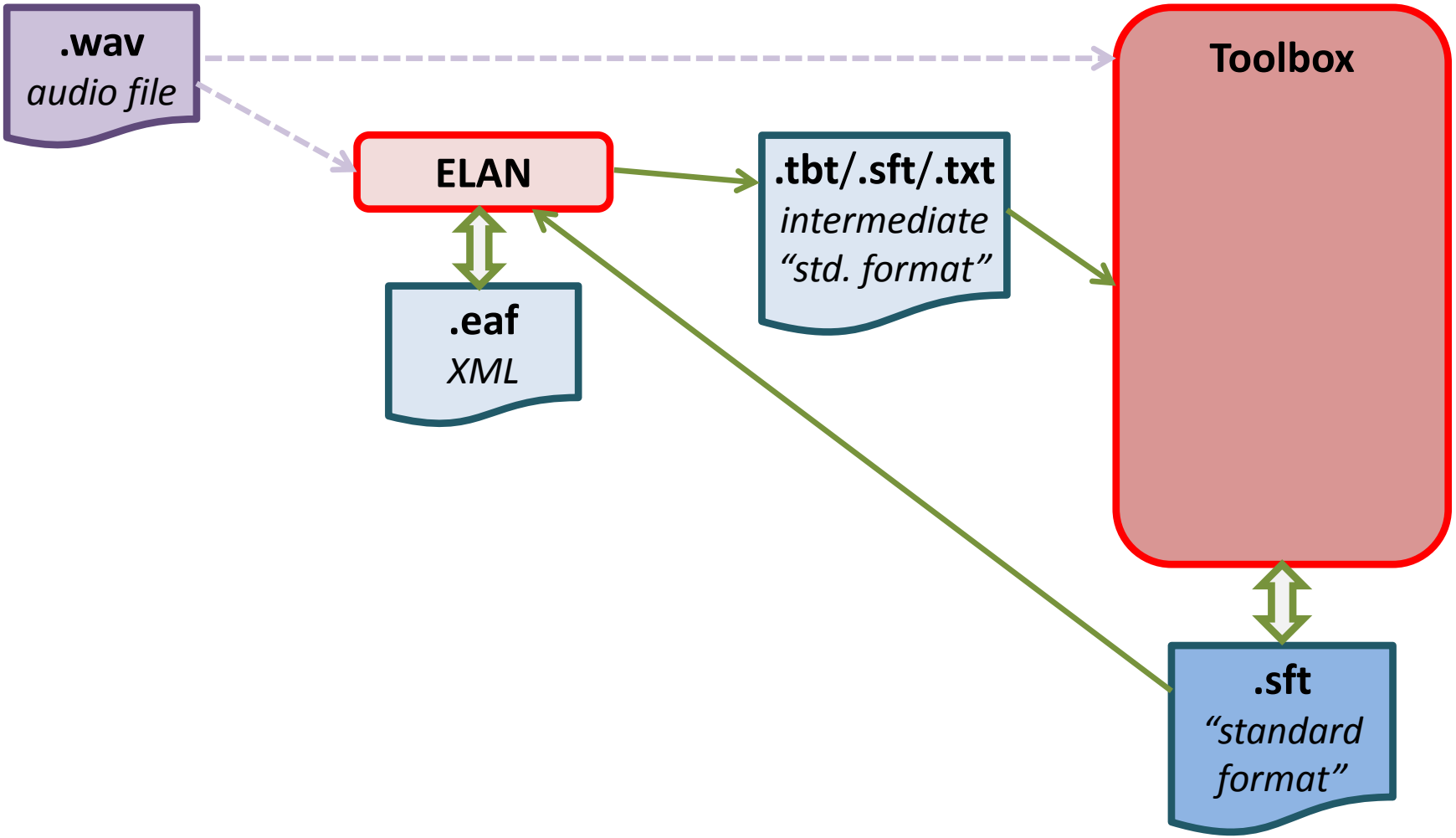
2: Using an external converter



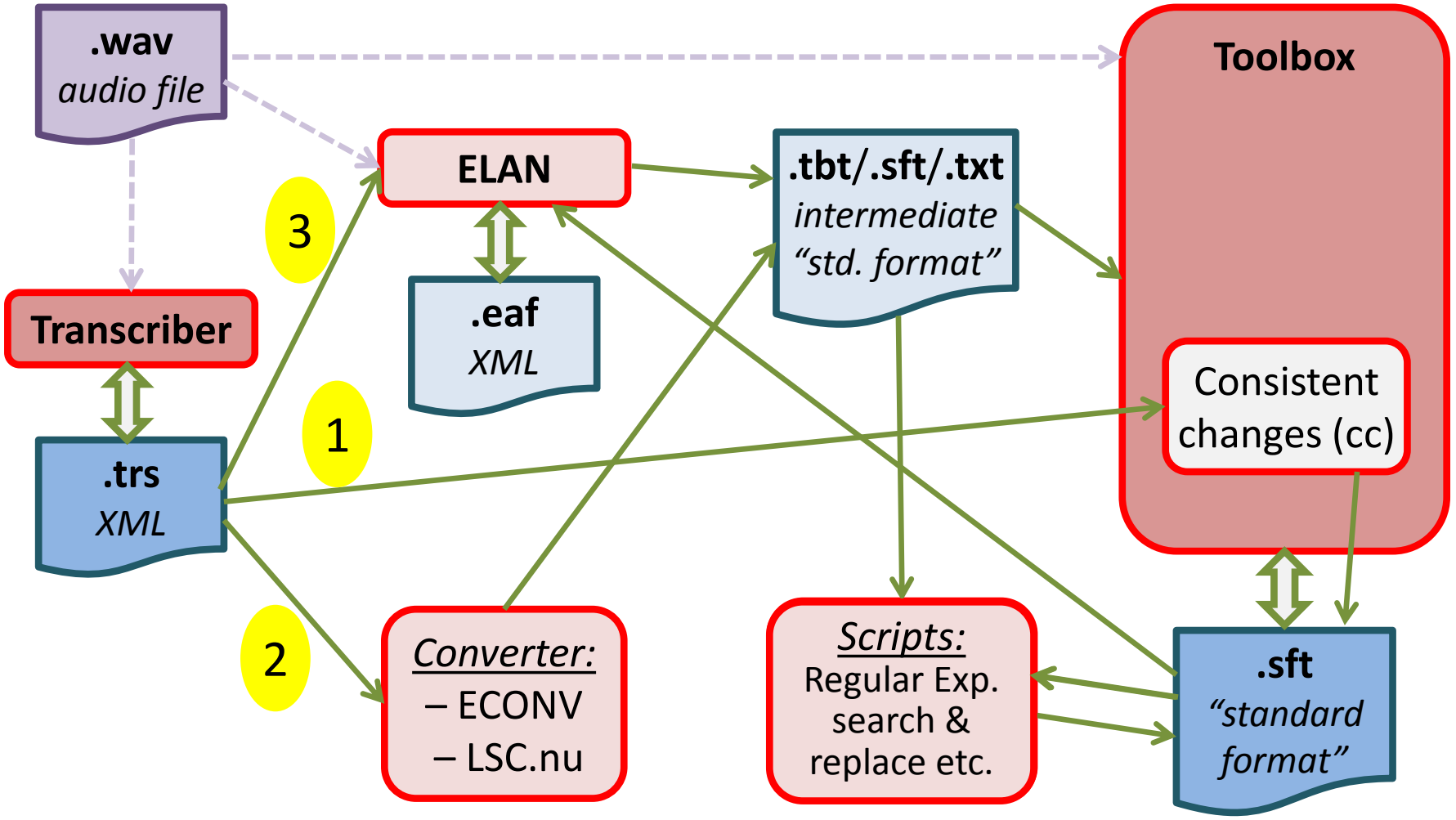
3: Using ELAN as a converter



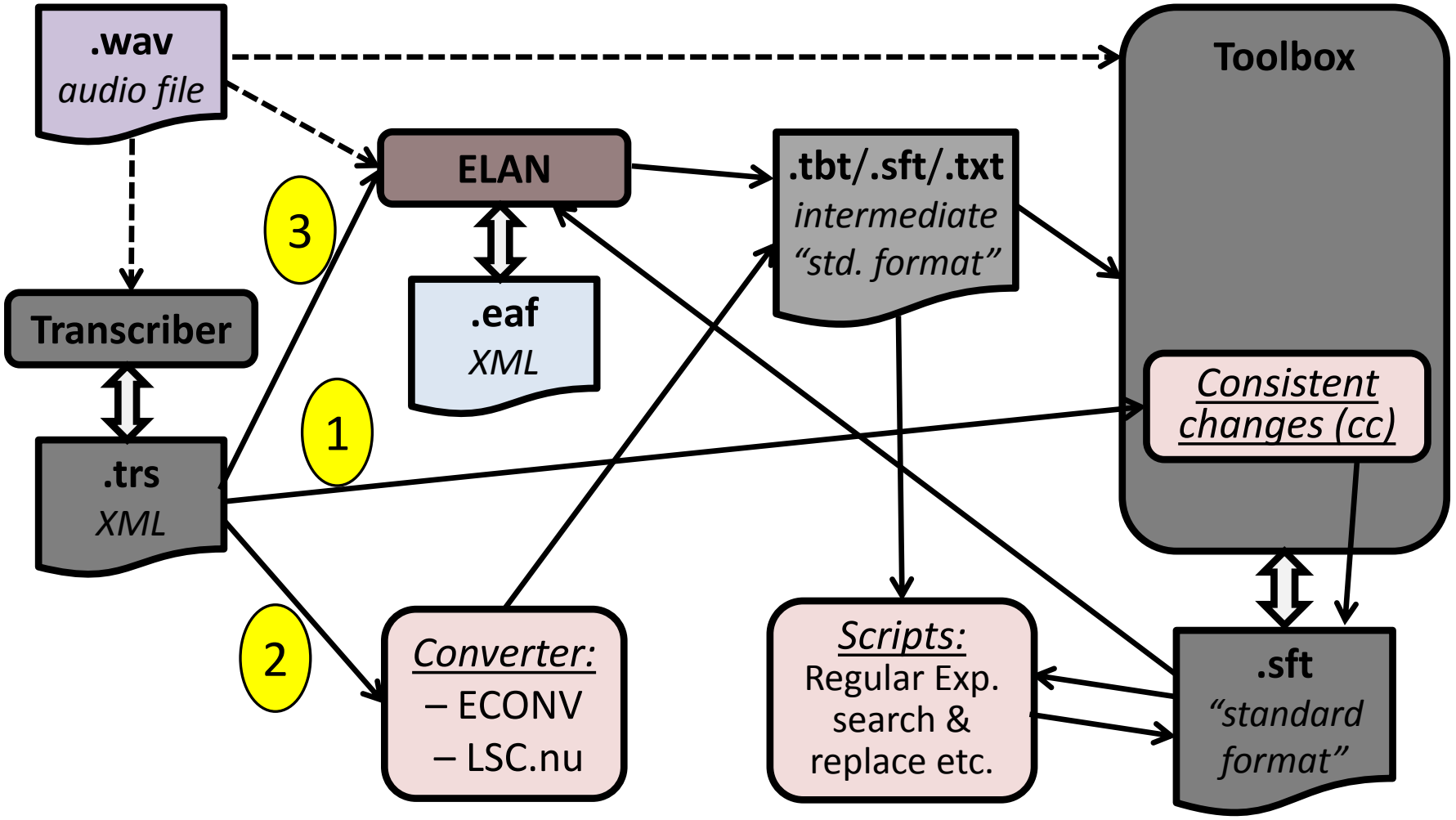
Round-trip Toolbox - ELAN



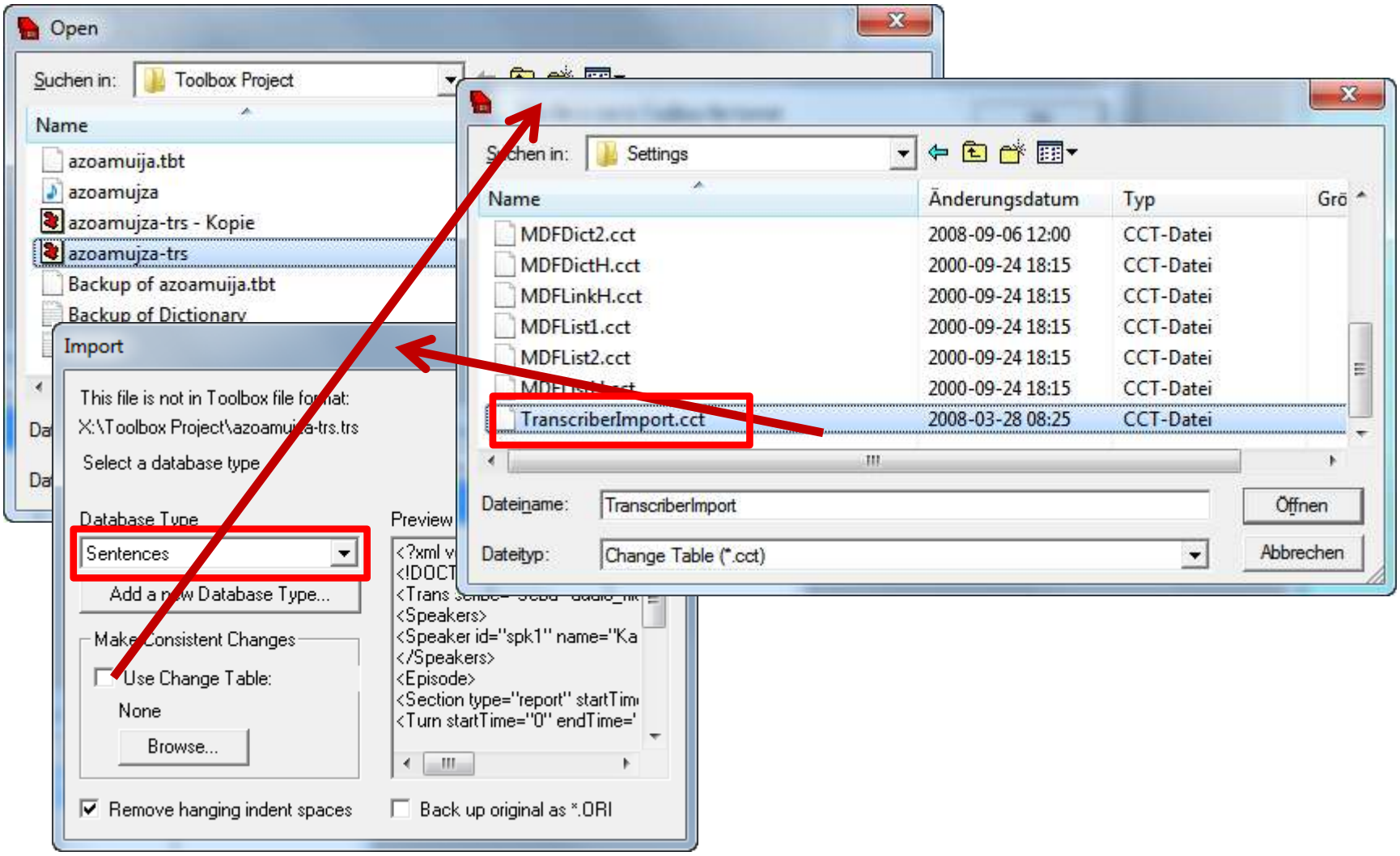
From Transcriber to Toolbox



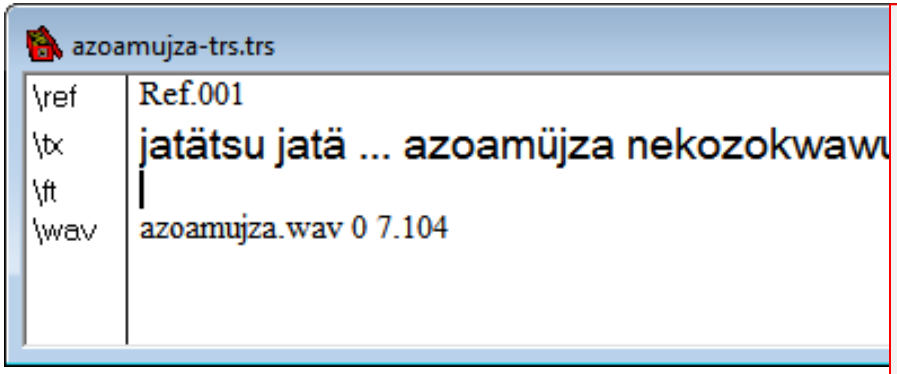
From Transcriber to Toolbox

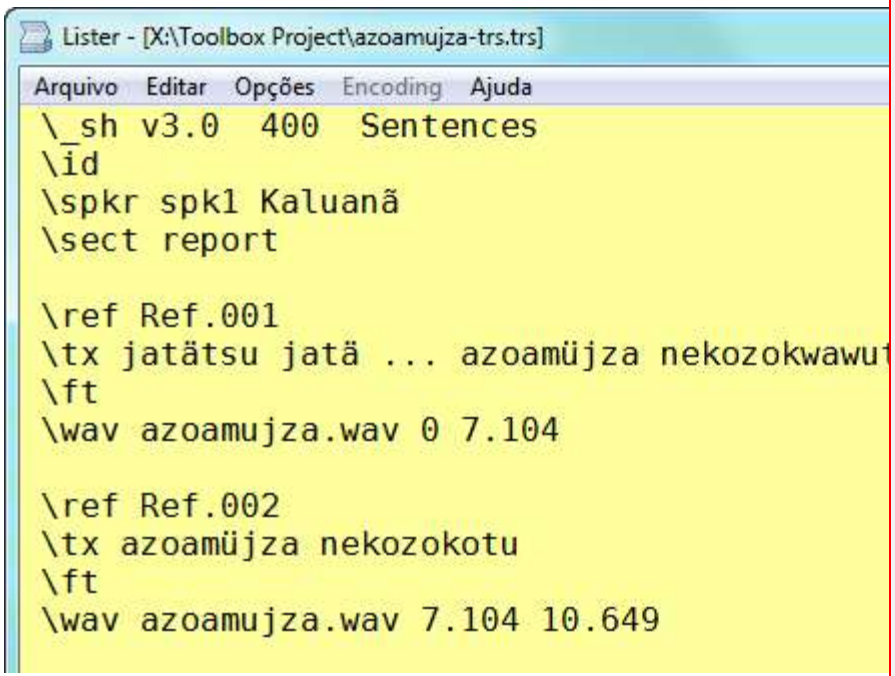


Toolbox: Direct import from Transcriber



Toolbox: Result from Transcriber import

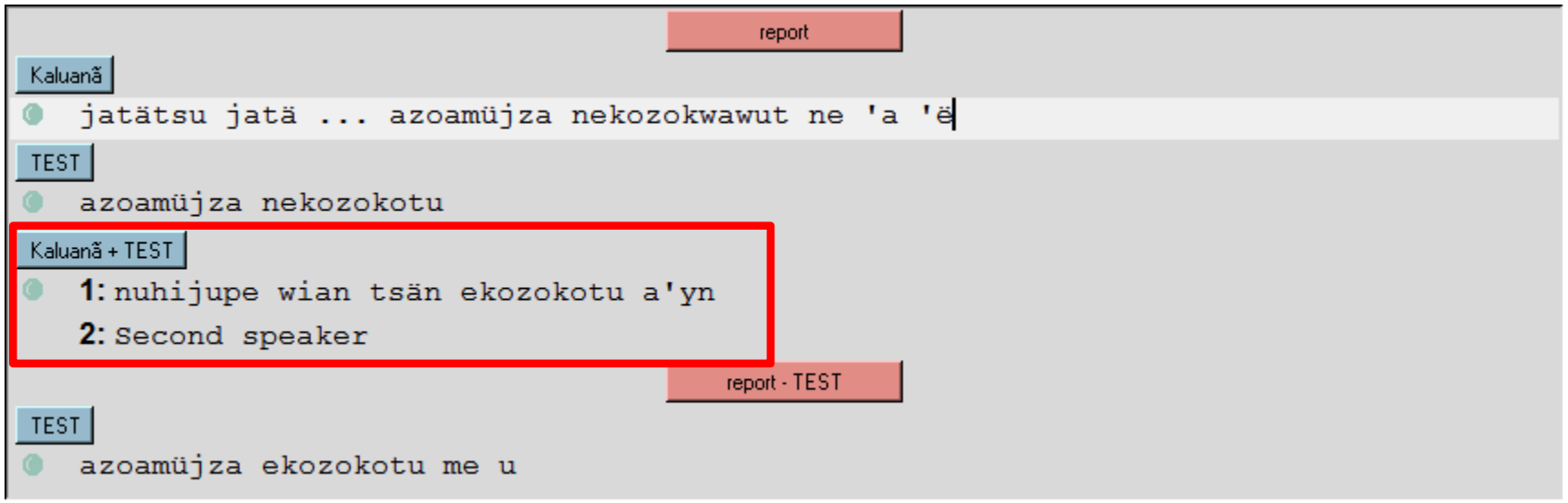




Problems:

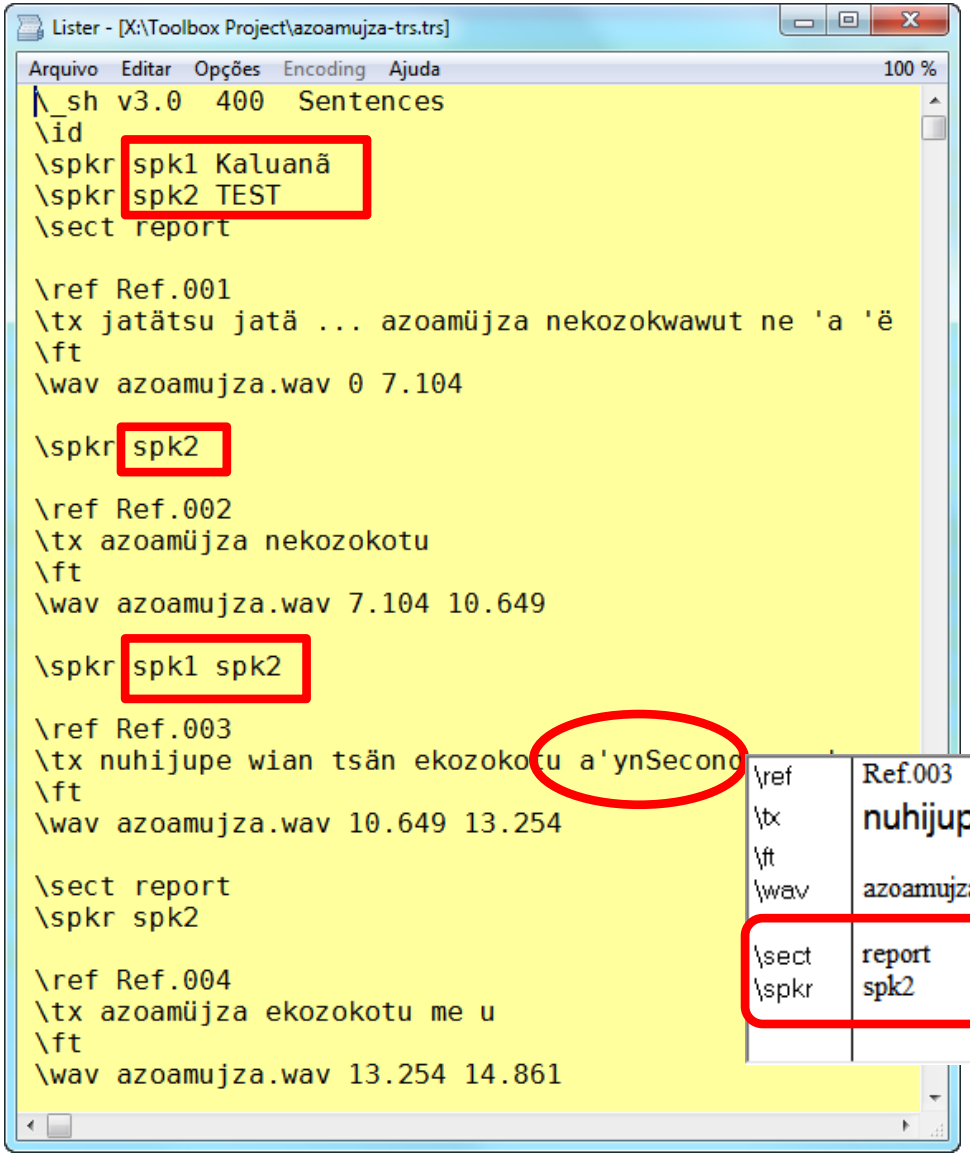
- The **\id** marker will be ignored (no problem)
- The **.trs** file is just overwritten without renaming (use a copy!)
- **\spkr** and **\sect** are at the wrong position in the hierarchy
- **\spkr** only appears with turn, not for each unit

Direct import from Transcriber: Tests with overlapping speech



report			TEST
Kaluanã	TEST	Kaluanã + TEST	TEST
jatätsu jatä ... azoamüjza nekozokwawut ne 'a 'ë	azoamüjza nekozokotu	nuhijupe wian tsän... Second speaker	azoamüjza... ... me u

Direct import from Transcriber: Tests



Problems:

- The speaker *names* are indicated only once, later “spk2”
- Overlapping speech is not preserved

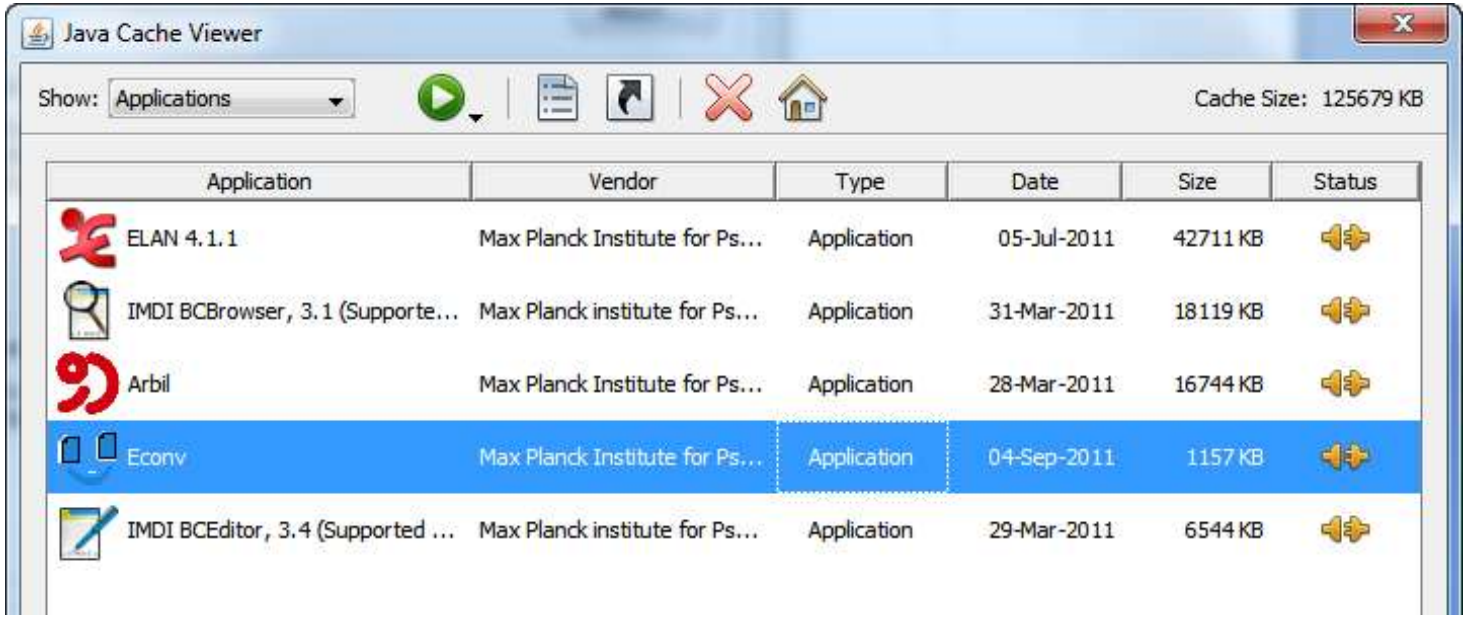
Transcriber > Toolbox: Use a converter

Converters:

- ECONV
- http://linguisticsoftwareconverters.zong.mine.nu/
- … write your own …

Transcriber > Toolbox: ECONV

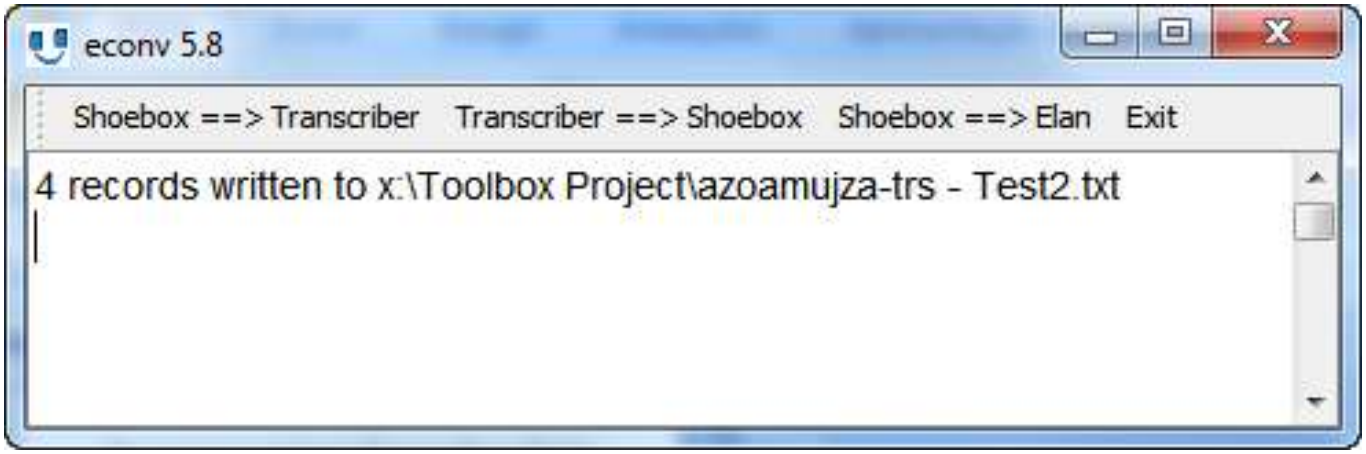
- There used to be a converter at the MPI: ECONV
- In fact, it is still online, but hidden: http://www.mpi.nl/tg/j2se/jnlp/econv/econv.jnlp
- Called with Java WebStart: **`Javaws -viewer`**



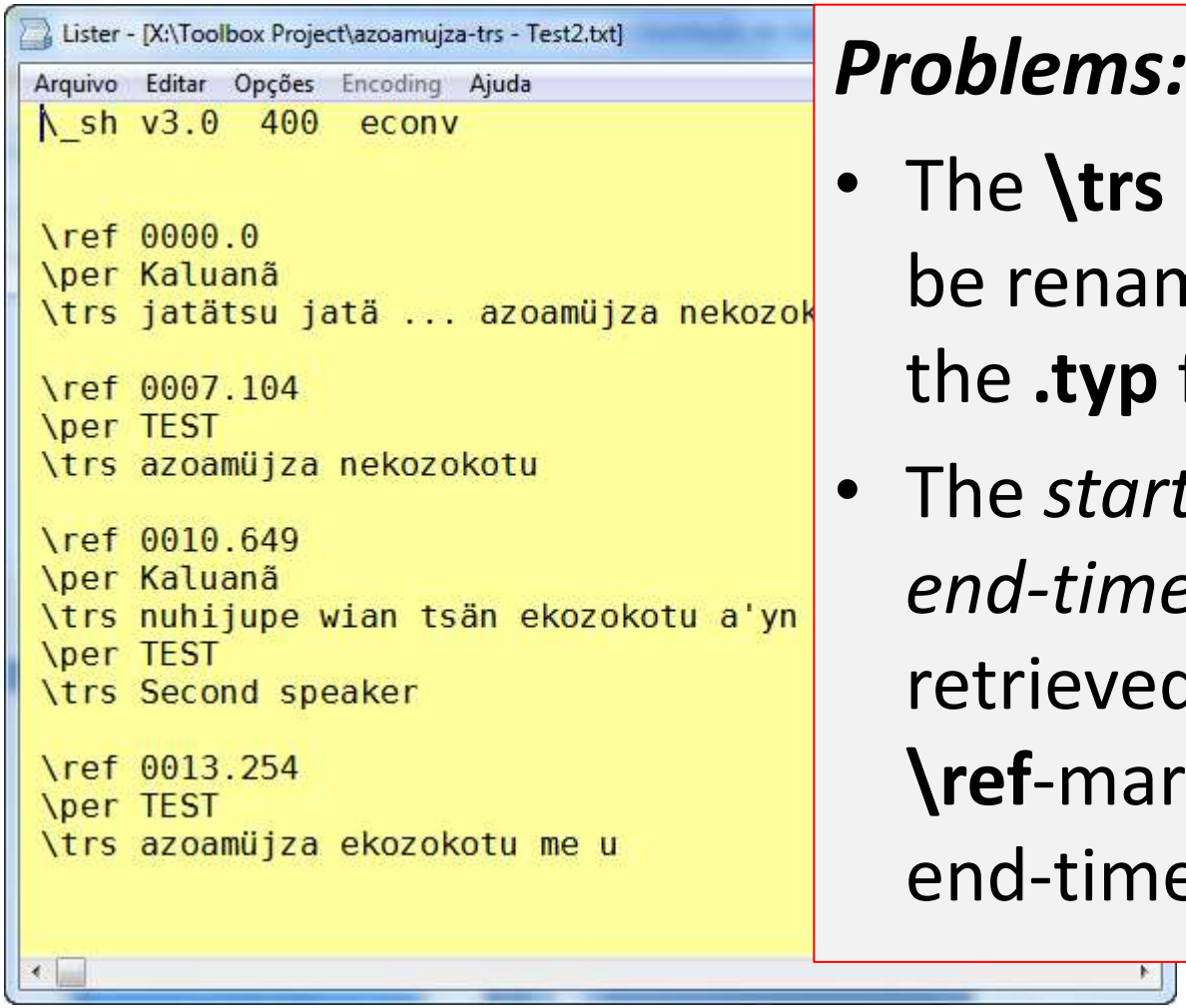
ECONV: Procedure

Several caveats:

- You need the file **`trans-14.dtd`** in the same directory as the file to be converted
- You must not use different ***sections***
- At least on speaker must be defined



ECONV: Problems

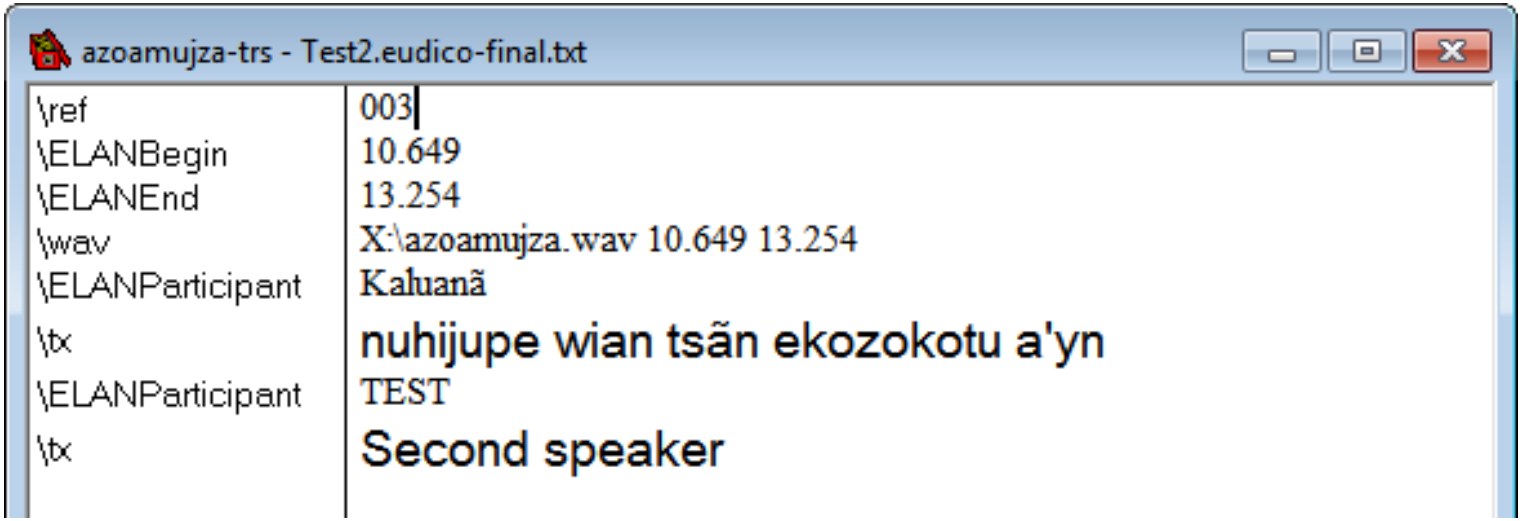


Problems:

- The **\trs** marker must be renamed to **\tx**, or the **.typ** file adjusted
- The *start-time* and *end-time* must be retrieved from the **\ref**-markers (last end-time is missing)

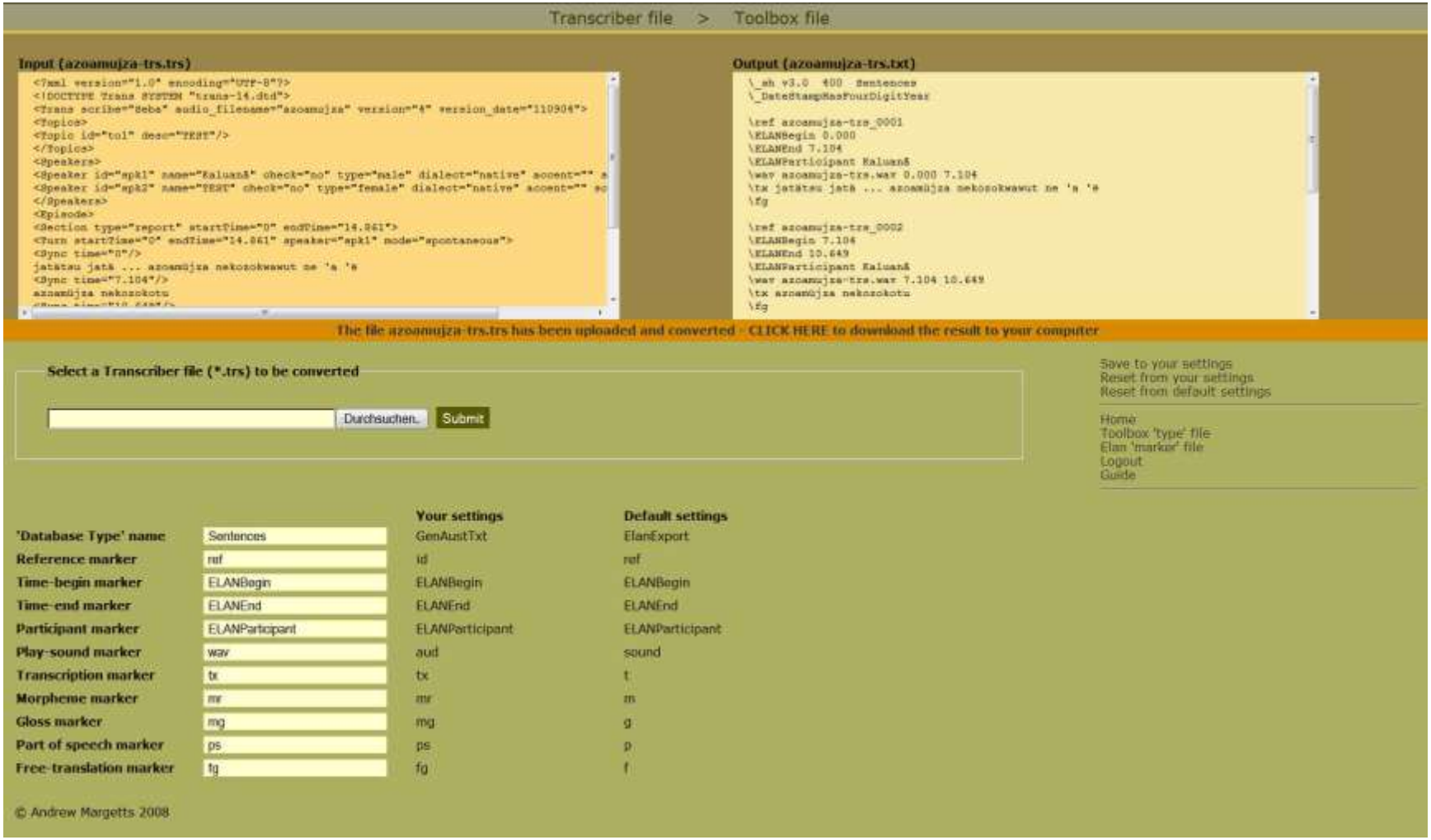
ECONV: Results

- All this can be done with a series of “scripts” which manipulate the std. fmt. text file
- The result is similar to the export from ELAN
- Overlapping speech: both **\tx** in one record

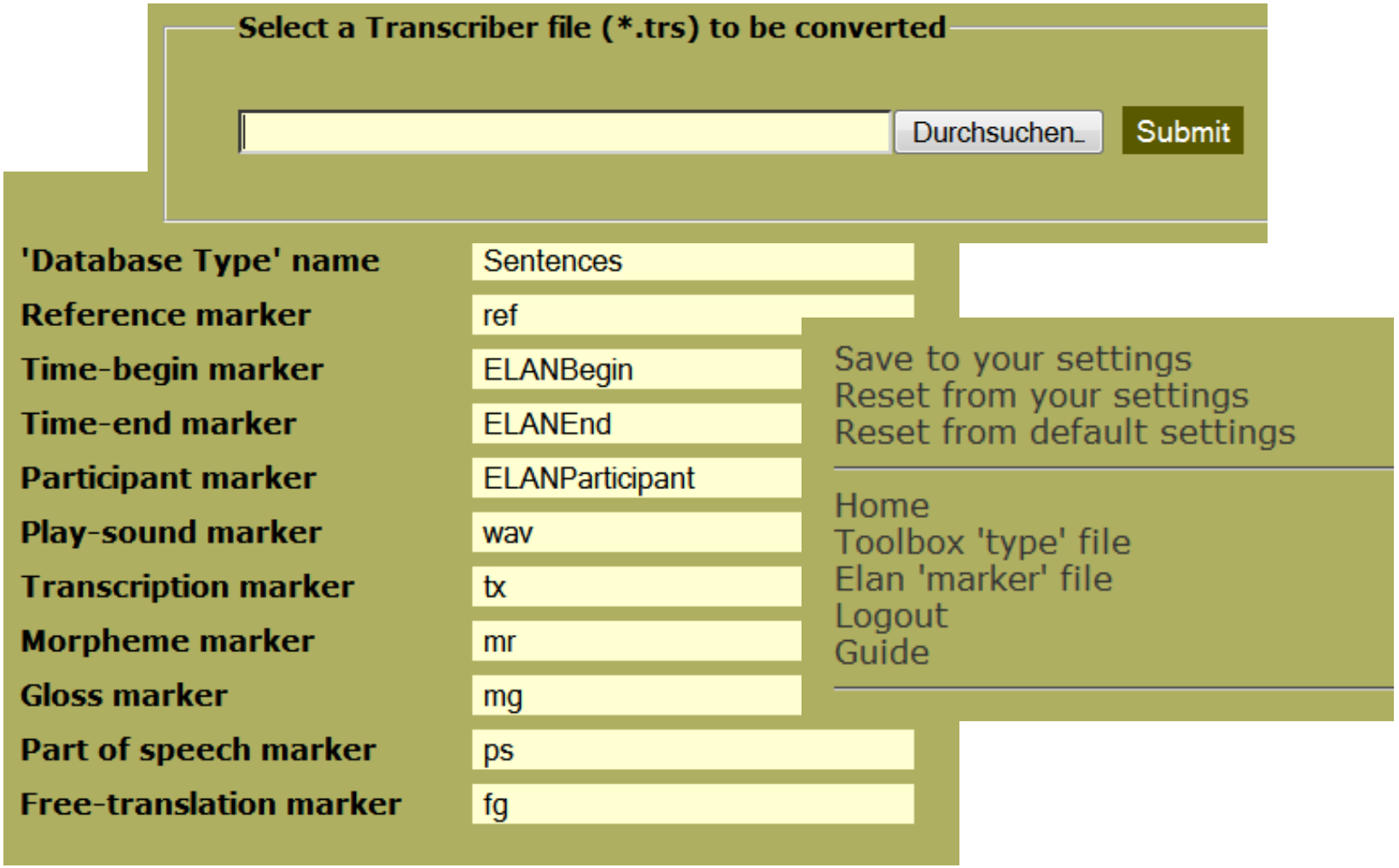


http://linguisticsoftwareconverters.zong.mine.nu

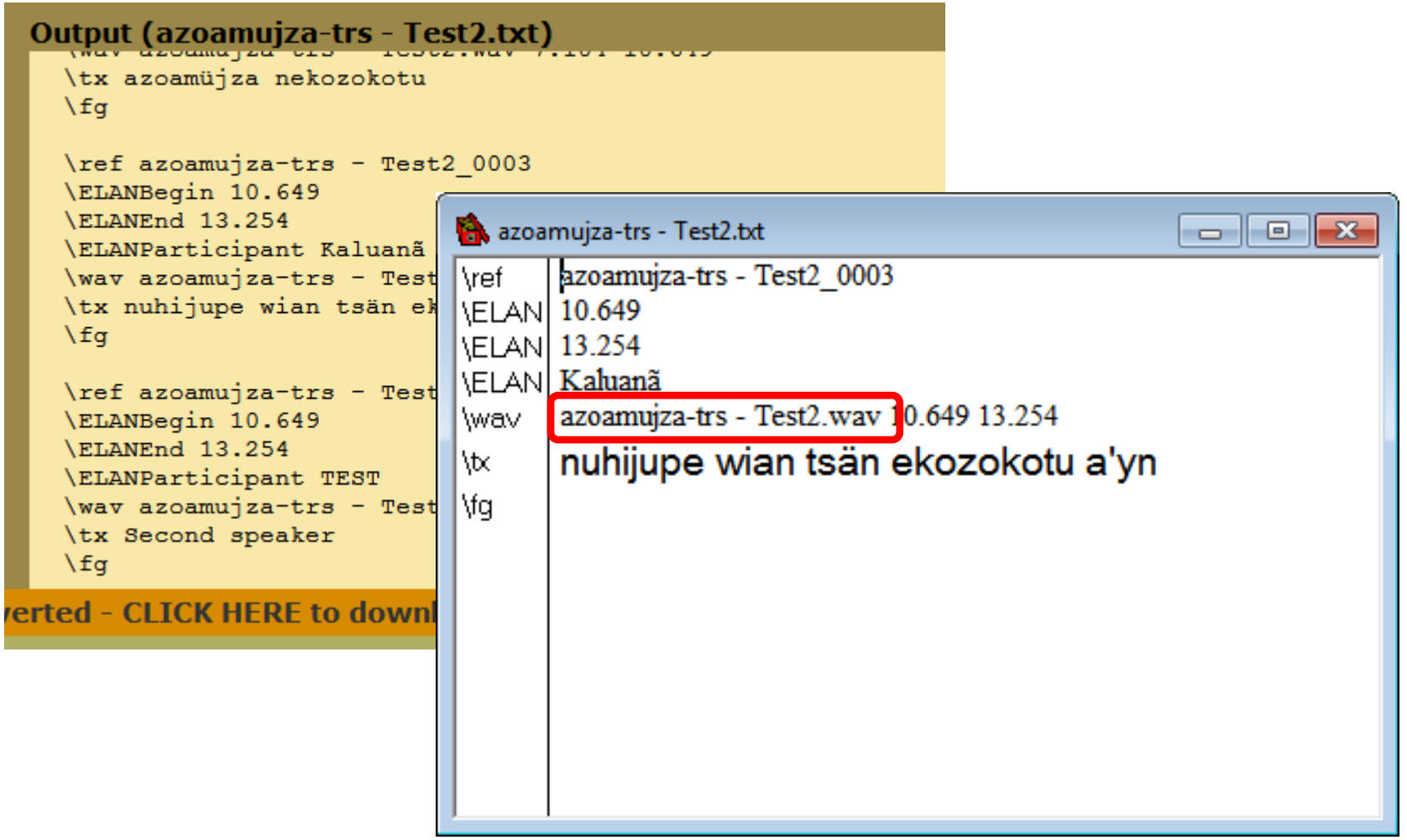
(by Andrew Margetts, DOBES)



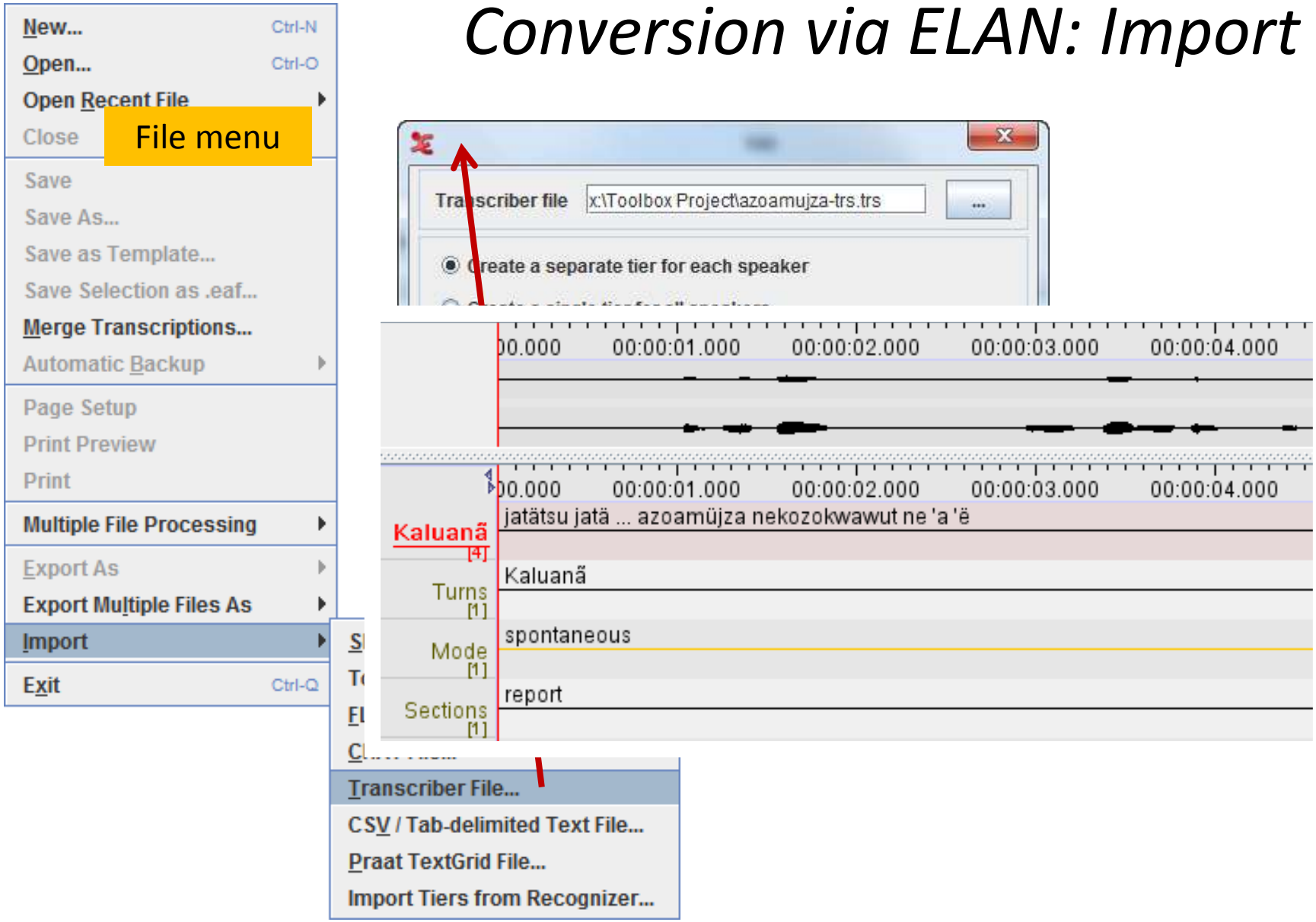
Linguistc Software Converters: Configuration



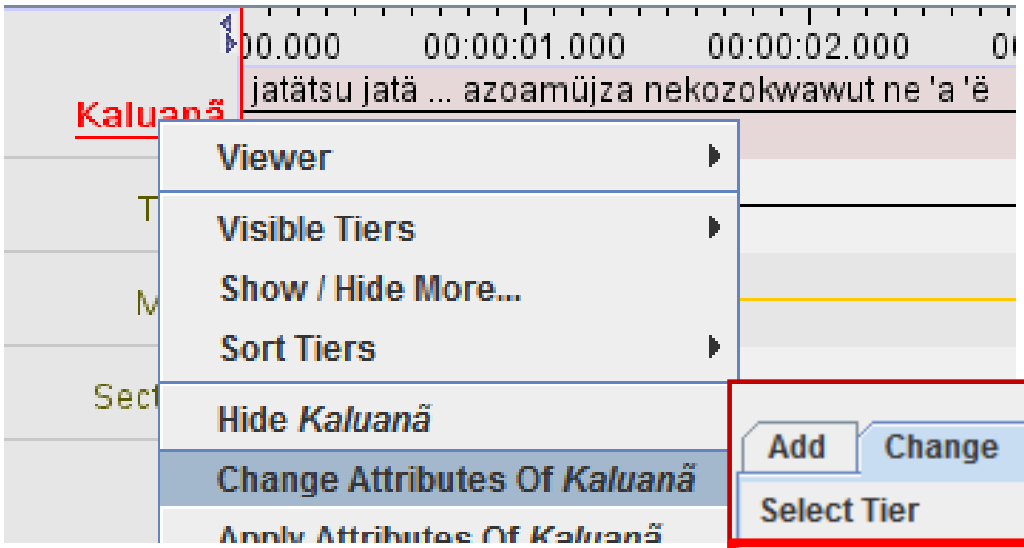
Linguistc Software Converters: Results



Conversion via ELAN: Import

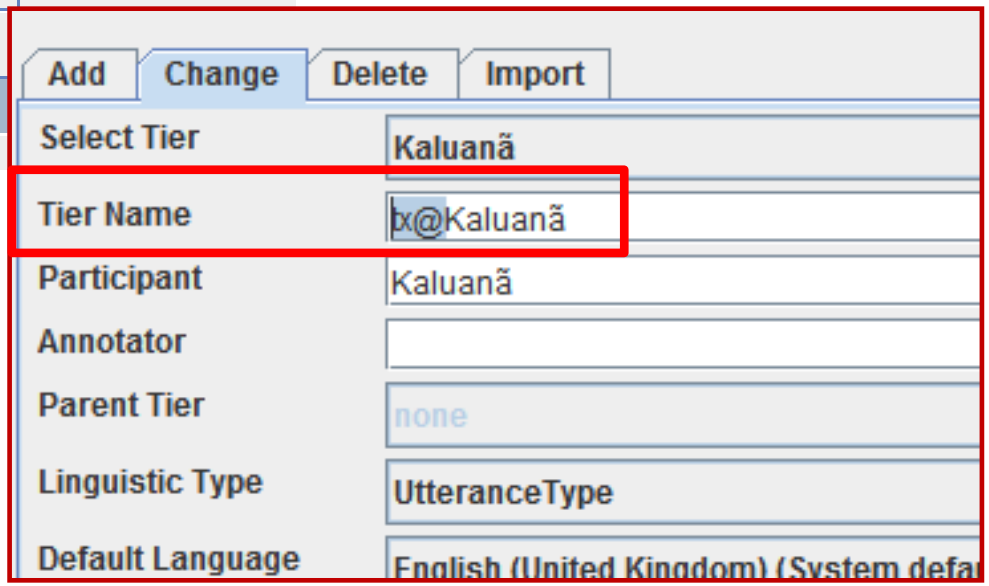


Adjustment in ELAN



Right-click
on the tier name

Choose “Change Attributes of …”



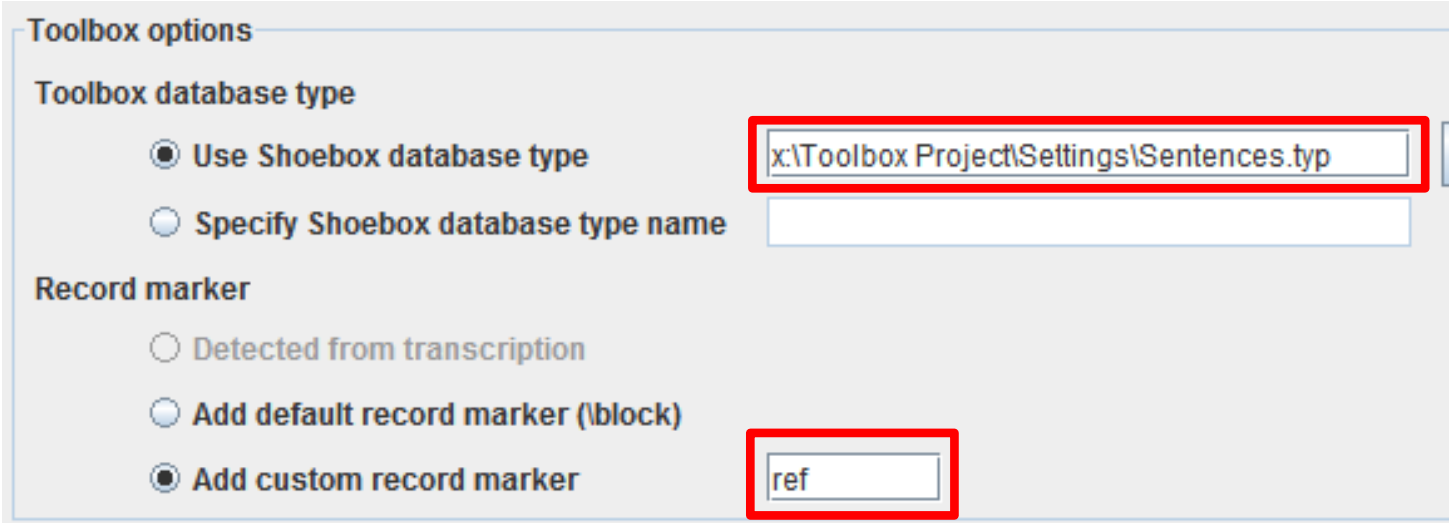
Add “**tx@**”
at the beginning
of the tier name

ELAN: Export to Toolbox

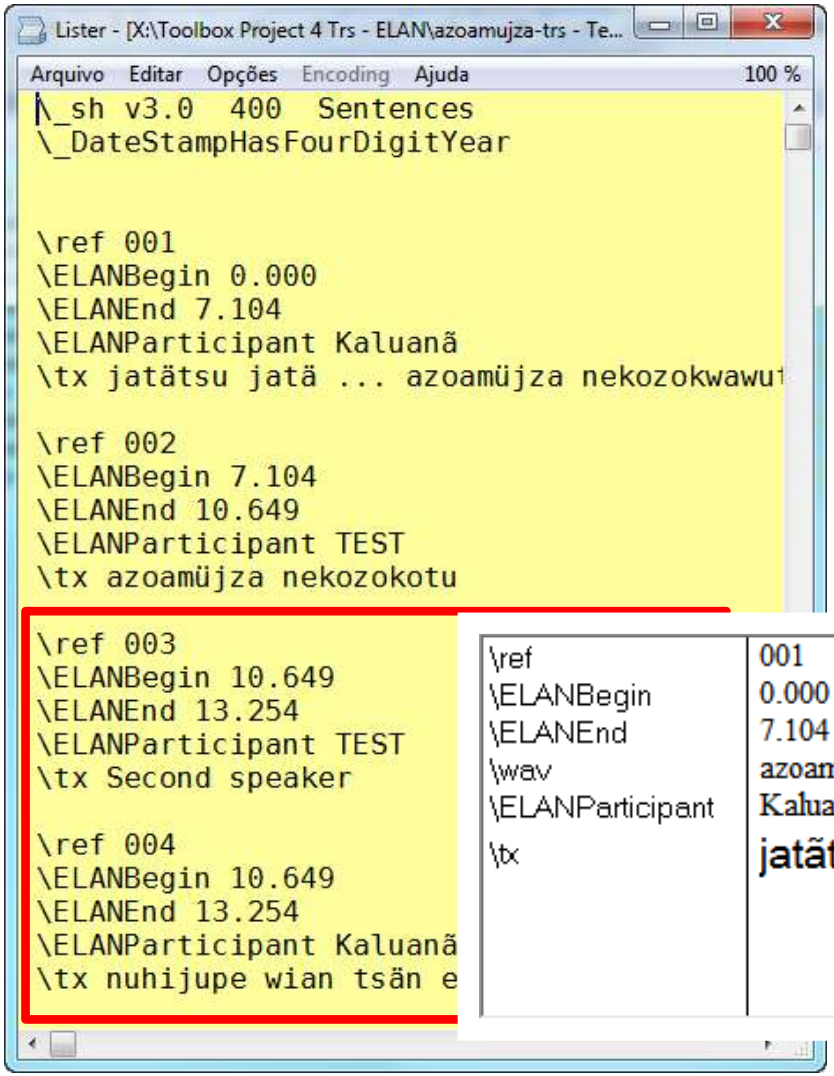


Do not export the additional tiers

Other settings are as before



Transcriber > ELAN > Toolbox: Result



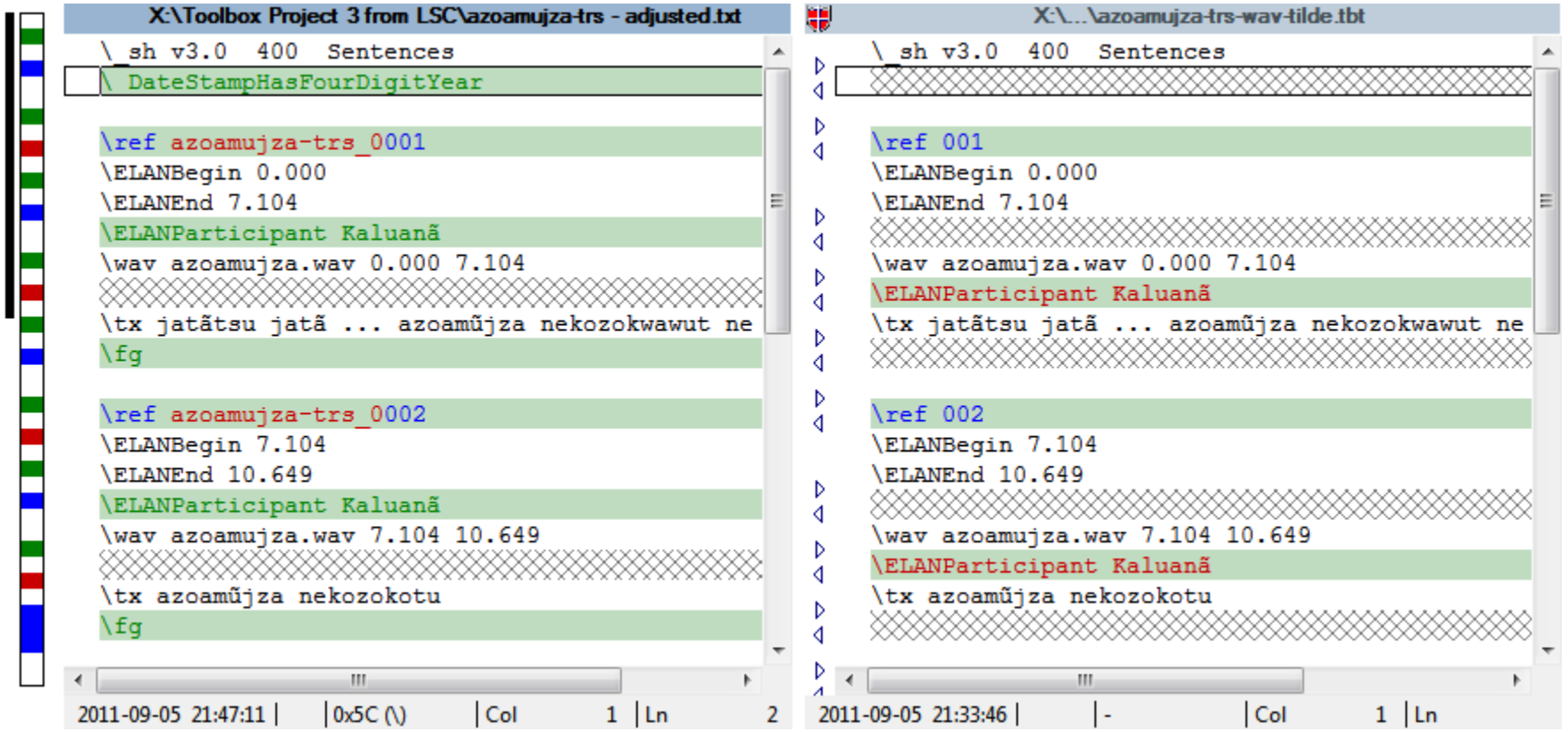
\ref	001
\ELANBegin	0.000
\ELANEnd	7.104
\wav	azoamujza.wav 0.000 7.104
\ELANParticipant	Kaluanã
\tx	jatãtsu jatã ... azoamũjza nekozokwawut ne 'a 'ẽ

Overlapping speech is represented in separate entries

After adding the wave field and replacing the umlaut by a tilde

LSC and ELAN as converter: comparison

Only the ref field and the order of fields are different



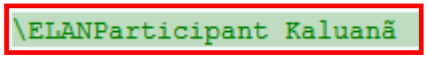
LSC and ELAN as converter: comparison

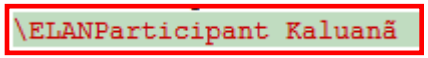
Only the ref field and the order of fields are different



LSC and ELAN as converter: comparison

Only the ref field and the order of fields are different





Topics

1. Data and Annotation (Theory)
2. Annotation Tools (Overview and Comparison)
3. Intro to Interlinearization (not time-aligned)

   Excurse: Text- vs. sentence-based databases
4. Time-aligned Annotation
   1. ELAN-Generated Annotation
   2. Play chunks of an audio file in Toolbox

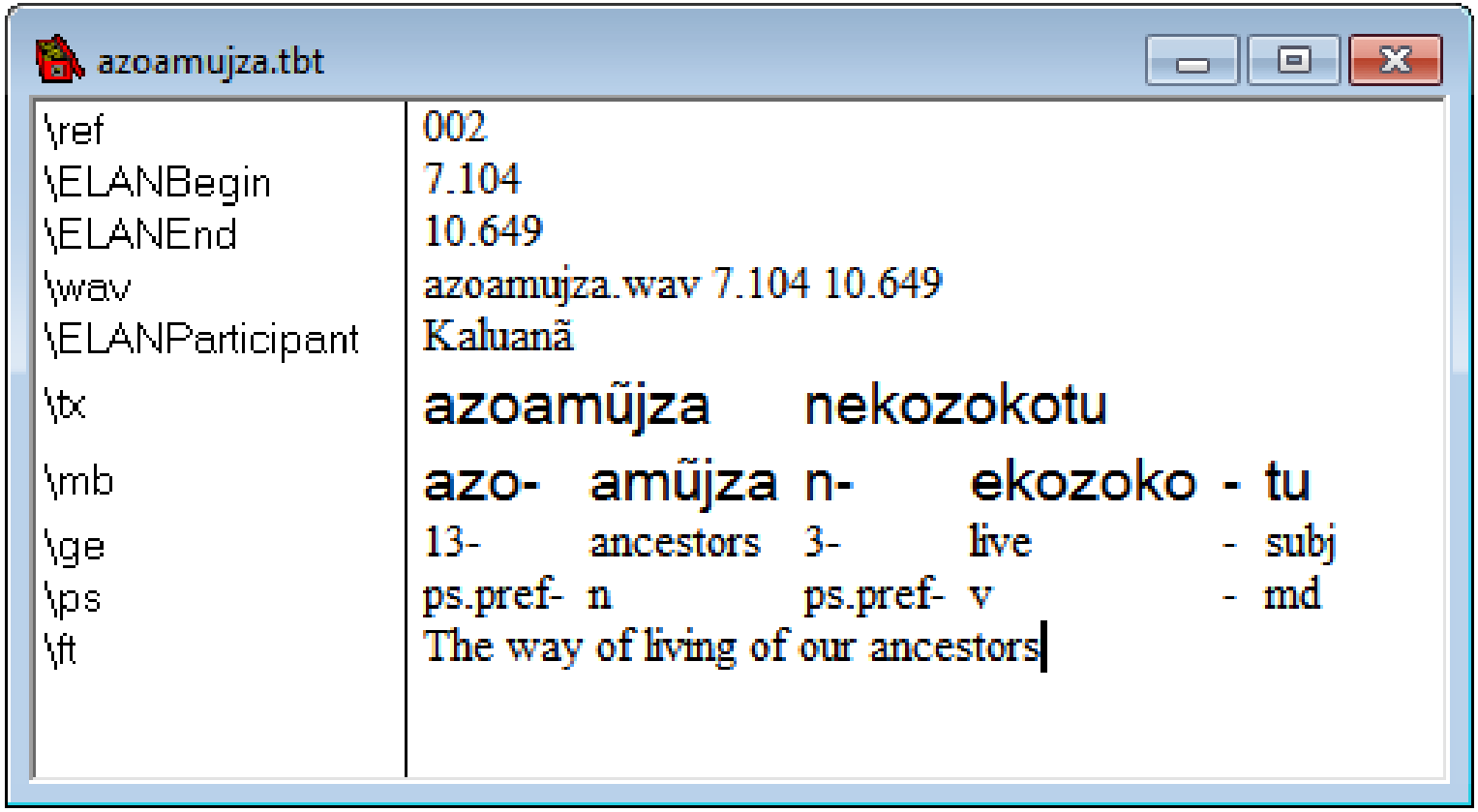
      Excurse: Regular Expressions

      Excurse: UNICODE and UTF-8
   3. Transcriber-Generated Annotation; Conversions
   4. **Round-trip Configuration ELAN--Toolbox**

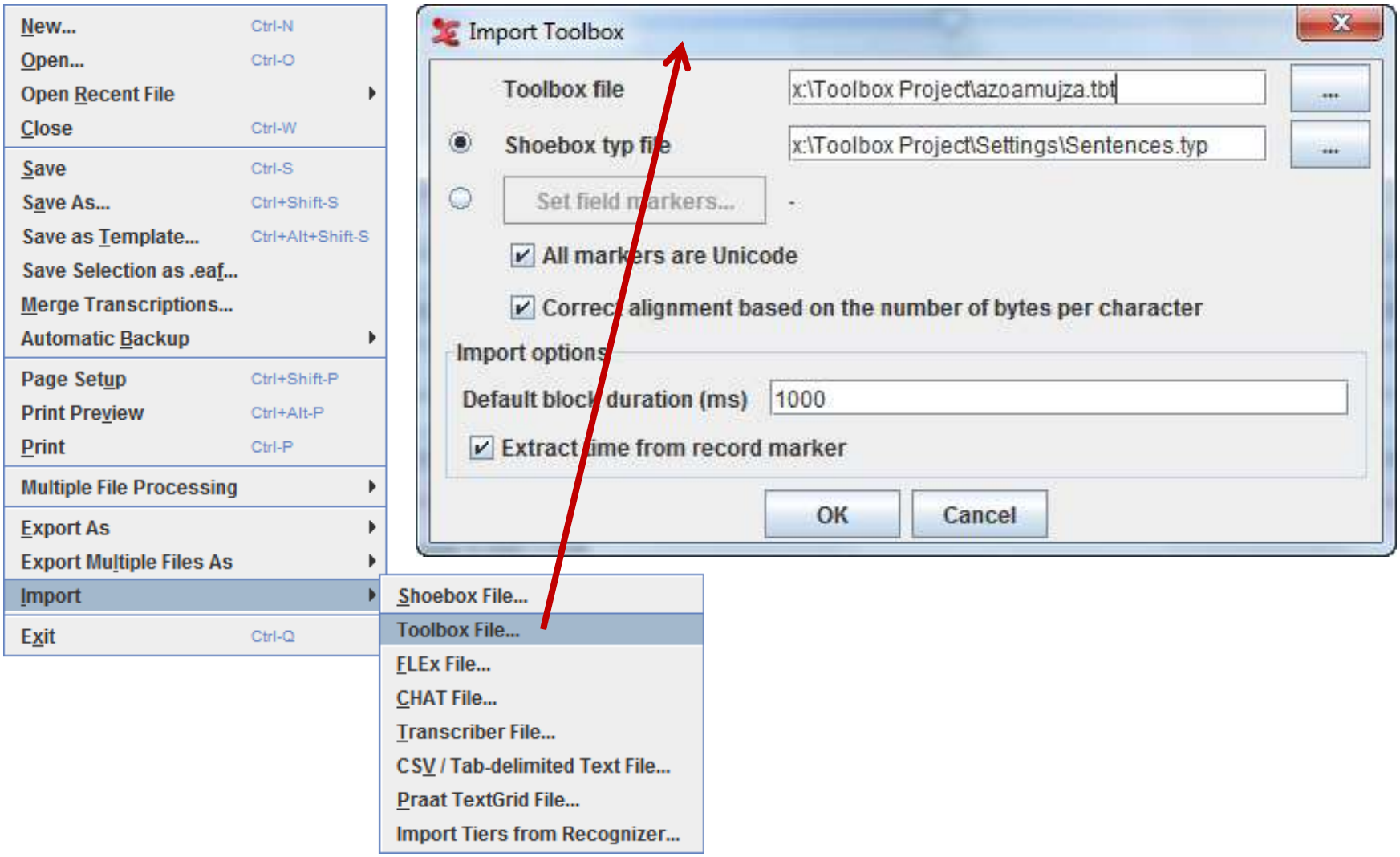
Interlinearize the time-linked transcription

- Use Toolbox to interlinearize the file with the time-marks and transcription generated with ELAN or Transcriber and imported to Toolbox
- The same settings as before with non-time-linked annotation should work
- After interlinarization, that file can be exported to other tools, e.g. to Audiamus or EOPAS, but in particular back to ELAN, for online-display with ANNEX

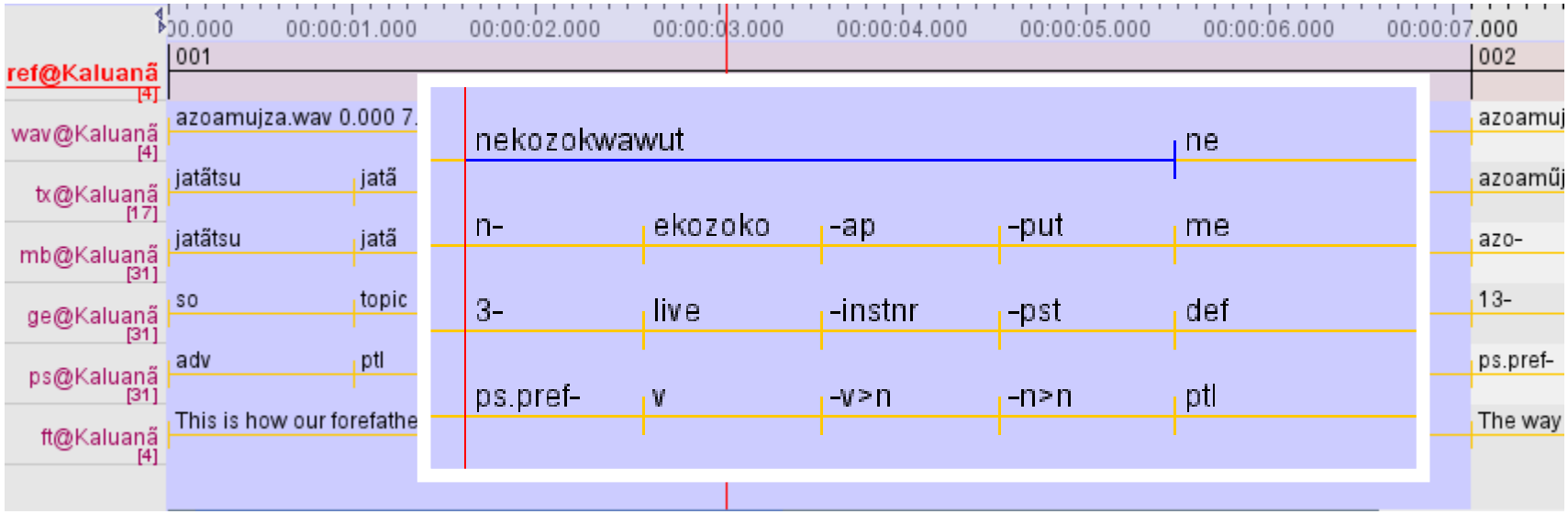
Interlinearized time-linked transcription



Importing interlinearized file into ELAN



Interlinearized file back to ELAN

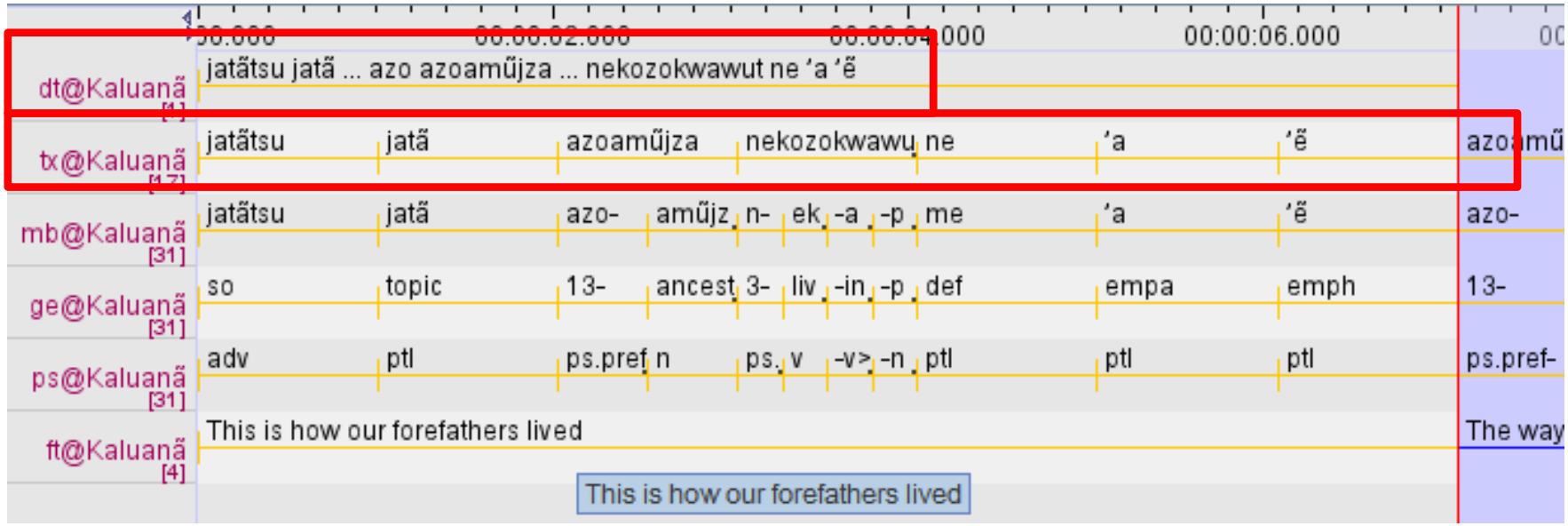


Usually, interlinearization is correctly preserved after loading the file in ELAN

Avoid using spaces in glosses or part of speech labels!! Use dots or hyphens or underlines

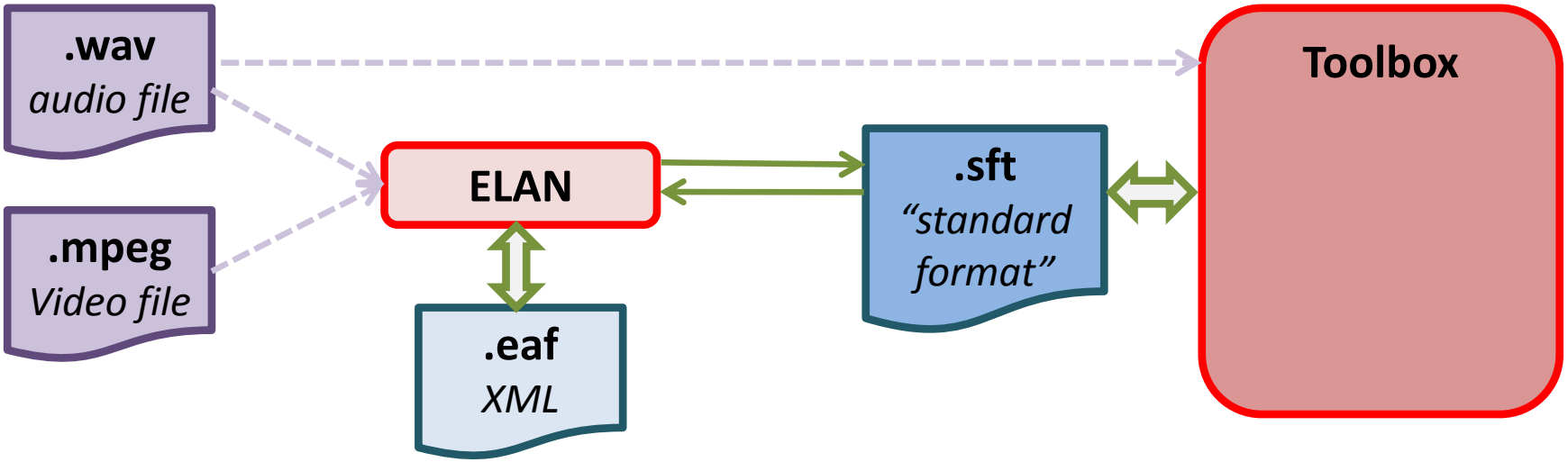
If things should go wrong, ask for help

Interlinearized file back to ELAN



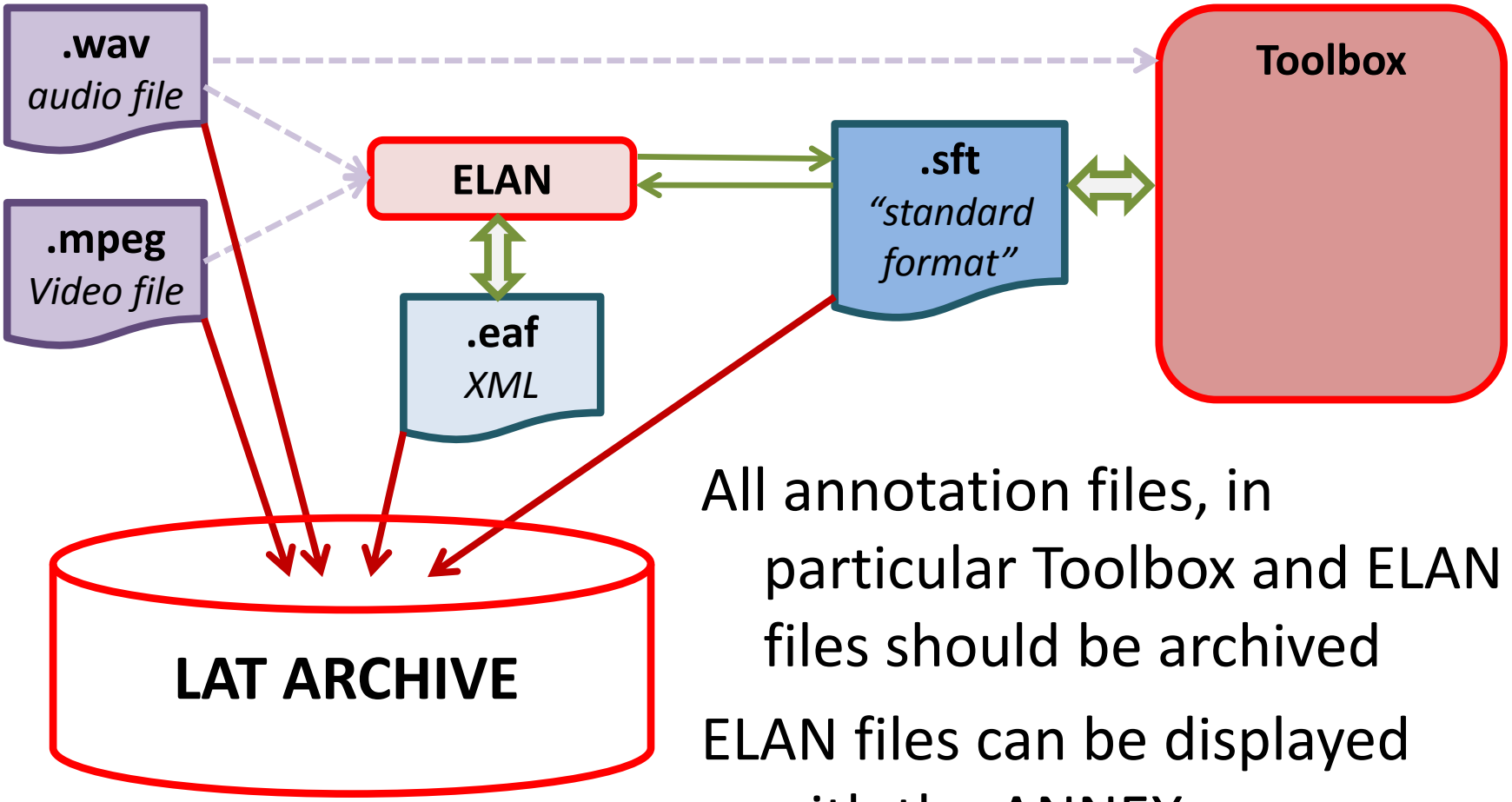
It may be useful to have TWO transcription lines, e.g. one “narrow” transcription, not used for interlinearization, and a normalized one for interlinearization. This facilitates reading.

"Round-trip" ELAN--Toolbox--ELAN--Toolbox



The goal is to have a working "round trip" setting, exchanging files between ELAN and Toolbox

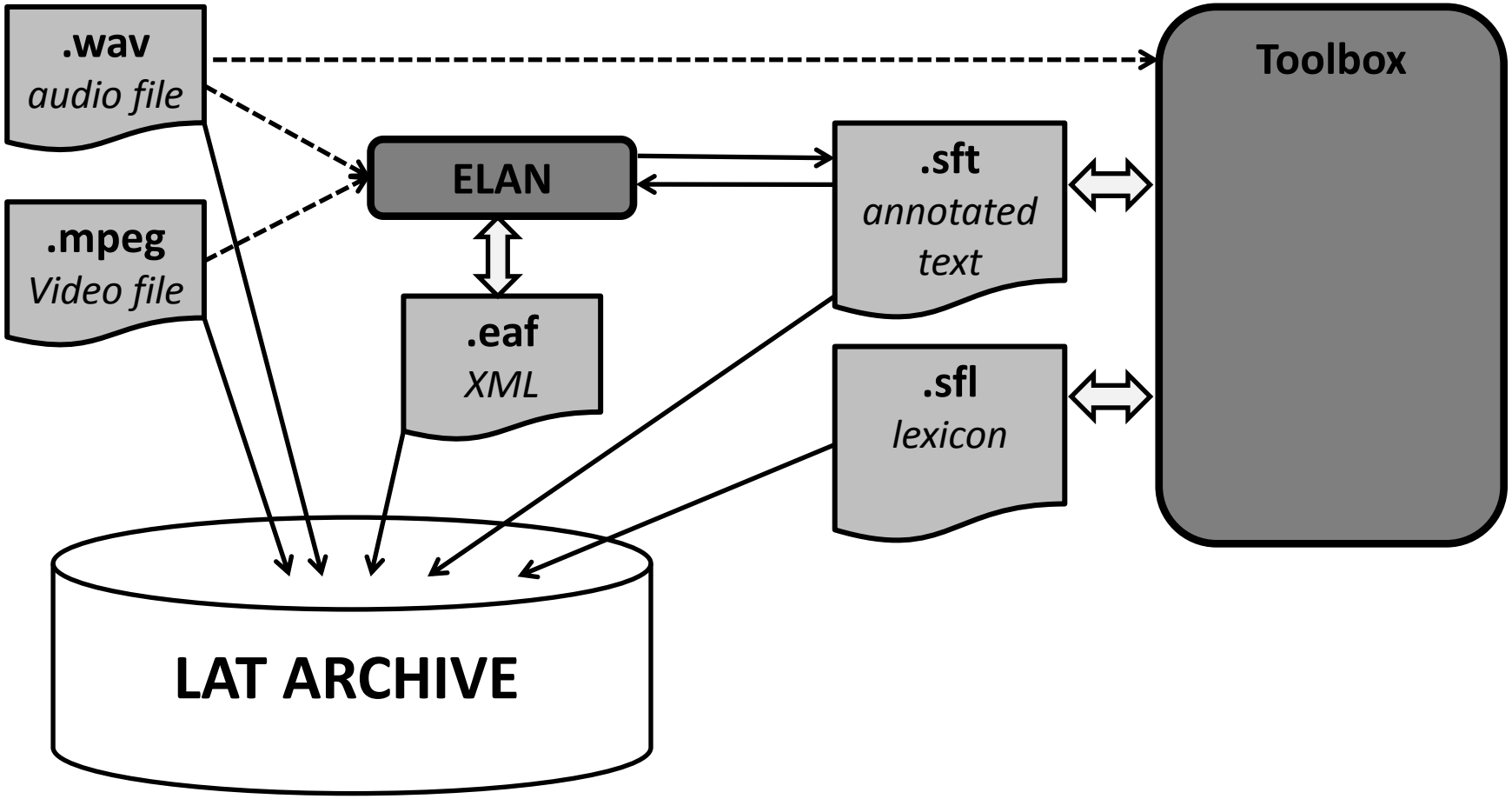
Archiving annotation files



All annotation files, in particular Toolbox and ELAN files should be archived

ELAN files can be displayed with the ANNEX program

Archiving annotation files



Annotation tool Toolbox

how to gloss/annotate in Toolbox

Language Documentation

SEBASTIAN DRUDE

2012-07