4. From Praat to ELAN

Important

Don’t forget the following steps in Praat before importing into ELAN:

In Praat, *Preferences*, check that *Text writing Preferences* is set to 'UTF-8'. If it isn’t, change it to UTF-8 and write again to Textgrid to save the new file in UTF-8; if you don’t do that, an Elan importation error will occur ("*operation interrupted...*").
This is to be checked each time you update the Praat version.

In the textGrid, the initial interval tiers (initially Mary John Bell) should be labelled

a) *ref@SP* (not *tx*), or *ref@SP1* and r*ef@SP2* if there are two speakers, etc. in case you have just one line of transcription (into grammatical words) per speaker,
b) *ref@SP* (for the broad phonetic transcription into phonological words) and *tmp@SP* (for the morpho-phonological transcription into grammatical words) for a two-line transcription (or *ref@SP1*, *tmp@SP1*, *ref@SP2, tmp@SP2*, etc. if there are several speakers). This case will not be treated here, but on the web site.

If for some reason the name of the Praat tier was not *ref*, rename it from Praat

- Open the textGrid in Praat, Select the corresponding object, Edit
- Tier, Rename Tier, Change the name of the tier to *ref* , Ok

If you have used the trigraph method to transcribe, convert the entire textGrid to Unicode

- Edit the textGrid object,
- Edit, Convert entire textGrid to Unicode.

To save the corrected textGrid

- File, Write TextGrid to text file

Before you start importing Praat documents, copy **Corpo1.etf**, **Corpo2.etf and Corpo3.etf** files into the *ELAN* folder, and if you are using Toolbox, copy **refCorp.typ** and **mdf.typ** in the *Toolbox\Settings* folder. This is done once and for all.

The Praat transcribed document including the intonation unit boundaries is now ready to be opened in ELAN in order to be prepared for the other annotations.

4.1. How to import a Praat document into Elan

Creating a new ELAN Document.

- File, New
- in *Files of type*, select: *Media files* (NOT *Template*), and choose in the left window the audio file you want to annotate.
- Click on the > > button between the 2 windows, then click on OK.

To give the new ELAN document a name:

- File, Save As: (enter the name of your file in the following format):

  LanguageCode_Author's Initials_type_num

  type = conv(ersation) or narr(ation); num = serial number of the file.

Importing the model of linguistic types

ELAN needs information on the hierarchical dependency of the tiers. To be consistent throughout the corpus, we will load a template for that.

- *Type, Import Types, Browse*, look for *Corpo1.etf* (*Corpo2.etf* if 2 speakers; *Corpo3.etf* if 3 speakers), *Import, Close*.

Deleting the *Default* tier

We don't need this default tier.

- right-Click on *Default*; Select *Delete Default*; press YES.

Importing the TextGrid file created with Praat

- *File, Import, Praat TextGrid, Browse*, look for and select the *TextGrid* file you want to import;
- Check the *Skip empty interval/annotations* box to avoid the creation of empty segments. *Next*;
- Make sure that *Linguistic type*: ***ref*** is selected under *Type Name*, NOT *default'*
- *Finish*
- *Operation completed*, OK

4.2. Preparing the mot *line in Elan from a one-tier transcription in Praat*

As we need a main labelled and numbered reference line *ref* for each annotation unit as well as a *tx* line, we will have to duplicate the *ref@SP* tier to create the *tx* tier. Then the *ref* tier will be labelled and numbered. Next, after the importation of the other tiers, the *mot* tier will be filled in by tokenizing the *tx* tier into it. Finally, the *tx* tier will be modified by hand to reflect the broad phonetics transcription of the sound file.

Creating a new *tx* tier

In order to create a new *tx* tier, let's duplicate the *ref* tier.

If there is only one speaker:

In the Tier menu :

- *Copy Tier*
- select *ref@SP*, *Next*
- once again: select *ref@SP*, *Next*
- as *Type Name*, choose *tx*, *Finish*
- *Operation completed*, OK

A tier *ref@SP-cp* was created.

If there are 2 speakers:

- *Copy, Tier*
- select *ref@SP1*, *Next*
- once again: select *ref@SP1*, Next
- select as *Linguistic type* : *tx*, *Finish*

Second speaker :

- *Copy, Tier*
- select *ref@SP2*, *Next*
- once again: select *ref@SP2*, *Next*
- select as *Linguistic type*: *tx*, *Finish*

And so on for other speakers.

Renaming the new *ref@SP-cp* tier as *tx@SP*

(resp. *ref@SP1-cp* as *tx@SP1, ref@SP2-cp* as *tx@SP2...* if multiple speakers)

In the Tier menu:

- *Change tier attributes*
- Select *ref@SP-cp*
- Type *tx@SP* as its new *Tier Name*
- Click on *Change*

For multiple speakers, do the same for each *ref@SP1-cp, ref@SP2-cp* or *ref@SP3-cp*

- *Close* the window when finished

Labelling and numbering the *ref* tier(s)

(for more than one speaker, do the same thing with *ref@SP1, ref@SP2* and *ref@SP3*)

- *Tier, Label and Number,* select *ref@SP*
  - *Include label part*:

1 speaker: LanguageCode_Author's Initials_type_num (= *name of the .wav file*)
more than 1 speaker*:* LanguageCode_Author's Initials_type_num_SPnumber

- *Insert other delimiter* : _ (underline symbol)
- OK, *Close*

Importing the remaining tiers

- *Tiers, Import Tiers, Browse*, look for *Corpo1.etf*, *Import, Close*.

(for two speakers, look for*Corpo2.etf*, for three speakers, look for *Corpo3.etf*)

Filling in the *mot* tier

We will just tokenize (i.e. split the words of the prosodic units into individual cells) the grammatical word tier(s) *tx@SP* into the *mot@SP* tier(s). (Respectively *tx@SP1* into *mot@SP1*; *tx@SP2* into *mot@SP2*; *tx@SPp* into *mot@SP3* for multiple speakers):

- *Tier, Tokenise tier*
- Source : *tx@SP*
- Destination *: mot@SP*
- *Start, Close*

Displaying the tiers in the right order

The imported tiers may appear in a mixed order

- Click-Drag and Drop the labels of the tiers you want to move

or

- Right-click on the labels area
- *Sort Tiers, sort by hierarchy*

Changing the transcription of the tx tier

The *tx@SP* line(s) contain(s) the morphophonological transcription. This line has to be changed in ELAN by hand, unit by unit, into a broad phonetic transcription closely mirroring the audio file (assimilations and dissimilations retained), and containing phonological words instead of grammatical words.

At the end the ELAN file is correctly prepared and the *mot* tier is ready to be segmented into morphemes and annotated with the help of the lexicon and the internal parser.

4.3. Exporting the transcription lines to Praat (for further prosodic investigations)

This is a parenthesis for those who are concerned about having both a broad phonetic transcription and a grammatical word transcription in Praat. Here is a way of obtaining that result through ELAN, instead of doing those two transcriptions in Praat. Indeed, it is easy to export tiers content and time delimitation from ELAN to Praat.

From ELAN:

- File, Export As, choose *Praat textGrid*
- Uncheck the *Show only root tiers* checkbox
- Choose *tx* and *mot* (even *ref* if you want*)*, OK
- Choose the directory where to save the textGrid, and give the file a name

Now, this textGrid can be opened in Praat. The first tier will be the broad phonetic translation, the second one the grammatical word transcription. Be aware that the time boundaries of each word of the *mot* tier are correctly inside the time boundaries of the *tx* unit they belong to, but they are not correctly related to their real time duration because ELAN just divides the duration of the parent *tx* unit into equal cells for each word contained in this unit. If you are concerned about the real time duration of each word, you will have to move the boundaries of each word to align them according to their proper duration, by playing the sound of the current word (click on the bar under it).

5. ELAN-CorpA: Elan for CorpAfroAs

You are now in Elan for CorpAfroAs .

Changing the ELAN preferences

While typing your annotations, if you want to save a cell in ELAN, the default method is CTRL+ENTER (or CMD+ENTER in Mac). But there is a faster way: in the *Edit/preferences* menu, there is an item *writing preferences,* which contains *editing* in which you can choose *enter key commits changes in the inline edit box*. Then, by pressing only the ENTER key, you save your changes.

When a virtual keyboard is used, system shortcuts may conflict with some ELAN shortcuts. The solution is to change the ELAN shortcuts in the menu: *Edit, Preferences, Edit shortcuts*.

Interlinearizing process into ELAN

Until now, ELAN was not able to generate the segmentation and glossing lines *mb, ge* and *rx* on its own. What was possible was:

- doing the job manually by splitting the segments and adding the gloss in the cells,

- exporting the data prepared in ELAN to Toolbox, then parsing and annotating using the functionalities of Toolbox, then re-importing the Toolbox file into ELAN.

The idea was to simplify this process by giving the user access to some Toolbox-like functionalities directly in ELAN, i.e to allow the segmentation of words by means of a lexicon containing affixes, and to propose glossing by looking up into the same lexicon .

To do this, an « Interlinearize » tab has been added .

Once a file is opened, Click on the *« Interlinearize »* tab*.*

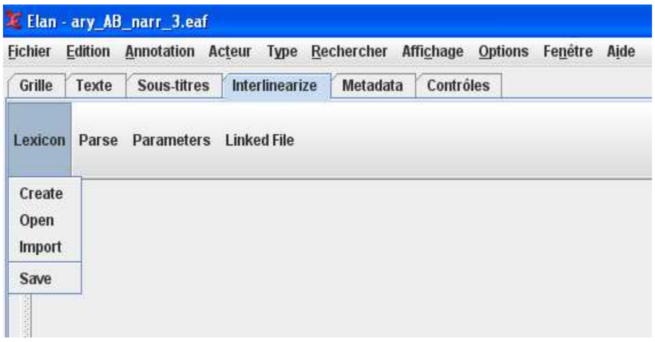
Creating, Opening, Importing a lexicon

The interlinearizing process relies on the existence of a lexicon which can be

- a brand new ELAN lexicon
- an existing ELAN lexicon
- a lexicon imported from Toolbox into ELAN

The « *Lexicon* » menu allows the choice

Click on the *«Lexicon »* tab


Elan - ary_AB_narr_3.eaf
Fichier Edition Annotation Acteur Type Rechercher Affichage Options Fenêtre Aide
Grille Texte Sous-titres Interlinearize Metadata Contrôles
Lexicon Parse Parameters Linked File
Create
Open
Import
Save

Creating an ELAN lexicon

When you choose *Create*, a file selection window will open. Choose the folder where you want to save your lexicon and give it a name. The **.eafl** extension will be automatically added.

On the left part of the screen, you will see a table with the different columns of the lexicon and a menu above, and on the right part of the screen, a display area with tabs and buttons relative to the interlinearizing process and the lexicon management (cf. figure below ary_AB_narr_3.eaf.)

Opening an ELAN lexicon

When you choose *Open*, a file selection window will appear. Choose the folder where your lexicon was saved (extension .eafl), select it, then open it.

Importing a Toolbox dictionary

When you choose *Import*, a file selection window will appear. Choose the folder where your Toolbox dictionary is saved, select it.

Not all the fields of a Toolbox dictionary are needed for the interlinearizing process (examples, definitions...). ELAN is aware of the following concepts (right box):

Lexeme (all kinds of entries in the lexicon: word entries, stems, word forms, affixes),

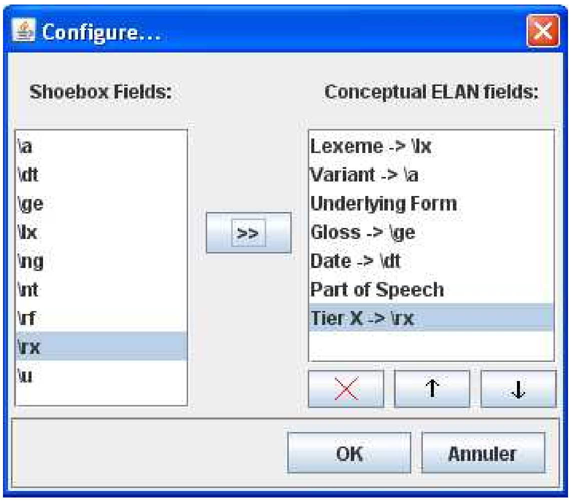
Variant (alternate form of a lexeme, principally depending on the context),
Underlying form (underlying segments of an entry or a variant of it),
Gloss (the meaning or sense of the lexeme),
Part of speech (grammatical category). This is not used in the CorpAfroAs format.
Tier X (category related to the entry, may be grammatical or other),
Date (last modified date of the entry).

Those ELAN lexicon concepts (right box) have to be related to the fields found in the Shoebox/Toolbox file (left box) for a correct importation of the dictionary data.

It is imperative for Lexeme, Gloss and Tier X to be related to a Shoebox/Toolbox field

If you don't have an \rx field in Toolbox, associate the Toolbox part of speech field (e.g \ps) to ELAN Part of Speech. This will copy the content of the Toolbox field into Tier X.".

You can define the relations you want by pairing the fields and concepts one by one, from the right box to the left and clicking on the > > button.



Configure...

Shoebox Fields:

\a
\dt
\ge
\lx
\ng
\nt
\rf
\rx
\u

Conceptual ELAN fields:

Lexeme -> \lx
Variant -> \a
Underlying Form
Gloss -> \ge
Date -> \dt
Part of Speech
Tier X -> \rx

- Select (click on) the concept in the right box
- Select the corresponding label field in the left box
- Click on the arrow button > > between the two boxes

Now, the concept selected from the right box has an arrow followed by the label of the corresponding Toolbox field.

To delete a correspondence, select the concept in the right box

- Click on the *red cross button*

To move a correspondence on top of the concept above

Click on the upward arrow button

To move a correspondence under the concept below

- Click on the *downward arrow button*

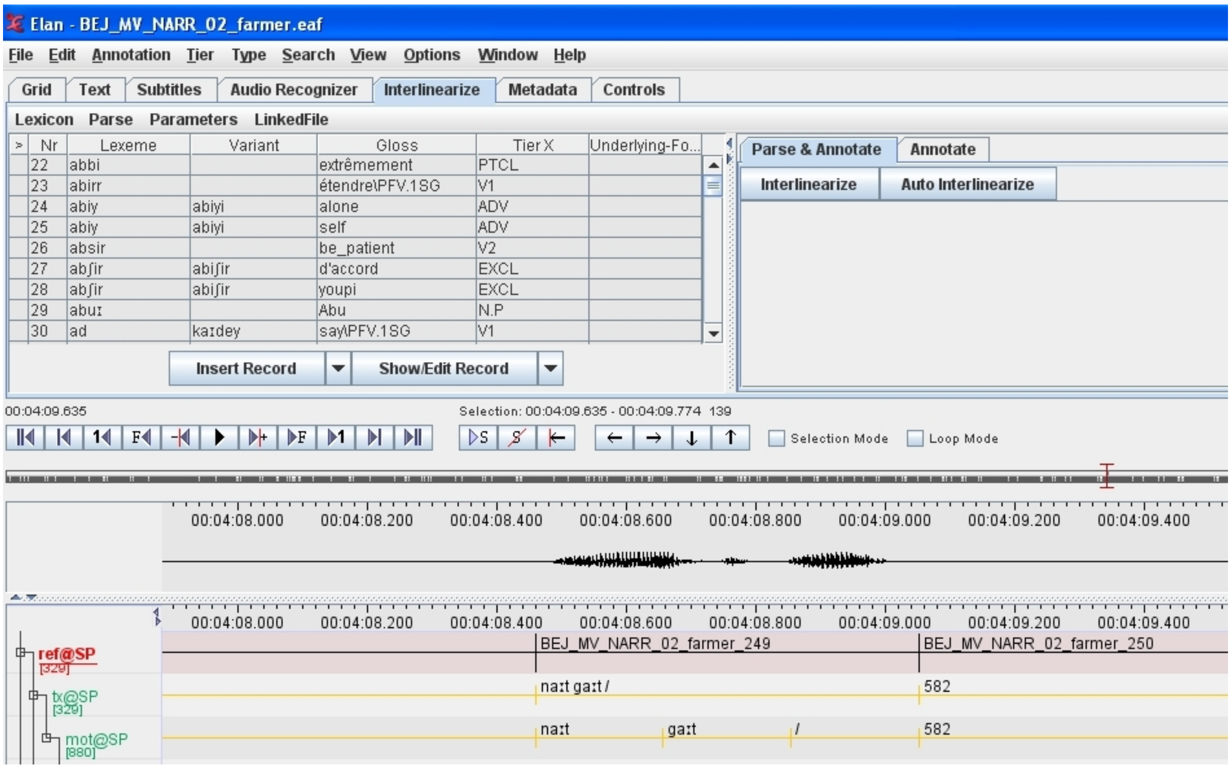
When all the correspondences are ok,

- Click on the « OK » button.

The program retrieves all the relevant data for interlinearization and copies them in a new ELAN lexicon file (that you will save); this newly created XML file (with **.eafl** extension) will be used as a lexicon for the interlinearizing process.

On the left part of the screen, a table showing the lexicon data is displayed with a menu above. On the right side, there are tabs and buttons for the interlinearizing process (*Segmentations*) and for the lexicon management (*Lexicon*).

Be aware that importation will not actually isolate the possibly multiple gloss of a lexeme, separated by a semicolon in the Toolbox gloss tier. Those entries must be edited in ELAN-CorpA to isolate each gloss and giving them their proper (rx) category.



>	Nr	Lexeme	Variant	Gloss	Tier X	Underlying-Fo...
	22	abbi		extrêmement	PTCL	
	23	abirr		étendre\PFV.1SG	V1	
	24	abiy	abiyi	alone	ADV	
	25	abiy	abiyi	self	ADV	
	26	absir		be_patient	V2	
	27	abʃir	abiʃir	d'accord	EXCL	
	28	abʃir	abiʃir	youpi	EXCL	
	29	abuː		Abu	N.P	
	30	ad	kaːdey	say\PFV.1SG	V1	

Self-opening of the lexicon

To the right of the *Lexicon* section menu, there is the *Linkedfile* menu. By default, the checkbox before the name of the lexicon is checked, so this lexicon will open automatically next time you open the ELAN file to which it is associated. If for some reason you don't want to open the lexicon automatically when opening the ELAN file, uncheck the lexicon in the *Linkedfile.*

Setting up the interlinearization process

Before launching the interlinearization process on the words of a tier, you have to choose this tier and define the associated annotation tiers. By default, those lines are « *mot* » for the line containing the words to be segmented and annotated, « *mb* » for the line containing the morpheme breaks, « *ge* » for the gloss of the morphemes, and

« *rx* » for the *grammatical* labels of the morphemes. If those tiers already exist, the current annotations will be overwritten during the interlinearization process.

If those tiers don't exist in the ELAN files, do the following:

Parameters, tier Parameters, configure interlinear Tiers

Choose the source tier to be segmented and annotated: *Choose interlinear tier* (*mot*)

Click on *OK*

Choose the labels for the morpheme breaks tier (*mb*), gloss (*ge*) and category (*rx*)

Click on the *Create tiers* button

The tiers are created, the process can start.

Remark that if a tier already exists with the same label as one of those you just enterede during the *configure interlinear tiers* process, a new tier will be created with this label ended by *–cp*, avoiding the loss of the original one. If you want to overwrite that existing tier, you should delete it beforehand.

Principles of annotation into ELAN

There are three kinds of entries (called here *Lexeme)* into the ELAN lexicon:

Lemma (base form chosen to represent the various forms of a word in context) – which may present alternate (contextual) forms known herein as *variants*,

Stem, which is a form which cannot appear on its own as a word; it needs a complementary affix. A stem may present a symbol (e.g _) to its left or right (or both) to distinguish it from a lemma if desirable,

or an *Affix*. Affixes represent all the morphemes that can be agglutinated to a lemma, a stem or another affix. By default, the affixes present a hyphen (-) to the left or to the right if they are respectively suffixes or prefixes. Clitics can be distinguished by the use of an equal sign (=) to the left or right, reduplication can also be represented by a tilde ~ at the beginning of the segment (cf. *parameters*)

Lookup at the words in the lexicon

The principle of the ELAN-CorpA annotation is, as a first step, to try and match the current word with the *lemma or stems* of the lexicon, or with their alternate forms (*variants*). If the word is found in the lexicon, the value of the fields *Lexeme*, *Gloss* and *Tier X* of the entry goes to the corresponding *mb*, *ge* and *rx* tiers under the current word in the annotation area. Notice that if the word corresponds to a variant of a lexeme, it is the underlying *lexeme* value that shows in the *mb* tier.

Now as a second step, if the word is not found, the parser tries to segment it using the affixes of the lexicon.

Segmentation

When a word is not found in the lexicon, the parsing process takes place, trying to match all the affixes (prefixes, suffixes, clitics, reduplications...) of the lexicon to the end and/or beginning of the word. When an affix matches, the parser isolates the affix, and the rest of the word is, in turn, searched in the lexicon, and so on. If the rest is not found, an asterisk will precede it, meaning it is a possible new entry. All the combinations are explored and the various segmentations are displayed in the *Segmentations* section. At this stage, to parse a new word, you should start by entering its affixes.

Affixes

To add a new affix in the lexicon, you can right-click the word containing this affix, in the segmentation area, and choose « *Insert a record* ». In the box where the word appears, delete everything but the affix. If it is a prefix, type a hyphen at the end, if it is a suffix, type the hyphen at the beginning.

When you launch the interlinearization process, the affix you entered is isolated from the word, then the rest is searched into the lexicon and if not found, the parser tries to find all the affixes that match the end or beginning of the rest, and so on. At the end, all the possible parsings of the word are displayed in the table of the segmentation area. If the parser did not give you the correct parsing, you have to add the (lexical or grammatical) morphemes that will fit this parsing, in the lexicon.

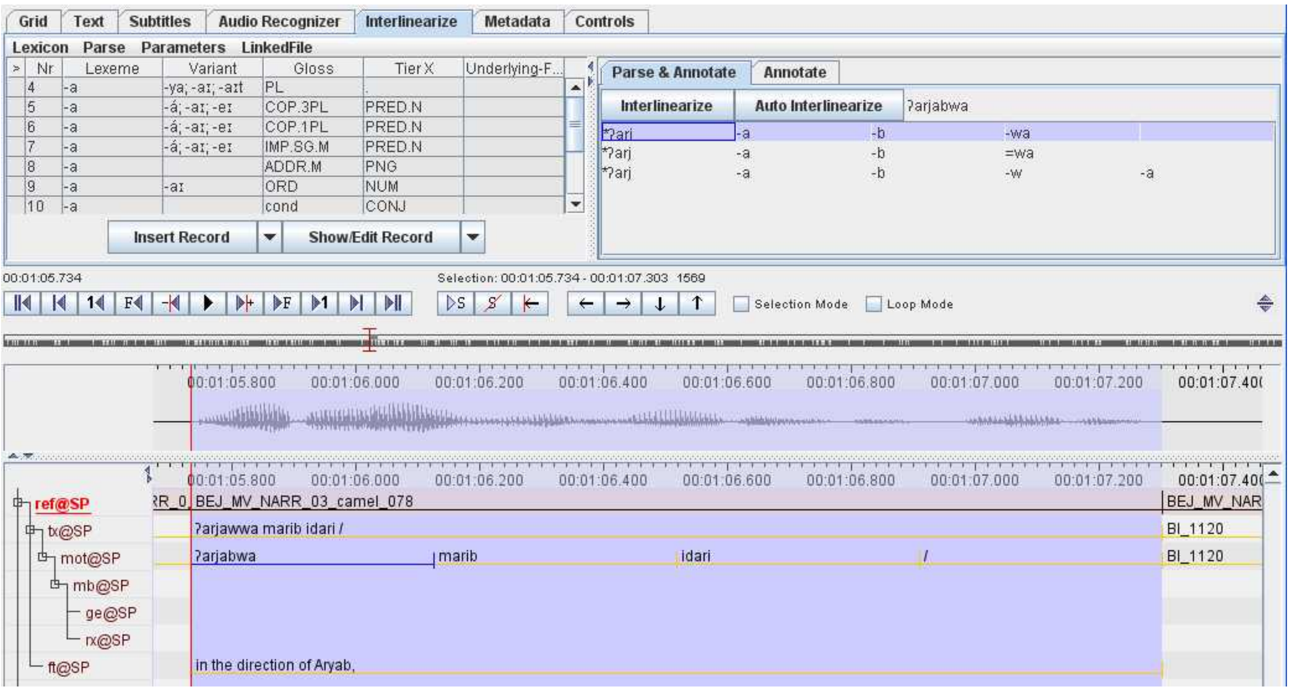
Launching the interlinearization process

The parser will search, one by one, all the words of the source tier in the lexicon, and if it doesn't find anything, it will try all the possible segmentations allowed by the current lexicon depending on the affixes it contains.

Click on the first word of the line to be annotated. Its segment will be underlined in blue.

Click on the « *Interlinearize* » button in the ***Segmentations*** section (to the right side of the screen).

The different possible morpheme breaks of the word are displayed in the ***Segmentations*** section, and now the lexicon will only show the entries that are involved in the morpheme break of the current word. The last unsegmentable segment is preceded by an asterisk, meaning that it has not been found in the lexicon.



>	Nr	Lexeme	Variant	Gloss	Tier X	Underlying-F...
	4	-a	-ya; -aɪ; -aɪt	PL	.	
	5	-a	-á; -aɪ; -eɪ	COP.3PL	PRED.N	
	6	-a	-á; -aɪ; -eɪ	COP.1PL	PRED.N	
	7	-a	-á; -aɪ; -eɪ	IMP.SG.M	PRED.N	
	8	-a		ADDR.M	PNG	
	9	-a	-aɪ	ORD	NUM	
	10	-a		cond	CONJ	

ʔarjabwa

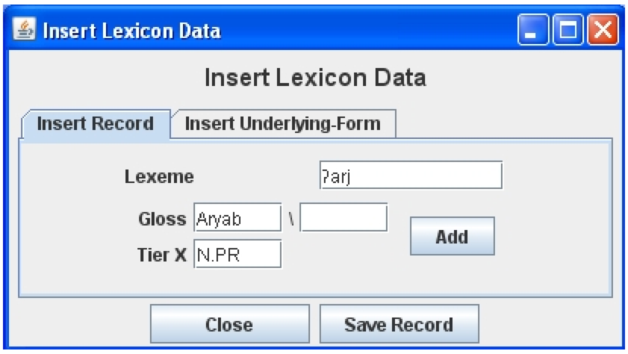
*ʔarj	-a	-b	-wa	
*ʔarj	-a	-b	=wa	
*ʔarj	-a	-b	-w	-a

00:01:05.734   Selection: 00:01:05.734 - 00:01:07.303 1569

ref@SP	R_0	BEJ_MV_NARR_03_camel_078			BEJ_MV_NAR
tx@SP		ʔarjawwa marib idari /			BI_1120
mot@SP		ʔarjabwa	marib	idari /	BI_1120
mb@SP					
ge@SP					
rx@SP					
ft@SP		in the direction of Aryab,			

In the above example, the word **ʔarjabwa** presents three possible segmentations. The suffixes **-a**, **-b** and **-wa** found in the lexicon lead to the isolation of a possible stem ***ʔarj**.

Adding an entry into the lexicon (*Insert record*)



Lexeme: ʔarj
Gloss: Aryab \
Tier X: N.PR

To add a new word to the lexicon, whether a lexeme, a stem or an affix, click on the « *Insert record* » button, in the *Segmentations* area, or here just Right-click on the word preceded by an asterisk (in the above example, ***ʔarj**)

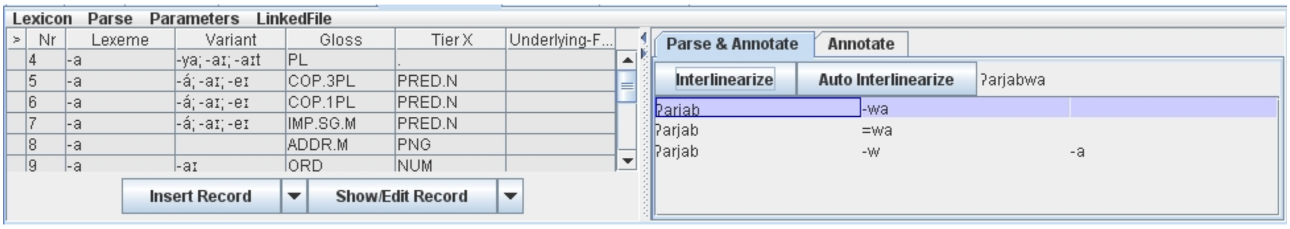
Selected tab: *Insert record*

A window appears with the selected word. It can be modified. For example, here the word to be added is ʔarjab which is glosed as the proper noun Aryab.

Click on *Save Record* button

If the morpheme you are glossing contains morphological features which cannot be segmented, or that you do not want to isolate as a separate morpheme, you can use the box to the right of symbol '\' to enter those features. Notice that you do not have to type the delimiter (\) before the grammatical label, it will be added automatically in the annotation line.

Once the entry is created, the process may be launched again with the « *Interlinearize* » button.



>	Nr	Lexeme	Variant	Gloss	Tier X	Underlying-F...
	4	-a	-ya; -aɪ; -aɪt	PL	.	
	5	-a	-á; -aɪ; -eɪ	COP.3PL	PRED.N	
	6	-a	-á; -aɪ; -eɪ	COP.1PL	PRED.N	
	7	-a	-á; -aɪ; -eɪ	IMP.SG.M	PRED.N	
	8	-a		ADDR.M	PNG	
	9	-a	-aɪ	ORD	NUM	

ʔarjabwa

ʔarjab	-wa	
ʔarjab	=wa	
ʔarjab	-w	-a

Here, as the new word *ʔarjab* has been entered into the lexicon, three new possible segmentations remain.

Selecting the segmentation and the gloss

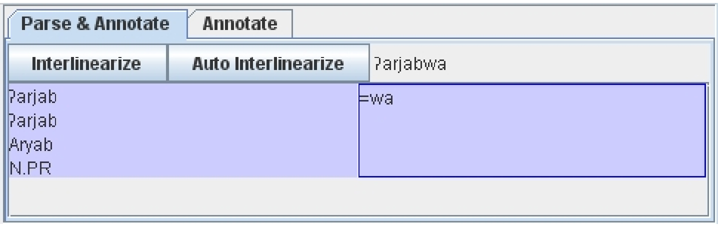
When there are several possible segmentations for the word, you have to choose the one that fits

Double-click on the appropriate first segment of the correct segmentation line (here *ʔarj on the first line).

The lexicon narrows down to the corresponding entries (which can be multiple in case of polysemy or homonymy).

Double-click on the correct entry in the lexicon area, depending on the gloss and the category.

The annotation of the first segment is displayed under it, in the *Segmentation* area, then the next segment is selected.

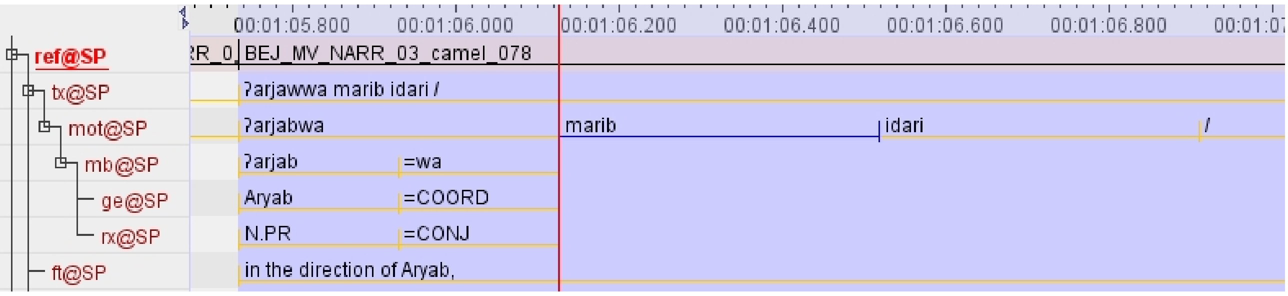


ʔarjabwa

ʔarjab ʔarjab Aryab N.PR	=wa

And so on: a double-click on the selected segment will narrow down the lexicon to the corresponding entries, then a double-click on the correct entry in the lexicon will display the values of this choice under the current segment. (Notice that, for saving clicks and time, when the next segment is automatically selected in the segmentation area, you can double-click directly on the correct lexicon entry without double-clicking on the current segment in the *Segmentation* section; in this case the lexicon will stay fully displayed.)

When the last segment of the current word is annotated, the chosen annotations are transferred under the word (in the annotation area), each in its own tier, and the next word is selected.

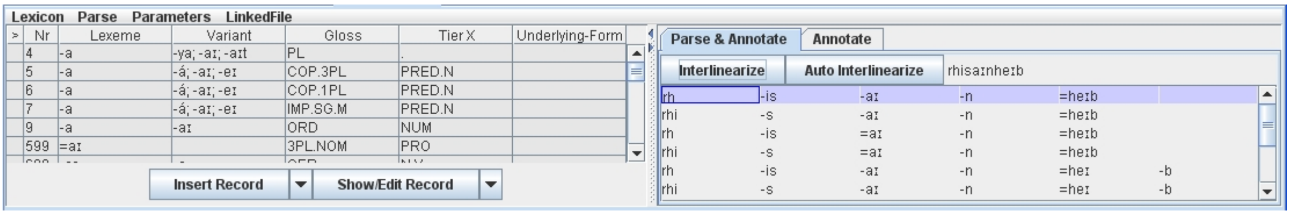


ref@SP	ʔR_0 BEJ_MV_NARR_03_camel_078	
tx@SP	ʔarjawwa marib idari /	
mot@SP	ʔarjabwa	marib | idari | /
mb@SP	ʔarjab	=wa
ge@SP	Aryab	=COORD
rx@SP	N.PR	=CONJ
ft@SP	in the direction of Aryab,	

Extended features of the parsing

Morphophonology (lemma and variant)

When a morphophonological change appears at the boundary of a stem and an affix (or of two successive affixes), you should always bear in mind that the parser searches for a match between what remains to be treated and the lexicon entries at *Lexeme* or *Variant* level.



>	Nr	Lexeme	Variant	Gloss	Tier X	Underlying-Form
	4	-a	-ya; -aː; -aːt	PL	.	
	5	-a	-á; -aː; -eː	COP.3PL	PRED.N	
	6	-a	-á; -aː; -eː	COP.1PL	PRED.N	
	7	-a	-á; -aː; -eː	IMP.SG.M	PRED.N	
	9	-a	-aː	ORD	NUM	
	599	=aː		3PL.NOM	PRO	

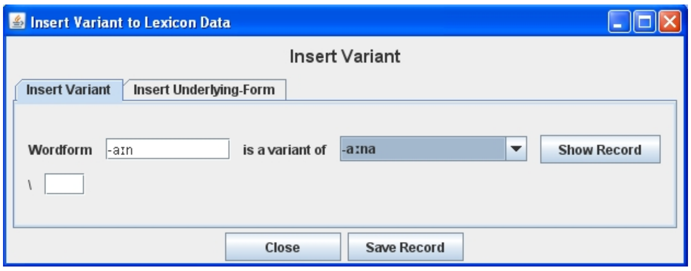
rhisaːnheːb

rh	-is	-aː	-n	=heːb	
rhi	-s	-aː	-n	=heːb	
rh	-is	=aː	-n	=heːb	
rhi	-s	=aː	-n	=heːb	
rh	-is	-aː	-n	=heː	-b
rhi	-s	-aː	-n	=heː	-b

In the example above, the parser cannot give the correct segmentation of the word 'rhisa:nhe:b' (should be rh –is -a:na =he:b) because of the collapse of the vowel 'a' of the suffixe '-a:na' before the last clitic '=he:b'. When '=he:b' is isolated, for the parser being able too correctly isolate the suffixe '-a:na ' already in the lexicon, we can enter '-a:n' as a variant of it..

Adding a variant to an entry

Change the « *Insert Record* » button into « *Insert Variant* » button with the downward little arrow, then click on it. Enter the variant form (-a:n) and select the associated entry (-a:na).

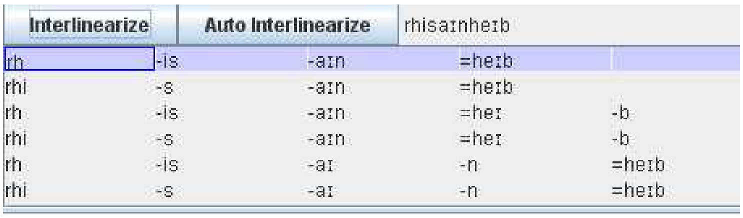


Wordform -aːn is a variant of -aːna

\

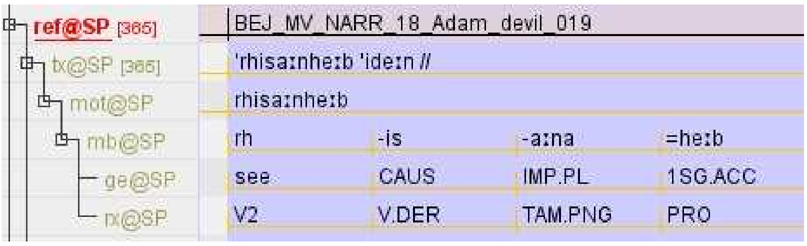
Save the record

As the parser searches for a match at the level of the *lexeme* or the *variant* level of the entries, it will now propose the variant '-a:n' of the entry '-a:na' as fitting the match.



rhisaːnheːb

rh	-is	-aːn	=heːb	
rhi	-s	-aːn	=heːb	
rh	-is	-aːn	=heː	-b
rhi	-s	-aːn	=heː	-b
rh	-is	-aː	-n	=heːb
rhi	-s	-aː	-n	=heːb

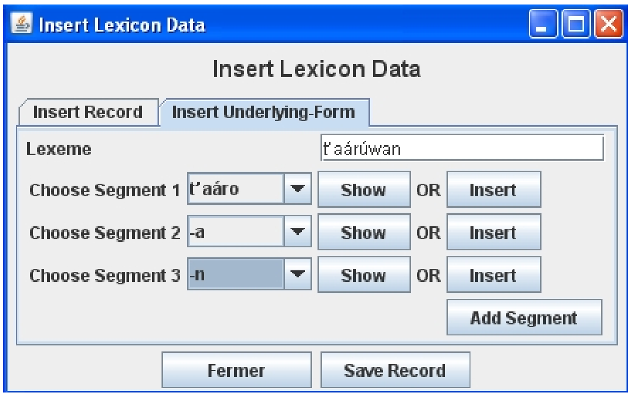


ref@SP [365]	BEJ_MV_NARR_18_Adam_devil_019			
tx@SP [366]	'rhisaːnheːb 'ideːn //			
mot@SP	rhisaːnheːb			
mb@SP	rh	-is	-aːna	=heːb
ge@SP	see	CAUS	IMP.PL	1SG.ACC
rx@SP	V2	V.DER	TAM.PNG	PRO

The annotation may then continue by validating the correct entries of the lexicon, and the morphem ‘-a:n’ will be returned with its *lexeme* base form value ‘-a:na’ to the *mb* annotation tier.

When the morphophonology is too complex for the parser to give the correct segmentation, even with alternate forms of affix or stem, it is always possible to give the correct segmentation directly into an entry of the lexicon. But be aware that the various segments composing the current entry have to already exist in the lexicon.

Inserting an underlying form



Lexeme: t'aárúwan
Choose Segment 1: t'aáro
Choose Segment 2: -a
Choose Segment 3: -n

Right-Click on the word to enter in the segmentation area,, or or Click the « *Insert Record* » button .

Select the « *Insert underlying form*» tab

Find the first segment (here *t’aáro*) in the drop list in front of *Choose Segment 1*

idem for segment 2, (here *-a*)

then add a segment if necessary by clicking the *Add button*, and choose the segment (here *–n*)

Validate with the *Save Record* button, then close the window.

In case of homonymy or polysemy of the lexical entries, it can be difficult to choose the right morpheme among several for the current segment. The *Show* button allows displaying the content of the lexical entry to verify if it is the correct one.

From this window, it is also possible to add an entry that is lacking in the lexicon and would be necessary for the segmentation.

Click on the *Insert* button on the same line as the current segment. A little window *Insert Morpheme* will open allowing you to add an entry in the lexicon. Validate with OK. This entry will constitute the new segment for the word to be segmented.

It should be noted that this method of giving the parser an ad hoc segmentation for a word, should be avoided as much as possible and only be used when the parser fails to give the correct segmentation with regard to the content of the lexicon (lemma, variants and affixes). As a matter of fact, this kind of specific entry of the lexicon only resolves the segmentation of one word (or maybe a complex combination of affixes). Recall that the principle of the parser consists in providing the lemma in one part and the affixes in the other part (with possible alternate forms), a method which is less time consuming and more consistent and less error-prone.

The auto-interlinearization function

To save time in the process of interlinearization, it is possible to choose the automatic process which will continue word after word, whenever the segmentation of the words is possible, unique and without ambiguity in the glossing.

Launching of the auto-interlinearization process

This function can be launched from any word in the annotation base tier (here the *mot* tier).

Click on the first word where the process must start (the base line of the word turns blue)

Click on the *Auto-Interlinearize* button

The segmentation starts, and will continue word by word until a word cannot be segmented or until an ambiguity arises.

Parse-lexicon

Once the annotation of a text is completed, another type of lexicon can be created with all the words of the text as entries and their glossed segmentations as data. This lexicon may be saved as a *Parse lexicon*, or merged with an older one. It can be used then for increasing the speed of the *auto-interlinearize* process.

Creating, merging, opening a *Parse lexicon*

To export the lexicon of the words and their glossed segmentation, go to the *Lexicon* area and choose the *Parse* menu

Parse, Export Parse data

Browse to the destination folder and give the file a name. The extension **.eafp** will be added.

To merge the current segmentations and annotations of the text with an older *Parse lexicon*, choose the *Parse* menu in the *Lexicon* area :

Parse, Export Parse data

Browse to the destination folder and select the parse file in which you want to merge the new parsing.

To open a *Parse* lexicon for the *auto-interlinearize* process, choose the *Parse* menu

Parse, Open Parse data

Self-loading of a *Parse lexicon*

By default, once a *parse* lexicon has been created for an ELAN annotation file, it will be automatically opened next time the annotation file is opened. If you want to avoid this, you have to delete the link between these two files,

Go to the *Linked File* menu in the *Lexicon* area
Uncheck the checkbox before the name of the *Parse* file
The *Parse* file will not be loaded next time the ELAN file is opened.

Saving the linked files

When you close the ELAN annotation file, a window will appear allowing you to unselect the linked files (lexicon and/or parse) you do not want to save (for any reason). Normally you should save the lexicons.

Anyway, it is advisable to save the ELAN lexicon regularly during the intelinearizing process with the *Save* item in the *Lexicon* menu, because the Ctrl/S shortcut in ELAN will not save the lexicons.