Interlinearization mode in ELAN 5.0.0-alpha: workflow, known issues, limitations

Modified: 17 Jan 2017

This document contains some additional information on the 5.0 alpha release (additional to what is in the manual, the primary source of information).

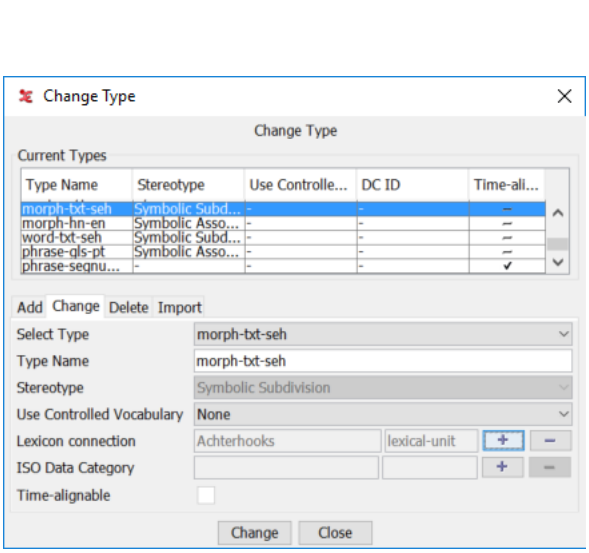
Connecting tier types to lexical entry fields

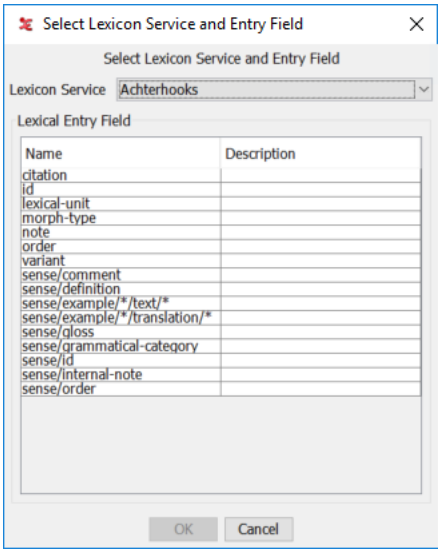
Part of setting up analyzers for your document is connecting tier types with a corresponding field within the structure of lexical entries. In the **Add** (or **Change**) **Tier Type** window, create or select a tier type, click the **+** button in the Lexicon connection section. This will produce a new window with at the top a dropdown box labelled Lexicon service. If you have a lexicon in ELAN's lexicon component format it will be listed here. After selecting a lexicon the entry fields are listed in the table. Select one and click **OK** to connect the tier type to the selected field.

In the case of assisted, semi-automatic parsing and glossing, the following connections can be established

Tier type	Entry field
morphemes, morpheme breakdown	lexical-unit (the main field (lemma, headword) of an entry)
part-of-speech	sense/grammatical category
gloss types	sense/gloss
…	…

See the Analyzers section for what is currently hardcoded, see the Lexicon Component section for the (fixed) structure of a lexical entry.





The Lexicon Connection property and the list of entry fields available in an ELAN lexicon (by default)

Lexicon Component + Integration

The lexicon component consists of a simple lexicon xml schema and a few user interface elements to create, import and export lexicons and to show, sort and edit lexical entries. At this point in time all usable fields are fixed but the schema allows for the (future) use of custom fields.

Lexicon structure

The names of the fields and the overall structure of entries is quite similar to the LIFT (Lexicon Interchange Format) structure, but simplified and more limited.

The main fields in a lexical entry are:

```
entry
      lexical-unit                    (1)
      morph-type                      (0 or 1)
      citation                        (0 or 1)
```

```
variant                          (0 or more)
     sense                       (1 or more)
          grammatical-category   (1)
          gloss                  (1 or more)
          order                  (1)
```

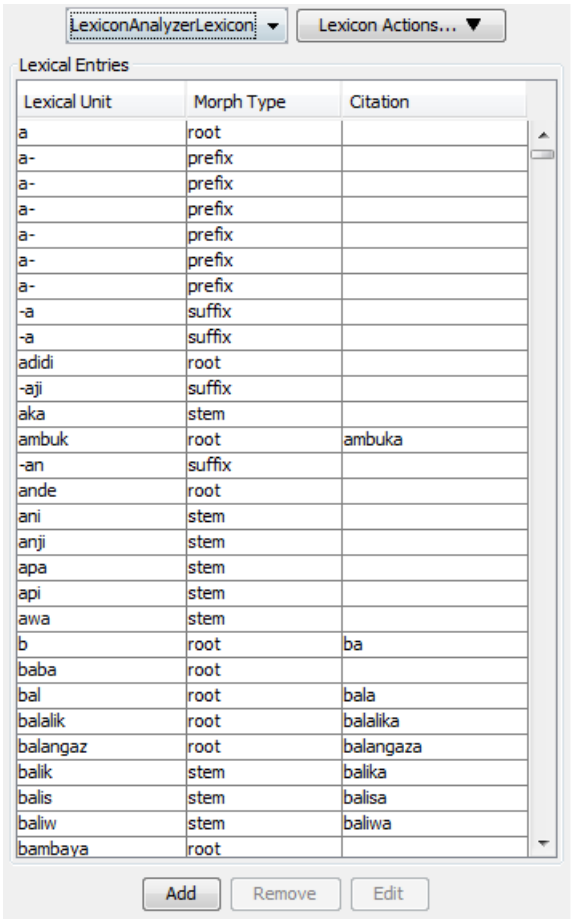
The main field is `lexical-unit` (equivalent to lemma, headword, the primary lexical form). `morph-type` indicates the word part (e.g. stem, prefix, suffix), analyzers can use this information when processing the input text. The `grammatical-category` field is the category of the lexical item, the part of speech.

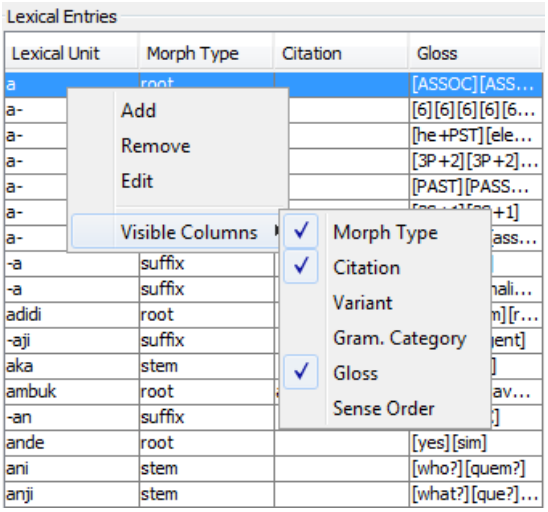
The Lexicon user interface

The user interface to the lexicon consists of a panel with a table listing the entries of the selected lexicon, a drop-down box listing all available lexicons and a pseudo-menu labelled **Lexicon Actions**.

The table displaying the lexical entries of the selected lexicon can be customized via the (right-mouse button) context menu where columns for (some) entry fields can be added or removed. These fields are in the **Visible Columns** sub-menu. The entries can be sorted, ascending or descending, by clicking on the header of a column. This only effects the display on screen, not the order in which entries are stored in the lexicon. Columns representing a field that can occur more than once in an entry will show each value within square brackets (e.g. [aa][bb]). A few columns can be edited directly in the entry table (the fields that can occur only once in an entry except for `lexical-unit`). All entry fields can be edited in a separate popup window which appears when choosing the Edit menu item in the context menu (or the Edit button). A new entry is first defined in a separate window which appears when choosing the **Add** option; the Add Entry window enforces that all mandatory fields are filled in. When a lexicon is edited the name of the lexicon will be shown in a different color in the list, until it is saved.

The **Lexicon Actions…** drop down menu contains items to create, open, close, save, import and export a lexicon. The **Save** and **Open** items concern lexicons in the Lexicon Component's xml format. Supported formats for importing are the LIFT (`.lift`) format and the CorpAfroAs (`.eafl`) format. The only export format is `.lift`.





Lexicon data

When a lexicon is created it is stored in a default folder in ELAN's data folder. When a Lexicon component's `.xml` file is opened from a different location (via the **Open Lexicon…** option) it is copied to the default folder and that will become the working copy. Similarly when a lexicon is imported (i.e. converted) from a different format, the result is stored in that same default folder. The lexicon folder is named **LexanLexicons**, the location of ELAN's data folder depends on the platform, see "1.1.2. Special ELAN data folder" in the manual. It will be important to back up the contents of this folder regularly for anyone who creates and edits lexicons in ELAN. The **Save Lexicon As** action can be used for this.
The XML schema can be found here: http://www.mpi.nl/tools/elan/LexiconComponent-1.0.xsd

Note

In ELAN it is possible to have multiple transcription windows open at the same time but lexicons are available "globally" or application wide. In principle changes to the lexicon made in one window will be visible in the same lexicon in another window. Problems can occur if ELAN has been launched multiple times (which is possible on all platforms); in that case changes in one window will not be visible in another and saving the lexicon in one window might overwrite changes saved from within another window.

To do (among other things)

- Support for import of Toolbox dictionaries
- Sorting a lexicon after import (and storing the entries in that order)
- Use of controlled vocabularies for specific entry fields
- Custom lexical entry sort order
- Support for custom fields
- Lexicon edit window for more convenient editing of multiple entries

Analyzers and API

Four analyzers are currently distributed with ELAN:

The **Whitespace analyzer** splits the input string using the white space as delimiter and returns multiple output strings. There is no support yet for special treatment of punctuation marks.

The **Parse analyzer** parses the input based on entries available in the lexicon. A lot is still hardcoded:

- Looks in the `lexical-unit` entry field when trying to parse the input
- Also looks in the `variant` field, if configured to do so
- Uses the `morph-type` field to determine the type of a unit and recognizes and supports the values `stem`, `root`, `suffix` and `prefix`. A hyphen ("-") is detected as part of prefixes and suffixes and is removed in the parsing process
- Supports fragment replacement in the parsing process and uses the `replace` field for the replacement
- Uses the string `++ABORT++` to indicate that too many parses are found and partial results are returned

The Parse analyzer produces suggestions to the user to choose from. Statistics on selected parses are stored and used to re-order suggestions the next time, placing the most frequently used parse on top.

This analyzer has at the moment two configurable options in its configuration panel; one to determine whether or not to include variants in the parsing process and one to set the maximum number of parses (default is 256).

The **Gloss analyzer** performs a lookup of an input string and returns the value of one or more fields of the found entries. The input string is (currently hardcoded) expected to be from a `lexical-unit` field and the returned values are taken from the `sense/gloss` field(s). If no matches found the string `***` is returned. Like the Parse analyzer the Gloss analyzer stores statistics on selected suggestions.

The **Lexicon analyzer** combines the functionality of the Parse and Gloss analyzer to produce suggestions for two target tiers. Potentially this can lead to (too) many suggestions.

Analyzers can store settings and statistics in the folder named **Analyzers** inside the ELAN data folder.

Analyzers are implemented and distributed as extensions of ELAN and it will be possible for others to implement their own analyzers. But the API for analyzers is not yet finalized and not yet documented.

To do (among other things)

- Make the analyzers more flexible, more customizable
- Extend and publish the API

Interlinear Viewer + Editor

The annotations are shown in an interlinear text style in this viewer and editor. There are only a few options to modify the appearance of the annotations; the font size can be changed, the visible tiers can be set via the context menu of the tier name area on the left and the width of the tier name area can be changed by dragging the divider left of the Interlinearize button in the panel above the annotation area.

To do (among other things)

- Visualization improvements (margins, bounding boxes optional etc.)
- Improved manual editing, keyboard shortcuts
- Option to add a manually created annotation to the lexicon
- Option to play the selected segment (if there linked media files)
- A mode to present parses incremental, left to right (first all candidates for the first morpheme, then after a choice the filtered set of possible second morphemes etc.)

Suggestion Window

The suggestions are currently presented as a list containing all full parses, the layout is similar to the layout of the Interlinear viewer, including the target tier names. How to interact with the suggestion window is described in the manual.

To do (among other things)

- Visualization improvements (margins, bounding boxes optional etc.)
- Tooltip per suggestion fragment to see the whole entry (or relevant parts of it)
- Option to show parses incremental, see above