Data transformations for processing interlinear glossed texts in SIL FLEx

Alexandre Arkhipov (Universität Hamburg / Lomonosov Moscow State University), 24/11/2020

In this talk, we will discuss import and export between FLEx (SIL Fieldworks LanguageExplorer) and other software (like ELAN, Praat and office suites).
We will focus on settings and workflows that facilitate seamless conversions, paying attention to specifics of FLEx built-in segmentation and mismatches between its input and output flextext documents as well as between export into different office formats.

1 RTFM

Software home pages:

FLEx:	https://software.sil.org/fieldworks/
ELAN:	https://archive.mpi.nl/tla/elan
Praat:	http://praat.org/
SayMore:	https://software.sil.org/saymore/

Generally, use **recent stable versions** of all the programs.

In case of problems, and also **before** you've got problems, it's always a good idea

- to **read Help**
- to **read Release notes / Known issues** on resp. websites
  E.g. https://software.sil.org/fieldworks/release-notes/,
  https://archive.mpi.nl/tla/elan/release-notes;
  https://software.sil.org/fieldworks/download/fw-9011/: Known Issues

  Specifically for FLEx:
- to search for similar issues described in FLEx user group:
  http://groups.google.com/group/flex-list
- to ask your question in the FLEx user group: flex-list@googlegroups.com
- if you seem to have met a bug / an error, write an email / send an automated error report to flex_errors@sil.org or FlexErrors@sil.org (Ann Bush)

2 FLEx main functions & features

- Lexicography + Interlinear text analysis (+ Grammar sketches + Discourse analysis…)
- **One project = One language / variety**
- Can be used collaboratively
- Manual or automated morphological analysis
- Import/export options connecting with other software
- Long history of development…

3 FLEx project & backup files

- Main project file: *.fwdata (holds almost everything). Do not touch.
- Changes are saved automatically (no Save button)
- Backup file: *.fwbackup (everything, zipped) > Backup & Restore
  Share with others (one way – no work in parallel!)
  Revert to a previous state
- Project/backup file as a "(lady's) handbag".
  Not the same as either exported texts or exported dictionaries!
- Data Model changes
  One user can upgrade anytime. In case of collaborative work (more than one user OR computer), everyone must upgrade together, unless there is no Data Model change – watch Release Notes! (but still, together is best).

4 The Writing System (WS)

- The most important small thing which can break things (or not)
- Identifier which encodes Language + Script + other variation parameters
  - Archi + IPA
  - Selkup + Cyrillic
  - English + Australia
- WS does not affect which characters can be used – it's Unicode anyway.
  It does affect which characters are counted as word-forming, the linked keyboard layout, the sorting order etc. And, most importantly, how FLEx treats what's written.
- One vernacular WS + many analysis WS
  - It is technically possible to have more than one vernacular WS. **DON'T do that!**
    Everything you want to analyze (gloss) should be in one WS.
    Everything else, even in the same language, should be in in an analysis WS. (E.g.: Cyrillic original, native speakers's unedited transcription, etc.)
  - Some fields only allow one WS inside. Add more fields to have them in alternative WSs (e.g. Russian gloss + English gloss + German gloss)
- Multi-WS fields
  - Some fields, including the text baseline and the translations, allow embedded WSs (i.e. a mixture of WSs). This can help but also can cause trouble.
  - Only text formatted with the vernacular WS in Baseline tab can be analyzed (glossed).
    - ⇨ If you can't analyze the text, check WS in baseline
    - ⇨ If you want to prevent text from being glossed, change its WS (to anything except vernacular)
  - Adding a "fake" WS is a workaround to have a custom tier.
    - ⇨ alternative transcriptions as "translations" (on phrases)
    - ⇨ covert categories as "glosses" (on roots) – e.g. gender/nominal class
  - FLEx 9.0+ allows for custom sentence-level tiers (not word-level), only in 1 analysis WS.
  - **Copy-Paste within FLEx** brings the WS along. Can be harmful e.g. with translation tiers.
    A known bug: when exporting as flextext XML, the "lang" attribute of a translation tier is determined by the **WS of the first character in the field** (not by what the field is meant to be). => can lead to two German translations and no English, etc.
- WS codes are crucial for import/export to function correctly

5 FLEx interlinear tiers & flextext XML

- [Bird, Bow & Hughes 2001] model
- interlinear-text > paragraph > phrase > word > morph
- At each hierarchy level, items with **type** and **lang** (=WS) attributes
  - interlinear-text: title, source,…
  - phrase: “**gls**”=Free translation, “**lit**”=Literal translation, “**note**”=Note, “**segnum**”=number +custom user types
  - word: “**gls**”=word gloss, “**ps**”=part of speech
  - morph: core interlinear tiers (“**txt**”=Morphemes, “**cf**”=Lex. Entries, “**gls**”=Lex. Gloss, “**msa**”=Lex. Gram. Info)
  - Where’s the Baseline? Type=“**txt**”, on phrase (import) or on word (export)
- **phrase** can bear time-alignment and speaker attributes (imported from ELAN)
- Word analyses are not stored with the text but stored apart and only referred to.
  So a text is assembled for display & export but does not exist as a single object.

6 ELAN > FLEx via flextext XML

- First **create a FLEx project**. Setup WSs and create a sample text. Export the text as flextext to be sure which codes to use where.
- **Only sentence-level tiers can be imported** (no words, no morphs/glosses).
- Create an ELAN template or use an existing file. Future Baseline (the line to be analyzed in FLEx) should be the **top-level tier** (directly time-aligned). All other tiers should be its children.
  - Alternatively, there can be a time-aligned reference tier at top level, and all the other tiers **including the future Baseline** should be its children.
  - Rename the tiers according to the pattern: <anything>-**<type>**-**<lang>**
  - E.g. A_phrase-**txt**-**sel** (Baseline), Rough_translation-**lit**-**ru**, ref-**segnum**-**en** (reference)
  - In case you have complex WS codes which contain a hyphen, e.g. **sel-Latn-x-source**, manual adjustments need to be done while exporting from ELAN (see below). Otherwise, default settings should work.
  - For dialogues, set speaker properties and create a set of tiers for each speaker (Tier > Add new participant).
- All kinds of text processing/editing are much easier in ELAN (or before) than in FLEx.
  - Splitting/merging sentences
  - Duplicating tiers
  - Transliterating (e.g. Cyrillic > Latin, project transcription > IPA)
  - Search & Replace

6.1 Segmentation, punctuation & time-alignment in FLEx

- **FLEx always splits sentences** at sentence-final punctuation, which is one of "**. ? ! §**" (possibly in combination with other punctuation such as quotation marks). If an ELAN annotation is split automatically by FLEx after import, the **time-alignment is lost** on this sentence.
  => You must do it before importing.
- Manually splitting or merging sentences in FLEx Baseline **also destroys time-alignment**.
  => You must do it before importing.

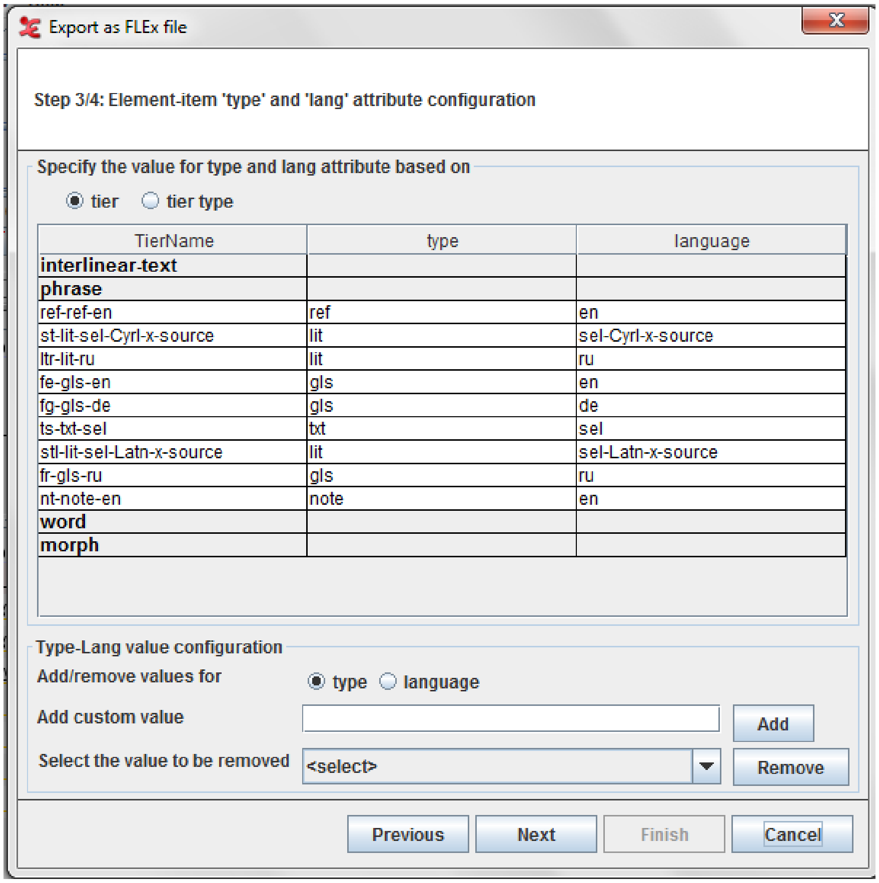
- Time-alignment can be fixed manually after exporting from FLEx, but it is (i) time-consuming and (ii) must be repeated each time you're exporting a particular text.
  => I guess you got it :-)
- Time-alignment and speaker info is not shown in FLEx interface (only stored). In dialogues, it might be worth to have speaker marked e.g. in notes while glossing (and remove after exporting from FLEx).
- By default, each ELAN annotation is imported as a paragraph consisting of one sentence (with a line break in the end, equivalent to pressing Enter in the Baseline tab).
- Decide on punctuation characters/delimiters as early as possible before you start importing (and also transcribing).
  - If you are using standard orthographical punctuation, including any of "**. ? ! §**", make sure it only appears at the end of an ELAN annotation. If any of them occur in the middle, split the annotation manually (split all tiers into two + remove extra half from each annotation). Note that three dots ("**...**") or a single ellipsis character ("**…**") are tolerated in the middle of a sentence.
  - If you are using some other segmentation marks, a "**§**" can be added at the end of each annotation and removed after processing in FLEx.

6.2 Other operations

- Duplicating tiers
  - E.g. leave a copy of native speaker's translation in parallel
- Transliterating (e.g. Cyrillic > Latin, project transcription > IPA)
  - E.g. with SIL Converters: TECkit maps can be run on selected elements in XML, including ELAN *.eaf.
- Search & Replace
  - Use regex!
  - Lookahead and lookbehind to handle replace with literals only.

6.3 Export ELAN > FLEx

1. Export as flextext file: File > Export as… > FLEx file, then **Next**, **Next**, (check screen). The screen on Step 3/4 **should have** all your tiers listed and all types and language codes treated correctly:



Step 3/4: Element-item 'type' and 'lang' attribute configuration

Specify the value for type and lang attribute based on

● tier ○ tier type

TierName	type	language
interlinear-text		
phrase		
ref-ref-en	ref	en
st-lit-sel-Cyrl-x-source	lit	sel-Cyrl-x-source
ltr-lit-ru	lit	ru
fe-gls-en	gls	en
fg-gls-de	gls	de
ts-txt-sel	txt	sel
stl-lit-sel-Latn-x-source	lit	sel-Latn-x-source
fr-gls-ru	gls	ru
nt-note-en	note	en
word		
morph		

Type-Lang value configuration

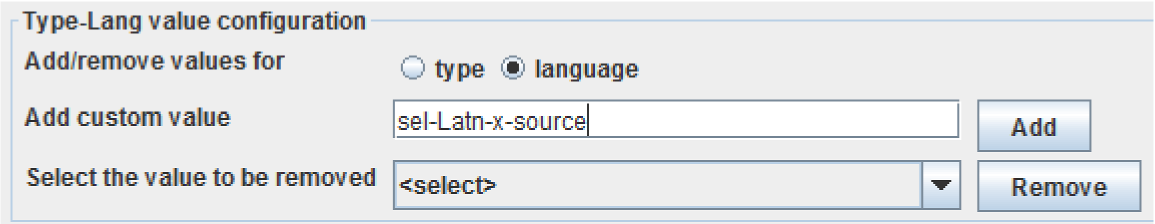
Add/remove values for ● type ○ language

Add custom value

Select the value to be removed <select>

2. **NB:** Since ELAN v5.4, the treatment of tier names changed, and complex “language” attributes need manual adjustment as follows:

E.g., **stl-lit-sel-Latn-x-source** will be by default split into **stl-lit-sel-Latn-x** and **source**, instead of **lit** and **sel-Latn-x-source**. So one needs to add **sel-Latn-x-source** as a “language” custom value (also **sel-x-source** if necessary):

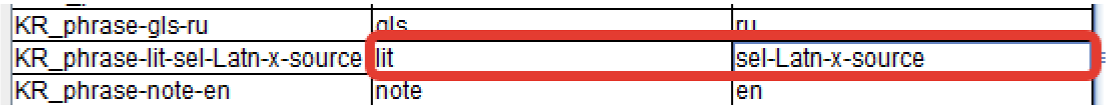

Type-Lang value configuration

Add/remove values for ○ type ● language

Add custom value sel-Latn-x-source

Select the value to be removed <select>

Then choose type “lit” and language “sel-Latn-x-source” for the **-lit-sel-Latn-x-source** tier (also “lit” and “sel-x-source” for **-lit-sel-x-source** if needed).



KR_phrase-gls-ru	gls	ru
KR_phrase-lit-sel-Latn-x-source	lit	sel-Latn-x-source
KR_phrase-note-en	note	en

3. Click **Next**, browse to folder and specify file name, **Finish**.
4. Use File > Import > FlexText Interlinear… in FLEx, browse to the flextext file, confirm.

7 Praat TextGrid > ELAN

- All imported tiers will be top-level time-aligned tiers.
- Create new tier type with Symbolic Association stereotype. Baseline shall remain at top-level, while all the others should be copied as its children with the new tier type.
- Proceed as with regular ELAN files (rename tiers, watch out for punctuation, export as flextext).

8 FLEx > ELAN via flextext

- Export interlinear from the Analyze tab.
    - The export will contain those and only those tiers which are currently displayed.
    - Make sure only one text is selected during export (bug in 9.0, fixed in 9.0.11).
- Normally, a flextext export from FLEx is ready to be imported back into ELAN.
- If time alignment has been lost, it needs to be fixed in the *.flextext XML
    - find phrase elements missing begin-time-offset, check next phrase
    - look in the original ELAN file for time values (in milliseconds)
    - fill in the missing attributes
- flextext export does not have a phrase-level “txt” element (baseline for the whole sentence).
    - Copy the word-level tier (e.g. **A_word-txt-sel**) as a Symbolic Association child under a phrase-level parent.
    - Remove extra spaces before punctuation with Search & Replace.
    - In a similar way, morpheme annotations can be joined into word-long annotations.