MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

WWW.MPI.NL

Interlinearization in ELAN

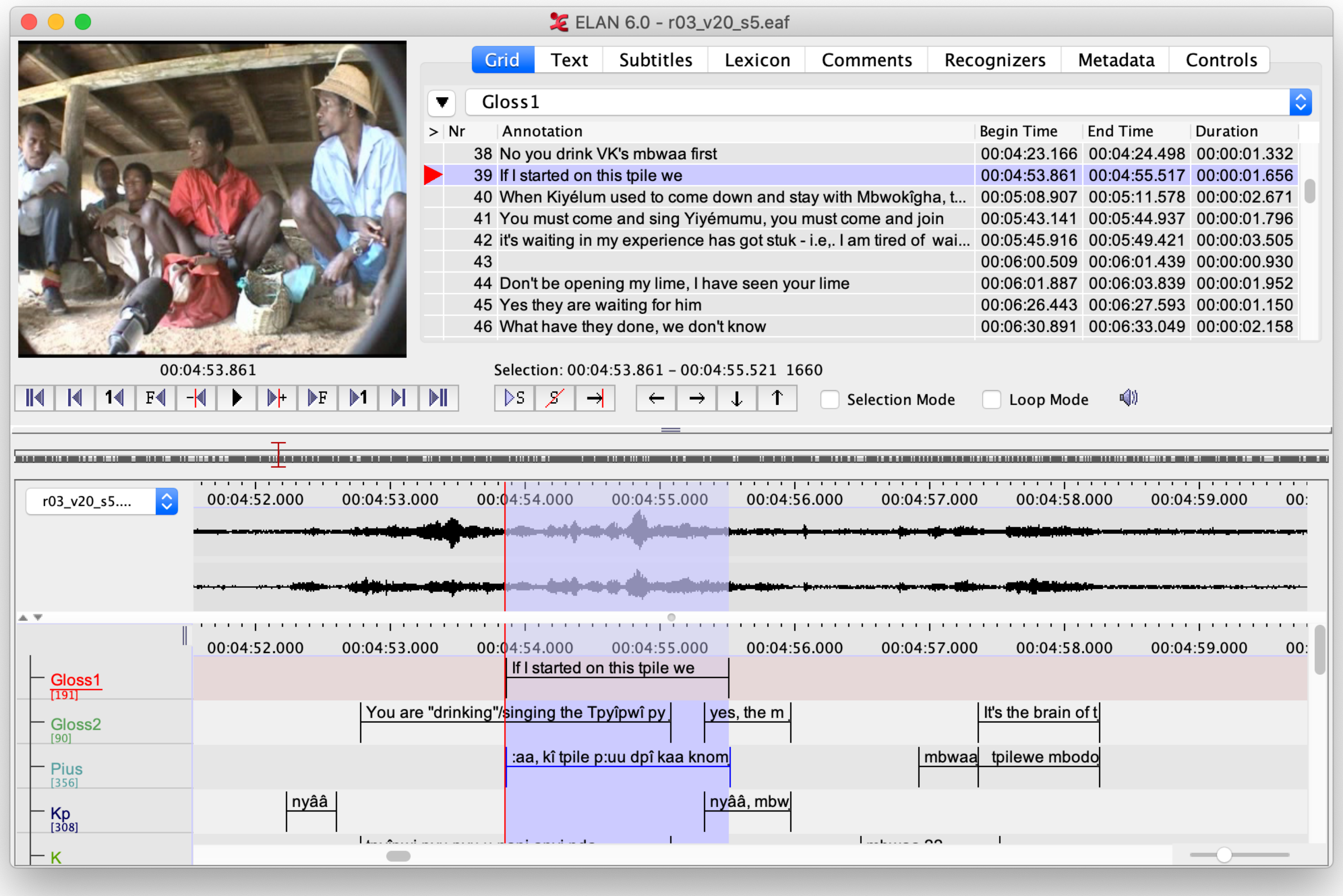
Han Slöetjes

han.sloetjes@mpi.nl

https://archive.mpi.nl/tla/elan

Introduction

- **ELAN** is a manual annotation tool developed by **The Language Archive/MPI for Psycholinguistics**. It supports multi-tier, multi-speaker, time-linked annotation of audio and video and is applied in many fields of research, language documentation being one of them.



ELAN 6.0 - r03_v20_s5.eaf

Nr	Annotation	Begin Time	End Time	Duration
38	No you drink VK's mbwaa first	00:04:23.166	00:04:24.498	00:00:01.332
39	If I started on this tpile we	00:04:53.861	00:04:55.517	00:00:01.656
40	When Kiyélum used to come down and stay with Mbwokîgha, t...	00:05:08.907	00:05:11.578	00:00:02.671
41	You must come and sing Yiyémumu, you must come and join	00:05:43.141	00:05:44.937	00:00:01.796
42	it's waiting in my experience has got stuk - i.e,. I am tired of wai...	00:05:45.916	00:05:49.421	00:00:03.505
43		00:06:00.509	00:06:01.439	00:00:00.930
44	Don't be opening my lime, I have seen your lime	00:06:01.887	00:06:03.839	00:00:01.952
45	Yes they are waiting for him	00:06:26.443	00:06:27.593	00:00:01.150
46	What have they done, we don't know	00:06:30.891	00:06:33.049	00:00:02.158

Selection: 00:04:53.861 – 00:04:55.521 1660

Background and history

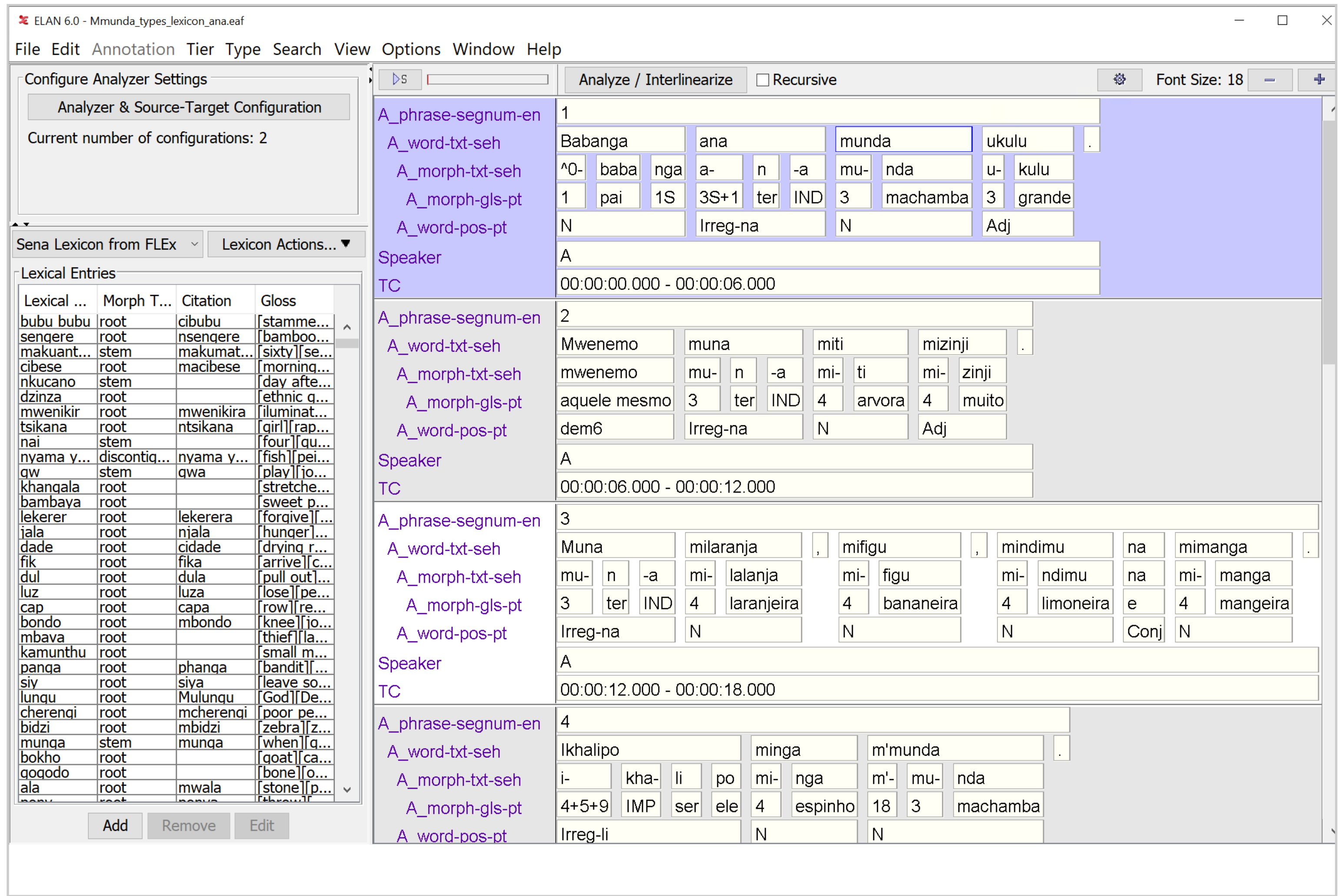
- in language documentation the transcription and translation steps are often followed by morphological parsing and glossing

- it is possible to perform these steps manually in **ELAN**, but the preferred approach is to apply computer-aided methods as offered by **Toolbox**[3] and **FLEx**[4]

- **ELAN** provides import and export functions for the file formats of these tools, so that users can create a toolchain and move their data from one tool to another for the task at hand

- **TLA** once developed the lexicon tool **LEXUS**[1] and prepared to combine it with **ELAN** by way of a new Natural Language Processing module, **LEXAN**[2]

Interlinearization mode



ELAN 6.0 - Mmunda_types_lexicon_ana.eaf

File Edit Annotation Tier Type Search View Options Window Help
Configure Analyzer Settings
Analyzer & Source-Target Configuration
Current number of configurations: 2
Sena Lexicon from FLEx
Lexicon Actions...
Lexical Entries
Add
Remove
Edit
Analyze / Interlinearize
Recursive
Font Size: 18

A_phrase-segnum-en	1									
A_word-txt-seh	Babanga			ana			munda		ukulu	.
A_morph-txt-seh	^0-	baba	nga	a-	n	-a	mu-	nda	u-	kulu
A_morph-gls-pt	1	pai	1S	3S+1	ter	IND	3	machamba	3	grande
A_word-pos-pt	N			Irreg-na			N		Adj	
Speaker	A									
TC	00:00:00.000 - 00:00:06.000									

A_phrase-segnum-en	2								
A_word-txt-seh	Mwenemo	muna			miti		mizinji		.
A_morph-txt-seh	mwenemo	mu-	n	-a	mi-	ti	mi-	zinji	
A_morph-gls-pt	aquele mesmo	3	ter	IND	4	arvora	4	muito	
A_word-pos-pt	dem6	Irreg-na			N		Adj		
Speaker	A								
TC	00:00:06.000 - 00:00:12.000								

A_phrase-segnum-en	3														
A_word-txt-seh	Muna			milaranja		,	mifigu		,	mindimu		na	mimanga		.
A_morph-txt-seh	mu-	n	-a	mi-	lalanja		mi-	figu		mi-	ndimu	na	mi-	manga	
A_morph-gls-pt	3	ter	IND	4	laranjeira		4	bananeira		4	limoneira	e	4	mangeira	
A_word-pos-pt	Irreg-na			N			N			N		Conj	N		
Speaker	A														
TC	00:00:12.000 - 00:00:18.000														

A_phrase-segnum-en	4									
A_word-txt-seh	Ikhalipo				minga		m'munda			.
A_morph-txt-seh	i-	kha-	li	po	mi-	nga	m'-	mu-	nda	
A_morph-gls-pt	4+5+9	IMP	ser	ele	4	espinho	18	3	machamba	
A_word-pos-pt	Irreg-li				N		N			

- new text oriented mode with an *Interlinear Glossed Text (IGT)* style of user interface
- the **LEXAN** modules became extensions of **ELAN** (after end of support for **LEXUS**)
- combined with a new lexicon component
- supports the process of (machine-assisted) parsing and glossing

40th ANNIVERSARY FOUNDED IN 1980

MAX PLANCK

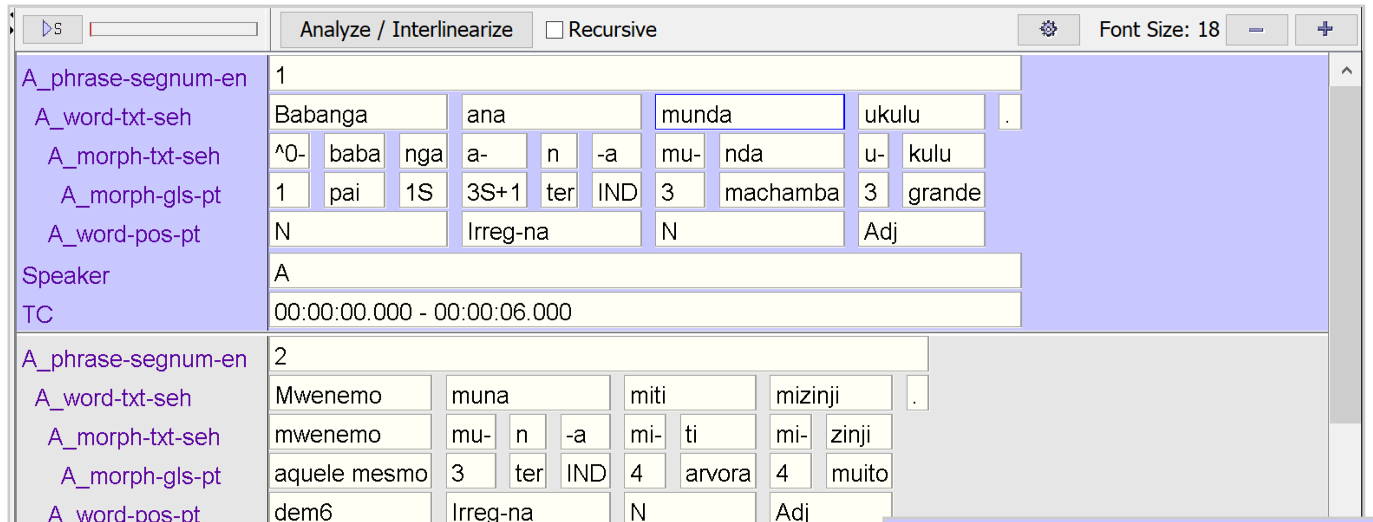
MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

WWW.MPI.NL

The Language Archive

Main characteristics

- text oriented (but still linked to the audio!)
- keyboard driven
- optimized for computer assisted morphological parsing and glossing
- text analyzer modules provide ‘*suggestions*’ to be disambiguated by the user
- contains a lexicon editor and viewer
- a flexible system, therefore some configuration is required

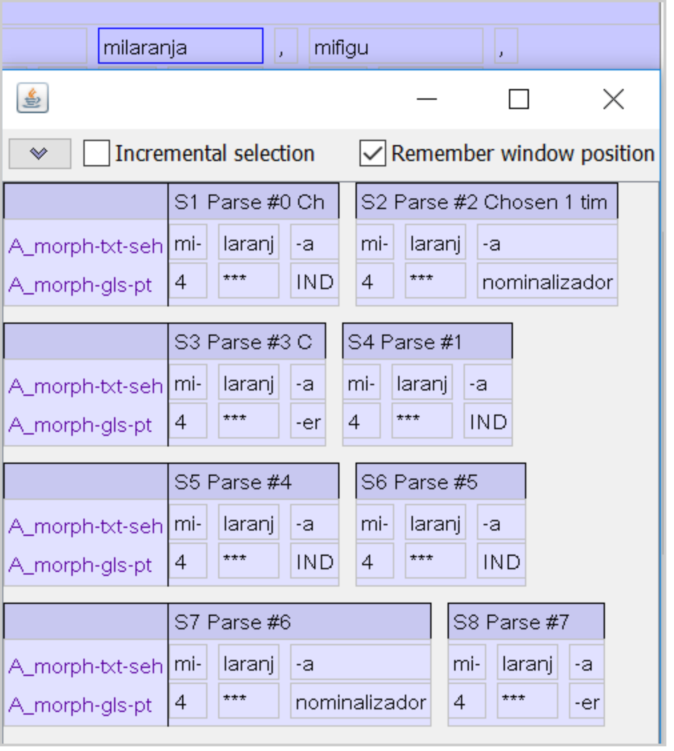


▷S Analyze / Interlinearize □ Recursive ⚙ Font Size: 18 − +

A_phrase-segnum-en	1								
A_word-txt-seh	Babanga			ana			munda		ukulu	.
A_morph-txt-seh	^0-	baba	nga	a-	n	-a	mu-	nda	u-	kulu
A_morph-gls-pt	1	pai	1S	3S+1	ter	IND	3	machamba	3	grande
A_word-pos-pt	N			Irreg-na			N		Adj	
Speaker	A									
TC	00:00:00.000 - 00:00:06.000									

A_phrase-segnum-en	2								
A_word-txt-seh	Mwenemo	muna			miti		mizinji		.
A_morph-txt-seh	mwenemo	mu-	n	-a	mi-	ti	mi-	zinji	
A_morph-gls-pt	aquele mesmo	3	ter	IND	4	arvora	4	muito	
A_word-pos-pt	dem6	Irreg-na			N		Adj		
Speaker	A								
TC	00:00:06.000 - 00:00:12.000								

Interlinear editor with a play selection button to play the current active phrase.



milaranja , mifigu ,

□ Incremental selection ☑ Remember window position

	S1 Parse #0 Ch			S2 Parse #2 Chosen 1 tim		
A_morph-txt-seh	mi-	laranj	-a	mi-	laranj	-a
A_morph-gls-pt	4	***	IND	4	***	nominalizador

	S3 Parse #3 C			S4 Parse #1		
A_morph-txt-seh	mi-	laranj	-a	mi-	laranj	-a
A_morph-gls-pt	4	***	-er	4	***	IND

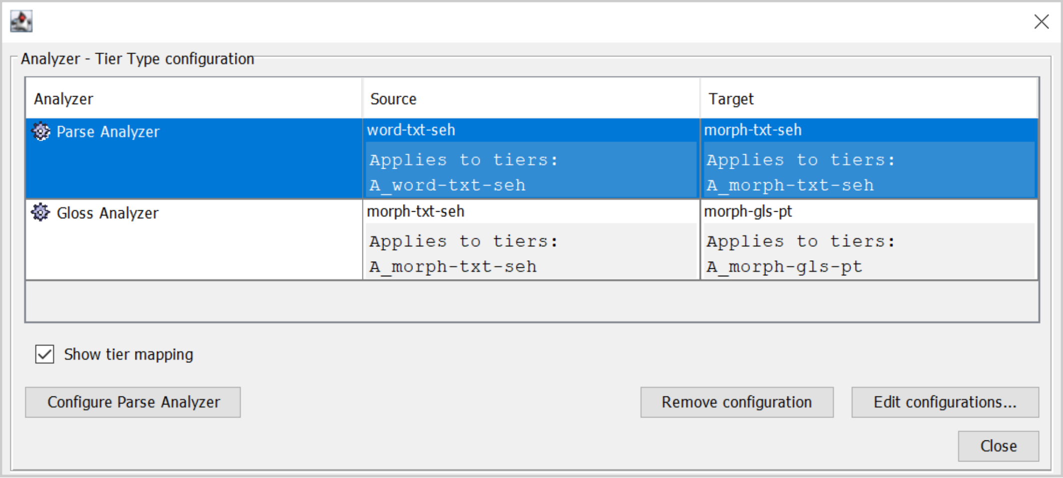
	S5 Parse #4			S6 Parse #5		
A_morph-txt-seh	mi-	laranj	-a	mi-	laranj	-a
A_morph-gls-pt	4	***	IND	4	***	IND

	S7 Parse #6			S8 Parse #7		
A_morph-txt-seh	mi-	laranj	-a	mi-	laranj	-a
A_morph-gls-pt	4	***	nominalizador	4	***	-er

The suggestions window displays the suggestions in the same layout as the editor. The most frequently selected parses are displayed at the top.

Text analyzers

- are modules that receive text as input (an annotation) and produce output for a single or for multiple annotations
- can be added as extensions through an API
- can be connected to a lexicon

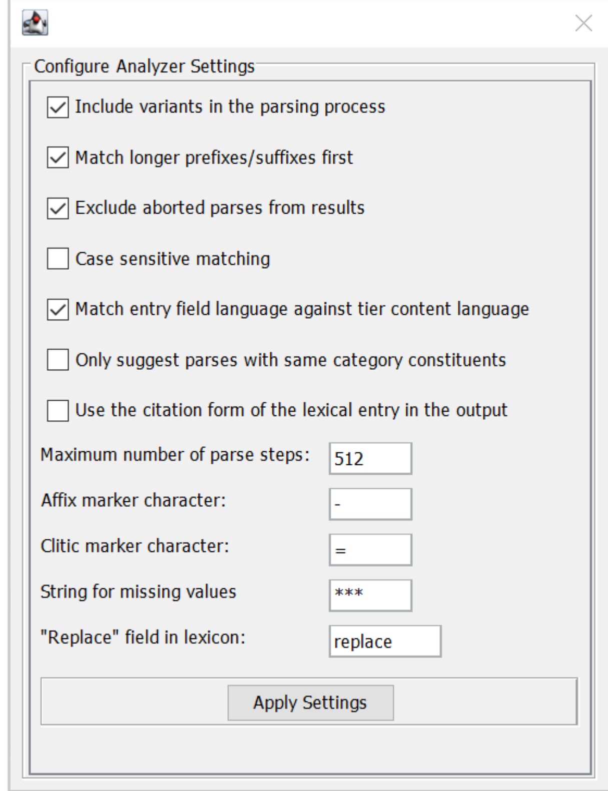


Analyzer - Tier Type configuration

Analyzer	Source	Target
Parse Analyzer	word-txt-seh Applies to tiers: A_word-txt-seh	morph-txt-seh Applies to tiers: A_morph-txt-seh
Gloss Analyzer	morph-txt-seh Applies to tiers: A_morph-txt-seh	morph-gls-pt Applies to tiers: A_morph-gls-pt

☑ Show tier mapping

Linking analyzers to tier types and configuration of an analyzer.



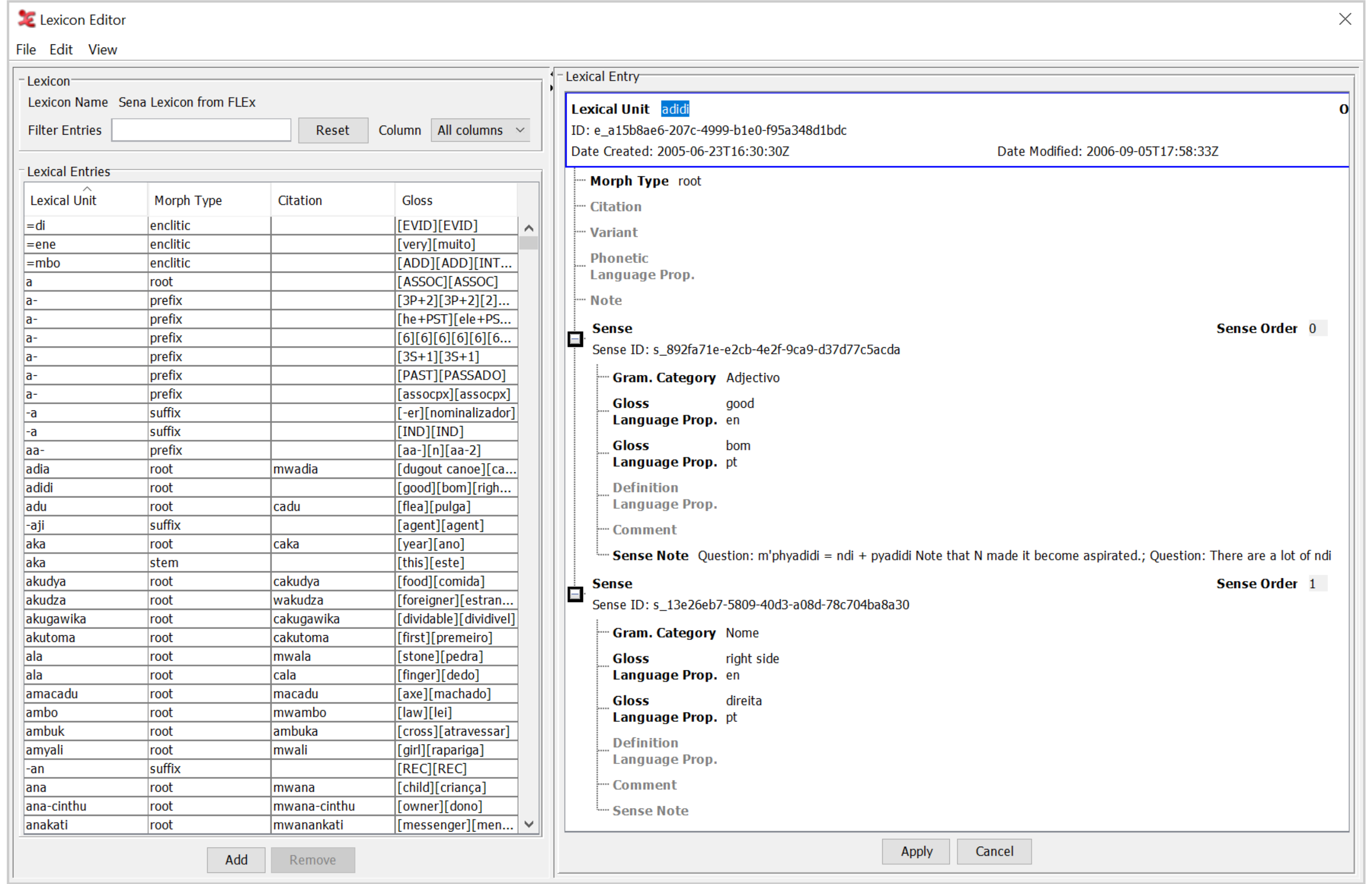
Configure Analyzer Settings

☑ Include variants in the parsing process
☑ Match longer prefixes/suffixes first
☑ Exclude aborted parses from results
□ Case sensitive matching
☑ Match entry field language against tier content language
□ Only suggest parses with same category constituents
□ Use the citation form of the lexical entry in the output

Maximum number of parse steps:	512
Affix marker character:	-
Clitic marker character:	=
String for missing values	***
"Replace" field in lexicon:	replace

- built-in analyzers:
  - whitespace splitter
  - morphological parser
  - gloss analyzer
- the parser and glosser:
  - require access to a lexicon
  - keep record of choices made to improve suggestions
  - are language independent
  - the parser is implemented as a finite state machine

Lexicon editor



Lexicon

Lexicon Name Sena Lexicon from FLEx

Filter Entries Reset Column All columns

Lexical Entries

Lexical Unit	Morph Type	Citation	Gloss
=di	enclitic		[EVID][EVID]
=ene	enclitic		[very][muito]
=mbo	enclitic		[ADD][ADD][INT...
a	root		[ASSOC][ASSOC]
a-	prefix		[3P+2][3P+2][2]...
a-	prefix		[he+PST][ele+PS...
a-	prefix		[6][6][6][6][6][6...
a-	prefix		[3S+1][3S+1]
a-	prefix		[PAST][PASSADO]
a-	prefix		[assocpx][assocpx]
-a	suffix		[-er][nominalizador]
-a	suffix		[IND][IND]
aa-	prefix		[aa-][n][aa-2]
adia	root	mwadia	[dugout canoe][ca...
adidi	root		[good][bom][righ...
adu	root	cadu	[flea][pulga]
-aji	suffix		[agent][agent]
aka	root	caka	[year][ano]
aka	stem		[this][este]
akudya	root	cakudya	[food][comida]
akudza	root	wakudza	[foreigner][estran...
akugawika	root	cakugawika	[dividable][dividivel]
akutoma	root	cakutoma	[first][premeiro]
ala	root	mwala	[stone][pedra]
ala	root	cala	[finger][dedo]
amacadu	root	macadu	[axe][machado]
ambo	root	mwambo	[law][lei]
ambuk	root	ambuka	[cross][atravessar]
amyali	root	mwali	[girl][rapariga]
-an	suffix		[REC][REC]
ana	root	mwana	[child][criança]
ana-cinthu	root	mwana-cinthu	[owner][dono]
anakati	root	mwanankati	[messenger][men...

Lexical Entry

Lexical Unit adidi 0
ID: e_a15b8ae6-207c-4999-b1e0-f95a348d1bdc
Date Created: 2005-06-23T16:30:30Z Date Modified: 2006-09-05T17:58:33Z

Morph Type root
Citation
Variant
Phonetic
Language Prop.
Note

Sense **Sense Order** 0
Sense ID: s_892fa71e-e2cb-4e2f-9ca9-d37d77c5acda
Gram. Category Adjectivo
Gloss good
Language Prop. en
Gloss bom
Language Prop. pt
Definition
Language Prop.
Comment
Sense Note Question: m'phyadidi = ndi + pyadidi Note that N made it become aspirated.; Question: There are a lot of ndi

Sense **Sense Order** 1
Sense ID: s_13e26eb7-5809-40d3-a08d-78c704ba8a30
Gram. Category Nome
Gloss right side
Language Prop. en
Gloss direita
Language Prop. pt
Definition
Language Prop.
Comment
Sense Note

- multiple lexicons can be created and linked
- lexicon entries have a few predefined fields
- custom fields can be added
- sorting of entries according to a custom sort order
- the structure of a lexical entry is similar to the **LIFT**[5] format
- import lexicons from **Toolbox**, **FLEx** (**LIFT**) and **CorpAfroAs**[6] format
- export a lexicon to **LIFT** format

Acknowledgements

The design and implementation of the Interlinearization mode have for a large part been the work of Herman Stehouwer (LEXAN) and Olaf Seibert.

1. Ringersma, J. and Kemps-Snijders, M. (2007). Creating multimedia dictionaries of endangered languages using LEXUS. Interdpeech 2007
2. Stehouwer, H. and Drude, S. (2012). Lexan: A lexical annotation framework for ELAN. Talk presented at LREC 2012
3. Field Linguist’s Toolbox, https://software.sil.org/toolbox/
4. FieldWorks Language Explorer, https://software.sil.org/fieldworks/
5. Lexicon Interchange Format, https://github.com/sillsdev/lift-standard
6. http://corpafroas.tge-adonis.fr