

GlossKit: An Integrated Protocol for Documentation, Data Processing, Archival, and Publication

Shuan Osman Karim

GlossKit LLC
Julius-Maximilians-Universität Würzburg
Endangered languages Documentation Programme
Berlin-Brandenburgische Akademie der Wissenschaft



GLOSS-KIT



UNTERSTÜTZT VON / SUPPORTED BY

Alexander von
HUMBOLDT
STIFTUNG

Outline

- 1 Introduction
- 2 My Background
- 3 ELAN-FLE_x-ELAN
- 4 GlossKit

Outline

1 Introduction

2 My Background

3 ELAN-FLE_x-ELAN

4 GlossKit

Introduction to GlossKit

- GlossKit is an integrated platform for Language Documentation.
- Modules include:
 - Segmentation/Transcription
 - Morphological Parsing
 - Translation
 - Archival
 - Publication
- GlossKit either replaces or complements standard documentary tools.
- GlossKit employs Machine Learning tools to aid in the process.

Outline

1 Introduction

2 My Background

3 ELAN-FLE_x-ELAN

4 GlossKit

My Background

- Historical Linguist: Traditional “Neo-Grammarian”-inspired Historical Linguistic
- Almost Documentarian: ELDP 2019 Training
- Machine Learning and AI Specialist?
- Full-stack Developer and GlossKit Founder (Agentic Engineering)

Outline

- 1 Introduction
- 2 My Background
- 3 ELAN-FLE_x-ELAN
- 4 GlossKit

ELAN-FLEx-ELAN I

ELAN (Wittenburg et al., 2006):

- Segmentation
- Annotation

The screenshot displays the ELAN-FLEx-ELAN software interface. At the top, a timeline bar shows time intervals from 00:00 to 00:09.000. Below this, a list of annotations is visible, including:

- Metadata-title-e** (1): Şa İsmâ'il ü Cemberta 5 | Recorded date: 2025-02-21
- M019_Transcrip** (8): lem guftûgoyea bûn taq yekêk le dergay ya
- M019_word-txt** (145): lem guftûgoyea bûn taq yekêk le dergay ya
- M019_morp** (75): l= em guftûgoyea bûn taq yekêk le dergay ya
- M019_mor** (75): in dem.pr conv dem in be.pst 3pl knock one indf from door 3sg glve.pst
- M019_mor** (75): procliti stem ste encl encl stem suffix stem stem suffix stem stem enclitic stem
- M019_word-txt** (145): PRON NOUN VERB NOUN NUM ADV NOUN VERB
- M019_Transla** (8): They were in this conversation and a knock came from the door.
- M019_Transla** (8): Estaban conversando y llamaron a la puerta.
- M019_Transla** (8): Ew di vê axafinê de bûn û lêdanek ji derî hat.
- M019_Particip** (8): M019

The interface also includes a search bar at the top left and a list of annotations on the left side.

ELAN-FLE_x-ELAN II

FLE_x (SIL International, 2022):

- As Skilton et al. (2025) show, FLE_x's design structures are deeply entwined with its origins in Bible translation and the Boasian “grammar-dictionary-text” model of documentation.
- Assumptions of FLE_x following Skilton et al. (2025):
 - Technical barriers:
 - FLE_x is resource-intensive
 - limited to Windows (Linux?)
 - performs poorly on low-RAM systems
 - Accessibility:
 - FLE_x lacks support for screen readers
 - customizable text sizes, or color adjustments
 - Mono-content bias:
 - FLE_x struggles with multilingual or multimodal data, often forcing users to “clean” code-switching or ignore gesture and prosody.
 - Collaboration constraints:
 - The program is built for single-user workflows.

ELAN-FLE_x-ELAN III

- Multi-user collaboration requires opaque sync protocols (via LanguageDepot or LanguageForge) that compromise Indigenous data sovereignty by granting SIL rights over uploaded content.
- Export fragility:
 - Dictionaries export only to rigid XHTML or bespoke XML formats.
 - Interlinear text export often destroys spacing or metadata, making corpora hard to repurpose.
 - The built-in Interlinear Export tool only handles examples (snippets) nicely. Whole narratives or texts export as fragmented or poorly formatted outputs.
 - Archiving only full backup files preserving databases, but they are opaque to future users.

ELAN-FLE_x-ELAN IV

5. The imported text should now appear in the list of texts, with the metadata as the title.

The screenshot shows the ELAN-FLE_x-ELAN interface. On the left, the 'Texts' panel lists the imported text with the title 'spa0001; Viaje a Dresden; Katherine Bolaños y Kelsey Neely'. On the right, the 'Text' editor shows the same text with the 'Info' tab selected, displaying the title and the text content: 'ah ya entonces mañana te vas a Dresden'.

6. You can now click on the *Analyze* tab to interlinearize the text.

The screenshot shows the ELAN-FLE_x-ELAN interface with the 'Text' editor. The 'Analyze' tab is selected, displaying the text 'ah ya entonces mañana'. The 'Word' column is highlighted in red, and the 'Morphemes' column shows the word 'ah'. The 'Lex. Entries' column shows the word 'ah'.

ELAN-FLE_x-ELAN V

ELAN-FLE_x-ELAN: If your tier structures are properly formatted and your FLE_x project is properly set up, you should be able to transfer back and forth between both tools.

Outline

- 1 Introduction
- 2 My Background
- 3 ELAN-FLE_x-ELAN
- 4 GlossKit

GlossKit

`https://app.glosskit.com`

Thank You

Zor spastan ekem!

References I

- SIL International (2022). *FieldWorks Language Explorer (FLEx)*.
<https://software.sil.org/fieldworks/>.
- Skilton, Amalia H. et al. (2025). "Digital infrastructure and its impacts on language work: A case study of FieldWorks Language Explorer". In: *Language Documentation & Conservation* 19, pp. 1–36.
- Wittenburg, Peter et al. (2006). "ELAN: A professional framework for multimodality research". In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*.