Towards the Integration of Cuneiform in the OntoLex-Lemon Framework

Timo Homburg[1], Thierry Declerck[2]
Mainz University Of Applied Sciences[1] and German Research Centre for AI (DFKI)[2]
timo.homburg@hs-mainz.de, declerck@dfki.de

Problem

Cuneiform artifacts have a rich paleography which changes over space and time. For Assyriologists, the conclusions they draw from a given text's paleography may determine a classification of the text in different time periods, may attest the scribe of the given text or give hints about the location where a text has been created. Data scientists and the NLP community may use paleographic descriptions for OCR and machine learning tasks. Currently, we lack a data model which could represent graphemes and grapheme variants and relates these grapheme variants to actual glyph representation on given writing media.

Foundations and Idea

Our approach adds an additional ontology in relation to the Ontolex-Lemon model [3] and the lemonETY model for etymology representation [2], which describes Graphemes, GraphemeVariants and can model the structure of single graphemes using their AtomicParts. We exemplify this ontology using the cuneiform script. Essential elements include:

- Paleography description module
- Etymology module
- Relation of Graphemes to Glyphs
- Connection of OntoLex-Lemon LexicalForms to Grapheme compositions
- Similarity and decomposition statements of graphemes

References

[1] Timo Homburg. Paleocodage - enhancing machine-readable cuneiform descriptions using a machine-readable paleographic encoding. *Digital Scholarship in the Humanities*, 36(Supplement$_2$), 112021.1

[2] Anas Fahad Khan. Towards the representation of etymological data on the semantic web. *Information*, 9(12), 2018. (document)

[3] John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. The ontolex-lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21, 2017. (document)

Acknowledgements

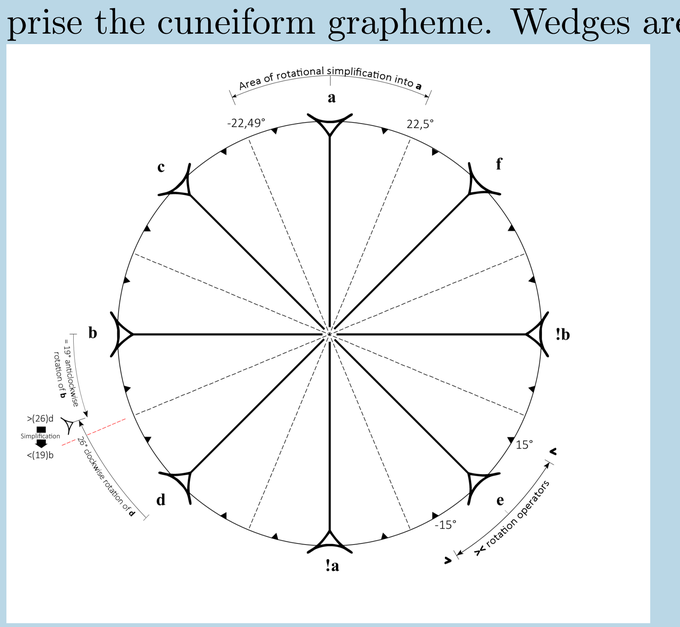
This paper is based upon work from the COST Action NexusLinguarum – European network for Web-centered linguistic data science (CA18209), supported by COST (European Cooperation in Science and Technology). It is also supported by the Horizon 2020 research and innovation program with the project Prêt-à-LLOD (grant agreement no. 825182).

More information

Github link: https://github.com/situx/graphemon

An ontology model to represent graphemes

Graphemes in cuneiform languages are comprised out of a set of interconnected wedges, which comprise the cuneiform grapheme. Wedges are defined by a size, length and angle on the unit circle and

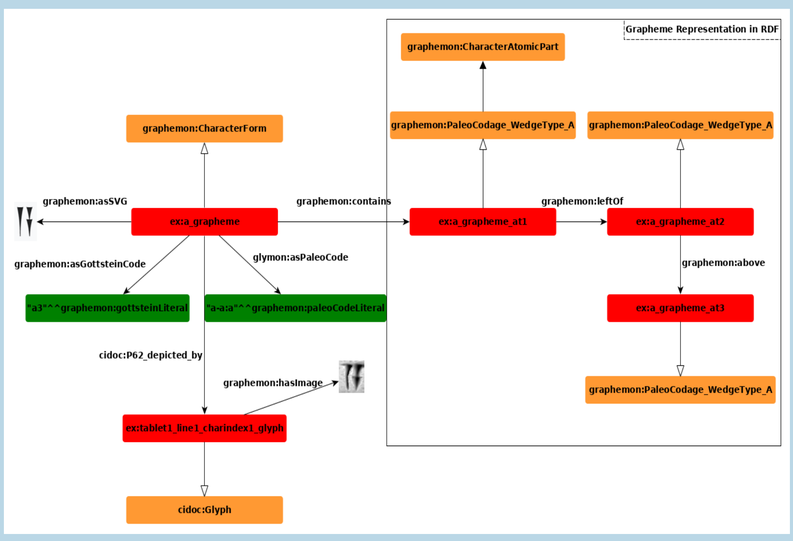


U+1208A E_v1	498	HT4	b::b:b:b:b_:a:a	
U+1208A E_v2	498	HT40,585,97	b:b:b_a-a:a	
U+1208A E_v3	498	HT88,291,37,298	b:b:b_a:a	
U+1208A E_v4	498	HT188	b::b_a-a:a	

Figure 1: Left: Cuneiform sign description mode using PaleoCodage [1]
Right: Sign variants in cuneiform as exemplified on the same cuneiform sign E in a corpus originating from the same location and time period

Grapheme and Glyph Description

- **Grapheme:** Link to Unicode, Sense, IDs
- Many **Grapheme variant** instances: Grapheme description as image, Character Description languages
- Each **Grapheme variant** instance connected to at least one **Glyph instance**
- **Grapheme variant** may be described by RDF subgraph of **AtomicParts**
- **Glyph**: Physical representation described by images and attributes (e.g. clay)

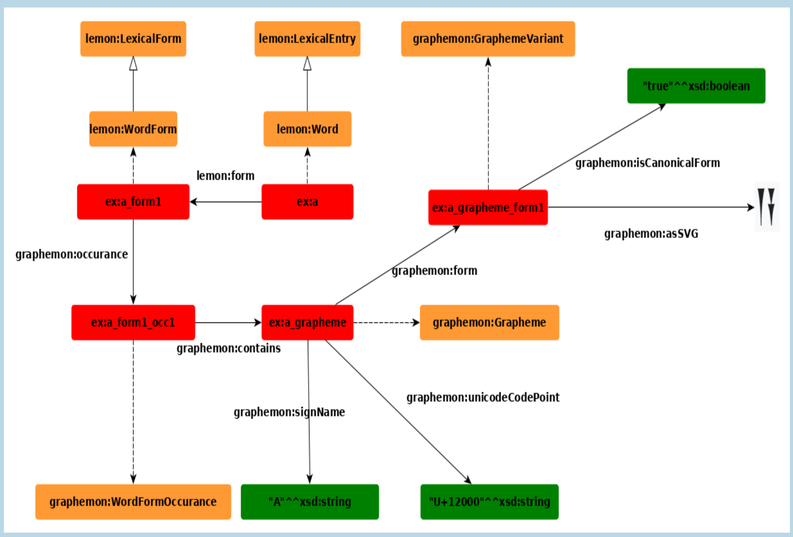


Relation to Ontolex-Lemon and Applications

- Word Form Occurrences connect to sets of GraphemeVariants

Applications

- Cuneiform Sign Variant Registry
- Preparation of Machine Learning datasets
- Solving research questions concerning
  - Linguistic features
  - Paleographic features



Sign Etymology and Similarity

- Grapheme and glyph similarity relations by
  - Shape similarity (e.g. comparison of grapheme image representations (SVG)
  - Encoding similarity (e.g. Levenshtein Distance between grapheme descriptions)
  - Semantic similarity (Etymology)
- Extension of the lemonETY vocabulary [2]
  - Include etymological relations between graphemes
  - Relate etymology of words to graphemes and possibly glyphs