New DH Tools and Methods for Ottoman Turkish

Rumi 1.1 ● DH-LAB ● TEI XML Viewer ● TEI XML Search Engine

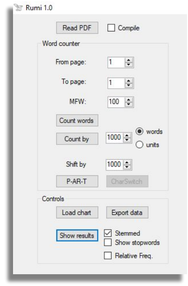
Tamás Kiss, PhD (Eötvös Loránd University of Sciences / Qulto)

Rumi 1.1: Word counter and lemmatizer for Ottoman Turkish

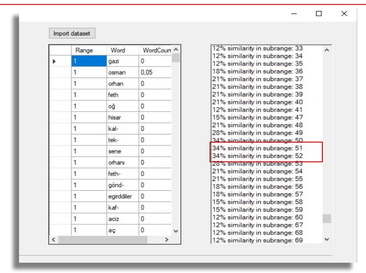
Software Architect and Developer: Tamás Kiss

Rumi 1.1 is a text analysis software optimized for the Ottoman language, which I have been developing since 2019. Essentially, Rumi is a lemmatizer and parameterizable word counter applicable to texts transliterated to the Latin script. Currently, *Rumi 1.1* enables users to perform the following processes:

- **Tokenization** (i.e. splitting the text into words)
- **Lemmatization of words** (i.e. removing suffices) **featuring Turkic morphology** (If a lemma does not already exist in the dictionary, the software asks the user to decide what the unrecognized word's lemma is. In subsequent analyses the software will recognize that word.)
- **Character normalization** (i.e. eliminating differences between transcription standards)
- **Word counting** with the following parameters
  - Individual document / various documents compiled
  - Number of most frequent words
  - Lemmatized / not lemmatized
  - Including / excluding stopwords
  - Absolute frequency / relative frequency
- **"Supervised" word counting** to count the elements of a predefined set of words within a document or documents with the following parameters
  - In subunits (i.e. substrings) defined by word count
  - In subunits defined by their number within the document
  - With subunits shifted by an adjustable number of words
- **P-AR-T analysis** to count the number of words of Persian, Arabic and Turkic origin within a document
- **Simple visualization** to display word counter results
- **Exporting** data in CSV

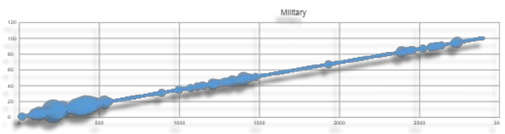


Results with Rumi 1.1



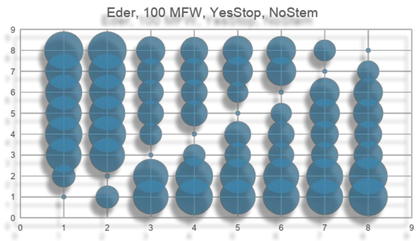
Theme identification

A substring's relative frequency of lemmatized types (i.e. elements of the lexicon) compared with every substring's relative frequency of lemmatized types in other documents



Narrative time / authorial intent analysis

The frequency of the elements of a set of tokens associable with a social group (e.g. the *askeri* class) counted in fixed-sized substrings throughout a document



Distance measurement / authorship attribution

Applying delta analysis (i.e. Eder's Delta, Canberra, Manhattan Burrows' Delta) to a set of texts to investigate similarities and identify possible authors of documents of anonymous authorship

DH-LAB Proofreading / Transcription module

Product Owner and Product Manager: Tamás Kiss
Software development: Zoltán Kanász-Nagy(DevLead), Dóra Horváth-Gyurcsik, Anett Mári

Digital Heritage National Laboratory (DH-LAB) is a research infrastructure with various functions and services for data replication, enrichment, processing and visualization, with a data repository at its core. DH-LAB is developed by **Qulto** in cooperation with **Eötvös Loránd University's Department of Digital Humanities**.

Functionalities and services of the DH-LAB infrastructure:

- InvenioRDM Data Repository
- Secondary GUI
- Batch data enrichment (i.e. external input of large amounts of data/metadata from file dialog or command-line)
- OCR service
- Proofreading and transcription
- Semantic indexing and grid building from metadata and plaintext
- NLP (tokenization, lemmatization, NER, etc.) optimized to English and Hungarian
- TEI XML search
- TEI XML visualization

DH-LAB's Proofreading and Transcription module enables users to

- process files existent in a data repository
- correct faulty OCR results
- transcribe documents from scratch (left-to-right and right-to-left text orientation is defined at line level, and conclusively transcribing "mixed" documents is supported
- save training data into separate models
- *HTR (Ottoman or other) documents with TrOCR (*under development*)

DH-LAB Transcription workflow

1. Upload image(s) of a scanned document into repository either directly or through DH-LAB's user interface (the latter supports batch imports)
2. OCR / *HTR image to either receive a character-recognized result or an empty XML (HTR under development)
3. Start Proofreading module for selected record
4. Set language and select a language model
5. Define blocks and lines and set line orientation (left-to-right or right-to-left) – adjusting line dimensions and 360° rotation are fully supported
6. Enter string into Page Editor window
7. Open Line Editor
8. Split line into words
9. Set word boundaries in the image layer
10. Save line and training data
11. Save page
12. Generate TEI XML
13. Enrich text with semantic, textological, linguistic and critical data in TEI XML

DH-LAB TEI XML Viewer and Search Engine

Product Owner and Product Manager: Tamás Kiss
Software development: Zoltán Kanász-Nagy(DevLead), Dóra Horváth-Gyurcsik, Anett Mári

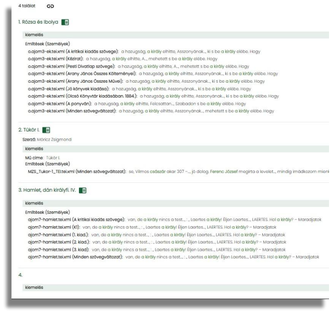
TEI XML Viewer is used to display the textual content and all added semantic, linguistic, textological, and critical information encoded into TEI XML files to make them easily explorable for human readers. It enables users to explore all versions of the same text at once or in parallel views, navigate OCR-ed / HTR-ed or transcribed documents in parallel views of the image layer and the text layer, and much more. The *TEI XML Search Engine* makes all unique features of TEI XMLs searchable at metadata, fulltext, and entity level. Searching and filtering for persons (such as editors, addressees of letters, people mentioned in the text, etc.) or place names (such as place of publication, geographical locations featured in the text, etc.) make investigating trends and features easier across entire corpora. By Named Entity Recognition the engine also enables users to search for people and place names that are not mentioned in the fulltext by name. (E.g. searching for Budapest will find "I went to the capital" if in the text "capital" refers to Budapest).

Key features of TEI XML Viewer:

- Basic search for records and files in linked GitHub repositories and data repositories
- Whole screen and split screen views
- One text version, all text versions, and simplified all text versions view
- Paginated and scrollable layout
- Annotations view
- Print view (i.e. printable critical edition with all textual and added information featured in footnotes)

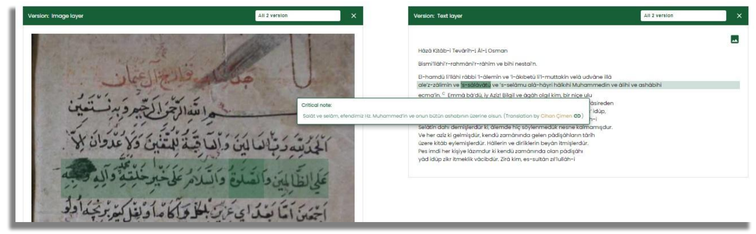
Search and filter categories in the TEI XML Search Engine:

- Metadata
  - Titles
  - People
  - Dates
  - Place names
  - Institutions
  - categories
  - People
  - Dates
  - Place names
  - Institutions
  - Manuscript features
  - Language
- Fulltext
  - in all text versions
  - in specific text versions
- Named entities
  - Persons
  - Places



DH-LAB TEI XML Viewer Examples

All versions view with a critical note and the occurrence of a word in specific versions displayed



Split screen view of image and text layer with a critical note displayed

tamas.mail@gmail.com

qulto for culture & knowledge

DH National Laboratory for Digital Heritage