New DH Tools and Methods for Ottoman Turkish

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 InvenioRDM Data Repository lemma is. In subsequent analyses the software will recognize that word.) Secondary GUI **Character normalization** (i.e. eliminating differences between transcription standards) • Word counting with the following parameters Rumi 1.0 command-line) Individual document / various documents compiled 0 Read PDF Compile OCR service Number of most frequent words Word counter Proofreading and transcription Lemmatized / not lemmatized 0 From page: 1 📫 Including / excluding stopwords To page: 1 🚔 Absolute frequency / relative frequency • TELXML search MFW: 100 🜲 • "Supervised" word counting to count the elements of a predefined • TEI XML visualization Count words set of words within a document or documents with the following words Count by 1000 parameters In subunits (i.e. substrings) defined by word count Shift by 1000 🜲 • process files existent in a data repository In subunits defined by their number within the document P-AR-T • correct faulty OCR results • With subunits shifted by an adjustable number of words Controls • P-AR-T analysis to count the number of words of Persian, Arabic and Load chart Export data Turkic origin within a document Show results Stemmed Show stopword Simple visualization to display word counter results Relative Freq. **Exporting** data in CSV Results with Rumi 1.1 - 🗆 Import dataset 12% similarity in subrange: 33 12% similarity in subrange: 34 12% similarity in subrange: 35 18% similarity in subrange: 36 21% similarity in subrange: 37 development) 21% similarity in subrange: 3 3. 21% similarity in subrange: 3 21% similarity in subrange: 40 12% similarity in subrange: 41 15% similarity in subrange: 47 21% similarity in subrange: 48 28% similarity in subrange: 49 kal 34% similarity in subrange: 5 sene 34% similarity in subrange: 52 6. Enter string into Page Editor window orhani 28% similarity in subrange: 21% similarity in subrange: 5 21% similarity in subrange: 5 Open Line Editor 8% similarity in subrange: Split line into words 15% similarity in subrange: 5 kaf-15% similarity in subrange: 5 12% similarity in subrange: 50 12% similarity in subrange: 60 12% similarity in subrange: 68 12% similarity in subrange: 68 aciz aç Save line and training data 10. % similarity in subrange: 6 11. Save page 12. Generate TEI XML Theme identification 13. 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 جملدر ودين يولنه غبر لولدر ودنيا بد مغر دُرٌ غُرُورًا هُلِنْدُنْ إِنْتِعَامَ آلِجُلَةُ وُوَغَرِيْبِلَوَى وَ National Laboratory for Digital Heritage 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 در وحقطو في in fixed-sized substrings throughout a document Eder, 100 MFW, YesStop, NoStem 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

tamas.mail@gmail.com

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

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

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:

Batch data enrichment (i.e. external input of large amounts of data/metadata from file dialog or

• Semantic indexing and grid building from metadata and plaintext

• NLP (tokenization, lemmatization, NER, etc.) optimized to English and Hungarian

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

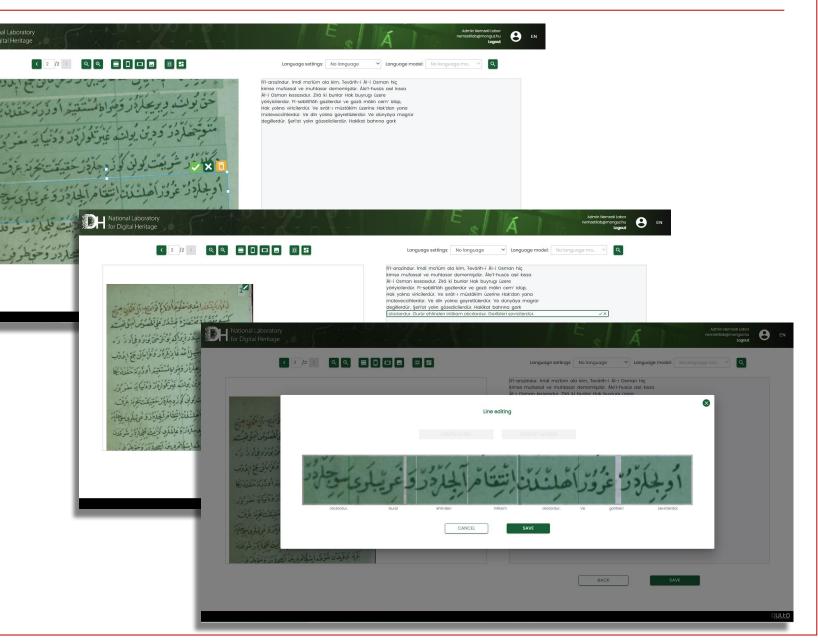
- transcribe documents from scratch (left-to-right and right-to-left text orientation is defined at line level, and conclusively transcribing "mixed" documents is supported
- save training data into separate models
- *HTR (Ottoman or other) documents with TrOCR (under development)

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
 - Start Proofreading module for selected record
- 4. Set language and select a language model
 - Define blocks and lines and set line orientation (left-to-right or right-to-left) adjusting line dimensions and 360° rotation are fully supported
 - Set word boundaries in the image layer



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

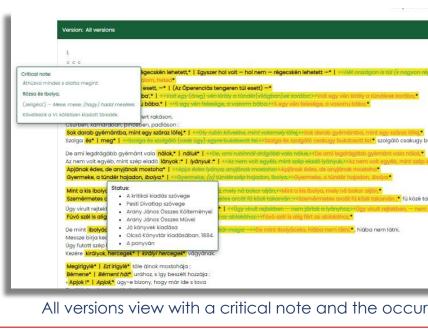
TEI XML Viewer is used to display the textual content and all added semantic, linguistic, textological, and critical information encoded into TEI XML files to make them easily explorable for human readers. It enables users to explore all versions of the same text at once or in parallel views, navigate 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 ion: Text layer Bismi'llähi'r-rahmäni'r-rähim ve bihi nestai'r ile'z-zálimîn ve <mark>'s-sálávátů</mark> ve 's-selámu alà-háyri hálkihi Muhammedin ve álihi ve ashábihi na'în. ^C Emmâ bâ'dû, iy Azîz! Bilgil ve âgâh olgil kim, bir niçe uli Ve her azîz ki gelmişdür, kendü zamânında gelen pâdişâhların târ ab eylemişlerdür. Hâllerin ve dirliklerin beyan itmişlerdi ád idüp zikr itmeklik vácibdúr. Zirá kim, es-sultán zil'lullál Split screen view of image and text layer with a critical note displayed

kiemelés	
Említések (Személyek)	
	. kritikai kiadás szövege): a hazugság, a király elhitte, Asszonyának…, ki s be a király elébe. Hogy
	ézirat): a hazugság, a király elhitte, A.,, mehetett s be a király elébe. Hagy
	esti Divatlap szövege): a hazugság, a király elhitte, A, mehetett s be a király elébe. Hogy
	rany János Összes Költeményei): a hazugság, a király elhitte, Asszonyának…, ki s be a király elébe. Hogy
	rany Jánas Összes Művei): a hazugság, a király elhitte, Asszonyának…, ki s be a király elébe. Hogy
	ó könyvek kiadása) : a hazugság <mark>, a</mark> király elhitte, Asszonyának, ki s be a király elébe. Hogy
	lleső Könyvtár kiadásában. 1884.): a hazugság, a király elhitte, Asszonyának…, ki s be a király elébe. Hogy
	ponyván); a hazugság, a király elhitte, Felcsattan, Szabadon s be a király elébe. Hogy
o.ajom3-ek.teixml (N	linden szövegváltozat): a hazugság, a király elhitte, Asszonyának, mehetett s be a király elébe. Hogy
2. Tükör I.	
Szerző: Móricz Zsigmond	
kiemelés	
Mű címe: Tükör I. Említések (Személyek)	
Említések (Személyek)	mi (Minden szövegyáltazat): se, Vilmos császár akar 307 – jó dolog. Ferenc Jázsef mogirta a levelet mindig imádkazom mieni
Említések (Személyek)	
Említések (Személyek) MZS_Tukor-1_TELteixi	mi (Minden szövegváltazat): se, Vilmos császár akar 307 – jó dolog. Ferenc Jázsef megirta a levelet mindig imádkozom mieni
Említések (Személyek) MZS_Tukor-1_TELteix	mi (Minden szövegváltazat): se, Vilmos császár akar 307 – jó dolog. Ferenc Jázsef megirta a levelet mindig imádkozom mieni
Említések (Személyek) MZS_Tukor-1_TELteixi Hamlet, dán királyfi.	mi (Minden szövegváttozat): sa, Vilmos császár okar 307 – jó dolog. Ferenc Jázsef mogitta a lovolet mindig imádkozom mieni IV. 🕒
Említések (Személyek) MZS_Tukor-1_TELteixi I. Hamlet, dán királyfi. kiemelés Említések (Személyek)	mi (Minden szövegváttozat): sa, Vilmos császár okar 307 – jó dolog. Ferenc Jázsef mogitta a lovolet mindig imádkozom mieni IV. 🕒
Említósek (Személyek) MZS_Tukor-1_TELteixa I. Hamlet, dán királyfi. kiemelés Említósek (Személyek) ajom7-hamlet.teixm	mi (Minden szövegváltozat): so, Vilmos császár okar 307 – " jó dolog. Ferenc Jászef mogirta a levelet mindig imádkozom mieni IV. 🔳
Emiltések (Személyek) MZS_Tukor-1_TElteixi Hamlet, dán királyfi. kiemelés Emiltések (Személyek) ajam7-hamlet.teixm ajam7-hamlet.teixm	mi (Minden szövegvátozat): se, Vilmos császár akar 307 – jó dolog. Ferenc Jázsef mogirta a lovelet mindig imádkazom mieni IV. 💼 I (A kritikai kiadás szövege): van, do a király nincs a test:, Laortos a király! Éljen Laortos (JARTES. Hol a király? – Maradjatok
Emiltősek (Szemőlyek) MZS_Tukor-1_TEiteixe Hamlet, dán királyfi. Kiemelős Emiltősek (Szemőlyek) ajam7-hamletteium ajam7-hamletteium ajam7-hamletteium	mi (Minden szövegvátozat): se, Vilmes császár okar 307 – jó dolog. Ferenc Jázsef megírta a levelet mindig imádkozom mieni IV. 💽 I (A kritikai kiadás szövege): van. de a kírály nincs a test : "Laertes a kírály! Eljon Laertes LAERTES. Hol a kírály? – Maradjatok I (KI): van. de a kírály nincs a test : "Laertes a kírály! Eljon Laertes LAERTES. Hol a kírály? – Maradjatok
Emiltések (Személyek) MZS_Tukor-1_TElteixe Hamlet, dán királyfi. kiemelés Emiltések (Személyek) ojom7-hamletteixm ojom7-hamletteixm ojom7-hamletteixm	mi (Mindan szövegvátozat): se, Vilmos császár akar 307 – jó dolog. Ferenc Jázsef mogirta a lovelet., mindig imádkazom mieni IV. (A kritikai kiadás szövege): (A kritikai kiadás szövege): van, do a király nincs a test., : Laortes a király Éjen Laortes, LAERTES, Hol a király? – Maradjatok (Kö): van, do a király nincs a test., : Laortes a király Éjen Laortes, LAERTES, Hol a király? – Maradjatok
Emiltések (Személyek) MZS_Tukor-1_TEiteixi B. Hamlet, dán királyfi. kiemelés Emiltések (Személyek) ajom?-hamletteixm ajom?-hamletteixm ajom?-hamletteixm ajom?-hamletteixm	II (Minden szövegvátozat): se, Vilmos császár akar 307 – jó dolog. Ferenc Jázsef mogirta a lovelet mindig imádkozom mieni IV. II. (A kritikai kiadás szövege): van, do a király nincs a test: Laertes a király Ejen Laertes LAERTES. Hol a király? – Maradjatok ((Q): van, do a király nincs a test: Laertes a király Ejen Laertes LAERTES. Hol a király? – Maradjatok ((Kad): van, do a király nincs a test: Laertes a király Ejen Laertes LAERTES. Hol a király? – Maradjatok ((Kad): van, do a király nincs a test: Laertes a király Ejen Laertes LAERTES. Hol a király? – Maradjatok





National Laboratory or Digital Heritage