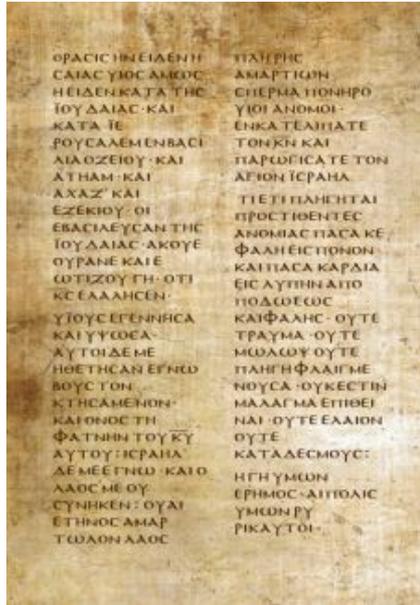


Halbautomatische Annotierung antiker Handschriften

Dr. Carina Geldhauser, MCML TUM

Ipek Tuncel, TUM & BAdW

HTR steps: pre-processing, segmentation, line recognition



ΟΡΑΣΙΣ ΗΝ ΕΙΔΕΝ Η
 ΣΑΙΑΣ ΥΙΟΣ ΑΜΕΩΣ
 ΗΕΙΔΕΝ ΚΑΤΑ ΤΗΣ
 ΤΟΥ ΔΑΙΑΣ· ΚΑΙ
 ΚΑΤΑ ΤΕ
 ΡΟΥΣΑΛΕΜ ΕΝ ΒΑΣΙ
 ΛΙΑ ΟΖΕΙΟΥ· ΚΑΙ
 ΑΤΗΑΜ· ΚΑΙ
 ΑΧΑΖ· ΚΑΙ
 ΕΖΕΚΙΟΥ· ΟΙ
 ΕΒΑΣΙΑΕΥΣΑΝ ΤΗΣ
 ΤΟΥ ΔΑΙΑΣ· ΑΚΟΥΕ
 ΟΥΡΑΝΕ ΚΑΙ Ε
 ΩΤΙ ΖΟΥ ΓΗ· ΟΤΙ
 ΚΣ ΕΛΑΛΗCΕΝ·
 ΥΙΟΥC ΕΓΕΝΝΗCΑ
 ΚΑΙ ΥΨΩCΑ·
 ΑΥΤΟΙ ΔΕ ΜΕ
 ΗΘΕΤΗCΑΝ ΕΓΝΩ
 ΒΟΥC ΤΟΝ
 ΚΤΗCΑΜΕΝΟΝ·
 ΚΑΙ ΟΝΟC ΤΗ
 ΦΑΤΗΝΗ ΤΟΥ ΚΥ
 ΑΥΤΟΥ· ΙCΡΑΗΛ·
 ΔΕ ΜΕ ΕΓΝΩ· ΚΑΙ Ο
 ΛΑΟC ΜΕ ΟΥ
 CΥΝΗΚΕΝ· ΟΥΔΙ
 ΕΤΗΝΟC ΑΜΑΡ
 ΤΩΛΟΝ ΛΑΟC

ΠΑΝΗΡΗC
 ΑΜΑΡΤΙΩΝ·
 CΠΕΡΜΑ ΠΟΝΗΡΟ
 ΥΙΟΙ ΑΝΟΜΟΙ·
 ΕΝΚΑ ΤΕΛΗΠΑΤΕ
 ΤΟΝ ΚΗ ΚΑΙ
 ΠΑΡΩΓΙCΑΤΕ ΤΟΝ
 ΑΓΙΟΝ ΙCΡΑΗΛ
 ΤΙ ΕΤΙ ΠΑΝΗΓΤΑΙ
 ΠΡΟCΤΙΘΕΝΤΕC
 ΑΝΟΜΙΑC ΠΑCΑ ΚΕ·
 ΦΑΛΗC ΠΟΝΟΝ·
 ΚΑΙ ΠΑCΑ ΚΑΡΔΙΑ
 ΕΙC ΑΥΠΗΝ ΑΠΟ
 ΠΟΔΩ ΕΩC
 ΚΑΙ ΦΑΛΗC· ΟΥΤΕ
 ΤΡΑΥΜΑ· ΟΥΤΕ
 ΜΩΛΩΨ· ΟΥΤΕ
 ΠΑΝΗΓΦΛΑΓΙΜΕ
 ΝΟΥCΑ· ΟΥΚ ΕCΤΙΝ
 ΜΑΛΑΓΜΑ ΕΠΙΘΕΙ
 ΝΑΙ· ΟΥΤΕ ΕΛΛΙΟΝ
 ΟΥΤΕ
 ΚΑΤΑΔΕCΜΟΥC·
 ΗΓΗ ΥΜΩΝ
 ΕΡΗΜΟC· ΑΙΠΟΛΙC
 ΥΜΩΝ ΡΥ
 ΡΙΚΑΥΤΟΙ·

ΟΡΑΣΙΣ ΗΝ ΕΙΔΕΝ Η
 ΣΑΙΑΣ ΥΙΟΣ ΑΜΕΩC
 ΗΕΙΔΕΝ ΚΑΤΑ ΤΗΣ
 ΤΟΥ ΔΑΙΑC· ΚΑΙ
 ΚΑΤΑ ΤΕ
 ΡΟΥCΑΛΕΜ ΕΝ ΒΑCΙ
 ΛΙΑ ΟΖΕΙΟΥ· ΚΑΙ
 ΑΤΗΑΜ· ΚΑΙ
 ΑΧΑΖ· ΚΑΙ
 ΕΒΑΣΙΑΕΥCΑΝ ΤΗC
 ΤΟΥ ΔΑΙΑC· ΑΚΟΥΕ
 ΟΥΡΑΝΕ ΚΑΙ Ε
 ΩΤΙ ΖΟΥ ΓΗ· ΟΤΙ
 ΚC ΕΛΑΛΗCΕΝ·
 ΥΙΟΥC ΕΓΕΝΝΗCΑ
 ΚΑΙ ΥΨΩCΑ·
 ΑΥΤΟΙ ΔΕ ΜΕ
 ΗΘΕΤΗCΑΝ ΕΓΝΩ
 ΒΟΥC ΤΟΝ
 ΚΤΗCΑΜΕΝΟΝ·
 ΚΑΙ ΟΝΟC ΤΗ
 ΦΑΤΗΝΗ ΤΟΥ ΚΥ
 ΑΥΤΟΥ· ΙCΡΑΗΛ·
 ΔΕ ΜΕ ΕΓΝΩ· ΚΑΙ Ο
 ΛΑΟC ΜΕ ΟΥ
 CΥΝΗΚΕΝ· ΟΥΔΙ
 ΕΤΗΝΟC ΑΜΑΡ
 ΤΩΛΟΝ ΛΑΟC

ΠΑΝΗΡΗC
 ΑΜΑΡΤΙΩΝ·
 CΠΕΡΜΑ ΠΟΝΗΡΟ
 ΥΙΟΙ ΑΝΟΜΟΙ·
 ΕΝΚΑ ΤΕΛΗΠΑΤΕ
 ΤΟΝ ΚΗ ΚΑΙ
 ΠΑΡΩΓΙCΑΤΕ ΤΟΝ
 ΑΓΙΟΝ ΙCΡΑΗΛ
 ΤΙ ΕΤΙ ΠΑΝΗΓΤΑΙ
 ΠΡΟCΤΙΘΕΝΤΕC
 ΑΝΟΜΙΑC ΠΑCΑ ΚΕ·
 ΦΑΛΗC ΠΟΝΟΝ·
 ΚΑΙ ΠΑCΑ ΚΑΡΔΙΑ
 ΕΙC ΑΥΠΗΝ ΑΠΟ
 ΠΟΔΩ ΕΩC
 ΚΑΙ ΦΑΛΗC· ΟΥΤΕ
 ΤΡΑΥΜΑ· ΟΥΤΕ
 ΜΩΛΩΨ· ΟΥΤΕ
 ΠΑΝΗΓΦΛΑΓΙΜΕ
 ΝΟΥCΑ· ΟΥΚ ΕCΤΙΝ
 ΜΑΛΑΓΜΑ ΕΠΙΘΕΙ
 ΝΑΙ· ΟΥΤΕ ΕΛΛΙΟΝ
 ΟΥΤΕ
 ΚΑΤΑΔΕCΜΟΥC·
 ΗΓΗ ΥΜΩΝ
 ΕΡΗΜΟC· ΑΙΠΟΛΙC
 ΥΜΩΝ ΡΥ
 ΡΙΚΑΥΤΟΙ·

ΟΡΑΣΙC ΗΝ ΕΙΔΕΝ Η
 ΣΑΙΑC ΥΙΟC ΑΜΕΩC
 ΗΕΙΔΕΝ ΚΑΤΑ ΤΗC
 ΤΟΥ ΔΑΙΑC· ΚΑΙ
 ΚΑΤΑ ΤΕ
 ΡΟΥCΑΛΕΜ ΕΝ ΒΑCΙ
 ΛΙΑ ΟΖΕΙΟΥ· ΚΑΙ
 ΑΤΗΑΜ· ΚΑΙ
 ΑΧΑΖ· ΚΑΙ
 ΕΒΑΣΙΑΕΥCΑΝ ΤΗC
 ΤΟΥ ΔΑΙΑC· ΑΚΟΥΕ
 ΟΥΡΑΝΕ ΚΑΙ Ε
 ΩΤΙ ΖΟΥ ΓΗ· ΟΤΙ
 ΚC ΕΛΑΛΗCΕΝ·
 ΥΙΟΥC ΕΓΕΝΝΗCΑ
 ΚΑΙ ΥΨΩCΑ·
 ΑΥΤΟΙ ΔΕ ΜΕ
 ΗΘΕΤΗCΑΝ ΕΓΝΩ
 ΒΟΥC ΤΟΝ
 ΚΤΗCΑΜΕΝΟΝ·
 ΚΑΙ ΟΝΟC ΤΗ
 ΦΑΤΗΝΗ ΤΟΥ ΚΥ
 ΑΥΤΟΥ· ΙCΡΑΗΛ·
 ΔΕ ΜΕ ΕΓΝΩ· ΚΑΙ Ο
 ΛΑΟC ΜΕ ΟΥ
 CΥΝΗΚΕΝ· ΟΥΔΙ
 ΕΤΗΝΟC ΑΜΑΡ
 ΤΩΛΟΝ ΛΑΟC

ΠΑΝΗΡΗC
 ΑΜΑΡΤΙΩΝ·
 CΠΕΡΜΑ ΠΟΝΗΡΟ
 ΥΙΟΙ ΑΝΟΜΟΙ·
 ΕΝΚΑ ΤΕΛΗΠΑΤΕ
 ΤΟΝ ΚΗ ΚΑΙ
 ΠΑΡΩΓΙCΑΤΕ ΤΟΝ
 ΑΓΙΟΝ ΙCΡΑΗΛ
 ΤΙ ΕΤΙ ΠΑΝΗΓΤΑΙ
 ΠΡΟCΤΙΘΕΝΤΕC
 ΑΝΟΜΙΑC ΠΑCΑ ΚΕ·
 ΦΑΛΗC ΠΟΝΟΝ·
 ΚΑΙ ΠΑCΑ ΚΑΡΔΙΑ
 ΕΙC ΑΥΠΗΝ ΑΠΟ
 ΠΟΔΩ ΕΩC
 ΚΑΙ ΦΑΛΗC· ΟΥΤΕ
 ΤΡΑΥΜΑ· ΟΥΤΕ
 ΜΩΛΩΨ· ΟΥΤΕ
 ΠΑΝΗΓΦΛΑΓΙΜΕ
 ΝΟΥCΑ· ΟΥΚ ΕCΤΙΝ
 ΜΑΛΑΓΜΑ ΕΠΙΘΕΙ
 ΝΑΙ· ΟΥΤΕ ΕΛΛΙΟΝ
 ΟΥΤΕ
 ΚΑΤΑΔΕCΜΟΥC·
 ΗΓΗ ΥΜΩΝ
 ΕΡΗΜΟC· ΑΙΠΟΛΙC
 ΥΜΩΝ ΡΥ
 ΡΙΚΑΥΤΟΙ·

Illustration: Ipek Tuncel

Vom basic XML zum custom XML

- **Testung verschiedener Transkriptionswerkzeuge:**
 - oft kostenpflichtig nach N Seiten
 - funktionieren nur für moderne Handschriften oder nur für Gedrucktes
 - Entscheidung für escriptorium: versatil, open-source
- escriptorium bietet bisher fast ausschließlich Transkription an
- einige Basisinformationen der Handschrift im “basic XML” vorhanden

Unser Ziel: halbautomatische Annotierung

Technische Herausforderungen bei der Annotierung

- “klassische Funktionen” oder “in den Modellparametern eingebaut”
 - feature-dependent
- Listen der zu annotierenden Wörter
 - custom-fetch aus andere Datenbanken?
 - Generierung mit LLMs?
- Export, Adaption und Weiterverarbeitung

Unser Beitrag: modularisierte Listen, Export als customXML

Short Demo

Please find the demo [here](#).

For questions & feedback: carina.geldhauser.math@gmail.com