

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
 ΩΤΙΖΟΥ ΓΗ· ΟΤΙ
 ΚΣ ΕΛΑΛΗΣΕΝ·
 ΥΙΟΥΣ ΕΓΕΝΝΗΣΑ
 ΚΑΙ ΥΨΩΣΑ·
 ΑΥΤΟΙ ΔΕ ΜΕ
 ΗΘΕΤΗΣΑΝ ΕΓΝΩ
 ΒΟΥΣ ΤΟΝ
 ΚΤΗΣΑΜΕΝΟΝ·
 ΚΑΙ ΟΝΟΣ ΤΗ
 ΦΑΤΗΝΗ ΤΟΥ ΚΥ
 ΑΥΤΟΥ· ΙΣΡΑΗΛ
 ΔΕ ΜΕ ΕΓΝΩ· ΚΑΙ Ο
 ΛΑΟΣ ΜΕ ΟΥ
 ΣΥΝΗΚΕΝ· ΟΥΔΑΙ
 ΕΤΗΝΟΣ ΑΜΑΡ
 ΤΩΛΟΝ ΛΑΟΣ

ΠΑΝΗΡΗΣ
 ΑΜΑΡΤΙΩΝ·
 ΣΠΕΡΜΑ ΠΟΝΗΡΟ
 ΥΨΙΟΙ ΑΝΟΜΟΙ·
 ΕΝΚΑ ΤΕΛΗΠΑΤΕ
 ΤΟΝ ΚΗ ΚΑΙ
 ΠΑΡΩΓΙΣΑΤΕ ΤΟΝ
 ΑΓΙΟΝ ΙΣΡΑΗΛ
 ΤΙ ΕΤΙ ΠΑΝΗΓΤΑΙ
 ΠΡΟΣΤΙΘΕΝΤΕΣ
 ΑΝΟΜΙΑΣ ΠΑΣΑ ΚΕ·
 ΦΑΛΗ ΕΙΣ ΠΟΝΟΝ·
 ΚΑΙ ΠΑΣΑ ΚΑΡΔΙΑ
 ΕΙΣ ΑΥΠΗΝ ΑΠΟ
 ΠΟΔΩ ΕΩΣ
 ΚΑΙ ΦΑΛΗΣ· ΟΥΤΕ
 ΤΡΑΥΜΑ· ΟΥΤΕ
 ΜΩΛΩΦ· ΟΥΤΕ
 ΠΑΝΗΓ ΦΛΑΓΙΜΕ
 ΝΟΥΣΑ· ΟΥΚ ΕΣΤΙΝ
 ΜΑΛΑΓΜΑ ΕΠΙΘΕΙ
 ΝΑΙ· ΟΥΤΕ ΕΛΛΙΟΝ
 ΟΥΤΕ
 ΚΑΤΑΔΕΣΜΟΥΣ·
 ΗΓΗ ΥΜΩΝ
 ΕΡΗΜΟΣ· ΑΙΠΟΛΙΣ
 ΥΜΩΝ ΡΥ
 ΡΙΚΑΥΤΟΙ·

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
 ΩΤΙΖΟΥ ΓΗ· ΟΤΙ
 ΚΣ ΕΛΑΛΗΣΕΝ·
 ΥΙΟΥΣ ΕΓΕΝΝΗΣΑ
 ΚΑΙ ΥΨΩΣΑ·
 ΑΥΤΟΙ ΔΕ ΜΕ
 ΗΘΕΤΗΣΑΝ ΕΓΝΩ
 ΒΟΥΣ ΤΟΝ
 ΚΤΗΣΑΜΕΝΟΝ·
 ΚΑΙ ΟΝΟΣ ΤΗ
 ΦΑΤΗΝΗ ΤΟΥ ΚΥ
 ΑΥΤΟΥ· ΙΣΡΑΗΛ
 ΔΕ ΜΕ ΕΓΝΩ· ΚΑΙ Ο
 ΛΑΟΣ ΜΕ ΟΥ
 ΣΥΝΗΚΕΝ· ΟΥΔΑΙ
 ΕΤΗΝΟΣ ΑΜΑΡ
 ΤΩΛΟΝ ΛΑΟΣ

ΠΑΝΗΡΗΣ
 ΑΜΑΡΤΙΩΝ·
 ΣΠΕΡΜΑ ΠΟΝΗΡΟ
 ΥΨΙΟΙ ΑΝΟΜΟΙ·
 ΕΝΚΑ ΤΕΛΗΠΑΤΕ
 ΤΟΝ ΚΗ ΚΑΙ
 ΠΑΡΩΓΙΣΑΤΕ ΤΟΝ
 ΑΓΙΟΝ ΙΣΡΑΗΛ
 ΤΙ ΕΤΙ ΠΑΝΗΓΤΑΙ
 ΠΡΟΣΤΙΘΕΝΤΕΣ
 ΑΝΟΜΙΑΣ ΠΑΣΑ ΚΕ·
 ΦΑΛΗ ΕΙΣ ΠΟΝΟΝ·
 ΚΑΙ ΠΑΣΑ ΚΑΡΔΙΑ
 ΕΙΣ ΑΥΠΗΝ ΑΠΟ
 ΠΟΔΩ ΕΩΣ
 ΚΑΙ ΦΑΛΗΣ· ΟΥΤΕ
 ΤΡΑΥΜΑ· ΟΥΤΕ
 ΜΩΛΩΦ· ΟΥΤΕ
 ΠΑΝΗΓ ΦΛΑΓΙΜΕ
 ΝΟΥΣΑ· ΟΥΚ ΕΣΤΙΝ
 ΜΑΛΑΓΜΑ ΕΠΙΘΕΙ
 ΝΑΙ· ΟΥΤΕ ΕΛΛΙΟΝ
 ΟΥΤΕ
 ΚΑΤΑΔΕΣΜΟΥΣ·
 ΗΓΗ ΥΜΩΝ
 ΕΡΗΜΟΣ· ΑΙΠΟΛΙΣ
 ΥΜΩΝ ΡΥ
 ΡΙΚΑΥΤΟΙ·

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
 ΩΤΙΖΟΥ ΓΗ· ΟΤΙ
 ΚΣ ΕΛΑΛΗΣΕΝ·
 ΥΙΟΥΣ ΕΓΕΝΝΗΣΑ
 ΚΑΙ ΥΨΩΣΑ·
 ΑΥΤΟΙ ΔΕ ΜΕ
 ΗΘΕΤΗΣΑΝ ΕΓΝΩ
 ΒΟΥΣ ΤΟΝ
 ΚΤΗΣΑΜΕΝΟΝ·
 ΚΑΙ ΟΝΟΣ ΤΗ
 ΦΑΤΗΝΗ ΤΟΥ ΚΥ
 ΑΥΤΟΥ· ΙΣΡΑΗΛ
 ΔΕ ΜΕ ΕΓΝΩ· ΚΑΙ Ο
 ΛΑΟΣ ΜΕ ΟΥ
 ΣΥΝΗΚΕΝ· ΟΥΔΑΙ
 ΕΤΗΝΟΣ ΑΜΑΡ
 ΤΩΛΟΝ ΛΑΟΣ

ΠΑΝΗΡΗΣ
 ΑΜΑΡΤΙΩΝ·
 ΣΠΕΡΜΑ ΠΟΝΗΡΟ
 ΥΨΙΟΙ ΑΝΟΜΟΙ·
 ΕΝΚΑ ΤΕΛΗΠΑΤΕ
 ΤΟΝ ΚΗ ΚΑΙ
 ΠΑΡΩΓΙΣΑΤΕ ΤΟΝ
 ΑΓΙΟΝ ΙΣΡΑΗΛ
 ΤΙ ΕΤΙ ΠΑΝΗΓΤΑΙ
 ΠΡΟΣΤΙΘΕΝΤΕΣ
 ΑΝΟΜΙΑΣ ΠΑΣΑ ΚΕ·
 ΦΑΛΗ ΕΙΣ ΠΟΝΟΝ·
 ΚΑΙ ΠΑΣΑ ΚΑΡΔΙΑ
 ΕΙΣ ΑΥΠΗΝ ΑΠΟ
 ΠΟΔΩ ΕΩΣ
 ΚΑΙ ΦΑΛΗΣ· ΟΥΤΕ
 ΤΡΑΥΜΑ· ΟΥΤΕ
 ΜΩΛΩΦ· ΟΥΤΕ
 ΠΑΝΗΓ ΦΛΑΓΙΜΕ
 ΝΟΥΣΑ· ΟΥΚ ΕΣΤΙΝ
 ΜΑΛΑΓΜΑ ΕΠΙΘΕΙ
 ΝΑΙ· ΟΥΤΕ ΕΛΛΙΟΝ
 ΟΥΤΕ
 ΚΑΤΑΔΕΣΜΟΥΣ·
 ΗΓΗ ΥΜΩΝ
 ΕΡΗΜΟΣ· ΑΙΠΟΛΙΣ
 ΥΜΩΝ ΡΥ
 ΡΙΚΑΥΤΟΙ·

Illustration: Ipek Tuncel

HTR: weitere Schritte: word splitting, post-processing

ΘΡΑΣΙΣ ΗΝ ΕΙΔΕΝ Η
ΣΑΙΑΣ ΥΙΟΣ ΑΜΩΣ
ΗΕΙΔΕΝ ΚΑΤΑ ΤΗΣ
ΤΟΥ ΔΑΙΑΣ ΚΑΙ
ΚΑΤΑ ΤΕ
ΡΟΥΣ ΑΕΜΕΝ ΒΑΣΙ
ΔΙΑΟΖΕΙΟΥ ΚΑΙ
ΑΤΗΑΜ ΚΑΙ
ΑΧΑΣ ΚΑΙ
ΕΖΕΚΙΟΥ ΟΙ
ΕΒΑΣΙΑΕΥΣΑΝ ΤΗΣ
ΤΟΥ ΔΑΙΑΣ ΑΚΟΥΕ
ΟΥΡΑΝΕ ΚΑΙ Ε
ΩΤΙ ΖΟΥ ΓΗ ΟΤΙ
ΚΕ ΕΑΛΛΗΣΕΝ
ΥΙΟΥΣ ΕΓΕΝΗΝ ΣΑ
ΚΑΙ ΨΥΧΑ
ΑΥΤΟΙ ΔΕ ΜΕ
ΗΘΕΤΗΣΑΝ ΕΓΜΩ
ΒΟΥΣ ΤΟΝ
ΚΙΝΗΣΑΜΕΝΟΝ
ΚΑΙ ΟΝΟΣ ΤΗ
ΦΑΤΗΝ ΤΟΥ ΚΥ
ΔΥΤΟΥ ΙΣΡΑΗΛ
ΔΕ ΜΕ ΕΓΜΩ ΚΑΙ Ο
ΔΔΟΣ ΜΕ ΟΥ
ΣΥΝΗΚΕΝ ΟΥΔΙ
ΕΤΙ ΗΝΟΣ ΑΜΑΡ
ΤΩΔΑΘΝ ΔΑΟΣ

ΠΑΙΡΗΣ
ΑΜΑΡΤΙΩΝ
ΣΠΕΡΜΑ ΠΟΝΗΡΟ
ΥΙΟΙ ΑΝΘΩΝ
ΕΝ ΚΑΤΕΛΗΝΑΤΕ
ΤΟΝ ΚΗ ΚΑΙ
ΠΑΡΩΣΙΣΑΤΕ ΤΟΝ
ΑΓΙΟΝ ΙΣΡΑΗΛ
ΤΙ ΕΤΙ ΠΑΝΗΤΑΙ
ΠΡΟΣΤΗΘΕΝΤΕΣ
ΑΝΟΜΙΑΣ ΠΑΣΑ ΚΕ
ΦΑΛΗΙΣ ΠΟΝΟΝ
ΚΑΙ ΠΑΣΑ ΚΑΡΔΙΑ
ΕΙΣ ΑΥΠΗΝΑ ΠΟ
ΠΟΔΩ ΕΩΣ
ΚΑΙ ΦΑΛΗΣ ΟΥΤΕ
ΤΡΑΥΜΑ ΟΥΤΕ
ΜΩΛΟΣ ΟΥΤΕ
ΠΑΝΗ ΦΑΛΗ ΜΕ
ΝΟΥΣ Α ΟΥΚ ΕΣΤΙΝ
ΜΑΛΛΑ ΓΜΑ ΕΠΗΘΕΙ
ΝΑΙ ΟΥΤΕ ΕΑΛΙΟΝ
ΟΥΤΕ
ΚΑΤΑ ΔΕ ΣΜΟΥΣ
Η ΠΗΛΜΩΝ
ΕΡΗΜΟΣ ΑΠΟΛΙΣ
ΥΜΩΝ ΡΥ
ΡΙΚΑΥΤΟΙ

ΘΡΑΣΙΣ ΗΝ ΕΙΔΕΝ Η
ΣΑΙΑΣ ΥΙΟΣ ΑΜΩΣ
ΗΕΙΔΕΝ ΚΑΤΑ ΤΗΣ
ΤΟΥ ΔΑΙΑΣ ΚΑΙ
ΚΑΤΑ ΤΕ
ΡΟΥΣ ΑΕΜΕΝ ΒΑΣΙ
ΔΙΑΟΖΕΙΟΥ ΚΑΙ
ΑΤΗΑΜ ΚΑΙ
ΑΧΑΣ ΚΑΙ
ΕΖΕΚΙΟΥ ΟΙ
ΕΒΑΣΙΑΕΥΣΑΝ ΤΗΣ
ΤΟΥ ΔΑΙΑΣ ΑΚΟΥΕ
ΟΥΡΑΝΕ ΚΑΙ Ε
ΩΤΙ ΖΟΥ ΓΗ ΟΤΙ
ΚΕ ΕΑΛΛΗΣΕΝ
ΥΙΟΥΣ ΕΓΕΝΗΝ ΣΑ
ΚΑΙ ΨΥΧΑ
ΑΥΤΟΙ ΔΕ ΜΕ
ΗΘΕΤΗΣΑΝ ΕΓΜΩ
ΒΟΥΣ ΤΟΝ
ΚΙΝΗΣΑΜΕΝΟΝ
ΚΑΙ ΟΝΟΣ ΤΗ
ΦΑΤΗΝ ΤΟΥ ΚΥ
ΔΥΤΟΥ ΙΣΡΑΗΛ
ΔΕ ΜΕ ΕΓΜΩ ΚΑΙ Ο
ΔΔΟΣ ΜΕ ΟΥ
ΣΥΝΗΚΕΝ ΟΥΔΙ
ΕΤΙ ΗΝΟΣ ΑΜΑΡ
ΤΩΔΑΘΝ ΔΑΟΣ

ΠΑΙΡΗΣ
ΑΜΑΡΤΙΩΝ
ΣΠΕΡΜΑ ΠΟΝΗΡΟ
ΥΙΟΙ ΑΝΘΩΝ
ΕΝ ΚΑΤΕΛΗΝΑΤΕ
ΤΟΝ ΚΗ ΚΑΙ
ΠΑΡΩΣΙΣΑΤΕ ΤΟΝ
ΑΓΙΟΝ ΙΣΡΑΗΛ
ΤΙ ΕΤΙ ΠΑΝΗΤΑΙ
ΠΡΟΣΤΗΘΕΝΤΕΣ
ΑΝΟΜΙΑΣ ΠΑΣΑ ΚΕ
ΦΑΛΗΙΣ ΠΟΝΟΝ
ΚΑΙ ΠΑΣΑ ΚΑΡΔΙΑ
ΕΙΣ ΑΥΠΗΝΑ ΠΟ
ΠΟΔΩ ΕΩΣ
ΚΑΙ ΦΑΛΗΣ ΟΥΤΕ
ΤΡΑΥΜΑ ΟΥΤΕ
ΜΩΛΟΣ ΟΥΤΕ
ΠΑΝΗ ΦΑΛΗ ΜΕ
ΝΟΥΣ Α ΟΥΚ ΕΣΤΙΝ
ΜΑΛΛΑ ΓΜΑ ΕΠΗΘΕΙ
ΝΑΙ ΟΥΤΕ ΕΑΛΙΟΝ
ΟΥΤΕ
ΚΑΤΑ ΔΕ ΣΜΟΥΣ
Η ΠΗΛΜΩΝ
ΕΡΗΜΟΣ ΑΠΟΛΙΣ
ΥΜΩΝ ΡΥ
ΡΙΚΑΥΤΟΙ

```
</ab>  
</div>  
<div id="G01.3">  
<ab id="G01.3.14">  
<ab id="f.56r-1.20">  
<span reason="sampling"/>  
<ab id="f.56r-1.21">  
<span reason="sampling"/>  
<seg type="equivalenceInit">  
<seg id="lang-grc">  
<ab id="r1">  
<span rend="larger">εϋε/ηι</span>  
</ab>  
</seg>  
<seg id="lang-lat">εϋε/ηι</seg>  
</seg>  
<ab id="f.56r-1.21">  
<span type="equivalenceInit">  
<seg id="lang-grc">  
<ab id="r2">  
<span rend="larger">εϋε/ηι</span>  
</ab>  
</seg>  
<seg type="equivalenceInit">  
<seg id="lang-grc">  
<ab id="r3">  
<choice ana="orthography">  
<span>εϋε/ηι</span>  
<span>εϋε/ηι</span>  
</choice>  
</ab>  
</seg>  
<seg id="lang-lat">  
<span>εϋε/ηι</span>  
</seg>  
</seg>  
<span type="equivalenceInit">  
<seg id="lang-grc">  
<ab id="r4">  
<span rend="larger">εϋε/ηι</span>  
</ab>  
</seg>  
<span type="equivalenceInit">  
<seg id="lang-grc">  
<ab id="r5">  
<choice ana="abbreviation">  
<span>εϋε</span>  
<span>εϋε</span>  
</choice>  
</ab>  
</seg>  
</div>  
</div>
```

Illustration: Ipek Tuncel, Juan Garcez

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