Munich Center for Machine Learning

Halbautomatische Annotierung antiker Handschriften

Dr. Carina Geldhauser, MCML TUM

Ipek Tuncel, TUM & BAdW

DHd 2024

HTR steps: pre-processing, segmentation, line recognition

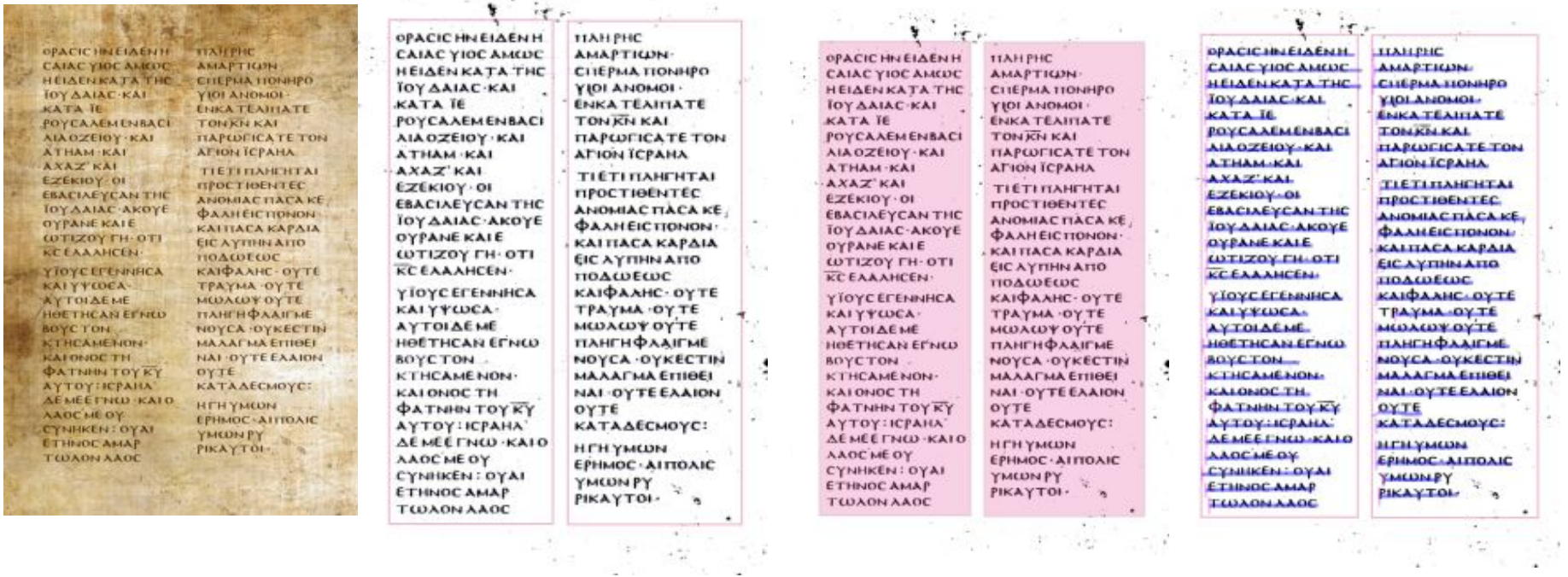

Illustration: Ipek Tuncel

HTR: weitere Schritte: word splitting, post-processing

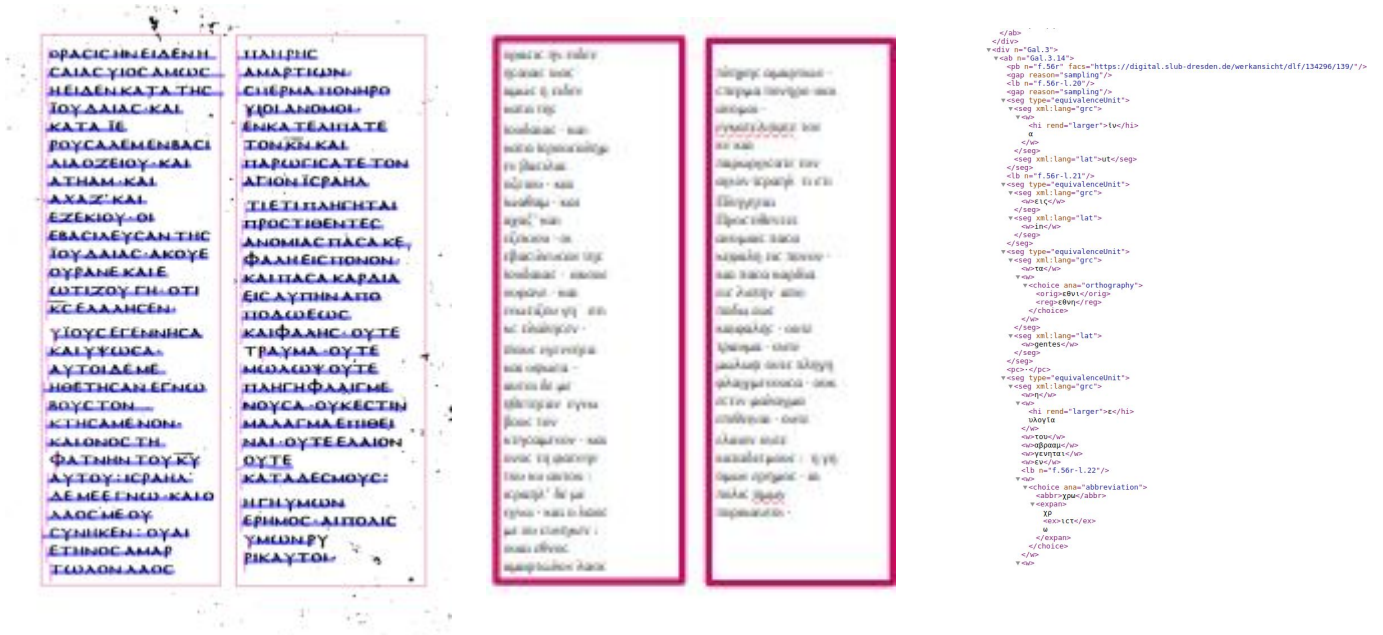

Illustration: Ipek Tuncel, Juan Garcez

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