A use-case in Digital Humanities – Some reflections on the project eAQUA

Gerhard Heyer

Universität Leipzig

heyer@informatik.uni-leipzig.de

UNIVERSITÄT LEIPZIG

Institut für Informatik

ASV

Automatische Sprachverarbeitung

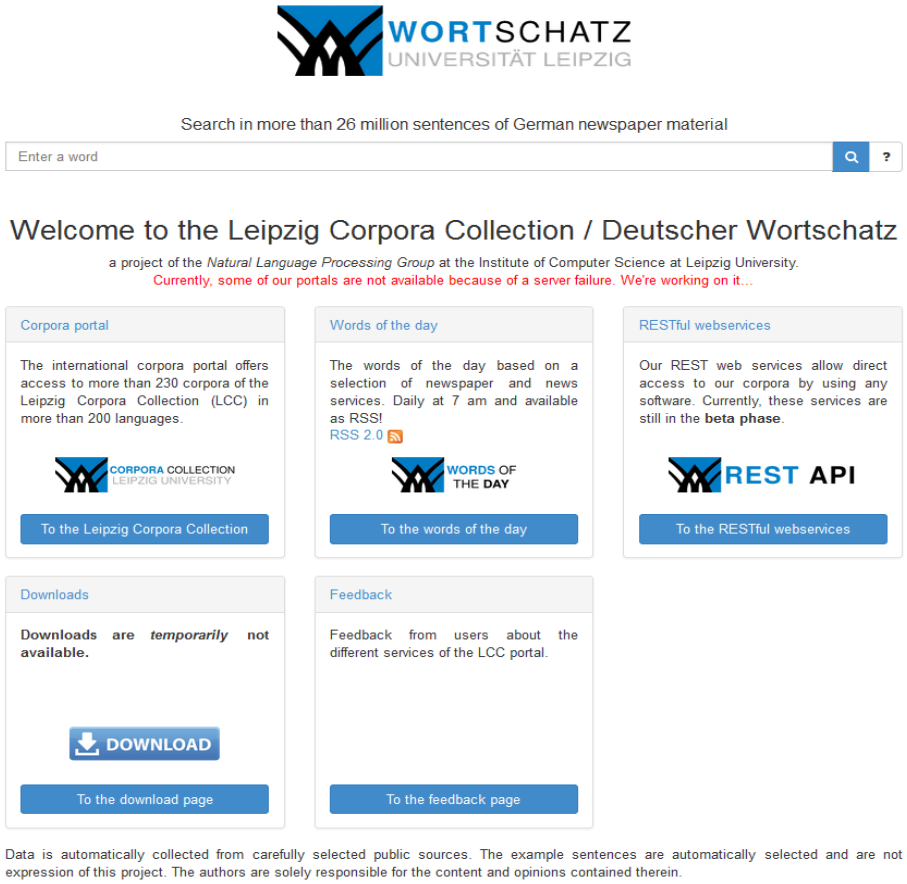
Automatische Sprachverarbeitung

The BMBF project eAQUA

- **Title**: „Extraktion von strukturiertem Wissen aus Antiken Quellen für die Altertumswissenschaft“
- **Duration**: 1.4.2008 – 31.3.2011
- **Portal:** still running under http://www.eaqua.net/, see also https://de.wikipedia.org/wiki/EAQUA
- **Participants**: ASV and classics department Leipzig University
- **Goals**: apply text mining to collections of classical text (Latin, Greek) in order to extract specific knowledge
- **Key developers**: Marco Büchler and Thomas Eckart

Automatische Sprachverarbeitung

Wortschatz http://wortschatz.uni-leipzig.de/en



Automatische Sprachverarbeitung

Corpora http://corpora.uni-leipzig.de/de?corpusId=pol_newscrawl_2011

Automatische Sprachverarbeitung

Subprojects

- **Atthidographers** – Classification of Atthidographers
- **Plato** - Aftermath of Plato's writings
- **Plautine metric** – Metrical analysis of Latin comedies
- **CAMENA** - Knowledge network of Latin CAMENA corpus
- **Inscriptions** - Extraction of significant templates for different kinds of inscriptions like release documents (slave trading)
- **Papyri** - Classification of papyri (e. g. slave trading contracts) and text completion of fragmentary texts
- **Mental maps** - Building of century based mental maps and highlighting differences

Challenges from Comuter Science point of view

- **Integrating textual ressources**

  Conversion of different formats and standards

- **Software engineering issues**

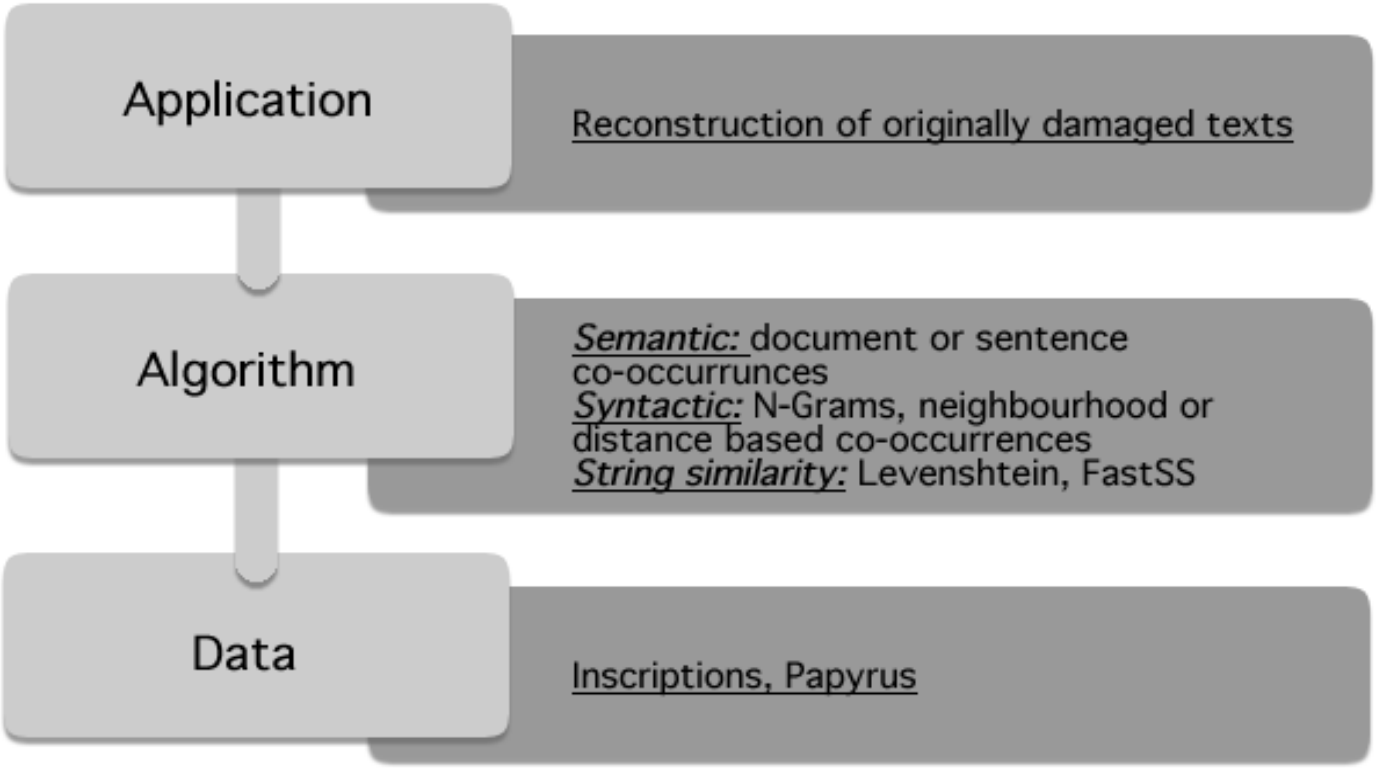
  Getting classicists to understand the intent and effects of text mining algorithms and clearly define their functional requirements

- **Semantic technologies that address the long tail**

  Modifying and amending statistical and pattern based methods to effectively deal with rare events

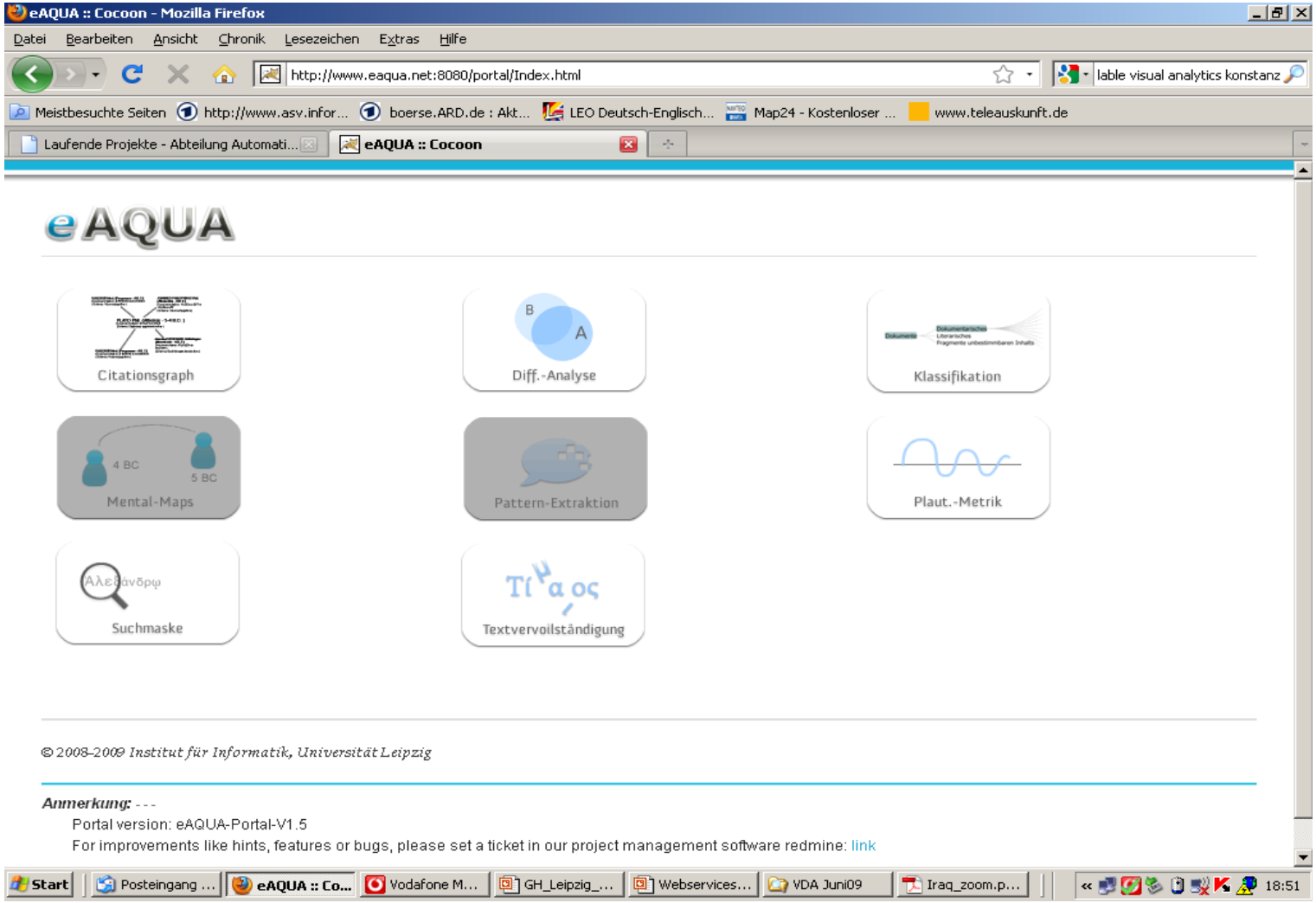
Automatische Sprachverarbeitung

Approach



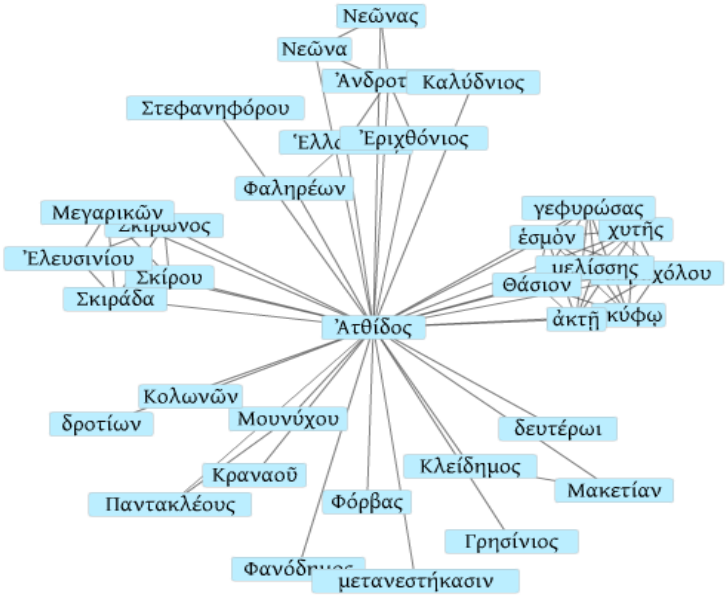
Automatische Sprachverarbeitung

Access to portal



ASV Automatische Sprachverarbeitung

Example: Searching for „Ἀτθίδος“



Automatische Sprachverarbeitung

Co-occurrences to find „interesting“ neighbours

Signifikante rechte Kookkurrenzen für *Ἀτθίδος*

ἀκραχόλου (5); σκύφῳ (5); Παντακλέους (5); Μουνύχου (5); F (12); συγκυρκανήσας (4); Σκίρου (5); χυτῆς (5); Πραξίων (4); Ἐλευσινίου (5); Νεῶνας (4); Κραναοῦ (5); FHG (9); μελίσσης (6); γεφυρώσας (5); ἐσμὸν (5); Σκίρωνος (4); ἀκτῇ (5); μετανεστήκασιν (3); Μεγαρικῶν (4); Ἀτθίδος (5); Φαληρέων (3); μάντεως (5); καύματος (5); I (9); κεκλῆσθαι (5); ὠνομάσθαι (5); φησίν (13); βιβλία (6); Ἀρκεσινεῦσιν (2); Ἀνδανίῃ (2); Ἁλικαρνάσσιός (2); Μινωίταις (2); κατῃσίμωκα (2); κατεστιλβωμένον (2); βουλῆς (2); ψιλοκιθαριστικήν (2); Δήμητρος (5); Μελίτῃ (3); δημιουργικὰς (3); πῶμα (4); FGrHist (4); Ἀριστάρχειόν (2); Ἀμοργίοις (2); Ἐπίγονον (2); διαδικασίαι (2); Κλείδημος (3); νομάδας (3); γλυκὺν (4); Μακετίας (2); Ἐλεγείαις (2); φησὶν (13); ἐξῳκίσθησαν (2); γυναικονόμοι (2); ὁμωνύμωι (2); τινὸς (9); Κεκροπίδην (2); συγγενομένης (2); ἀπεχειροτονήθη (2); ἑάλωσάν (2); ἐῆς (2); Ἄγραυλον (2); Ἁλωπεκόννησον (2); διεξῆλθεν (3); ἐπεποιήκη (2); κιθάρισιν (2); μαγείρων (3); λύσιν (5); λίθου (5); πειθήμονα (2); κεδνῇ (2); ὁμογάλακτας (2); θέσμιον (2); Αἰγιαλὸν (2); Κῆρυξ (2); Ἁλιρρόθιον (2); κόβαλον (2); εὐογκον (2); σποράδην (3); Θυμοίτης (2); Ἀλκίππην (2); δράματι (3); βωμολόχον (2); νομισθὲν (2); Εὔδοξος (3); διαλέκτου (3); αὐτόχθονος (2); Ἀρεοπαγιτῶν (2); Δείναρχος (3); οἰκήσεις (3); θεωρικὸν (2); κιθαριστὴς (2); ἀπέτισε (2); θυγατρός (3); ἡγεμονεύων (2); Ξούθου (2); Ἐριχθονίου (2); φῦλον (3); Κεραμεικῷ (2); ἰδίωμα (3);

- νομάδας **is an interesting candidate**
- **Reference to quotation from the Atthis of Philochorus (3. / 2. century BC) in a normal word search hard to find**
- **Co-occurrence indicates that atthidographers may have invented a phase in Athenian history to show an evolutionary development from *nomadism* to settled life**

Citation analysis: Plato and Neo-Platonism

Automatische Sprachverarbeitung

Assessment – What can we learn from the project?

- **What we did well**
  - Sustainable portal, well documented
  - Demonstration of feasibility to apply text mining in the domain of classics

[Marco Büchler, Gerhard Heyer und Sabine Gründer: eAQUA - Bringing modern Text Mining approaches to two thousand years old ancient texts. In: Proceedings of the 4th International Conference on e-Science (IEEE08), 2008]

What we did not do well

- Balanced integration of requirements analysis and implementation
  - Relying on waterfall model of SE resulted in unclear specifications, ill-defined roles, and substantial delays
  - Agile paradigm would have been much more appropriate

What we did not do well

- Apply tools available at that time without designing solutions to particular research questions
  - co-occurrences, differential analysis, …
  - no word embeddings, topic models, vector space semantics, ...
- Taking tools offered by CS as a solution to research questions instead of a challenge to redesign the research question and rethink the research question in terms of computational requirements

Automatische Sprachverarbeitung

Thanks for listening ...

Automatische Sprachverarbeitung

Thanks for listening ...

Automatische Sprachverarbeitung

Thanks for listening ...

Questions please!