Workshop: "Mehr Personen - Mehr Daten - Mehr Repositorien"
04.–06. März 2013
Berlin-Brandenburgische Akademie der Wissenschaften

Jurish, Bryan; Thomas, Christian (Berlin-Brandenburgische Akademie der Wissenschaften)

Vortrag:

Named Entity Recognition (NER) im Deutschen Textarchiv (DTA).
Computerlinguistisch gestützte Identifikation von Personen- und Ortsnamen in den Korpora des DTA.

Themen: Kooperation, Elektronische Biografik, Digitale Ressourcen, Semantische Technologien, Normdateien/Standards, Personenforschung, Digital Humanities Wissenschaftliche Nutzbarkeit, Sprach- und Literaturwissenschaft
Stichworte: Named Entity Recognition (NER), Corpus Linguistics, Normalization, TEI-XML

Zusammenfassung: Der Beitrag gibt einen Einblick in die automatisierte sowie die manuelle, computerlinguistisch gestützte Identifikation von Personen- und Ortsnamen in den Korpora des Deutschen Textarchivs (DTA, www.deutschestextarchiv.de). Diese Arbeiten erfolgen unter anderem im Rahmen des DFG-geförderten Projekts AEDit Frühe Neuzeit. Eine Auswahl von Texten aus dem DTA-Kernkorpus und aus dem AEDit-Korpus wurde zunächst manuell getaggt, um eine Grundlage für die Evaluation (semi-)automatischer Verfahren zur NER in deutschsprachigen historischen Texten zu schaffen. Die Evaluation schloss den Vergleich des im DTA entwickelten Ansatzes – einer Kombination des moot-Taggers und des synaktischen Parsers SynCoPe – mit zwei weiteren NER-Tools. Die Ergebnisse dieser Evaluation und die bestehenden Herausforderungen, die die angestrebte, weitgehende Automatisierung des NE-Tagging gerade in heterogenen historischen Textkorpora bedeutet, werden diskutiert.

Folien: http://deutschestextarchiv.de/dtaq/files/DTAE-NER_vortrag-2013-03-06.pdf