Mit der FinderApp durch Goethes Faust: Treffer im Faksimile visuell hervorgehoben und multimediale Ausgabe in Videoaufführung und Hörbuch.

,
maximilian@cis.uni-muenchen.de
Ludwig Maximilians Universität München, Deutschland

,
e.eder@campus.lmu.de
Ludwig Maximilians Universität München, Deutschland

,
r.capsamun@campus.lmu.de
Ludwig Maximilians Universität München, Deutschland

,
nora.eichfeldt@campus.lmu.de
Ludwig Maximilians Universität München, Deutschland

,
s.herteis@campus.lmu.de
Ludwig Maximilians Universität München, Deutschland

,
m.lindinger@live.de
Ludwig Maximilians Universität München, Deutschland

,
R.Hoeps@campus.lmu.de
Ludwig Maximilians Universität München, Deutschland

,
stefan@schweter.eu
Ludwig Maximilians Universität München, Deutschland

Einleitung

In unserem Poster möchten wir unsere neueste FinderApp GoetheFind vorstellen, die mit computerlinguistischen Methoden die Originalausgabe von Goethes Faust durchsucht. GoetheFind hat einen neuen browser-basierten Faksimile-Viewer, der die Schaltstelle der multimedialen Ausgabe der Suchtreffer darstellt: Die Treffer werden im Faksimile der Originalausgaben gehighlighted dargestellt und mit einer gesprochenen Faust-Hörbuchausgabe verlinkt. Bei Faust I werden die Treffer mit der entsprechenden Szene des Videos der Bühnenaufführung vom Hamburger Schauspielhaus mit Gustav Gründgens (1960) verlinkt. GoetheFind entstand aus unserer FinderApp WiTTFind (Hadersbeck / Pichler; Hadersbeck et al. 2014), die den öffentlich zugänglichen Teil des Nachlasses von Ludwig Wittgenstein durchsucht und mit der wir im Sommer 2014 den EU-AWARD des EU-Projekt Digitised Manuscripts to Europeana (cf. Ploeger 2014) gewannen.

In unserer neuen FinderApp GoetheFind, setzen wir Ideen des „Standoff-Markups“ um, damit „overtagged“-XML vermieden wird. Wir entwickelten eine reduzierte „XML-TEI-P5 anchor-key“ Edition und speichern die Metainformatione in einer „NoSQL-mongo“-Datenbank. Alle relevanten Editions-, OCR- und Transkriptionsinformationen zur multimedialen Trefferausgabe sind in der Datenbank gespeichert.

Modellierung der Editionsdaten

Anstatt „overtagged“ XML ein reduziertes „anchor-key“ XML-TEI-P5 mit „NoSQL“-Database

Grundlage unserer FinderApp GoetheFind ist die XML-TEI-P5 Textedition im DTABf Format vom Deutschen Textarchiv (DTA) der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW 2013; Haaf et al. 2015), dazu die Bilddigitalisate der Staatsbibliothek zu Berlin (BBAW 2013) und dem freiem Hochstift Frankfurt des Frankfurter Goethe-Hauses (Signatur: III B / 23). Da wir zur multimedialen Ausgabe der Suchtreffer zahlreiche Metainformationen benötigen und "overtagged" XML des Editionstexts vermeiden wollten, verwenden wir Ideen des „Standoff-Markups“ und lagern alle notwendigen Meta-Informationen in der „NoSQL“ mongo-Datenbank GoetheDB aus. Eine eindeutige Referenz der Datenbankeinträge zum Editionstext lösen wir über den XML-TEI-P5 Tag <anchor/>, der an Seitenanfängen in die Edition eingefügt ist. Die Trefferpositionen werden über ein XML-Attributtrippel (Seite, Zeile, Token) genau spezifiziert.

Vorverarbeitung der Edition, Faksimile, Videoaufführung und Audioaufnahme

Da unsere FinderApp die Suchanfragen regelbasiert mit Hilfe von lokalen Grammatiken im Kontext eines Satzes realisiert, verwenden wir als wichtigste

Strukturierungsebene Sätze. Goethes Faustdrama bettet Sätze in Rede und Gegenrede, sogenannte Repliken ein, die die zweite Strukturierungsebene darstellt. Zur visuellen Hervorhebung und multimedialen Ausgabe der gefundenen Textstellen im Faksimile ermitteln wir für die Replike geometrische Informationen mit Hilfe eines von uns entwickelten semiautomatischem OCR-Correction Tools. Die Bühnen- und Audioaufnahme werden mit Hilfe des Clarin-Tools: „Munich Automatic Segmentation System WebMAUS“ (CLARIN) semiautomatisch transkribiert.

Computerlinguistische Methoden zur Textinterpretation

Mit Hilfe unseres Speziallexikons GoetheLEX, angereichert um historische Sprachvarianten, Part of Speech Tagging und lokalen Grammatiken implementierten wir eine Partikelverb- und Semantische Suche. Bei der Eingabe von Suchanfragen verwenden wir eine symmetrische Autovervollständigung mit Informationen zur Häufigkeit des Auftretens im Text.

Treffer im Faksimile-Viewer und multimediale Ausgabe

Ähnlich wie bei Google-Docs entwickelten wir einen Browser basierter Faksimile-Viewer mit dem man in einem doppelseitigen Buchlesemodus durch das Dokument blättern kann und die gefundenen Textstellen farblich hervorgehoben werden. GoetheFind vernetzt alle Treffer mit der entsprechenden Replik in der Bühnenaufführung und der Hörbuchausgabe. Sobald der Nutzer auf die Mulitmediabuttons des Treffers drückt, startet im Browser ein Videoviewer oder eine Ausdioausgabe ab dieser Stelle.

Danksagung

Wir danken dem Deutschen Textarchiv für die gute Zusammenarbeit und die freundliche Verfügungsstellung der Editionsdaten von Goethes Faust (BBAW 2013). Der Staatsbibliothek zu Berlin – Preußischer Kulturbesitz und dem Freien Deutschen Hochstift, in Frankfurt danken wir für die Wiedergaberechte der Bilddigitalisate der Originalausgabe Goethes Faust.

Bibliographie

BBAW (2013): *Das DTA-Basisformat DTABf*. Berlin-Brandenburgische Akademie der Wissenschaften (BBAW) http://www.deutschestextarchiv.de/doku/basisformat [letzter Zugriff 09. Januar 2016].

CLARIN (o. J.): *CLARIN-D WebMAUS*. Automatic Segmentation of Speech. https://www.clarin.eu/movies/clarin-d-webmaus-automatic-segmentation-speech [letzter Zugriff 08. September 2015].

Haaf, Susanne / Geyken, Alexander / Wiegand, Frank (2015): "The DTA 'Base Format': A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources", in: *Journal of the Text Encoding Initiative* 8 https://jtei.revues.org/1114 [letzter Zugriff 09. Januar 2016].

Hadersbeck, Max / Bruder, Daniel / Capsamun, Roman / Conforti, Costanza / Eder, Elisabeth / Eichfeldt, Nora / Fink, Florian / Herteis, Simeon / Höps, Raphael / Lindinger, Matthias / Ling, Jennifer / Mittelhammer, Katharina / Schmidt, Katharina / Schweter, Stefan *FinderApp GoetheFind*. Centrum für Informations- und Sprachverarbeitung (CIS), Ludwig Maximilians Universität München http://goethefind.cis.uni-muenchen.de/ [18.02.2016].

Hadersbeck, Max / Pichler, Alois (eds.) (o. J.) *FinderApp WiTTFind*. Centrum für Informations- und Sprachverarbeitung (CIS), Ludwig Maximilians Universität München & Wittgenstein Archives, University of Bergen http://wittfind.cis.uni-muenchen.de/ [letzter Zugriff 09. Januar 2016].

Hadersbeck, Max, Pichler, Alois, Fink, Florian , Gjesdal, Øyvind Liland (2014): "Wittgenstein's Nachlass. WiTTFind and Wittgenstein advanced search tools (WAST)", in: *Digital Access to Textual Cultural Heritage (DaTeCH 2014)*, Madrid 91-96 http://dblp.uni-trier.de/db/conf/datech/datech2014.html#HadersbeckPFG14 [letzter Zugriff 09. Januar 2016].

Hadersbeck, Max / Pichler, Alois / Fink, Florian / Bruder, Daniel / Arends, Ina / Baiter, Johannes (2015): "Wittgensteins Nachlass. Erkenntnisse und Weiterentwicklung der FinderApp WiTTFind", in: *2. Tagung der Digital Humanities im deutschsprachigen Raum 23.-27.2.2015*, Graz.

Ploeger, Lieke (2014): "Open Humanities Awards round 2 – Winners announced", in: *DM2E. Digitised Manuscripts to Europeana* http://dm2e.eu/open-humanities-awards-round-2-winners-announced/ [letzter Zugriff 09. Januar 2016].

TEI (2015): "Stand-off Markup", in: *TEI Guidelines for Electronic Text Encoding and Interchange*, Version 2.9.1, 16.9 http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SA.html#SASO [letzter Zugriff 09. Januar 2016].