Workshop: Automatisierte Handschriftenerkennung mit der Transcription & Recognition Platform (TRP)

Hindergrund

Im Rahmen des Projekts tranScriptorium, das sich mit der automatisierten Erkennung historischer Handschriften beschäftigt, entwickelt die Projektgruppe „Digitalisierung und Elektronische Archivierung“ (DEA) am Institut für Germanistik der Universität Innsbruck eine Plattform (TRP) mit deren Hilfe handschriftliche Dokumente in neuartiger Weise erschlossen werden können. (Sánchez et al. 2013)

Wie die Vorträge bei der International Conference on Frontiers in Handwriting Recognition (2014) gezeigt haben, handelt es sich dabei um eine Technologie, die am Sprung zum Praxiseinsatz steht. Das größte Hindernis stellt jedoch die ungenügende Menge an Referenzdaten dar, die daher bei den eingesetzten statistischen Verfahren oftmals zu unbefriedigenden Ergebnissen führen. (Plötz & Fink 2009)

Die von uns im Rahmen es Workshops vorgestellte Plattform will diesem Mangel nun dadurch begegnen, dass sie Geisteswissenschaftlern die Möglichkeit bietet, die Transkription eines handschriftlichen Textes in einer besonders komfortablen und teilweise mit automatisierten Methoden unterstützten Art und Weise durchzuführen. Also besonderer Anreiz ist hierbei die enge Verbindung zwischen Text und Bild auf Block, Zeilen und Wortebene zu nennen, zum anderen die standardisierten Exportformate: TEI (Text Encoding Initiative) sowie PDF (Portable Document Format) zur lokalen Benutzung, aber auch METS (Metadata Encoding and Transmission Standard) für die Integration in Repositorien wie etwa FEDORA. (Mühlberger 2014)

Gleichzeitig können nun die von Geisteswissenschaftlern produzierten Transkriptionen auch für das Training von Handwritten Text Recognition (HTR) Maschinen genützt werden. Mithilfe der automatisierten Erkennung kann nicht nur die Transkription selbst unterstützt werden, sondern können auch noch nicht transkribierte, größere Mengen von Dokumenten automatisiert erkannt werden.

Die zu erzielenden Genauigkeiten hängen von vielerlei Faktoren ab. Erste Experimente zeigen jedoch, dass bei einem nicht zu komplexen Layout und gerader Linienführung durchaus Ergebnisse von unter 30% Word Error Rate zu erzielen sind. (Romero et al. 2013)

Ziele des Workshops

Die Teilnehmer des Workshops erhalten die Möglichkeit mit einer Betaversion der Plattform zu arbeiten und die automatisierte Erkennung an Beispieldokumenten durchzuführen. Die Software wird Anfang 2015 öffentlich von der Webseite des Projekts zum Download angeboten. Ab diesem Zeitpunkt haben die Teilnehmer folgende Möglichkeiten.

- Sich als Benutzer in der Plattform zu registrieren
- Eigene Bilddateien auf die Plattform hochzuladen
- Eine manuelle Transkription durchzuführen
- Eine automatisierte Block- und Zeilenerkennung durchzuführen
- Blöcke, Zeilen, Wortsegmente edieren

- Ausgewählte Beispieldokumente (Schriften des Jeremy Bentham aus dem Transcribe Bentham Projekt) automatisiert zu erkennen
- Die entsprechenden Dokumente in den angebotenen Standardformaten (TEI, METS, PAGE, PDF) zu exportieren

Ablauf des Workshops

Der Workshop wird aus drei Teilen bestehen:

1. Einführung in das Thema Handwritten Text Recognition (Vortragender: Joan Andreu Sanchez) – ca. 30‘
   Vorgestellt werden die grundlegenden Technologien und Tools, die der automatisierten Handschriftenerkennung zugrunde liegen. Joan Andreu Sanchez ist wissenschaftlicher Koordinator des EU Projekts tranScriptorium und Professor für Comuter Science an der Technischen Universität Valencia.

2. Vorstellung der Transcription & Recogniton Platform (Vortragender: Günter Mühlberger) – ca. 30‘
   Hier wird auf das grundlegende Konzept der Plattform eingegangen und die Idee einer digitalen Infrastruktur zur Erkennung von Handschriften im Detail erläutert. Günter Mühlberger leitet die Gruppe „Digitalisierung und elektronische Archivierung“ (DEA) am Institut für Germanistik der Universität Innsbruck und ist für das Arbeitspaket „Datenmanagement“ im oben genannten Projekt verantwortlich.

3. Einführung in das Tool „Transcribus“ (Vortragende: Sebastian Colutto und Philip Kahle) – ca. 30‘
   Das für den Geisteswissenschaftler wichtigste Interface zur HTR Technologie im Rahmen der Transcription & Recognition Platform ist das Tool „Transcribus“. Es ist mit JAVA und SWT programmiert und muss lokal installiert werden. Allerdings werden die Bilder und Daten mittels einer Remoteverbindung zum TRP Server geladen und gespeichert. Auf diese Weise kann ein sehr flüssiges Arbeiten mit einem „Rich Client“ erzielt werden. Sebastian Colutto und Philip Kahle sind seit mehreren Jahren Projektmitarbeiter in der DEA Gruppe und arbeiten seit knapp zwei Jahren intensiv an dem vorliegenden Prototypen.

4. Selbständiges Arbeiten mit der Plattform bzw. Transcribus - ca. 2,5h
   Die Teilnehmer sollen die Möglichkeiten und Grenzen der Technologie in allen Einzelheiten an ihrem PC ausprobieren können und werden dabei von den vier Vortragenden unterstützt.

Zielgruppen für den Workshops

1. Geisteswissenschaftler, die mit der wissenschaftlichen Edition handschriftlicher Texte befasst sind. Sie erhalten mit Transcribus bzw. TRP ein mächtiges Werkzeug, das sie in ihrer täglichen Arbeit unterstützen soll.
2. Archive und Bibliotheken die interessiert sind, ihre digitalisierten handschriftlichen Bestände für die Öffentlichkeit zu öffnen. In diesem Fall kann TRP benutzt werden, um ein Team von ehrenamtlichen Mitarbeitern koordinieren zu können.

Literatur

Mühlberger, G., 2014. Die automatisierte Volltexterkennung historischer Handschriften als gemeinsame Aufgabe von Archiven, Geistes- und Computerwissenschaftlern. Das Modell einer

zentralen Transkriptionsplattform als virtuelle Forschungsumgebung. In *Digitalisierung im Archiv. Neue Wege der Bereitstellung des Archivguts. Beiträge des 18. Archivwissenschaftlichen Kolloquiums am 26. und 27. November 2013*.

Plötz, T. & Fink, G. a., 2009. Markov models for offline handwriting recognition: a survey. *International Journal on Document Analysis and Recognition (IJDAR)*, 12(4), pp.269–298. Available at: http://link.springer.com/10.1007/s10032-009-0098-4 [Accessed August 20, 2013].

Romero, V.V. et al., 2013. The ESPOSALLES database: An ancient marriage license corpus for off-line handwriting recognition. *Pattern Recognition*, 46(6), pp.1658–1669. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0031320312005080.

Sánchez, J.A. et al., 2013. tranScriptorium: an European Project on Handwritten Text Recognition. *DocEng'13, September 2013, Florence, Italy*, pp.227–228.