

Workshop „Automatisierte Erkennung handschriftlicher Dokumente mit der Transcription & Recogniton Plattform (TRP)

Call Text für den Workshop

Hintergrund

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.

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.

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.

Gleichzeitig können nun die von Geisteswissenschaftlern produzierten Transkriptionen auch für das Training von 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.

Call für Teilnehmer

Der Workshop richtet sich an

1. Geisteswissenschaftler (Historiker, Literaturwissenschaftler, Theologen, Juristen, ...), die mit der Transkription und Edition historischer Dokumente befasst sind. Besonders interessiert sind wir an Teilnehmern aus laufenden Transkriptionsprojekten, um die Plattform an die Bedürfnisse der Benutzer anpassen zu können.
2. Archivare und Bibliothekare (Sondersammlungen), die entweder selbst mit der Transkription handschriftlicher Texte zu tun haben, oder aber daran denken, dies im Rahmen eines Crowd-Sourcing Projekts zu organisieren. Die vorgestellte Plattform verfügt über ein Rechte- und Rollenmanagement, sodass mehrere Personen an einem Dokument arbeiten können, bzw. Dokumente auch für die Öffentlichkeit zur Verfügung gestellt werden können.
3. Studenten der Geistes-, Archiv- oder Bibliothekswissenschaften mit einem besonderen Interesse an der Transkription historischer Handschriften.

Wir erwarten uns von den Teilnehmern eine kurze Beschreibung der von Ihnen bisher durchgeführten Transkriptionsprojekte bzw. der geplanten Vorhaben in diesem Bereich. Alle Teilnehmer sind darüber hinaus aufgefordert, eigene digitalisierte Dokumente zum Workshop

mitzunehmen, bzw. diese bereits im Vorfeld auf die Plattform hochzuladen. Entsprechende Zugangsdaten werden von uns nach der Zusage zur Teilnahme zugestellt.

Ablauf des Workshops

Der Workshop wird aus drei Teilen bestehen:

1. Einführung in das Thema Handwritten Text Recognition (Vortragender: Joan Andreu Sanchez) – ca. 30'
Vorgelegt 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 Computer Science an der Technischen Universität Valencia.
2. Vorstellung der Transcription & Recognition Platform (Vortragender: Günter Mühlberger) – ca. 30'
Hier wird auf das grundlegende Konzept der TRP 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 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 am eigenen PC ausprobieren können und werden dabei von den vier Vortragenden unterstützt.