Automatische Volltexterkennung mit Transkribus im Projekt OCR-BW

Grundlage einer Edition, ob nun gedruckt oder digital, bleibt weiterhin die Transkription des Textes vom Original in ein zur Weiterverarbeitung geeignetes Textformat. Für diesen Umwandlungsprozess gibt es mittlerweile technische Hilfsmittel, die den Editorinnen und Editoren die Arbeit wesentlich erleichtern können. In diesem Beitrag soll mit der Transkriptionsplattform Transkribus ein Werkzeug zur automatischen Handschriftenerkennung vorgestellt werden.

Das 2019 im Rahmen des Projekts OCR-BW als Service der Universitätsbibliotheken Tübingen und Mannheim neu eingerichtete Kompetenzzentrum „Volltexterkennung von handschriftlichen und gedruckten Werken“ unterstützt Wissenschaftlerinnen und Wissenschaftler sowie Bibliotheken, Archive und andere Institutionen in Baden-Württemberg bei der Anwendung von automatischer Texterkennungs- und Transkriptionssoftware. Bei einer Volltexterkennung werden textliche Bildinhalte in digitale Textformate übersetzt. Erkannte Texte können durchsucht, kopiert, bearbeitet und für eine Extraktion von Forschungsdaten verwendet werden.

Die UB Tübingen evaluiert Transkribus mit seinen Tools unter anderem für die Layoutanalyse und HTR (Handwritten Text Recognition). Es wird untersucht, inwieweit bzw. für welche Text- und Schriftarten Transkribus nutzbar ist. Angestrebt wird nach Vorgaben der DFG eine CER (Character Error Rate) von unter 5%. Hierfür werden Ground-Truth-Daten (korrekte Transkriptionen anhand derer die Software trainiert wird) und Texterkennungsmodelle für verschiedene forschungsrelevante Korpora der Handschriftenabteilung und des Universitätsarchivs der UB Tübingen erstellt. Im bisherigen Projektverlauf hat sich jedoch auch gezeigt, dass für bestimmte Materialgruppen generische Texterkennungsmodelle bereits mit sehr gutem Ergebnis eingesetzt werden können. Ziel ist also, auf Basis der Projektergebnisse Handlungsempfehlungen für unterschiedliche Materialgruppen zu etablieren, um den Zugang zu diesen Texten zu erleichtern und neue wissenschaftliche Fragestellungen und Auswertungsmöglichkeiten zu schaffen. Anhand der bisher bearbeiteten Dokumente – vom mittelalterlichen Gebetbuch bis zum paläontologischen Expeditionstagebuch des 20. Jahrhunderts – soll beispielhaft gezeigt werden, welche Möglichkeiten die automatische Texterkennung mit Transkribus als Grundlage für eine weitere Verarbeitung des Textes bietet.