Digitalisierung und Volltexterkennung der ehemals Reichenauer Inkunabeln

Das Projekt

Anlässlich des Jubiläums zum 1.300-jährigen Bestehen des Benediktinerklosters auf der Reichenau digitalisiert die Badische Landesbibliothek die ca. 240 Titel umfassende Inkunabelsammlung aus der ehemaligen Klosterbibliothek und erschließt diese in maschinenlesbarer Form. Ermöglicht wird dieses Projekt, das nicht nur von hoher landesgeschichtlicher Bedeutung ist, sondern mit einem Umfang von ca. 70.000 Seiten auch quantitativ neue Maßstäbe hinsichtlich der computergestützten Volltexterkennung von Wiegendrucken setzt, durch die Förderung der Stiftung Kulturgut Baden-Württemberg. Es ist angesiedelt in den Abteilungen Regionalia und Sammlungen.

Im Rahmen der Arbeiten der letzten Jahre mit und an OCR wurde der Bereich der Inkunabeln bislang nur in klein angelegten Pilotprojekten angegangen. Auch das groß angelegte Projekt OCR-D widmet sich in erster Linie der Entwicklung von Verfahren zur Volltexterkennung der im deutschen Sprachraum erschienenen Drucke des 16. bis 18. Jahrhunderts. Im Vergleich zu späteren Drucken zeichnen sich Inkunabeln jedoch durch eine größere Variabilität der Schrifttypen und Layoutkonventionen, eine dichtere Verwendung des aus der Manuskripttradition stammenden Abbreviaturensystems, häufiger handschriftlich nachgetragene Rubrizierungen und teils schlechtere Erhaltungszustände aus. Diese Eigenschaften machen nicht nur eine Bearbeitung mit herkömmlicher OCR-Software unmöglich, sie führen auch dazu, dass auf spätere Drucke der frühen Neuzeit zugeschnittene Texterkennungsmodelle für Inkunabeln nur bedingt anwendbar sind.

Daher entwickelt die Badische Landesbibliothek im Reichenau-Projekt unter Verwendung der Software Transkribus eine Serie von Texterkennungsmodellen, die eine zuverlässige Volltexterschließung lateinisch- und deutschsprachiger Inkunabeln ermöglichen sollen. Das zu Grunde liegende Korpus und den beispielhaften Anwendungsfall stellt die genannte Sammlung ehemals Reichenauer Inkunabeln dar.

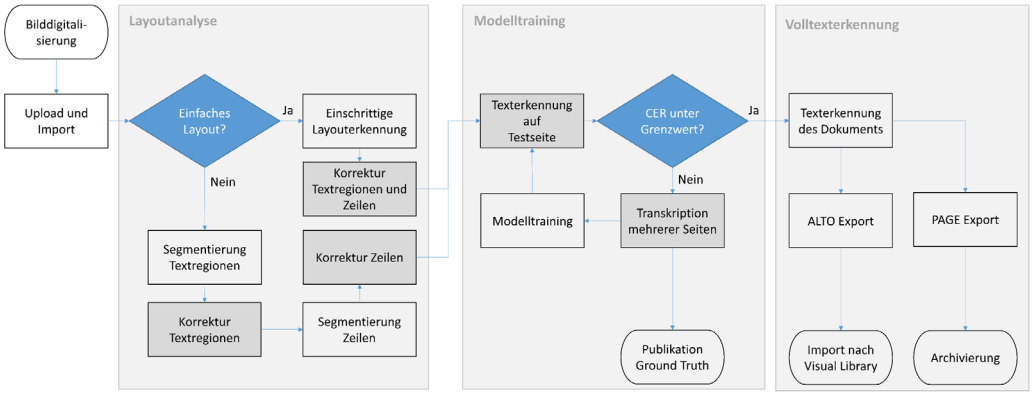
Die Software Transkribus

Die für das Projekt verwendete Software Transkribus bietet einen relativ niedrigschwelligen Zugang zu der Erstellung von Texterkennungsmodellen, die unter Verwendung modernster Methoden maschinellen Lernens auf spezifisches gedrucktes oder handschriftliches Textmaterial zugeschnitten sind. Während die von Transkribus verwendeten Software-Komponenten großenteils frei verfügbar sind, liegt der Mehrwert dieser Plattform in deren nahtloser Integration in zwei Benutzeroberflächen (Web App und Expert Client) und dem Betrieb der zugehörigen Serverinfrastruktur. Da sämtliche Dokumente auf den Servern der READ-COOP-Kooperative in Innsbruck verarbeitet werden, entsteht im Vergleich zu konkurrierenden Lösungen wie eScriptorium oder OCR4all kein lokaler Administrationsaufwand. Erkauft wird diese Einfachheit mit der Bindung an die Transkribus-Plattform, die zwar einen Export der erzeugten Volltexte ermöglicht, den Download der Texterkennungsmodelle selbst aber aus wirtschaftlichen Gründen verwehrt – diese können

nur innerhalb der Plattform betrieben und ausschließlich mit anderen Nutzerinnen und Nutzern von Transkribus geteilt werden. Abzuwarten bleibt, wann und wie eine angekündigte Revision des Gebührenmodells für die Nutzung von Transkribus umgesetzt werden wird; diese könnte insbesondere die Verarbeitung von gedrucktem Material deutlich verteuern.

Der Arbeitsprozess

Um das Volumen des Projektkorpus von ca. 70.000 Seiten bewältigen zu können, folgen wir einem dreischrittigen Arbeitsprozess:



Der Volltexterkennungs-Workflow des Projekts. Prozessschritte mit hohem manuellem Arbeitsaufwand sind grau hinterlegt.

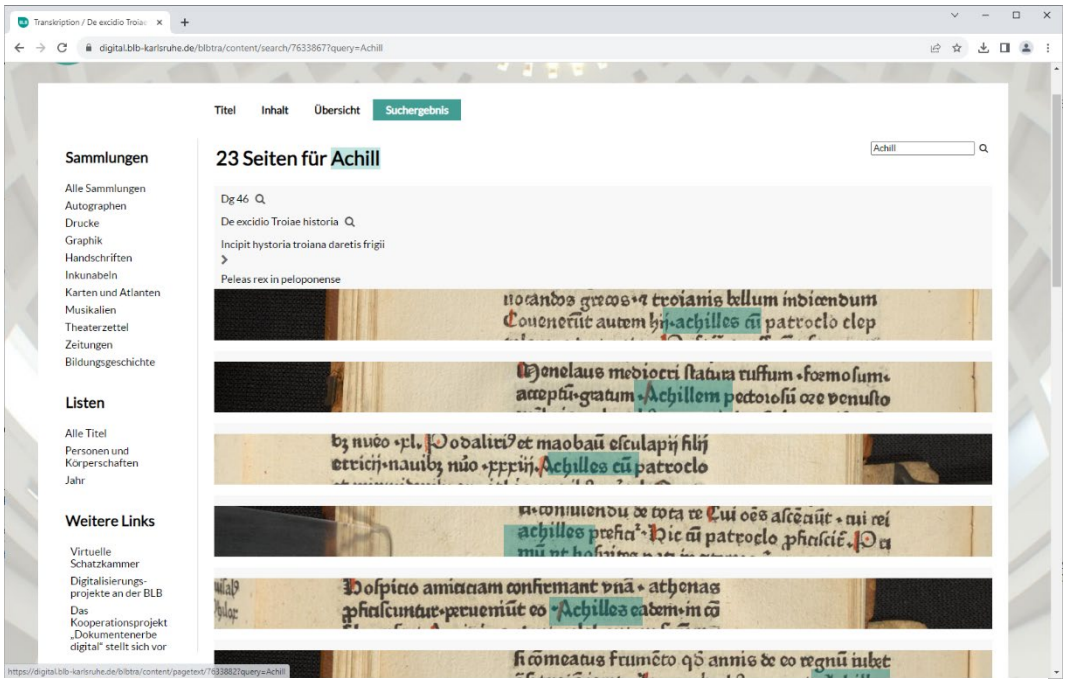
In Abstimmung mit der Bestandserhaltung digitalisiert die Digitalisierungswerkstatt der Badischen Landesbibliothek die Inkunabeln und stellt ihre Digitalisate dem Projekt als hochauflösende TIF-Dateien bereit. Über eine FTP-Schnittstelle importieren wir diese auf die Transkribus-Server, sodass wir mit ihnen innerhalb der Softwareumgebung arbeiten können.

Da die von Transkribus unterstützten Texterkennungsmodelle jeweils den Bildausschnitt einer einzelnen Zeile als Input erwarten, müssen die Bilddateien im Zuge einer vorbereitenden Layoutanalyse in Textregionen und Zeilen segmentiert werden. Dieser Vorgang ist relativ fehleranfällig, für die Qualität der anschließenden Texterkennung aber von hoher Bedeutung. Aufgrund der besonderen Beschaffenheit von Inkunabeln müssen die Ergebnisse der automatisierten Layoutanalyse daher zumindest stichprobenartig gesichtet und ggf. manuell korrigiert werden. Je nach Komplexität des Materials können Textregionen und Zeilen in einem oder in zwei Verarbeitungsschritten erkannt werden.

Anschließend bestimmen wir anhand einer exemplarischen Seite aus dem Dokument, ob unser Texterkennungsmodell dessen Schriftbild bereits hinreichend zuverlässig erkennt oder ob eine neue Modellversion trainiert werden muss. Hierzu korrigieren wir das automatisch erstellte Transkript manuell und lassen die zugehörige Zeichenfehlerrate (Character Error Rate, CER) bestimmen. Liegt diese oberhalb des angestrebten Grenzwerts von 0,6%, erzeugen wir zusätzliches Trainingsmaterial (sog. Grundwahrheit, Ground Truth), indem wir weitere Transkripte manuell korrigieren. Diese werden anschließend in

das Training der nächsten Modelliteration miteinbezogen. Zumeist stellt sich der erwünschte Lerneffekt schon mit wenigen Seiten zusätzlichen Trainingsmaterials ein.

Ist die Zeichenfehlerrate der Stichprobe akzeptabel, wenden wir das Texterkennungsmodell auf den restlichen Teil des Dokuments an. Eine abschließende manuelle Korrektur der Transkripte erfolgt an dieser Stelle nicht mehr. Der Export im ALTO XML-Format erlaubt es uns, die Volltexte anschließend in unsere Digitalisierungsplattform Visual Library zu importieren und den Nutzerinnen und Nutzern unserer digitalen Bibliothek in verschiedenen Formaten zur Verfügung zu stellen (Seite-an-Seite-Ansicht von Digitalisat und Volltext, Volltextsuche, Download von PDF-Dateien mit hinterlegtem Volltext). Das mit zahlreichen Texterkennungsprogrammen kompatible PAGE-XML-Format nutzen wir für die Archivierung des Projektertrags sowie für die angestrebte Publikation der im Rahmen des Projektes erzeugten Ground-Truth-Daten über die Plattformen Zenodo und RegionaliaOpen.



Volltextsuche in Dares Phrygius, De excidio Troiae historia, Köln: Johann Schilling, bis 1472 (BLB, Dg 46).

Erkenntnisse aus dem Projekt

Im Gegensatz zu anderen Projekten, die druckereispezifische Texterkennungsmodelle für Inkunabeln trainierten, arbeiten wir erfolgreich mit allgemeineren, lediglich sprachspezifischen Modellen (lateinisch, deutsch oder bilingual lateinisch-deutsch). Selbst die Einbeziehung von ca. 10% Antiquamaterial in unser lateinisches, großenteils auf Frakturtexten basierendes Texterkennungsmodell führte zu keiner Verschlechterung in der Qualität der Texterkennung.

Da Transkribus nicht Zeichen für Zeichen, sondern zeilen- und wortbasiert arbeitet, können Texterkennungsmodelle darauf trainiert werden, Abbreviaturen unter einer gewissen Berücksichtigung des (Zeilen-)Kontextes auszuschreiben. Für Abkürzungen, die eher

schwach verkürzen, häufig auftreten und relativ kontextunabhängig sind, funktionierte dieses Verfahren in unseren anfänglichen Tests recht gut. Da aber große Teile unseres Materials eine hohe Dichte komplexer Abbreviaturen aufweisen und wir keine Möglichkeit haben, 70.000 Seiten automatisiert erstellter Transkripte Korrektur zu lesen, entschieden wir uns dafür, einen möglichst zuverlässigen und interpretationsarmen Text zu erzeugen, der die Abbreviaturen in Unicode abbildet.

Auf diese Weise erreichen wir hervorragende Zeichenfehlerraten von 0,6% auf unserem ehemals Reichenauer Material (Validation Set) und von 0,8% auf einer Stichprobe des übrigen Inkunabelbestands der Badischen Landesbibliothek (Test Set).

1-1 # ¶ Primus.
2-1 # Altiſſimi doctoris Antonij andree ſeraphici
2-2 # ordinis minorum queſtiones ſubtiliſſime ſup
2-3 # duodecī libros methaphiſice Aꝛ. felicit̓ īcipiūt.
2-4 # ~~Arum~~ Irum celi cir
2-5 # cuiui ſola. eccleſiaſtici. xxiiij.
2-6 # Secunduꝫ doctrinam Aꝛ. 7
2-7 # ~~Ieuꝫ~~ euꝫ cōir ſequētiuꝫ ſcīa ~~mecha.~~ metha.
2-8 # que ~~theologia~~ theologya phoꝛ 7 ſapien¬
2-9 # tia noīatur: verſat circa totū ens. 7 ſignāt circa
2-10 # ſbſt̄atias ſepatas: vt circa nobiliores ptes ſui
2-11 # ſubi prī. Et iō qꝛ ē circa nobiliſſima entia: nobi
2-12 # liſſima ſcīa ē īter oēs ſcīas naturaĺr adinuētas.
2-13 # Nobilitas .n. ſcīaꝛ ex nobilitate oritur ſubie¬
2-14 # ctorū. ex primo de aīa. Iō in pſona huius ſcīe
2-15 # ꝯgrue pōt dici Girum celi. 7c̄. ¶ Ubi deſcribit
2-16 # eius dignitas admirāda q̄tum ad quattuor .ſ.
2-17 # q̄tum ad amplexū ābitōis magnifice. Influxū
2-18 # correctōis autētice. Actū īqſitionis amplifice.
2-19 # Et ꝯdum p̄latōis mirifice. ¶ Primū ꝑbat ip̄i⁹
2-20 # giri ꝯtinētia generalis: cū p̄mittit girū ad ~~inſta¬~~ inſtar
2-21 # enim cuiuſlibet giri vl̓ circuli q ē figura ~~capaci¬~~ capaciſ
2-22 # ſima ſm geometricos ſcīa metha. oīa ābit. ~~Na~~ Nā
2-23 # vt dicitur p⁹ huius ī ꝑlogo ſapiētis .i. methaph̓.
2-24 # ē oīa ſcire vt ꝯtingit. Scīa .n. metha. girat oīa
2-25 # entia ſiue ſint īmobilia 7 īcorruptibilia: ſiue mo

Ausschnitt aus einem automatisch erzeugten Volltext mit einer Zeichenfehlerrate von 0,62% (Antonius Andreas, Quaestiones super XII libros Metaphysicae Aristotelis, Venedig: Antonius de Strata 1481, BLB Pb 27a, Bl. a2a).

Trotz der hohen Qualität, die unsere Texterkennungsmodelle auszeichnet, verhindern Schwächen im Bereich der Layouterkennung, dass das Ziel einer vollautomatischen, wirklich „unbeaufsichtigten“ Texterkennung auf Inkunabeln in naher Zukunft erreichbar scheint: Während automatische Segmentierungsmethoden mit einfachen Seitenlayouts gut zurechtkommen, erfordern gedruckte Marginalien, mittelalterliche Tabellen und Diagramme oder auch verzerrte Zeilen am Seitenfalz stets eine Abwägung zwischen hohem

manuellem Arbeitsaufwand und deutlichen Abstrichen im Ergebnis. Während technische Fortschritte in einzelnen Punkten durchaus erkennbar sind – so ermöglichen etwa die in der Testphase befindlichen Field-Modelle eine automatische Erkennung der weit verbreiteten Klammerglossen-Layouts –, ist aus unserer Sicht keine grundsätzliche Änderung dieser Problemkonstellation absehbar.

Ausblick

Das Vorliegen qualitativ hochwertiger Volltexte eröffnet im bibliothekarischen Bereich zahlreiche Möglichkeiten zur besseren Aufbereitung und informativen Anreicherung historischer Drucke: Die großen Fortschritte im Bereich der maschinellen Sprachverarbeitung lassen hier zum Beispiel an das syntax- und kontextsensible Ausschreiben von Abbreviaturen denken, an die Möglichkeit, Personen-, Orts- und Objektnamen automatisch erkennen zu lassen (Named Entity Recognition) und mit Normdatenbanken zu verknüpfen, oder auch Übersetzungen fremdsprachlicher Texte auf Nutzerwunsch zu erzeugen. Aber auch für die wissenschaftliche Nutzung eröffnen Volltexte neue Perspektiven, beispielsweise im softwaregestützten Vergleich des Textbestands verschiedener Ausgaben und Auflagen eines Werkes. Für uns stellt die Volltexterschließung historischer Bestände somit einen notwendigen ersten Schritt dar, um diese für neue und zukunftswiesende Nutzungsmöglichkeiten zu erschließen. Sie bleibt jedoch vorerst arbeitsaufwändig, eine Vollautomatisierung ist bisher ohne erhebliche Qualitätsabstriche nicht möglich.

Katharina Ost, Badische Landesbibliothek, Karlsruhe