idw - Informationsdienst Wissenschaft
Nachrichten, Termine, Experten

Press release

Julius-Maximilians-Universität Würzburg
Robert Emmerich

01/19/2022
http://idw-online.de/en/news786866

Research projects
Cultural sciences, History / archaeology, Information technology, Language / literature
transregional, national

Historische Schriften digital erkennen

Die Texterkennungssoftware OCR4all kommt bei historischen Drucken mit sehr gutem Erfolg zum Einsatz. Jetzt wird sie auf alte Handschriften trainiert.

Heutige Standardschriften wie Calibri oder Times New Roman einzulesen, ist für moderne Texterkennungssoftware, kurz OCR, kein Problem. Schwieriger wird es bei historischen Drucken. Denn je weiter man in die Geschichte zurückblickt, desto variantenreicher werden die Schriften – bis hinein in eine Zeit, in der jeder Drucker seine eigenen Schriftsets schnitzte.

Darum gibt es eine gute Nachricht für alle, die mit derartigem historischem Material arbeiten: Das Programm OCR4all ist eine Texterkennungssoftware, die historische Druckschriften erkennt und in computerlesbaren Text umwandelt. Um es zu bedienen, sind keinerlei Programmierkenntnisse nötig.

OCR4all steht seit 2019 im Web weltweit kostenlos zur Verfügung. Rund 5.000 Mal wurde es inzwischen heruntergeladen; ein vergleichbares Angebot im Open-Source-Bereich gab es bis dato nicht. Entwickelt wurde das Tool von einem interdisziplinären Team um Dr. Christian Reul, Leiter der Digitalisierungseinheit am Zentrum für Philologie und Digitalität „Kallimachos“ (ZPD) der Julius-Maximilians-Universität (JMU).

OCR4all ging aus dem vom Bundesforschungsministerium geförderten Kallimachos-Verbundprojekt der JMU hervor. Dieses Projekt schlug Brücken zwischen den Geisteswissenschaften, der Informatik und den Digital Humanities. Anfangs ging es bei OCR4all darum, im Teilprojekt Narragonien digital Sebastian Brants Narrenschiff digital aufzubereiten, eine Moralsatire aus dem 15. Jahrhundert.

Werkspezifische Modelle sind sehr genau

Seither ist das Projekt deutlich gewachsen und auch im Ausland in Fachkreisen bekannt. „Das Schöne an Open-Source-Projekten: Es ist immer ein Geben und Nehmen“, sagt Reul. Damit die Software bestimmte Schrifttypen später möglichst genau erkennt, werden Modelle trainiert. Dafür braucht es möglichst viel Trainingsmaterial, bestehend aus Zeilenbildern und der korrekten Transkription des darauf zu sehenden Texts, und das wird häufig von den Software-Nutzerinnen und -Nutzern selbst zur Verfügung gestellt.

Diese Form der Kooperation trägt Früchte, wie Reul erklärt: So lassen sich bei so genannten werkspezifischen Modellen inzwischen sehr genaue Erkennungsergebnisse erzielen, selbst auf den ältesten existierenden Drucken aus der Inkunabelzeit (vor 1500). Dies sind Modelle, die wie im Falle des Narrenschiffs speziell für die Erkennung einer Drucktype trainiert werden.

Förderung durch die Vogel Stiftung

Das ZPD arbeitet nun verstärkt daran, gemischte Modelle weiterzuentwickeln, die im Idealfall auf möglichst viele Drucktypen angewendet werden können. Während es zum Beispiel für deutschsprachige Frakturschriften des 19. Jahrhunderts bereits sehr gute Modelle gab, fehlte es bislang an einem noch breiter aufgestellten Modell, das guten Gewissens auf Drucke aus mehreren Jahrhunderten angewendet werden kann. Dafür brauchte es laut Reul vor allem weitere Trainingsdaten.

Entsprechend glücklich war er deshalb über eine Förderung durch die Vogel Stiftung Dr. Eckernkamp (Würzburg): „Vor allem bei historischen Frakturschriften gab es Lücken in den Trainingsdaten, die wir durch die Förderung gezielt schließen konnten", sagt der Informatiker.

Auszeichnung mit Best Paper Award

Bei der Fachkonferenz HIP'21 (6th International Workshop on Historical Document Imaging and Processing) im September 2021 in Lausanne (Schweiz) präsentierte Reul erstmals eine Publikation zu einem gemischten Modell, das lateinische Schrift aus der Zeit von 1450 bis 1900 abdeckt.

„Wir waren seinerzeit bei einer Zeichengenauigkeit von mehr als 98 Prozent gelandet, das übertraf den bisherigen State-of-the-Art deutlich", sagt der JMU-Informatiker. Kaum erstaunlich also, dass die Veröffentlichung von der HIP-Konferenz mit dem Best Paper Award ausgezeichnet wurde.

350.000 Euro von der DFG

Als Meilenstein bezeichnet Reul zudem das im Juli 2021 von der Deutschen Forschungsgemeinschaft (DFG) genehmigte und mit 350.000 Euro geförderte Zwei-Jahres-Projekt OCR4all-libraries. „Wir verheiraten nun OCR4all mit OCR-D", freut er sich.

Das Hauptziel des DFG-geförderten OCR-D-Projekts ist die konzeptionelle und technische Vorbereitung der Volltexttransformation der im deutschen Sprachraum erschienenen Drucke des 16. bis 18. Jahrhunderts. Dazu wird die automatische Volltexterkennung in einzelne Prozessschritte zerlegt, die dann jeweils mit unterschiedlichen Werkzeugen bearbeiten werden können. Dies zielt darauf ab, optimale Workflows für die zu prozessierenden alten Drucke zu erstellen und damit wissenschaftlich verwertbare Volltexte zu generieren.

Ein Zusatznutzen der Software aus Würzburg im Zuge der Volltexterkennung der historischen Sammlung: OCR4all ermöglicht die Anwendung durch technisch weniger versierte Nutzenden und dient weiterhin auch erfahreneren Nutzenden als Handwerkszeug, um den Workflow zu analysieren und zu optimieren.

Reul hofft im Zuge des Projekts OCR4all-libraries auf eine umfassende Weiterentwicklung der Software, speziell durch die stark wachsende Anzahl der verfügbaren Werkzeuge. Zusammenarbeiten wird das ZPD dabei mit dem Leibniz-Institut für Bildungsmedien | Georg-Eckert-Institut in Braunschweig und dem JMU-Lehrstuhl für Mensch-Computer-Systeme.

Historische Handschriften: eine Herausforderung

Texterkennungssoftware für alte Drucke ist das eine. Doch wie steht es um historische Handschriften?

„Vom Prinzip her ist die Herangehensweise ähnlich, aber wegen der Unregelmäßigkeit der Schriften meist deutlich anspruchsvoller", sagt Reul. Außerdem können Handschriften erheblich älter sein als Drucke, decken somit eine noch größere Zeitspanne ab und sind häufiger schlecht erhalten.

Kein Grund für das ZPD, sich nicht auch dieser Herausforderung zu stellen. „Der Bedarf bei Handschriften ist riesig – hier findet man wie gedruckt wirkende Buchschriften bis hin zu Texten, die nahezu unlesbar sind.“, weiß Reul.

Angesichts dieser Herausforderung bleibt er gelassen: „Wir brauchen jetzt erstmal viel Training für eine solide Grundlage.“ Eine erste Kooperation kam im Frühjahr 2021 zustande mit Dr. Stefan Tomasek vom JMU-Lehrstuhl für deutsche Philologie, ältere Abteilung: Er stellte dem ZPD im Zuge seiner Neuedition der Kindheit Jesu Konrads von Fußesbrunnen Daten für das Modelltraining zur Verfügung. Seitdem wird in Kooperation zwischen dem ZPD und dem Lehrstuhl der Bestand an Trainingsdaten und somit das Modell stetig weiterentwickelt. Auf mittelalterlichen Handschriften konnten dadurch bereits hervorragende Ergebnisse erzielt werden. Erste Modelle sollen noch in den kommenden Wochen online frei zur Verfügung gestellt und das zugehörige Paper noch im Januar 2022 eingereicht werden. Ein gemeinsamer DFG-Antrag ist ebenfalls in Vorbereitung.

Auch andere Forschende greifen inzwischen bei ihren Drittmittelprojekten vermehrt auf das Knowhow des Würzburger ZPD und auf OCR4all zurück. So befinden sich, neben dem bereits laufenden DFG Projekt Camerarius digital, zahlreiche Projektanträge in Vorbereitung, sowohl für die Erkennung von Handschriften als auch von Drucken.

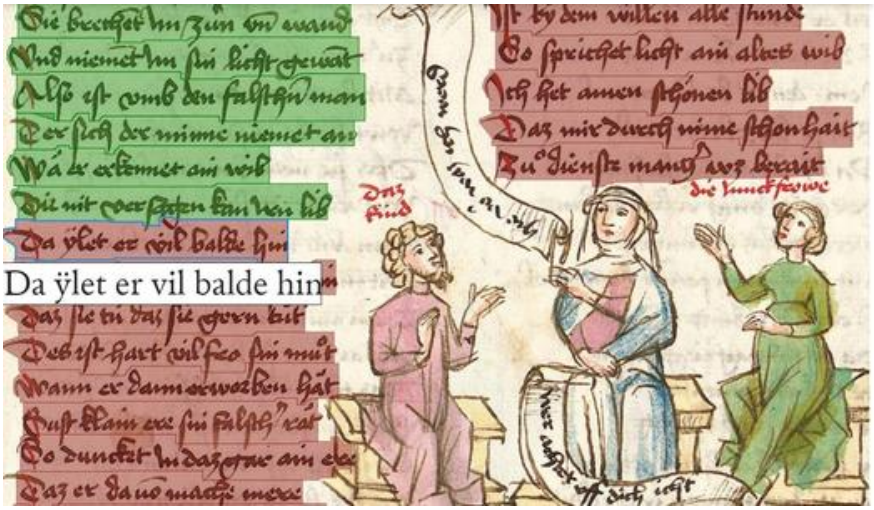
Neubau für das Zentrum für Philologie und Digitalität

Das Zentrum für Philologie und Digitalität „Kallimachos“ (ZPD) ist eine zentrale wissenschaftliche Einrichtung der Universität Würzburg. Es verfolgt seit seiner Gründung 2019 den Zweck, die geisteswissenschaftliche Forschung im digitalen Zeitalter bestmöglich zu unterstützen und weiterzuentwickeln.

Das ZPD erhält auf dem Campus Nord einen dreigeschossigen Forschungsneubau mit 2.500 Quadratmetern Nutzfläche. Voraussichtlich Ende 2022 ist er bezugsfertig. Dann sollen möglichst viele Forschungsprojekte einziehen, um Wissen und Kompetenzen aus Geistes-, Kultur- und Humanwissenschaften und Informatik auch räumlich zu bündeln.

Das ZPD ist dennoch bereits jetzt voll einsatzfähig und steht für Kooperationen zur Verfügung. Weitere Schwerpunkte, neben der maschinellen Texterkennung auf historischen Drucken und Handschriften, bilden die kollaborative Erstellung digitaler Editionen in einer virtuellen Forschungsumgebung sowie die Modellierung und Realisierung semantischer Datenbanken.

contact for scientific information:
Dr. Christian Reul, Zentrum für Philologie und Digitalität, Universität Würzburg, T +49 931 31-80722, christian.reul@uni-wuerzburg.de



Wå er erkennet ain wib
Wå er erkennet ain wib
Die nit verſagen kan Irn lib
Die nit verſagen kan Irn lib
Da ylet er vil balde hin
Da ylet er vil balde hin
Vnd wirbet vaſt daz iſt ſin ſin
Vnd wirbet vaſt daz iſt ſin ſin

Der Ausgangstext einer historischen Handschrift kann in verschiedenen Ansichten der Transkription in computerlesbaren Text zeilengenau gegenübergestellt und bei Bedarf korrigiert werden. Das ist nur eine der zahlreichen OCR4all-Funktionen.
Christian Reul
Universität Würzburg