KI-GESTÜTZTE TEXTERKENNUNG (OCR/HTR) BEI „KLEINEN“ SPRACHEN ZENTRAL- UND OSTEUROPAS

Wie lassen sich historische handgeschriebene Texte einfacher entziffern, noch dazu solche, die in „kleinen“, zur Zeit ihrer Entstehung nur unzureichend standardisierten Sprachen verfasst wurden? Dieser Frage gingen die aus so unterschiedlichen Disziplinen wie der Geschichtswissenschaft, der Linguistik, der Informatik und den Bibliothekswissenschaften stammenden Teilnehmenden des Workshops „KI-gestützte Texterkennung (OCR/HTR) bei ‚kleinen‘ Sprachen Zentral- und Osteuropas“ nach. Er fand am 14. September 2023 am Collegium Carolinum im Rahmen des von der DFG geförderten Projektes „Die Entdeckung der Muttersprache oder wie man spricht, so schreibt man? Normierungsstrategien ‚kleiner‘ Sprachen in Europa: Das Okzitanische, Jiddische und Belarusische“ statt. Geboten wurde ein instruktiver Blick in die Gegenwart und Zukunft der automatisierten Texterkennung.

Nach einer Begrüßung durch die beiden Organisatorinnen der Veranstaltung, Dr. Martina Niedhammer, wissenschaftliche Mitarbeiterin am Collegium Carolinum (München), und Prof. Dr. Jana Osterkamp von der Universität Augsburg startete ein „Rundgang“ durch die an verschiedenen Häusern im deutschsprachigen Raum derzeit praktizierten Formen automatisierter Texterkennung. Den Auftakt bildete eine Keynote von Michael Moser (Wien), der die Möglichkeiten und Grenzen herkömmlicher Texterkennungsprogramme am Beispiel des Ruthenischen, wie es im habsburgischen Galizien vom Ende des 18. Jahrhunderts bis 1918 verwendet wurde, aufzeigte. Eine wesentliche Schwierigkeit liege, so Moser, in den großen Abweichun-

gen vom heutigen Standardukrainischen, die in der Vergangenheit immer wieder zu falschen bibliothekarischen Erfassungen von Titeln geführt habe und daher die Sichtbarkeit auch großer „kleiner" Sprachen, wie des Ukrainischen, noch immer einschränke. Bei entsprechendem Training könnten KI-basierte Programme helfen, die Auffindbarkeit und damit die potentielle Rezeption solcher Texte zu erhöhen. Diese Perspektive ergänzte und erweiterte Annette Höslinger-Finck, die ebenfalls aus Wien angereist war und über ein 2022 gestartetes Projekt zur Neuerfassung der Ruthenica-Sammlung an der Österreichischen Nationalbibliothek berichtete. Ziel sei es dabei, die Visibilität der aufgrund langjähriger bibliothekarischer Standards häufig nicht als ukrainisch, sondern lediglich als „slawisch" deklarierten ruthenischen Bestände aus dem historischen Kronland Galizien im Sinne von Mosers Eröffnungsvortrag zu erhöhen.

Im folgenden Panel stand das Sorbische im Mittelpunkt; außerdem kam die Informatik ins Spiel. Wito Böhmak von der Sorbischen Zentralbibliothek in Bautzen und Kay-Michael Würzner von der Sächsischen Landes-, Staats- und Universitätsbibliothek in Dresden erörterten, wie quelloffene Workflows zur automatischen Texterkennung für ressourcenarme Sprachen am Beispiel des Obersorbischen geschaffen werden konnten. Bereits die scheinbar so simple Digitalisierung eines Buches sollte sich dabei als komplex erweisen, da ein Programm zur automatischen Texterkennung (OCR) eines umfangreichen Trainings sowie einer ausführlichen Ergebniskontrolle bedürfe. Nicht nur vor diesem Hintergrund warben beide Referenten für größere Aufgeschlossenheit gegenüber Open-Source-Lösungen, stelle sich doch bei diesen, anders als bei kommerziellen OCRs, nicht die Frage nach dem Ort der Datenhoheit sowie der generellen, auch finanziellen, Abhängigkeit.

Am frühen Nachmittag nahmen die Teilnehmenden das Format des „Workshops" wörtlich und beschäftigten sich in einer Hands-on-session mit der unmittelbaren Erprobung des bisher Gehörten, das heißt, sie versuchten, angeleitet von Niklas Platzer (Florenz) und Florian Langhanki (Würzburg), eigene Modelle zu trainieren und einen kleinen Workflow zu erstellen. Grundlage war die an der Universität Würzburg entwickelte Open-Source-Software OCR4all, in die Florian Langhanki einführte. Einen Blick in eine teilweise vergleichbare kommerzielle Lösung gewährte eine Kurzpräsentation von Niklas Platzer, der Transkribus vorstellte, mit dem er im Rahmen eines Projekts am Collegium Carolinum zur digitalen Erfassung sudetendeutscher, tschechischer und slowakischer Exilzeitschriften gearbeitet hat.

Den zweiten Teil des Nachmittags füllten drei Kurzvorträge, die aus der forschungspraktischen Arbeit mit drei „kleinen" Diasporasprachen im östlichen Europa stammten: Armenisch, Armeno-Kipčak und Jiddisch. Arpine Maniero (München), Jürgen Heyde (Leipzig) und Daria Vakhrushova (Düsseldorf/München) reflektierten über den Stand der Digitalisierung und benannten potentielle Schwierigkeiten für eine automatisierte Handschriftenerkennung (HTR) in „ihren" Sprachen. Während es im Armenischen vor allem morphologische Eigenheiten sind, die auch aus dem historischen Nebeneinander zweier Standards (West- und Ostarmenisch) resultieren, sind es in der jiddischen Kursive etliche Buchstaben, die in individuellen Handschriften große Ähnlichkeit aufweisen und so Hindernisse für HTR darstellen können, da sie sich nur über den Kontext – und somit nicht ohne

Weiteres automatisiert – erschließen lassen. Relativ gut bestellt ist es dagegen um die materiellen Grundlagen, insbesondere für das Training von OCR: So liegen zahlreiche jiddische und alle armenischen Zeitschriften des 19. und frühen 20. Jahrhunderts in digitaler Form vor, und auch für das einst in L'viv/Lemberg gesprochene Armeno-Kipčak ändert sich die Situation gerade, wenn auch aus einem traurigen Anlass: Infolge des brutalen Angriffskrieg, den Russland seit dem Februar 2022 gegen die Ukraine führt, hat die Digitalisierung entsprechender Dokumente erheblich an Fahrt aufgenommen, um das kulturelle Erbe vor der Zerstörung zu schützen.

Abschließend präsentierte Aleksej Tikhonov (Freiburg i. Breisgau/Zürich) mehrere Programme für HTR, an deren Entwicklung er gemeinsam mit Kolleginnen und Kollegen arbeitete, darunter Anwendungen für slawische Sprachen in nicht-lateinischen Alphabeten (Glagolitisch und vormodernes Kyrillisch) sowie das Jiddische. Als Partner fungiert dabei Transkribus und damit ein kommerzieller Anbieter, wobei das Projekt explizit auch eine breitere Öffentlichkeit und deren Wünsche an Texterkennung als „Alltagstool", etwa als Hilfe bei der Entzifferung alter Dokumente in Familienbesitz, einbezieht. Dies schlägt sich in der Zusammenstellung der Trainingsdaten, aber auch in der Auswahl derjenigen Sprachen und Schriftformen nieder, für die Modelle trainiert werden sollen – so etwa jüngst auch für Stenographie.

In der darauffolgenden Abschlussdiskussion kamen nochmals zentrale Diskussionspunkte des Workshops zur Sprache, so die Frage nach der Wahl eines Open-Source Programmes oder eines kommerziellen Produktes. Deutlich wurde dabei die hohe Aktualität des Themas, die ganz konkrete Bedeutung etwa bei der Planung und Beantragung von Forschungsprojekten und den dafür benötigen Ressourcen erlangt.

München Daniel Kastan