**Umfrage zu OCR für nichtlateinische Sprachen**

Wir interessieren uns für die Nutzung von OCR (Optical Character Recognition) für nichtlateinische Sprachen und führten dazu eine Umfrage, bestehend aus 12 Fragen, durch. Insgesamt nahmen 38 Personen daran teil, von denen 27 bereits Erfahrungen mit OCR haben. Unter den Teilnehmenden befanden sich Bachelor- und Masterstudierende, Professor\*innen sowie wissenschaftliche Mitarbeitende. Einige dieser Teilnehmenden haben OCR nur im Rahmen eines Projekts einmalig genutzt, die Mehrheit verwendet OCR jedoch regelmäßig, sei es täglich, wöchentlich oder monatlich.

Die Teilnehmenden arbeiteten mit verschiedenen Schriften, darunter Persisch, Syrisch, Arabisch in hebräischer Schrift, Garshuni (Arabisch in aramäischer Schrift) und Maya-Schriften. OCR wurde in verschiedenen Forschungsbereichen eingesetzt, darunter die Bearbeitung islamischer Handschriften und Archivdokumente, Handschriftenforschung, Fallstudien zur automatischen Zeichenerkennung, Präsentation von Quellentexten in Vorträgen oder anderen Präsentationen, Material für das Sprachtraining, Textverarbeitung für wissenschaftliche Publikationen und Texteditionen, Koran- und Bibelwissenschaft, vergleichende Religionswissenschaft, interreligiöser Dialog, Literaturwissenschaft, Katalogisierung, Textidentifizierung und die Bearbeitung von Handschriftenkatalogen.

Uns interessierte besonders, für welche Zwecke OCR genutzt wird und ob die generierten Texte ohne manuelle Korrektur übernommen und weiterverwendet werden können. Mehr als die Hälfte der Teilnehmenden berichteten, dass sie die generierten Texte zur Weiterverwendung korrigieren müssen. Dies zeigt sich insbesondere bei den Verwendungszwecken wie der Textverarbeitung für Zitation, Zusammenstellung und Indexierung, der Beschleunigung des Transkriptionsprozesses bestehender Handschriften, als erster Schritt zur Nutzung der Übersetzung, der Erleichterung des Findens von Textstellen bei nicht elektronischen Büchern, der Analyse von Handschriften durch Tagging-Systeme, der Einarbeitung des OCR in TEI (Text Encoding Initiative) und der sinnvollen Nachnutzbarkeit von Forschungsdaten.

Die Teilnehmenden verwendeten eine Vielzahl von OCR-Tools, darunter Acrobat, Word, PDF24, Google Lens, Greek OCR , Sakhr Automatic Reader und Readiris PRO 11 Corporate Ed. Es zeigte sich jedoch, dass die Ergebnisse überwiegend unzufriedenstellend waren. Auf die Antwort, welche Tools verwendet werden, wurden von den Teilnehmenden ebenfalls Angaben über die OCR-Engines gemacht. Es wurde namentlich erwähnt, dass Tesseract verwendet wird, auf welchem auch unser Workflow basiert.

Besonders bei der Texterkennung von biskriptualen Texten gab es viele unzufriedene Rückmeldungen, da die Texterkennung mehrerer Schriften sehr schlecht ausfiel, obwohl ein Bedarf nach einer verbesserten Texterkennung für biskriptuale Texte besteht.

Abgesehen von biskriptualen Texten sahen die Teilnehmenden alle vorangegebenen Herausforderungen als zutreffend an. Diese Herausforderungen bleiben weiterhin aufgrund von schlechter Erkennungsrate, der Notwendigkeit nach manueller Korrektur, hohem Zeitaufwand, aufkommenden Kosten und der Konvertierung in bestimmte Dateiformate bestehen.

Diese Umfrageergebnisse geben einen umfassenden Einblick in die vielfältigen Anwendungen und Herausforderungen der Nutzung von OCR für nichtlateinische Sprachen und verdeutlichen die Notwendigkeit manueller Korrekturen zur Sicherstellung der Genauigkeit und Verwendbarkeit der generierten Texte.