Abstract:

Optical Character Recognition (OCR) ist ein zentrales Instrument für die digitale Erschliessung historischer Bibliotheksbestände, indem sie Volltextsuche und computergestützte Analysen ermöglicht. Allerdings erschwert die Fehleranfälligkeit von OCR-Transkriptionen alter Drucke die Anwendung. Vorliegende Arbeit untersucht die OCR-Qualität frühneuzeitlicher deutschsprachiger Bücher aus dem Bestand der Zentralbibliothek Zürich auf e-rara.ch. Zur automatisierten Qualitätsevaluation werden drei Messgrössen getestet: 5-gramm-Analyse, Pseudo-Perplexität und lexikalischer Lookup. Eine Regressionsanalyse zeigt, dass die 5-gramm-Analyse die Zeichenfehlerrate am zuverlässigsten schätzt ($R^2 = 0.80$) und als Qualitätsindikator für den Gesamtbestand verwendet werden kann. Gemäss Untersuchung weisen rund 43 Prozent der OCR-Texte eine Zeichenfehlerrate (CER) von über 10 Prozent auf, wodurch Auffindbarkeit und Weiterverarbeitung dieser Texte erheblich eingeschränkt ist. Basierend auf den Ergebnissen werden Empfehlungen für eine gezielte Nachkorrektur, den Einsatz trainierter OCR-Modelle und eine transparente Kennzeichnung der OCR-Qualität für Nutzende formuliert.

Marco Geissbühler