Ludwig Boltzmann Institut
Neulateinische Studien

Dr. Uwe Springmann

(Centrum für Informations- und Sprachverarbeitung München)

Neue Methoden für die OCR historischer Drucke am Beispiel des Lateins

Monday, 06.06.2016, 18:00
Zentrum für Alte Kulturen
Langer Weg 11, SR 5

Mit der fortschreitenden Digitalisierung (verstanden als fotografische Abbildung von Buchseiten) historischer Buchbestände wird eine enorme Menge neulateinischen Schrifttums zu Tage gefördert, dessen Existenz allenfalls Spezialisten bekannt war. Eine umfassende Sichtung und Erschließung setzt jedoch die Umwandung der Bildseiten in elektronischen Text voraus, der dann indiziert, annotiert und einer Suche zugänglich gemacht werden kann. Maschinelle Verfahren der Optical Character Recognition (OCR) versagen jedoch bisher angesichts der wenig normierten Typografien, vielen Sonderzeichen und dem schlechten Verhältnis von Signal (Schrift) zu Rauschen (Seitenhintergrund) aufgrund Alterungs- und Nutzungsprozessen, so dass lediglich Erkennungsraten von etwa 85% korrekt erkannten Zeichen erreicht werden konnten. Seit 2013 konnten jedoch neue Verfahren maschinellen Lernens mittels rekurrenter neuronaler Netze, die derzeit im Bereich der Mustererkennung Furore machen, für die OCR historischer Drucke nutzbar gemacht werden. Der Vortrag gibt einen Überblick über die erreichbaren Zeichengenauigkeiten von bis zu 99% selbst bei den ältesten historischen Drucken der Inkunabelzeit (vor 1500) und gibt einen Ausblick auf die künftig zu erwartende großflächige Verwandlung bereits gescannten neulateinischen Schrifttums in elektronischen Text.