Zeitschrift des Max-Planck-Instituts für europäische Rechtsgeschichte
Journal of the Max Planck Institute for European Legal History

Rechts Rg geschichte

Rechtsgeschichte
Legal History

www.rg.mpg.de

http://www.rg-rechtsgeschichte.de/rg24 Rg 24 2016 349 – 350

Thorsten Schlauwitz

Chancen und Grenzen der automatischen Schriftanalyse und -erkennung

Thorsten Schlauwitz

Chancen und Grenzen der automatischen Schriftanalyse und -erkennung

Die Anwendung der Verfahren der Automatischen Mustererkennung auf historische Dokumente ist ein Teilgebiet der Digital Humanities, welches bislang relativ wenig Beachtung erfährt. Dabei liegen in näherer Zukunft hier die größten Potentiale. Die Mustererkennung kann in mehreren geisteswissenschaftlichen Bereichen wie z. B. bei Audio- und Bilddokumenten eingesetzt werden. Die folgenden Ausführungen werden sich auf die Einsatzmöglichkeiten bei (mittelalterlichen) Textdokumenten konzentrieren. Mittels computergestützter Verfahren können zunächst Layoutanalysen (z. B. Verhältnisse von Textblock zur Dokumentengröße, Anzahl, Höhe und Abstand von Zeilen) durchgeführt und somit Fragen zur Präzision des Schreibers sowie zum optisch-symbolischen Charakter der Schriftdokumente beantwortet werden. Ebenso ist es möglich, Phänomene der Schriftentwicklung über einen langen Zeitraum aufzuzeigen, was zum Verständnis der Ursachen und der Einflussfaktoren von Veränderungsprozessen beiträgt. Im Zusammenhang damit steht das Forschungsinteresse an einer Schreiberidentifizierung, wofür die Bestimmung von Autographen oder die Untersuchung von Verwaltungsvorgängen nur einzelne Anwendungsbeispiele sind (vgl. die Projekte »Schrift und Zeichen«, Erlangen/München und »eCodicology«, Darmstadt).

Daneben ist die Entwicklung einer automatischen Schrifterkennung von erheblichem Wert, nicht zuletzt vor dem Hintergrund wohl allgemein abnehmender Paläographie-Fähigkeiten. Die zunehmende Schriftlichkeit ab dem späten Mittelalter verbunden mit dem Aufkommen der Stadtbücher und das heranziehende Aktenzeitalter führen seit dieser Zeit zu einer wachsenden Diskrepanz zwischen den überlieferten und den edierten Texten. Mit klassischen Verfahren wird sich an diesen Zuständen in absehbarer Zeit nicht viel ändern, was zwangsläufig dazu führt, dass das vollständige Erkenntnis-Potential nicht ausgeschöpft wird. Gleichzeitig steht eine zunehmende Anzahl an Digitalisaten im Internet zur Verfügung, die aber nicht im wünschenswerten Umfang durch Regesten oder gar Volltext erschlossen sind.

Die Texterkennung von mit der Hand geschriebenen Dokumenten ist allgemein recht weit fortgeschritten. Bei modernen Schriften können somit bereits sehr hohe Erkennungsraten erreicht werden. Bei historischen Texten jedoch liegt die Quote auch bei einer intensiven Vorbereitung zumindest derzeit noch wesentlich niedriger, was u. a. mit dem Schreibmaterial und dem Erhaltungszustand der Dokumente im Zusammenhang steht. Eine Vorverarbeitung der Digitalisate kann diese Fehlerquellen minimieren, aber nicht gänzlich beseitigen. Auch die Erkennung der Textteile (Textdetektion) und die Zerlegung in einzelne Bestandteile (Segmentierung in Zeilen, Wörter, Buchstaben) stellen ein Problem dar, wofür aber zumindest teilweise bereits technische Lösungen und Ansätze gefunden wurden. Erkennungsraten wie bei Druckwerken und modernen Handschriften werden wohl dennoch nie erreicht werden. Als ein dringendes Desiderat erscheint ein verbesserter Umgang mit Korrekturen im Text. Bei den in Reinschrift verfassten Buchschriften, wo erwartungsgemäß die besten Erkennungsraten zu verzeichnen sind, handelt es sich häufig um Texte, die allgemein bekannt, erforscht und ediert sind. Größeren Erkenntnisgewinn versprechen dagegen Schriften des alltäglichen Verwaltungsschriftgutes, welches durch Rand- und Interlinearglossen sowie Durchstreichungen geprägt ist.

Nach dieser Skizze der Einsatzmöglichkeiten und dem technischen Stand der automatischen Handschriftenanalyse und -erkennung sind im Folgenden die Implikationen auf die Geisteswissenschaft zu betrachten. Grundsätzlich ist beim derzeitigen Forschungsstand der Einsatz dieser Technik nur bei einem größeren und weitgehend homogenen Quellenkonvolut sinnvoll. Da optimale Ergebnisse im Bereich der Schrifterkennung nur erzielt werden, wenn für jede Schreiberhand (oder doch zumindest sehr ähnliche Schreiberhände) ein eigener Algorithmus trainiert wird, ist der Anfangsaufwand recht hoch und rentiert sich somit nur in Fällen mit entsprechend umfangreicher Textmasse. Gleiches trifft für die Schreiberidentifizierung und Schriftanalyse zu. Da in diesem Bereich aber technische Verbesserungen zu erwar-

ten sind und die Zahl der vorhandenen Algorithmen und Trainingsdatensätze zunimmt, ist tendenziell mit einer abnehmenden Grenzschwelle zu rechnen. Voraussetzung ist eine zentrale Sammlung dieser Daten mit einer entsprechenden Infrastruktur, wofür das Projekt READ (Innsbruck) die besten Rahmenbedingen zu bieten scheint. Perspektivisch wäre es sogar vorstellbar, mittels einer computergestützten Analyse ein Dokument anhand seiner Schrift einer Region, einem Zeitraum, einer Kanzlei und vielleicht sogar einem einzelnen Schreiber zuzuordnen.

Eine breitflächige Durchsetzung dieser Technik hätte Konsequenzen: Sie würde vermutlich eine Konzentration auf Editionsvorhaben mit größeren, formal und schrifttechnisch homogenen Quellen zur Folge haben, während kleinere und individuelle Texte weniger Berücksichtigung fänden. Dies könnte aber ein derzeit bestehendes Missverhältnis ausgleichen. Zudem dürfte dieses Vorgehen die Entwicklung hin zu digitalen Publikationen mit einer wünschenswerten Text-Bild-Verlinkung verstärken. Damit ständen diese auch als paläographisches Schulungsmaterial zur Verfügung. In diesem Rahmen könnten weiterhin kollaborative Arbeitsumgebungen geschaffen werden, mit der automatisch transkribiertes, aber nicht manuell kontrolliertes Material durch die Forschergemeinschaft bearbeitet wird.

Die geisteswissenschaftliche Expertise ist bei all diesen Einsatzmöglichkeiten unerlässlich. Dies betrifft bereits die Auswahl des Quellenmaterials, die entscheidend das folgende Ergebnis beeinflussen kann. Paläographische Fähigkeiten bleiben daher weiter unerlässlich. Weiterhin müssen die Trainingsdaten für die Computeranalyse erstellt werden, vor allem aber die Ergebnisse geprüft, eventuelle Fehlerquellen benannt und getilgt sowie v. a. die Ergebnisse interpretiert werden. Hierfür ist auf Seiten des Geisteswissenschaftlers ein Verständnis für die Verfahren der Mustererkennung vonnöten, ohne damit den Fachinformatiker ersetzen zu können. So ist zu bedenken, dass in der Mustererkennung mit Wahrscheinlichkeiten gearbeitet wird. Vom Computer wird daher der Wert ausgegeben, der nach den ihm vorliegenden Informationen am ehesten zutrifft. Lässt man sich stattdessen die drei Treffer mit den höchsten Wahrscheinlichkeiten anzeigen, erhöht sich die Erkennungsrate erheblich. Dies führt zu einer höheren Validität der Ergebnisse, hat jedoch nicht nur einen vermehrten Arbeitsaufwand für den Geisteswissenschaftler, der aus den Vorschlägen den korrekten Wert auswählen muss, zur Folge, sondern erweitert gleichzeitig den Interpretationsrahmen.

Die automatisch generierten Ergebnisse sind daher stets kritisch zu hinterfragen. Der methodisch richtige Umgang mit der hohen Suggestivkraft des vom Computer vorgegebenen Ergebnisses ist zu erlernen. Bei der Transkription ist bei den derzeitigen Wortfehlerraten immer ein folgender Korrekturgang notwendig. Bei der Schriftanalyse und Schreiberidentifizierung sollten die computergestützten Angaben vielmehr als Hinweise und Wegweiser in großen Quellenkonvoluten betrachtet werden und nicht als unumstößliches Ergebnis. Ergänzende Methoden wie die Stilometrie können die Ergebnisse untermauern oder widerlegen. Die automatische Mustererkennung hilft also bei der Bearbeitung großer Quellenkorpora und beschleunigt Arbeitsprozesse. Gleichzeitig muss man anfangs jedoch in das Verfahren investieren. Die zu behandelnden Fragen sind dabei nicht neu, doch können manch offene Punkte durch die automatischen Verfahren erstmals beantwortet werden.

■