„Wenn der Computer Opas Tagebuch transkribiert...“ Historisches und philologisches Arbeiten mit Transkribus

Barbara Denicolò, Universität Innsbruck (8. Juni 2017)

In meinem Vortrag habe ich ein an der Universität Innsbruck im Rahmen eines EU-Projekts entwickeltes Computerprogramm zur Transkription historischer Handschriften und Drucke vorgestellt, das ich über meine Mitarbeit an einem Drittmittelprojekt kennengelernt habe und jetzt mittlerweile auch individuell für Folgeprojekte und eigene Ideen und Projekte nütze.

„Wenn der Computer Opas Tagebuch transkribiert...“, dieser etwas flapsige und reißerische Titel stammt nicht von mir, sondern war die Schlagzeile eines Zeitungsartikels in der Südtiroler Tageszeitung Dolomiten, der im Herbst 2016 über das Projekt erschienen ist. Er zeigt aber m. E. sehr gut, welche Vorstellungen und Erwartungen von Außenstehenden an ein Programm zur automatischen Handschriftentranskription mittels neuronaler Netze herangetragen werden.

In den meisten Fällen müssen sie aber relativiert und die Fähigkeiten künstlicher Intelligenz etwas entzaubert werden. HistorikerInnen und PhilologInnen schaffen sich also nicht selbst ab, wenn sie solche Möglichkeiten unterstützen und nützen, vielmehr erhalten sie nützliches Werkzeug und viele neue Perspektiven und Möglichkeiten, große Daten- und Textmengen zu bearbeiten und auszuwerten. Denn an die intellektuellen Auswertungs- und Interpretationskompetenzen des Menschen wird die Maschine noch lange nicht herankommen.

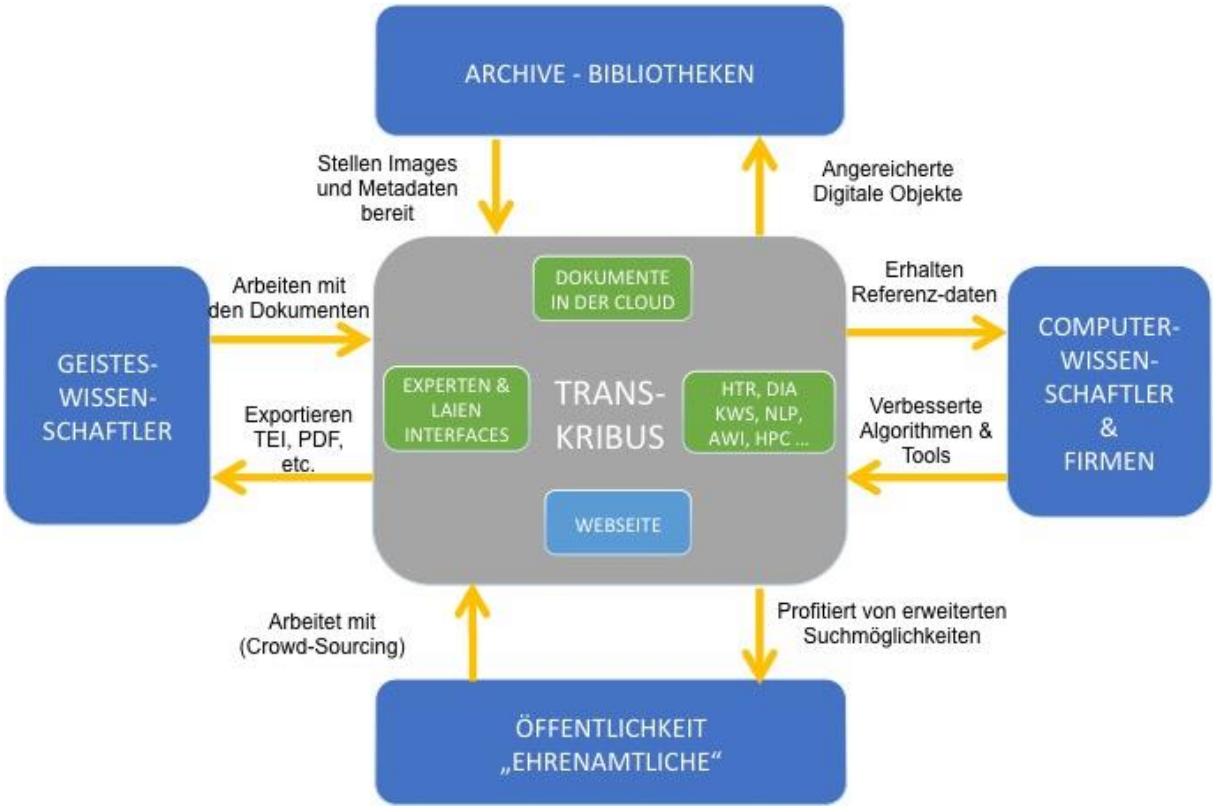
Editionen waren und sind das “Kerngeschäft” von HistorikerInnen und PhilologInnen. Durch den „digital turn“ und den dadurch entstandenen gesellschaftlichen Auftrag zur Massendigitalisierung an Bibliotheken und Archive liegt der Fokus nun zunehmend auf „digitalen Editionen“ stark gewachsener Korpora. Dieser Umstand bietet viel Potenzial für neue Forschungsfragen und neue Entdeckungen, stellt ForscherInnen aber auch vor neue Herausforderungen und Probleme.

Denn auf der anderen Seite sind Forschungsförderungen heute stets knapp bemessen und nur mehr auf wenige Jahre beschränkt. Editionen größeren Umfangs „im alten Stil“ sind daher kaum mehr möglich.

Große Quellenbestände stellen die ForscherInnen vor verschiedene, ganz unterschiedliche Hürden: Wie bearbeite ich diese Fülle an Material? Wo und wie lagere ich die großen Datenmengen? Können mehrere Personen an verschiedenen Orten gleichzeitig daran arbeiten? Wie bringe ich ihre Arbeitsleistung zusammen? Durch welche Speicherformate, etablierte Codierungen etc. lässt sich eine breite und dauerhafte Nutzung gewährleisten? Wie umgehen mit Worttrennungen, Streichungen, Einfügungen, Marginalien? Was tun mit Abkürzungen? Wie kann ich sog. “Named Entities” (Namen von Personen, Völkern, geographischen Orten etc.) und Fachtermini markieren und normalisieren? Wie gehe ich mit unterschiedlichen Schreibweisen von Eigennamen um, gerade im Hinblick auf eine spätere Durchsuchbarkeit der Daten? Wie präsentiere ich das Material nach Abschluss der Arbeiten? Digital oder gedruckt?

Die zunehmende Digitalisierung und Vernetzung sowie die stärkere Öffnung der Archive und Bibliotheken gegenüber der Öffentlichkeit führen schließlich auch dazu, dass sich qualifizierte Freiwillige aus der breiten Öffentlichkeit an wissenschaftlichen Projekten und an der Erhaltung von Kulturerbe beteiligen möchten.

Um all diese Interessen miteinzubeziehen, die verschiedenen Bedürfnisse zu befriedigen und Synergien zu schaffen, wurde daher die Forschungsplattform „Transkribus“ konzipiert, von der das oben erwähnte Programm nur ein Teilbereich ist, mit dem Ziel, einerseits die Grundlagenforschung in den Computerwissenschaften zu befördern, andererseits für Archive, Bibliotheken, GeisteswissenschaftlerInnen und die Öffentlichkeit die Technologie nutzbar zu machen.



Über Transkribus, dessen verschiedene Tools und die daran angliederten Clouddienste lassen sich die verschiedenen Interessen und Bedürfnisse von GeisteswissenschaftlerInnen, der Gesellschaft und interessierten Freiwilligen, ComputerwissenschaftlerInnen und Firmen sowie Bibliotheken und Archiven zusammenbringen, sodass alle von einander profitieren können.

Gerade die automatische Handschriftenerkennung (HTR für Handwritten Text Recognition) ist einerseits auf die Bereitstellung von Images verschiedener Provenienzen und Epochen aus den Archiven und Bibliotheken angewiesen, andererseits aber auch auf möglichst viel Trainingsmaterial, also transkribierten und korrigierten Text, um die neuronalen Netze trainieren zu können. Denn der Computer muss jede Handschrift zu erst erlernen, um dann dem Menschen durch automatisierte, schnelle Transkriptionsleistungen behilflich sein zu können.

Dazu wurde das Crowdsourcingprojekt „Bozner Ratsprotokolle transkribiert“ in Zusammenarbeit mit dem Stadtarchiv Bozen gestartet, wo Teile der Bozner Stadtratsprotokolle vergangener Jahrhunderte durch interessierte Freiwillige (meist AhnenforscherInnen, Studierende, RentnerInnen, ChronistInnen usw.) in Transkribus bearbeitet und transkribiert werden, um Trainingsmaterial für die HTR zu schaffen und dem Stadtarchiv digital angereichertes und daher besser strukturiertes und durchsuchbares Archivgut zurückzugeben.