Mit neuen Suchstrategien vom isolierten Text zu 'illuminierten Urkunden'

Bürgermeister, Martina
martina.buergermeister@uni-graz.at
Universität Graz, Österreich

Bartz, Gabriele
Gabriele.Bartz@oeaw.ac.at
ÖAW, Wien

Gneiß, Markus
Markus.Gneiss@oeaw.ac.at
ÖAW, Wien

Immer mehr mittelalterliche Urkunden sind heute digital verfügbar, doch viele nur als Text in Urkundenbüchern. Die hohe Anzahl von online verfügbaren Abbildungen hat die Erforschung von ‚illuminierten Urkunden', die sich in unterschiedlichsten Archiven befinden, erst möglich gemacht. Unter ‚illuminierte Urkunden' versteht man Urkunden mit bildlichem Dekor, die in spezifischen Fällen einen starken wechselseitigen Bezug von Bild und Text aufweisen. Eine Suchabfrage nach Illuminationen führt aber kaum zu Treffern. Das heißt, mit klassischen Suchstrategien sind ‚illuminierte Urkunden' wenn überhaupt nur mit sehr hohem Zeitaufwand auffindbar.

Deshalb scheint es zielführende, dass Expertinnen und Experten der Diplomatik, Kunstgeschichte und Digital Humanities zusammenarbeiten, um neue Suchstrategien für ‚illuminierte Urkunden' zu erproben. Urkunden mit Bildschmuck als interdisziplinäres Forschungsfeld sind noch nicht lange im Fokus der Geisteswissenschaften (Roland/Zajic 2013, Bartz/Gneiß 2018). Das erklärte Ziel besteht darin, Hinweise auf Illuminationen in den Metadatentexten aufzufinden. Dazu werden in einem mehrstufigen Textanalyseprozess aus bereits erschlossenen Urkunden automatisiert Stichwörter extrahiert („keyword extraction"), gerankt und zur API-Abfrage von online Archivdatenbanken aufbereitet, um zu neuen Treffern zu führen.

Der Einsatz von textstatistische Methoden zur Beantwortung von Forschungsfragen der digitalen Diplomatik ist state of the art (vgl. Tilahun/Feuerverger/Gervers 2012, Perreaux 2014, HIMANIS-Projekt: www.himanis.org). Neu jedoch ist deren Anwendung, um über Metadaten zu einer innovativen Suchstrategie nach Urkunden mit Bildschmuck zu gelangen. Unser Beitrag ist als Reaktion auf das wahrgenommene Defizit der isolierten Betrachtung von Text- und Bilddaten zu verstehen.

Daten

In unserem Fokus steht eine weit verbreitete Gruppe von ‚illuminierten Urkunden': den Wappenbriefen. Bei dieser Urkundengattung gewährt ein Herrscher das Recht ein bestimmtes Wappen zu tragen, das normalerweise auf der Urkunde gemalt ist. Wappenbriefe sind in großer Zahl ausgestellt worden und sind deshalb einheitlich in ihren Textbestandteilen. In Vorgängerprojekten (FWF, Go!Digital) wurden bereits etwa 150 Wappenbriefe mit historischen Methoden gesammelt, beschrieben und über das Urkundenarchiv „monasterium.net" (monasterium.net/mom/collection/illuminierteUrkunden) veröffentlicht.

Zum ersten Mal sollen nun unter Anwendung textstatistischer Methoden eine große Zahl an Wappenbriefen gefunden, in Hinblick auf die gesamteuropäische Entwicklung untersucht und über monasterium.net veröffentlicht werden. Die multimodalen Erschließungsmöglichkeiten in monasterium.net schaffen die Voraussetzung zur komparativen Untersuchung von textlichen Eigenschaften und Illuminationen von zentral-, süd- und westeuropäischen Beispielen. Insgesamt werden durch das Projekt neue Impulse für die Erforschung von Wappenbriefen gesetzt.

Methoden

Die geringe Anzahl an bisher gesammelten und tiefenerschlossenen Wappenbriefen, sowie der Umstand, dass es sich um Kurztexte handelt, ist für das Verfahren der „keyword extraction" eine besondere Herausforderung. Um jene Wörter in den bereits erschlossenen Dokumenten zu finden, die konstitutiv sind, werden ausschließlich statistische Methoden, welche an kein umfangreiches Trainingsset gebunden sind, eingesetzt. Grundlage der Verfahren ist die numerische (vektorisierte) Repräsentation aller zu untersuchenden Wörter eines Dokuments. In einem ersten Schritt soll das TF-IDF-Maß gewichtete Stichwörter liefern. Dann werden zwei weitere methodische Ansätze (clustering, entropy), die sich zur Stichwortextraktion eignen (vgl. Herrera/Pury 2008, Carretero-Campos et al. 2013, Jamaati/Mehri 2018) auf Texte der Wappenbriefe angewandt. Zur Bestimmung der Wortrelevanz spielt sowohl beim Clustering also auch bei der Entropie-Methode die Verteilung der Wörter im Text eine entscheidende Rolle. Beim Clustering wirkt sich die Streuung gemessen an unterschiedlichen Wortabständen (Vorkommen, Position) auf das Ranking der so generierten Stichwörter aus. Dabei gilt, je größer die Abweichung, desto relevanter das Wort (vgl. Zhou/Slater 2003, Carretero-Campos et al. 2013). Die mithilfe Shannons Entropie (Shannon/Weaver 1963) errechneten relevanten Wörter werden auch an ihrer Verteilungsheterogenität gemessen: Je ungleichmäßiger ihre Verteilung desto relevanter der Inhalt (vgl. Herrera/Pury 2008). Schließlich werden die Ergebnisse aller Messungen evaluiert und

für die Datenbankabfrage aufbereitet. Die ermittelten Schlüsselwörter sollen die Qualität des Information Retrieval erhöhen und über die Abfrage von öffentlichen Datenbank-APIs, wie Regesta Imperii (www.regesta-imperii.de) und Archivportal-D (www.archivportal-d.de/) zum Auffinden von weiteren Wappenbriefen führen.

Schluss

Die Suchstrategie geht auf eine neue Art mit isoliertem Text um. Die Multimedialität von ‚illuminierten Urkunden' wird durch die automatisierte Extraktion von Stichwörtern zu einem serialisierten Abfragemuster für öffentliche Archiv-APIs, die so zu umfangreicheren Quellenfunden führen. Die quantitative, statistische Methodologie des Projektes bringt zudem eine neue Perspektive auf Wappenbriefe und deren formelhafte Struktur, die über Zeit und Ort verglichen werden kann. Was im vorliegenden Projekt an Wappenbriefen getestet wird kann auf andere Urkundentypen ausgedehnt werden und einen grundlegenden DH-Beitrag zum Thema „keyword extraction" bei Kurztexten liefern.

Bibliographie

Bartz, Gabriele / Gneiß, Markus (2018) (eds.): *Illuminierte Urkunden. Beiträge aus Diplomatik, Kunstgeschichte und Digital Humanities/Illuminated Charters*. Essays from Diplomatic, Art History and Digital Humanities (= Archiv für Diplomatik, Schriftgeschichte, Siegel-und Wappenkunde 16), Wien: Böhlau Verlag.

Carretero-Campos, Concepcion / Bernaola-Galvan, Pedro / Coronado Jiménez, Ana Victoria / Carpena, Pedro (2013): *Improving statistical keyword detection in short texts: Entropic and clustering approaches* in: Physica A 392: 1481-1492.

Herrera, Juan P. / Pury, Pedro A. (2008): *Statistical keyword detection in literary corpora* in: The European Physical Journal B 63/1: 135-146.

Jamaati, Maryam / Mehri, Ali (2018): *Text mining by Tsallis entropy*, in: Physica A 490: 1368-1376.

Perreaux, Nicolas (2014): *De l'accumulation à l'exploitation? Expériences et propositions pour l'indexation et l'utilisation des bases de données diplomatiques* in: **Ambrosio/Barret/Vogeler (eds.)** *Digital diplomatics. The computer as a tool for the diplomatist?* (= Archiv für Diplomatik Beihefte 14), Köln/ Weimar/Wien: Böhlau Verlag 187-210.

Roland, Martin / Zajic, Andreas (2013): *Illuminierte Urkunden des Mittelalters in Mitteleuropa*, in: Archiv für Diplomatik 59: 241-432.

Shannon, Claude Elwood / Weaver, Warren (1963): *The Mathematical Theory of Communication* Champaign: Illinois University Press.

Tilahun, Gelila / Feuerverger, Andrey / Gervers, Michael (2012): *Dating Medieval English Charters*, in: The Annals of Applied Statistics 6/4: 1615-1640.

Zhou, H. / Slater, G.W. (2003): *A metric to search for relevant words*, in: Physica A 329 (2003) 309–327.