NLP meets RegNLP meets Regesta Imperii

Blessing, Andre
andre.blessing@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

Kuczera, Andreas
andreas.kuczera@adwmainz.de
Akademie der Wissenschaften und der Literatur, Mainz

Dieser Posterbeitrag veranschaulicht die Interaktion zwischen computerlinguistischen Methoden und Regestenforschung. Es wird eine Anwendung vorgestellt, die bereits in einem graphbasierten Format vorliegendene Regesten webbasiert anzeigt und es erlaubt, Registereinträge im Text zu verorten. Die daraus entstandene Datenbasis hilft dabei neues Wissen zu generieren, so können z.B. Verwandtschaftsbeziehungen automatisch erkannt und in den Regesten-Graph integriert werden.

Das im Rahmen des Bund-Länder-geförderten Akademienprogramms angesiedelte Grundlagenforschungsprojekt Regesta Imperii erstellt deutschsprachige Inhaltsangaben (sog. Regesten) von Kaiser-, Königs und Papsturkunden, begonnen von Karl dem Großen bis hin zu Kaiser Maximilian. Seit Projektbeginn 2001 wurden 1829 Regesten erstellt und digitalisiert und stehen inzwischen als Volltext im Internet zur Verfügung. Die Publikation der digitalisierten Register befindet sich gerade in Vorbereitung.

Neben den Regesta Imperii sind immer mehr Editionen und Regestenwerke als Volltext im Internet zugänglich und können über Suchmasken abgefragt und genutzt werden. Die Nutzungsart unterscheidet sich zumeist aber nicht grundlegend von einer analogen Nutzung des Buches: Das Register wird aufgeschlagen und man kann anschließend die einem Registereintrag zugeordneten Urkunden oder Regesten aufrufen und lesen.

Mit der Nutzung von Graphentechnologien in den digitalen Geisteswissenschaften werden neue Nutzungs- und Analyseformen der bereits vorhanden digitalen Editions- und Regestenwerke möglich. Die Digitale Akademie, Mainz (*www.digitale-akademie.de*) hat auf ihrer Seite *www.graphentechnologien.de* einige beispielhafte Anwendungsszenarien für die Nutzung von Graphdatenbanken zur Erschließung von Onlineregesten vorgestellt. Für dieses Beispielprojekt wurden die Regesten Kaiser Friedrichs III. in eine Graphdatenbank konvertiert, anschließend das zugehörige Register digitalisiert und in die Graphdatenbank integriert. Im Graphenmodell ist es über die Abfrage nun möglich herauszufinden, in welchem Regest eine Person genannt wird und eine Analyse der gemeinsam mit ihr im Regest genannten Personen zu veranlassen (Abbildung 1). Das Graphenmodell erlaubt zudem die weitere Ergänzung von Kanten zwischen den Registerknoten. So ist es beispielsweise möglich, dass zwei Personenknoten, die Vater und Sohn darstellen, durch eine KIND-Kante ergänzt werden, um so deren Verwandtschaftsbeziehung explizit im Graphen zu repräsentieren. Mit solchen Zusatzinformationen kann das Register als Erschließungswerkzeug immer weiter wachsen.

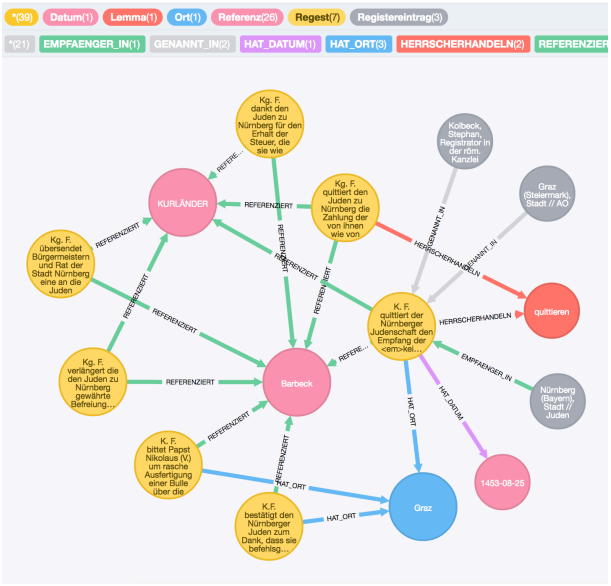

Abbildung 1 Graphbasierte Repräsentation von Regesten (gelb) mit zugehörigen Registereinträgen (rot,blau,grau).

Methoden aus der Computerlinguistik helfen diesen manuell sehr aufwendigen Grapherweiterungsschritt semi-automatisch durchzuführen. Dazu werden im ersten Schritt alle Regesten mit einer auf der Clarin-D Infrastruktur basierenden Sprachverarbeitungs-Pipeline (Malow 2012) maschinell verarbeitet: Auflösung von Abkürzungen, Tokenisierung, Part-of-Speech Tagging, Parsing und Entitätenerkennung. Im zweiten Schritt werden die Texte der einzelnen Regesten in einer interaktiven Webanwendung ausgewertet. Für jedes Regest werden alle Registereinträge (Personen, Orte, usw.), die während der Digitalisierung manuell mit Regesten verbunden wurden, angezeigt. Abbildung 2 stellt rechts den Regestentext und links alle verbundenen Registereinträge dar. In den Ausgangsdaten ist nicht gespeichert wie die Regeistereinträge im Text erwähnt werden, z.B. dass sich „der Stadt“ im Text auf den Eintrag „Aachen“ bezieht. Unsere Anwendung ermöglicht es hingegen jede im Text erkannte Entität mit den Registereinträgen per einfachem Mausklick zu verbinden. Mittels dieses Verfahrens konnten bereits ca. 4000 Registereinträge mit passenden Textstellen verbunden werden.

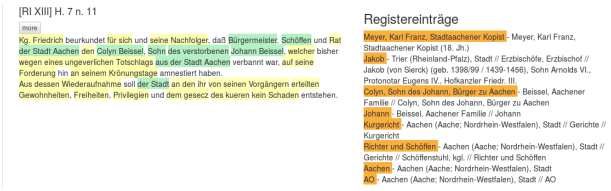

Abbildung 2: links: Regesttext, erkannte Entitäten sind hervorgehoben (bereits neu verortete Textstellen zum Eintrag sind grün); rechts: verknüpfte Registereinträge zum Regest

Die neu geschaffene annotierte Datenmenge bildet somit einerseits eine wichtige Grundlage, für die Regestenforschung, da nun sehr einfach Wissen aus Texten mit den verknüpften Registereinträgen automatisch abgeleitet werden kann: z.B. die Vater-Sohn-Relation zwischen Colyn Beisse und Johann Beissel (Abbildung 3).

Andererseits bietet dieser Datensatz für die computerlinguistische Forschung weitere Anknüpfungspunkte. Mittels Distant Supervision (Blessing 2012) können aus den verknüpften und im Text verankerten Relationen zu den Registereinträgen neue Modelle trainiert werden, die wiederum auf nicht manuell annotierten Textstellen der Regesten Anwendung finden. Durch diesen iterativen Ansatz können sukzessive große Regestensammlungen mit immer neuem Wissen angereichert werden und somit eine Grundlage für neue Analyseformen bieten.

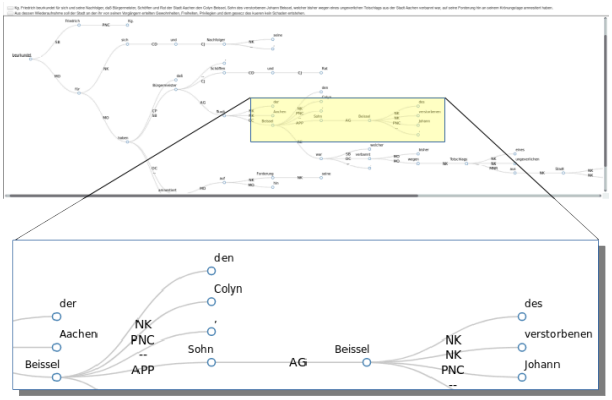

Abbildung 3: Depedenzanlayse, die zeigt wie Verwandtschaftsrelationen direkt aus der Analyse extrahiert werden können. In diesem Beispiel ist Colyn Beissel der Sohn von Johann Beissel.

Bibliographie

Blessing, Andre / Schütze, Hinrich (2012) Crosslingual Distant Supervision for Extracting Relations of Different Complexity. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*

Kuczera, Andreas (2017) Herrscherhandeln in den Regesta Imperii. Beispielprojekt an den Regesten Kaiser Friedrichs III. URL: http://www.digitale-akademie.de/forschung/graphentechnologien/beispielprojekte/ (abgerufen am 14.09.2017)

Kuczera, Andreas (2016): Graphdatenbanken für Historiker. Netzwerke in den Registern der Regesten Kaiser Friedrichs III. mit neo4j und Gephi, in: Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte, 05.05.2015. URL: *http://mittelalter.hypotheses.org/5995* .

Mahlow, Cerstin / Eckart, Kerstin / Stegmann, Jens / Blessing, Andre / Thiele, Gregor / Gärtner, Markus / Kuhn, Jonas (2014) Resources, Tools, and Applications at the CLARIN Center Stuttgart in Proceedings of the 12th Konferenz zur Verarbeitung natürlicher Sprache 11-21