Ketzan, Erik
eketzan@gmail.com
Birkbeck, University of Lonfon

Wildgans, Julia
j.wildgans@googlemail.com
IDS Mannheim; Universität Mannheim

Witt, Andreas
witt@ids-mannheim.de
IDS Mannheim; Universität zu Köln

Wissenschaftler im Bereich der Digital Humanities sind ständig auf einen Zugang zu vertrauenswürdigen und zuverlässigen rechtlichen Informationen angewiesen. Die entscheidenden rechtlichen Herausforderungen stellen sich vor allem im Immaterialgüterrecht (insbesondere in Bezug auf das Urheberrecht, das sui generis-Recht für Datenbanken und das verwandte Schutzrecht für den Verfasser von wissenschaftlichen Ausgaben) und im Datenschutzrecht. Daher ist es sinnvoll, beides bereits in der Anfangsphase jedes Projekts zu berücksichtigen, um rechtliche Probleme in späteren Projektphasen und das Scheitern von Forschungsprojekten zu vermeiden.

Allerdings erscheint eine ständige Information über die rechtlichen Rahmenbedingungen vor dem Hintergrund der ständigen Änderungen der Gesetze, die die neuen Technologien betreffen, sehr schwierig. Auch in 2018 wird es sowohl im deutschen als auch im europäischen Datenschutzrecht wesentliche Änderungen geben, die Auswirkungen auf die Erhebung, den Zugang und die Verwendung von Forschungsdaten haben werden. Darüber hinaus wird derzeit über den Entwurf einer neuen Richtlinie im Urheberrecht diskutiert, die möglicherweise schon bald verabschiedet wird. All diese Änderungen im Blick zu behalten erfordert jedenfalls regelmäßigen Zugang zu aktuellen rechtlichen Informationen.

Daher haben Pawel Kamocki und Erik Ketzan im Jahr 2012 die CLARIN-D Legal Information Plattform für DH Forscher in Deutschland aufgesetzt: Sie ist sowohl in deutscher als auch in englischer Sprache verfügbar. 2016 folgte die CLARIN Legal Information Plattform für Wissenschaftler aus den übrigen CLARIN Consortium Ländern, die bisher lediglich in englischer Sprache abrufbar ist. Beide Webseiten stellen in verschiedenen Artikeln und Formaten (derzeit insgesamt ca. 25.000 Wörter) rechtliche Informationen für den Bereich der Digital Humanities bereit und streben dabei danach, die umfangreichste und aktuellste Wissensressource für Wissenschaftler zu sein.

Sie enthalten Erklärungen zu den grundlegenden rechtlichen Prinzipien und Konzepten im Bereich des Urheberrechts (Gegenstand, Rechteinhaberschaft, Umfang und Reichweite des Schutzes und Schrankenregelungen insbesondere für wissenschaftliche Zwecke) und des sui-generis-Rechts für Datenbanken, zur Lizenzierung (einschließlich der Nutzung öffentlicher Lizenzen für Daten und Software) und zum Datenschutz. Darüber hinaus werden Wissenschaftler bei Bedarf auch zu praktischen Lizenzauswahlinstrumenten weitergeleitet, wie z.B. dem “Public License Selector” (http://ufal.github.io/public-license-selector/), der 2014 im Rahmen einer Kooperation zweier CLARIN-Zentren von Kamocki, Stranak und Sedlak entwickelt wurde. Zusätzlich bieten die Plattformen Zugriff auf die CLARIN Legal Issues Committee (CLIC) White Paper Series, die eine Open Access Publikation von Kommentaren und Forschungsergebnissen bezüglich rechtlicher Fragestellungen im Bereich der Sprachwissenschaft unter der redaktionellen Leitung des CLIC ermöglichen.

Das Legal Helpdesk ist der “direkte Draht” zu einer persönlichen Hilfestellung: Dieses ermöglicht eine Kontaktaufnahme mit einem Teammitglied des CLARIN-Teams, das Wissenschaftler zu hilfreichen Ressourcen und Informationen bezüglich ihrer Forschungsfrage leiten kann.

Die Plattformen sind frei im Internet verfügbar und werden in regelmäßigen Abständen aktualisiert. Beide werden häufig im Rahmen von CLARIN-D und CLARIN-EU-Projekten genutzt.

Unser Poster wird diese hilfreichen CLARIN Ressourcen anhand von Graphiken und Text vorstellen und aktuelle Updates darstellen, die der DH Community möglicherweise noch unbekannt sind.

Delta vs. N-Gram-Tracing: Wie robust ist die Autorschafts-attribuierung?

Proisl, Thomas
thomas.proisl@fau.de
Friedrich-Alexander-Universität Erlangen-Nürnberg, Deutschland

Evert, Stefan
stefan.evert@fau.de
Friedrich-Alexander-Universität Erlangen-Nürnberg, Deutschland

Die Autorschaftsattribuierung, also die Zuweisung von Texten unbekannter oder umstrittener Autorschaft zu ihrem wahren Autor, hat vielfältige Anwendungen beispielsweise in der Literatur- und Geschichtswissenschaft oder der forensischen Sprachwissenschaft. Eine populäre Methode zur Autorschaftsattribuierung ist die Anwendung von Deltamaßen (Burrows 2002; Argamon 2008) wie zum Beispiel Cosine-Delta (Smith und Aldridge 2011). Deltamaße verwenden die n häufigsten Wörter im Korpus, standardisieren die Frequenzen auf z-Werte und wenden ein Abstandsmaß, im Fall von Cosine-Delta den Kosinusabstand, an. Typischerweise schließt sich die Anwendung eines (hierarchischen) Clusterverfahrens an, das Texte des selben Autors zusammengruppiert.

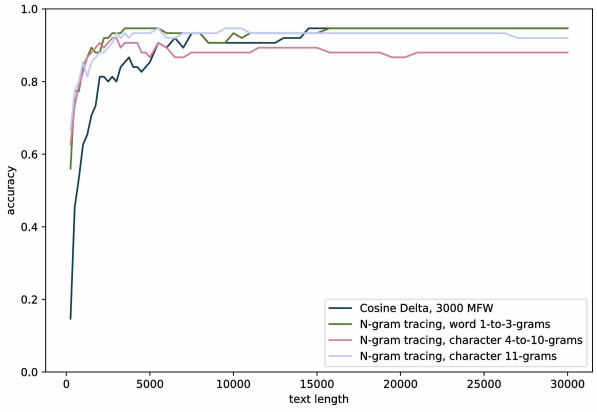
Eine neue Methode zur Autorschafts-attribuierung ist das sogenannte N-Gram-Tracing (Grieve et al., in Begutachtung). Hierbei werden aus dem zu klassifizierenden Text alle Wort- oder Buchstaben-N-Gramme einer bestimmten Länge extrahiert. Der Text wird dann dem Autor zugewiesen, der im Vergleichskorpus die meisten dieser N-Grammtypen verwendet. Die Häufigkeit der N-Gramme spielt dabei keine Rolle, es geht nur darum, wie viele N-Gramme aus dem zu klassifizierenden Text auch im Vergleichskorpus auftauchen.

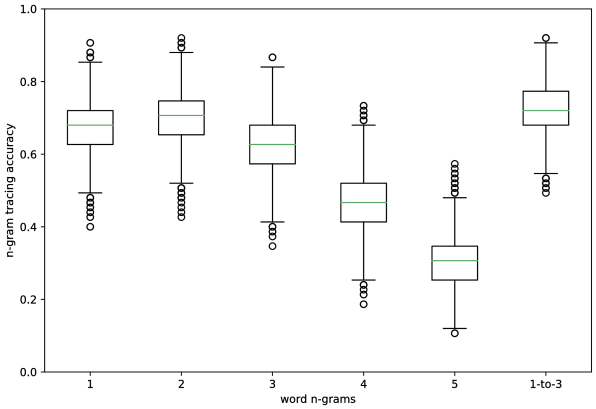
Wenn Methoden zur Autorschaftsattribuierung angewandt werden sollen um tatsächlich eine strittige Autorschaftsfrage zu klären, ist es sehr wichtig die Zuverlässigkeit und Robustheit der Verfahren abschätzen zu können, schließlich gibt es eine ganze Reihe von Einflussfaktoren. Kritisch sind zum Beispiel die folgenden Fragen: Welchen Einfluss haben die Länge des zu klassifizierenden Textes und die Größe des Vergleichskorpus auf die Genauigkeit der Autorschaftsattribuierung? Gibt es für die beiden Verfahren eine Mindesttextlänge, die nicht unterschritten werden sollte? Wie stark werden die Verfahren durch autor- und werkspezifische Eigenheiten beeinflusst? Ist die Genauigkeit der Autorschaftsattribuierung robust in Bezug auf die Zusammensetzung des Vergleichskorpus oder kann die Auswahl der Autoren und Texte das Ergebnis beeinträchtigen?

Um diese Fragen zumindest teilweise beantworten zu können, führen wir eine Reihe von Evaluationsexperimenten durch. Um die Ergebnisse des N-Gram-Tracings besser mit denen von Delta vergleichen zu können, führen wir auf den Deltaabständen zwischen den Texten kein Clustering sondern eine nearest-neighbor-Klassifikation durch, d.h. wir weisen den zu klassifizierenden Text dem Autor des Textes mit dem geringsten Abstand zu. Im Einzelnen handelt es sich um zwei Kürzungs- und zwei Samplingexperimente. Datengrundlage für die Kürzungsexperimente sind die deutschen, englischen und französischen Romankorpora, die unter anderem von Jannidis et al. (2015) und Evert et al. (2017) verwendet wurden. Jedes Korpus besteht aus je drei Romanen von 25 Autoren, also aus 75 Romanen. Für das erste Kürzungsexperiment wird die Größe des Vergleichskorpus stabil gehalten und nur der zu klassifizierende Text gekürzt. Für Delta wird zusätzlich die Anzahl der verwendeten häufigsten Wörter variiert. Im zweiten Kürzungsexperiment werden sowohl der zu klassifizierende Text als auch das Vergleichskorpus gekürzt. Über ein leave-one-out-Verfahren werden alle Texte im Korpus klassifiziert um die Genauigkeit der Verfahren zu ermitteln.

Für die Samplingexperimente verwenden wir eine Sammlung von 1018 deutschen Romanen aus dem langen 19. Jahrhundert. Alle Texte wurden von Muttersprachlern verfasst. Für das erste Samplingexperiment ziehen wir 5000 zufällige Stichproben von 25 Autoren und je drei Romanen (die Zusammensetzung der einzelnen Stichproben ist also vergleichbar mit den oben erwähnten Romankorpora). Für das zweite Samplingexperiment beschränken wir uns auf die 25 Autoren, die in unserer Sammlung mit den meisten Romanen vertreten sind, und ziehen 5000 zufällige Stichproben von je drei Romanen pro Autor (also ebenfalls 25×3 Texte). Für jede Stichprobe ermitteln wir über ein leave-one-out-Verfahren die Genauigkeit der beiden Verfahren.

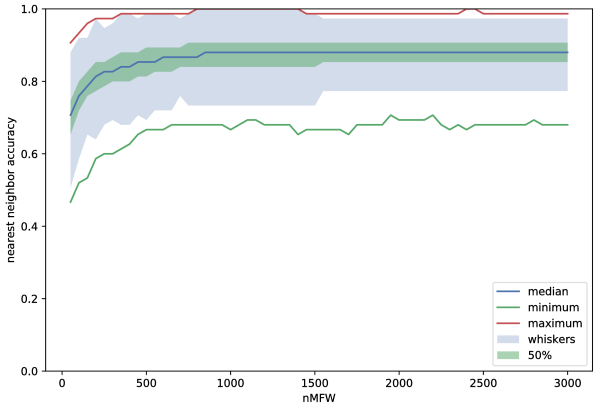
Aus Platzgründen berichten wir an dieser Stelle nur knapp die Ergebnisse des ersten Kürzungsexperiments und des ersten Samplingexperiments und beschränken und dabei auf auf die deutschen Daten. Die Ergebnisse des ersten Kürzungsexperiments, in dem nur der zu klassifizierende Text gekürzt wird, sind in der folgenden Abbildung dargestellt:





Wir vergleichen Cosine-Delta auf Basis der 3000 häufigsten Wörter mit N-Gram-Tracing auf Basis von Wort-1-bis-3-Grammen, Zeichen-4-bis-10-Grammen und Zeichen-11-Grammen. Bis zu einer Textlänge von 5000 Tokens liefern alle N-Gram-Tracing-Varianten bessere Ergebnisse als Delta, für längere Texte funktionieren Delta und N-Gram-Tracing auf Wort-N-Grammen am besten. Bei weniger als 2000 Tokens brechen die Ergebnisse für Delta ein, bei weniger als 1000 Tokens auch die für N-Gram-Tracing.

Die Ergebnisse des ersten Samplingexperiments zeigen, dass die Klassifikationsgenauigkeit bei beiden Verfahren großen Schwankungen unterworfen ist. Hier die Ergebnisse für Cosine-Delta:



Die Grafik zeigt, dass ab ca. den 1000 häufigsten Wörtern zwar im Mittel eine Klassifikationsgenauigkeit von rund 85% erreicht wird, allerdings mit enormen Schwankungen zwischen knapp über 60% und knapp unter 100%.

Die Ergebnisse für N-Gram-Tracing auf Basis von Wort-N-Grammen sehen ähnlich aus:

Durch die Kombination von Wort-1- bis Wort-3-Grammen wird zwar eine mittlere Klassifikationsgenauigkeit von über 70% erreicht, aber auch hier mit enormen Schwankungen.

Die Ergebnisse zeigen, dass N-Gram-Tracing auf kurzen Texten besser funktioniert als Cosine-Delta, allerdings werden für beide Verfahren längere Texte benötigt, als häufig verwendet werden. Die Wahl der Autoren im Vergleichskorpus und auch, wie das Poster zeigen wird, die Wahl der einzelnen Werke haben einen enormen und schwer vorhersehbaren Einfluss auf die Qualität der Autorschaftszuschreibung, deren Genauigkeit ohne weiteres um 20 Prozentpunkte schwanken kann. Im Licht dieser Erkenntnisse ist es durchaus fraglich, wie valide und generalisierbar bisherige Forschungsergebnisse auf dem Gebiet der Autorschaftsattribuierung sind.

Bibliographie

Argamon, Shlomo (2008): „Interpreting Burrows' delta: Geometric and probabilistic foundations“. In: *Literary and Linguistic Computing* 23/2: 131–47. https://doi.org/10.1093/llc/fqn003

Burrows, John (2002): „'Delta'—A measure of stylistic difference and a guide to likely authorship“. In: *Literary and Linguistic Computing* 17/3: 267–87. https://doi.org/10.1093/llc/17.3.267.

Evert, Stefan / Proisl, Thomas / Jannidis, Fotis / Reger, Isabella / Pielström, Steffen / Schöch, Christof / Vitt, Thorsten (2017): „Understanding and explaining Delta measures for authorship attribution.“ In: *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqx023.

Grieve, Jack / Carmody, Emily / Clarke, Isobelle / Gideon, Hannah / Heini, Annina / Nini, Andrea / Waibel, Emily (in Begutachtung): „Attributing the Bixby Letter using n-gram tracing“. Eingereicht bei *Digital Scholarship in the Humanities* am 26. Mai 2017.

Jannidis, Fotis / Pielström, Steffen / Schöch, Christof / Vitt, Thorsten (2015): „Improving Burrows' Delta – An empirical evaluation of text distance measures". In: *Digital Humanities 2015: Conference Abstracts*. http://dh2015.org/abstracts.

Smith, Peter W. H. / Aldridge, W. (2011): „Improving authorship attribution: Optimizing Burrows' delta method". In: *Journal of Quantitative Linguistics* 18/1: 63–88. https://doi.org/10.1080/09296174.2011.533591.

Denkmalpflege in der DDR. Analoge Netzwerke digital – Chancen und Möglichkeiten

Klemstein, Franziska
f.klemstein@gmail.com
Technische Universität Berlin, Deutschland

Die klassische Kunstgeschichte verzichtet noch heute weitestgehend auf die Möglichkeiten, die unsere digitale Welt uns bietet. Zwar werden digitale Werkzeuge bereits vielfältig genutzt, jedoch bisher häufig ohne ausreichende Reflexion und Rückkopplung in die Lehre. [1]

Innerhalb meines Dissertationsprojektes zum Thema „Denkmalpflege zwischen System und Gesellschaft – Netzwerke der Denkmalpflege im Sozialismus" habe ich es mir zum Ziel gesetzt sowohl eine technikgeschichtliche Methode zur Darstellung von Handlungen und Strukturen zu nutzen als auch analoge Netzwerke digital abzubilden.

Die Zielsetzung ist es, zum einen die Komplexität der denkmalpflegerischen Aufgaben abbilden zu können und zum anderen – und dies ist das Ziel des gesamten Dissertationsprojektes – unzutreffende Verkürzungen und Verallgemeinerungen in Bezug auf die Denkmalpflege in der DDR zu vermeiden, in dem die unterschiedlichen Akteure und Zeitphasen im Zusammenhang mit den konkreten Objekten und der Darstellung des Erfolgs oder Misserfolgs der Denkmalpfleger und Denkmalpflegerinnen innerhalb der DDR erfasst, dargestellt und abgefragt werden können. Fragekomplexe, die hierbei in den Blick genommen werden, sind u.a.: Welche Akteure haben an welchen denkmalpflegerischen Projekten gearbeitet oder waren involviert? Welche Bauaufgaben wurden zu welchen Zeiten besonders stark gefördert, diskutiert, unter Schutz gestellt oder zum Abriss freigegeben? Welche Akteure konnten zu welchem Zeitpunkt erfolgreich Belange der Denkmalpflege umsetzen?

Die technikhistorische Methode basiert auf dem von Wolfgang König entwickelten Akteur-Struktur-Modell (ASM), das eine Kombination von Handlungs- und Strukturtheorie darstellt. Innerhalb der Kunstgeschichte und Denkmalpflege fand diese Methode bisher jedoch kaum Beachtung. Die Anwendung dieses Modells innerhalb einer architekturhistorischen Arbeit, soll den Blick auf einen Themenbereich weiten, der bislang häufig nur auf Teilaspekte oder regionale Entwicklungen beschränkt wurde. Das Akteur-Struktur-Modell stellt dabei den Versuch dar, Handlungen und Strukturen strikt symmetrisch zu behandeln, da Strukturen aus Handlungen hervorgehen und Handlungen aus Strukturen. (König 2013a : 514) Dabei wird zwischen verschiedenen Handlungsebenen (Makro-, Meso-, Mikroebene) unterschiedenen. Strukturen stehen hingegen „für Tradition und für Dauer, für soziokulturelle Verfasstheiten, in denen sich die Akteure bewegen und bewegen müssen." (König 2013a : 512) Strukturen bilden somit den Handlungsrahmen oder Spielraum der handelnden Personen (Mikroebene), Organisationen (Mesoebene) oder auch der Regierungen (Makroebene), wobei deren Handlungen bestehende Strukturen sowohl stabilisieren als auch destabilisieren können.

Zugleich sollen mit der Anwendung des Modells auch seine Grenzen und Probleme aufgezeigt werden, die sich ergeben, wenn ein Modell aus einem anderen Wissenschaftsbereich für die Kunstgeschichte nutzbar gemacht wird. Obwohl Wolfgang König auf den Begriff des Netzwerkes verzichtet, möchte ich diesen innerhalb meines Dissertationsprojektes verwenden und folge hierbei Christoph Hubig, welcher vorschlägt, die Dynamik zwischen Akteuren und Strukturen mit Hilfe der Netzwerkmetapher zu modellieren. (König 2013b : 605 und Hubig 2013 : 546f.) Dies erscheint sinnvoll, da die Protagonisten der Denkmalpflege der DDR formale Beziehungen[2] miteinander unterhalten haben, welche die Dynamik innerhalb der scheinbar festen Strukturen, welche das sozialistische System geprägt beziehungsweise festgelegt hat, überhaupt erst möglich werden ließ. In diesem Sinne möchte ich auch die graphbasierte Datenbank neo4j nutzbar machen und den Netzwerkbegriff nicht nur als Metapher verwenden.

Jeder Teil unseres Lebens wird von zahlreichen Verbindungen geprägt, so auch die institutionellen wie auch persönlichen Netzwerke innerhalb