Introduction to Digital Humanities

Manuel Burghardt, Thomas Efer, Andreas Niekler

Computational Humanities Group

In wiefern verarbeitet Kafka seine Beziehung zu seinem Vater in seinen Werken? Wie werden Vater-Sohn-Beziehungen von Kafka dargestellt und welche Parallelen finden sich in seinen Briefen an den Vater? Mithilfe von Topic Modelling werde ich einige Werke Kafkas und die Briefe an seinen Vater hinsichtlich der Vater-Sohn-Beziehung analysieren und dabei durch Konnotationen und stilistische Mittel in Zusammenhang stellen.

Von Sophia Schrödter

qu19orel

1. Einführung

Im Rahmen der digitalen Revolution, welche Ende des 20. Jahrhunderts entstand und sich seither stetig weiterentwickelt, veränderten sich die Wissenschaften und Forschungsmethoden stark. Nach G. Lauer, seien Computer in den Natur- und Lebenswissenschaften schon lange ein großer Bestandteil der wissenschaftlichen Arbeit. Viele Werkzeuge zum Auswerten oder Erstellen von Daten sind heutzutage digital. Diese Veränderungen im Vorgehen der Forschung fanden nicht nur in den Naturwissenschaften statt. Auch in den geisteswissenschaftlichen Disziplinen finden digitale Methoden mittlerweile immer mehr Anwendung.

Die Digital Humanities sind ein interdisziplinäres Forschungsfeld, welches Fragestellungen und Inhalte der Geisteswissenschaften mit digitalen Methoden zur Erstellung, Auswertung und Vermittlung verbindet (nach Burdick et al.). In den traditionellen Geisteswissenschaften ist es sehr schwer große Datenmengen auszuwerten. Der Fokus liegt hier eher auf qualitativen Analysen von einzelnen Werken, was dazu führt, dass vorrangig zu als wichtig anerkannte Werke und Autoren wissenschaftliche Arbeiten entstehen. Die digitalen Methoden, die mit den Digital Humanities nun aufkommen, führen dazu, dass auch große Mengen an Texten auf gestellte Fragestellungen analysiert werden können. Zudem entstehen immer mehr Medien, die Teil des Betrachtungsbereichs der Geisteswissenschaften sind, digital, wovon viele auch nur rein digital verfügbar sind, wie z.B. Videospiele oder Podcasts. C. Roth beschreibt drei verschiedene Wege als die die Digital Humanities gedeutet werden können. Dazu zählen die „numerical humanities“, auch „computational humanities“ genannt, die „digitized humanities“ und die „humanities oft he digital“, welche zusammen mit den „public humanities“ die Forschungsdisziplin Digital Humanities bilden.

In meinem Projekt untersuche ich, inwiefern sich Franz Kafkas Beziehung zu seinem Vater in seinen Werken wiederspiegelt. Dabei ziehe ich einen Vergleich zu den Brief an seinen Vater, welcher nicht fiktional ist und so Kafkas Gefühle seinem Vater gegenüber realitätsgetreu wiedergibt. Diese Fragestellung ist eindeutig den Geisteswissenschaften zuzuordnen, um genau zu sein ist es Teil der Literaturwissenschaften bzw. der Germanistik.

2. Forschungsagenda

Für mein Projekt habe ich mir folgende Forschungsfrage gestellt:

In wie fern verarbeitet Kafka seine Beziehung zu seinem Vater in seinen Werken? Wie werden Vater-Sohn-Beziehungen von Kafka dargestellt und welche Parallelen finden sich in seinen Briefen an den Vater?

Dabei werde ich die Häufigsten Themen im Zusammenhang mit einem Vater, bzw. einer Vater ähnlichen Figur herausarbeiten. Dies werde ich einerseits mit einer Auswahl an Kafkas Werken machen, zusätzlich mit den Werken aus dieser Auswahl, die explizit einen Vater beinhalten, und anschließend mit dem Brief an den Vater. Die Ergebnisse werde ich in Zusammenhang stellen, wobei Konnotationen der Wörter eines Topics Aufschluss über die Gefühle zum Vater geben sollen, um meine Forschungsfrage beantworten zu können.

Die Untersuchung der Vater-Sohn-Beziehung geschieht mit der Programmiersprache R, in welcher ich Topic Modeling auf meine Datensätze anwende. Somit verbinde ich also eine geisteswissenschaftliche Fragestellung mit einer digitalen Methode, wodurch das Projekt den Digital Humanities zuzuordnen ist.

3. Datenübersicht

- Brief an den Vater (Kafka): https://www.projekt-gutenberg.org/kafka/vater/vater.html
- „Betrachtung“, „Der Heizer: Ein Fragment“, „Ein Hungerkünstler“, „Der Mord“, „In der Strafkolonie“ (Kafka): https://www.gutenberg.org/ebooks/search/?query=franz+kafka&submit_search=Go%21
- „Der Prozess“, „Die Verwandlung“, „Das Urteil“ (Kafka): https://www.deutschestextarchiv.de/search?q=franz+kafka&in=text

Die Verwendeten Daten sind allesamt literarische Werke von Franz Kafka, die digitalisiert wurden. Ich habe acht Werke, bestehend aus Novellen, Kurzgeschichten und einer Prosasammlung, ausgewählt, welche ich in einem Ordner namens „texte“ gesammelt habe. Dabei habe ich von jedem Werk die text-Datei verwendet. Den Brief an den Vater habe ich ebenfalls als text-Datei verwendet, diesen jedoch in einem separaten Ordner namens „Brief“ gespeichert. Einen Teil der Texte habe ich von der Website des Deutschem Textarchiv, einen anderen Teil von der Website des Project Gutenberg. Den Brief habe ich von der deutschen Seite des Projekt Gutenberg. Bei den Texten, die ich vom Projekt Gutenberg verwendet habe, habe ich Metadaten zum Verlag, sowie zu Lizenzen ausgelassen, da diese für das Projekt weder zielführend noch relevant wären und somit die Ergebnisse verfälschen könnten. Die Texte stammen aus den Jahren zwischen 1912 bis 1925, womit also Texte von Kafkas ersten Texten bis hin zu seinem Tod 1924 und darüber hinaus verwendet werden. Somit sollten viele Jahre in Kafkas Leben und damit auch sein Empfinden seines Vaters gegenüber abgedeckt sein. Die Daten sind weder annotiert noch ausreichend bereinigt, sodass dies in meinem R-Programm umgesetzt werden muss.

Ich habe keine externe Datensätze für die Lemmatisierung und die Stopwords verwendet, sondern stattdessen die Bibliotheken „quanteda“ für eine deutschsprachige Stopwordliste und „udpipe“ für die deutschsprachige Lemmatisierung.

Da ich nicht alle Texte untersuche, die Kafka geschrieben hat, kann dies durchaus die Arbeit beeinflussen. Da es eine willkürliche Auswahl an Texten ist, ist es weniger mein eigenes Bias der hier Einfluss auf die Ergebnisse hat. Dennoch kann es sein, dass Kafka gerade in den Texten, die ich nicht in meine Datensammlung aufgenommen habe, ein ganz anderes Bild der Vaterfigur darstellt, als sich in meinen Ergebnissen zeigt. Da jedoch vor allem der Vergleich zu der Darstellung des Vaters in Kafkas Brief gezogen werden soll, halte ich die Menge an Texten als ausreichend. Besonders wichtig erachte ich zudem die Texte in dem es eine eindeutige Vaterfigur gibt. Dies sind „Das Urteil“(1912), „Die Verwandlung“(1915) und „Der Brief an den Vater“(1919).

Eine andere Art der Verfälschung die auftreten könnte ist, dass andere autoritäre Strukturen den Vater indirekt darstellen könnten, was sich jedoch in der Untersuchung nicht direkt erkennen ließe.

4. Methodenübersicht

Als Methode zur Auswertung meiner Daten verwende ich Topic Modeling. Topic Modeling identifiziert Themenfelder eines Korpus an Texten, welche durch Wort-Kookkurrenzen Begriffe sammeln, die häufig im Zusammenhang miteinander auftreten. Ein Thema (Topic) beinhaltet also eine bestimmte Anzahl an Wörtern, die in einer vom Autor bewussten oder unbewussten Verbindung miteinander stehen. Dies ist besonders für große Textmengen nützlich, bzw. meist auch erst auf diese Anwendbar, da eine gewissen Häufigkeit an Wiederholungen von Wörtern nötig ist. Topic Modeling erzeugt probabilistische graphische Modelle, die auf mathematischen Berechnungen basieren und schließlich graphisch für den Nutzer dargestellt werden.

David Blei beschreibt es so, dass mit Topic Modeling latente semantische Strukturen aus einem Text ermittelt werden können.

Dabei ist „Latent Dirichlet Allocation“ (LDA) das am häufigsten verwendetet Modell für Topic Modeling, welches auf dem Bag-Of-Words-Modell basiert, was also bedeutet, dass die Reihenfolge der Wörter irrelevant ist. Dieses Modell verwende ich auch für meine Analyse. Für das Topic Modeling verwende ich also die Bibliotheken „topicmodels“ und „LDAvis“.

Durch die Anwendung von Topic Modeling für mein Projekt kann ich verschiedene Themenfelder darstellen lassen und so explizit nach den Themen suchen, welche einen Vater bzw. eine ähnliche Bezeichnung beinhalten. Dadurch kann ich mir anschauen, welche Wörter oder Themenbegriffe Kafka am häufigsten mit einem Vater in Verbindung setzt. Dies kann ich einerseits für meine Texte Sammlung, sowie für den Brief an den Vater machen. Zusätzlich werde ich dies noch einmal mit den Texten machen, die eine explizite Vaterfigur beinhalten.

Probleme die durch die Verwendung von Topic Modeling auftreten könnten, wären einerseits, dass Kafka die Vaterfiguren nicht immer mit dem Begriff Vater benennt, sondern evtl. mit bestimmten

Namen, oder als Herr o.ä.. Hier würden also Wort-Kookkurrenzen nicht alle unter dem gleichen Namen erzeugt werden und dadurch evtl. gar nicht in ein Topic mit aufgenommen werden. Dies könnte dazu führen, dass in meiner Analyse weniger Topics zur Vaterfigur aufkommen, als eigentlich zustande kommen könnten.

5. Verwandte Arbeiten

In ihrer Arbeit „Topic Modeling and Figurative Language“ befasst sich Lisa M. Rhody damit, inwiefern sich Topic Modeling dafür eignet, figurative Sprache in Gedichten zu analysieren. Die prüft dabei, ob man mit Topic Modeling Themen ermitteln kann, die Aufschluss auf Metaphern und bildliche Sprache geben können. Dabei prüft sie dies anhand einer Sammlung an Gedichten, auf diese sie Topic Modeling anwendet.

Ein weiterer Artikel in dem Topic Modeling als Methode verwendet wurde, ist die Arbeit von David Mimnos unter dem Namen „Computational Historiography: Data Mining in a Century of Classics Journals“. Dieser untersucht anhand digitaler Methoden, wie z.B. Topic Modeling, wie sich die bestimmten Inhalte und Themen der klassischen Altertumswissenschaft über die Jahre verändert haben. Dabei fragt sich auch er, ob Topic Modeling eine geeignete Methode ist um diese Inhalte darzustellen und zu ermitteln. Er analysiert dazu Artikel der klassischen Philologie und Archäologie die über einen Zeitraum von hundert Jahren veröffentlicht wurden.

Auch der Artikel „Topic Modeling and Digital Humanities“ von David M. Blei, den ich zuvor schon referenziert habe, beschäftigt sich mit der Verwendung von Topic Modeling in den Digital Humanities. Dabei stellt sich Blei darin die Frage, inwiefern Topic Modeling dazu geeignet ist, große Textmengen zu analysieren. Dabei erforscht er, welche Muster und Themenbereiche sich entdecken lassen und im Besonderen, wie sich die Methode zum klassischen Close Reading der traditionellen Geisteswissenschaften unterscheidet. Dabei wendet Blei Topic Modeling nicht direkt auf Datensätze an, gibt jedoch Beispiele inwiefern dies verwendet werden könnte, beispielsweise die Analyse von historischen Texten.

Topic Modeling wird in allen Artikeln als Methode zum Ermitteln von Themenfeldern und draufbasierender tiefergehende Analyse von größeren Sammlungen an Texten angewendet. Im Gegensatz zu meinem Datensatz sind die hier verwendeten meist größer im Umfang. Dennoch zielen sie mit der Verwendung von Topic Modeling auf die gleiche Art an Ergebnisse ab, wie ich in meiner Untersuchung.

6. Versuchsaufbau

Zu Beginn meines Versuches habe ich die schon zuvor genannten Bibliotheken „quanteda“, „topicmodels“, „LDAvis“, sowie „Rtsne“ und „udpipe“ heruntergeladen. Dabei habe ich für „udpipe“ das Language Model für Deutsch verwendet.

Für die Textvorverabeitunf habe ich zwei verschiedene textdata Variablen erstellt, eine die die Text-Datei für den Brief an den Vater speichert und eine, die die Sammlung der verwendeten Werke, ausgeschlossen des Briefes, speichert. Durch Auskommentieren kann ich dadurch mein Programm entweder auf den Brief, oder die Texte Sammlung anwenden. Dies hätte ich auch machen können indem ich zwei verschieden Variablen erstellt hätte, jedoch hätte ich dann die weiteren Schritte doppelt auf beide Variablen anwenden müssen. Zunächst tokenisiere ich die Datensätze mit „quanteda“, wobei Satzzeichen, Symbole und Zahlen entfernt werden. Anschließend verwende ich „udpipe“ dazu um eine Lemmatisierung durchzuführen, sodass alle Wörter einheitlich in ihre jeweiligen Grundformen umgewandelt wurden. Auch Stopwords werden mit „quanteda“ entfernt. Anschließend verwende ich textstat_collocations, damit Häufige Wortkombinationen ermittelt werden. Hierbei habe ich 100 als Anzahl für die Häufigsten Kollokationen verwendet. Tokens_compound wird nun angewendet um Mehrwertausdrücke als Einheit zu betrachten. In diesem Schritt kam das Problem auf, dass es im Brief an den Vater NA-Werte gab, die noch bereinigt werden mussten, sodass ich hier noch einen extra Schritt nur für den Datensatz ausführen musste.

Anschließend führe ich die Model Calculation durch. Hierbei wird eine Dokument-Term-Matrix (DTM) erstellt. In dieser entferne ich häufig vorkommende Wörter, wie z.B. „er“ oder „sagen“.

Auf dieser Basis wird im nächsten Schritt die LDA mit 20 Themen trainiert. Hier musste ich den Wert auf die Größe meiner Textsammlung anpassen, sodass es genug Themen für die Erstellung gibt, aber auch nicht zu wenige erstellt werden, die möglicher Weise nicht relevant sind.

Zur Visualisierung der Themen habe ich nun eine json-Datei erstellt. Hierfür wollte ich zuerst die Bibliothek „tsne“ verwenden, welche allerdings von R nicht mehr aktiv unterstützt wird. Ich habe stattdessen auf die Bibliothek „Rtsne“ zurückgegriffen.

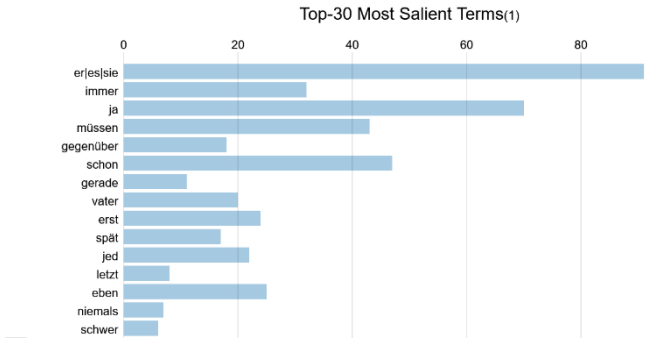
Im letzten Schritt habe ich die ermittelten Topics gefiltert, indem ich die Topics rausgesucht habe, welche beispielsweise den Begriff „Vater“ beinhalten.

Probleme bei der Verwendung von Topic Modeling hier könnten sein, dass Wörter die den Begriff Vater umschreiben oder eher unbekannte Synonyme sind, nicht dem Begriff Vater in den Topics zugeordnet werden könnten. Dadurch würden hier Zusammenhänge außeracht gelassen werden, die für die Analyse relevant gewesen wären, oder das dies die Ergebnisse verfälscht, da der Begriff Vater in anderen Kontexten als ein Begriff der nicht erkannt wurde. Außerdem könnte es passieren das durch die Stopword Entfernung relevante Begriffe fälschlicherweise entfernt wurden. Zudem ist die Anwendung

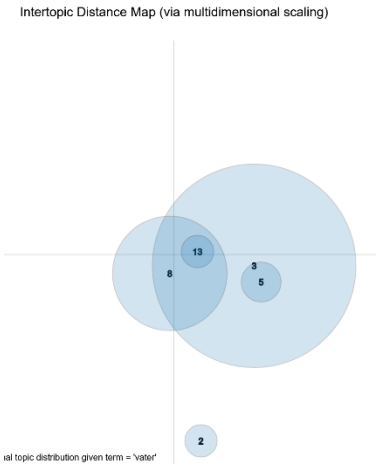
von Topic Modeling auf einen einzelnen Text eventuell nicht die beste Methode, da zu wenige Kollokationen zwischen Wörtern entstehen könnten um eine zuverlässige Aussage auf die Bedeutung und Gewichtung der Zusammenhänge treffen zu können.

7. Ergebnisse und Diskussion

Bei der Anwendung des Codes auf den „Brief an den Vater“ wird schnell deutlich, dass Vater mit unter den häufigsten Verwendeten Wörtern ist. Davor stehen nur Wörter, welche wenig Aussagen, bzw. keine bestimmte Konnotation mit sich tragen. Dies lässt sich ebenfalls in der Texte Sammlung erkennen, bei der der Begriff „Vater“ an fünfter Stelle steht. Der Vater scheint also durch Kafkas Texte hinweg häufig eine zentrale Figur zu sein. Im folgenden habe ich mir ermitteln lassen in, welchem Topic Vater vorkommt. Dabei wurde mir die Topics 14 und 20 ausgegeben. In der json-Datei stachen zu Topic 20 besonders die Begriffe „Rettung“, „Selbstvertrauen“, „Undankbarkeit“ und „Stärke“ hervor. In der Visualisierung des Topics traten Begriffe wie „Einzelheit“, „Mensch“, „allein“ und „Hass“ hervor. Viele dieser Begriffe haben eine negative Konnotation und kamen in erhöhter Kollokation mit dem Begriff „Vater“ auf. Allgemein enthielten viele der Topics besonders negativ konnotierte Wörter. Dies steht im Zusammenhang zum Inhalt des Briefes, in dem Kafka die Erziehung seines Vaters kritisiert und ihm gegenüber seine Gefühle wie Unterlegenheit und Angst darstellt.



Im zweiten Teil meiner Analyse habe ich meinen kompletten Datensatz durch das Programm auswerten lassen. Besonders häufig kam der Begriff Vater in den Topics 2, 3, 8, 13 und 14 vor, welche untereinander auch viele Überschneidungen vorweisen. Bei genauerer Betrachtung dieser Topics lassen sich jedoch nur wenige direkt aussagekräftigen Begriffe erkennen, welche bestimmte Konnotationen im Bezug auf den Vater erkennen lassen würden. In Topic 13 steht der Begriff „Vater“ besonders im Zusammenhang mit „Schwester“, „Mutter“, „Familie“ und „Eltern“. Bei Topic 8 hingegen findet man Begriffe wie „darunter“, „angewiesen“, „Organisation“ und „verfolgen“.



Hier lassen sich wiederrum negative Konnotationen erkennen. Die Begriffe „darunter“ und „angewiesen“ zeigen eine Unterlegenheit, die Kafka hier im Bezug zum Vater stellt und seine Vorwürfe aus dem Brief widerspiegeln könnten. Der Begriff „Organisation“ kann ebenfalls als übergeordnete Kraft gedeutet werden, die ein bestimmtes Machtverhältnis repräsentiert. Auch „verfolgen“ kann ich Kontext von den Erwartungen des Vaters an Kafka gedeutet werden.

Die anderen Topics weisen jedoch keine relevanten Begriffe auf, die parallelen zu Kafkas Brief an seinen Vater aufweisen würden. Daher habe ich letztendlich noch einmal eine kleinere Textsammlung erstellt, welche aus den Texten „Das Urteil“ und „Die Verwandlung“ besteht, in denen es eine eindeutige Vaterfigur gibt. Hiermit wollte ich überprüfen, ob hier die Vaterfiguren stärker auf eine bestimmte Weise beschrieben werden, bzw. in bestimmten Kontexten häufiger vorkommen. Hier ließ sich direkt auch wieder deutlich erkennen, wie bedeutsam die Figur des Vaters in Kafkas Werken ist. Nach „Gregor“ ist der Begriff „Vater“ das häufigste relevante Wort in den erstellten Topics. Behandelt wird der Vater jedoch nur in den Topics 1, 2, 3, 8 und 15. Topic 1 beinhaltet wieder die Begriffe rund um Familie. Topic 8 hingegen beinhaltet Begriffe die die Forschungsfrage stützen könnten.

Begriffe wie „groß“ und „klein“ zeigen einen deutlichen Widerspruch, der das Machtverhältnis zwischen Kafka und seinen Vater zeigen könnten. „Sollen“, „müssen“ und „lassen“ sind wiederrum Befehle, die das Handeln einer Person beschreiben. Dennoch zieht sich auch hier das Familienthema durch und viele Worte sind wenig, bis gar nicht aussagend.

https://github.com/sophiascer/DH_Projekt

8. Fazit

Abschließend lässt sich sagen, dass nur wenige Anhaltspunkte mithilfe von Topic Modeling gefunden werden konnten, die meine Forschungsfrage bestätigen würden. In manchen Hinsichten taucht der Vater in häufiger Verbindung mit anderen Worten auf, die negativ konnotiert sind. Zwar ist der Vater immer eine wichtige Figur und kommt häufig in den Topics vor, dennoch oft im familiären Kontext.

Die Werke von Kafka scheinen nicht umfangreich genug zu sein, um eine zuverlässige und aussagekräftige Analyse mithilfe von Topic Modeling durchzuführen. Da eine väterliche Figur nicht in allen Werken als „Vater" benannt wird, sondern häufig in Form von Übergeordneten Instanzen auftritt, bekommt man mit der Analyse nach dem Begriff „Vater" zu wenige Kollokationen mit anderen Wörtern. Vermutlich würde es helfen in der Lemmatisierung alle bekannten väterlichen Strukturen aus Kafkas Werken unter einem Begriff zu vereinen, jedoch wäre hierfür eine genau Kenntnis mit allen von Kafka geschriebenen Texten notwendig, was nur allein durch Topic Modeling nicht möglich ist herauszufinden.

Im „Brief an den Vater" konnten eindeutig negative Konnotationen im Bezug auf den Vater festgestellt werden. Generell unterliegt der Brief einem eher traurig klingenden Ton.

Die Frage, inwiefern Kafka seine Beziehung zu seinem Vater in seinen Texten verarbeitet, kann also nur so weit beantwortet werden, als dass der Vater eine wichtige Rolle in vielen Werken einnimmt. Dieser tritt häufig im familiären Ramen auf und ist teilweise negativ konnotiert, taucht jedoch auch in eher irrelevanten Kontexten auf. Der Brief an den Vater hingegen zeigt deutlicher die Vater-Sohn-Beziehung auf.

9. Referenzen

Lauer, G. (2013). Die digitale Vermessung der Kultur. *Geiselberger, H., Moorstedt, T. Big Data: Das neue Versprechen der Allwissenheit.* Berlin: Suhrkamp.

Burdick, A., et al. (2012). Digital_Humanities. *MIT Press.*

Roth, C. (2019). Digital, digitized, and numerical humanities. *Digital Scholarship in the Humanities,* 34(3), 616-632.

Blei, D. (2012). Topic Modeling and Digital Humanities. *Journal of Digital Humanities.* Available from http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/

Rhody, L. (2012). Topic Modeling and Figurative Language. *CUNY Academic Works.* Available from https://academicworks.cuny.edu/cgi/viewcontent.cgi?article=1557&context=gc_pubs

Mimno, D. (2012). Computational Historiography: Data Mining in a Century of Classics Journals. *Journal on Computing and Cultural Heritage (JOCCH), Volume 5, Issue 1.* Available from https://dl.acm.org/doi/pdf/10.1145/2160165.2160168

Weitere Quellen:

https://stackoverflow.com/questions/65664123/lemmatization-of-german-words-capital-letters-and-lower-case-letters