Universität Basel | Departement Geschichte | FS2024

Übung: *Grand Tour gone digital. Computergestützte Analyse frühneuzeitlicher Reiseberichte*

Leitung: Ina Serif

Arbeitsbericht von Stefan Heinen, Anna Scognamiglio und Nora Holländer

Eingereicht am 27.08.2024

Quantitative Textanalyse frühneuzeitlicher Reiseberichte

1. Einleitung

Die Leitfrage dieses Projekts war, ob und wie sich eine vergleichende quantitative Textanalyse gewinnbringend für die Untersuchung eines historischen Textkorpus einsetzen lässt. Wir wollten herausfinden, ob sich die Texte im Hinblick auf Autorschaft, Entstehungszeit und Entstehungskontext in ihrer Wortwahl und Ausdrucksweise signifikant unterscheiden. In Bezug auf das Genre der frühneuzeitlichen Reiseberichte stand hier insbesondere die Frage im Zentrum, wie die Stationen der beschriebenen Reisen – die von den Autoren besuchten Städte und Ortschaften – sowie deren Zweck – sind die Autoren Forschungsreisende, politische Gesandte oder Abenteurer? – die Wortwahl beeinflussten. Darauf aufbauend stellten wir uns in methodologischer Hinsicht die Frage, was die Ergebnisse einer solchen Untersuchung für eine weitergehende qualitative Analyse eines historischen Textkorpus bedeuten könnten. Lassen sich mit den angewandten Methoden bereits sinnvolle Aussagen über den Inhalt der einzelnen Texte treffen? Oder dient eine solche «Vogelschau» mittels *Distant Reading* eher dem Zweck, gewisse Anhaltspunkte in den Texten zu finden, auf Grundlage derer eine Vertiefung stattfinden könnte? Zur Beantwortung dieser Fragen stellten wir in einem ersten Schritt ein Textkorpus zusammen, welches wir dann für einen maschinellen Vergleich aufbereiteten. Im zweiten Schritt folgte die Untersuchung des Korpus mittels verschiedener Methoden der quantitativen Textanalyse.

2. Textgrundlage

Als Ausgangskorpus wählten wir die Reiseberichte des Augsburger Kaufherrn Philipp Hainhofer (1578 – 1647). Diese sind in handschriftlicher Form überliefert und werden im Rahmen eines digitalen Editionsprojektes der *Herzog August Bibliothek* digitalisiert und transkribiert. Die Transkripte werden auf der Webseite des Projekts als TXT-Dateien frei zur Verfügung gestellt.[[1]](#footnote-1) Philipp Hainhofer war Kunstsammler, Kunsthändler und Agent und arbeitete in dieser Rolle für verschiedene Herzogen. Er unternahm zahlreiche Reisen, zumeist in diplomatischem Auftrag, die ihn hauptsächlich nach Süddeutschland, aber auch nach Österreich führten.[[2]](#footnote-2) Für die vorliegende Arbeit wurden elf Einzeltexte analysiert, die während seiner Reisen zwischen den Jahren 1603 – 1636 entstanden sind. Im Konkreten handelt es sich um fünf Reiseberichte aus München, zwei aus Neuburg jeweils einen aus Nürnberg und Regensburg und einen, der von Wildbad über Heidelberg nach Durlach führt. Die Sammlung enthält noch weitere Texte, die bis ins Jahr 1594 zurückgehen. Diese werden noch bearbeitet und schrittweise der Öffentlichkeit zugänglich gemacht, standen zum Zeitpunkt der Projektarbeit jedoch noch nicht zur Verfügung.

Neben einem Vergleich der einzelnen Texte von Hainhofer sollte ein Vergleich mit weiteren Texten erfolgen, um die quantitativen sowie qualitativen Ergebnisse der angewandten Methoden besser einordnen zu können. Daher wurden im weiteren Prozess die Reiseberichte von drei zusätzlichen Autoren ausgewählt und die Texte Hainhofers gesammelt als vierter Vergleichskorpus betrachtet. Es handelt sich dabei um die Reiseberichte von Georg Christoph von Neitzschitz (ca. 1600 – 1636), der von 1630 bis 1636 als Forschungsreisender den Orient bereist hatte, und kurz nach seiner Rückkehr verstarb[[3]](#footnote-3), dem Zürcher Kunsttheoretiker und Pädagogen Johann Georg Sulzer (1720 – 1779)[[4]](#footnote-4), und dem Magdeburger Schriftsteller Friedrich Schulz (1762 – 1798).[[5]](#footnote-5) Im Gegensatz zu den Handschriften Hainhofers handelt es sich hier um Werke, die im Druck erschienen waren. Sie liegen im *Deutschen Textarchiv (DTA)* in digitalisierter Form sowie als TXT-Dateien vor. Neben dem Genre und der digitalen Verfügbarkeit dienten als Auswahlkriterien erstens die Länge der Texte, da uns ein vergleichbarer Datenumfang angesichts einer quantitativen Datenanalyse wichtig erschien, um Verzerrungen in den Ergebnissen weitestgehend zu vermeiden. Zweitens sollte eine zeitliche Distanz des Entstehungszeitraumes einen historischen Vergleich ermöglichen. Mit den Texten / Textsammlungen von Hainhofer und Neitzschitz, sowie von Sulzer und Schulz bestand das Korpus schliesslich aus vier Quellen, von denen jeweils zwei bezüglich ihrer Entstehungszeit relativ nahe beieinander liegen.

3. Vorgehen

-> Hier könnte man allenfalls eine kurze Zusammenfassung des ‘Projektdesigns’ voranstellen:

*Programmierung mit Python / Präsentation mit Jupyter Notebook*

*Aufbereitung der Texte für einen maschinellen Vergleich*

*Auswahl der Verfahren:*

*Statistische Gewichtung (Tf-idf-Mass)*

*Vergleich absoluter Worthäufigkeiten*

*-> Stoppwörter*

*Natural language processing (spaCy)*

*Trainiertes Spachmodell (‘de\_core\_news\_lg’)*

*Visualisierung der Ergebnisse*

* *Was wir vielleicht tatsächlich erwähnen sollten, ist, dass die einzelnen Schritte jeweils im publizierten Notebook nachvollzogen werden können.*

Als erste Analysemethode wählten wir die Kalkulation des Tf-idf-Masses für die einzelnen Wörter im Korpus. Tf-idf steht für «*term frequency* (‚Vorkommenshäufigkeit‘)– *inverse document frequency* (‚inverse Dokumenthäufigkeit‘)». Mit dieser Methode werden Wortvorkommnisse innerhalb eines Textkorpus statistisch gewichtet. So kann die relative Häufigkeit von Wörtern in Dokumenten errechnet werden. Als Ergebnis einer Tf-idf-Analyse erhält man einen Wert pro Wort, der angibt, wie spezifisch dieses Wort für ein Dokument ist, verglichen mit anderen Dokumenten im Korpus. Kommt also ein Wort in einem Dokument sehr häufig vor, im ganzen Korpus jedoch nur sehr selten, hat es einen hohen Tf-idf-Wert im jeweiligen Dokument. Kommt ein Wort in einem Dokument zwar sehr häufig vor, aber auch in allen anderen Dokumenten des Korpus, dann hat es einen tieferen Tf-idf-Wert. Somit kann herausgefunden werden, welche Wörter charakteristisch für ein Dokument sind, während häufig vorkommende, nicht-bedeutungstragende Wörter, wie Funktionswörter oder Partikel, nicht ins Gewicht fallen.

Zur Vorbereitung haben wir als Erstes den Blogbeitrag zu Tf-idf von Programming Historian gelesen und uns mit dem dort zur Verfügung gestellten Code vertraut gemacht.[[6]](#footnote-6) Nachdem wir das Jupyter Notebook und Python installiert hatten, haben wir ein Programm erstellt, mit dem alle Quellen einheitlich in ein TXT-Format gebracht und in einer CSV-Datei zusammgenführt wurden. Auf diese Weise sind Analysen mit Python-Modulen leicht durchführbar. Zunächst haben wir uns auf die Hainhofer-Texte konzentriert und die Tf-idf-Werte für jedes Wort von allen elf Texten des Hainhofer-Korpus errechnet. Als Ergebnis erhielten wir eine Liste/CSV-Tabelle je Dokument, welche die Wörter nach Tf-idf-Wert geordnet aufgelistet hat. Da die elf Texte elf unterschiedlichen Reisen entsprechen, die sich hinsichtlich des Zeitpunkts, des Anlasses und z.T. der Destinationen unterscheiden, wurde erwartet, dass sich diese Differenzen in unterschiedlichen Tf-idf-Werten ausdrücken. In der Tat wurden für alle Reiseberichte verschiedene Wörter als besonders spezifisch für den jeweiligen Text identifiziert, wodurch sich ein eigenes Wörter- bzw. Tf-idf-Profil pro Dokument darstellte. Es gab jedoch auch einzelne Wörter, welche nicht aussagekräftig waren wie «euer» oder «bei». Bei der Tf-idf-Methode wird die Relevanz von Wörtern so gewichtet, dass häufig auftretende, wenig bedeutungstragende Wörter einen niedrigeren Wert erhalten und somit als wenig relevant eingestuft werden. Da nun dennoch solche bedeutungsschwachen Wörter z.T. hohe Werte erhalten haben, könnten diese mit manuell definierten Stop-Word-Listen herausgefiltert werden, damit sie in der Berechnung nicht berücksichtigt werden. Dies wurde für die weiterführende Vergleichsanalyse der vier verschiedenen Autoren berücksichtigt und angewendet.

Wir haben uns nun etwas genauer mit den Quellen selbst und deren Inhalt auseinandergesetzt, also das Distant Reading durch ein Close Reading ergänzt. Hierfür wurden für jedes Dokument die 20 Begriffe mit den höchsten Tf-idf Werten extrahiert und mit der online zur Verfügung stehenden Kurzzusammenfassung des jeweiligen Reiseberichtes von Hainhofer verglichen. Wir wollten herausfinden, wie repräsentativ die errechneten Tf-idf-Werte und die ausgegebene Spezifität der Begriffe für jedes Dokument sind. Dabei konnte festgestellt werden, dass die ermittelten Begriffe durchaus als semantisch richtungsweisend anzusehen sind. Im Reisebericht von München 1613 beispielsweise geht Hainhofer nach München, um eine Hochzeit zu besuchen. Unter den ersten sieben Wörtern befinden sich «ehestand, braut, bräutigam, breutigam, hochzeiterin» Auch wenn die Liste die Zusammenfassung nicht ersetzt, wird das Thema also in der Tf-idf-Liste semantisch gut fassbar. Ein solches Profil ist bei allen Wortlisten erkennbar, auch wenn die Qualität unterschiedlich zu sein scheint. Zum Teil tauchen Wörter auf, die wenig bedeutungstragend sind.

In einem nächsten Schritt haben wir die drei weiteren Reiseberichte hinzugefügt, um Vergleichswerte ermitteln zu können.

*Da sich die bei den Texten aus dem Deutschen Textarchiv die Repräsentation der Druckschrift in den publizierten TXT-Dateien wesentlich von derjenigen der Handschriften Hainhofers unterschied, mussten diese zuerst angeglichen werden, um einen sinnvollen maschinellen Vergleich zu ermöglichen. Weiter wurde das Textmaterial von Hainhofer und Sulzer, das aus mehreren Dateien bestand, jeweils zu einer TXT-Datei zusammengefasst.*

* *Wie konkret steht im Notebook, weiss nicht ob wir das hier nochmal erklären müssen?*
* Ab hier wird’s irgendwie unverständlich, bzw. werden die einzelnen Schritte auch nicht mehr Korrekt beschrieben. Bisschen ergänzen bezüglich spaCy und der generierten Resultate könnte man wohl auch noch:

Wir haben Stoppwortlisten errechnet und für jedes Wort (Lemma) die Wortgruppe bestimmen lassen. Auf diese Weise konnten wir anschliessend alle Adjektive sowie Nomen herausfiltern und einer Ähnlichkeitsanalyse unterziehen. Hierfür haben wir Tf-idf Listen nur mit den Adjektiven bzw. Nomen für jedes der vier Dokumente erstellt und diese mit Balkendiagrammen visualisiert.

Parallel dazu haben wir den gesamten Korpus mithilfe von Voyant Tools etwas genauer betrachtet. Um die Bearbeitung in Voyant zu erleichtern, haben wir dafür auch manuell eine Stoppwortliste angelegt. Auch in Voyant haben wir eine Tf-idf Liste und andere interessante Werkzeuge entdeckt, die zur Verfügung gestellt werden. Bei den Stoppwortlisten ist jedoch nicht transparent, wie genau die Werte zustandekommen und die Liste scheint ungenauer zu sein.

Auch die weiteren Tools von Voyant halfen dabei, einen groben Überblick zu erhalten, aber auch sie verhalfen uns nicht zu einer konkreten Fragestellung.

* Würd ich streichen…

4. Schwierigkeiten

Während der Projektarbeit haben sich uns verschiedene Probleme/ Herausforderungen ergeben, welche aber größtenteils lösbar waren. Der erste Themenkomplex war die Aufarbeitung der Texte. Beispielsweise kam manchmal dasselbe Wort mit unterschiedlicher Schreibweise in den Listen mehrmals vor. Als Lösung für dieses Problem haben wir eine Lemmatisierung durchgeführt.

Ein anderes Problem dieser Kategorie war, dass auch viele irrelevante Wörter wie «einem» oder «bej» im ersten Teil der Tf-idf-Listen vorkamen. Um diesem Problem entgegenzuwirken, haben wir Stoppwortlisten erstellt.

*Nicht zuletzt stellte uns die zeitspezifischen Ausdrücke und Schreibweisen vor Probleme; zumindest das an modernem Deutsch trainierte Sprachmodell war für die historischen Texte nur bedingt geeignet.*

Die Qualität der einzelnen Listen hängt stark vom Gesamtkorpus ab. Eine der Listen war unbrauchbar, da das Dokument auch einen italienischen Brief enthielt und Italienisch in den anderen Berichten gar nicht vorkam. Somit wurden diese italienischen Begriffe nicht als «irrelevant» identifiziert und daher nicht herausgefiltert. Die ermittelten Werte sind in diesem Fall unbrauchbar, da hier die relative Häufigkeit äquivalent ist mit der absoluten Häufigkeit der Wörter. Dies kann damit gelöst werden, dass das Dokument übersetzt oder ausgelassen wird.

* *Umformulieren oder streichen. Siehe Kommentar von Ina.*

Es stellte sich als sehr schwierig heraus, von einer Methode zu einer Fragestellung zu gelangen. In einem nächsten Projekt würden wir von einer Fragestellung ausgehen und uns dann in einem zweiten Schritt überlegen, mithilfe von welchen Techniken und Methoden diese Fragestellung beantwortet werden kann.

* *Würd ich streichen: Die Fragestellung ist halt eher Methodologischer Natur, ist oben in der Einleitung gewissermassen als ‘Leitidee’, die dem Projekt ja durchaus zugrunde lag, ausformuliert.*

Trotzdem haben wir im Verlauf dieses Projekts sehr viel gelernt und sind nun viel vertrauter mit Tf-idf und den notwendigen Parametern und Aufbereitungsschritten. Wir können uns nun besser vorstellen, für welche Projekte Tf-idf sinnvoll sein kann. Es ist sicherlich ein gutes Tool, um einen ersten Eindruck eines Textkorpus zu erhalten.

* *Ist gar allgemein gehalten. An diesem Punkt geht es ja nicht mehr nur um Tf-idf, sondern um die Analyse als Ganzes, wo wir ja verschiedene Methoden kombiniert haben.*

Allenfalls würde dafür auch die Tf-idf Funktion von Voyant Tools ausreichen.

* *Für was denn die ganze Arbeit? Finde wir konnten das schon gut präsentieren mit dem Notebook.*

Für eine weitergehende Arbeit mit Tf-idf ist es sicherlich sinnvoll, selbst Code zu schreiben, um Parameter individuell und gezielt anpassen zu können.

* *Haben wir ja genau gemacht*.

5. Fazit

Die angewandten Methoden geben einen guten ersten Einblick und wären auch für grössere Korpora anwendbar. Um tatsächlich zu gehaltvollen Resultaten zu kommen, müsste man allerdings «tiefer rein» in die Textstruktur, oder aber diese mit *Close Reading* verbinden.

* *Gerne Ergänzen. Im Notebook hab ich schon einige (Zwischen-) Ergebnisse ausformuliert, die kann man hier sonst auch übernehmen*.

*Vergleich von Wortgruppen (n-grams) könnte spannend sein. Also in welchem Kontext werden welche Wörter gewählt? Noun-chunks funktioniert eher schlecht mit spaCy (bis gar nicht, mit den älteren Texten. Vielleicht funktioniert das aber auch nur auf Englisch richtig).*

*Will heissen, das Repertoire ist noch nicht ausgeschöpft. Ich sehe auf diesem Weg schon Möglichkeiten für eine tiefergehende Analyse. Insofern bin ich der Meinung, dass wir durchaus zu passablen Resultaten gelangt sind.*

6. Bibliografie

5.1 Quellen

Hainhofer, Philipp: Reiseberichte und Sammlungsbeschreibungen 1594–1636. Edition und Datensammlung zur Kunst- und Kulturgeschichte der ersten Hälfte des 17. Jahrhunderts [Wolfenbütteler Digitale Editionen, Nr. 4], hrsg. und eingeleitet von Michael Wenzel, Transkription und Kommentar von Ursula Timann und Michael Wenzel, Wolfenbüttel: Herzog August Bibliothek 2020ff.

Neitzschitz, Georg Christoph von: Sieben-Jährige und gefährliche WeltBeschauung Durch die vornehmsten Drey Theil der Welt Europa/ Asia und Africa. Bautzen, 1666. In: Deutsches Textarchiv <https://www.deutschestextchiv.de/neitschitz\_reise\_1666>, abgerufen am 19.07.2024, URN: urn:nbn:de:kobv:b4-20878-1.

Schulz, Friedrich: Reise eines Liefländers. 3 Bde. Berlin, 1795.

Bd. 1, H. 1: <https://www.deutschestextarchiv.de/schulz\_reise0101\_1795>, abgerufen am 19.07.2024, URN: urn:nbn:de:kobv:b4-200905197358.

Bd. 1, H. 2: <https://www.deutschestextarchiv.de/schulz\_reise0102\_1795>, abgerufen am 19.07.2024, URN: urn:nbn:de:kobv:b4-17297-4.

Bd. 2, H. 3: <https://www.deutschestextarchiv.de/schulz\_reise0201\_1795>, abgerufen am 19.07.2024, URN: urn:nbn:de:kobv:b4-200905197368.

Bd. 2, H. 4: <https://www.deutschestextarchiv.de/schulz\_reise0202\_1795>, abgerufen am 19.07.2024, URN: urn:nbn:de:kobv:b4-17298-9.

Bd. 3, H. 4 u. 5: <https://www.deutschestextarchiv.de/schulz\_reise03\_1795>, abgerufen am 19.07.2024, URN: urn:nbn:de:kobv:b4-200905197374.

Sulzer, Johann Georg: Tagebuch einer von Berlin nach den mittäglichen Ländern von Europa in den Jahren 1775 und 1776 gethanen Reise und Rückreise. Leipzig, 1780. In: Deutsches Textarchiv <https://www.deutschestextarchiv.de/sulzer\_reise\_1780>, abgerufen am 19.07.2024, URN: urn:nbn:de:kobv:b4-25231-5.

5.2 Tutorials

Matthew J. Lavin: "Analyzing Documents with TF-IDF," Programming Historian 8 (2019), <https://doi.org/10.46430/phen0082>.

Die zahlreichen Beiträge auf *Stack Overflow* (<https://stackoverflow.com/>) die massgeblich zum gelingen dieses Projektes beigetragen haben, können an dieser Stelle leider nicht einzeln gelistet werden. Unser Dank geht an die Python-Community!

5.3 Software

-> kann man allenfalls auch weglassen. Steht ja im Notebook…

1. Philipp Hainhofer: Reiseberichte und Sammlungsbeschreibungen 1594–1636. Edition und Datensammlung zur Kunst- und Kulturgeschichte der ersten Hälfte des 17. Jahrhunderts [Wolfenbütteler Digitale Editionen, Nr. 4], hrsg. und eingeleitet von Michael Wenzel, Transkription und Kommentar von Ursula Timann und Michael Wenzel, Wolfenbüttel: Herzog August Bibliothek 2020ff.

   Die Links zum Download finden sich unter: <https://hainhofer.hab.de/informationen-zur-edition/downloads>, zuletzt abgerufen am 13.09.2024. [↑](#footnote-ref-1)
2. Vgl. dazu: <https://hainhofer.hab.de/informationen-zur-edition/ueber-philipp-hainhofer>, zuletzt abgerufen am 12.09.2024. [↑](#footnote-ref-2)
3. Neitzschitz, Georg Christoph von: Sieben-Jährige und gefährliche WeltBeschauung Durch die vornehmsten Drey Theil der Welt Europa/ Asia und Africa. Bautzen, 1666. In: Deutsches Textarchiv <https://www.deutschestextarchiv.de/neitschitz\_reise\_1666>, abgerufen am 19.07.2024, URN: urn:nbn:de:kobv:b4-20878-1. [↑](#footnote-ref-3)
4. Sulzer, Johann Georg: Tagebuch einer von Berlin nach den mittäglichen Ländern von Europa in den Jahren 1775 und 1776 gethanen Reise und Rückreise. Leipzig, 1780. In: Deutsches Textarchiv <https://www.deutschestextarchiv.de/sulzer\_reise\_1780>, abgerufen am 19.07.2024, URN: urn:nbn:de:kobv:b4-25231-5. [↑](#footnote-ref-4)
5. Schulz, Friedrich: Reise eines Liefländers. 3 Bde. Berlin, 1795.

   Bd. 1, H. 1: <https://www.deutschestextarchiv.de/schulz\_reise0101\_1795>, abgerufen am 19.07.2024, URN: urn:nbn:de:kobv:b4-200905197358.

   Bd. 1, H. 2: <https://www.deutschestextarchiv.de/schulz\_reise0102\_1795>, abgerufen am 19.07.2024, URN: urn:nbn:de:kobv:b4-17297-4.

   Bd. 2, H. 3: <https://www.deutschestextarchiv.de/schulz\_reise0201\_1795>, abgerufen am 19.07.2024, URN: urn:nbn:de:kobv:b4-200905197368.

   Bd. 2, H. 4: <https://www.deutschestextarchiv.de/schulz\_reise0202\_1795>, abgerufen am 19.07.2024, URN: urn:nbn:de:kobv:b4-17298-9.

   Bd. 3, H. 4 u. 5: <https://www.deutschestextarchiv.de/schulz\_reise03\_1795>, abgerufen am 19.07.2024, URN: urn:nbn:de:kobv:b4-200905197374. [↑](#footnote-ref-5)
6. <http://programminghistorian.org/en/lessons/analyzing-documents-with-tfidf>, zuletzt abgerufen am 12.09.2024. [↑](#footnote-ref-6)