

Intervalle, Konflikte, Zyklen. Modellierung von Makrogenese in der Faustedition

Vitt, Thorsten

thorsten.vitt@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Deutschland

Brüning, Gerrit

bruening@faustedition.de
Freies Deutsches Hochstift / Frankfurter Goethe-Museum

Pravida, Dietmar

dpravida@goethehaus-frankfurt.de
Freies Deutsches Hochstift / Frankfurter Goethe-Museum

Wissenbach, Moritz

wissenbach@faustedition.de
Julius-Maximilians-Universität Würzburg, Deutschland

Die Faustedition versammelt 573 Textzeugen mit ca. 730 datierbaren Objekten zu Goethes „Faust“ in einem digitalen Archiv. Für die Benutzer sollen sich die repräsentierten historischen Objekte nicht isoliert voneinander, sondern in einem sinnvollen Zusammenhang darstellen. Eine solcher Zusammenhang kann dadurch entstehen, dass die Objekte genetisch geordnet werden. Genetische Einordnung bedeutet zuallererst, Objekte zeitlich zu situieren. Dabei geht es teils um die Ermittlung von Kalenderdaten, teils um die Bestimmung des relativen zeitlichen Verhältnisses mehrerer Objekte. Die Ermittlung chronologischer Systeme ist eine Grundfrage in vielen Disziplinen. Je nach Sachlage können dabei etwa statistische Verfahren (z.B. Bayliss 2015 zur Archäologie) oder evolutionäre Modelle (Trovato 2014: 189–200 zur mediävistischen Philologie) eingesetzt werden; die Verwendung disziplinfremder Ansätze kann aber auch zu Problemen führen (Pereltsvaig/Lewis 2015 zur Glottochronologie). Beim derzeitigen Stand scheint es am aussichtsreichsten, das in einzelnen Disziplinen oder bei einzelnen Projekten geübte Vorgehen zu formalisieren, um potentiell generalisierbare Verfahren zu entwickeln.

Besonders schwierig ist es, nicht bloß einzelne Objekte zu datieren, sondern eine große Menge in eine chronologische Ordnung zu bringen, wenn die Objekte genetisch voneinander abhängig sind, nur wenige absolute Daten zur Verfügung stehen und sonst nur lokale Anhaltspunkte für relative Chronologien gegeben sind (klassisches Beispiel: die antike Chronographie; Grafton 1993, Burgess/Witakowski 1999). Dies ist auch bei umfangreichen

genetischen Handschriftendossiers neuzeitlicher Autoren und Werken mit komplexer Entstehungsgeschichte der Fall. Hier können einzelne Teilentwürfe in relativer zeitlicher Beziehung stehen, vereinzelt sind Datierungen verfügbar; doch die Rekonstruktion der *Makrogenese*, d.h. der chronologischen Ordnung des Gesamtbestands (zum Begriff Van Hulle 2018: 47–48), kann sich als außerordentlich komplex erweisen.

Und so liegen die Dinge auch bei „Faust“. Nur wenige der makrogenetischen Objekte sind genau datierbar; stattdessen gibt es eine große Menge relativer, aber strikt lokaler Chronologien für jeweils nur einige Objekte. Den bislang einzigen Versuch, Einzelaussagen zu aggregieren und alle Objekte in zeitlich-stemmatische Beziehung zueinander zu setzen, macht Fischer-Lamberg für zwei Akte des „Faust II“. Ihre Stemmata (Fischer-Lamberg 1955: 150–166) markieren die praktische Grenze dessen, was an Einzelinformationen mit menschlichen Mitteln aggregiert werden kann. Die Rekonstruktion einer (theoretisch beliebig großen) Makrogenese verlangt nach maschineller Verarbeitung und visueller Aufbereitung vorhandener Einzelinformationen.

Datengrundlage

Um dies zu ermöglichen, wurde der Informationsgehalt der verfügbaren einschlägigen Aussagen zur Datierung¹ in XML erfasst:

- bibliographische Quelle
- datierter Textzeuge
- absolute Datierung
- relative Datierung
- (ungefähre) Gleichzeitigkeit von Textzeugen

Eine *Relative Datierung* setzt Zeugen² in eine zeitliche Reihenfolge. Es wird ausgesagt, dass ein Zeuge vor einem anderen entstanden ist.

```
<relation name="temp-pre">
  <source uri="faust://bibliography/fischer-lamberg1955">S. 160</source>
  <item uri="faust://document/wa/2_III_H.5"/>
  <item uri="faust://document/wa/2_III_H.8"/>
</relation>
```

Abbildung 1. Kodierte Aussage: Laut Fischer-Lamberg 1955: 160 (source), ist 2 III H.5 vor 2 III H.8 (temp-pre) (vereinfacht).

Eine *Absolute Datierung* ordnet einen oder mehrere Textzeugen konkreten Datumsangaben zu. Tagesgenaue Datierungen bilden die Ausnahme; häufig sind Aussagen, dass ein Zeuge nicht vor oder nicht nach einem Zeitpunkt entstanden sei. Typisch sind unscharfe Angaben wie 1800/1801 oder Frühsommer. Solche Angaben werden nach dokumentierten Regeln auf Aussagen der Form nicht vor und nicht nach mit jeweils normierten tagesgenauen Angaben abgebildet.

```
<date notBefore="1825-02-26" notAfter="1825-04-05">
<source uri="faust://bibliography/bohnenkamp1994">S. 573</source>
<item uri="faust://document/wa/2_III_H.5"/>
</date>
```

Abbildung 2. Laut Bohnenkamp 1994: 208, ist 2 III H.5 auf den Zeitraum vom 26. Februar zum 5. April 1825 zu datieren (vereinfacht).

Modellierung als Graph

Die verschiedenen Aussagen werden in einem gerichteten Graphen modelliert³: Die Zeugen bilden Knoten, die relativen Datierungen Kanten des Graphen. Zur Integration der absoluten Datierungen werden die Datumsangaben (Tage) ebenfalls als Knoten in den Graphen integriert: Aus einer Datierung wie 2 III H.5 wurde nicht vor dem 26. Februar 1825 und nicht nach dem 5. April 1825 geschrieben wird so ein Teilgraph 1825-05-25 # 2 III H.5 # 1825-04-06, in dem die Knoten für Ereignisse (Daten oder Zeugen) und die Kanten für zeitlich vor stehen. Ergänzt wird der Graph durch »Zeitstrahlkanten«, die von jedem Datum zum nächstfolgenden führen.

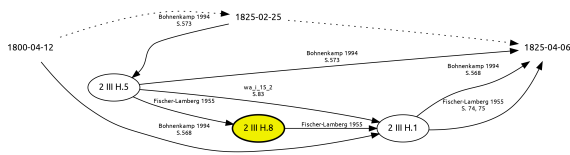


Abbildung 3. Aus Einzelaussagen gebildeter Graph zu 2 III H.8, unmittelbar benachbarten Zeugen und Datierungen (vereinfacht).

Aus diesem Graphen lassen sich Informationen ableiten, die sich erst aus dem Zusammenspiel der Einzelaussagen ergeben: Betrachtet man einen Zeugen z , so sind alle von z aus entlang gerichteter Kanten erreichbaren Zeugen (hier: 2 III H.1) nach z entstanden, und die von z aus erreichbaren Daten bilden Grenzen des *terminus ante quem* (des Zeitpunkts, vor welchem ein Zeuge entstand). Der Graph bietet damit die Grundlage für eine ungefähre Datierung auch derjenigen Zeugen, für die keine direkte absolute Datierung vorliegt.

Ist die Gesamtheit der Aussagen nicht widerspruchsfrei, so ergeben sich Zyklen im Graphen. Dies induziert einen Teilgraphen, in dem (vgl. Abb. 4 mit relativen Datierungen einiger Handschriften) jeder Knoten von jedem anderen erreichbar ist (der Teilgraph ist *stark zusammenhängend*).

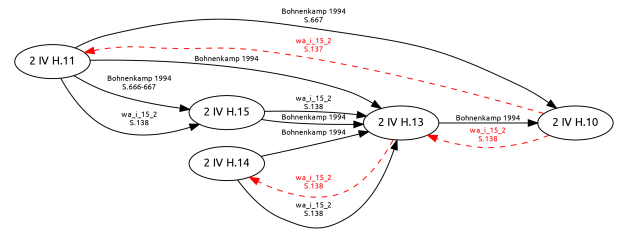


Abbildung 4. Relative Datierungen für einige Handschriften des 4. Akts. Erst die Entfernung aller rotgestrichelten Kanten führt zu einem zyklensfreien Graphen.

Um den Graphen aus Abb. 4 zyklensfrei zu machen, ist die Entfernung von wenigstens drei Kanten notwendig (gestrichelt). Der komplette Makrogenesegraph enthält eine stark zusammenhängende Komponente mit 477 Dokumenten und 2136 Kanten – zu umfangreich, um die Konflikte manuell zu eliminieren.

Konfliktbehandlung

Eine möglichst kleine Menge von Kanten zu entfernen, um einen zyklensfreien Graphen zu erhalten, ist ein als *Minimum Feedback Arc Set* bekanntes, NP-vollständiges (Karp 1972) Problem der Graphentheorie; um eine optimale Lösung zu finden, ist die größte der stark zusammenhängenden Komponenten zu groß (und die aus graphentheoretischer Sicht optimale Lösung muss auch nicht die philologisch korrekte sein, es wäre nur diejenige, die die wenigsten Datierungsaussagen verwirft). Verwendet wird stattdessen eine Heuristik, z.Z. das Verfahren von Eades (1993, Implementierung von igraph⁴), wobei zuviel entfernte Kanten nach Möglichkeit wieder hinzugefügt werden.

Entfernt man alle Konfliktkanten, so erhält man einen zyklensfreien gerichteten Graphen (DAG), der die Basis für die automatisierte Weiterverarbeitung ist. Dessen Knoten können in eine *topologische Ordnung* gebracht werden, d.h. eine Reihenfolge, die mit allen Kanten des Graphen konsistent ist (Manber 1989: 199).

Um die aus einer Vielzahl teils widersprüchlicher Aussagen mechanisch gezogenen Schlüsse nachvollziehbar und verbesserbar zu machen, werden die Daten in einer Reihe verlinkter Darstellungen mit GraphViz (Gansner/North 2000) visualisiert. Die Grundlage bildet der Gesamtgraph mit allen Aussagen, in dem algorithmisch entfernte Aussagen (Konflikte) rotgestrichelt visualisiert werden.

Zu jedem Zeugen gibt es einen Teilgraphen, der seine Nachbarn, die nächsten erreichbaren absoluten Datierungen und alle Aussagen dazwischen visualisiert. Darunter werden die Aussagen tabellarisch aufgelistet. Zu jeder (entfernten) Konfliktkante zeigt eine separate Visualisierung einen Pfad in der Gegenrichtung, mit dem die Kante in Konflikt stand, so dass der Konflikt erkennbar

wird und zu den beteiligten Zeugen weternavigiert werden kann.⁵ Daneben gibt es aufbereitete Darstellungen für jede Szene und jede Quelle.

Anhand der Visualisierungen können die vorliegenden Reihenfolgeentscheidungen nachvollzogen, aber auch Datierungskontroversen und -lücken identifiziert und in den Quelldateien behoben werden:

- Aussagen können als *zu ignorieren* markiert werden, um sie bei der Bildung des Graphen auszuschließen.
- Quellen können nach ihrer Zuverlässigkeit bewertet werden, um zu beeinflussen, wie leicht oder schwer entsprechende Kanten als Konfliktkanten entfernt werden.
- Eigene Erkenntnisse können mit entsprechend hohem Gewicht einbezogen werden.

Die visualisierten Ergebnisse erfüllen so einen mehrfachen Zweck: Sie dienen zur systematischen Erschließung der einschlägigen Forschung, zur Klärung der genetischen Verhältnisse für den gesamten Faust und als Hilfsmittel zur Überprüfung, Vervollständigung und Verbesserung der Datenlage, d.h. zur Erweiterung des Forschungsstand. Darüber hinaus wird die ermittelte Reihenfolge in der Faustedition verwendet, um die Zeilensynopse sowie das Balkendiagramm zu sortieren.

Ausblick

Neben der manuellen Nachbearbeitung der Daten kommen zur Verbesserung des Verfahrens alternative Heuristiken für das Minimum-Feedback-Arc-Problem in Frage (z.B. Even et al. 1998).

Bei der vorgestellten Methode werden in der relativen Chronologie die Entstehungsintervalle nur in eine disjunkte vor-nach-Beziehung gesetzt. Möglich wäre es, hier zu prüfen, ob eine weitere Ausdifferenzierung der Beziehungen zwischen Entstehungsintervallen realisierbar ist, so dass Beziehungen der Form „wurde begonnen vor Fertigstellung von“ ausgedrückt werden können. Komplexere Relationen als eine einfache zeitliche Abfolge werden bereits von Allen (1983) in einem Graphmodell beschrieben, allerdings ist hier die globale Konfliktfreiheit ein Problem.

Eine Alternative zu der oben beschriebenen Abbildung unscharfer Aussagen auf scharf begrenzte Intervalle ist etwa die Modellierung über Fuzzy-Mengen (vgl. z.B. Barro et al. 1994). Dies erfordert jedoch auch die Neudefinition der Relationen (Schockaert/De Cock 2008) und der darauf aufbauenden Verfahren etwa zur Konfliktauflösung.

Der vorgestellte Ansatz basiert auf bereits vorhandenen, mit traditionellen philologischen Mitteln gewonnenen Datierungsaussagen. In Wissenbach/Pravida/Middell (2012) wird ein Verfahren vorgestellt, mit der kodierte, textinhärente Eigenschaften von Fassungen für die regelbasierte Bildung genetischer Hypothesen genutzt werden, um auf diesem Weg generelle Hypothesen zur

Arbeitsweise des Autors zu verifizieren. Nachdem die genetische Analyse des Korpus nun weiter vorangeschritten ist, kann dieser Ansatz auf einer breiteren Datenbasis evaluiert werden.

Fußnoten

1. Zu den ausgewerteten Quellen siehe faustedition.net/macrogenesis/sources.
2. Das Datenmodell der Faustedition sieht eine konzeptuelle Unterscheidung zwischen Zeugen und Inskriptionen (Niederschriften) vor: Die eigentlichen Objekte der Datierung sind Inskriptionen. Ein Zeuge kann eine oder mehrere unterschiedlich datierte Inskriptionen enthalten. Im folgenden wird der Einfachheit halber nur von Zeugen gesprochen.
3. Mit der Graphbibliothek NetworkX (Hagberg/Schult/Swart 2008).
4. <http://igraph.org/>
5. Beispiel: faustedition.net/macrogenesis/H_P93--1825-01-01.

Bibliographie

- Allen, James F. (1983):** *Maintaining knowledge about temporal intervals*, in: *Communication of ACM* 21(11): 832–843 doi: 10.1145/182.358434.
- Barro, Senén / Marín, Roque / Mira, José / Patón, Alfonso R. (1994):** *A model and a language for the fuzzy representation and handling of time*, in: *Fuzzy Sets and Systems* 61: 153–175. doi: 10.1016/0165-0114(94)90231-3.
- Bayliss, Alex (2015):** *Quality in Bayesian chronological models in archaeology*, in: *World Archaeology* 47(4): 677–700. doi: 10.1080/00438243.2015.1067640
- Bohnenkamp, Anne (1994):** ... das Hauptgeschäft nicht außer Augen lassend. *Die Paralipomena zu Goethes Faust*. Frankfurt am Main / Leipzig: Insel.
- Burgess, Richard W. / Witakowski, Witold (1999):** *Studies in Eusebian and Post-Eusebian Chronography* (= *Historia. Einzelschriften*; 135). Stuttgart: Steiner.
- Eades, Peter / Lin, Xue-Min / Tamassia, Roberto (1993):** *A fast and effective heuristic for the feedback arc set problem*, in: *Information Processing Letters* 47(6): 319–323. doi: 10.1016/0020-0190(93)90079-O.
- Even, Guy / Naor, Joseph / Schieber, Baruch / Sudan, Madhu (1998):** *Approximating Minimum Feedback Sets and Multicuts in Directed Graphs*, in: *Algorithmica* 20(2): 151–174. doi: 10.1007/PL00009191.
- Fischer-Lamberg, Renate (1955):** *Untersuchungen zur Chronologie von Faust II 2 und 3. Diss. phil. (masch.)*, Humboldt-Universität Berlin.
- Gansner, Emden R. / North, Stephen C. (2000):** *An open graph visualization system and its applications to software engineering*, in: *Software: Practice and Experience* 30(11): 1203–

1233. doi: 10.1002/1097-024X(200009)30:11<1203::AID-SPE338>3.0.CO;2-N.

Grafton, Anthony (1993): *Joseph Scaliger. A Study in the History of Classical Scholarship. Vol. II: Historical Chronology* (= Oxford-Warburg Studies). Oxford: Clarendon.

Hagberg, Aric A. / Schult, Daniel A. / Swart, Pieter J. (2008): *Exploring Network Structure, Dynamics, and Function using NetworkX*, in: **Varoquaux, Gael / Vaught, Travis / Millman, Jarrod (eds):** Proceedings of the 7th Python in Science Conference (SciPy2008) in Pasadena, CA <https://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-UR-08-05495> [letzter Zugriff 14. Oktober 2018].

Karp, Richard M. (1972): *Reducibility Among Combinatorial Problems*, in: **Miller, Raymond Edward / Thatcher, James W. (eds.):** *Complexity of Computer Computations*. New York: Plenum 85–103. doi: 10.1007/978-3-540-68279-0_8.

Manber, Udi (1989): *Introduction to Algorithms: A Creative Approach*. Reading, MA: Addison-Wesley.

Pereltsvaig, Asya / Lewis, Martin (2015): *The Indo-European Controversy. Facts and Fallacies in Historical Linguistics*. Cambridge: Cambridge University Press.

Schockaert, Steven / De Cock, Martine (2008): *Temporal Reasoning about Fuzzy Intervals*, in: *Artificial Intelligence* 172(8): 1158–1193. doi: 10.1016/j.artint.2008.01.001.

Trovato, Paolo (2014): *Everything You Always Wanted to Know About Lachmann's Method. A Non-Standard Handbook of Genealogical Textual Criticism in the Age of Post-Structuralism, Cladistics, and Copy-Text*. Padova: libreriauniversitaria.it

Van Hulle, Dirk (2018): *Editing the Wake's Genesis: Digital Genetic Criticism*, in: **Sartor, Genevieve (ed.):** *James Joyce and Genetic Criticism. Genesis Fields* (= European Joyce Studies; 28). Leiden, Boston: Brill Rodopi 37–54.

Wissenbach, Moritz / Pravida, Dietmar / Middell, Gregor (2012): *Reasoning about Genesis or The Mechanical Philologist*, in: **Meister, Jan Christoph (ed.):** *Digital Humanities 2012. Conference Abstracts*. Hamburg: Hamburg University Press 418–422 <http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/reasoning-about-genesis-or-the-mechanical-philologist.1.html> [letzter Zugriff 14. Oktober 2018].