ISSN 2568-9967

BIIIE

THEMENHEFT

12

Elisabeth Lienert, Joachim Hamm,
Albrecht Hausmann und
Gabriel Viehhauser (Hrsg.)

DIGITALE MEDIÄVISTIK

WWW.ERZAEHLFORSCHUNG.DE

BmE

THEMENHEFT 12

Elisabeth Lienert / Joachim Hamm
Albrecht Hausmann / Gabriel Viehhauser (Hrsg.)

Digitale Mediävistik

Perspektiven der Digital Humanities für die Altgermanistik

Publiziert im November 2022.

http://www.erzaehlforschung.de – Kontakt: herausgeber@erzaehlforschung.de
ISSN 2568-9967

Zitiervorschlag für dieses Themenheft:
Lienert, Elisabeth / Hamm, Joachim / Hausmann, Albrecht / Viehhauser, Gabriel (Hrsg.): Digitale Mediävistik. Perspektiven der Digital Humanities für die Altgermanistik, Oldenburg 2022 (BmE Themenheft 12, online).

Dank

Der vorliegende Band enthält das Gros der Beiträge der Internationalen digitalen Tagung ›Digitale Mediävistik. Perspektiven der Digital Humanities für die Altgermanistik‹ (9.-11. Februar 2022), teilweise erweitert und um Diskussionsberichte ergänzt. Für Tagungsunterstützung danken wir Sonja Kerth, Katharina Biskup, Julia Gardlo, Ramona Theßmann und Elvira Vogt, den drei Letztgenannten auch für Protokollnotizen. An der redaktionellen Einrichtung der Beiträge haben Elvira Vogt und Ramona Theßmann mitgewirkt; den ›Satz‹ hat Albrecht Hausmann übernommen. Ihnen gilt besonderer Dank. Dank gebührt darüber hinaus den Herausgeber*innen der Beiträge zur mediävistischen Erzählforschung, Anja Becker und wiederum Albrecht Hausmann, für die Aufnahme des Bandes in die Reihe der Themenhefte und für die Betreuung der Publikation.

Elisabeth Lienert, auch im Namen der Mitherausgeber

Inhaltsverzeichnis

Gabriel Viehhauser / Joachim Hamm
Albrecht Hausmann / Elisabeth Lienert

Einführung

Obwohl das Schlagwort der Digitalität immer noch mit dem Topos der Neuheit und des Innovativen, manchmal auch des Zeitgeistigen belegt ist, gibt es die Digital Humanities nun schon ziemlich lange. Das Projekt, das als ihr Gründungsereignis gilt, war bekanntlich ein mediävistisches, nämlich der ab 1949 vom Jesuitenpater Roberto Busa in Zusammenarbeit mit der Firma IBM erstellte Index Thomisticus, also eine digitale Konkordanz zu den Werken des Thomas von Aquin (vgl. hierzu Jones 2016 sowie Unsworth 2011). Und auch und gerade die Altgermanistik kann durchaus zu den *early adopters* der digitalen Geisteswissenschaften gezählt werden (Gärtner 2016). Genannt seien hier stellvertretend die Pioniere Roy Wisbey, Kurt Gärtner und Klaus Schmidt mit seiner Mittelhochdeutschen Begriffsdatenbank (MHDBDB). Wenn man sich diese Projekte und Namen ansieht, dann zeigt sich ein erfreulicher, vielleicht auch wieder dem Klischee der digitalen Ephemerität widersprechender Befund, nämlich, dass es manche von diesen Urgestein-Projekten tatsächlich heute auch noch gibt und diese nicht in unlesbare Magnet-Datenträger oder nicht mehr aufrufbare Internet-Seiten abgetaucht sind. Den Index Thomisticus kann man heute noch auf einer durchaus eleganten Webseite benutzen, und Vertreter*innen der mittelhochdeutschen Begriffsdatenbank oder aus Trier waren auf der Tagung und sind im Sammelband vertreten.

Die digitale Konferenz »Digitale Mediävistik: Perspektiven der Digital Humanities für die Altgermanistik« (9.-11.2.2022, veranstaltet von Elisabeth Lienert, Bremen; Joachim Hamm, Würzburg; Albrecht Hausmann,

Oldenburg; Gabriel Viehhauser, Stuttgart; in Kooperation mit dem Institut für Mittelalter- und Frühneuzeitforschung am Fachbereich 10 der Universität Bremen und dem Verbund Mittelaltergermanistik Nord) und der hier vorgelegte Online-Sammelband hatten und haben auch eine sammelnde und zusammenfassende Zielsetzung, nämlich aufzuzeigen, was es schon gibt und in welchen Bereichen digitale Altgermanistik möglich ist, war und wäre. In erster Linie geht es aber um Perspektiven im Sinne eines zukunftsgewandten Blicks. Denn trotz der gar nicht so kurzen Vorgeschichte gibt es die Digital Humanities auf einer institutionell eingespielten Ebene in ausgeprägter Form nun doch noch gar nicht so lange. Erst in jüngerer Zeit sind in größerer Zahl Professuren für Digital Humanities und entsprechende Studiengänge eingeführt worden.[1] Diese zum Teil noch anhaltende institutionelle Etablierung geht einher mit den Ausweitungen der digitalen Möglichkeiten und einem grundsätzlichen Data Turn in der Gesellschaft, der nicht nur die Forschung, sondern alle Lebensbereiche so umgestaltet, dass man an ihm gar nicht vorbeigehen kann, selbst wenn man es wollte, auch in den Geisteswissenschaften nicht.

Die Digital Humanities gibt es also, und doch scheint das Verhältnis von digitalen Geisteswissenschaften und ›traditionellen‹ Fächern ausbaufähig. Tatsächlich spielt in den Digital Humanities als Fach zurzeit die Methodenentwicklung eine große Rolle; diese ist ja auch der gemeinsame Nenner der so unterschiedlichen Disziplinen, die sich unter diesem Dach versammeln. Es gibt eigene, gut besuchte DH-Konferenzen, an denen man sich sehr gut über neueste Methoden und durchaus auch deren Reflexion austauschen kann, aber nur selten über fachdisziplinäre Inhalte. Vor diesem Hintergrund stell(t)en die Bremer Tagung und der vorliegende Tagungsband sehr konkret und konzentriert aus der Perspektive der Altgermanistik die Frage »Was braucht das Fach?«. Was können digitale Methoden und Infrastrukturen also konkret zum Erkenntnisinteresse der germanistischen Mediävistik beitragen? Diese dezidiert aus der Germanistik kommende Frage soll keinesfalls die Notwendigkeit bestreiten, dass digitale Geisteswissenschaft

und insbesondere digitale Mittelalterstudien immer auch die interdisziplinäre Perspektive im Blick haben müssen. Dennoch erschien und erscheint uns dieser Bottom-up-Zugang sinnvoll.

Die Frage, was die germanistische Mediävistik braucht, erschöpft sich dabei nicht im Aufbereiten und Verstehen von Texten. Neben klassischen Feldern wie digitaler Editorik und Interpretation spielen nämlich auch Bereiche wie Infrastrukturen und Repositorien, aber auch das digitale Publizieren eine große Rolle. Wenn man im digitalen Bereich arbeitet, sind Theorie und Praxis untrennbar miteinander verbunden; trotz der Auflösung der Objekte in Zahlen spielen die technischen Grundlagen und das handfeste Materielle gerade eine besondere Rolle. Denn was nutzt es z. B., wenn man im Digitalen zwar potentiell wunderbare Editionen erstellen könnte, die die ganze Varianz von Texten in den Blick nehmen und diese zudem auch noch vielfältig multimodal kontextualisieren können, wenn man aber keine entsprechenden Infrastrukturen zu Verfügung hat, die dabei helfen, innovative Darstellungsformen jenseits des Buchparadigmas zu ermöglichen und auch langfristig nachhaltig sichtbar zu machen?

Nicht intendiert war und ist eine bloße Projekt-Revue. Konkrete Beispiele sollen nur Ausgangspunkte, die Stoßrichtung immer eine grundsätzliche sein: Bei welchen mediävistischen Forschungsfragen können digitale Verfahren und Methoden helfen? Welche digitale Infrastruktur braucht man dazu? Wie positioniert man sich im Spannungsfeld von Differenzierung und Standardisierung? Und schließlich auch, wie vermittelt man zwischen den spezifischen Kompetenzen der traditionellen und der digitalen Geisteswissenschaftler und Geisteswissenschaftlerinnen?

Dieses Themenheft der BmE enthält das Gros der Tagungsbeiträge, teils in erweiterter Form und ergänzt um, für Sektion 1 und 5 vertreten durch Diskussionsberichte. Sechs Sektionen galten und gelten einigen Bereichen der Digitalität, die für die Mittelaltergermanistik von Bedeutung sind und eine besondere Dynamik aufweisen, ihrem Nutzen für das Fach, ihrer Anwendbarkeit und Eignung, ihrer technologischen und methodischen Inno-

vation und spezifischen Problematik, ihrer wünschenswerten Fortentwicklung: Digitalisierung von Handschriften und frühen Drucken sowie OCR (Sektion 1); Digitale Edition (Sektion 2); Digitale Infrastruktur und Forschungsdatenmanagement (Sektion 3); Repositorien und Datenbanken (Sektion 4); Online publizieren und digitale Wissenschaftskommunikation (Sektion 5); Stilometrie und Textanalyse (Sektion 6). Ziel war und ist nicht die vollständige Bestandsaufnahme, sondern der Anstoß zum Austausch über Aspekte der Digitalität, die für unser Fach bedeutsam sind und werden – die Diskussion reicht über diesen Rahmen hinaus und sollte auch über ihn hinaus weitergeführt werden.

Sektion 1: Die Digitalisierung der handschriftlichen Überlieferung wird bekanntlich seit Jahren vorangetrieben. So hat die DFG im Jahr 2018 ein LIS-Förderprogramm (Wissenschaftliche Literaturversorgungs- und Informationssysteme) und einen Masterplan aufgelegt, um die insgesamt über 60.000 abendländisch-mittelalterlichen Handschriften, die in öffentlichen Einrichtungen in Deutschland verwahrt werden, zu digitalisieren und zu erschließen. Die ›Manuscripta mediaevalia‹ verzeichnen bereits über 14.300 digitalisierte Handschriften, und das neue DFG-Projekt ›Handschriftenportal‹ wird dies weiterführen. Auch die Digitalisierung von Inkunabeln und Frühdrucken ist weit vorangeschritten: Die etwa 26.000 Ausgaben von Wiegendrucken in Deutschland sind zum Großteil digitalisiert (vgl. die Verlinkungen im Gesamtkatalog der Wiegendrucke [GW]). Von den ca. 120.000 deutschen Druckausgaben des 16. Jahrhunderts sind etwa 106.000 im VD16 erfasst, und 68.000 davon haben Links zu Digitalisaten. Um diese Bilddaten durchsuchen, annotieren, analysieren, ja überhaupt weiterverarbeiten zu können, müssen sie in computerlesbaren Text umgesetzt werden. Hier kommt die Texterkennung ins Spiel. Gerade für frühe Drucke und für Handschriften hat man hier in den letzten Jahren enorme Fortschritte erzielt, die – über das Edieren hinaus – ganz neue Horizonte eröffnen. Gabriel Viehhauser resümiert die Vorträge von Günter Mühlberger und Christian Reul über die beiden in der Altgermanistik am meisten

genutzten Texterkennungs-Tools Transkribus und OCR4all sowie die Diskussion. Die Genauigkeit der maschinellen Texterkennung wäre (ggf. über Nachkorrektur-Tools oder Verlinkungen mit Wörterbüchern und Grammatiken) noch weiter verbesserungsfähig, doch eröffnet eine gewisse Fehlertoleranz die Möglichkeit, vergleichsweise rasch große Textmengen verfügbar zu machen, womöglich gar eine Volltexterfassung aller mittelalterlichen deutschen Handschriften in öffentlichem Besitz.

Sektion 2: Digitale Editionsformen bieten vielfältige neue Möglichkeiten, werfen aber auch altbekannte Grundsatzfragen der Philologien aufs Neue auf (was ist ein Text, was eine Edition?) und generieren ihrerseits Schwierigkeiten, die es zu lösen gilt. Ihre Relevanz für das Fach ist unübersehbar, nicht nur, weil die DFG längst für alle Editionsprojekte zumindest »eine Sicherung und Bereitstellung der Textdaten in digitaler Form« (DFG: Informationen für Geistes- und Sozialwissenschaftler/innen. Förderkriterien für wissenschaftliche Editionen in der Literaturwissenschaft, S. 2) verlangt. Patrik Sahle hat das Feld der digitalen Editorik schon vor Jahren vermessen, und seitdem ist die Zahl der Projekte, Konzepte und Realisierungen stetig angewachsen: Sahles Online-Katalog verzeichnet mehr als 780 digitale Editionsprojekte. Dass dabei technologisch nicht jedes Mal das Rad neu erfunden werden müsste, liegt auf der Hand, ist aber durchaus keine Selbstverständlichkeit. Umso wichtiger sind digitale Großprojekte, die Infrastrukturen ausarbeiten, Arbeitsumgebungen schaffen, Standards vorbereiten. Zwei Beispiele werden von Jakub Šimek (»heiEDITIONS – eine Heidelberger Infrastruktur für Editionen (nicht nur) mittelalterlicher Texte«) und Angila Vetter (»ediarum.mediaevum – Eine Arbeitsumgebung zur Edition mittelalterlicher (Prosa)Texte«) vorgestellt: Die komplexe Editionsinfrastruktur heiEDITIONS der Universitätsbibliothek Heidelberg bietet digitalen Editionen die multidimensionale Erschließung ihrer Gegenstände durch Datenmodellierung und Visualisierung sowie die Möglichkeit langfristiger Aufbewahrung. Um der Breite der Überlieferung und der Komplexität der Prosatexte wie auch den

Ausgabeformaten als Print- und Webedition gerecht zu werden, wurde im Projekt ›Der Österreichische Bibelübersetzer‹ auf der Basis von ediarum die Arbeitsumgebung ediarum.mediaevum entwickelt. Es gibt freilich auch die Kehrseite, die digitale Editionsruine, das Altgemäuer im Internet, das keiner mehr pflegt, dem mit jeder neuen PHP-Version der Einsturz, besser: der Absturz droht. Eine eher desillusionierte Bilanz hinsichtlich der Lebensdauer digitaler Editionen zieht der Beitrag von Sonja Glauch (»Welche Lebenserwartung haben digitale Editionen?«): Vor allem durch veraltete digitale Frameworks, misslungene Migrationen und ins Leere laufende Links droht digitalen Editionen die Unbenutzbarkeit. Eine Lösungsperspektive für das Langfristaufbewahrungsproblem könnten im Aufbau befindliche Forschungsdateninfrastrukturen wie Text+ bieten, ohne dass jedoch ein konkreter Königsweg in Sicht ist. Albrecht Hausmann berichtet über die Diskussion, die sich vor allem auf Standardisierungserfordernisse und deren Kehrseite, die Gefahr von Beschränkungen aufgrund von Standardisierung, sowie die Archivierungsproblematik und die Rolle der Bibliotheken bei der Pflege digitaler Editionen bezog.

Sektion 3 behandelt digitale Infrastrukturen und Forschungsdatenmanagement, den Umgang mit digitalen Daten in der Forschung von ihrer Planung und Generierung über ihre Nutzung und Verarbeitung bis hin zu Aufbereitung, ggf. Veröffentlichung, Speicherung, Archivierung, Nachnutzung – oder eben auch Löschung. Ziel ist es, Datenbestände unterschiedlicher Disziplinen für das deutsche Wissenschaftssystem systematisch zu erschließen, zu vernetzen und nachhaltig sowie qualitativ in Forschung und Lehre nutzbar zu machen. Andrea Rapp (»Digitale Infrastruktur und Forschungsdatenmanagement«) stellt das Konsortium Text+ vor, das im Rahmen der seit 2020 im Aufbau befindlichen Nationalen Forschungsdateninfrastruktur (NFDI) für Sprach- und Textdaten zuständig und auch für mediävistische Bedarfe offen ist, Thomas Burch (»Infrastrukturprojekte zur digitalen Lexikographie. Vorgestellt am Beispiel des Zentrums für Historische Lexikographie«) das Trierer Zentrum für Histo-

rische Lexikographie ZHistLex mit dem Projekt ›European Lexicographic Infrastructure‹ (ELEXIS). Über die Diskussion berichtet Albrecht Hausmann: Hier geht es um konzeptionelle Herausforderungen, vor allem bezüglich der Heterogenität der Daten, der Unterscheidung von Forschungsergebnissen und Forschungsdaten, der Notwendigkeit (und Problematik) von Standardisierung, aber auch um praktische Fragen der dezentralen Organisation, der Benutzbarkeit der jeweiligen Infrastrukturen, der Vernetzung und Entwicklung von Datenschnittstellen, nicht zuletzt auch der Förderpolitik.

Sektion 4 präsentiert ausgewählte Repositorien und Datenbanken: Jürgen Wolf (»Handschriftencensus (HSC). Von der Handschrift zu den Metadaten«) erläutert das vielfältige Potential des Handschriftencensus (HSC) als Kompetenzzentrum zur deutschsprachigen Textüberlieferung (einschließlich Handschriften-, Werk- und Autoridentifikation, Verlinkungen zu Digitalisaten, Editionen und Katalogen, Möglichkeit zur schnellen Publikation von Neufunden in der Online-Zeitschrift ›Maniculae‹). In ihrem Vortrag stellte Stefanie Dipper das Referenzkorpus Mittelhochdeutsch (ReM) vor; der weiterführende Beitrag mit Simone Schultz-Balluff (»ReM für Mediävist*innen. Perspektiven des Referenzkorpus Mittelhochdeutsch (1050-1350) für germanistisch-mediävistische Fragestellungen«) illustriert exemplarische mediävistische Auswertungen (Attribuierungen von Personennamen, Personifikationen, Metaphorik) unter Nutzung des Korpussuchtools ANNIS. Katharina Zeppezauer-Wachauer (»50 Jahre Mittelhochdeutsche Begriffsdatenbank (MHDBDB). Eine Jubiläums-Zeitreise zwischen Lochkarten, Pixel-Drachen, relationaler Datenbank und Graphdaten«) stellt die Geschichte der Mittelhochdeutschen Begriffsdatenbank, aktuelle Nutzungsmöglichkeiten (Suche, Metadaten, Annotation) sowie Kooperationen und Zukunftsperspektiven vor. Für den Gesamtkatalog der Wiegendrucke GW, dessen Anwendungsaspekte auf der Tagung der Vortrag Oliver Duntzes umfassend demonstrierte, sei hier auf die vorzügliche Website verwiesen. Über die Diskussion berichtet Joachim

Hamm; sie bezog sich in erster Linie auf Desiderata der Nutzer*innen sowie auf Aspekte der Vernetzung. Der Nutzen der Repositorien und Datenbanken steht außer Frage.

Sektion 5 wurde in Form einer Podiumsdiskussion durchgeführt, deren Ergebnisse hier durch den Bericht von Gabriel Viehhauser dokumentiert sind. Der Bereich von Online-Publizieren und digitaler Wissenschaftskommunikation weist bekanntlich eine hohe technologische Dynamik auf, man denke nur an die diversen Kanäle des digitalen Austauschs, von eher schweigsamen Mailinglisten über Mediaevum und engagierte Mittelalterblogs bis hin zu Twitteraccounts und Online-Publikationen wie den ›Beiträgen zur mediävistischen Erzählforschung‹ (BmE) oder dem ›Archivum Medii Aevi Digitale‹ (AMAD). Gerade dieses Thema berührt vitale Belange der Wissenschaften wie der Verlage, betrifft Traditionen und Zukunftsperspektiven des Fachs, ist für jede Wissenschaftlerin und jeden Wissenschaftler von Belang. Die Podiumsdiskussion (mit Albrecht Hausmann, BmE; Robert Forke, De Gruyter; Karoline Döring, Archivum Medii Aevi Digitale – AMAD; Leitung: Henrike Lähnemann) kreiste um die Fragen nach den Veränderungen der Publikationslandschaft durch Digitalisierung und damit verbundenen Chancen und Risiken für Qualitätssicherung und Zugänglichkeit, nach Publikationsinteressen der Wissenschaft, nach der Aufgabenverteilung zwischen Wissenschaft, Verlagen und Bibliotheken. Angesprochen wurden ferner Probleme der Finanzierung und der Reputation von (Online-)Publikationen.

Sektion 6 widmet sich Stilometrie und Textanalyse, Bereichen der Digital Humanities also, die philologisch-literaturwissenschaftliches Arbeiten durch neue Technologien ergänzen und dadurch im Kern verändern. Stilistische Merkmale und ihre Häufigkeit im Text lassen sich digital erheben und dazu nutzen, um Ähnlichkeiten zwischen Texten zu bestimmen und Texte zu klassifizieren und zu clustern. Autorschaftsattribution ist eine der Anwendungen in der Praxis; mittlerweile kommen zahlreiche weitere Fragestellungen, etwa zu Stilmitteln und Figurenanalyse hinzu. Gabriel

Viehhauser (»Digitale Methoden der Textanalyse für die Altgermanistik«) stellt am Beispiel eines Minnesangkorpus einige grundlegende Methoden digitaler Textanalyse vor (Frequenzanalyse, Principal Component Analysis, Lexical Diversity, Topic Modeling) und plädiert für eine multiperspektivische Verbindung von quantitativer digitaler Makroanalyse mit qualitativer Detailanalyse. Phillip Brandes, Sophie Marshall und Felix Schneider (»Stilfiguren aus der Distanz gelesen. Zur automatischen Detektion von Wortstellungsfiguren und deren Nutzen für die qualitative Analyse«) erläutern die quantitative Analyse von rhetorischen Stilmitteln am Beispiel der Parallelismus- und Chiasmusdetektion im Projekt *Anomaly-based large-scale analysis of style and genre reflected in the use of stylistic devices in medieval literature* (Leitung: Sophie Marshall) sowie ihre Auswertung mit Blick auf die Frage nach der Korrelation von Stilmittelhäufigkeit und Gattung. Der Beitrag von Friedrich Michael Dimpel, Andre Blessing, Peter Hinkelmanns, Nora Ketschik und Katharina Zeppezauer-Wachauer (»Figuren und ihr Handeln – eine computergestützte Untersuchung von Figurenaktivitäten im Kontext von Figurenreferenzen mit Hilfe des Begriffssystems der MHDBDB«) stellten Verfahren und Ergebnisse automatischer Annotierung und Erfassung von Figurenaktivitäten und ihrer Zuordnung zu Figurentypen in mittelhochdeutschen Erzähltexten dar. Die Diskussion referiert Elisabeth Lienert: Als Grundlage für digitale Textanalysen werden (möglichst einheitlich) normalisierte und nach Möglichkeit annotierte Ausgaben (einschließlich öffentlicher Förderung für solche Ausgaben) sowie Vernetzung mit (annotierten) digitalen Wörterbüchern und Datenbanken gefordert. Trotz beschränkter Aussagekraft digitaler Methoden für außergewöhnliche Einzelwerke können Distant und Close Reading sich sinnvoll ergänzen. Problematisiert wurden die Gefahr der Zirkularität und Bestätigung von Vorannahmen und Altbekanntem sowie ein Missverhältnis zwischen Aufwand und Ergebnis; hier sind vor allem neue erkenntnisgeleitete Fragestellungen zu entwickeln.

Die Abschlussdiskussion (zusammengefasst von Elisabeth Lienert) bezog sich vor allem auf Status und Akzeptanz von Digitalität, insbesondere bei der Anwendung digitaler Methoden auf genuin literaturwissenschaftliche Fragestellungen. Mit den Digital Humanities ändern sich auch Forschungsgegenstände und Methoden; insbesondere Verknüpfungen und Vernetzungen ermöglichen neue Fragen, mit denen auch der Gefahr der Komplexitätsreduktion zu begegnen ist. Gewinn liege freilich auch im Quantitativen – selbst die vollständige OCR aller mittelalterlichen Handschriften rückt in den Bereich des Möglichen. Als zentrale Aufgaben wurden die Langfristaufbewahrung von Daten sowie Normierung und Standardisierung benannt; gefordert wurde hierfür ein zentrales Datenrepositorium.

Tagung und Sammelband dokumentieren exemplarisch das im Fach bereits durch die Digital Humanities Erreichte: mächtige Textverarbeitungstools, Modellprojekte für digitale Editionen, nützliche Forschungsinfrastrukturen, vielfach nutzbare und zunehmend vernetzte Repositorien und Datenbanken, neue Methoden quantitativer Analyse. Benannt wurden aber auch Desiderata, insbesondere adäquate Standardisierung, eine zentrale Anlaufstelle und eine entsprechend langfristige Förderpolitik unter Einschluss der Nationalbibliotheken.

Die Resonanz der Tagung und ihre engagierten Diskussionen deuten darauf hin, wie lohnend, ja notwendig es ist, den Austausch über das Thema »Was braucht das Fach?« in der Mediävistik über die Bremer Initiative hinaus fort- und weiterzuführen. Die eigenen Bedarfe und Anliegen im Digitalen zu erfassen, adäquate Methoden und technische Verfahren zu entwickeln und zu reflektieren und die fachinterne Kommunikation über Fragen des Digitalen zu fördern und zu intensivieren, sind neue Aufgaben eines Fachs, das eben auch eine »Digitale Mediävistik« ist.

Anmerkungen

1 Im deutschsprachigen Raum z. B. in Basel, Bern, Berlin, Bielefeld, Darmstadt, Erlangen, Göttingen, Graz, Hamburg, Köln, Leipzig, Paderborn, Passau, Stuttgart, Trier, Wien, Wuppertal und Würzburg. Nicht wenige dieser Professuren sind dabei mit Mediävist*innen besetzt. Einen Überblick bietet die Seite ›kleinefaecher.de‹.

Literaturverzeichnis

Sekundärliteratur

Gärtner, Kurt: Die Anfänge der Digital Humanities, in: Akademie Aktuell 56,1 (2016), S. 18-23.

Jones, Steven: Roberto Busa, S. J., and the Emergence of Humanities Computing. The Priest and the Punched Cards, London 2016.

Unsworth, John: Medievalists as Early Adopters of Information Technology, in: Digital Medievalist 7 (2011) (online).

Online-Ressourcen

AMAD (Archivum Medii Aevi Digitale): http://www.amad.org/.

ANNIS (Annotation of Information Structure): https://corpus-tools.org/annis/.

BmE (Beiträgen zur mediävistischen Erzählforschung): http://www.erzaehlforschung.de/.

Der Österreichische Bibelübersetzer: https://bibeluebersetzer.badw.de/das-projekt.html.

DFG: Informationen für Geistes- und Sozialwissenschaftler/innen. Förderkriterien für wissenschaftliche Editionen in der Literaturwissenschaft: https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/foerderkriterien_editionen_literaturwissenschaft.pdf.

ediarum: https://www.ediarum.org/.

ELEXIS (European Lexicographic Infrastructure): http://www.elex.is/.

GW (Gesamtkatalog der Wiegendrucke): https://www.gesamtkatalogderwiegendrucke.de/.

Handschriftenportal: https://handschriftenportal.de/.

HSC (Handschriftencensus): https://handschriftencensus.de/.

heiEDITIONS (Heidelberger Infrastruktur für Editionen): https://heieditions.github.io/.
Index Thomisticus: https://www.corpusthomisticum.org/.
LIS (Wissenschaftliche Literaturversorgungs- und Informationssysteme): https://www.dfg.de/foerderung/programme/infrastruktur/lis/.
Manuscripta mediaevalia: http://www.manuscripta-mediaevalia.de/.
Mediaevum: https://www.mediaevum.de/.
MHDBDB (Mittelhochdeutsche Begriffsdatenbank): http://mhdbdb.sbg.ac.at/.
NFDI (Nationale Forschungsdateninfrastruktur): https://www.nfdi.de/.
OCR4all: http://www.ocr4all.org/.
Portal Kleine Fächer, Digital Humanities: https://www.kleinefaecher.de/kartierung/kleine-faecher-von-a-z.html?tx_dmdb_monitoring%5BdisciplineTaxonomy%5D=140&cHash=c5b76ccd171ecce8fe0ed45c4afaa5bc.
ReM (Referenzkorpus Mittelhochdeutsch): https://www.linguistics.rub.de/rem/.
Sahle, Patrick [u. a.]: A catalogue of Digital Scholarly Editions: https://www.digitale-edition.de.
Text+: https://www.text-plus.org/.
Transkribus: https://readcoop.eu/transkribus/.
VD16 (Verzeichnis der im deutschen Sprachbereich erschienenen Drucke des 16. Jahrhunderts): https://www.bsb-muenchen.de/sammlungen/historische-drucke/recherche/vd-16/.
ZHistLex (Zentrum für Historische Lexikographie): http://www.zhistlex.de/.

Anschriften der Autorin und Autoren:

Prof. Dr. Gabriel Viehhauser
Universität Stuttgart
Institut für Literaturwissenschaft
Digital Humanities
Herdweg 51
70174 Stuttgart
E-Mail: viehhauser@ilw.uni-stuttgart.de

Prof. Dr. Joachim Hamm
Julius-Maximilians-Universität Würzburg
Institut für deutsche Philologie
Am Hubland
97074 Würzburg
E-Mail: joachim.hamm@uni-wuerzburg.de

Prof. Dr. Albrecht Hausmann
Carl von Ossietzky Universität Oldenburg
Institut für Germanistik
26111 Oldenburg
E-Mail: albrecht.hausmann@uni-oldenburg.de

Prof. Dr. Elisabeth Lienert
Universität Bremen
Fachbereich 10
Universitäts-Boulevard 13
28359 Bremen
E-Mail: elienert@uni-bremen.de

Sektion 1: Digitalisierung von Handschriften und frühen Drucken, OCR

Gabriel Viehhauser

Digitalisierung von Handschriften und frühen Drucken, OCR (Bericht über Kurzvorstellungen und Diskussion Sektion 1)

Die automatische Erkennung von Handschriften und Drucken gehört wohl zu jenen Bereichen auf dem Gebiet der digitalen Methodik, in denen die Fortschritte, die in den letzten Jahren erzielt worden sind, am offenkundigsten (und auch für Laien ersichtlich) ins Auge fallen: Ähnlich wie bei der ebenso in jüngerer Zeit merkbar verbesserten Übersetzungssoftware werden diese Fortschritte weit über den akademischen Bereich hinaus wahrgenommen und sind letztlich dem *data-* bzw. *deep-learning-turn* geschuldet. Erst der Einsatz von elaborierten maschinellen Lernverfahren, denen große Mengen an Trainingsmaterial zugrunde liegen, hat hier zu den entscheidenden Verbesserungen im Vergleich zu früheren Versuchen geführt.

Für die Mediävistik ist die Handschriften- und Frühdruckerkennung neben ihrer praktischen Bedeutung von nicht zu unterschätzender grundlegender Signifikanz und hätte durchaus das Potential, einen Paradigmenwechsel in der Vorstellung von mittelalterlicher Schriftlichkeit an sich hervorzurufen: So erscheint es nun absolut in Greifweite, Handschriften in großer Zahl im Volltext zu digitalisieren und auch bisher unbeachtete Texte (oder Textvarianten) für literaturwissenschaftliche, linguistische und kulturwissenschaftliche Auswertungen zugänglich zu machen. Damit ergibt sich die Chance, auch solche Texte ins wissenschaftliche Blickfeld zu bringen, deren manuelle Erschließung bislang als zu aufwendig oder nicht

lohnend erschien. Bekanntermaßen beruhen solche Aufwandseinschätzungen nicht selten auf ästhetischen Werturteilen und berühren Fragen der Kanonbildung. Auch wenn es zu diskutieren bleibt, ob Kanonisierungsprozesse durch einen *Distant Reading*-Zugang im Sinne Franco Morettis (Moretti 2016) überwunden werden können (bzw. überhaupt sollten), so dürfte alleine die bloße Möglichkeit der Verfügbarkeit einer großen Masse von Texten zu Verwerfungen im Feld kanonischer Selbstverständlichkeiten führen. Der in der digitalen Editorik bereits ersichtliche Trend zur Überlieferungsnähe (und die sich daran anschließenden Debatten, inwieweit diese auch literaturwissenschaftlich relevant ist), dürfte sich jedenfalls, bedingt durch die Verbesserungen auf dem Gebiet der Handschriftenerkennung, weiter fortsetzen.

Zugleich zeichnet sich angesichts der Möglichkeiten dieser Verfahren aber auch ein Paradigmenwechsel oder, weniger radikal gedacht, eine Erweiterung im Spektrum philologischer Zugangsweisen ab: Automatische Handschriftenerkennung wird selbst bei besttrainierten Modellen zu einer gewissen Fehlerrate in der Erkennung führen. Ohne nachträgliche manuelle Korrektur (*post-processing*) wird also kein philologisch einwandfreier Text entstehen. Es stellt sich nun die Frage, ob sich die Altgermanistik darauf einlassen kann und will, schnell und kostengünstig automatisch erstellte, aber in Details womöglich fehlerhafte Textversionen zu akzeptieren, natürlich nicht als unbesehen übernommene Grundlage für eine philologisch exakte Edition, aber doch ergänzend zur überblicksmäßigen Auswertung und zur sinngemäßen Erschließung von Textkonvoluten, bei denen es vielleicht nicht in jedem Aspekt auf hundertprozentige Genauigkeit in der Textwiedergabe ankommt; oder aber als Ausgangsbasis für Volltextsuchen, bei denen dann ergebnisbezogen Fehler an Einzelstellen im Transkriptionstext verbessert werden können.

Auch wenn man nicht so weit gehen möchte, so bringen die Fortschritte in der Handschriften- und Frühdruckerkennen jedenfalls unbestreitbar neue Potentiale für die Texterschließung und Zugänglichmachung mit sich

und können die Arbeit etwa an digitalen Editionen und Textrepositorien erheblich erleichtern bzw. kostengünstiger machen. In der entsprechenden Sektion auf unserer Tagung wurden zwei aktuelle Systeme im Bereich der Handschriften- und Frühdruckerkennung vorgestellt, die sich besonders für den Einsatz in der Altgermanistik eignen und bereits in einigen Projekten zur Anwendung kommen, nämlich Transkribus und OCR4all.

Im Vortrag von Günter Mühlberger zu Transkribus standen insbesondere dessen Geschichte, das gegenwärtige Geschäftsmodell und typische Workflows im Vordergrund. Mühlberger berichtete von den Anfängen des OCR in der Frakturerkennung und davon, dass in den 2000er-Jahren noch ›traditionelle‹ Formen des OCR vorherrschten, bis dann der *data turn* das Feld revolutioniert habe. Ab 2013 habe auch das Vorgängerprojekt von Transkribus (tranScriptorium) *machine-learning*-Methoden zum Einsatz gebracht, mittlerweile sei das System in der Digital-Humanities-Szene (und darüber hinaus) stark verbreitet: Laut einer Auswertung von in der Fachliteratur erwähnten Tools gehörte Transkribus 2019 zu den meistverwendeten Programmen in den digitalen Geisteswissenschaften (vgl. https://weltliteratur.net/dh-tools-used-in-research/). 77000 User seien bislang registriert, pro Tag arbeiteten derzeit ca. 100 Nutzer und Nutzerinnen mit dem Werkzeug. Diese Zahl sei stetig am Steigen und wachse über den engeren Kreis der Digital Humanities hinaus (etwa in kommerzielle Bereiche oder in das Gebiet der Ahnenforschung).

Aus dieser Sachlage ergäben sich Anforderungen, die ab 2013/14 zur Erstellung eines neuen, in den Digital Humanities noch wenig verbreiteten Geschäftsmodells geführt haben, nämlich zur Einrichtung einer Genossenschaft. Diese Genossenschaft, die heute 102 (auch außereuropäische) Teilnehmer zählt, arbeite zwar profitorientiert, aber vor allem für den Selbsterhalt; es gebe in diesem Sinne keinen *shareholder value*. In Rechnung gestellt werde insbesondere die Texterkennung, dazu würden Einnahmen durch Kundenprojekte lukriert.

Transkribus stelle große, vortrainierte Modelle zur Verfügung, daneben könnten eigene Modelle trainiert werden, die auch öffentlich freigegeben werden könnten - aber nicht müssten. Mehr als 12.000 Modelle seien bereits von Nutzer und Nutzerinnen erstellt worden, der Schwerpunkt liege allerdings eher im 19. Jahrhundert und nicht in der Mediävistik.

Zum Abschluss seines Vortrags skizzierte Mühlberger Empfehlungen für Workflows bei der Anwendung von Transkribus. So sollten am Anfang der Projektarbeit umfangreiche Vorüberlegungen dazu angestellt werden, welches Ergebnis letztendlich erzielt werden soll: Welcher Grad an Genauigkeit der Transkription wird erwartet, ist eine Strukturauszeichnung mit TEI geplant bzw. wie tief soll diese erfolgen und schließlich, als Grundlage für all diese Entscheidungen, wieviel Aufwand kann und soll für die Erstellung der Umschriften betrieben werden? Sollen spezielle Zeichen und Abkürzungen verwendet werden (neuerdings ermöglicht Transkribus auch das Mittrainieren von Abkürzungszeichen und deren Auflösungen)? Wie viele Handschriften bzw. Drucke umfasst das Projekt und wie viel Text beinhalten diese jeweils? Vor der Arbeit sollte jedenfalls überprüft werden, ob öffentliche Modelle zugänglich bzw. deren Einsatz für die gewünschte Transkriptionsqualität ausreichend sind. Das selbständige Trainieren eines auf die eigenen Anforderungen spezialisierten Modells beginne sich bei Einzelmanuskripten ab 100 Seiten Länge zu lohnen.

In der an den Vortrag anschließenden Diskussion (Leitung: Freimut Löser) standen insbesondere Fragen zu den Modellen im Vordergrund, etwa welche speziellen Modelle (z. B. für Koberger) es bereits gebe (Martin Schubert), ob sich Modelle über unterschiedliche Editionsprojekte hinweg aggregieren ließen (Michael Stolz; aufgrund von überschneidenden Richtlinien ist dies problematisch) und ab wann es sinnvoll werde, ein eigenes Modell zu veröffentlichen (Gabriel Viehhauser; sinnvoll ab ca. 100.000 trainierten Wörtern). Bereits hier wurde deutlich, dass es wünschenswert wäre, den Austausch von mediävistischen Modellen durch verstärkte Koordination innerhalb des Fachs zu befördern. Schließlich richtete sich eine

Frage auf die Einsatzbarkeit von Transkribus für ideographische Sprachen (Meihui Yu; alle horizontal angeordneten Schriftzeichen können mittrainiert werden).

Im Vortrag von Christian Reul zu OCR4all standen vor allem arbeitspraktische Aspekte im Vordergrund. Im Gegensatz zu Transkribus ist OCR4all eine Open-Source-Software, die durch die unbeschränkte Replizierbarkeit nachhaltig und ausbaubar bleibe, dieser Vorteil werde aber durch die nicht-triviale Entwicklung erkauft. OCR4all soll insbesondere eine niederschwellige Plattform zur Verfügung stellen, welche in einer Live-Demo präsentiert wurde. Das Tool bietet neben der Schrifterkennung vor allem eine interaktive Oberfläche für die Nach-Korrektur des maschinell erkannten Textes. Maschinelle Erkennung und manuelles *post-processing* gehen damit Hand in Hand. Auch hier zeige sich, dass beim Einsatz von digitalen Methoden insbesondere durch die Kombination von automatischen und manuellen Verfahren die besten Ergebnisse erzielt werden können und Interfaces, die diesen Zusammenhang berücksichtigen, einen besonderen Nutzen erbringen.

In Hinblick auf die Leitfrage der Tagung („Was braucht das Fach?“) gab Reul zu bedenken, dass dies mitunter dem Fach selbst nicht klar sei. Jedenfalls sei immer ein *trade-off* zwischen Qualitätsanspruch und Aufwand zu berücksichtigen. Nicht zuletzt, um Verluste dieses *trade-offs* zu minimieren, solle, wenn möglich, zumindest auf gemischte Modelle zurückgegriffen werden, also eigenes Training von vorhandenen Modellen ausgehen.

Auch dieser Vortrag führte somit letztlich auf das Desiderat der Zusammenarbeit bei der Modellerstellung, denn je mehr Modelle vorhanden und für die Community verfügbar sind, desto positiver fällt die Kosten-Nutzen-Rechnung beim Einsatz von Schrifterkennungssystemen aus. In der allgemeinen Diskussion der beiden Vorträge wurde daher zunächst nach der Austauschbarkeit der trainierten Modelle zwischen Transkribus und OCR4all gefragt (Viehhauser). Zwar sind die beiden Systeme verschieden und in der Tiefe nicht kompatibel, Mühlberger wies jedoch darauf

hin, dass die zum Training erstellten Daten (also Digitalfaksimiles und deren zeilengenaue manuelle Umschrift) freilich identisch seien und ausgetauscht werden könnten.

Einen weiteren Diskussionspunkt stellte die Frage dar, ob sich die Systeme durch Einbeziehung von linguistischen Daten (wie etwa durch Wörterbuchabgleich oder Informationen über die Sprachstruktur) verbessern ließen (Elisabeth Lienert, Albrecht Hausmann). Automatische Nachkorrekturen schienen aber eher zu Verschlimmbesserung zu führen; auch der Einsatz von *transformer*-Sprachmodellen, die Kontextwahrscheinlichkeiten berücksichtigen, führe derzeit nur zu geringen Verbesserungen.

Damit war in der Diskussion schließlich die Frage nach dem Perfektionsgrad erreicht, der mit Hilfe von automatischen Verfahren angestrebt werden sollte: Inwieweit ist es sinnvoll, die mittlerweile ohnedies hohen Erkennungsraten aufwändig noch weiter zu verbessern, oder sollte nicht eher das Augenmerk auf die Verbesserung von Nachkorrektur-Tools gesetzt werden (Joachim Hamm)? Dass solche *post-correction*-Prozesse sinnvollerweise mit Erkennungsalgorithmen zusammengedacht werden sollten, demonstriere ja insbesondere das OCR4all-Projekt. Kurt Gärtner wies darauf hin, dass sich eine solche Postkorrektur auf alle Fälle lohne, und Andrea Rapp plädierte dafür, mehr Mut zur Publikation von Zwischenergebnissen zu zeigen, auch wenn diese nicht perfekt sind. Freimut Löser erinnerte daran, dass Transkribus (ebenso wie OCR4all) den Vorteil biete, die Texttranskription mit dem Digitalfaksimile zeilengetreu zu verbinden (was die Stellen in ihrer korrekten handschriftlichen Textgestalt leicht auffindbar macht). Ein fehlerfreier, perfekter Text sei ohnedies nie zu erreichen, und es wäre zu erwägen, ob die rasche Verfügbarmachung von großen Textmassen für die Volltextsuche einen perfekten Text überhaupt voraussetze. Unter Umständen könne also eine solche rasche Verfügbarmachung sinnvoller sein. Stephan Müller wies in diesem Zusammenhang darauf hin, dass die DFG Texterkennungsprojekte nach durchaus unterschiedlichen Kriterien für förderungswürdig erachtet. Wichtig sei jedenfalls, Güte-

klassen und Fehlerquotienten anzugeben. Auch Hausmann schloss sich dem Plädoyer für die Fehlertoleranz mit dem Hinweis an, dass für korpuslinguistische Explorationen Rohdaten und Textmassen ausschlaggebender seien als perfekte Editionen. Simone Schultz-Balluff berichtete, dass sie bei ihren eigenen Editionsprojekten (wie von Rapp angeregt) Daten schnell öffentlich zur Verfügung gestellt habe; diese Daten seien auch mit semantischen Auszeichnungen versehen worden.

Die Diskussion wurde schließlich durch Hinweise zu weiterführenden Perspektiven abgerundet: Jakub Simek fragte danach, wie sich Transkriptions-Tools in einen Editions-Workflow einbinden ließen. Mühlberger schlug die Integration über eine API vor und wies darauf hin, dass sich die Qualität der Erkennung nicht nur aufgrund der Datenmengen, sondern auch aufgrund technischer Weiterentwicklung in den nächsten zehn Jahren nochmals verbessern werde (woraus sich die Frage ergibt, ob zum Teil bereits automatisch erfasste Texte nochmals erschlossen werden sollen). Torsten Schaßan stellte schließlich das Projekt einer umfassenden digitalen Erschließung sämtlicher Handschriften, die sich im Besitz öffentlicher Einrichtungen befinden, im Rahmen der Handschriftenzentren in Aussicht und fragte danach, ob dabei eine Volltexterfassung von der Community gewünscht sei bzw. wie eine solche geplant und wo die Ergebnisse gespeichert werden sollten. Eine solche Perspektive macht nochmals deutlich, welches Potential sich aus der Anwendung digitaler Methoden im Bereich der Handschriften- und Frühdruckerkennung für die Altgermanistik ergeben könnte.

Literaturverzeichnis

Sekundärliteratur

Moretti, Franco: Distant Reading, Konstanz 2016.

Online-Ressourcen

OCR4all: http://www.ocr4all.org/.
tranScriptorium: https://cordis.europa.eu/project/id/600707.
Transkribus: https://readcoop.eu/transkribus/.
weltliteratur.net: https://weltliteratur.net/dh-tools-used-in-research/.

Anschrift des Berichterstatters:

Prof. Dr. Gabriel Viehhauser
Universität Stuttgart
Institut für Literaturwissenschaft
Herdweg 51
70174 Stuttgart
E-Mail: viehhauser@ilw.uni-stuttgart.de

Sektion 2: Digitale Edition

Jakub Šimek

heiEDITIONS – eine Heidelberger Infrastruktur für Editionen (nicht nur) mittelalterlicher Texte

Abstract. Die seit 2018 an der Universitätsbibliothek Heidelberg unter dem Namen heiEDITIONS entwickelte Infrastruktur für digitale Editionen baut auf Erfahrungen in der Digitalisierung, in der virtuellen Rekonstruktion von Sammlungen und in der Handschriftenerschließung auf und fügt sich als Teil der Heidelberg Research Infrastructure (heiRIS) zu den in Heidelberg strategisch vorangetriebenen Maßnahmen zur Förderung wissenschaftlichen Publikationswesens im Open Access. Besonderer Wert wird darauf gelegt, dass die Datenmodellierung und Visualisierung den vielfältigen Dimensionen der edierten Gegenstände gerecht werden. Gerade den Bedarfen der Altgermanistik kommt dabei eine wichtige Rolle zu.

Die Universitätsbibliothek Heidelberg ist seit 2014 auf dem Gebiet digitaler Editionen aktiv und baut dafür seit 2018 mit heiEDITIONS eine dedizierte Infrastruktur auf. Dieses Engagement ist eingebettet in die langfristige Entwicklungsstrategie des Hauses und hat damit u. a. die folgenden Hintergründe:

In den meisten digitalen Editionsprojekten kommt den Digitalisaten der Textträger eine wesentliche Rolle zu. Bei der Anfertigung von Digitalisaten – der Digitalisierung historischen Schriftguts – verfügt die Universitätsbibliothek über reiche Erfahrungen, die bis 2001 zurückreichen.[1] Unser Digitalisierungsmanagementsystem DWork (Heidelberger Digitalisierungsworkflow) ist seit Langem etabliert und wird auch von anderen Institutionen genutzt. Die Digitalisierung von historischen Textträgern, besonders

auch von den deutschsprachigen Handschriften der Bibliotheca Palatina, gehört an der UB zu den Schwerpunkten ihrer langfristigen Ausrichtung. Digitalisate der Textträger sind für die meisten digitalen Editionsprojekte ein wichtiger Bestandteil der Edition, nicht nur weil sie als offen einsehbare Editionsgrundlage für nachhaltige editorische Transparenz sorgen: Sie werden auch selbst zu einem Gut an sich und sichern die Textüberlieferung angesichts nicht auszuschließender künftiger Verschlechterung der Originale (oder gar eines Restrisikos ihres Verlusts) zusätzlich ab.

Ein anderer Aspekt sind unsere Erfahrungen mit der digitalen Zusammenführung historisch zusammenhängender Textsammlungen. Dabei handelt es sich einerseits um heute dislozierte Bestände ehemaliger und in der Vergangenheit zerstreuter Bibliotheken, die digital rekonstruiert werden können (hier sind die digitale Zusammenführung der Bibliothek des Reichsklosters Lorsch[2] und die kürzlich abgeschlossene Digitalisierung der heute außerhalb Heidelbergs aufbewahrten Handschriften der ehemaligen Bibliotheca Palatina[3] zu nennen). Andererseits geht es um Zusammenstellungen von historischen Textträgern eines Werkes, sodass die Parallelüberlieferung (die sich örtlich durchaus auch auf viele Standorte verteilen kann) in einer ›virtuellen Bibliothek‹ von einer zentralen Stelle aus überblickt werden kann. In solchen Fällen ist es ein ideell (literaturhistorisch) ausgemachtes ›Werk‹, das den Zusammenhang einer digitalen Sammlung stiftet. Als prominente Beispiele sind die virtuellen Bibliotheken der Editionsprojekte ›Welscher Gast digital‹ oder ›Iwein – digital‹ zu nennen.[4] Manchmal gelingt es dabei ebenfalls, heute örtlich zerstreute Teile eines ursprünglichen Textzeugen wieder zusammenzuführen.[5] Begleitend bietet es sich auch an, solche Sammlungen um Digitalisate moderner Abschriften und Editionen des jeweiligen Werkes zu ergänzen, was besonders dann von großem Wert ist, wenn der ursprünglich zugrundeliegende historische Textträger nicht mehr greifbar ist.

Des Weiteren sieht sich die Universitätsbibliothek Heidelberg in der Pflicht, ihre eigenen historischen Bestände nicht nur sicher aufzubewahren

und für die Zukunft zu erhalten, sondern sie auch kontinuierlich zu erschließen. Das erfolgt traditionell über die katalogisierende Beschreibung und neuerdings über die bereits erörterte Digitalisierung, darüber hinaus aber auch über die Förderung von Editionen. Dies als Teil des eigenen Auftrags zu betrachten, ist vielleicht nicht selbstverständlich, kann aber als ein konsequenter Folgeschritt nach der Bild-Digitalisierung gesehen werden, der neuerdings durch maschinelle Texterkennungsverfahren zusätzlichen Aufwind bekommt, da mit der Unterstützung von *machine learning* gewonnene Rohtextdaten zur Grundlage von Editionen werden können. Wenn wissenschaftliche Bibliotheken sich editorisch engagieren, kann man sich an das alte Selbstverständnis klösterlicher Skriptorien erinnert fühlen, die als ›Abteilungen‹ von Klosterbibliotheken für die Bewahrung und Weitergabe von Texten zu sorgen hatten; der steigende Bedarf der Wissenschaft an nachhaltigen und den Steuerzahler durch nicht-gewinnorientierte Preisgestaltung entlastenden Anbietern von digitalen Open-Access-Plattformen ist jedoch ganz und gar aktuell.

So wird unser Einsatz für Editionen dadurch begünstigt und fügt sich auch deswegen organisch zu unserer übrigen Hausstrategie, weil einer unserer Entwicklungsschwerpunkte das wissenschaftliche Publikationswesen ist. Die Verlage heiUP (Heidelberg University Publishing), heiBOOKS (Heidelberger E-Books), die Publikationsplattformen unserer Fachinformationsdienste arthistoricum.net (Fachinformationsdienst Kunst, zusammen mit der Sächsischen Landesbibliothek – Staats- und Universitätsbibliothek Dresden), Propylaeum (Fachinformationsdienst Altertumswissenschaften, zusammen mit der Bayerischen Staatsbibliothek München) und FID4SA (Fachinformationsdienst Südasien) sowie die zahlreichen von uns gehosteten Zeitschriften sind nur einige Belege für diesen Fokus unserer Arbeit.

Die genannten Umstände mögen das Engagement der Universitätsbibliothek Heidelberg, mit dem Aufbau ihrer Editionsinfrastruktur heiEDITIONS digitale Editionen zu ermöglichen, zu fördern und letztlich auch

auf Dauer zu betreiben, plausibel erscheinen lassen. Sie würden aber vielleicht nicht vollumfänglich einen solchen Vorstoß rechtfertigen, gäbe es für digitale Editionen nicht gewichtige wissenschaftsimmanente Gründe. Die bisherige praktische Erfahrung zeigt, dass die Erstellung digitaler Editionen durchaus teurer ist, länger dauert und mit größeren Arbeitsaufwänden verbunden ist als das Edieren über herkömmliche Druckausgaben. Eine Universitätsbibliothek, die wie die Heidelberger den Wert des gedruckten Buches nach wie vor hochhält und angesichts der Verantwortung für ihre historischen Bestände täglich daran erinnert wird, ihre Arbeit nicht an den nächsten paar Jahren, sondern an Jahrhunderten auszurichten, kann auch den Vorwurf des Kurzlebigen und des Ephemeren, der Online-Publikationen durchaus noch anhaften kann, nicht leichtfertig abtun – zumal die Edition generell als Publikationsart, welche überhaupt erst die Grundlagen für die meisten historischen Wissenschaften schafft, indem sie mit erheblichem, andere wissenschaftliche Publikationen meist weit übersteigendem Aufwand Primärtexte vorlegt, folgerichtig und allein schon aus ökonomischen Gründen die langlebigste Publikationsgattung sein müsste. Die berechtigte Forderung nach Zuverlässigkeit und Dauerhaftigkeit, die an zitierfähige Texteditionen gestellt wird, wird bei digitalen Editionen aktuell noch durch die Tatsache konterkariert, dass keine digitale Edition bisher den Nachweis erbringen konnte, über mehrere Jahrzehnte lang zuverlässig ihre Aufgabe zu erfüllen. Digitalen Editionen wird, wo auch immer sie vorangetrieben werden, ein Vertrauensvorschuss entgegengebracht, ihre Förderer müssen stärker als etwa bei klassischen Druckkostenzuschüssen in Vorleistung treten. Das Risiko großer Ressourcenverschwendung wäre untragbar, wenn nicht glaubhaft gemacht werden könnte, dass digitale Editionen einen Mehrwert haben (oder das Potenzial eines Mehrwerts), der die Risiken einer solchen paradigmatischen Neuentwicklung aufwiegt. Ein Teil dieses Mehrwerts rührt dabei von den informationslogistischen Vorteilen von Open Access und restriktionsfreier

weltweiter Erreichbarkeit, und auch dieser praktische Aspekt der Informationsversorgung sollte als intrinsischer Wert gewürdigt werden.

Eine Universitätsbibliothek, die als Dienstleister für die Wissenschaft agiert, sollte jedoch die Güte ihrer Dienstleistung nicht nur an Äußerlichkeiten der ›Lieferwege‹ messen, sondern in dem Grad, in dem sie selbst an der Methodik der wissenschaftlichen Leistung beteiligt ist, auch die wissenschaftsinhärenten Qualitäten der von ihr bereitgestellten Dienste auf den Prüfstand stellen, also fragen, inwiefern ein Dienst dem wissenschaftlichen Erkenntnisgewinn förderlich ist. Nach bisheriger Erfahrung ist die methodische Involvierung unserer Bibliothek in die digitalen Editionsprojekte, die wir als Partner der Wissenschaft betreiben, beträchtlich, da wir stärker als bei Druckpublikationen häufig am gesamten ›Lebenszyklus‹ einer digitalen Edition beteiligt sind, nicht zuletzt durch die Erfordernisse der Datenmodellierung an editionsmethodischen Entscheidungen partizipieren und diese durch vorbereitende Beratung und laufende Unterstützung mitprägen. Eine Infrastruktur wie heiEDITIONS, die vielen editorischen Gegenständen und Ansätzen ein Obdach bieten will, könnte ihre Aufgabe aber nicht auf skalierbare, zukunftsorientierte und letztlich wirtschaftliche Weise erfüllen, wenn sie nicht über einen einheitlichen konzeptuellen Rahmen verfügen würde, der die Heterogenität der Gegenstände und Ansätze in einer einheitlichen Matrix aufnehmen könnte. Eine Editionsinfrastruktur für viele ließe sich nicht betreiben, wenn jede Edition *a class of its own* wäre. Dadurch sind wir als Anbieter für die konzeptuelle Gestaltung des Dienstes verantwortlich, den wir für digitale Editionen entwickeln, und müssen solche Konzepte, die nicht selten genuin wissenschaftliche editionsmethodische Konsequenzen mit sich bringen, selbst kritisch hinterfragen.

Digitale Editionen haben die Chance, sich im Vergleich zu herkömmlicher Editorik in einem inhärent anspruchsvolleren Editionsparadigma zu verorten, das nicht nur zu einem besseren Arbeitsprozess des Edierens führt, sondern das neuen und höheren Ansprüchen zu genügen verhilft, die

an Editionen von den sie nutzenden Wissenschaften heute und künftig herangetragen werden. Mit ›digitaler Edition‹ meine ich nicht in erster Linie die Publikationsweise im *World Wide Web*, sondern beziehe mich vor allem auf die Freiheiten der digitalen Arbeitsweise sowie auf die Flexibilität und Ausdrucksmächtigkeit digitaler Datenmodelle, die sich nicht oder nicht ausschließlich von einer Endpräsentation ableiten, weder von einer druckorientierten typographischen Gestaltungsform noch von der einen oder anderen digitalen Anzeigemöglichkeit: Mein Fokus liegt auf einem Selbstverständnis der Edition, das sich primär an den Erfordernissen der Materie ausrichtet, also an der Frage, wie das zu Edierende am sachgerechtesten dokumentiert, erfasst, wiedergegeben, durchdrungen, erschlossen und natürlich letztendlich auch präsentiert werden kann – das Letztere aber nur als ein Aspekt von vielen. Sicher sind die *interfaces*, die Präsentationsformen, entscheidend für die Rezeption der Ergebnisse einer Edition, jedoch muss das Edieren – anders als vielfach bei der herkömmlichen Erstellung nur gedruckter Ausgaben – nicht von vornherein durch sie vorgegeben werden. Digitale Datenmodelle für Editionen können mehr fassen, als jede einzelne Präsentationsform wiedergeben kann. Das digitale Edieren eröffnet einem Herausgeber mehr als zuvor die Chance, sich bei der Konzeption einer Edition primär durch die Beschaffenheit der edierten Materie leiten zu lassen, indem womöglich mehr Dimensionen des Edierten zu ihrer vollen Geltung kommen können, als wenn eine vorgegebene Präsentationsform von Anfang an die editorische Methodik steuert. Wenn die New Philology in den um 1990 mehr erhofften als tatsächlich verfügbaren Möglichkeiten der Computertechnik begeistert eine Wende begrüßte, durch die es möglich würde, die Materialmenge einer Edition ohne einen durch den Druck vorgegebenen Auswahlzwang vollumfänglich zu präsentieren (also etwa alle Textzeugen in Digitalisaten und Transkripten),[6] so erkennen wir heute deutlicher, dass die Ausdrucksfähigkeit digitaler Datenmodelle Editionen neue qualitative Möglichkeiten eröffnet, ihrem Gegenstand umfassender gerecht zu werden.

Bei der Erstellung herkömmlicher gedruckter Ausgaben bestimmt häufig eine für lineare Lektüre zu optimierende, nach semantisch-logischen Einheiten anzuordnende Textgestaltung die editorische Arbeitsweise, wobei ein primär als sinntragender Sprachausdruck aufgefasster Text vom visuellen und materiellen Substrat seiner historischen Überlieferung abzulösen ist und Letzteres allenfalls beschreibend in die Edition Eingang finden kann. Alternativ kann auch in herkömmlicher Arbeitsweise – von der Faksimilierung einmal abgesehen – ›diplomatisch‹ ediert werden, wobei viel von den graph(emat)ischen und auch visuellen Eigenschaften der historischen Quelle abgebildet werden kann, die semantisch-logische Textebene aber unweigerlich zu kurz kommt. Auch eine linguistische Erschließung ist für herkömmliche druckorientierte Editionen etwa über (ggf. lemmatisierte) Konkordanzen möglich; die Benutzung bleibt dabei aber freilich meist statisch auf eine alphabetische Listenanordnung der Konkordanz angewiesen. Für inhaltliche Erschließung von *named entities*, intertextuellen Bezügen und Sachbegriffen haben sich in Druckeditionen Register als nützliches Werkzeug etabliert; ihre Funktionalität beschränkt sich aber vor allem auf eine Nutzungsrichtung (Suche im Register als Weg zu Fundstellen im Text), und die Tragweite von Registern kann kaum über ein einzelnes ediertes Werk hinausreichen. Für die Abbildung der Überlieferungsvarianz verfügen herkömmliche Editionen über das Mittel des *apparatus criticus*, dessen Grenzen bekannt sind: Sollen sie ihren Nutzer nicht durch unüberblickbaren Umfang erdrücken, werden sie zu Schaukästen ausgewählter Lesarten, bei deren Rezeption der Leser dem Herausgeber einen erheblichen Vertrauensvorschuss gewähren muss; das Nachvollziehen einer Lesart im Originalkontext eines Textzeugen ist schwer möglich.

Die gerade aufgezählten Ebenen des Edierten – neben der semantisch-logischen Ebene der Sinneinheiten die visuell-materielle Ebene des Textträgers, die linguistische Ebene, die extensionale Ebene der Verweise auf textexterne Dinge (›Extrareferenzialität‹) sowie die Ebene der Überlieferungspluralität und somit der Varianz – sind auch in herkömmlichen

Editionen bekannt. Die traditionellen Lösungen für den Umgang mit ihnen werden ihnen aber jeweils nur partiell gerecht, weil es weder in einem herkömmlichen Editionsmanuskript des Herausgebers noch in einer einzigen statischen Druckform möglich ist, ihnen jeweils umfassend Rechnung zu tragen.

Digitale Editionsdatenmodelle können vollwertige Räume für mehrdimensionale Texterschließung schaffen: Die Qualität einer Editionsinfrastruktur, die für viele Editionsgegenstände und -paradigmen ein angemessenes Zuhause bieten soll, wird sich daran messen müssen, wie sachgemäß und ausdrucksfähig die Lösungen sind, welche die Infrastruktur für einen integrierten Umgang mit den genannten Ebenen bereitstellen kann, sowohl in der Datenmodellierung als auch in der Präsentation und nicht zuletzt in der Datenerarbeitung, also beim *authoring* der Edition, in Hilfsmitteln für den Arbeitsprozess der Herausgeber.

Die Heidelberger Editionsinfrastruktur heiEDITIONS wird mit Blick auf solche Ansprüche entwickelt, wobei für einige der genannten Aspekte bereits Lösungen existieren oder in Sicht sind, andere noch nicht über Experimente (wie den Umgang mit der linguistischen Annotation) oder erste Schritte (Erleichterungen beim *authoring*) hinausreichen.

Ein grundlegendes Problem ist die Unterscheidung der genannten (traditionell primären) semantisch-logischen Textebene, in der ein Text meist als eine mehr oder weniger hierarchisch strukturierte, geordnete Abfolge von Sinneinheiten wie Kapiteln, Strophen, Prosaabsätzen oder Versen aufgefasst wird (wobei in Verstexten die Situation zusätzlich dadurch verkompliziert wird, dass formal-poetische Einheiten wie Verse auch nicht zwingend mit Sinneinheiten zusammenfallen), von der Ebene der visuellen Einheiten von Textträgern, die im Wesentlichen durch Seitenflächen, Layoutbereiche und Zeilen konstituiert wird. Diese beiden Ebenen interagieren miteinander auf vielfältige Weise. Häufig gehen sie miteinander Symbiosen ein (Beginn eines Kapitels auf einer neuen Seite; Absetzung von Versen) und machen gegenseitig voneinander ›Gebrauch‹, in anderen Fällen schei-

nen sie sich neutral zueinander zu verhalten und überlappen einander ohne erkennbare Rücksicht, gelegentlich laufen sie aber auch einander zuwider (Zeilenüberläufe, bei denen der Text eines inhaltlichen Segments in einer ›fremden‹ Zeile fortgesetzt wird; Briefe, deren Text nach der letzten Seite auf den eigentlich vorangegangenen Seiten fortgeschrieben und am Rand der ersten Seite abgeschlossen wird). Aus mittelalterlichen Handschriften zur Genüge bekannt sind auch diverse Formen von Marginalien, welche die Linearität des Textstroms durch Anmerkungen unterbrechen oder durch Einschübe bewirken, dass der inhaltlich zu lesende Text visuell immer wieder von einem Layoutbereich in einen anderen springt.

Wenn die visuelle Struktur des Textträgers ernst genommen werden soll, um etwa Seite für Seite – in Korrelation zum Digitalisat – eine diplomatisch zeilengetreue Textanzeige zu bewerkstelligen oder um überhaupt erst einen Rohtext mithilfe von Texterkennungssystemen[7] zu erarbeiten und das dabei gewonnene Text-Bild-Alignment auch bei weiterer editorischer Textaufbereitung nicht zu verlieren, braucht es Mechanismen, mit denen die visuelle Struktur unabhängig von der Sinnstruktur erfasst werden kann. Auch für die editorische Neuanordnung der zu lesenden Textabfolge im potenziellen Widerspruch zur Textreihenfolge der Quelle sind Lösungen notwendig.

Die Datenmodellierung in heiEDITIONS basiert zum allergrößten Teil auf der Text Encoding Initiative (TEI), bei der die semantisch-logische Textstruktur grundsätzlich (und auch historisch) Vorrang hat (vgl. Burnard 2013, Abschnitt 13). Um eine von Sinnstrukturen unabhängige Erfassung visueller Einheiten zu ermöglichen, sind in heiEDITIONS TEI-konforme Erweiterungen vorgesehen, die eine vollständige Trennung der Ebenen ermöglichen. Darauf basierend sieht auch die Visualisierung in heiEDITIONS zwei grundverschiedene Ansichten (*views*) der Editionstexte vor (die selbstredend aus einer einzigen Datenbasis erzeugt werden): In der ›Leseansicht‹ wird ein zur Lektüre sinnvoll angeordneter und nach Sinneinheiten strukturierter Text präsentiert, der keinerlei Layoutaspekte der histori-

schen Quelle nachbildet, nur punktuell die Seitenanfänge der Quelle im Text anzeigt und zu diesen im Digitalisat verlinkt. Die Navigation im Text orientiert sich hier an inhaltlichen Abschnitten wie Kapiteln. In der ›Quellenansicht‹ hingegen bestimmen die visuellen Strukturen des Textträgers die Navigation und die Textanzeige. Der Leser ›blättert‹ durch das Digitalisat des Textträgers und bekommt seitenweise einen Text präsentiert, der wie in der Quelle in Layoutbereichen und Zeilen untergebracht ist, einschließlich der positionsgetreuen Anordnung von Spalten und Marginalien.

Die jeweils spezifische Leistungsfähigkeit der ›Quellenansicht‹ und der ›Leseansicht‹ geht weit darüber hinaus, den Text einmal mit Originalumbruch und einmal ohne diesen einzublenden, und der Unterschied zwischen den beiden Ansichten hat auch wenig damit zu tun, dass der Text auf Zeichenebene einmal originalgetreu und einmal z. B. normalisiert oder mit übernommenen editorischen Korrekturen erscheint. Die Textdarstellung auf Zeichenebene ist (je nach editorischer Erschließungstiefe) grundsätzlich für beide ›Ansichten‹ dynamisch einstellbar. Die ›Quellenansicht‹ und die ›Leseansicht‹ unterscheiden sich vielmehr in der Makroanordnung des Textes oberhalb der Zeichenebene, und sie sind in der Lage, den Text nicht nur jeweils anders hierarchisch strukturiert, sondern auch in einer jeweils anderen Reihenfolge, ggf. auch komplett umgeordnet, zu präsentieren. Somit wird es künftig auch für Editionen, die mit Texterkennungssystemen arbeiten und dabei zeilenweise Koordinateninformationen im Digitalisat gewinnen, möglich sein, diese enge Kopplung an das Digitalisat der Quelle beizubehalten und den Text bei Bedarf dennoch nach inhaltlichen Gesichtspunkten editorisch neu zu sortieren.

Für die inhaltliche Erschließung der in Texten erwähnten Personen, Orte, Ereignisse, Werke, Begriffe u. Ä. bietet heiEDITIONS eine Registerfunktionalität, die an herkömmliche Buchregister angelehnt ist, aber im Vergleich zu ihnen wesentliche Mehrwerte mit sich bringt. Zum einen ist die Nutzungsrichtung nicht auf eine Suche nach Textfundstellen in einer Liste beschränkt, da annotierte Textstellen auch als solche in der visuali-

sierten Edition direkt sichtbar sind und beim Anklicken weiterführende Informationen bieten (perspektivisch auch direkt eine Auflistung weiterer Fundstellen derselben Sache, die momentan noch lediglich über einen ›Listeneinstieg‹ des Registers zugänglich ist). Zum anderen können heiEDITIONS-Register zu einer Entität viel mehr als nur eine Ansetzungsform des Namens enthalten: Personen können mit Lebensdaten, Orte mit Lageinformationen, jeder Eintrag außerdem mit beliebigen Erläuterungen, Links (etwa zu Wikipedia oder zu speziellen Internetressourcen), bibliographischen Angaben und Bildern versehen werden, sodass solche Register auch zu kleinen Nachschlagewerken oder Glossaren werden und bei geographischen Daten die Form von Gazetteers annehmen können. Darüber hinaus sind Verknüpfungen zu Normdateien wie der GND oder GeoNames möglich und erwünscht, über die weitere Angaben aus solchen kollaborativ erarbeiteten zentralen *knowledge bases* zur Anzeige gebracht werden können und die inhaltliche Erschließung von Editionstexten künftig auch für projektübergreifende Fragestellungen verwertet werden kann.

Dem Aspekt der Überlieferungspluralität edierter Texte begegnet heiEDITIONS mit zwei Ansätzen, die beide an traditionelle typographische Lösungen anknüpfen, dabei aber deutliche funktionale Verbesserungen mitbringen: Synopsen und textkritischer Variantenapparat.

(1) Die synoptische Nebeneinanderstellung mehrerer Volltexte, die an sich in der Schriftkultur sogar noch vor der Erfindung des Codex als Mittel zur Darbietung von Textversionen und zur Erleichterung der Rezeption von Textvarianz angewandt wurde – man denke an die ›Hexapla‹ des Origenes –, ist auch in der modernen editorischen Druckwelt weit verbreitet, bleibt aber durch physische Formate und die Notwendigkeit, statisch endgültig festzulegen, was miteinander visuell korrelieren und daher nebeneinander gesetzt werden soll, eingeschränkt. Die digitale Datenmodellierung macht es möglich, die Korrelation zwischen Teilen von Textversionen präziser zum Ausdruck zu bringen, etwa dann, wenn ›asymmetrische‹ Korrelationen zwischen jeweils ungleicher Anzahl von Sinneinheiten (z. B.

ein Vers korreliert mit einem Verspaar) beschrieben werden sollen oder wenn es darum geht, die hypothetische Position eines Textsegments in einem Textzeugen, in dem das betroffene Textsegment fehlt, im Vergleich zu anderen Textversionen sinnvoll zu verorten. Die dynamische digitale Präsentation gestattet es wiederum überhaupt erst, einerseits die überlieferte Textabfolge eines Textzeugen in originaler Anordnung zu präsentieren und andererseits auf Nutzerwunsch für jede Sinneinheit die damit in anderen Textversionen jeweils korrelierenden Stellen direkt nebeneinander anzuzeigen – für Texte mit stark variabler Anordnung wie den ›Armen Heinrich‹[8] (aber grundsätzlich für jede Mehrfachüberlieferung mit Umstellungen) mit Mitteln statischer Drucksynopsen ein Ding der Unmöglichkeit.

heiEDITIONS bietet für diese Herausforderung eine Lösung in Form der sog. ›Fokuszeile‹:[9] Der Nutzer wählt in einer beliebigen Textversion in einer synoptischen Spalte ein Textsegment, woraufhin sich alle anderen Versionen (in anderen Spalten) an diesem Segment mit ihren jeweils korrelierenden Textstellen vertikal ausrichten (einschließlich identifizierter vermuteter Fehlstellen) und so gleichsam eine Zeile bilden, in der die gegenseitige Korrelation gewährleistet ist, während oberhalb und unterhalb dieser ›Fokuszeile‹ die Textversionen in ihrer jeweiligen Originalanordnung bleiben, sodass sich bei Umstellungen, Hinzufügungen oder Fehlstellen außerhalb der Fokuszeile nicht mehr zwingend eine Korrelation im Nebeneinander ergibt. Die Festlegung der ›Auflösungsstufe‹ der Synopse (metaphorisch in Analogie zur Bildauflösung), also der hierarchischen Ebene der Sinneinheiten, deren gegenseitige Korrelation durch die Synopse ausgedrückt werden soll (z. B. Verse versus Strophen), obliegt dem Herausgeber.

(2) Auch für die digitale Modellierung des Variantenapparats wäre eine präzise Beschreibung der Korrelation möglichst auf Tokenebene die eigentlich idealerweise anzustrebende Lösung, denn dann wäre für jedes beliebige Textsegment in jedem Textzeugen aus den Editionsdaten abzulesen, welche Tokens in jedem anderen Textzeugen die jeweilige Entsprechung

bilden. Allerdings wäre so nur gesagt, was einander entspricht und nicht, wie beschaffen eine Entsprechung ist: Die eigentlich relevante Frage nach der Art der etwaigen Varianz wäre damit noch nicht beantwortet. Für die digitale Modellierung einer derart feinen Korrelationsbeschreibung zwischen Textversionen auf Tokenebene wäre der Ansatz des *variant graph*[10] auch für heiEDITIONS vielversprechend, da erst das Graphmodell (im Gegensatz zu tabellarischen Strukturen) die erforderliche Ausdrucksmächtigkeit angesichts von Umstellungen in Aussicht stellt. Die Art der Entsprechung und damit die Beschaffenheit der Varianz zu beschreiben, ist aber schwieriger; dies könnte auf Tokenebene bis zu einem gewissen Grad durch linguistische Annotation möglich gemacht werden (was eine maschinelle Auswertung der Varianz nach linguistischen Kategorien wie der Lexik oder *part of speech* erlauben würde). Eine vollwertige semantische Beschreibung der Varianz – sei es von größeren variierenden Wortgruppen oder in inhaltlichen Nuancen – ist so aber kaum vorstellbar (wobei diese Einschränkung den Wert der linguistischen Annotation nicht mindern soll, die auch in heiEDITIONS perspektivisch angestrebt wird).

Für die greifbare Zukunft realistischer erscheint uns für den digitalen editorischen Umgang mit Varianzaufbereitung ein an herkömmliche Formen angelehnter Variantenapparat, der einerseits von den bekannten typographischen Konventionen des *apparatus criticus* ausgeht, sich andererseits aber die digitale Verfügbarkeit der vollständig transkribierten Paralleltexte (neben einem zum Leittext gewählten Textzeugen) voll zunutze macht – natürlich nur dann, wenn eben tatsächlich die miteinander kollationierten Texte als Volltranskripte vorliegen (und somit in der Synopse auf einer gröberen Ebene gewählter Sinneinheiten sozusagen schon ein positiver Variantenapparat gegeben ist, z. B. versweise). Wenn man dann nach erfolgter philologisch-intellektueller Kollation als Herausgeber die für die Präsentation im Apparat (nach eigenen Kriterien) ausgewählten Lesarten angäbe, würde man diese Lesarten nicht als Zeichenketten aus dem Paralleltext abtippen oder kopieren, sondern man würde eine Verknüpfung

zwischen der als Lemma festgelegten Passage des Leittextes und einem Segment des Paralleltextes setzen.

Daraus ergäbe sich nicht nur eine effiziente Arbeitsweise, die gleichzeitig unnötige Fehler von selbst vermiede, sondern die so mit dem Leittext verknüpften Lesarten könnten ebenso dynamisch anpassbar zur Anzeige gebracht werden wie der Leittext selbst (mit Darstellungsoptionen wie dem einstellbaren Umgang mit Abkürzungen). Die angezeigten Lesarten könnten zudem in einer Online-Visualisierung mit Verlinkungen in die Volltexte der Parallelversionen unterlegt werden, sodass man beim Anklicken einer Lesart zur entsprechenden Stelle im Volltranskript des abweichenden Textzeugen gelänge.[11]

Da die Qualität der TEI-basierten Datenmodelle in heiEDITIONS direkt mit der Eignung der Infrastruktur zusammenhängt, die skizzierten editionsrelevanten Textdimensionen editorisch abzubilden, kommt der Datenmodellierung und dem Datenmodellmanagement ein hoher Stellenwert zu. Zudem müssen Aufgaben der Datenmodellierung auch den ökonomischen Anforderungen an Beherrschbarkeit und Skalierbarkeit selbst bei großer Heterogenität editorischer Paradigmen und Gegenstände standhalten. Schließlich sollen die in Editionsprojekten erarbeiteten TEI-Daten zuverlässig und robust visualisiert werden können, ohne dass für jedes einzelne Projekt *ad hoc*-Lösungen erfunden werden müssten – solche sind vielmehr nach Möglichkeit zu vermeiden, zugunsten einer zentralen und allen Projekten gemeinsamen technischen Infrastruktur. Nur eine Plattform, deren modulare Komponenten mit ihren Verarbeitungs- und Visualisierungstools vergleichbaren Prinzipien folgen, kann selbst bei wachsender Komplexität und Anzahl der Editionen beherrschbar bleiben. Dank einer möglichst homogenen Technik werden einzelne Projekte auch von künftigen Weiterentwicklungen des Systems profitieren und ggf. in neue Formate migriert werden können. Die Einheitlichkeit der Systeme erfordert dann ihrerseits eine Einheitlichkeit der Daten, die sich der Datenmodellierung verdankt.

In heiEDITIONS wird für diese Aufgabe ein dediziertes Set an Tools eingesetzt, mit deren Hilfe ein TEI-konform aufgebautes und für die eigenen Bedarfe angepasstes XML-Schema (eine sog. *customization* der TEI: heiEDITIONS Schema), ein als Ontologie im Sinne des Semantic Web strukturiertes Kategoriensystem (heiEDITIONS Concepts) sowie Anleitungen und eine technische Dokumentation (heiEDITIONS Dokumentation) gepflegt werden.[12] Zum Einsatz kommen dabei in erster Linie die von der TEI selbst bereitgestellten Werkzeuge; eine wesentliche Erweiterung stellt die Einbindung des auf RDF (Resource Description Framework) und OWL (Web Ontology Language) basierenden Kategoriensystems heiEDITIONS Concepts dar, das in den konzeptuellen Modellen CIDOC CRM und FRBRoo[13] verankert ist und für eine breite Palette überlieferungsbezogener und editorischer Phänomene normierte Definitionen und Identifikatoren bietet, auf die in TEI-Dokumenten direkt Bezug genommen wird. Somit wird das in TEI-Daten zu verwendende taxonomische System an einer zentralen Stelle gepflegt; bei immer wieder erforderlichen Ergänzungen werden neue Konzepte in das bestehende Kategoriensystem so eingepflegt, dass sie sich möglichst in den bestehenden Rahmen fügen.

Ziel der Datenmodellmanagement-Komponente von heiEDITIONS ist eine interne Interoperabilität[14] innerhalb des Hauses, damit einmal entwickelte Verarbeitungs- und Präsentationswerkzeuge unter Vermeidung unnötiger Redundanzen über Projektgrenzen hinweg genutzt werden können und Neuerungen bei kontinuierlicher Weiterentwicklung möglichst vielen Projekten zugute kommen.

heiEDITIONS ist an der Universitätsbibliothek Heidelberg eingebettet in heiRIS (Heidelberg Research Infrastructure) – ein breit gefächertes Ökosystem miteinander lose gekoppelter Dienste, die von der DOI-Vergabe über die Speicherung audio-visueller Medien bis hin zur etwaigen Buchpublikation, zur bibliothekarischen Katalogisierung einer Edition und zur Langzeitarchivierung der Editionsdaten reichen (zu heiEDITIONS im Kontext von heiRIS vgl. Effinger [u. a.] 2019). Alle Editionen nutzen für die Er-

werbung, Verwaltung und Präsentation von Digitalisaten das bewährte, anfangs bereits genannte Heidelberger Digitalisierungssystem DWork.

Nicht alle Aspekte von heiEDITIONS sind für die Altgermanistik relevant, da die Infrastruktur ein breites Spektrum editorischer Gegenstände bedient, zu denen auch moderne Gelehrtenkorrespondenzen oder Editionen aus dem Umfeld der Kunstgeschichte gehören. Dennoch hat die Altgermanistik in heiEDITIONS eine besondere Stellung inne, und die Weiterentwicklungsperspektiven der Infrastruktur orientieren sich – meist induktiv in strategisch ausgewählten Pilotprojekten – ausdrücklich auch am Bedarf des Faches.

Anmerkungen

1 Zum ersten Digitalisierungsprojekt ›Digitalisierung spätmittelalterlicher Handschriften aus der Bibliotheca Palatina‹, das mit DFG-Förderung in den Jahren 2001–2002 stattfand, und den damaligen Perspektiven vgl. Effinger [u. a.] 2000; Saurma-Jeltsch/Effinger 2001; sowie Effinger [u. a.] 2003.

2 ›Bibliotheca Laureshamensis – digital: Virtuelle Klosterbibliothek Lorsch‹.

3 ›Bibliotheca Palatina – digital: Virtuelle Rekonstruktion der einst berühmtesten Büchersammlung Deutschlands‹; vgl. Probst 2017.

4 Im altgermanistischen Kontext sind ferner die Sammlungen der Editionsprojekte ›Kaiserchronik – digital‹, ›Der arme Heinrich – digital‹, ›Lübisches Recht – digital‹, und ›Nikolaus von Jeroschin – digital‹ zu nennen. Im letztgenannten Fall handelt es sich sogar eher um eine Autorsammlung, da neben der ›Krônike von Pruzinlant‹ auch das Fragment der ›Adalbert-Vita‹ zum Angebot gehört.

5 So etwa bei Fr1, Fr2 und Fr8 der ›Krônike von Pruzinlant‹ des Nikolaus von Jeroschin (https://digi.ub.uni-heidelberg.de/nvjd/jeroschin/handschriften.html).

6 Vgl. Cerquiglini 1989, S. 112–116; kritisch dazu Stackmann 1993, S. 8f., und ders. 1994, S. 418.

7 Besonders relevant für Editionen sind aktuell im Bereich der Texterkennung (OCR – *optical character recognition* bzw. HTR – *handwritten text recognition*) die Programme Transkribus und eScriptorium. Die UB Heidelberg erwägt perspektivisch eine Integration der Open-Source-Software eScriptorium in ihre Texterkennungsworkflows und betreibt eine Testinstanz des Systems. Nicht

zuletzt die Gewinnung des Text-Bild-Alignments (Koordinaten von Textsegmenten auf digitalen Bildern) halten wir dabei auch im Editionskontext für attraktiv.

8 Vgl. Fernández Riva/Millet 2022.

9 Unser Konzept der ›Fokuszeile‹ wurde unter maßgeblicher Beteiligung von Leonhard Maylein, der auch für die technische Umsetzung verantwortlich zeichnete, im Rahmen der Edition ›Der arme Heinrich – digital‹ erstmalig entwickelt und wird derzeit weiter verbessert.

10 Zum *variant graph* s. Schmidt/Colomb 2009. In abgewandelter Form ging dieses Graphmodell in die Softwareprojekte CollateX und Stemmaweb ein, vgl. dazu z. B. Dekker [u. a.] 2014 und Andrews 2014. Dabei werden Tokens, die quer über mehrere Textzeugen als gleich betrachtet werden, jeweils durch einen Knoten repräsentiert; die Kanten drücken aus, in welchen Textzeugen die durch eine Kante verbundenen Knoten vertreten sind; die Richtung der Kanten steht für den Textfluss. Gewisse Schwächen des Modells in den bisher vorgestellten Implementierungen bestehen aus meiner Sicht darin, dass bei asymmetrischer Korrelation von Tokens (wenn jeweils eine ungleiche Anzahl von Tokens einander entspricht) nicht präzise zum Ausdruck gebracht wird, welche Tokens bzw. Tokengruppen einander entsprechen, wenn in den verglichenen Textversionen in direkter Nachbarschaft zu einer Korrelationsstelle gleichzeitig auch unbeteiligte Hinzufügungen oder sonstiges Fremdmaterial (Interpunktion, Glossen, Paratexte) vorlägen; ferner dass keine ›Schachtelung‹ innerhalb miteinander korrelierender Textsegmente abgebildet werden kann (Korrelation auf mehreren Ebenen bzw. in unterschiedlichen Präzisionsstufen); und schließlich dass noch nicht an die Abbildung von Varianz unterhalb der Tokenebene gedacht wurde (vgl. hypothetisch ›hellblau‹ vs. ›hellrot‹ vs. ›dunkelblau‹ vs. ›dunkelrot‹). Dessen ungeachtet gibt ein *variant graph* die Variation zwischen Textversionen in ihrem Textfluss sehr gut wieder.

11 Die skizzierte Form des Variantenapparats soll im Rahmen des Editionsprojekts ›Boners *Edelstein* – digital‹ realisiert werden, das die UB Heidelberg zusammen mit Gerd Dicke als wissenschaftlichem Herausgeber plant. Die Digitalisate aller Textzeugen sind bereits im Open Access verfügbar.

12 Die sukzessive erarbeitete Dokumentation von heiEDITIONS ist in Teilen bereits unter https://heieditions.github.io zugänglich.

13 CIDOC CRM ist ein konzeptuelles Referenzmodell für die Beschreibung von Gegenständen des kulturellen Erbes; FRBRoo ist eine mit CIDOC CRM harmonisierte Form des bibliothekarischen Referenzmodells FRBR (›Functional Re-

quirements for Bibliographic Records‹). U. a. die FRBR-Kategorien ›Werk‹, ›Expression‹, ›Manifestation‹ und ›Exemplar‹ erweisen sich auch bei der Modellierung von Editionsdaten als operational nützlich, etwa wenn es zu unterscheiden gilt zwischen singulären Phänomenen einer materiellen Quelle und dem, was einen Text auch unabhängig von einem Träger ausmacht, oder wenn multiple Textversionen einem gemeinsamen ›Werk‹ zugeschrieben werden. Freilich entscheidet immer letztlich der Herausgeber, was er editionspraktisch für ein ›Werk‹ hält.

14 Die Interoperabilität (direkte technische Nachnutzbarkeit) der heiEDITIONS-gemäß gestalteten TEI-Daten wird tatsächlich nur im Rahmen der hauseigenen Infrastruktur bezweckt, darüber hinaus ist sie kaum realistisch, weil die TEI große Spielräume bei konkreter Anwendung zulässt und Anpassungen durch ihre Nutzer geradezu vorsieht. Gleichzeitig legen wir großen Wert darauf, dass die TEI-Daten von heiEDITIONS mit anderen Interessierten austauschbar sind, d. h. dass sie transparent und verständlich sind und damit von Dritten mit vernünftigem Aufwand wiederverwendet werden könnten. Zur Diskussion der Begriffe *interoperability* und *interchange* im TEI-Kontext vgl. Holmes 2016.

Literaturverzeichnis

Sekundärliteratur

Andrews, Tara L.: Analysis of variation significance in artificial traditions using Stemmaweb, in: Digital Scholarship in the Humanities 31 (2016) [im Advance Access bereits 2014], S. 523–539 (online).

Burnard, Lou: The evolution of the Text Encoding Initiative. From research project to research infrastructure, in: Journal of the Text Encoding Initiative 5 (2013) (online).

Cerquiglini, Bernard: Éloge de la variante. Histoire critique de la philologie, Paris 1989.

Dekker, Ronald Haentjens/van Hulle, Dirk/Middell, Gregor/Neyt, Vincent/van Zundert, Joris: Computer-supported collation of modern manuscripts. CollateX and the Beckett Digital Manuscript Project, in: Digital Scholarship in the Humanities 30 (2015) [im Advance Access bereits 2014], S. 452–470 (online).

Effinger, Maria/Maylein, Leonhard/Pietzsch, Eberhard/Spyra, Ulrike: Per Mausklick ins Spätmittelalter. Digitalisierung und Erschließung spätmittelalterlicher Bilderhandschriften aus der Bibliotheca Palatina, in: BIT online 6 (2003), S. 235–247 (online).

Effinger, Maria/Maylein, Leonhard/Šimek, Jakub: Von der elektronischen Bibliothek zur innovativen Forschungsinfrastruktur. Digitale Angebote für die Geisteswissenschaften an der Universitätsbibliothek Heidelberg, in: Bibliothek Forschung und Praxis 43 (2019), S. 311–323 (online).

Effinger, Maria/Saurma-Jeltsch, Lieselotte E./Pietzsch, Eberhard: Deutsche Forschungsgemeinschaft fördert Projekt »Digitalisierung spätmittelalterlicher Bilderhandschriften aus der Bibliotheca Palatina«, in: Theke (2000), S. 47–50.

Fernández Riva, Gustavo/Millet, Victor: ›Verschiedenheit‹ der Handschriften. Über Varianz in der Überlieferung des ›Armen Heinrich‹ Hartmanns von Aue. Mit einer vollständigen Verskonkordanz, in: ZfdA (2022), S. 291–321 (online).

Holmes, Martin: Whatever happened to interchange?, in: Digital Scholarship in the Humanities 32 (2017) [im Advance Access bereits 2016], S. i63–i68 (online).

Probst, Veit: Digitization at the Heidelberg University Library. The digital Bibliotheca Palatina project, in: Digital Philology 6 (2017), S. 213–233 (online).

Saurma-Jeltsch, Lieselotte E./Effinger, Maria: Forschung per Mausklick, in: Ruperto Carola 3 (2001), S. 4–12.

Schmidt, Desmond/Colomb, Robert: A data structure for representing multi-version texts online, in: International Journal of Human-Computer Studies 67 (2009), S. 497–514 (online).

Stackmann, Karl: Die Edition – Königsweg der Philologie?, in: Bergmann, Rolf/Gärtner, Kurt (Hrsg.): Methoden und Probleme der Edition mittelalterlicher deutscher Texte, Tübingen 1993 (Beihefte zu editio 4), S. 1–18.

Stackmann, Karl: Neue Philologie?, in: Heinzle, Joachim (Hrsg.): Modernes Mittelalter. Neue Bilder einer populären Epoche, Frankfurt a. M./Leipzig 1994, S. 398–427.

Online-Ressourcen

arthistoricum.net. Fachinformationsdienst Kunst: https://www.arthistoricum.net.

Bibliotheca Laureshamensis – digital: Virtuelle Klosterbibliothek Lorsch: https://www.bibliotheca-laureshamensis-digital.de.

Bibliotheca Palatina – digital: Virtuelle Rekonstruktion der einst berühmtesten Büchersammlung Deutschlands: https://digi.ub.uni-heidelberg.de/de/bpd/index.html.

Boners *Edelstein* – digital: https://doi.org/10.11588/edition.bed.

CIDOC CRM (Conceptual Reference Model): https://www.cidoc-crm.org.

Der arme Heinrich – digital: https://doi.org/10.11588/edition.ahd.

DWork (Heidelberger Digitalisierungsworkflow): https://www.ub.uni-heidelberg.de/helios/digi/dwork.html.

eScriptorium: https://escripta.hypotheses.org.
FID4SA (Fachinformationsdienst Südasien): https://www.fid4sa.de.
FRBRoo (Functional Requirements for Bibliographic Records): https://cidoc-crm.org/frbroo.
GeoNames: https://www.geonames.org/.
GND (Gemeinsame Normdatei): https://gnd.network.
heiBOOKS (Heidelberger E-Books): https://books.ub.uni-heidelberg.de/heibooks.
heiEDITIONS (Heidelberger digitale Editionen): https://heieditions.github.io.
heiUP (Heidelberg University Publishing): https://heiup.uni-heidelberg.de.
Iwein – digital: https://doi.org/10.11588/edition.iwd.
Kaiserchronik – digital: https://doi.org/10.11588/edition.kcd.
Lübisches Recht – digital: https://doi.org/10.11588/edition.lrd.
Nikolaus von Jeroschin – digital: https://doi.org/10.11588/edition.nvjd.
Propylaeum. Fachinformationsdienst Altertumswissenschaften: https://www.propylaeum.de.
TEI (Text Encoding Initiative): https://tei-c.org.
Transkribus: https://readcoop.eu/transkribus.
Welscher Gast digital: https://doi.org/10.11588/edition.wgd.

Anschrift des Autors:

Dr. Jakub Šimek
Universitätsbibliothek Heidelberg
Plöck 107–109
69117 Heidelberg
E-Mail: simek@ub.uni-heidelberg.de

Angila Vetter

ediarum.MEDIAEVUM

Eine Arbeitsumgebung zur Edition mittelalterlicher (Prosa)Texte

Abstract. Das Interakademische Langzeitprojekt ›Der Österreichische Bibelübersetzer. Gottes Wort deutsch‹ widmet sich der kritischen editorischen Erschließung und Kommentierung der Schriften des sogenannten ›Österreichischen Bibelübersetzers‹. In den Arbeitstellen der Bayerischen Akademie der Wissenschaften an der Universität Augsburg und der Berlin-Brandenburgischen Akademie der Wissenschaften zu Berlin wird das Gesamtwerk ediert und für die geplante Hybridedition vorbereitet. Die Fülle des überlieferten Materials und die Komplexität der Texte lassen sich nur mit digitalen Methoden bewältigen. Die in den letzten Jahren an den beiden Arbeitstellen entwickelte Arbeitsumgebung ediarum.MEDIAVUM soll den Ansprüchen einer Edition der breit überlieferten Prosatexte ebenso gerecht werden wie den beiden geplanten Ausgabeformaten als Print- und Webedition.

Der folgende Beitrag möchte einen Einblick in die Entwicklung und Nutzung der Arbeitsumgebung *ediarum.MEDIAEVUM* innerhalb des Langzeitprojekts ›Der Österreichische Bibelübersetzer. Gottes Wort deutsch‹ geben.[1]

Der Österreichische Bibelübersetzer. Gottes Wort deutsch

Seit 2016 arbeiten Arbeitsgruppen der Bayerischen Akademie der Wissenschaften (BAdW) an der Universität Augsburg und der Berlin-Brandenburgischen Akademie der Wissenschaften zu Berlin (BBAW) an einer

hybriden Edition des Œuvres des sogenannten »Österreichischen Bibelübersetzers« (vgl. Löser/Stöllinger-Löser 1989, S. 251), eines bislang namenlosen Autors, der in der ersten Hälfte des 14. Jahrhunderts große Teile der lateinischen Vulgata in die deutsche Sprache übertrug und kommentierte. Die größeren Werkzusammenhänge wurden erst in den letzten Jahrzehnten offenbar. Bekannt sind umfangreiche Übersetzungen und Kommentierungen verschiedener Bücher des Alten Testaments, das sogenannte ›Alttestamentliche Werk‹ (sechs Textzeugen), eine Evangelienharmonie, das sogenannte ›Evangelienwerk‹ (30 Textzeugen) sowie ein ›Psalmenkommentar‹ (72 Textzeugen). Hinzu treten weitere kleine Schriften und Traktate. Sein Gesamtwerk nimmt hinsichtlich des Umfangs der übersetzten biblischen Bücher und der Übersetzungsleistung eine Sonderstellung in der langen Tradition deutschsprachiger Bibelübersetzungen ein. Er ist zweifellos einer der bedeutendsten deutschen Bibelübersetzer der Zeit vor Luther. Doch seine Arbeitsweise unterscheidet sich deutlich von der Luthers: Wie fast alle mittelalterlichen Übersetzer in Europa orientiert er sich an der lateinischen Bibel, der Vulgata. Außerdem interpretiert er die übersetzten Wörter, er erklärt, wie die Passagen zu verstehen sind und zerlegt sie dazu wörtlich in ihre Einzelteile. Seine Kommentare und Interpretationen richten sich dezidiert an ein bestimmtes Publikum: die Laien. Sie sollen die Bibel auch außerhalb klerikaler Deutungshoheit verstehen können (dazu ausführlich vgl. Löser/Stöllinger-Löser 1989).

Das Werk und seine Edition – Problemstellung

Der Bibelübersetzer greift bei der Kommentierung seiner Werke auf unterschiedliche Quellen zurück und kombiniert die Versatzstücke gemäß seinem Anliegen neu (ausführlich zum Vorgehen des Österreichischen Bibelübersetzers vgl. neuerdings Zinsmeister 2021). Das derzeit in einem gemeinsamen Vorhaben der Bayerischen Akademie der Wissenschaften mit einer Arbeitsstelle an der Universität Augsburg und der Berlin-Branden-

burgischen Akademie der Wissenschaften edierte ›Evangelienwerk‹ ist von kommentierenden oder predigthaften Glossen durchsetzt, häufig werden lateinische Legenden und Apokrypha eingestreut. Der Bibelübersetzer möchte hier alles, was es über Jesu Leben und Nachleben zu berichten gibt, vereinigen (vgl. dazu Gärtner 1983, Sp. 1251; sowie Schubert 2019, S. 211–226). So finden sich neben dem rund zwei Drittel des Textes einnehmenden Leben Jesu auch der Beginn der Apostelgeschichte sowie die Pilatus-Veronika-Legende. Das Ende bildet die Zerstörung Jerusalems. Das ›Evangelienwerk‹ ist in derzeit 30 bekannten Textzeugen, davon fünf Vollhandschriften, zwölf Fragmente und 13 Exzerpthandschriften, und zwei Fassungen auf uns gekommen: die angenommene Erstfassung *Gö und die Bearbeitung *SK.[2]

Die Edition der beiden Fassungen wird nach dem Leithandschriftenprinzip erstellt. Es wird also nicht versucht, einen Urtext zu rekonstruieren, sondern für jede Fassung wird jeweils eine Handschrift ausgewählt, deren Text der Fassungsedition zugrunde gelegt und in den nur bei sinnentstellenden Fehlern eingegriffen wird. Die inhaltlichen Fassungsunterschiede definieren sich zum einen durch die unterschiedliche Gestaltung der Kapitelübergänge und zum anderen durch Änderungen in der Anordnung des Textes sowohl bei der Reihenfolge der Kapitel als auch in einzelnen Kapiteln bei der Reihenfolge der Absätze (vgl. Kornrumpf 1991, hier S. 123–124; ausführlich dazu vgl. Zinsmeister 2021). Die Harmonisierung und Auslegung der Evangelien folgt in *Gö (eine Vollhandschrift, acht Fragmente, zehn Exzerpthandschriften) in der Regel den chronologischen Abläufen der Perikopen. In der Fassung *Gö gliedert sich der Text in rund 250 Kapitel.[3] Der umfassendste Teil des ›Evangelienwerks‹, das Leben Jesu, folgt keiner stringenten Chronologie:

> Vor allem im Teil über Jesu Geburt und Kindheit bis hin zur Versuchung Jesu sowie in der Passionsgeschichte sind immer wieder Kapitel eingeschoben, die alttestamentliche Prophezeiungen enthalten, die sich meist auf die im nachfolgenden Kapitel übersetzte Evangelienstelle beziehen. (Zinsmeister 2021, S. 474)

Die umfangreichere Bearbeitungsfassung *SK (vier Vollhandschriften, zwei Fragmente, drei Exzerpthandschriften), die aufgrund des Alters des ältesten Textzeugen nicht lange nach der Erstfassung entstanden sein kann, nimmt hingegen intensiven Eingriffe in Abfolge und Gewichtung vor. So tritt in *SK der seltene Fall auf, dass nicht vorrangig der Inhalt, sondern vor allem die Reihenfolge der Kapitel und deren Überleitungen abgeändert wurden, was die Textstruktur, die originär auf die Erstfassung zurückzuführen ist, in *SK ebenfalls grundlegend im Vergleich zu *Gö verändert (vgl. Zinsmeister 2021, S. 476). Einige Kapitel wurden zusammengezogen, andere wiederum umgestellt. Dabei führen manche Verweise innerhalb des Textes durch die Umstellung ins Leere, oder aber sie wurden entsprechend korrigiert (vgl. Zinsmeister 2021, S. 477f.). Die Kapitelzählung passte der Bearbeiter daraufhin ebenfalls an, wodurch zwei unterschiedliche Kapitelzählungen vorliegen, die kaum noch miteinander zu vergleichen sind. Beispielhaft sei das am Kapitel 35 der Bearbeitung *SK vorgeführt. Hier werden alle Jüngerberufungen zusammengezogen (vgl. *Gö, Kapitel 45, 55, 59) und mit der Bibelstelle über den Ernst der Nachfolge (Mt 8,19–22; Lc 9,57–62) zu einem Kapitel vereinigt (vgl. *Gö, Kapitel 60 und 94).[4] Die fünf Abschnitte bleiben aber getrennt; Bibeltext und Glosse wechseln sich ab (Abb.1):

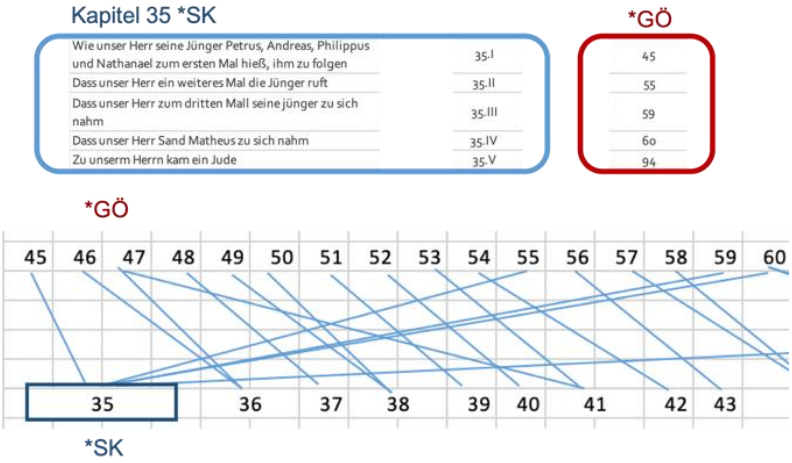


Abb. 1: Umstellungen im ›Evangelienwerk‹, Kapitel 35 der Bearbeitung *SK im Vergleich zur Erstfassung *Gö.[5]

Zwischen Print und Web – Hybridedition

Diese erheblichen Textumstellungen lassen eine Fassungssynopse im Druck impraktikabel erscheinen, sinnvoll darstellen lässt sie sich allein in der digitalen Edition. Es braucht also ein Datenmodell, das die verschiedenen Ebenen der Edition bzw. den ›einen Text‹ in seinen verschiedenen Erscheinungsformen (Handschriftenimage, Transkription, kritischer Text inklusive Metadaten sowie Fassungssynopse) über dynamische Verknüpfungen aufrufbar und transparent hält und die angestrebte Hybridedition ermöglicht. Dafür wird wiederum eine digitale Arbeitsumgebung benötigt, die es erlaubt, die Texte, d. h. die Transkriptionen und die Fassungseditionen, in XML/TEI einzugeben und zu bearbeiten, wobei sie gleichzeitig Hilfestellungen geben sollte für die korrekte Dateneingabe. Auch müssen in ihr die nötigen Register und Apparate angelegt und verwaltet werden können, so z. B. Namen- oder Quellenregister bzw. Lesarten-, Lese-

hilfen- oder Quellenapparate. Die Daten müssen zudem mit den für die Darstellung in der digitalen Edition nötigen internen und externen Verknüpfungen versehen werden können, sodass z. B. der Lesartenapparat mit den Transkriptionen verlinkt ist; die Kapitel-/Absatz-/Satzzählung der beiden Fassungseditionen miteinander, um die Fassungssynopse generieren zu können; der Lesehilfeapparat mit dem Wörterbuchnetz und das Namenregister mittels der GND (Gemeinsame Normdatei) mit weiteren Online-Ressourcen. Außerdem muss die Arbeitsumgebung ein kollaboratives Arbeiten unterstützen, das die gemeinsame Arbeit an den Dateien von den beiden Standorten aus zulässt.

Datenerfassung und digitale Infrastruktur – Workflow

In der BBAW ist in einer ganzen Reihe von Vorhaben die digitale Arbeitsumgebung *ediarum* im Einsatz. Entwickelt wurde sie von ›TELOTA – The Electronic Life of the Academy‹ (BBAW) ursprünglich für das Schleiermacher-Projekt, in dem allerdings mit neuzeitlichen Briefen, Tageskalendern und Vorlesungsmitschriften Material bearbeitet wird, das deutlich andere Ansprüche stellt als unseres, das aus umfangreichen, breit überlieferten Prosatexten des Spätmittelalters besteht. In Zusammenarbeit mit Nadine Arndt vom ›Otto von Passau-Projekt‹ wurde die digitale Arbeitsumgebung *ediarum.MEDIAEVUM* entwickelt, einer auf *ediarum* basierenden Arbeitsumgebung. Für Anpassung an die Projektbedürfnisse, etwa die notwendigen Strukturen zur Synoptisierung der Fassungen des ›Evangelienwerks‹, das Anlegen mehrerer Kommentarebenen und einer Datenbank sowie das Einfügen der CEI (Charters Encoding Initiative) konnten wir uns dabei auf die Zusammenarbeit mit den Kolleg*innen von Telota und dem *Deutschen Textarchiv* (DTA, ebenfalls BBAW) stützen, dessen Basisformat *ediarum* zugrunde gelegt ist (vgl. Vetter/Zinsmeister 2020, v. a. S. 127–130).

Aus *ediarum.MEDIAEVUM* heraus werden sowohl die Printeditionen – für das ›Evangelienwerk‹ die Editionen der beiden Fassungen – als auch die digitale Edition generiert, wobei unser Fokus klar auf der digitalen Edition liegt. In ihr sollen alle Textzustände und Forschungsdaten zugänglich gemacht werden, also die Digitalisate zu allen Textzeugen, die dazugehörigen Transkriptionen, Kollationen, die beiden Fassungseditionen und schließlich die Fassungssynopse.

Die erarbeiteten Konventionen für das Markup in XML/TEI (P5) sind auf die digitale Edition ausgerichtet. Für das Layout der Printausgabe sind die Richtlinien der Reihe *Deutsche Texte des Mittelalters* (DTM) Vorbild, in der die Editionen der Werke des Bibelübersetzers erscheinen werden.

Das Erfassen der Daten erfolgt in XML sowohl aus der ›codexbasierten‹ als auch aus der ›werkbasierten‹ Perspektive, um die dynamische Darstellung der kritischen Fassungstexte, ihrer Textzeugen in Form von Digitalisaten und Volltranskriptionen sowie der angestrebten Fassungssynopse umsetzen zu können (dies bereits bei Šimek 2014, hier § 3; Plattform: ›Welscher Gast digital‹). Das jeweilige Verhältnis von Textualität und Materialität wird dabei je unterschiedlich gewichtet, dennoch stehen die über diese Perspektiven erfassten Texte gleichberechtigt nebeneinander, ergänzen und erweitern sich (vgl. Sahle 2013, S. 244f.). Die werkbasierte Perspektive strukturiert die Daten primär nach hierarchischen Einheiten des abstrakt gedachten Werks: Büchern, Kapiteln, Absätzen, Sätzen, Wörtern etc. Die codexbasierte Perspektive legt hingegen jeweils eine physische Handschrift zugrunde. In den Transkriptionen werden aus der codexbasierten Perspektive so neben der Varianz und den Lesarten der einzelnen Textzeugen auch die kodikologischen und paläographischen Parameter sowie etwaige Illustrationen erfasst. Die Auszeichnung der den Text strukturierenden Mittel wie Initialen, Majuskeln oder Paragraphen- und Verweiszeichen legt hier wichtige Grundlagen für die Herstellung des kritischen Texts in werkbasierter Perspektive. Teilweise können Aspekte der Werkperspektive schon im Markup in den Transkriptionen der Fas-

sungsleithandschriften berücksichtigt werden, wie etwa Namen und Ortsangaben, aus welchen etwa im Fall der Printedition die Registereinträge generiert werden können.

In der Berliner Arbeitsstelle kam bis zur Fertigstellung der digitalen Arbeitsumgebung *ediarum.MEDIAEVUM* das *Tübinger System von Textverarbeitungsprogrammen* (TUSTEP) für die Erstellung der Transkriptionstexte wie des Editionstexts zum Einsatz.[6] In der Arbeitsgruppe Augsburg werden seit Januar 2017 die Transkriptionen der Textzeugen der Fassung *SK mit Hilfe der gleichnamigen Softwarekomponente der Plattform *Transkribus* von den studentischen und wissenschaftlichen Hilfskräften erstellt. Die maschinell ausgelesene Transkription wird von einer Hilfskraft kontrolliert und ggf. verbessert. Zugleich wird das Markup in werkbasierter Perspektive vorgenommen. Mittels der Transformationsskripte (XSLT, [Extensible Stylesheet Language Transformations]) werden die in TUSTEP und *Transkribus* hergestellten Transkriptionen der Textzeugen in die Leitfaden-konforme TEI überführt und in der existDB gespeichert.

Der kritische Text entsteht mit dem Framework *ediarum.MEDIAEVUM* in werkbasierter Perspektive. Ihm werden insgesamt vier Apparate beigegeben: Apparat 1 verzeichnet die Lesarten der weiteren Textzeugen; Apparat 2 gibt Lesehilfen, Apparat 3 verzeichnet die Quellennachweise und Apparat 4 gibt die Selbstzitate und Querverweise des Bibelübersetzers an.

TUSTEP bietet zudem weiterhin die einzig adäquate Umgebung für das Herstellen der benötigten Kollationsdateien sowie für die Printausgabe der Korrekturfahnen und die spätere Drucklegung des kritischen Textes. Damit ergibt sich folgender Workflow (Abb. 2):

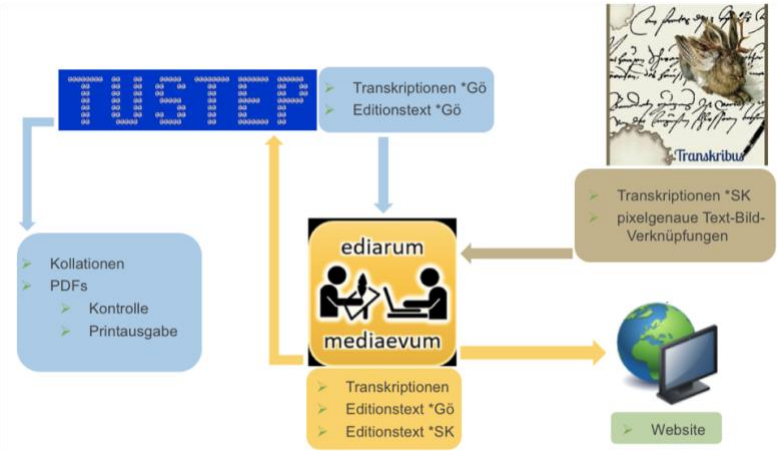


Abb. 2: Workflow im Vorhaben ›Österreichischer Bibelübersetzer‹, Edition ›Evangelienwerk‹.

Zur Aufbereitung der XML-Dateien der Transkription der Fassung *SK für die Kollation in TUSTEP wurde ein Makro geschrieben, das die einfache Umwandlung der Annotationen und Sonderzeichen in *plain text* durchführt. Die so entstandenen Word-Dateien können dann nachfolgend in TUSTEP eingespeist werden, wo die Kollationen erstellt werden.

Der in *ediarum.MEDIAEVUM* hergestellte kritische Text basiert für *SK auf der transformierten Transkriptionsdatei der Leithandschrift, der ›tT‹. Als ›Aussagen über den handschriftlichen Text‹ sind die Annotationen der den handschriftlichen Text strukturierenden Mittel wie Initialen, Majuskeln oder Paragraphen- und Verweiszeichen zwar wichtige Anhaltspunkte für die Herstellung des kritischen Textes, werden aber in Codierung hier nicht benötigt.

Über ein XSLT-Script werden diese Angaben in <comments> ausgelagert, Abbreviaturen aufgelöst und geringfügige Normalisierungen auf Zeichenbasis vorgenommen; zurück bleibt die ›transformierte Editionstranskription‹, die ›tE‹, die den Bearbeiter*innen das Arbeiten in werkbasierter Perspektive erleichtert, ohne dass dabei die technische Ver-

knüpfung zu Transkriptions- und Bilddatei verloren ginge. Die tE stellt lediglich eine Zwischenstufe in eben genannter Funktion dar; aus ihr wird der betreffende Kapiteltext herauskopiert und als eigenständige Kapiteldatei abgespeichert. Das Aufsplitten des ODD (One Document Does it all) in verschiedene Datenpakete sichert, indem darüber Verschachtelungen von verschiedenen Strukturen innerhalb eines Dokuments vermieden werden, die technische Funktionalität der Dateien ebenso wie es die leichtere Bearbeitung über verschiedene Perspektiven auf den Text ermöglicht (vgl. Abb. 3).

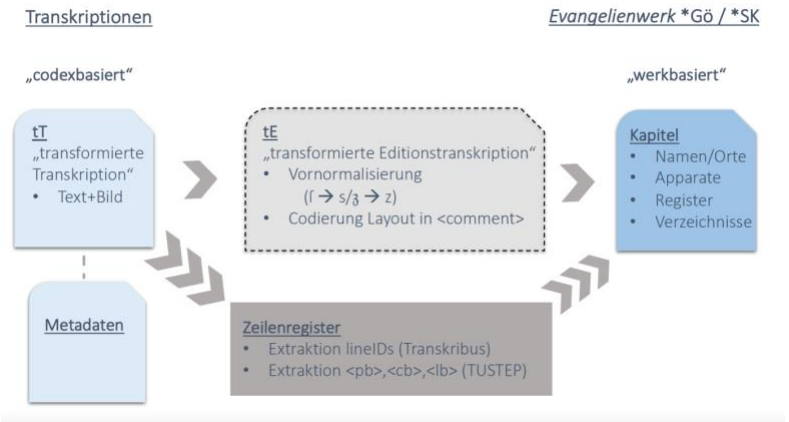


Abb. 3: Datenmanagement ›Evangelienwerk‹.

Analog wurde für *Gö die Transkription der Leithandschrift Gö in TUSTEP durch eine Austauschroutine in eine der tE entsprechende Datei verwandelt, in der die weitere Bearbeitung als Editionstext erfolgt. Die technischen Verknüpfungen mit dem Digitalisat und der Transkription werden hier nachträglich in *ediarum.MEDIAEVUM* eingefügt und erfolgen seiten- bzw. spaltenweise.

Die Arbeitsumgebung erleichtert das Erstellen des kritischen Texts dahingehend, dass die Bearbeiter*innen in *oXygen XML Author* nicht in ei-

ner Codeansicht, sondern in einer benutzerfreundlichen ›Autorenansicht‹ arbeiten, die über Cascading Stylesheets (CSS) gestaltet wird (Abb. 4):

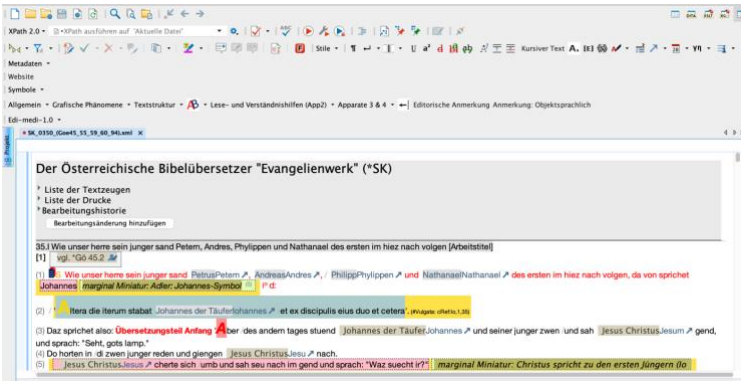


Abb. 4: Autoransicht in ediarum.MEDIAEVUM am Beispiel von Kapitel 35, ›Evangelienwerk‹, Fassung *SK.

Den Bearbeiter*innen stehen dabei mehrere Ansichten zur Auswahl, so dass per Mausklick die für den Arbeitsschritt geeignetste ausgewählt werden kann, bspw. die ›codexbasierte Perspektive‹, die den Text in den materiellen Gegebenheiten des Textzeugen zeigt, der der Datei jeweils zugrunde liegt (also die Seiten-, Spalten- und Zeilenumbrüche) und so auch innerhalb des bereits edierten Prosatextes eine rasche Orientierung in der Transkription – etwa zur Kontrolle – ermöglicht, wobei die Angabe der jeweilig genutzten Leithandschrift sowie deren Spalten- und Foliowechsel auch in der Edition konsequent mitgeführt werden und immer sichtbar sind.

Auszeichnungen können die Bearbeiter*innen über eine eigene Werkzeugleiste per Knopfdruck vornehmen. So lassen sich etwa Einträge in die vier Apparate per schrittweiser Abfrage mit der entsprechenden TEI-Auszeichnung versehen. Soll bspw. ein Vulgata-Nachweis in ›Apparat 3: Quellen‹ vorgenommen werden, so muss lediglich der betreffende Abschnitt markiert und dann über die Funktion ›Apparat 3: Vulgata-Nachweis‹ der Abfrage gefolgt werden: 1. Art des Nachweis (Zitation –

Paraphrase – Vergleich); 2. Titel des Buchs (als Kürzel, bspw. Mt); 3. Kapitel, 4. Verse(e). Die Angabe weiterer Bibelstellen ist möglich. Ebenso das ›Fortsetzen‹ des vorgenommenen Eintrags über das im Hintergrund erstellte Bibelstellenregister, wenn bspw. eine Verschachtelung in XML vermieden werden muss.

Unter ›Stile‹ lässt sich mit ›Leseansicht‹ eine überarbeitete Autoransicht anzeigen, die sich noch stärker an den traditionellen Lesegewohnheiten orientiert. Diese Ansicht ist insbesondere für die Korrektur oder das Lesen des Textes hilfreich.[7] Die Einträge bzw. Apparate lassen sich durch Anklicken der betreffenden Stelle im Text nach dem Einfügen ausklappen und – ohne in die Textansicht wechseln zu müssen – bearbeiten, korrigieren und auch löschen (Abb. 5):

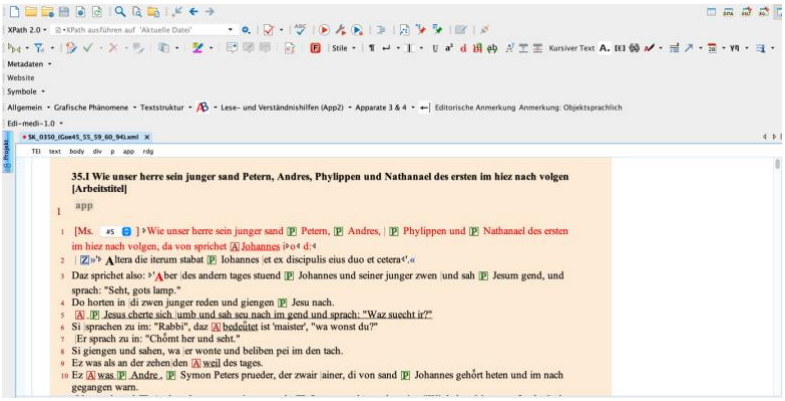


Abb. 5 Angepasste Autoransicht ›Lese- und Korrekturansicht‹ in ediarum.MEDIAEVUM, am Beispiel von Kapitel 35, ›Evangelienwerk‹, Fassung *SK.

Die XML-Datenbank *eXist-db* dient in der digitalen Arbeitsumgebung als zentrales Repositorium für die XML-Dokumente. Die Datenbank ist auf einem Server installiert und online global mit entsprechenden Zugangs-

daten zugänglich. Dadurch können alle Projektmitarbeiter*innen auf ein und denselben Datenbestand zugreifen und zusammenarbeiten.

bibeluebersetzer.digital – Ein Ausblick

Diese Form der Codierung verknüpft so nicht allein die verschiedenen Bestandteile der Edition miteinander, sondern öffnet die Edition nach ihrer Veröffentlichung auch für das Semantic Web. Die digitale Edition verfügt damit über einen beträchtlichen Mehrwert. Die Darstellung der digitalen Edition des ›Evangelienwerks‹ im Web befindet sich derzeit im Entstehen. Die grundsätzliche Funktionalität der Daten hinsichtlich einer dynamischen Darstellung des Editionstextes der beiden Fassungen, der kritischen Apparate, Register, Literaturangaben, Transkriptionen und Digitalisate lässt sich aber jetzt schon realisieren und hier zumindest in Abbildung zeigen (Abb. 6):

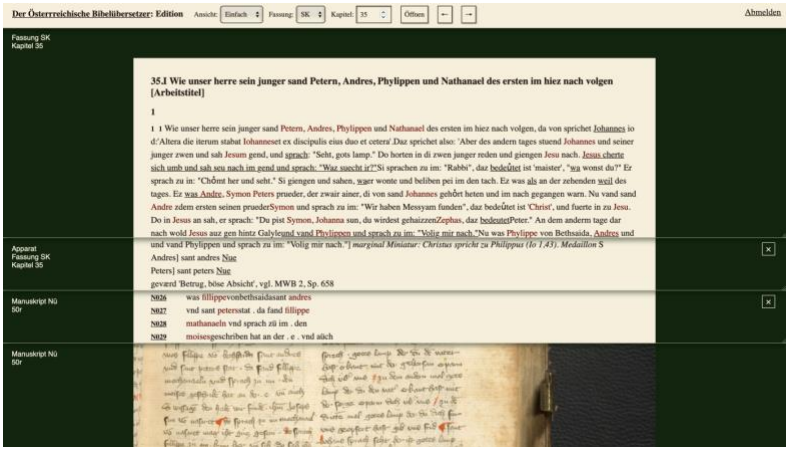


Abb. 6: Beta-Ansicht der digitalen Edition, hier am Bsp. von Kapitel 35, ›Evangelienwerk‹, Fassung *SK.

Die synoptische Darstellung der beiden Fassungen ist bereits möglich. Eine optische Überarbeitung wird vor Veröffentlichung noch erfolgen (Abb. 7):

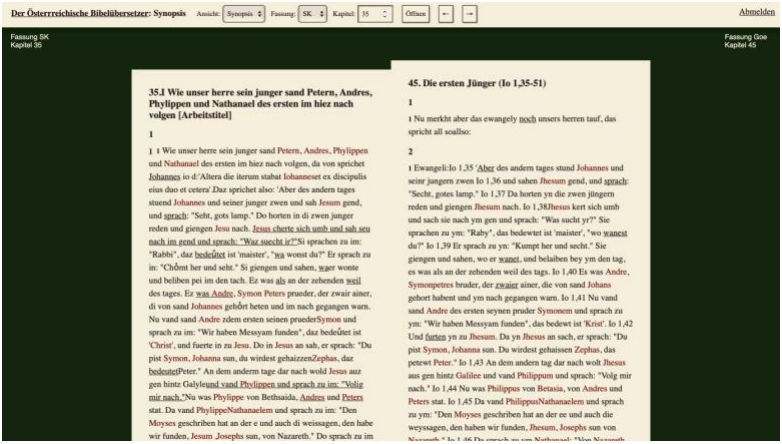


Abb. 7: Beta-Ansicht der Fassungssynopse des ›Evangelienwerks‹ am Beispiel von Kapitel 35, Fassung *SK und Kapitel 45, Fassung *Gö.

Auch in dieser Ansicht lassen sich sowohl die Apparate als auch die Transkriptionen und Digitalisate hinzuschalten.

ediarum.MEDIAEVUM – Einer für alle?

Das verfolgte Konzept bei der Datenerfassung ist reproduktionsorientiert. Es sollen möglichst zugleich menschen- und maschinenlesbare Daten erhoben werden, um sowohl den Austausch und die Zusammenarbeit mit anderen (mediävistischen) Editionsprojekten als auch die Interoperabilität mit verschiedenen Applikationen zu gewährleisten, wobei stets versucht wird, soweit wie möglich ›standardnah‹ zu arbeiten. Dies soll vor allem der Austausch mit unseren Partnern und Kolleg*innen anderer (mediävistischer) Editionsvorhaben sichern. Zu bedenken bleibt, dass der Mehrwert

der in dieser Form erhobenen Daten nur durch einen deutlich höheren Aufwand und höhere Kosten als bei einer reinen Printedition zu erreichen ist.

Die Anpassung und Wartung der eingesetzten Softwares, die notwendige Einarbeitung der Mitarbeiter*innen sowie immer wieder auftretende technische Probleme im Bereich Performance der Softwares oder der Server sind Faktoren, die dauerhaft zu berücksichtigen sind.

Ob *ediarum.MEDIAVUM* sich tatsächlich als Arbeitsumgebung für andere Vorhaben und Texte eignet, wird die Zukunft zeigen. Das Datenmodell und die Texterschließung im Vorhaben ›Der Österreichische Bibelübersetzer‹ sind komplex und vielschichtig, und die Funktionen in *ediarum.MEDIAEVUM* mögen daher für manches Vorhaben in ihrer Fülle eher überfordern. Grundsätzlich ist es generell einfacher, das Codierungsset zu verkleinern, als es zu erweitern. Die Anpassung an die jeweils spezifischen Bedürfnisse eines Textes und seiner Edition bleibt notwendig, doch die Funktionsweise des Varianten- oder auch des Worterläuterungen-Apparats (Apparat 1 und 2) sind sicher für die meisten Vorhaben nutzbar. Auch die Quellen- und Querverweisapparate (Apparat 3 und 4) lassen sich modifizieren und für individuelle Bedürfnisse anpassen, ohne dass eigene Referenz- und Registerstrukturen aufgebaut werden müssen. Als eine Art ›mediävistische Basis-Erweiterung‹ zu *ediarum.basis* ist *ediarum.MEDIAEVUM* durchaus denkbar.[8]

Anmerkungen

1 Ausführliche Informationen zum Projekt ›Der Österreichische Bibelübersetzer. Gottes Wort deutsch‹: Projektauftritt Berlin-Brandenburgische Akademie der Wissenschaften; Projektauftritt Bayerische Akademie der Wissenschaften.

2 Die Erstfassung *Gö ist bezeichnet nach ihrer Leithandschrift Göttweig, Stiftsbibliothek, Cod. 222 (rot) / 198 (schwarz) (früher C 2) aus dem sechsten Jahrzehnt des 15. Jahrhunderts. Für die Bearbeitung ist die Leithandschrift die prächtige, mit über 429 erhaltenen Federzeichnungen illustrierte Pergamenthandschrift S (Schaffhausen, Stadtbibliothek, Cod. Gen. 8) von um 1340, die da-

mit einen sehr autorzeitnahen Sprachstand bezeugt. Wo sie wegen Blattverlust ausfällt, springen die Handschriften K2 (Klosterneuburg, Stiftsbibliothek, Cod. 51) von 1415 oder K1 (Klosterneuburg, Stiftsbibliothek, Cod. 4) von etwa 1410 ein. Nach diesen Handschriften wird die Bearbeitung als Fassung *SK bezeichnet. Siehe zur Überlieferung ausführlich: Zinsmeister 2021, S. 470–473; sowie Vetter/Zinsmeister 2020, S. 125–140, hier S. 125f.

3 Aufgrund von Textverlust beinhaltet die Handschrift Gö nur 249 Kapitel, ein Vergleich mit den Handschriften aus *SK macht es wahrscheinlich, dass noch ein bis zwei Kapitel folgten.

4 Beschreibung folgt Zinsmeister 2021, S. 477, die beiden in Kapitel 60 der Erstfassung zusätzlich enthaltenen Bibelstellen füllen in *SK jeweils ein eigenes Kapitel (47 bzw. 48).

5 Für die Erlaubnis zur Nutzung der Darstellung danke ich herzlich meiner Kollegin Edith Kapeller.

6 TUSTEP hat sich seit langem für die Aufbereitung von Textdaten für die Printedition in der Reihe DTM bewährt. Für eine Reihe wichtiger Arbeitsschritte (z. B. u/v-Ausgleich, Kollationen) liegen in der Arbeitsstelle erprobte Routinen vor.

7 Die Erfahrungswerte zeigen, dass sich zur Arbeit am Text dennoch weiterhin die Standardansicht empfiehlt, die die über farbige Hervorhebung der von den Bearbeiter*innen vorgenommenen Einträge – und damit die dahinterliegende Codierungsstruktur – transparent hält.

8 Kontakt bei Interesse an *ediarum.MEDIAEVUM*: bibeluebersetzer@philhist.uni-augsburg.de.

Literaturverzeichnis

Handschriften

Gö Göttweig, Stiftsbibliothek, Cod. 222 (rot) / 198 (schwarz) (früher C 2)
K1 Klosterneuburg, Stiftsbibliothek, Cod. 4
K2 Klosterneuburg, Stiftsbibliothek, Cod. 51
S Schaffhausen, Stadtbibliothek, Cod. Gen. 8

Sekundärliteratur

Gärtner, Kurt: ›Klosterneuburger Evangelienwerk‹, in: ²VL, Bd. 4 (1983), Sp. 1248–1258.

Schubert, Martin: *Ander heilige geschrifft*. Die Haltung zu Apokryphen im ›Evangelienwerk‹ des Österreichischen Bibelübersetzers, in: Weigand, Rudolf Kilian/Schiewer, Regina Dorothea/Haustein, Jens/Schubert, Martin (Hrsg.): Traditionelles und Innovatives in der geistlichen Literatur des Mittelalters, Stuttgart 2019 (Meister-Eckhart-Jahrbuch, Beiheft 7), S. 211–226.

Kornrumpf, Gisela: Das ›Klosterneuburger Evangelienwerk‹ des österreichischen Anonymus. Datierung, neue Überlieferung, Originalfassung, in: Deutsche Bibelübersetzungen des Mittelalters. Beiträge eines Kolloquiums im Deutschen Bibel-Archiv, unter Mitarbeit von Nikolaus Henkel hrsg. von Heimo Reinitzer, Bern [u. a.] 1991 (Vestigia Bibliae 9/10 [1987/1988]), S. 115–131.

Löser, Freimut/Stöllinger-Löser, Christine: Verteidigung der Laienbibel. Zwei programmatische Vorreden des österreichischen Bibelübersetzers der ersten Hälfte des 14. Jahrhunderts, in: Kunze, Konrad/Mayer, Johannes G./Schnell, Bernhard (Hrsg.): Überlieferungsgeschichtliche Editionen und Studien zur deutschen Literatur des Mittelalters, Tübingen 1989 (Texte und Textgeschichte 31), S. 245–313.

Sahle, Patrick: Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels, Teil 3: Textbegriffe und Recodierung, Norderstedt 2013 (Schriften des Instituts für Dokumentologie und Editorik 9).

Šimek, Jakub: Welscher Gast Digital. TEI-Handbuch. Version 0.6. 15.1.2014 (online).

Vetter, Angila/Zinsmeister, Elke: Die Bibel für alle. Der Österreichische Bibelübersetzer auf dem Weg ins Web, in: Fischer, Martin (Hrsg.): Digitale Methoden und Objekte in Forschung und Vermittlung der mediävistischen Disziplinen, Bamberg 2020 (Bamberger interdisziplinäre Mittelalterstudien 15), S. 125–140.

Zinsmeister, Elke: Sinnbildende Umstrukturierung? Fassungsunterschiede im ›Evangelienwerk‹ des Österreichischen Bibelübersetzers, in: Leppin, Volker (Hrsg.): Schaffen und Nachahmen. Kreative Prozesse im Mittelalter, Berlin 2021, S: 469–484 (Das Mittelalter. Perspektiven mediävistischer Forschung, Beiheft 16).

Online-Ressourcen

BAdW (Bayerische Akademie der Wissenschaften): https://badw.de/die-akademie.html.

BBAW (Berlin-Brandenburgische Akademie der Wissenschaften): https://www.bbaw.de/.

CSS (Cascading Stylesheets): https://www.w3.org/Style/CSS/.

›Der Österreichische Bibelübersetzer. Gottes Wort deutsch‹, Projektauftritt Bayerische Akademie der Wissenschaften, https://bibeluebersetzer.badw.de/das-projekt.html.
›Der Österreichische Bibelübersetzer. Gottes Wort deutsch‹, Projektauftritt Berlin-Brandenburgische Akademie der Wissenschaften, https://www.bbaw.de/forschung/der-oesterreichische-bibeluebersetzer-gottes-wort-deutsch.
DTA (Deutsches Textarchiv) http://www.deutschestextarchiv.de.
DTM (Deutsche Texte des Mittelalters): https://dtm.bbaw.de.
ediarum (TELOTA, Berlin-Brandenburgische Akademie der Wissenschaften): http://www.bbaw.de/telota/software/ediarum.
eXist-db: http://exist-db.org/.
›Otto von Passau, Die 24 Alten‹: https://www.bbaw.de/forschung/otto-von-passau.
oXygen XML Author: https://www.oxygenxml.com/xml_author.html.
TELOTA (The Electronic Life Of The Academy): https://www.bbaw.de/bbaw-digital/telota.
Transkribus: https://readcoop.eu/transkribus/.
TUSTEP (Tuebingen System of Text Processing tools): http://www.tustep.uni-tuebingen.de/.
›Welscher Gast digital‹: https://digi.ub.uni-heidelberg.de/wgd/.
XSLT (Extensible Stylesheet Language Transformations): https://www.w3.org/TR/xslt-30/.

Anschrift der Autorin:

Dr. Angila Vetter
Universität Augsburg
Forschungsprojekt ›Der Österreichische Bibelübersetzer‹
Werner-von-Siemens-Str. 6
86159 Augsburg
E-Mail: angila.vetter@philhist.uni-augsburg.de

Sonja Glauch

Welche Lebenserwartung haben digitale Editionen?

Abstract. Der Beitrag unterscheidet zunächst vier Hinsichten, in denen digitale Editionen altern und gegebenenfalls sterben können: 1. Datenträgertechnologien und Laufumgebungen, 2. Kodierungen, Verlinkungen, Metadatenformate, 3. softwaretechnische Aspekte der Präsentationsschicht, 4. Aspekte der Ästhetik und des Bedienungskomforts. Er wendet sich dann einer aktuellen Lösungsperspektive für das Langfristbewahrungsproblem zu, nämlich den im Aufbau befindlichen Forschungsdateninfrastrukturen, und erörtert die Frage, ob digitale Editionen Forschungsdaten sind und inwiefern ihre Dauerhaftigkeit im Rahmen von Initiativen wie Text+ gewährleistet werden könnte. Abschließend nennt der Beitrag vier Optionen, die konkret in den letzten Jahren für die Bewahrung von digitalen Editionen ins Spiel gebracht worden sind. Da es sich bei diesen Ansätzen zum größten Teil um Vorüberlegungen oder (teils gescheiterte) Prototypen handelt, ist zum jetzigen Zeitpunkt nicht entscheidbar, welcher dieser Wege konkret für jetzt entstehende komplexe, dynamische Editionen dauerhaft(er)en Bestand gewährleisten könnte.

Wie lang bleiben die digitalen Editionen funktionsfähig, an denen manche von uns arbeiten? Das ist eine Frage, die wir uns morgens und abends stellen. Nachhaltigkeit und Langzeitverfügbarkeit sind als (noch) ungelöste Probleme allseits erkannt und im Bewusstsein, nicht zuletzt, weil sie von den Drittmittelgebern eingefordert werden.

Ich möchte vorneweg betonen, dass ich an diese Frage aus der Perspektive von jemandem herangehe, die an einer konkreten einzelnen Edition arbeitet, und nicht aus der Perspektive von Informationswissenschaftlern und Datenmanagement-Spezialisten. Ich stelle sie also aus der Perspektive

›Was braucht das Fach, was brauchen die Digitalphilologen?‹, also von der Nachfrageseite, und bin nicht die, die das Angebot überblickt, das dieser Nachfrage gegenübersteht, weil die Angebotslandschaft für den einzelnen Philologen überhaupt nicht überschaubar ist.

Zunächst eine kleine Reflexion darüber, in welchen Hinsichten digitale Editionen altern und das Zeitliche segnen können:

1. Das Handgreiflichste sind zunächst einmal Datenträgertechnologien und Laufumgebungen, die außer Mode kommen. Das führt zum Exitus, falls es nicht durch eine Migration des Projekts aufgefangen werden kann. Ein Beispiel bieten Kiernan/Iacob (o. J.) mit einem Bericht zu den Zwangsmigrationen des ›Electronic Beowulf‹ [1]2000, [2]2003, [3]2011 (CD-ROM/DVD-ROM). Wir haben in der kurzen Geschichte der digitalen Edition schon mehrere Medienbrüche erlebt, so besonders von Editionen auf CD bzw. als lauffähige Programme für ein bestimmtes Betriebssystem hin zu Online-Editionen. Dieser Faktor dürfte mit der allgemeinen Umstellung auf browserbasierte Angebote für die Zukunft keine Rolle mehr spielen. Hier findet die Obsoleszenz stattdessen in einem anderen Bereich statt, der auf einer CD ausgeschlossen ist, nämlich:

2. Kodierungen, Verlinkungen, Metadatenformate: Das Auftauchen von falschen Sonderzeichen bzw. Ersatzzeichen für falsch kodierte Sonderzeichen dürfte ärgerlich, aber nicht letal sein; tödlicher ist dann schon ein 404-er, aus welchem Grund auch immer. Häufig müssen im Laufe der Jahre Daten und Ressourcen auf neue Server umgezogen werden; dabei werden oft interne Verzeichnisstrukturen umgestellt, was oft zu solchen Fehlern führt. Dergleichen hat nicht unbedingt das Ableben einer Edition zur Folge, sondern – je nachdem, wie zentral die fehlenden Daten sind – manchmal auch bloß eine etwas eingeschränkte Benutzbarkeit.

3. Diese Quelle von Vitalitätsverlusten lässt sich kaum trennen vom nächsten Punkt, nämlich technischen Aspekten der Präsentationsschicht. Mit zunehmend komplexer werdenden Editionen, die meist keine statischen Inhalte mehr präsentieren, wird auch die eingesetzte Software viel-

fältiger. Wir haben oft auf der Serverseite eine Datenbank laufen mit ihrer Software sowie eine Skriptsprache wie PHP (Hypertext Preprocessor) oder Python zur Generierung der Seiten, dazu weitere Skriptsprachen auf der Clientseite wie JavaScript, XSLT (Extensible Stylesheet Language Transformations) und dazu Stylesheet-, also Gestaltungselemente. Jedes dieser Elemente ist eventuell wiederum eingebettet in oder angebunden an ein Framework oder größere Komplettsysteme wie Bootstrap, WordPress, Mediawiki, IIIF (International Image Interoperability Framework) usw. Dazu kommt der Einsatz fertiger Module für schwierige Aufgaben wie einen Bildviewer, eine Textsynopse usw. Jeder dieser softwaretechnischen Bausteine stellt eine Abhängigkeit dar, jedes dieser Systeme kann theoretisch außer Betrieb gehen oder von seinen Entwicklern aufgegeben werden. Ein auf diese Weise erzwungener Umbau der Gesamtarchitektur kann natürlich brutal sein. Auch wenn man mit so etwas nicht unbedingt rechnen muss – auf die Dauer kann jeder der genannten Bausteine in dem Sinne veralten, dass er nicht mehr als *state of the art* gilt und eigentlich durch etwas anderes ersetzt werden müsste. Ein solcher Ersatz bedeutet meist einen kompletten Relaunch der Präsentation, aber auch unterhalb dieser Ebene und unsichtbar für die Benutzer müssen die ständigen Versionsanpassungen all dieser Bausteine und Pakete nachvollzogen werden, es sei denn, man hätte es geschafft, sein Gesamtpaket so zu schnüren, dass alle Programmier- und Anpassungsarbeit bei dem außer Haus gewarteten Framework liegen und man selbst für alle spezifischen Inhalte der Edition ein vollständig standardisiertes Datenformat gefunden hat, das in 50 Jahren immer noch gültig und von dem Framework oder einem Nachfolgerframework interpretierbar sein wird. Das wäre der ›heilige Gral‹, damit wäre auch die Frage der Langzeitlauffähigkeit gelöst. Anders gesagt: Das gibt es meines Wissens nicht oder nur in sehr günstigen institutionellen Konstellationen. Unterzieht man sein System den nötigen Aktualisierungen nicht, dann droht eben doch der Exitus, je nachdem wie zentral der Baustein und wie unglücklich im Nachhinein die ursprüngliche Ent-

scheidung für und wider eine Technik war. Sicher haben sich einige der verschwundenen digitalen Editionen auch aus solchen Gründen aus dem Netz verabschiedet.

Es hat schon einen Grund, warum professionelle Webanwendungen, etwa der Banken oder der Presse- und Medienhäuser, kaum je länger als drei Jahre existieren, bevor wieder ein kompletter Relaunch, eine komplette Neuprogrammierung nötig wird. Das Tempo der Entwicklungen in der Webprogrammierung ist enorm. Bei den Neugestaltungen solcher Profi-Webpräsentationen spielt freilich ein letzter Faktor der Obsoleszenz stark mit herein:

4. Aspekte der Ästhetik der Präsentationsschicht, des Bedienungskomforts, der Nutzeransprüche, denen nicht mehr entsprochen wird. Layout, Typographie, Bedienung von Suchfeldern und Eingabemasken, Look-and-Feel, das sind weiche Kriterien. Den aktuellen *state of the art* nicht mehr liefern zu können, führt nicht dazu, dass ein elektronisches Angebot nicht mehr funktioniert. Ob es sich in der Benutzungsfrequenz niederschlägt, wäre interessant zu wissen, aber vermutlich besteht in Wissenschaftskreisen eine recht hohe Toleranz gegenüber ältlichem Webdesign, Unübersichtlichkeit und umständlicher Funktionalität, sofern die Inhalte nützlich sind. Für manche Zwecke haben statische Seiten der 90er vielleicht sogar gewisse Vorteile. Jedenfalls hat sich die Optik von Textdarbietung im Web selbst innerhalb der letzten zehn Jahre noch einmal deutlich geändert; das Tempo, in dem Editionen veralten, nimmt in dieser Dimension also eher zu als ab. Ich möchte betonen, dass es dabei nicht allein um Layout und Optik geht, sondern auch um Werkzeuge wie Suchfunktionen, Filter und Visualisierungen. In diesem Bereich überschneiden sich Aspekte reiner Bequemlichkeit mit Aspekten der Funktion, die sehr wohl zum Kern einer Edition beitragen und ihre Leistungsfähigkeit bestimmen können. In anderen Beiträgen in diesem Band werden schon so viele mögliche künftige Anbindungen von Editionen an Tools angesprochen, dass es evident sein dürfte,

dass auch auf dieser Ebene fertige Editionspräsentationen veralten, wenn sie solchen künftigen Standards nicht mehr entsprechen.

Der Rückblick zeigt so viele damals nicht antizipierbare Entwicklungen, dass wir uns davor hüten sollten, Prognosen über die nächsten zehn oder zwanzig Jahre abzugeben. Natürlich waren die Pioniere immer besonders betroffen, weil sie nicht wissen konnten, was sich in der Folge zum Standard entwickeln würde; was lässt uns glauben, dass dies in den nächsten zwanzig Jahren anders sein wird?

Das Problem als solches ist natürlich erkannt. So adressieren die neueren Bemühungen um eine dauerhafte Bewahrung von Forschungsdaten und um Repositorien für Forschungsdaten genau diese Frage. Das neue Konsortium Text+ im Rahmen der großen Initiative zum Aufbau einer nationalen Forschungsdateninfrastruktur (NFDI) beschreibt seine Zuständigkeit damit, »text- und sprachbasierte Forschungsdaten langfristig [zu] erhalten und ihre breite Nutzung in der Wissenschaft [zu] ermöglichen« (www.text-plus.org). Text+ läuft seit Herbst 2021 für fünf Jahre und will die Wissenschaft nicht nur beraten und bei der Weiterentwicklung von Standards und Normdaten unterstützen, sondern offenbar auch Dienste und Routinen installieren, mit denen Daten gehostet und archiviert werden können. Editionen sind eine von drei Domänen, also Zuständigkeitsbereichen, von Text+.

Nun erhebt sich freilich die Frage, ob mit ›Forschungsdaten‹ im Kontext von digitalen Editionen eigentlich die digitalen Editionen selbst gemeint sind. Diese Sichtweise fände ich nicht selbstverständlich, weil wir Editionen als Publikationen und damit als Resultate unserer Forschungsarbeit (nicht als anfallende Begleit- und Rohdaten) verstehen und weil wir den Begriff ›Daten‹ dafür etwas befremdlich finden, während viele Definitionen von Forschungsdaten das Instrumentelle hervorheben: Daten, die im Forschungsprozess anfallen und für die Forschung (nach-)genutzt werden. Andererseits sind Definitionen des Begriffs ›Forschungsdaten‹ aber meist so allgemein, dass sie Editionen sehr wohl erfassen sollten – diese sind

digitale Repräsentationen von Texten, deren zentraler Zweck darin besteht, für weitere Forschung zur Verfügung zu stehen, somit sind sie Daten aus der und für die Forschung, also Forschungsdaten. Die Frage mag gesucht klingen, aber sie entscheidet schließlich darüber, in welcher Hinsicht sich etwa eine nationale Forschungsdateninfrastruktur zuständig fühlt für eine digitale Edition.

Wenn man hier etwas weiter die momentane Lage sondiert, dann zeigt sich, dass die Frage doch nicht ganz müßig ist. Ein zentraler Baustein einer Infrastruktur, die Nachhaltigkeit herstellen und garantieren soll, sind Repositorien: Orte, an denen Daten abgelegt, aufbewahrt und zugänglich gehalten werden. Jedoch, sieht man sich einmal in einem großen Verzeichnis von solchen Repositorien um, re3data.org, einem am KIT (Karlsruher Institut für Technologie) angesiedelten »global registry of research data repositories« (www.re3data.org/about), dann stellt man fest, dass dort Projekte als Repositorien eingetragen sind, die man ebenso als Editionsprojekte begreifen kann; so etwa eine Handschriftenpräsentation des Codex Sinaiticus mit Digitalisaten und Transkription: www.re3data.org/repository/r3d100010560.

Das lässt erahnen, dass Editionsprojekte sich selbst als Archiv oder Repositorium von Originaldokumenten verstehen können. Ich kann nicht ermessen, ob das nur fragwürdige Grenzfälle einer Eintragung in diese Repositorienliste sind, es scheint mir aber sichtbar machen zu können, dass es zwischen digitalen Daten und den Strukturen, in denen die Daten aufbewahrt werden, keine so klaren Unterscheidungen gibt wie in der analogen Welt zwischen dem Buch und der Bibliothek. Ein Buch kann sich zwar ›Archiv‹ betiteln, aber niemand wird wirklich die Dienstleistungen eines Archivs von ihm erwarten. Eine digitale Edition leistet dagegen immer in gewissem Grad sowohl das, was ein Buch als auch was eine Bibliothek leistet, und folglich besteht eine Grauzone in der Zuordnung dessen, was Editionsprojekte sind und tun. Wenn nun eine digitale Edition wie die des Codex Sinaiticus sich selbst als ein Forschungsdatenrepositorium begreift,

was genau wären (in technischer Hinsicht) dann die dort deponierten Forschungsdaten? Und welche Infrastruktur wäre zuständig für den Erhalt des Repositoriums, also der Edition als ganzer?

Das mag jetzt als sehr künstliche Zuspitzung erscheinen. Sicher spielen auch Verständigungsschwierigkeiten eine Rolle, weil die Datenmanagement-Leute ihr Tun mit Begriffen beschreiben, die für uns Mediävisten sehr ungewohnt sind: Da sind wir Stakeholder, da werden Daten und Software kuratiert etc. Es macht aber auf das Problem aufmerksam, das ich hinsichtlich der Dauerhaftigkeit von digitalen Editionen als das zentrale wahrnehme: Editionen bestehen aus Inhalten wie Texten und Bildern sowie einer Darbietungsoberfläche und Zugangsstruktur. Das ist in der Tat eine sehr ähnliche funktionale Zweiteilung wie bei einem Datenrepositorium, denn das muss genauso eine Zugangsstruktur und eine Benutzeroberfläche haben, damit die Daten, die in ihm deponiert sind, nutzbar sind. Der Unterschied zum Repositorium besteht darin, dass weite Teile der Darbietungsoberfläche und der Zugangsstrukturen integrale und wesenhafte Bestandteile der Edition bilden. Die Funktionsschicht ist mit den Daten aufs engste verzahnt, und ohne ihre Funktionsschicht sind die Daten weitgehend nutzlos. Folglich bietet die Idee, die Daten allein (also z. B. die XML/TEI-Dateien) in einem Repositorium abzulegen, nicht wirklich Abhilfe gegen das schnelle Altern von digitalen Editionen; da sage ich nichts Überraschendes.

Es gibt nun verschiedene Ansätze, dem Problem zu begegnen:

- generische Viewer für XML/TEI, womit die Wartung und Weiterentwicklung des Viewers in institutioneller Hand liegen könnte und dem einzelnen Projekt erspart bliebe. Das ist auch von Seiten der Infrastrukturanbieter erwogen worden, wurde aber nicht als erfolgversprechend betrachtet (vgl. Buddenbohm [u. a.] 2016): »[…] stellte sich aber heraus, dass mittelfristig wohl nur eine Anwendungskonservierung eine breitere Akzeptanz bei den Nutzerinnen und Nutzern erlangen kann, wenn es um die Langzeitverfügbarkeit einer digitalen Edition als Webanwendung geht.«

- ›Anwendungskonservierung‹: Wenn ich es richtig verstehe, dann handelt es sich dabei um die Emulation eines gesamten Systems, wie es zum Zeitpunkt der Übergabe an das Datenzentrum beschaffen ist, als virtueller Server. Das klingt genial, und ein solches Angebot scheint auch bereits im Portfolio des Göttinger Humanities Data Centre vorhanden. Es gibt auch schon zwei solche einbalsamierten Projekte auf der Göttinger Seite aufzurufen (humanities-data-centre.de/?page_id=688). Leider klappt aber der Zugriff nicht; man landet in einer Dauerschleife von Verbindungsfehlern. Ganz so einfach ist der Umgang mit den lebenden Leichen oder das Channelling in die Cloud also nicht. Selbst wenn es technisch funktionieren sollte, wäre das doch nur eine unveränderliche Archivierung, abgekoppelt vom Internet in dem Sinne, dass Links und Webservices nicht mehr funktionieren. Das ist nicht das, was man mit Nachhaltigkeit meint, und es verlängert die Lebenserwartung einer Edition nicht.
- Ein anderer Ansatz ist die Entwicklung einer generalisierten Arbeits- und Publikationsplattform für Editionen, was eine Homogenisierung von Editionsprojekten *ab ovo* bedeutet. Zu nennen ist hier das schweizerische Projekt ›Nationale Infrastruktur für Editionen‹ (NIE): »Die Plattform soll den spezifischen Bedürfnissen umfangreicher und komplexer Editionsprojekte gerecht werden und insbesondere die elektronische Publikation und die langfristige Verfügbarkeit von Forschungsdaten und -ergebnissen in einem zentralen Bereich der nationalen geisteswissenschaftlichen Forschung gewährleisten« (fee.unibas.ch/de/nie-ine/). 2016 bis 2019 wurden 5 Mio. Franken dafür bewilligt, aber die Homepage der NIE ist im April 2021 zum letzten Mal gesichtet worden (web.archive.org/web/20210423073819/https://www.nie-ine.ch/). Das Unterfangen dürfte also gescheitert sein. Lebendig ist dagegen GAMS, das ›Geisteswissenschaftliche Asset Management System‹ der Uni Graz, wo im Augenblick in der Größenordnung von fast 100 Einzelprojekten gehostet werden (gams.uni-graz.at/context:gams.projekte?locale=de).

Die dort gewählte Herangehensweise bringt es mit sich, dass Editionsprojekte bereits im Benehmen mit der künftigen Publikationsumgebung konzipiert und entwickelt werden müssen. Dies dürfte auch bedeuten, dass dieses Verfahren nicht skalierbar ist: Es können also nicht beliebig viele Projekte betreut werden. Wenn Digitalisierung (im Sinne eines gesamtkulturellen Prozesses) aber nicht eine verbesserte Skalierbarkeit zeitigt, darf man sie kontraproduktiv nennen.

- Eine weiterte Option läge in einer stärkeren Formalisierung und Abstraktion, indem nämlich das »funktionale[] Zusammenspiel aller Komponenten von Text, Struktur, Layout, Schnittstelle und Metadaten« standardisiert beschrieben wird, und zwar im XML-Universum (Stäcker 2020) oder mit Linked Open Data. Das »Desiderat eines abstrakten Modells für Strukturen und Inhalte digitaler Editionen« hat auch Patrick Sahle bei dem Workshop ›Nachhaltigkeit digitaler Editionen‹ 2018 formuliert (Dängeli 2019).

Ich habe bei dem letztgenannten Ansatz den Eindruck, *this is the way*. Aber ich kann nicht abschätzen, wie aufwendig es bei einer komplexen Edition eigentlich wäre, nachträglich formalisiert und maschineninterpretierbar zu beschreiben, wie die Gesamtheit der algorithmischen Prozesse beschaffen sind, die aus den Textrohdaten eine Benutzererfahrung generieren, und mit welchen Werkzeugen man eine solche Modellierung und Beschreibung überhaupt angehen würde, zumal bestehende Editionen in ihrer Funktionsschicht viel Wildwuchs zeigen dürften, weil es eben bislang keine Standardisierung für diese Prozesse gibt. Mir ist auch nicht klar, wie man sich die Software vorstellen soll, auf der das Paket dann quasi ›läuft‹, das aus all den Beschreibungsdaten eine dynamische Benutzeroberfläche macht.

Solange also keiner dieser Ansätze wirklich umsetzbar ist, wird es wohl noch eine Weile in punkto Lebenserwartung dabei bleiben, was schon seit den 90ern gilt: Eine Edition lebt so lange, wie sich jemand aktiv um die Softwarebausteine der Funktionsschicht kümmert, plus schätzungsweise fünf bis fünfzehn Jahre.

Literaturverzeichnis

Primärliteratur

Electronic Beowulf. Student Edition, hrsg. von Kevin Kiernan. 3. Aufl. London 2011 (DVD-ROM).

Electronic Beowulf 4.0, hrsg. von Kevin Kiernan. 2015 (online).

Sekundärliteratur

Buddenbohm, Stefan/Engelhardt, Claudia/Wuttke, Ulrike: Angebotsgenese für ein geisteswissenschaftliches Forschungsdatenzentrum, in: Zeitschrift für digitale Geisteswissenschaften 1 (2016) (online).

Dängeli, Peter: Die Nachhaltigkeitsproblematik digitaler Editionen – Workshop-bericht (online).

Kiernan, Kevin/Iacob, Emil: Going Online (online erreichbar über: https://web.archive.org/web/ 20181004083649/http://ebeowulf.uky.edu:80/)

Stäcker, Thomas: ›A digital edition is not visible‹ – some thoughts on the nature and persistence of digital editions, in: Zeitschrift für digitale Geisteswissenschaften 2020. text/html Format (online).

Online-Ressourcen

Bootstrap: https://getbootstrap.com/.

Codex Sinaiticus: https://codexsinaiticus.org/de/.

Electronic Beowulf: http://ebeowulf.uky.edu:80/.

GAMS (Geisteswissenschaftliche Asset Management System): https://gams.uni-graz.at/.

HDC (Humanities Data Centre): https://humanities-data-centre.de/.

IIIF (International Image Interoperability Framework): https://iiif.io/.

Mediawiki: https://www.mediawiki.org/wiki/MediaWiki.

NFDI (Nationale Forschungsdateninfrastruktur): https://www.nfdi.de/.

PHP (Hypertext Preprocessor): https://www.php.net/.

Python: https://www.python.org/.

re3data: https://www.re3data.org/repository/r3d100010560.

Text+: https://www.text-plus.org/.

WordPress: https://wordpress.com/de/.

Anschrift der Autorin:

Prof. Dr. Sonja Glauch
Friedrich-Alexander-Universität
Institut für Germanistik
Bismarckstraße 1
91054 Erlangen
E-Mail: sonja.glauch@fau.de

Albrecht Hausmann

Digitale Edition (Diskussionsbericht Sektion 2)

Die beiden vorgestellten Editionsprojekte wurden in der Diskussion (Leitung: Michael Stolz) zunächst auf ihr gegenseitiges Verhältnis hin befragt: Handelt es sich bei heiEditions und ediarum.mediaevum um (möglicherweise sogar konkurrierende) Parallelentwicklungen, zwischen denen man sich als künftiger Editor entscheiden muss (Elisabeth Lienert)? Jakub Šimek und Angila Vetter sahen eine solche Konkurrenz nicht: Zum einen kooperierten die Projekte auf verschiedenen Ebenen und tauschten Fähigkeiten aus, zum anderen fokussierten sie auch unterschiedliche Ziele. Ein grundsätzliches Problem bei einer Editionsumgebung wie heiEditions könnte sein, so Stephan Müller, dass die damit verbundenen Standardisierungsansprüche und überhaupt die Auffassung von Editionen als ›Forschungsdaten‹ zu konzeptionellen Beschränkungen führen könnten. So könne man fragen, ob der in heiEditions zugrunde gelegte Werkbegriff nicht hinter den Forschungsstand zurückfalle, den die germanistische Mediävistik hierzu in den letzten Jahrzehnten erreicht habe. Ebenso sei zu fragen, ob eine solche Editionsumgebung offen genug sei, um neue Entwicklungen aufnehmen zu können (Müller). Im Zusammenhang damit stand auch die Frage, ob und wie die Rohdaten der entsprechenden Editionen (hier bei heiEditions) auch für neue und nach ganz anderen Editionsprinzipien verfahrende Projekte verwendet werden könnten (etwa auf der Grundlage einer Datenschnittstelle oder zumindest der klaren Trennung von Daten und GUI, Albrecht Hausmann). Beide Fragen zeigten auch den Bedarf einer dauerhaften aktiven Pflege der z. B. bei heiEditions gehosteten digitalen Editionen an (Šimek). In diesem Zusammenhang wurde die künftige Rolle von Bibliothe-

ken diskutiert, die als auf Dauer angelegte Institutionen bei der Bereitstellung und kontinuierlichen Pflege von digitalen Editionen eine wichtige Rolle spielen könnten. Für Šimek könnten Bibliotheken sogar in die Rolle von Mitherausgebern rücken. Systematisch gesehen benötigten digitale Editionen eine ähnliche Bereitstellungsinfrastruktur wie gedruckte Editionen; die Herausforderung bestehe deshalb im Aufbau eines Datenmanagementsystems, das ebenso wie das analoge Bibliothekssystem nicht ohne finanzielle Ressourcen auskommen werde (Andrea Rapp).

Damit war die Problematik der Nachhaltigkeit digitaler Editionen angesprochen (Vortrag Sonja Glauch), die einen weiteren Schwerpunkt der Diskussion bildete. Für Freimut Löser gehöre es zum Wesen der Edition, dass sie veraltet – das sei bei gedruckten Editionen nicht anders als bei digitalen. Allerdings zähle auch die konzeptionell veraltete Edition zum Archiv einer Wissenschaft; wie aber seien dann digitale Editionen so zu archivieren, dass sie funktional bleiben? In diesem Zusammenhang machte Brigitte Bulitta auf den grundsätzlichen Unterschied zwischen digitalen Editionen (die digitale Funktionalitäten bewusst nutzen und z. B. dynamische Darstellungsweisen durch Einsatz von Software während der Benutzung bieten) und nur digitalisierten Editionen (bei denen lediglich analoge Editionen ins digitale Medium transferiert werden) aufmerksam. Für Michael Stolz werde sich die Frage der Nachhaltigkeit auch über die Nutzung klären: Was gebraucht und genutzt werde, werde auch bleiben. Gabriel Viehhauser wies darauf hin, dass Nachhaltigkeitsprobleme nicht zuletzt Standardisierungsfragen berührten; durch die Schaffung einheitlicher Datenstandards könne auch im Bereich der digitalen Editionen Nachhaltigkeit erzeugt werden.

Anschrift des Berichterstatters:

Prof. Dr. Albrecht Hausmann
Carl von Ossietzky Universität Oldenburg
Institut für Germanistik
26111 Oldenburg
E-Mail: albrecht.hausmann@uni-oldenburg.de

Sektion 3: Digitale Infrastruktur und Forschungsdatenmanagement

Andrea Rapp

Digitale Infrastruktur und Forschungsdatenmanagement

Abstract. Seit 2020 wird in Deutschland mit einer Bund-Länder-Förderung die Nationale Forschungsdateninfrastruktur (NFDI) aufgebaut. Sie hat zum Ziel, Datenbestände aller Wissenschaftsbereiche entlang der FAIR-Prinzipien und ohne Disziplinen- und Ländergrenzen zu erschließen und langfristig zu sichern. Der Beitrag stellt das NFDI-Konsortium Text+ vor, das die digitale Infrastruktur und das Forschungsdatenmanagement für Sprach- und Textdaten organisiert. Hierbei liegt der Fokus zum einen auf Anwendungsbeispielen aus der Mediävistik, zum anderen auf den Möglichkeiten der Beteiligung der Fachcommunity, der zukünftigen Weiterentwicklung sowie den bereits artikulierten konkreten Bedarfen.

1. Vorbemerkung/Einleitung

Anders als in vielen Natur- und Ingenieurwissenschaften drängt sich ein Bezug zu ›Infrastrukturen‹ für viele Geisteswissenschaften zunächst nicht auf. Dennoch sind gerade die Geistes- und Kulturwissenschaften[1], die sich mit den kulturellen Erzeugnissen des Menschen (im weiten Sinne) befassen, seit jeher auf umfangreiche analoge Infrastrukturen angewiesen und haben ihren Aufbau, ihre Pflege und ihren Betrieb seit Jahrhunderten professionalisiert. Archive, Bibliotheken und Museen und die dazugehörigen Wissenschaften und Ausbildungsberufe sind für geisteswissenschaftliche Forschung und Lehre unabdingbar. Aus der umfassenden digitalen Transformation dieser Einrichtungen und des gesamten Wissenschaftsbetriebs erklären sich die Notwendigkeit und der Anspruch der Geistes-

wissenschaften auf professionelle digitale Infrastrukturen. Zum einen liegen immer mehr – wenn auch noch lange nicht alle – Quellen und Ressourcen geisteswissenschaftlicher Forschung als digitales Surrogat vor, zum anderen wird mit digitalen Methoden geforscht. In diesem Prozess entstehen annotierte, mit Expertise und Wissen angereicherte digitale Derivate der ursprünglichen Forschungsobjekte als Ergebnisse und Grundlagen weiterer geisteswissenschaftlicher Forschung, deren Bewahrung und Pflege – das Forschungsdatenmanagement – derselben Professionalität bedürfen wie analoge Archive, Bibliotheken und Museen.

Im Folgenden möchte ich zeigen, wie sich das NFDI-Konsortium Text+ der Aufgabe annimmt, die digitale Infrastruktur und das Forschungsdatenmanagement für Sprach- und Textdaten in diesem Sinne aufzubauen und zu organisieren. Ich fokussiere dabei für den Beitrag in unserem Rahmen exemplarisch auf die mediävistische Philologie. Nach einer Definition dessen, was geisteswissenschaftliche Forschungsdaten umfassen und einer kursorischen Vorstellung der Nationalen Forschungsdateninfrastruktur NFDI gehe ich vor allem auf das Konsortium Text+ ein. Hierbei liegt der Fokus auf den Möglichkeiten der Beteiligung der Fachcommunity, der zukünftigen Weiterentwicklung sowie den bereits artikulierten konkreten Bedarfen.

2. Was sind Forschungsdaten in den Geisteswissenschaften?

Ganz vereinfacht gesagt, sind alle Daten, die im Rahmen des (geistes-) wissenschaftlichen Forschungsprozesses erzeugt und benötigt werden, Forschungsdaten. Mit dem Vorschlag von Johanna Drucker, solche Daten nicht als »data« (something given), sondern als »capta« (something actively taken) zu sehen, erhalten sie ihre epistemische Qualität (Drucker 2011, S. 3). Sowohl der Rat für Informationsinfrastrukturen (2016, Anhang, S. A–13) als auch das DARIAH-Stakeholdergremium ›Sammlungen‹

(Oltersdorf/Schmunk 2016, S. 3) haben Definitionen für Forschungsdaten erarbeitet.

Im Hinblick auf den Vorgang der Datenerhebung umfassen geisteswissenschaftliche Forschungsdaten demnach sämtliche digitale Daten, die bei Quellenerschließungen, Transkriptionen, Editionen, Erhebungen, Grabungen, Experimenten, Messungen, Simulationen, Aufnahmen oder Umfragen usw. entstehen oder deren Ergebnis sind. So vielfältig wie die Erhebungsmethoden sind auch ihre Formen, Medien und Formate, sie erscheinen als Texte, Bilder, mehrdimensionale Modelle, Audio, Video, Tabellen, Datenbanken, Dokumentationen, Software, Verzeichnisse usw. Sie liegen disziplinspezifisch in unterschiedlichen Aggregationsstufen vor und verwenden zumeist verschiedene digitale Formate. Sie sind eingebunden in den Forschungsdatenlebenszyklus, d. h. Ergebnis und Ausdruck einzelner Schritte des Forschungsprozesses bzw. Grundlage und Ausgangspunkt weiterer Schritte (vgl. forschungsdaten.info).

3. Die Geisteswissenschaften in der Nationalen Forschungsdateninfrastruktur

Die drängende Notwendigkeit, den nachhaltigen Betrieb digitaler Infrastrukturen für das Forschungsdatenmanagement für die gesamte Wissenschaftslandschaft in Deutschland zu organisieren, führte 2018 zu einer Bund- und Ländervereinbarung zum zunächst auf zweimal fünf Jahre befristeten Aufbau einer Nationalen Forschungsdateninfrastruktur (Bund-Länder-Vereinbarung). Als Ziel wurde die Erschließung und langfristige Sicherung von Datenbeständen entlang der FAIR-Prinzipien und ohne Disziplinen- und Ländergrenzen formuliert. Dafür sollten 30 Konsortien für eine breite Abdeckung der Wissenschaft eingerichtet werden.[2] Zentral für Akzeptanz und Bedarfsorientierung ist dabei, dass »[d]ie NFDI [...] in einem aus der Wissenschaft getriebenen Prozess als vernetzte Struktur eigeninitiativ agierender Konsortien aufgebaut [wird].« (Deutsche Forschungs-

gemeinschaft 2020). Das bedeutet, dass die Konsortien (und mithin die NFDI) von den Wissenschaftler:innen selbst in Kooperation mit passenden Infrastruktureinrichtungen geformt, betrieben und gesteuert werden. Durch diese Konstruktion wird gewährleistet, dass die Infrastruktur bedarfs- und wissenschaftsgerecht angelegt wird, sie erfordert jedoch auch von den Forschenden einen hohen Aufwand an Ressourcen und Zeit und nicht zuletzt an Kenntnissen in einem Maße, das vermutlich doch deutlich über das auch vorhandene Engagement für analoge Infrastrukturen hinausgeht. Die Nachhaltigkeit des Konstrukts wird sich ebenfalls noch erweisen müssen, denn als Förderung sind zunächst zweimal fünf Jahre vorgesehen, evaluiert wird der NFDI-Prozess durch die DFG und den Wissenschaftsrat (vgl. Deutsche Forschungsgemeinschaft 2018).

Für die Geisteswissenschaften stellen die Vielfalt und Heterogenität ihrer Disziplinen einerseits und die Überschneidungen bei den Quellen und Ressourcen bei gleichzeitiger Heterogenität der Zugriffsmethoden andererseits eine große Herausforderung dar. Dem sind sie durch einen intensiven Vernetzungs- und Abstimmungsprozess begegnet, der in einem Memorandum of Understanding dokumentiert ist, das für die erste Antragsrunde 2019 publiziert und anschließend für die weiteren Runden aktualisiert wurde (vgl. Brünger-Weilandt [u. a.] 2020 und forschungsinfrastrukturen.de). Dieses Memorandum legt den Grundstein für einen gemeinsamen Aufbau einer NFDI in den Geistes- und Kulturwissenschaften, indem sich die Initiativen zur Kooperation verpflichten und Verantwortlichkeiten und die Form ihrer Zusammenarbeit definieren.

Gemäß der Anlage der NFDI-Struktur als Community-getragene Institution haben sich auch die Geisteswissenschaften an den Bedarfen spezifischer Communitys organisiert. Das seit 2020 geförderte Konsortium NFDI4Culture adressiert die Architektur, die Kunstgeschichte, die Musik-, Theater- und Tanzwissenschaft sowie die Medienwissenschaften, während sich das seit 2021 geförderte Konsortium Text+ an alle sprach- und textbasierten Disziplinen wendet, insbesondere an die Linguistik, die Alt- und

Neuphilologien, die Literaturwissenschaft, die Philosophie, die Anthropologie sowie die nichteuropäischen Kulturen und Sprachen. Die aktuell (Stand Mai 2022) in der Begutachtung befindlichen Konsortien NFDI4memory und NFDI4objects richten sich auf die Geschichte, die Wirtschafts- und Sozialgeschichte, die Area Studies, die Religionswissenschaft, die historische Philosophie sowie auf Gedächtnisinstitutionen bzw. auf die Archäologie, die Archäometrie, die Archeological Sciences, die Anthropologie, die Antiken Architekturen, die Cultural Heritage Studies und die Conservation Science. Als gemeinsame Querschnittsbereiche wurden Metadaten, Normdaten, Terminologien, Provenienz, Recht und Ethik sowie Data Literacy identifiziert.

Sprache und Text spielt in zahlreichen Bereichen eine wichtige, ja eine zentrale Rolle, nicht allein in den Geisteswissenschaften. Der Fokus von Text+ lässt sich dennoch gut eingrenzen durch das auf Sprache und Text selbst gerichtete Erkenntnisinteresse, das seine Community eint.

4. Das NFDI-Konsortium Text+

Wie oben bereits benannt, baut Text+ eine auf Sprach- und Textdaten ausgerichtete Forschungsdateninfrastruktur auf. Das Konsortium orientiert sich dabei in seiner Struktur dezidiert an Datendomänen, nicht an Fach-Disziplinen, um der Vielfalt und Heterogenität der Fächer und ihrer Nutzungscommunitys gerecht zu werden. Der aktuelle Fokus der ersten Förderphase liegt auf den Datendomänen digitale Sammlungen/Korpora, lexikalische Ressourcen und Editionen. Sie haben eine für die Disziplinen häufig konstitutive Funktion, eine lange Forschungstradition und sind mit ausgereiften methodologischen Paradigmen verknüpft, die charakteristische, aber auch bereichsübergreifende Praktiken der Erzeugung, Kuratierung und des Managements von Daten erfordern. Bevor diese Datendomänen gleich ausführlicher vorgestellt werden, soll kurz erläutert werden, wie bereits im Vorfeld bzw. bei der Profilierung des Konsortiums und

der Ausarbeitung des Antrags Bedarfe der Forschenden erhoben wurden. Ihre Berücksichtigung ist durch demokratische Einbindung in die Governance-Struktur und im Monitoring durch gewählte Wissenschaftler:innen fest in Text+ verankert.[3]

Im Sommer 2020 wurde ein initialer und thematisch offener Call for User Stories gestartet, mit denen eine große Bandbreite an Disziplinen, Datendomänen und Forschungsfragen erfasst werden sollte. Dieser Call erbrachte innerhalb kurzer Zeit 120 Rückmeldungen mit ausgearbeiteten User Stories erbrachte. Aufgrund der Vorgaben, möglichst eine inhaltliche Struktur einzuhalten, die die Bereiche 1. Motivation, 2. Ziele, 3. Herausforderungen, 4. Lösung, 5. Evaluation durch die Community umfasste, konnten die Stories sehr gut verglichen und ausgewertet werden und sind in großem Umfang in den Antrag eingegangen. Alle User Stories wurden nach Datendomänen sowie nach DFG-Fachsystematik klassifiziert und unter den Namen und mit Zustimmung der Autor:innen auf der Text+-Homepage veröffentlicht. Die Auswertung, die klare Prioritäten der Nutzenden ergeben hat (z. B. Interoperabilität und Nachnutzbarkeit sowie Zugänglichkeit von Daten), sowie die Daten für die Analyse sind ebenfalls veröffentlicht (Rißler-Pipka [u. a.] 2021). Diese Calls sollen in regelmäßigen Abständen wiederholt werden, um auf aktuelle Entwicklungen und Bedarfe eingehen zu können und die Nutzungscommunity erweitern zu können.

Im Folgenden wird nun beschrieben, welche Datentypen und -angebote zu den drei Datendomänen gerechnet werden, im Anschluss werden einige User Stories aus der Mediävistik exemplarisch vorgestellt und eingeordnet.

Sammlungen

Zu dieser umfangreichen, sehr offen gestalteten und vielfältigen Datendomäne zählen Sprach- und textbasierte Sammlungen aller Art. Sie umfasst Sammlungen geschriebener, gesprochener oder gebärdeter Sprache und Texte sowie sprach- und textbezogene Experimental- oder Messdaten, die auf Grundlage wissenschaftlicher Kriterien gesammelt wurden, z. B.

Textsammlungen, mono- und multimodale Aufnahmen, Sensordaten, Befragungen, Verzeichnisse, Kataloge usw. Zu dieser Domäne gehören 61 User Stories, sie stellt damit mit über 50 Prozent Anteil den weitaus größten Bereich und zeigt den immensen Bedarf am Aufbau, insbesondere an der nachhaltigen und offenen Verfügbarkeit digitaler Sammlungen aller Art für die geisteswissenschaftliche Forschung auf. Drei Beispiele aus der Mediävistik demonstrieren zum einen die Bandbreite der Datentypen und der damit verbundenen Anforderungen, zum anderen nicht zuletzt die Interdisziplinarität der ›Mediävistik‹ im Sinne eines *umbrella term*.

Die User Story ›Reference Corpus Middle Low German/Low Rhenish (1200–1650) (ReN)‹[4] beschreibt eine in der Forschung gängige Situation, wie Korpora mit viel Aufwand aufgebaut und mit erschließenden Annotationen angereichert werden. Eine Verknüpfung mit weiteren Korpora und niedrigschwelligen Abfrage- und Nutzungsmöglichkeiten einerseits könnten den Wert für die philologische Forschung noch deutlich erhöhen, eine nachhaltige Bereitstellung durch technologische Aktualisierungen andererseits schonte Ressourcen und sorgte für wissenschaftliche Transparenz und Kontinuität. Projektförmige Strukturen und Personalmangel und -fluktuation führen häufig dazu, dass solche wertvollen Ressourcen verwaisen – eine zentrale Infrastruktur könnte hier Abhilfe schaffen.

Die Sammlung ›Roman Seals and Inscriptions‹[5] adressiert eine Quellengruppe, die nur selten in regionalen Inschriftenkorpora aufgenommen wird bzw. dort nachgewiesen ist, dadurch wird die Forschung erschwert. Insbesondere der philologischen Forschung sind wertvolle Quellen nur schwer zugänglich und damit kaum mit anderen Inschriftenquellen vergleichbar, so dass hier kein Gesamtüberblick erreichbar ist. Eine Koordination der weltweit verstreuten Quellen und Interoperabilität der Erschließungsinstrumente würde die Forschung also entscheidend erleichtern.

Die ›Bibliotheca Arabica‹[6] erschließt die Geschichte der Arabischen Literatur von den Anfängen bis ins 19. Jahrhundert und mit ihr eine reichhaltige und vielfältige Handschriftenkultur. Eine besondere Herausforde-

rung liegt hier in der Verarbeitung der nichtlateinischen Schrift. Auch hier werden klar Unterstützung und Koordination beim Thema Interoperabilität und Standardisierung adressiert.

Diese (und weitere) User Stories im Feld ›Sammlungen‹ eint, dass zum einen zentrale Anlaufstellen und Orte für nachhaltige Sicherung und Pflege von Daten gefordert werden, weil Einzelforschende oder zeitlich befristete Projekte diese Nachhaltigkeit nicht leisten können, zum anderen der Wunsch nach Standardisierung und Interoperabilität – auch auf internationalem Level.

Lexikalische Ressourcen

Unter Lexikalische Ressourcen fassen wir Daten, die die Verwendung von Wörtern in Sätzen, Texten und multimodaler Kommunikation beschreiben, z. B. Wörterbücher und Enzyklopädien, Normdaten, terminologische Datenbanken, Wortnetze, Ontologien, Wortkarten und linguistische Atlanten. Obwohl die Wörterbucherstellung und die Lexikographie seit langem und umfassend digital transformiert sind und Normdaten und Ontologien zu den Schlüsseltechnologien eines FAIRen Forschungsdatenmanagements gehören, sind hier mit 19 User Stories die wenigsten Rückmeldungen eingegangen. Dies ist darauf zurückzuführen, dass beispielsweise Wörterbucharbeit in vergleichsweise wenigen Vorhaben und Arbeitsstellen gebündelt ist, anders als Sammlungen und Korpora, die fast flächendeckend in den Wissenschaften entstehen.

Die User Story ›Old High German Dictionary: Supplements‹[7] spricht ein klassisches, langlaufendes Akademiewörterbuch an, das als Sprachstadienwörterbuch und Thesaurus die Aufgabe hat, das Althochdeutsche gesamthaft zu erfassen und zu dokumentieren. Die digitale Version ermöglicht zwar leicht die Aufnahme von Neufunden und Korrigenda, dennoch muss der Konnex zum gedruckten Wörterbuch und damit die Zitierfähigkeit gewahrt werden. Neben Datenhaltung und Versionierung tut sich damit die

Anforderung auf, Datenaggregation zu steuern und die Daten insgesamt persistent zu adressieren.

Dagegen geht es bei der User Story ›Word family database of historical German‹[8] um ein Vorhaben, das auf den vorhandenen historischen Wörterbüchern und Referenzkorpora zum Deutschen aufbaut und diese Daten nachnutzt. Bedarf besteht insbesondere im Hinblick auf ein zentrales Repository sowie bei der Harmonisierung vor allem der Metadaten, aber auch bei der Entwicklung von Tools.

Die Verbindung von gedruckten und digitalen Ergebnispublikationen und das dazugehörige Spannungsfeld von statischer, aber recht nachhaltiger Buchpublikation und dynamischer Webpublikation ist ein Aspekt vieler zentraler Nachschlagewerke und Wörterbücher – allein aufgrund des Umfangs des dort bearbeiteten Materials und der damit zusammenhängenden langen Laufzeit von Vorhaben, die zum Teil ja noch vor der Digitalen Transformation begonnen wurden. Zugleich wird in Feld Lexikalische Ressourcen besonders deutlich, welche Potenziale in der Aggregation, Interoperabilität, Nachnutzung und übergreifenden Auswertungsmöglichkeit dieser Daten liegen.

Editionen

Editionen sind kritische Repräsentationen historischer Dokumente. Sie bestehen aus der zuverlässigen methodengeleiteten Bewahrung, Präsentation und Kommentierung aller Arten von Texten in verschiedenen Sprachen und Schriftsystemen. Sie liefern damit unabdingbare Grundlagen für (geistes-)wissenschaftliche Forschungen. Die Relevanz zeigt sich bereits durch einfache Recherchen in Forschungsinformationssystemen; so liefert die Datenbank GEPRIS rund 1500 ›Editionen‹, das Informationssystem der Akademien AGATE meldet 111 mit dem Schlagwort ›Edition‹ getaggte Vorhaben.[9] In den 23 User Stories dieser Datendomäne wird der Bedarf nach nachhaltigen Infrastrukturen für die Sicherung der Ergebnisse ganz besonders deutlich.

›Editing a Medieval Hispanic Poetry Corpus with TextGrid‹[10] befasst sich mit der Edition mittelalterlicher Literatur, wobei die Forschenden an ihrem Standort in Argentinien nicht auf ein Umfeld gleichartiger Vorhaben vor Ort bauen konnten, sondern Kooperation im internationalen Bereich suchten. Die Nutzung von TextGrid-Komponenten war aufgrund der Offenheit zwar leicht möglich, dennoch ergaben sich im Detail Schwierigkeiten bei der Nutzung eigener TEI-Schemata und avancierterer X-Technologien. Auch eine Diversifizierung von Dokumentationssprachen (Multilingualität von Software und ihrer Dokumentation) wird angemahnt.

Die User Story ›Digital Humanities and Medieval Literary Studies‹[11] geht generisch auf die Problematik ein, dass für die Erstellung von Editionen neben philologischen auch technologische Kompetenzen erforderlich sind. Daher besteht ein hoher Bedarf an Beratung, Weiterbildung und Training bzw. spezifischer Kooperation. Bei Editionen entstehen wertvolle Forschungsdaten als ›Zwischen- und Neben-Ergebnis‹ wie etwa Transkriptionen, für die ein zentraler Nachweis ebenso fehlt wie ein zuverlässiger Speicherort. Und nicht zuletzt sind das Interface einer Edition und seine Funktionalität ein zentrales Ergebnis, das ebenso wie die ›(Roh-)Daten‹ gesichert werden muss.

›Inscriptions in Germany from the Middle Ages to Early Modern Times (Die Deutschen Inschriften)‹[12] verdeutlicht die Besonderheit der Quellengattung Inschriften, die u. a. darin besteht, dass die Verbindung von Schrift, Schriftträger und Aufstellungsort nicht gelöst werden kann, zugleich sind diese Objekte ungeheuer vielfältig (Gebäude, Skulpturen, Grabsteine, Glocken, liturgische Objekte usw.). Die Erstellung von kontrollierten Vokabularen, das Aggregieren von Daten und die Aushandlung und Propagierung von Standards sind daher unabdingbar für die Inschriftenforschung und könnten über verschiedene geisteswissenschaftliche Konsortien hinweg adressiert werden.

Auch im Feld Editionen zeigt sich die spezifische Vielfalt der Anforderungen und Bedarfe: zum einen gibt es sehr viel Abstimmungsbedarf in

einem in vielfältiger Hinsicht und aus Forschungsgründen auch zurecht heterogenen Bereich, zum anderen muss nach organisatorischen und technischen Lösungen für die nachhaltige Sicherung von Editionen, insbesondere für die Roh- und ›Peripheriedaten‹ sowie für die Interfaces gesucht werden.

5. FAIR-Prinzipien und Vernetzung

Obwohl die Aufgliederung in die genannten drei Datendomänen aufgrund ihrer je spezifischen Anforderungen im Datenmanagement notwendig und sinnvoll ist, gibt es auch viele Berührungs- und Überschneidungspunkte, die nicht aus dem Blick geraten dürfen, sondern die erst eine echte wechselseitige Erschließung erlauben. Daher ist der systematische Ausbau von Verlinkungen zwischen den Datendomänen ein zentrales Ziel von Text+-. Editionen und Korpora liefern gesicherte Belege für lexikalische Ressourcen, während wiederum Lexikalische Ressourcen Editionen und Korpora strukturiert erschließen. Korpora liefern Rohdaten und Grundlagen für Editionen, während Editionen ›korpusfähig‹ und in Korpora integriert werden. Grundlage dafür sind zum einen persistente Identifikatoren und normierte Vokabulare bzw. allgemeine, standardisierte Normdaten wie die Gemeinsame Normdatei (GND), GeoNames sowie beispielsweise der Vokabular-Server DANTE. Hier finden sich persistent adressierbare Normdaten zu Personen, Körperschaften, Geografika, Ereignissen, Werken und Sachbegriffen, deren Verlinkung – z. B. in den Apparaten von Editionen – den wissenschaftlichen Kommentar von Routineangaben entlastet. Für die GND gilt beispielsweise, dass durch die Verbindung mit Thesauri und Fachdatenbanken der verschiedenen wissenschaftlichen Disziplinen einerseits und weiteren offenen Normdaten andererseits zu jeder GND-Entität nach Bedarf dann auch Daten aus der Linked Data Cloud wie Wikidata, Wikipedia, VIAF usw. ›automatisch‹ verlinkt werden können. Durch solche Verlinkungen werden Wissensbasen auf- und ausgebaut, so dass nicht

allein die Auffindbarkeit von Forschungsdaten (findability), sondern auch die Möglichkeiten von automatisierten Verfahren wie Text and Data Mining (reusability) entscheidend befördert werden. In Deutschland wird die GND als kollaboratives zentrales Tool breit disziplinübergreifend genutzt und kooperativ ausgebaut und gepflegt. Darüber hinaus ist sie international mit Normdaten (authority files) weiterer Nationalbibliotheken verlinkt und damit ein wichtiger Hub in der Linked Open Data Cloud, die auch in Text+ eine entscheidende Rolle spielt.[13] Die GND soll daher für sprach- und textbasierte Forschungsdaten weiterentwickelt und ausgebaut werden, was nicht allein den Textdisziplinen zugutekommen wird, sondern als Beitrag zu einem domänenübergreifenden Wissensgraphen der NFDI insgesamt sowie weiten Bereichen der Kultur und Wissenschaft zu sehen ist. Im Rahmen von Text+ wird daher eine entsprechende GND-Agentur eingerichtet, um niedrigschwellige qualitätsgesicherte Beteiligungsmöglichkeiten für Forschende zu schaffen und zugleich den Vernetzungsgrad der GND auch durch Terminologie-Mappings zu erweitern.

Diese effiziente, das Wissensnetz erweiternde und verdichtende Erschließung und Sicherung unserer Quellen gelingt jedoch nur, wenn möglichst alle wissenschaftlichen Forschungsdaten im Sinne der FAIR- und CARE-Prinzipien verfügbar gemacht werden (Blumesberger 2021). Wenn wir erwarten, dass wir für unsere Forschung auf Daten des Kulturellen Erbes und geisteswissenschaftliche Forschungsergebnisse dauerhaft zugreifen können, müssen wir nicht allein Aspekte von Individualität/Besonderheit und Standardisierung aushandeln, sondern auch und vor allem unsere eigenen Daten und Ergebnisse zur Verfügung stellen. Es ist jedoch auch klar, dass dies nur in einer vertrauenswürdigen und nachhaltig abgesicherten digitalen Forschungsinfrastruktur erfolgen kann – insofern muss die Chance NFDI von den Forschenden genutzt und mitgestaltet werden.

6. Zusammenfassung: Community-Beteiligung erwünscht!

Text+ bietet verschiedene und vielfältige Möglichkeiten der Partizipation und der Mitgestaltung. Da ist zum einen das Angebot, das Text+ in Form von Daten, Tools und Services an die Community macht, zum anderen die Einladung zum Engagement und zur Mitbestimmung im Konsortium. Hier greife ich nur einige Dinge exemplarisch heraus, über die Website und die Kommunikationsaktivitäten z. B. in den Sozialen Medien oder auch direkte Anfragen kann man sich über alle Angebote informieren und sich einbringen.

Tools alleine ohne Dokumentationen, Tutorials und Schulungen haben kaum Chancen auf Nutzung, daher sind Aktivitäten im Bereich Education and Training auch für Text+ zentral. Es werden eigene Workshops und Tagungen entwickelt, aber auch Beiträge zu anderen Workshops, Verbandstagungen und Summerschools usw. geleistet. Die Struktur und damit die Bündelung von Expertise nach Datendomänen ist besonders gut geeignet, um adressatenorientierte und zielgruppenspezifische Angebote zu entwickeln. Darüber hinaus wird die Community (mindestens) jährlich zu Vollversammlungen eingeladen, das Text+ Plenary 2022 findet am Montag, den 12. September 2022 in Mannheim statt.

Text+ verfügt über die Möglichkeit, das Portfolio zu erweitern, indem neue Partner und Angebote aufgenommen werden. Im April 2022 erfolgte eine erste Ausschreibung zur Einreichung von Kooperationsprojekten zur weiteren Integration von Daten und Services in die Text+-Infrastruktur, die über die sogenannten Flexfunds des Antrags ermöglicht werden und deren Auswahl von den Mitgliedern der SCCs vorgenommen wird.[14] Nicht zuletzt soll noch einmal darauf hingewiesen werden, dass gerade die SCCs und das OCC als zentrale Mitbestimmungsorgane des Konsortiums vom Engagement der Community leben. Bei der Vollversammlung im September 2022 werden die Vertreter*innen für die nächsten Jahre gewählt (Weimer 2022).

Der oben beschriebene Call for User Stories sowie die Ausschreibung für Kooperationsprojekte werden regelmäßig wiederholt.

Es liegt also auch an den Mediävist*innen, ihrer Disziplin und ihren Anliegen in der NFDI eine starke Stimme zu geben.

Anmerkungen

1 Ich folge in unserem Zusammenhang einer weiten Definition von Geistes- und Kulturwissenschaften und werde im Folgenden Geisteswissenschaften als ebensolchen weit gefassten Terminus verwenden und nicht weiter ausdifferenzieren; vgl. abouthumanities.sagw.ch/02-was-geisteswissenschaften.html. Alle im Beitrag angegebenen URLs wurden zuletzt eingesehen am 26.05.2022.

2 Stand Mai 2022 werden 19 Konsortien gefördert (www.nfdi.de/konsortien/); weitere 15 bewerben sich in der dritten Ausschreibungsrunde um Förderung. Informationen zum Prozess hat aktuell www.dfg.de/foerderung/programme/nfdi/.

3 Siehe dazu die Governancestruktur mit Scientific Coordination Committees und dem Operations Coordination Committee, die den Task Areas zugeordnet und deren Mitglieder u. a. von einschlägigen Fachverbänden nominiert und gewählt sind; www.text-plus.org/ueber-uns/governance/.

4 Ingrid Schröder, Robert Peters: www.text-plus.org/en/research-data/user-story-339/.

5 Marietta Horster: www.text-plus.org/en/research-data/user-story-308/.

6 Boris Liebrenz: www.text-plus.org/en/research-data/user-story-346/.

7 Uwe Kretschmer, Brigitte Bulitta: www.text-plus.org/en/research-data/user-story-517/.

8 Jost Gippert, Sarah Ihden, Ralf Plate, Ingrid Schröder: www.text-plus.org/en/research-data/user-story-516/.

9 Auswertungen Stand Mai 2020.

10 Gimena del Rio Riande: www.text-plus.org/en/research-data/user-story-407/.

11 Gabriel Viehhauser: www.text-plus.org/en/research-data/user-story-409/.

12 Arbeitsstelle Inschriften, Nordrhein-Westfälische Akademie der Wissenschaften und der Künste (Universität Bonn): www.text-plus.org/en/research-data/user-story-420/.

13 Vgl. die Grafik Folie 9 Semantische Interoperabilität in: Hinrichs [u. a.] 2020.

14 www.text-plus.org/forschungsdaten/kooperationsprojekte/.

Literaturverzeichnis

Sekundärliteratur

Blumesberger, Susanne: Forschungsdaten in den Geisteswissenschaften. Bereits selbstverständlich oder doch noch etwas exotisch? In: o-bib. Das offene Bibliotheksjournal 8–4 (2021) (online).

Brünger-Weilandt, Sabine/Bruhn, Kai-Christian/Busch, Alexandra W./Hinrichs, Erhard/Maier, Gerald/Paulmann, Johannes/Rapp, Andrea/von Rummel, Philipp/Schlotheuber, Eva/Schmidt, Dörte/Schrade, Torsten/Simon, Holger/Stein, Regine/Teich, Elke: Memorandum of Understanding by NFDI Initiatives from the Humanities and Cultural Studies, Version 28.09.2020 (online).

Bund-Ländervereinbarung zu Aufbau und Förderung einer Nationalen Forschungsdateninfrastruktur (NFDI) vom 26. November 2018, in: Bundesanzeiger Amtlicher Teil, 21.12.2018 B10 (online).

Deutsche Forschungsgemeinschaft: Nationale Forschungsdateninfrastruktur (NFDI): DFG übernimmt Auswahl und Evaluation der Konsortien, in: Pressemitteilung Nr. 58 (07. 12.2018) (online).

Deutsche Forschungsgemeinschaft: Nationale Forschungsdateninfrastruktur – Ausschreibung 2020 für die Förderung von Konsortien, in: Information für die Wissenschaft Nr. 29 (25.05.2020) (online).

Drucker, Johanna: Humanities Approaches to Graphical Display. In: DHQ: Digital Humanities Quarterly 5.1 (2011) (online).

Hinrichs, Erhard/Henrich, Andreas/Rapp, Andrea/Stein, Regine: Text+: Sprach- und Textbasierte Forschungsdateninfrastruktur. Text+ Präsentation bei der NFDI-Konferenz 2020. Leibniz-Institut für Deutsche Sprache (IDS) Mannheim 2020 (online).

Oltersdorf, Jenny/Schmunk, Stefan: Von Forschungsdaten und wissenschaftlichen Sammlungen. Zur Arbeit des Stakeholdergremiums »Wissenschaftliche Sammlungen« in DARIAH-DE. Bibliothek Forschung und Praxis 40 (2016), S. 179-185 (online).

Rat für Informationsinfrastrukturen: Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland. Göttingen 2016 (online).

Rißler-Pipka, Nanette/Barthauer, Raisa/Buddenbohm, Stefan/Calvo Tello, José/Friedrichs, Sonja/Weimar, Lukas: Community Involvement in Research Infrastructures: The User Story Call for Text+. 2021 (online).

Weimer, Lukas: Bitte mitmachen! DHd-Blog 31. Januar 2022 (online).

Online-Ressourcen

About Humanities. Was sind Geisteswissenschaften?: https://abouthumanities.sagw.ch/02-was-geisteswissenschaften.html.
AGATE. A European Gateway for the Academies of Sciences and Humanities: https://agate.academy/.
Bund-Ländervereinbarung zu Aufbau und Förderung einer Nationalen Forschungsdateninfrastruktur (NFDI) vom 26. November 2018: https://www.gwk-bonn.de/fileadmin/Redaktion/Dokumente/Papers/NFDI.pdf.
DANTE (DAtendrehscheibe für Normdaten und TErminologien): https://dante.gbv.de/search.
DARIAH-DE (Digital Research Infrastructure for the Arts and Humanities): http://dx.doi.org/10.20375/0000-000E-67ED-4.
Forschungsdaten.info: https://www.forschungsdaten.info/themen/informieren-und-planen/datenlebenszyklus/.
Geistes- und kulturwissenschaftliche Forschungsinfrastrukturen e. V.: http://forschungsinfrastrukturen.de/.
GeoNames: http://www.geonames.org/.
GIDA (Global Indigenous Data Alliance): https://www.gida-global.org/care.
GND (Gemeinsame Normdatei): https://gnd.network/Webs/gnd/DE/Home/home_node.html.
GO FAIR: https://www.go-fair.org/fair-principles/.
NFDI (Nationale Forschungsdateninfrastruktur): https://www.nfdi.de/.
RfII (Rat für Informationsinfrastrukturen): https://rfii.de/de/start/
TextGrid: https://textgrid.de/.
Text+: https://www.text-plus.org/
https://www.text-plus.org/ueber-uns/governance/;
https://www.text-plus.org/forschungsdaten/kooperationsprojekte/;
https://www.text-plus.org/forschungsdaten/user-stories/.

Anschrift der Autorin:

Prof. Dr. Andrea Rapp
Technische Universität Darmstadt
Institut für Sprach- und Literaturwissenschaft
Residenzschloss, Marktplatz 15
64283 Darmstadt
E-Mail: andrea.rapp@tu-darmstadt.de

Thomas Burch

Infrastrukturprojekte zur digitalen Lexikographie

Vorgestellt am Beispiel des Zentrums für Historische Lexikographie

Abstract. Forschungsinfrastrukturen stellen Instrumente, Ressourcen und Dienstleistungen zur Verfügung, deren Ziele darin bestehen, Forschung, Lehre und Nachwuchsförderung zu unterstützen oder erst zu ermöglichen. Dieser aus den technischen Disziplinen stammende Ansatz hat sich innerhalb der letzten zehn bis fünfzehn Jahre auch im Bereich der Geistes- und Sozialwissenschaften etabliert. Hier bieten sich gerade geisteswissenschaftliche Grundlagenwerke wie Editionen und Wörterbücher, die einerseits Forschungsgegenstand, andererseits aber auch Forschungsinstrument sind, als Basis entsprechender Infrastrukturen an. Im Beitrag wird mit dem vom BMBF geförderten Zentrum für Historische Lexikographie eine dieser Initiativen und die dort erzielten Ergebnisse vorgestellt.

Forschungsinfrastrukturen stellen Instrumente, Ressourcen und Dienstleistungen zur Verfügung, »die speziell für wissenschaftliche Zwecke errichtet, mittelfristig bis tendenziell permanent bereitgestellt werden und für deren sachgerechte Errichtung, Betrieb und Nutzung in der Regel spezifische fachwissenschaftliche oder interdisziplinäre (Methoden-)Kompetenzen erforderlich sind. Ihre Funktion ist es, Forschung, Lehre und Nachwuchsförderung zu ermöglichen oder zu erleichtern. Sie sind örtlich fixiert, auf mehrere Standorte verteilt oder werden ohne definierte physische

Anlaufstelle ausschließlich virtuell bereitgestellt. Sie werden nicht ausschließlich von einzelnen Personen oder Gruppen genutzt, sondern stehen prinzipiell einer internationalen Fachgemeinschaft oder mehreren Fachgemeinschaften offen.« (WR 2017) Diese Charakterisierung aus dem ›Bericht zur wissenschaftsgeleiteten Bewertung umfangreicher Forschungsinfrastrukturvorhaben für die Nationale Roadmap‹ stellt eine davor eher auf die technischen Disziplinen gewendete Definition auf eine breitere Basis und bezieht mit Hinblick auf nachfolgende Förderprogramme insbesondere die geistes- und sozialwissenschaftlichen Fächer ein.

Gerade geisteswissenschaftliche Grundlagenwerke wie Editionen und Wörterbücher, die einerseits Forschungsgegenstand, andererseits aber auch Forschungsinstrument sind, bieten sich als Basis zur Konzeption und Entwicklung entsprechender Infrastrukturen an. Wörterbücher und lexikologische Darstellungen zur Entwicklung des Wortschatzes und des Wortgebrauchs sind zentrale wissenschaftliche Dokumentationsformen im Schnittbereich zwischen Philologie, Sprach-, Literatur- und Kulturwissenschaft sowie den Forschungsdisziplinen, die auf die sprachliche Beschreibung einzelner Gegenstandsbereiche, ihrer geschichtlichen Entwicklung und ihrer Erforschung bezogen sind. Beim Übergang ins digitale Zeitalter waren vor allem im Bereich der historischen Lexikographie und Lexikologie zwei zeitlich versetzte Strategien zu beobachten. Zunächst wurden die Informationsbestände der vorhandenen Wörterbücher retrodigitalisiert (entweder als Volltext oder in Form reiner Imagedigitalisate), die digitalen Versionen waren zum Teil direkte Gegenstücke zu den gedruckten Fassungen. Eine zweite Strategie bestand und besteht darin, vorhandene und hinzukommende Daten mit neuen Darstellungs-, Interaktions- und Vernetzungsmöglichkeiten anzureichern, die erst durch digitale Techniken und Infrastrukturen ermöglicht werden. Diese Initiativen stießen jedoch zunehmend an die Grenzen einer disparaten Verteilung der Arbeiten zur historischen Wortforschung, die mit jeweils eigenem rigidem Zeitregime und fixen Arbeitsplänen durchgeführt werden, dabei aber

informationstechnisch kaum in nennenswerter Weise aufeinander abgestimmt sind. Besonders bemerkbar machte sich dieses Problem im Fehlen integrativer und interoperabler eHumanities-Ansätze. Eine Konsequenz bestand daher in der Entwicklung von Initiativen, die nationalen und internationalen Bestrebungen im Bereich der Lexikographie zu integrieren, zu erweitern und zu harmonisieren, mit dem Ziel, eine nachhaltige Infrastruktur zu schaffen, die einerseits einen effizienten Zugang zu hochwertigen lexikalischen Daten im digitalen Zeitalter ermöglicht und andererseits die Kluft zwischen fortgeschritteneren und weniger gut ausgestatteten wissenschaftlichen Gemeinschaften, die an sprachwissenschaftlichen Ressourcen arbeiten, überbrücken hilft (vgl. Burch [u. a.] 2020).

Der Bedarf an lexikographisch ausgerichteten Infrastrukturen hatte sich bereits als ein Ergebnis aus der Gemeinschaft im Rahmen der COST-Action European Network of e-Lexicography ENeL ergeben, die 2017 beendet wurde (vgl. Declerck [u. a.] 2015). Aus dieser eher institutionell vernetzenden Initiative sind Nachfolgevorhaben auf nationaler und europäischer Ebene wie das vom Bundesministerium für Bildung und Forschung (BMBF) 2017 eingerichtete eHumanities-Zentrum für Historische Lexikographie ZHistLex und die im Rahmen des Horizon2020-Programms von der EU geförderte European Lexicographic Infrastructure ELEXIS entstanden. Beide Initiativen verfolgen im Wesentlichen gleiche Ziele, die sich auf die Definition und Bereitstellung gemeinsamer Interoperabilitätsstandards, Arbeitsabläufe, konzeptioneller Modelle und Datendienste sowie Schulungs- und Ausbildungsaktivitäten mit einem Schwerpunkt im Bereich lexikographischer Nutzungsszenarien sowie der disziplinübergreifenden Anwendung beziehen.

Beide Initiativen werden von Konsortien an verteilten Standorten getragen, die sich aus Institutionen und Forschern mit komplementärem Hintergrund – Lexikographie, digitale Geisteswissenschaften, Informatik – zusammensetzen, die neben der jeweiligen Expertise auch lexikographische Ressourcen einbringen. Sowohl für die nationale wie die interna-

tionale Landschaft der digitalen Lexikographie gilt, dass sie sich recht heterogen darstellt. Sie ist gekennzeichnet durch eigenständige Datenbestände, die in der Regel in inkompatiblen Strukturen kodiert sind, da die Arbeiten oft isoliert durchgeführt werden. Dies verhindert die Wiederverwendung dieser wertvollen Daten in anderen Bereichen, wie beispielsweise der Verarbeitung natürlicher Sprache, der Verknüpfung offener Daten, die Einbindung ins Semantic Web sowie ihre breite Nutzung im Kontext der digitalen Geisteswissenschaften.

Vor diesem Hintergrund werden Strategien, Werkzeuge und Standards für die Extraktion, Strukturierung und Verknüpfung lexikographischer Ressourcen entwickelt, um deren volles Potenzial hinsichtlich Linked Open Data (LOD) und Semantic Web sowie im allgemeinen Kontext der digitalen Geisteswissenschaften zu erschließen. Gleichzeitig sollen Wissenschaftler dabei unterstützt werden, homogene Datenformate über nationale Grenzen hinweg zu erstellen, zu teilen, zu verknüpfen, zu analysieren und zu interpretieren, was den Weg für transnationale, datengesteuerte Fortschritte in diesem Bereich ebnet und gleichzeitig Doppelarbeit über disziplinäre Grenzen hinweg deutlich reduzieren hilft. Im Folgenden werden die Arbeiten und Ergebnisse des Zentrums für Historische Lexikographie vorgestellt, die sich in ähnlicher Form mit dem erweiterten Blick auf internationale lexikographische Strukturen auch in ELEXIS zeigen (Krek [u. a.] 2018).

Gerade für die historische Lexikographie des Deutschen gilt, dass sie sich in besonderer Weise als Ausgangspunkt für ein virtuelles Zentrum eignet, das auf breiter Grundlage neue Wege digitaler Wörterbuchproduktion und -präsentation ebnen sowie ihre Nutzung in unterschiedlichen Feldern wortgeschichtsbezogener Forschung ermöglichen kann. Die historische Lexikographie des Deutschen ist einerseits gekennzeichnet durch eine reich entfaltete und differenzierte Wörterbuchlandschaft, andererseits durch avancierte Initiativen im Sinne der eHumanities, der Computer- und der Internetlexikographie. ZHistLex stellte sich die Aufgabe, solche Ansätze zu entwickeln bzw. auszubauen, auch über ihre An-

wendung für das Deutsche hinaus. Die Ergebnisse dieser Arbeiten sind eine wesentliche Voraussetzung und ein Bezugspunkt für die zukünftige Forschung und die zentrale Ergebnisdokumentation in vielen Bereichen wortgeschichtlicher Forschung. Sie sind damit auch ein Stimulans für die Erarbeitung neuer lexikographischer Materialien und für den Aufbau und die Erschließung einschlägiger digitaler Quellentexte. Das Zentrum verstand sich als Kristallisationspunkt für abgeschlossene, laufende und zukünftige Vorhaben in der historischen Lexikographie sowie für empirische Forschungen in der Sprachgeschichte, insbesondere der historischen Lexikologie und anderer wortgeschichtlicher Arbeitsfelder.

Die Ziele von ZHistLex lassen sich in drei Dimensionen charakterisieren, indem Beiträge zu einer integrierten Dateninfrastruktur im Bereich der digitalen historischen Lexikographie geleistet werden, eine konzeptionelle Systematisierung und Erweiterung von digital unterstützten Untersuchungs-, Erschließungs- und Präsentationsverfahren erstellt sowie die Einrichtung einer Kollaborationsstruktur angestrebt wird. Die Umsetzung der Ziele spiegelt sich in den auf drei Ebenen angesiedelten Arbeitsgebieten wider, die getrieben von den zugrundeliegenden lexikographischen Daten deren Codierung, den Zugriff darauf sowie geeignete Nutzungsszenarien betrachten und damit eine zunehmende Abstraktion von den jeweiligen digitalen Repräsentationen der Wörterbücher erlauben und zu einer interoperablen, wörterbuch-übergreifenden Verwendung führen (vgl. Burch [u. a.] 2020).

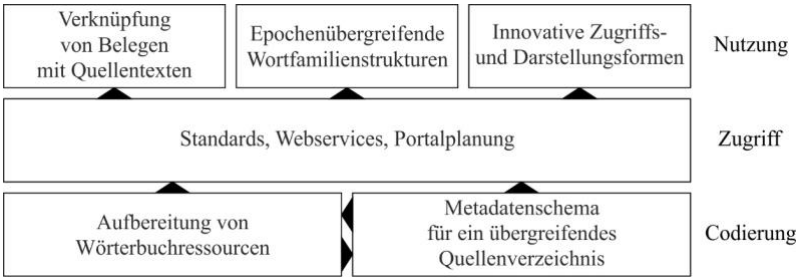


Abb. 1: Ebenen der Datenbearbeitung in ZHistLex

So werden auf unterster Ebene Basistechnologien für die Harmonisierung von Ressourcen hinsichtlich der gemeinsamen Anwendung von Standards für die Zusammenführung von Wörterbüchern durch interne und externe Vernetzung, für die Zusammenführung bislang unabhängig geführter Belegquellenverzeichnisse sowie für die Verknüpfung von Belegen mit verfügbaren Volltexten geschaffen. Diese Techniken bilden das Rückgrat des Zentrums und sind Leitlinie für laufende lexikographische Vorhaben sowie für kommende Vorhaben, seien sie mit den aktuellen Vorhaben verbunden oder nicht. Der Ausgangspunkt der technisch-informatischen Entwicklungen war hier gegeben durch eine beträchtliche Heterogenität der Ansätze und der technischen Lösungen innerhalb der beteiligten Einrichtungen. Während sich eine XML-Kodierung im Rahmen des TEI-P5-Standards mittlerweile durchgesetzt hat, so erlaubt dieser Standard doch eine Fülle von Anwendungsszenarien, die eine interoperable Adressierung von Wörterbuchdaten erschwert. Die Anwendung der TEI-Guidelines zur Codierung der in ZHistLex betrachteten Wörterbücher zeichnete sich durch eine hohe Variabilität hinsichtlich Textperspektiven, Annotationsebenen und Annotationstiefen aus. Eine grundsätzliche Vereinheitlichung der Codierung folgte hier den Vorschlägen der internationalen Initiative TEI Lex-0 (Romary/Tasovac 2019), die ein sogenanntes Baseline-Encoding für Wörterbuchdaten erarbeitet hat. Darüber lassen sich die Einheiten in der Makro- und Mesostrukturebene von Wörterbüchern einheitlich codie-

ren. Die teilweise sehr spezifischen Elemente der Mikrostrukturebene werden nach den allgemeinen TEI-Guidelines abgebildet. Eine entsprechende Vereinheitlichung erfolgt hier über die Spezifikation einer gemeinsamen offenen Schnittstelle.

Die zweite strukturelle Ebene wird durch die Entwicklung von standardisierten Kommunikationsmechanismen abgedeckt, die die Interoperabilität von digitalen Wörterbuchangeboten überhaupt erst ermöglichen bzw. vorhandene Zugriffsformen verbessern. Geleitet wurde die Implementierung dieser offenen Schnittstellen von den Zielen, über lexikographische Webservices eine übergreifende Recherche in und den Zugriff auf die Ressourcen (Wörterbücher, Quellenverzeichnisse, Textcorpora) zu unterstützen. Die Implementierung selbst setzte auf verbreiteten Entwurfsmustern für derartige APIs (Application Programmable Interfaces) auf, indem als Vorlage und Best-Practice das Konzept einer REST (Representational State Transfer)/SOAP (Simple Object Access Protocol)-Schnittstelle (W3C 2000; Open API 2021) zugrunde gelegt und eine gemeinsame Schnittmenge an Funktionen und Parametern definiert wurde. Über die Syntax der Schnittstellen können beispielsweise Abfrageszenarien wie ›Gib alle Wörterbuchartikel zurück, die im lexikographischen Kommentar die beschreibungssprachliche Wortform Haus enthalten‹, ›Gib alle Wörterbuchartikel zurück, deren Stichwörter länger als 200 Jahre belegt sind‹ oder ›Welche Artikel des althochdeutschen Wörterbuchs führen etymologische Belege zu mittelhochdeutschen Lemmata‹ formuliert werden können.

Entscheidend für eine generische Umsetzung und damit die Übertrag- und Erweiterbarkeit der Schnittstelle ist eine formale Spezifikation der Abfrageszenarien und deren Überführung in eine einheitliche Syntax, die hier standardisiert über eine YAML-Beschreibung (YAML 2021) erfolgte und damit einerseits direkt überprüfbar und andererseits auch entsprechend modular anpassbar ist. Die Schnittstelle erlaubt nur einen lesenden Zugriff auf die Daten (sogenannte GET-Requests), da die Wörterbuchinhalte nicht von den Nutzer*innen geändert werden sollen. Für jede Ressource wird ein

spezifischer API-Endpoint nach dem RESTful-Paradigma implementiert, wodurch die einzelnen APIs sehr einfach gehalten werden können, weil beispielsweise auf Funktionen zur Wörterbuchauswahl an den Endpoints verzichtet werden kann. Auch für das Rückgabeformat des Webservice wird mit JSON (JavaScript Object Notation) ein standardisiertes Austauschformat eingesetzt, über welches Daten in strukturierten Formen zur maschinellen Weiterverarbeitung bereitgestellt werden können.

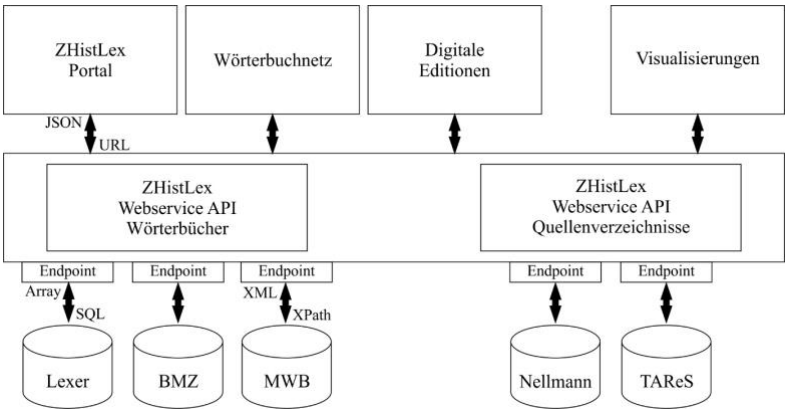


Abb. 2: vereinheitlichte Sicht auf die Daten durch ZHistLex-Webservices

Im Rahmen von ZHistLex erfolgte eine Implementierung der Schnittstelle für den Zugriff auf das neue Mittelhochdeutsche Wörterbuch (MWB) und dessen Quellenverzeichnis (der Endpoint der Service-Implementationen ist http://tares.uni-trier.de/ZHistLex/API/). Die implementierten Services werden produktiv in der Wörterbucharbeit genutzt, einerseits durch das Redaktionssystem TAReS (TriererArtikelRedaktionsSystem), indem es Informationen verschiedener Ressourcen dynamisch bündelt, sowie andererseits für die Vernetzung des MWB mit dem Wörterbuchnetz. Die Implementierung der Webservices seitens des Trierer Wörterbuchnetzes erfolgt über den Endpoint api.woerterbuchnetz.de/open-api/. Hierüber können die über die Schnittstelle erreichbaren Ressourcen sowie die für die einzel-

nen Wörterbücher implementierten Zugriffsmechanismen abgefragt werden. So liefert beispielsweise die Anfrage https://api.woerterbuchnetz.de/open-api/dictionaries/Lexer als Ergebnis die Struktur


```
{"result_type": "method_list",
 "result_count": 3,
 "query": "/open-api/dictionaries/Lexer",
 "result_set": [{
   "methodid": "fulltext",
  "comment": "Gesamter Text",
  "path": "/open-
api/dictionaries/Lexer/fulltext/:searchpattern"},
 {"methodid": "lemmata",
  "comment": "Stichwort",
  "path": "/open-
api/dictionaries/Lexer/lemmata/:searchpattern"},
 {"methodid": "definition",
  "comment": "Definitionen",
  "path":"/open-
api/dictionaries/Lexer/definition/:searchpattern"
 }
]}
```


und beschreibt damit, dass für das Mittelhochdeutsche Handwörterbuch von Matthias Lexer neben einer üblichen Volltextsuche (`"methodid": "fulltext"`) auch eine Suche über die Artikelstichwörter (`"methodid": "lemmata"`) und die Bedeutungserläuterungen (`"methodid": "definition"`) möglich ist. Somit ist eine Anfrage der Form

https://api.woerterbuchnetz.de/open-api/dictionaries/Lexer/definition/haus

korrekt im Sinne der Spezifikation und liefert 64 Lexer-Artikel als Ergebnis.

Die zentrale technische Idee in ZHistLex war es, die einzelnen beteiligten Ressourcen des Zentrums (Wörterbücher und Quellentexte) über ein standardisiertes System von Webservices interoperabel zu machen. Die Entwicklung einer übergreifenden Oberfläche (GUI, Graphical User Interface) für menschliche Nutzung war dabei von nachgeordneter Bedeutung.

Das Potenzial der Infrastruktur zeigte sich neben dem Einsatz im MWB und im Wörterbuchnetz aber auch anhand prototypisch implementierter Abfrage- und Visualisierungsszenarien. Ein einfacher Demonstrator zur Quellenbibliographie führt eine gleichzeitige Anfrage an alle implementierten Services durch, bereitet das Ergebnis auf, visualisiert die Überschneidung der Quellennutzung in den beteiligten Wörterbüchern und illustriert damit exemplarisch die Möglichkeit, die Interaktion mit den ZHistLex-Webservices über Suchinterfaces auch menschlichen Nutzern zu erschließen. Er verwirklicht bewusst nur ein Basis-Konzept und realisiert nicht alle spezifizierten Suchoptionen, verdeutlicht aber, dass die gewählte Architektur geeignet ist, die Ressourcen der beteiligten Standorte interoperabel zu machen. Da alle Services ohne Zugangsbeschränkungen verfügbar sind, können von beliebiger Seite weitere Anwendungen auf dieser Basis entwickelt werden, wie beispielsweise eine föderierte Suche über eine Menge von unterschiedlichen Wörterbüchern mit anschließenden komplexen Aggregationen, Synopsen und Darstellungen der Ergebnismengen.

Der Forschung werden durch diese Zugriffsformen auf Wörterbücher mit ihrer sorgfältigen, je spezifischen lexikographischen Auswahl und Bearbeitung sowie auf Belegquellen mit einem reichhaltigen Kontext, den die lexikographischen Belege allein nicht bieten können, neue Möglichkeiten eröffnet. Ebenso wird die Begrenzung einer Suchanfrage auf den Dokumentationszeitraum eines einzelnen Wörterbuchs überwunden. Eine Vielzahl von Anforderungen, die die Forschung ebenso wie das gebildete Laienpublikum an die Suche und Präsentation von sprachhistorischen Daten hat, kann auf diese Weise besser und umfassender bedient werden (vgl. Burch [u. a.] 2020).

Die in ZHistLex und ELEXIS gewonnenen Erfahrungen fließen ein in die Arbeiten des Konsortiums Text+ im Rahmen der Nationalen Forschungsdateninfrastruktur (NFDI), die damit in entscheidendem Maße zu einer langfristigen Nachhaltigkeit der erzielten Ergebnisse beiträgt und sie insbesondere weiterentwickelt. Im Bereich der Datendomäne Lexical Re-

sources (neben den Domänen Text Collections und Editions) werden hier die spezifisch für die historische Lexikographie erarbeiteten Konzepte auf weitere Arten von lexikalischen Ressourcen wie Wörterbücher zu verschiedenen Sprachen, verschiedenen regionalen Varietäten sowie auf Enzyklopädien, aber auch auf maschinell lesbare Wörterbücher, terminologische Datenbanken, Ontologien, Wortlisten, Wortkarten und linguistische Atlanten ausgeweitet. Daher beschäftigt sich das Konsortium auch mit der Frage, wie wissenschaftliche Kommunikation in und durch ihre Forschungsdaten nachnutzbar, verfügbar und geordnet werden kann. Die hierbei entstehenden Werkzeuge und Kompetenzen sind für Forschungen aus allen Bereichen von Wissenschaft relevant, die ihre Erkenntnisse oder Prozesse in Sprach- und Textformen vermitteln.

Literaturverzeichnis

Sekundärliteratur

Burch, Thomas/Gloning, Thomas/Harm, Volker/Herold, Axel/Hoenen, Armin/Plate, Ralf/Recker-Hamm, Ute: Schlussbericht des Verbund-Projekts ZHistLex (»eHumanities-Zentrum für Historische Lexikographie«). Gießen 2020 (online).

Declerck, Thierry/Wandl-Vogt, Eveline/Mörth, Karlheinz: Towards a Pan European Lexicography by Means of Linked (Open) Data, in: Kosem, I[ztok]/Jakubíček, M[iloš]/Kallas, J[elena]/Krek, S[imon] (Hrsg.): Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Ljubljana/Brighton 2015, S. 342–355 (online).

Krek, Simon/Kosem, Iztok/McCrae, John P./Navigli, Roberto/Pedersen, Bolette S./Tiberius, Carole/Wissik, Tanja: European Lexicographic Infrastructure (ELEXIS), in: Čibej, Jaka/Gorjanc, Vojko/ Kosem, Iztok/ Krek, Simon (Hrsg.): Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts. Ljubljana 2018, S. 881–892 (online).

Romary, Laurent/Tasovac, Toma: TEI Lex-0 — A baseline encoding for lexicographic data. (online).

WR 2017: Wissenschaftsrat: Bericht zur wissenschaftsgeleiteten Bewertung umfangreicher Forschungsinfrastrukturvorhaben für die Nationale Roadmap. (Drs. 6410-17), Juli 2017 (online).

Online-Ressourcen

COST (European Cooperation in Science and Technology): https://www.cost.eu/.
ENeL (European Network of e-Lexicography): https://www.elexicography.eu/.
JSON (JavaScript Object Notation): https://www.json.org/json-en.html.
MWB (Mittelhochdeutsches Wörterbuch): www.mhdwb-online.de/.
NFDI (Nationale Forschungsdateninfrastruktur): https://www.nfdi.de/.
OpenAPI (OpenAPI Specification v3.1.0): https://spec.openapis.org/oas/latest.html.
SOAP (Simple Object Access Protocol): http://www.w3.org/TR/2000/NOTE-SOAP-20000508.
TAReS: http://www.mhdwb.uni-trier.de/TAReS/index.html.
Text+: http://www.text-plus.org/.
Wörterbuchnetz: www.woerterbuchnetz.de.
YAML (YAML Ain't Markup Language): https://yaml.org/spec/1.2.2/.
ZHistLex (Zentrum für Historische Lexikographie): www.zhistlex.de.

Anschrift des Autors:

Dr. Thomas Burch
Universität Trier
Kompetenzzentrum – Trier Center for Digital Humanities
Universitätsring 15
54286 Trier
E-Mail: burch@uni-trier.de

Albrecht Hausmann

Digitale Infrastruktur und Forschungsdatenmanagement (Diskussionsbericht Sektion 3)

Einen der Schwerpunkte der Diskussion (Leitung: Stephan Müller) bildete die Frage nach dem Verhältnis zwischen Einzelprojekten, in denen laufend Daten generiert werden, und der jetzt im Aufbau befindlichen nationalen Dateninfrastruktur (etwa durch das NFDI-Konsortium Text+). Für die Nachhaltigkeit von Einzelprojekten sei die Möglichkeit essentiell, die generierten Daten dauerhaft in einer institutionalisierten Infrastruktur abzulegen (Jürgen Wolf). Problematisch erschien (und erscheint) die nach wie vor nur mittelfristig gesicherte Finanzierung dieser Infrastruktur; erforderlich sei, so Wolf, stattdessen eine dauerhafte und vom Staat finanzierte ›Datennationalbibliothek‹, die ähnlich gut mit Ressourcen ausgestattet sein sollte wie die vorhandene analoge Bibliotheksinfrastruktur. Andrea Rapp wies demgegenüber darauf hin, dass der Grundgedanke hinter dem Projekt einer nationalen Dateninfrastruktur allerdings gerade dezentral sei. Entscheidend sei jedenfalls, so Rapp, dass sich die betroffenen Wissenschaftlerinnen und Wissenschaftler immer wieder zu Wort meldeten und dieses wichtige Interesse an den entsprechenden Stellen sichtbar machten. Thematisiert wurde auch der internationale Aspekt; zum einen könnten bereits vorhandene Strukturen, z. B. in den Niederlanden, als Modell dienen (Michael Stolz), zum anderen stelle sich auch die Frage, wie z. B. die Schweiz und Österreich in das NFDI-Konzept einzubeziehen seien. Aus österreichischer Perspektive könne das in Deutschland bereits Erreichte als vorbildlich bezeichnet werden (Katharina Zeppezauer-Wachauer).

Von mehreren Diskutanten wurde das Problem der Nutzerfreundlichkeit der zentralen Dateninfrastruktur thematisiert: Wie könnten Frustrationserlebnisse bei der Arbeit mit den entsprechenden Angeboten im Netz vermieden werden – insbesondere auf Seiten derer, die nicht IT-affin sind (Albrecht Hausmann)? Was konkret biete z. B. Text+ einem digitalen Projekt wie etwa ›Lyrik des deutschen Mittelalters‹ (Manuel Braun)? Welche Arbeiten nehme Text+ den Wissenschaftlerinnen und Wissenschaftlern ab (Elisabeth Lienert)? Für Rapp war die Vorstellung, dass Text+ am Ende eine Art Server darstellen könnte, auf dem man seine Daten einfach abgeben könne, illusorisch; vielmehr stelle auch das Forschungsdatenmanagement einen sich entwickelnden Prozess dar, der keineswegs schon abgeschlossen sei, sondern weiterhin z. B. durch die Analyse von User-Stories verbessert werden müsse. In diesem Zusammenhang merkte Klaus Kipf an, dass innerhalb der entsprechenden Infrastrukturprojekte dauerhaft Stellen für Service geschaffen werden sollten, die die Verbindung zwischen Wissenschaft und Datenmanagement herstellten. Dies sei, so Kipf, »vielleicht wichtiger als neue Digital-Humanities-Professuren«. Rapp wies aber auch darauf hin, dass auch von Seiten der Wissenschaftlerinnen und Wissenschaftler, die entsprechende Dienste in Anspruch nähmen, gewisse Kompetenzen eingebracht werden müssten: »Wenn Sie Auto fahren wollen, müssen Sie zwar keinen Motor reparieren können, aber Sie sollten eben doch einen Führerschein haben.«

Einen grundsätzlichen Aspekt thematisierte Stephan Müller: Die operationale Umsetzung des Forschungsdatenmanagements werde z. B. aufgrund von Standardisierungsanforderungen immer mit einer gewissen Komplexitätsreduktion einhergehen, die den bereitstellenden Wissenschaftlerinnen und Wissenschaftlern problematisch erscheinen könne; entscheidend sei hier, dass man sich dieser Problematik ständig bewusst sei; jedenfalls dürften Digitalisierung und Standardisierung nicht einschränkend auf Forschungsprozesse zurückwirken.

Anschrift des Berichterstatters:

Prof. Dr. Albrecht Hausmann
Carl von Ossietzky Universität Oldenburg
Institut für Germanistik
26111 Oldenburg
E-Mail: albrecht.hausmann@uni-oldenburg.de

Sektion 4: Repositorien und Datenbanken

Jürgen Wolf

Handschriftencensus (HSC)

Von der Handschrift zu den Metadaten

Abstract. Das Akademieprojekt Handschriftencensus (HSC) erfasst die gesamte deutschsprachige Handschriftenüberlieferung des Mittelalters (~750-1520) in einer Online-Datenbank. In dieser Eigenschaft fungieren der HSC bzw. das HSC-Team zugleich als Kompetenzzentrum zur deutschsprachigen Textüberlieferung inklusive der Handschriften-, Werk- und Autoridentifikation. Zudem werden im HSC die entsprechenden Digitalisate, Editionen und Kataloge erfasst und, wo möglich, verlinkt. Mit der neu ins Leben gerufenen Online-Zeitschrift ›Maniculae‹ besteht ergänzend die Möglichkeit, neue Funde und/oder Informationen zu Handschriften und Fragmenten schnell und unbürokratisch zu publizieren.

Im folgenden Beitrag werden in einer ersten Rubrik Daten, Datenmodelle und Features des Handschriftencensus (HSC)[1] vorgestellt, in einer zweiten Rubrik geht es um Nutzungs- und Innovationsszenarien – mit einem Schwerpunkt auf ›Normdaten‹

1. Basisinformationen

Die Idee des Handschriftencensus ist es, die gesamte deutschsprachige Handschriften- und Fragmentüberlieferung des Mittelalters zu erfassen und frei online zugänglich zu machen und zugleich ein vollständiges Repertorium der deutschsprachigen Autoren und Werke des Mittelalters bereitzustellen. Die Weiternutzung der Daten erfolgt unter der Creative Com-

mons Lizenz CC BY-SA 3.0 DE. Die Startseite des Online-Portals fasst die wesentlichen Inhalte und historischen Entwicklungen kurz zusammen:

> Der Handschriftencensus ist eine Online-Datenbank zu sämtlichen deutschsprachigen Handschriften des Mittelalters (750–1520) weltweit. Er vereint basale Informationen zu Autoren, Werken und ihrer Überlieferung. Darüber hinaus bietet er zu jedem Textzeugen eine überlieferungsgeschichtlich einschlägige Literaturauswahl und den Zugang zu Digitalisaten. Der Handschriftencensus versteht sich als zentrale Anlaufstelle zum Verzeichnen von Handschriften in ihren vielfältigen Ausprägungen, ihrer Datierung, Provenienz und ihrer inhaltlichen Ausrichtung, er ist außerdem ein professionelles Instrument für die wissenschaftliche Erforschung deutschsprachiger Schriftzeugnisse der Vergangenheit. Neuigkeiten vom Handschriftencensus finden Sie über den Kurznachrichtendienst Twitter. Seit 2017 ist der Handschriftencensus ein Vorhaben der Akademie der Wissenschaften und der Literatur Mainz. Gefördert von Bund und Ländern im Rahmen des Akademienprogramms der Union der Deutschen Akademien der Wissenschaften.

Technische Basis[2] ist eine historisch gewachsene und auf MySQL basierende relationale Datenverwaltung. Für die Online-Präsenz werden PHP (Hypertext Preprocessor) und das CakePHP-Framework genutzt. Für die Datennachnutzung wird ein Export via JSON (JavaScript Object Notation) angeboten, weitere Schnittstellen und Exportformate sind in Vorbereitung. Vor allem hinsichtlich einer angedachten – und vielfach nachgefragten – Portalfunktion des Handschriftencensus werden Schnittstellen und Austauschformate weiterentwickelt, aber auch verteilte Lösungen mit dezentral organisierten Netzwerken angedacht. Erste Schritte in diese Richtung wurden mit der Datenintegration der GND (Gemeinsame Normdatei) in den HSC erprobt (s. u.).

1.1 Zahlen, Daten, Fakten

In die Datenbank des Handschriftencensus eingeflossen sind neben der Arbeit der ›Arbeitsgemeinschaft Handschriftencensus‹[3] die Erträge diver-

ser Drittmittel-Projekte. Zu nennen wären einerseits die DFG-geförderten Marburger Repertorien deutschsprachiger Handschriften des 13. und 14. Jahrhunderts und andererseits das Thyssen-geförderte Paderborner Repertorium der deutschsprachigen Textüberlieferung des 8. bis 12. Jahrhunderts. Unmittelbar ergänzend kamen u. a. Erträge aus dem Freidank-Repertorium und dem Marburger Repertorium zur Übersetzungsliteratur im deutschen Frühhumanismus hinzu. Ausgewertet sind zudem Datenbanken, Online-Ressourcen, Handschriftenkataloge, Editionen und Forschungsliteratur aller Art. Das Gesamtspektrum ist über das Literaturverzeichnis des Handschriftencensus und die Katalogisate selbst erschließbar. Zur Zeit sind dies über 20.000 Titel und/oder Datenbanken.

Mit Stand 04/2022 bietet der Handschriftencensus in den Rubriken Katalogisate, Werke und Autoren und Literatur als Kerngerüst Katalogisate zu ca. 26.000 Textzeugen (Handschriften und Fragmente) mit knapp 30.000 Signaturen in 25.000 Beschreibungen mit HSC-IDs.[4] Von diesen Textzeugen sind ca. 1.500 Einheiten verbrannt, verschollen oder firmieren unter »Verbleib unbekannt«, d. h. sie sind ausschließlich durch ältere Beschreibungen, Hinweise oder im Idealfall Fotografien sekundär dokumentiert, aber nicht physisch greifbar. 11.000 Beschreibungen sind mit einem Nachweis von Abbildung versehen, wovon wiederum gut 8.500 Beschreibungen Links zu einer oder mehreren Online-Abbildungen enthalten – vorzugsweise zu Volldigitalisaten. Der Handschriftencensus ist damit das weltweit umfangreichste Nachweisinstrument für digitalisierte deutsche Handschriften.[5]

Die Textzeugen verteilen sich auf 36 Länder.[6] Erfasst sind weltweit ca. 800 Bibliotheksorte mit 1.600 Bibliotheken, Archiven oder anderen Sammlungsinstitutionen sowie ein halbes Tausend Privatbesitzer (https://handschriftencensus.de/hss/Privat), wobei das Gros dieser Privatbesitzer auf historische Besitzzustände rekurriert. Häufig sind die entsprechenden Stücke bzw. Privatsammlungen im Laufe des 20. und 21. Jahrhunderts in

öffentliche Bibliotheken übergegangen, was in den Katalogisatköpfen nachgewiesen wird:

Abb. 1: Nachweis der Besitzzustände (https://handschriftencensus.de/14903)

Mit ca. 6.800 Werken sind im HSC noch deutlich mehr mittelalterliche deutsche Werke angelegt als im Verfasserlexikon bzw. in der Verfasserdatenbank. Rund 600 dieser mittelalterlichen Werke sind jedoch offline, weil sie nur in neuzeitlichen Abschriften und/oder Drucken tradiert oder überhaupt nur indirekt bezeugt sind. Rund ein Viertel der abrufbaren Werke ist mit Normdatenstatus angelegt und mit der Gemeinsamen Normdatei (GND[7]) verlinkt. Hinzu kommen knapp 1.800 mittelalterliche Autoren, die nahezu komplett normiert und GND-verlinkt sind.

Zu den Autoren und Werken sind fast 3.000 Ausgaben bzw. Editionen in den entsprechenden Übersichten nachgewiesen. Eine noch weitaus größere Zahl von (vorzugsweise) Transkriptionen bzw. Abdrucken ist im Literaturverzeichnis der jeweiligen Katalogisate genannt. Erhebliche Nach-

weis-, aber auch generelle Editionsdefizite gibt es bei der geistlichen Prosaliteratur und der pragmatischen Literatur.

Eine aktuell noch nicht freigeschaltete Option wird es in Zukunft den Nutzern erlauben, all diese Autoren und Werke nach Textsorten zu kategorisieren. Wie sich die einzelnen Textsorten verteilen, sei prototypisch in einem Vorabeinblick in den internen Bereich der Datenbank skizziert:

Epik	1.627
Geistliche Literatur	2.751
Geschichtsschreibung	547
Rechtsliteratur	321
Lyrik	466
Fachliteratur	1.160
Drama (Spiel etc.)	187

(HSC-interne Erhebung)

Die verwendeten ›Textsortencluster‹ lehnen sich an die Vorgaben der GND an, d. h. sie sind an der modernen Literatur/Literaturwissenschaft und der gedruckten Literatur orientiert und damit für das Mittelalter mit den Besonderheiten der handschriftlichen Tradierung nur bedingt aussagekräftig. In diesem hochkomplexen Feld wird aktuell in enger Zusammenarbeit mit der DNB (Deutsche Nationalbibliothek, GND) und der Bayerischen Staatsbibliothek in München und in Absprache mit dem Handschriftenportal an mittelaltergerechteren Textsortenzuordnungen, Gattungsclustern sowie einer ›Erfassungshilfe mittelalterliche Werke‹ gearbeitet.

In der begleitenden Literaturdatenbank werden alle in den Katalogisaten genannten Titel sowie Datenbanken erfasst. Rund 20.000 Titel bzw. Online-Ressourcen sind geprüft und freigeschaltet. Davon sind wiederum mehr als 5.500 Titel mit Internet-Links versehen, d. h. sie sind frei online zugänglich.

Last but not least sei auf die seit Jahren intensiv genutzte Mitteilungsfunktion zu den einzelnen Katalogisaten verwiesen. Über die unter allen Beschreibungen verlinkte Rubrik ›Mitteilung (Ergänzung/Korrektur)‹

kann jeder Nutzer ohne Einschränkung zu den Beschreibungen kommentieren, ergänzen und korrigieren.

Abb. 2: Mitteilungsfeld (https://handschriftencensus.de/11773/mitteilung#show_mitteilung_form)

Das Mitteilungsfeld ist in die Abteilungen ›Kontaktdaten‹ und ›Mitteilung‹ und die Rubriken ›Literaturhinweise‹ oder ›Allgemeine Mitteilungen‹ aufgeteilt, um den Datenstrom (aktuell weit über 20.000 archivierte Mitteilungen) zu kanalisieren.

J e d e dieser Mitteilungen wird redaktionell geprüft und erst nach diesem Prüfdurchgang werden ggf. Daten, Hinweise, Ergänzungen oder Links aus der Mitteilung in ein Katalogisat oder die Übersichten übernommen. Aus Gründen der Datenhygiene wurde auf ein zunächst angedachtes Wiki-System verzichtet, denn nur so scheint uns die Datenreinheit dauerhaft gewährleistet. Bezahlt wird dieser – intellektuelle und personelle – Aufwand mit einer ›gewissen‹ Bearbeitungsdauer. Andererseits ist gerade dieses Redaktionssystem für die weltweite Community offensichtlich so attraktiv, dass trotz Wartezeiten die Mitteilungszahlen, aber auch die Qua-

lität der Mitteilungen stetig wachsen. Von besonderer Attraktivität scheint dabei zu sein, dass Mitteiler von ›Nachrichten mit substantiellem Wert‹ in den Katalogisaten in der Rubrik ›Mitteilungen‹ namentlich aufgeführt bzw. bei grundlegenden Mitteilungen sogar in der Verfasserzeile der Katalogisate aufgenommen werden. Man kann in diesen Fällen von ›Mikropublikationen‹ sprechen.

1.2 Credo ›einfach und intuitiv‹

Die Philosophie des Handschriftencensus ist, dass jeder Nutzer, egal ob Handschriftenexperte, Fachforscher oder ›normaler‹ Mensch, alles einfach und intuitiv bedienen kann. Das Frontend des Handschriftencensus ist deshalb bewusst dem Minimalismus verpflichtet und mit dem responsiven Layout auch auf mobilen Geräten ohne Einschränkungen nutzbar.

Abb. 3: Frontend (https://handschriftencensus.de/)

Hinter dem sichtbaren Minimalismus steht eine vergleichsweise komplexe Datenbankstruktur und vor allem eine geradezu apodiktische Datenidee: Neue Daten gelangen nur nach redaktioneller Prüfung und dem Vieraugenprinzip in den HSC! Die der historischen Datenstruktur geschuldeten Freitextfelder enthalten allerdings noch nicht normierte und nicht atomisierte Daten, was bislang noch die Datenextraktion und Datennormierung erschwert.

2. Zwischen Gegenwart und Zukunft

2.1 Normdaten

In der Vernetzung von Daten, Ressourcen und Wissen (Forschung) liegt ein entscheidender Mehrwert. Für die Vernetzung sind allerdings durchaus ›schmerzliche‹ Grundvoraussetzungen unumgänglich. Dazu gehören die Atomisierung und die Normalisierung des Datenmaterials bzw. letztlich des Wissens insgesamt. Beides erwies und erweist sich für den Handschriftencensus als nicht zu unterschätzende Hürde, nicht zuletzt, weil man über einen in weiten Strecken historisch gewachsenen Datenbestand verfügt. Teile des digitalen Materials gehen auf Forschungsprojekte bis in die 1990er Jahre zurück (MR13, MR14, MR-Freidank), viele Kerndaten sogar auf gedruckte Handschriftenkataloge oder sogar handschriftliche (Berliner Handschriftenarchiv) Inventare und Verzeichnisse.

Dieses noch weitgehend analoge bzw. einer analogen Beschreibungsphilosophie verpflichtete Datenmaterial für eine moderne Datenbank aufzubereiten, birgt Herausforderungen, die letztlich nur in der engen Verbindung von Technik und menschlicher Expertise zu lösen sind: Vieles muss geprüft, normiert, atomisiert, aber vor allem qualifiziert validiert, aktualisiert und fortgeschrieben werden. Im Handschriftencensus läuft seit mehreren Jahren ein entsprechender Datenbereinigungs- und Datenumformungsprozess.

Die geschilderten Arbeitsprozesse werden durch ein vernetztes Team von Experten durchgeführt und kontrolliert. Wegweisend sind dabei die Verbindung von fachlicher bzw. handschriftenkundlicher Expertise (HSC-Mitarbeiter*innen), IT-Expertise (HSC-IT-Abteilung), bibliothekarischer Expertise (UB Marburg in Verbindung mit BSB München) und Normdatenexpertise (DNB/GND in Verbindung mit BSB München).

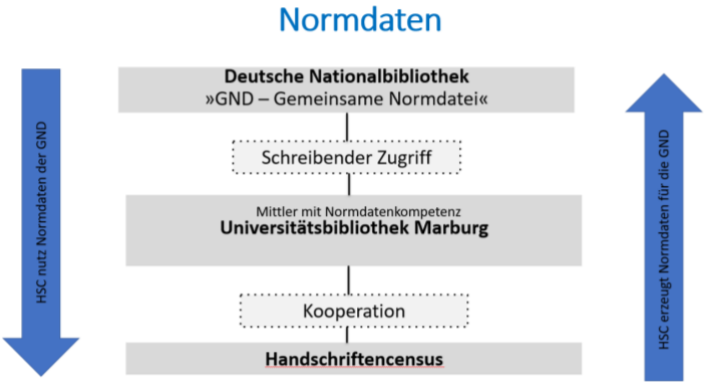


Abb. 4: Normdaten – Vernetzungsoptionen

Genau an diesem Punkt ist die Vernetzung zwischen Forschung (HSC) und Normdateninstitution (GND) allerdings noch nicht optimal austariert. Die GND ist nur bedingt auf Anforderungen aus Wissenschaft und Forschung zugeschnitten, d. h. alte Wissenstatbestände – häufig aus Lexika gewonnen – stehen nicht selten gegen aktuelle Forschungsergebnisse. Ein weiteres Problem stellt sich bei ›Massendaten‹, denn die GND ist qua Definition nur an ›Basisdaten‹ interessiert, aber nicht an ›Massendaten‹ etwa zu ›normalen‹ Menschen, Einzeltexten, ›komplizierten‹ Werkverbünden etc. Eine Lösung könnten hier Vernetzungen mit weiteren Normdatenverwaltern sein:

- ›Autorités‹ des französischen Verbundkatalogs SUDOC (Système universitaire de documentation)
- Library of Congress Authority Files (LCAF)
- Virtual International Authority File (VIAF)

- Getty Union List of Artist Names
- GeoNames
- Wikidata
- Basic Register of Thesauri, Ontologies & Classifications (BARTOC)

Oder die Etablierung völlig neuer, entgrenzter Normdatenverbünde.

Im Zentrum der Normierung innerhalb des Handschriftencensus stehen zunächst gut kalkulierbare Datenfelder zu Autoren, Werken, Institutionen und kodikologischen Basisdaten. Bei den mittelalterlichen Autoren ist ein nahezu 100%iger Normierungsgrad und bei den mittelalterlichen Werken eine Vollnormierung von ca. 25% (HSC und GND in Vollverknüpfung) und eine Teilnormierung von ca. 60–70% (Werke hier wie da angelegt, aber noch nicht verknüpft) erreicht.

Als besonders fruchtbar hat sich in diesem Kontext die enge Zusammenarbeit von HSC und GND herauskristallisiert, wobei HSC und GND wechselseitig voneinander profitieren. So hat der HSC Normdaten aus der GND verlinken und für Such- und Sortierungsfunktionen integrieren können (s. o.), andererseits wurden aus dem HSC heraus von einer in der Marburger UB verorteten Diplombibliothekarin des Projekts GND-Daten ergänzt, korrigiert und neu angelegt.

Die Normierungsarbeiten konzentrieren sich aktuell auf den Bereich der kodikologischen Daten (Überlieferungsform, Beschreibstoff, Blattgröße, Schriftraum, Spaltenzahl, Zeilenzahl), was für Suche und Sortierung, aber vor allem auch für weiterführende literar- und kulturhistorische Auswertungen neue Dimensionen eröffnen wird, denn erstmals erreichen die auszuwertenden Daten bzw. Datenmengen einen so hohen Grad an Validität, dass z. B. visualisierte ›Aussagen‹ nicht mehr beliebig sind.

Die Normierung hat darüber hinaus erhebliche Folgen für die Organisation, aber vor allem für die Nutzung der HSC-Datenbank, denn Normdaten sind ›Hilfsmittel‹ und ›Grundlage‹ für Linked Open Data, Semantic Web sowie allgemein für die Optimierung der Dateninfrastruktur. So konnten Suchoptionen durch die Verbindung von HSC- und GND-Daten

entscheidend verbessert werden: In die HSC-Suche wurden (zunächst nur für bestimmte Bereiche) die GND-Normdaten samt aller in den entsprechenden GND-Datensätzen vorgehaltenen Informationen etwa zu Alternativansetzung integriert. In der HSC-Suche kann man beispielsweise nach ›Der Nibelunge Noth‹ suchen und erhält zielgerichtete Treffer, obwohl es im HSC eigentlich keinen Werkeintrag ›Der Nibelunge Noth‹ gibt. Über die aus der GND implementierten Normdaten zum Werkeintrag ›Nibelungenlied‹ werden jedoch alle Werktitelvarianten in die HSC-Suche miteinbezogen. Entsprechende GND-Integrationen sind außer bei den Werken bei Autornamen und Aufbewahrungsorten realisiert; angestrebt ist eine flächendeckende Normierung und GND-Integration.

Schon jetzt bietet der HSC darüber hinaus diverse weiterführende Anzeigeoptionen mit

- Informationen in der GND (Gemeinsame Normdatei)
- Informationen bei LOBID (Linking Open Bibliographic Data)
- Informationen in GiN (Germanistik im Netz)
- Standort des Aufbewahrungsortes auf OpenStreetMap

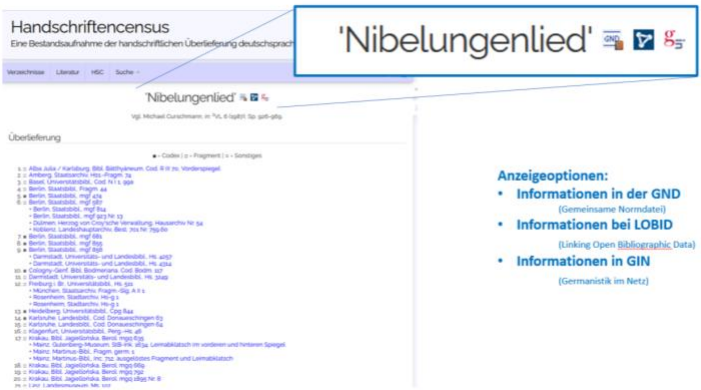


Abb. 5: Anzeigeoptionen

In der nahen Zukunft wird es über die GND-ID als Referenznummer auch möglich sein, weitere Daten aus anderen Projekten anzeigen zu können, die nicht in das Forschungsfeld des HSC fallen.

2.2 Nutzungsszenarien

Der Handschriftencensus bedient mit seinen Angeboten ein fächerübergreifendes Nutzerprofil. Angeboten werden sowohl kodikologische, bestandstechnische wie literarhistorische, sprachhistorische und kulturhistorische Basisinformationen. Gleichzeitig werden für die einzelnen Textzeugen der Forschungsstand, für die Werke die Editionslage und für die Bestandserfassung die Katalogisierung dokumentiert. Aus diesem Angebotsprofil ergeben sich zahlreiche ›traditionelle‹ literar- und kulturhistorische Nutzungsszenarien. So können einzelne Textzeugen und die komplette Werküberlieferung oder Überlieferungsverbünde in den Katalogisaten direkt per Klick oder per Suche abgefragt werden. Auch Fragen nach der Datierung und Lokalisierung werden beantwortet. Zentrale Schwerpunkte des Angebots liegen zudem auf den kodikologischen Dimensionen jeder einzelnen Handschrift, der Werkidentifikation und der Autoridentifikation.

Für die Recherche stehen dabei zwei traditionelle Zugriffswege zur Verfügung: A) Die direkte Ansteuerung via Menü und die Suche. B) Zusätzlich werden über eine mehrfach geschichtete Suche globale, aber auch zielgerichtete Suchabfragen ermöglicht (https://handschriftencensus.de/search/repertory). Kombinationsoptionen erlauben es, selbst komplexe Zusammenhänge gezielt zu recherchieren. So ist es z. B. möglich, sich alle ›Parzival‹-Handschriften eines bestimmten Zeitraums mit einer bestimmten Blattgröße, einer bestimmten Einrichtung und Ausstattung aus einem bestimmten Dialektraum ausliefern zu lassen. Ein Problem an dieser erweiterten Suche ist allerdings der noch nicht vollständig kategorisierte Datenbestand.

Der Datenexport via JSON ermöglicht auch ›großflächigere‹ Nutzungsszenarien. Exemplarisch sei die Visualisierung herausgehoben, beispielsweise um Werkcluster, personale Netzwerke, Sammlungszentren oder Manuskriptwanderungen sichtbar zu machen. Entsprechende Musterszenarien mit HSC-Material haben u. a. Gustavo Fernández Riva (Heidelberg; Riva 2019), ein Team um Mike Kestemont und Elisabeth de Bruijn (Antwerpen; Kestemont [u. a.] 2022) sowie ein Team um Andrea Rapp (Darmstadt) erprobt.

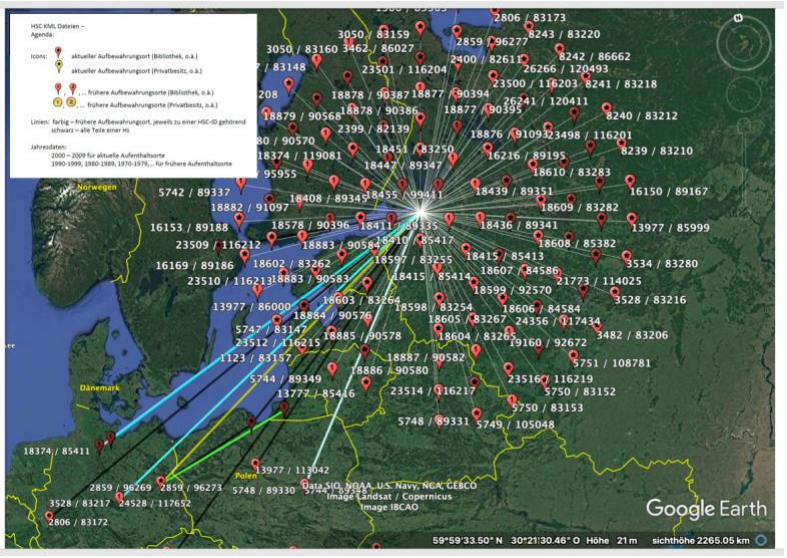


Abb. 6: Wanderungswege der heute in St. Petersburg aufbewahrten dt. Handschriften (AG Rapp, Darmstadt, 2022)

Nicht unterschlagen werden dürfen dabei allerdings die limitierenden Faktoren, denn was ›schön‹ aussieht und eine einzigartige Einblicktiefe suggeriert, steht nicht selten auf tönernen Füßen. Erinnert sei an die extrem hohen Verlustraten (von dem, was einstmals an Handschriften existierte, haben wir heute nur wenige Prozent[8]) und an die Fragmentproblematik[9],

denn ein manchmal nur wenige Quadratzentimeter großes Reststück einer einstmals vielleicht umfänglichen Sammelhandschrift verrät nichts über die in diesem Buch ehemals überlieferten Texte, über Werkverbünde, über Ausstattungs- und Einrichtungsspezifika, Miniaturen, historisierte Initialen, Schreiber und Nutzer usw.

2.3 Zukunft

Neben diesen Formen der Nutzung und Auswertung der HSC-Daten eröffnen sich via Normierung auch und gerade über die einzelne Datenbank bzw. Anwendung hinausreichende Nutzungsszenarien. So können über vernetzte Normdaten unterschiedlichste Angebote in größeren dezentralen Verbünden ›zusammengebunden‹ werden. Das ist insbesondere für Forschungs- und Erfassungsprojekte von höchster Relevanz, müssen dann doch nicht mehr in jedem Einzelprojekt alle Einzeldaten etwa zu einem Werk und dessen Tradierung erhoben, sondern ›nur noch‹ verknüpft werden. Wie attraktiv solche Verknüpfungsoptionen sind, lassen zahlreiche Anfragen an den Handschriftencensus mit der Bitte um Vernetzung oder Aufnahme in den unmittelbaren HSC-Kontext erahnen. Das Spektrum der Anfragen reicht vom Einzelforscher über Forschungsprojekte, Forschungsdatenbanken, nationale/internationale Handschriftenportale bis hin zu bestandshaltenden Institutionen, Repertorien und Wörterbüchern. Einige Verbünde sind bereits prototypisch etabliert bzw. vereinbart u. a. mit der Deutschen Nationalbibliothek (GND), manuscripta.at, manuscripta.pl, dem Berliner Repertorium, der Rechtsbuchdatenbank DRD (Deutsche Rechtsbücher Digital), dem Gesamtkatalog der Wiegendrucke (GW), dem Bamberger Glossen-Projekt BStK Online und der LegIT-Datenbank, dem Katalog der deutschsprachigen illustrierten Handschriften in München (KdiH), dem Kolophon-Projekt in Kiel, der CoReMA-Datenbank in Graz (Cooking Recipes of the Middle Ages) und dem Handschriftenportal der Deutschen Handschriftenzentren.

Kommen wir damit zu einem Zukunftsfeld, das schon Realität geworden ist: Online-Publikationen und Social Media. Um den Handschriftencensus hat sich eine rege ›Mitmach-Kultur‹ etabliert. Weit über 20.000 Mitteilungen sprechen eine deutliche Sprache. Nicht selten handelt es sich bei diesen Mitteilungen um Fundmeldungen: Irgendwo in der Welt hat jemand ganz physisch einen neuen Textzeugen gefunden, hält ihn vielleicht sogar im Augenblick der Meldung selbst in Händen. Häufig sind es aber auch Funde in alter Forschungsliteratur.

Bis dato setzt sich bei Neufunden eine eher schwerfällige Forschungsmaschinerie in Bewegung. Das entscheidende Bindeglied auf dem Weg des Fundes in die Forschungsöffentlichkeit war dessen Publikation. Die sollte in der Regel in einer der angesehenen Fachzeitschriften – vorzugsweise der ZfdA[10] – erfolgen. Eine solche Publikation setzte hohe Hürden: Oft fühlten sich die Finder nicht kompetent genug, oder sie hatten keine Zeit, einen solchen Fundbeitrag in einem wissenschaftlichen Publikationsorgan zu verfassen. Die Publikation zog sich dann länger hin oder erschien schlicht nie, oder der Fund wurde in nur regional zugänglichen Medien veröffentlicht. In den genannten Fällen war der Forschungscommunity der Zugriff auf den Fund lange, im schlimmsten Fall komplett verwehrt. Hier setzt Maniculae an:

> Die Gründung der Open-Access-Zeitschrift Maniculae geht aus dieser alltäglichen Erfahrung der Arbeit des Akademievorhabens ›Handschriftencensus‹ hervor. Kurze, prägnant formulierte Beiträge informieren in Maniculae rasch und verlässlich über Neuigkeiten auf dem Gebiet der Handschriftenforschung. (AG Handschriftencensus [2020], S. 1)

Maniculae ist so konzipiert, dass sie besonders niederschwellige Publikationsangebote macht, d. h. ein Fundbericht kann – durch das Redaktionsteam begleitet – schnell, kurz und ohne Text-/Formbeschränkungen innerhalb weniger Tage publiziert werden.

Abb. 7: Maniculae-Startseite (https://maniculae.de/index.php/maniculae/)

Dieses Angebot wurde hervorragend angenommen. Kurz nach dem Start sind mittlerweile zwei Jahrgänge mit insgesamt sieben (Jahrgang 2020) und 15 (Jahrgang 2021) Fundberichten online. Für 2022 liegen zahlreiche Fundberichte vor, noch weit mehr sind angekündigt.

Eine weitere Dimension der Öffentlichkeitsarbeit eröffnen die sozialen Netzwerke. Bald nach Projektstart wurde ein Twitter-Account für den Handschriftencensus angelegt (@HSCensus). Überraschend schnell etablierte sich eine Handschriftencensus-Twitter-Gemeinde. Aktuell sind über 1.000 Follower dabei. Der Kanal wird für wissenschaftliche Fund- und Forschungsmeldungen, Werkidentifikationen bis hin zu Nachrichten aller Art, aber vor allem auch für die Kommunikation zu Forschungsdetails genutzt. So dauerte es beispielsweise nur Minuten, bis der sensationelle Fund eines noch ins 13. Jahrhundert datierten und damit die gesamte Sexualgeschichte des Mittelalters revolutionierenden ›Rosendorn‹-Fragments im Kloster Melk in der ganzen Welt bekannt war.[11]

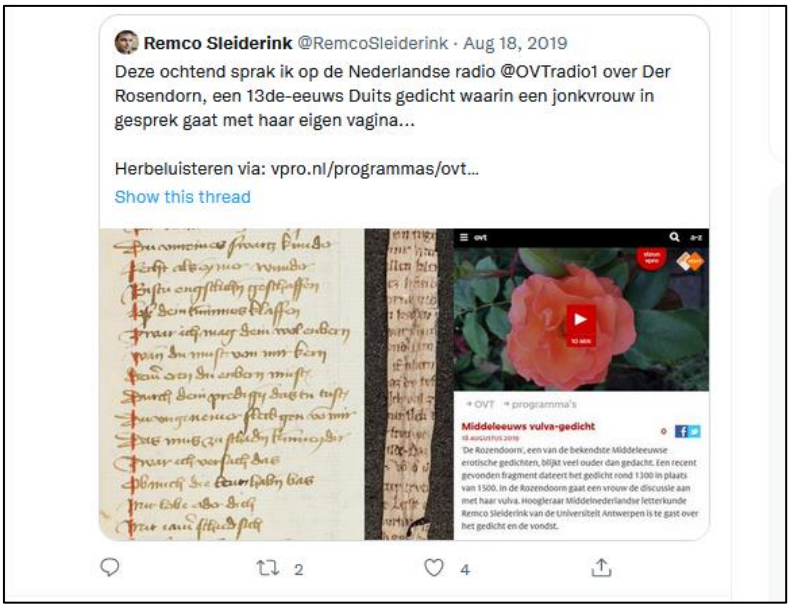


Abb. 8: Tweet zum ›Rosendorn‹-Fund

3. Eine Bilanz: Von der Handschrift zu den Metadaten

Der kurze Streifzug durch den Handschriftencensus zeigt, was bereits da und realisiert ist: ein riesiger Datenberg, leidlich strukturiert, auch teilweise normiert, aber längst noch nicht voll erschlossen. Er hat mehr noch gezeigt, wo die Gegenwartsdesiderate, aber zugleich auch wo Zukunftschancen liegen. Als Stichworte seien für die Desiderate nur genannt:

- Vervollständigung und Atomisierung der Daten
- Validierung
- Normierung
- Schnittstellen / Austauschformate

Woraus sich beinahe automatisch auch die Zukunftschancen speisen:

- Vernetzung

- Visualisierung
- Popularisierung
- Portalfunktion

Dass sich in der Zukunft en passant daraus nicht nur tiefere Forschungseinsichten, sondern vielfach auch völlig neue Forschungsfragen und Forschungsansätze ergeben werden, versteht sich von selbst. Wenn heute Editions- und Analyseprojekte, Digitalisierung und Bestandserfassung sowie allgemein die Fachforschung(en) – noch – weitgehend nebeneinander agieren – und das war eine durchaus schmerzvolle Erkenntnis der Bremer Tagung –, werden normierte und vernetzte Daten in der Zukunft völlig andere Formen der Zusammenarbeit nicht nur möglich machen, sondern Forschungsalltag werden. Ansätze in diese Richtung hat die Bremer Tagung in vielfältiger Form gezeigt bzw. in den Diskussionen zumindest eingefordert. Nun gilt es, diesen Weg zu beschreiten – und wir werden ihn erfolgreich nur gemeinsam gehen können, d. h. wir müssen aus den unterschiedlichsten Richtungen zueinander finden. Zusätzlich, und das war ebenfalls eine Bremer Grunderkenntnis, wird es nicht ohne ein ›institutionelles Dach‹ gehen.

Anmerkungen

1 Zu Genese und Geschichte des Handschriftencensus vgl. Klein 2009; Busch 2012; Heinzle/Klein 2001; Wolf 2009; Gamper/Glaßner 2013; Schwanitz 2019; Wolf 2019; Busch/Wolf 2019a; Busch/Wolf 2019b; Runzheimer 2019; Busch [u. a.] 2019; Wikipedia-Artikel ›Handschriftencensus‹.

2 Vgl. kursorisch Runzheimer 2019, hier bes. S. 92–95.

3 AG Handschriftencensus: Dr. Astrid Breith (Österreichische Akademie der Wissenschaften Wien); Prof. Dr. Nathanael Busch (Philipps-Universität Marburg); Dr. Rudolf Gamper (St. Gallen/Winterthur); Dr. Christine Glaßner (Österreichische Akademie der Wissenschaften Wien); Dr. Karl Heinz Keller (Wien); Dr. Daniel Könitz (Philipps-Universität Marburg); Dr. Wolfgang Metzger (Universitätsbibliothek Heidelberg); Dr. Diana Müller (Universitätsbibliothek Marburg); Tobias Müllerleile (Universitätsbibliothek Marburg); Dr. Monika Studer

(Universitätsbibliothek Basel); Dr. Bettina Wagner (Staatsbibliothek Bamberg); Prof. Dr. Jürgen Wolf (Philipps-Universität Marburg); Dr. Karin Zimmermann (Universitätsbibliothek Heidelberg); Dr. Elke Zinsmeister (Berlin-Brandenburgische Akademie der Wissenschaften).

4 Die HSC-IDs haben zugleich den Status von Permalinks. So setzen sich die URLs aller HSC-Beschreibungen und aller Materialien bzw. Übersichten aus einem Basisteil »https://handschriftencensus.de«, ggf. »Spezifizierungskürzel« und »HSC-ID« zusammen. Für die Handschriftenkatalogisate ergibt sich daraus z. B. folgende URL-Tektonik: https://handschriftencensus.de/12616. Für die Werke tritt das Kürzel »werke« hinzu (z. B. https://handschriftencensus.de/werke/2448); für die Autoren das Kürzel »autoren« (z. B. https://handschriftencensus.de/autoren/14) und für die Aufbewahrungsorte das Kürzel »hss« (z. B. https://handschriftencensus.de/hss/ Aachen).

5 Aktuell ist es noch nicht möglich, die Volldigitalisate grosso modo aus dem Gesamtbestand herauszufiltern. Eine entsprechende Listenfunktion ist in Vorbereitung.

6 Armenien, Belgien, Dänemark, Deutschland, Estland, Finnland, Frankreich, Großbritannien, Irland, Island, Italien, Kanada, Kroatien, Lettland, Liechtenstein, Litauen, Luxemburg, Monaco, Neuseeland, Niederlande, Norwegen, Österreich, Polen, Rumänien, Russland, Schweden, Schweiz, Slowakei, Slowenien, Spanien, Südafrika, Tschechien, USA, Ukraine, Ungarn, Vatikanstadt (vgl. https://handschriftencensus.de/hss/laender). In der Übersicht ›Verzeichnisse‹ kann mit einem Klick umgeschaltet werden zwischen der Ansicht ›Orte: alphabetisch‹ und ›Orte: nach Ländern‹.

7 Bei der GND handelt es sich um eine Sammlung von Normdaten zu Personen, Körperschaften, Konferenzen, Geographika, Sachschlagwörtern und Werktiteln mit ca. 9 Mio. Datensätzen. Host ist die Deutsche Nationalbibliothek. Entstanden ist die GND im Jahr 2012 aus GKD, PND und SWD. Sie basiert auf dem internationalen Regelwerk RDA (Ressource Description and Access). Der Zugang ist frei, die Bearbeitung eingeschränkt. Für mittelalterliche deutsche Werke verfügt der HSC über die höchste Zugangsberechtigung.

8 Exemplarisch Wolf 2008 (mit weiterführender Literatur).

9 Für die Epenüberlieferung im 12. und 13. Jahrhundert liegt der Anteil der Fragmente z. B. bei über 80% (vgl. HSC).

10 Vgl. dort die Rubrik ›Handschriftenfunde zur Literatur des Mittelalters‹ mit mittlerweile mehr als 250 Fundberichten (http://zfda.de/inhalt.php?mode=hssfunde).

11 Busch 2019 (HSC-Eintrag: https://handschriftencensus.de/26081).

Literaturverzeichnis

Sekundärliteratur

AG Handschriftencensus: Editorial, in: Maniculae 1 (2020), S. 1 (online)

Busch, Nathanael: www.handschriftencensus.de. Eine Datenbank sammelt Informationen zu deutschsprachigen Handschriften aus Hessen, in: Archiv-Nachrichten aus Hessen 12/1 (2012), S. 28–30 (online).

Busch, Nathanael: Höfische Obszönitäten? Ein ›Rosendorn‹-Fund und seine Folgen, in: ZfdA 148 (2019), S. 331-347.

Busch, Nathanael/Wolf, Jürgen: Radiobeitrag: »Faszination mittelalterliche Manuskripte«, SWR2 Kultur neu entdecken (3/2019a) (online).

Busch, Nathanael/Wolf, Jürgen: Radiobeitrag: »Kulturelles Erbe – Forscher untersuchen Handschriften aus dem Mittelalter«, Deutschlandfunk »Aus Kultur- und Sozialwissenschaften« (8/2019b) (online).

Busch, Nathanael/Gamper, Rudolf/Glaßner, Christine/Wolf, Jürgen: Radiobeitrag: Kulturelles Erbe aus dem Mülleimer, SRF2 »Kultur Kontext« (11/2019) (online).

Gamper, Rudolf/Glaßner, Christine: ›Handschriftencensus‹ – An Inventory of German Medieval Manuscripts, in: Golob, Nataša (Hrsg.): Medieval autograph manuscripts. Proceedings of the XVIIth Colloquium, Turnhout 2013, S. 291–295.

Heinzle, Joachim/Klein, Klaus: Die Marburger Repertorien zur Überlieferung der deutschen Literatur des Mittelalters, in: ZfdA 130 (2001), S. 245f. (online).

Kestemont, Mike/Karsdorp, Folgert/de Bruijn, Elisabeth/Driscoll, Matthew/Kapitan, Katarzyna A./Ó Macháin, Pádraig/Sawyer, Daniel/Sleiderink, Remco/Chao, Anne: Forgotten Books: The Application of Unseen Species. Models to the Survival of Culture, in: Science 375, 765 (2022) (online).

Klein, Klaus: Grundlagen auf dem Weg zum Text: www.handschriftencensus.de, in: Hofmeister, Wernfried/Andrea Hofmeister-Winter (Hrsg.): Wege zum Text. Überlegungen zur Verfügbarkeit mediävistischer Editionen im 21. Jahrhundert. Grazer Kolloquium 17.–19. September 2008 (Beihefte zu editio 30), Tübingen 2009, S. 113–119.

Riva, Gustavo Fernández: Network Analysis of Medieval Manuscript Transmission. Basic Principles and Methods, in: Journal of Historical Network Research 3 (2019), S. 30–49.

Runzheimer, Bernhard: »Das ist nicht ganz trivial… «. Die Anpassung gewachsener Projektstrukturen an moderne IT-Standards am Beispiel des Handschriftencensus, in: Huber, Martin/Krämer, Sybille/Pias, Claus (Hrsg.): Forschungsinfrastrukturen in den digitalen Geisteswissenschaften. Wie verändern digitale

Infrastrukturen die Praxis der Geisteswissenschaften? (Symposienreihe ›Digitalität in den Geisteswissenschaften‹), Frankfurt a. M. 2019 (online).
Schwanitz, Lara: Auf digitaler Spurensuche im Mittelalter, in: Avenue 7,1 (2019), S. 39–41.
Wolf, Jürgen: Buch und Text. Literatur- und kulturhistorische Untersuchungen zur volkssprachigen Schriftlichkeit im 12. und 13. Jahrhundert (Hermaea N.F. 115), Tübingen 2008, S. 20–27.
Wolf, Jürgen: Handschriftencensus – Eine Bestandsaufnahme, in: ZfdA 138 (2009), S. 279f. (online).
Wolf, Jürgen: Radiobeitrag: Interviewgespräch, WDR3 Mosaik (2/2019) (online).

Online-Ressourcen

BARTOC (Basic Register of Thesauri, Ontologies & Classifications): https://bartoc.org/.
Berliner Repertorium: https://www.literatur.hu-berlin.de/de/forschung/forschungsprojekte/berliner-repertorium.
BStK Online: Datenbank der althochdeutschen und altsächsischen Glossenhandschriften: https://www.uni-bamberg.de/germ-ling/forschung-und-lehre/forschungsprojekte/glossenhandschriftendatenbank/.
CoReMA (Cooking Recipes of the Middle Ages): https://gams.uni-graz.at/context:corema.
DNB (Deutsche Nationalbibliothek): https://dnb.de.
DRD (Deutsche Rechtsbücher Digital [Version 0.9.0]): http://magdeburger-recht.org/drd.
GND (Gemeinsame Normdatei): https://gnd.network/Webs/gnd/DE/Home/home_node.html.
GeoNames: https://www.geonames.org/.
GiN (Germanistik im Netz): https://www.germanistik-im-netz.de/.
GW (Gesamtkatalog der Wiegendrucke): https://www.gesamtkatalogderwiegendrucke.de.
Getty Union List of Artist Names: https://getty.edu/research/tools/vocabularies/ulan.
HSC (Handschriftencensus): https://handschriftencensus.de.
Handschriftenarchiv (Berlin-Brandenburgische Akademie der Wissenschaften, Arbeitsstelle Deutsche Texte des Mittelalters. HSA-Beschreibungen): https://www.bbaw.de/forschung/dtm/HSA/hsa-index.html.

Anschrift des Autors:

Prof. Dr. Jürgen Wolf
Philipps-Universität Marburg
Fachbereich 09, Institut für Deutsche Philologie des Mittelalters
Deutschhausstraße 15
35032 Marburg
E-Mail: juergen.wolf@staff.uni-marburg.de

Stefanie Dipper / Simone Schultz-Balluff

ReM für Mediävist*innen

Perspektiven des Referenzkorpus Mittelhochdeutsch (1050–1350) für germanistisch-mediävistische Fragestellungen

Abstract. In diesem Beitrag wollen wir illustrieren, wie die historischen Referenzkorpora für germanistisch-mediävistische Fragestellungen genutzt werden können. Wir tun dies anhand von drei beispielhaften Fragestellungen, für die wir das Referenzkorpus Mittelhochdeutsch auswerten: (i) Merkmalszuschreibung über Attribuierungen der Personennamen; (ii) Personifizierung; (iii) Metaphorisierung. Der Beitrag zeigt, wie das Referenzkorpus Mittelhochdeutsch und seine Annotationen (Lemma, Wortart) mit dem Korpussuchtool ANNIS durchsucht werden kann und wie die entsprechenden Treffer auch quantitativ ausgewertet werden können.

In den letzten Jahren entstanden eine Reihe von Referenzkorpora für verschiedene Sprachstufen und -räume des Deutschen, siehe die Übersicht in Tabelle 1. Die Korpora sind reichhaltig annotiert, sowohl mit Metadaten wie auch mit linguistischen Annotationen, darunter Lemma, Flexionsmorphologie und Wortart. Sie sind u. a. über das Korpussuchtool ANNIS (Annotation of Information Structure; Krause/Zeldes 2016) abfragbar, das einfache Suchanfragen wie z. B. nach den Vorkommen eines bestimmten Lemmas bis hin zu komplexen Suchanfragen, die mehrere Annotationsebenen involvieren, unterstützt. Die Referenzkorpora werden aktuell auf verschie-

denen Servern bereitgestellt, die Seite www.deutschdiachrondigital.de verlinkt auf die jeweiligen Internetseiten.

Korpus	**Zeitraum**	**Sprachraum**	**Größe**
ReA	750–1050	Althochdeutsch, Altsächsisch	0,5 Mio
ReM	1050–1350	Mittelhochdeutsch	2,5 Mio
ReF	1350–1650	Frühneuhochdeutsch	3,5 Mio
ReN	1200–1650	Mittelniederdeutsch, Niederrheinisch	2,3 Mio

Tab. 1: Überblick über die Referenzkorpora, ihre Abdeckung und ihre Größe (in Anzahl Tokens)

Die Referenzkorpora zu einzelnen Zeit- und Sprachstufen des Deutschen (Altdeutsch, Mittelhochdeutsch, Mittelniederdeutsch/Niederrheinisch, Frühneuhochdeutsch) zeigen eine ausgewogene Präsenz der Textsorten und der Textform (Vers, Prosa) und legen Wert auf eine gleichmäßige sprachräumliche und zeitliche Verteilung der Texte. Für einen möglichst identischen Umfang werden längere Texte nur in Ausschnitten erfasst (gezählt werden dabei die Wortformen). Mit der lexematischen, morphologischen, syntaktischen und in Teilen auch graphematischen Annotation bieten sich die Referenzkorpora in erster Linie für linguistische Fragestellungen an; der strukturierte Aufbau hinsichtlich der Textsortenverteilung macht die Referenzkorpora jedoch auch für philologische Fragestellungen attraktiv.

Nutzungsszenarien der Referenzkorpora für Fragestellungen der germanistischen Mediävistik

Die Referenzkorpora sind bisher für genuin linguistische Fragestellungen herangezogen worden und bilden z. B. die Grundlage für grammatikographische Untersuchungen und Darstellungen.[1] Gelegentlich wurden Untersuchungen an der Schnittstelle von Linguistik und Literaturwissenschaft im Bereich der historischen Semantik bzw. Konzeptforschung vorgenommen (Schultz-Balluff 2009, 2014, 2018, 2020), dies jedoch mit hohem Einsatz manueller und nicht mittels automatischer Analyseverfahren.

Mit Abschluss der Aufbereitung der Textdaten des Referenzkorpus Mittelhochdeutsch (1050–1350; abgekürzt ReM) können nun verstärkt auch automatische Verfahren für nicht rein linguistische Fragestellungen angewandt werden. Im Folgenden werden drei Nutzungsszenarien entworfen, die an der Schnittstelle von Linguistik und Literaturwissenschaft angesiedelt sind und dazu anregen möchten, germanistisch-mediävistische Fragestellungen konsequent von der sprachlichen Gemachtheit von Literatur bzw. Text allgemein her zu denken.

Für die konkreten Fragestellungen werden (mögliche) methodische Zugänge vorgestellt, die abschließend im Rahmen einer Methodenreflexion kritisch betrachtet werden. Grundsätzlich geht es um das Zusammenwirken qualitativer und quantitativer Verfahren, dabei spielt vor allem die Validierung von Aussagen über Gegenproben eine Rolle, für die annotierte Korpora genutzt werden können.

Figurenzeichnung mittels Attribuierung

Die Figurenkonzeption und Figurenzeichnung beschäftigt die mediävistische Literaturwissenschaft intensiv, vor allem geht es häufig darum, wie Figuren widersprüchlich aufgebaut oder inszeniert werden und welche Auswirkungen dies auf den Handlungsverlauf hat. Die Widersprüchlichkeit

liegt oftmals in einem Auseinandertreten von Positionierungen der Figur (durch sich selbst oder durch den Erzähler) oder dem Figurenhandeln begründet.

Nicht zuletzt wird die grundlegende Konzeptionierung einer Figur stark durch unmittelbare Kennzeichnungen in Form von Attributen bestimmt. D. h. wenn der Name z. B. immer mit einem oder einer bestimmten Auswahl an Adjektiven kombiniert wird, prägt dies das Figurenkonzept. Möglich ist daher, dass hier Stereotype transportiert und auch zementiert werden, die vielleicht gar nicht (mehr) mit dem Verlauf der Erzählung korrelieren. Bislang liegt – soweit wir das überblicken – keine Untersuchung vor, die die sprachlichen Verfahren untersucht, mit denen Figuren konturiert werden. Einen ersten und naheliegenden Zugriff bietet die Untersuchung der Merkmalszuschreibung über Attribuierungen der Personennamen. Das Ziel wäre herauszufinden, inwiefern sprachliche Verfahren zur Figurenzeichnung beitragen. Eine quantitative Untersuchungsmethode könnte Aussagen auf der Ebene der Textanalyse stützen oder grundsätzlich ermöglichen. Im Folgenden zeigen wir, wie dies anhand des Korpus ReM mit Hilfe von ANNIS untersucht werden kann.

Die Referenzkorpora sind mit Wortarten (»part of speech«, abgekürzt »pos«) gemäß dem Schema HiTS annotiert (Dipper [u. a.] 2013). Dieses Schema kodiert nicht nur die Wortart als solche, sondern spezifiziert zusätzliche morphosyntaktische Information. Z. B. wird bei allen Vorkommen von Adjektiven zusätzlich angegeben, ob sie attribuierend vorangestellt (pos=»ADJA«), attribuierend nachgestellt (pos=»ADJN«) oder prädikativ (pos=»ADJD«) genutzt werden. (1) zeigt Beispiele aus ReM für diese drei Typen; die Adjektive sind jeweils Instanzen des Lemmas *guot*.[2]

(1)
a. ADJA: *alſo ſint diu guoten werch niecht ane rehte geloube·*
Path: 11-12_1-obd-PV-X > M010-N1 (tok_dipl 209 - 220)
ANNIS: https://annis.linguistics.rub.de/?id=234aceb4-711b-4d45-a89a-2e296b452ffb

b. ADJN: *So nemich einen helit got vnde balt.*
Path: 12-13_1-mdnd-PV-X > M206-N1 (tok_dipl 13234 - 13245)
ANNIS: https://annis.linguistics.rub.de/?id=05a15456-caa2-40ca-a1f8-0aaccc5e3e1f

c. ADJD: *Scelle wurze ſoch iſt got. den tunchelen ogen.*
Path: 11-12_1-obd-PV-X > M126-N1 (tok_dipl 81 - 91)
ANNIS: https://annis.linguistics.rub.de/?id=63ec52af-436f-4bfd-9126-4f0656a80159

Diese Annotation kann also genutzt werden, um spezielle sprachliche Formen der Attribuierung zu untersuchen. Dazu suchen wir in ANNIS nach Vorkommen von Eigennamen (pos=»NE«) mit entweder vorangestelltem ADJA oder nachfolgendem ADJN.[3]

Wir suchen zunächst nach vorangestellten Adjektiven. Abb. 1 zeigt das Such-Interface von ANNIS: Die Suchanfrage (»Adjektiv gefolgt von Eigennamen«) ist links oben eingegeben, direkt darunter steht die Anzahl der Belege, auf die die Suchanfrage zutrifft: insgesamt 4401 in 266 verschiedenen Dokumenten. Unten links sind die Teilkorpora von ReM ausgewählt, in denen gesucht werden soll (z. B. 11-12_1-obd-PV-X) — im aktuellen Beispiel sind alle Teilkorpora ausgewählt, so dass im gesamten Korpus gesucht wird. Rechts im Hauptfenster werden die Belege im Kontext angezeigt. Die gesuchten Ausdrücke (Adjektiv und Eigenname) werden von ANNIS farblich hervorgehoben.[4]

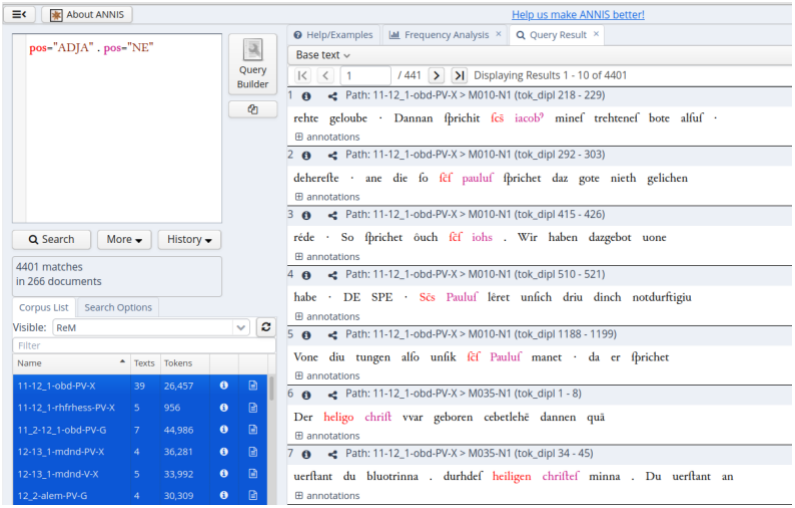


Abb 1: Das Such-Interface des Korpustools ANNIS.

Man sieht schon an den ersten sieben Treffern, die in Abb. 1 angezeigt werden, dass *sancte* ein beliebtes Adjektiv zur Attribuierung von Eigennamen darstellt. ANNIS bietet eine einfache statistische Analyse, mit der Kreuztabellen mit beliebigen Merkmalen erstellt werden können. Abb. 2 zeigt die am häufigsten vorkommenden Kombinationen von Adjektiv und Eigennamen (jeweils als Lemmata) mit ihren Frequenzen.[5] Demnach ist *hèilig* + *Krist* die häufigste Kombination mit 308 Vorkommen in ReM. Nach *sancte* mit insgesamt 2854 Vorkommen[6] gehören *hèilig* (351 Vorkommen), *guot* (148) und *süèze* (95) zu den häufigsten Adjektiven in dieser Kombination.

Download as CSV

1073 items with a total sum of 4401 (query on 13_2-bairalem-PV-G, 12-13_1-mdnd-PV-X, 12_2-wmd-PV-G, 13_1-alem-PV-G, 14_1-alem

rank	#1|lemma	#2|lemma	count
1	hèilig	Krist	308
2	sancte	Johannes	283
3	sancte	Pêter	217
4	sancte	Maria	158
5	sancte	Paulus	146
6	sancte	Johann	105

Abb. 2: Die häufigsten Kombinationen aus vorangestelltem Adjektiv und Eigennamen (Lemmata) in ReM mit ihren Frequenzen.

Die Kreuztabellen lassen sich auch alphabetisch nach dem Eigennamen sortieren, so dass dessen Attribuierungen verglichen werden können. Abb. 3 zeigt das Beispiel *Alexander*, der häufig als *wunterlich* und eher selten als *küène* oder *stolz* charakterisiert wird.

rank	#1|lemma	#2|lemma	count
76	wunter-lich	Alexander	8
205	kriëchisch	Alexander	3
263	müè-lich	Alexander	2
341	wîs(e)	Alexander	2
658	küène	Alexander	1
678	stolz	Alexander	1
689	riche	Alexander	1
864	vri	Alexander	1
885	tumb	Alexander	1

Abb 3: Attribuierungen des Eigennamens *Alexander*.

Schaut man sich zum Vergleich Eigennamen mit nachgestellten Adjektiven an, so zeigt sich, dass es davon in ReM nur 90 Vorkommen in 29 Texten gibt. Über die Hälfte enthält das Adjektiv *sælig* (mit einer Frequenz von 51), dann folgen *wîse* (7) und *rèine* (3).

Nachgestellte Adjektive kommen generell überraschend selten in ReM vor, mit insgesamt 5402 Instanzen in 208 Dokumenten. Zum Vergleich: Es gibt 70.782 Instanzen von vorangestellten Adjektiven in 378 Dokumenten. Insgesamt enthält ReM 395 Dokumente, d. h. nachgestellte Adjektive kommen nur in rund der Hälfte der Dokumente vor. Zwei naheliegende Hypothesen sind, dass nachgestellte Adjektive eher in frühen Texten vorkommen sowie eher in Versen als in Prosa. Diese Hypothesen lassen sich überprüfen, indem man in die Kreuztabellen geeignete Metadaten integriert.

Eine Kreuztabelle mit dem Meta-Merkmal »time«, das die grobe Datierung der Handschrift wiedergibt, zeigt ein sehr uneinheitliches Bild mit einer ungleichmäßigen Verteilung. Das Merkmal »genre«, das zwischen Versen, Prosa und Urkunden unterscheidet, zeigt hingegen ein sehr klares Bild: 4457 (d. h. 83%) der 5402 nachgestellten Adjektive stammen aus

Versen, 795 (15%) aus Prosa und 115 (2%) aus Urkunden, der Rest aus gemischten oder nicht klassifizierbaren Dokumenten. Insofern bestätigt sich zumindest die zweite Hypothese.

Schaut man sich allerdings nur die Kombinationen von Eigennamen mit nachgestelltem Adjektiv an, so ändert sich das Bild: Zunächst scheint es, als ob es hier eine klare diachrone Entwicklung hin zu mehr Vorkommen von nachgestellten Adjektiven gäbe, mit gerade mal 15 Vorkommen im Zeitraum 1100–1250, verglichen mit 75 Vorkommen im (kürzeren) Zeitraum 1250–1350. Allerdings sieht man, wenn man das Merkmal »genre« hinzunimmt, dass es sich bei den späten Vorkommen ganz überwiegend um Urkunden handelt und eher selten um Verse. Fast alle Vorkommen in den Urkunden enthalten das Adjektiv *sælig*, von dem bereits oben die Rede war. Ein typisches Beispiel dafür ist (2).

(2)
das fro anne Hern Johannef felígen def Níenerf tohter.
Path: 13_2-alem-PU-G > M346-G1 (tok_dipl 4416 - 4427)
ANNIS: https://annis.linguistics.rub.de/?id=5c464117-6fe5-4afc-9b6a-baa9333ce662

Fazit aus den bisherigen Beobachtungen ist, dass in den Texten von ReM nachgestellte Adjektive generell, aber auch insbesondere bei Eigennamen sehr selten sind und es sich bei den Eigennamen vorwiegend um Urkunden-spezifische Verwendungsweisen handelt. Daher sollte sich eine tiefergehende Untersuchung auf vorangestellte Adjektive konzentrieren.

Abschließend noch eine allgemeinere Überlegung: Wie »typisch« sind eigentlich Attribuierungen durch Adjektive, einerseits bei Nomen appelativa, andererseits bei Eigennamen? D. h. wir machen eine »Gegenprobe« und ermitteln dazu die Anzahl unattribuierter Nomen bzw. Eigennamen. Die Frequenzen in Tab. 2 zeigen, dass adjektivische Attribuierungen in der Minderheit sind und Eigennamen verglichen mit Nomen appelativa sogar seltener attribuiert werden. Allerdings muss man hier berücksichtigen,

dass nicht nur Personenbezeichnungen, sondern z. B. auch Ortsnamen unter die Eigennamen fallen.

	gesamt	**ohne ADJ-Attribution**	**mit ADJA**	**mit ADJN**	**mit ADJA und ADJN zugleich**
Nomen appelativa	363.928	310.484	49.593	3478	373
	100%	85%	14%	1%	0,1%
Eigennamen	44.686	40199	4397	86	4
	100%	90%	10%	0.2%	0%

Tab. 2: Absolute und relative Frequenzen von Nomen bzw. Eigennamen mit und ohne adjektivische Attribuierung.

Natürlich gibt es im Deutschen noch diverse andere Konstruktionen der Attribution, z. B. durch Appositionen (wie in *Sifrit der vil chune*[7]), Relativsätze (*Helena die daz kriuze vant*[8]), Genitivphrasen oder Präpositionalphrasen. Diese lassen sich jedoch ohne eine syntaktische Annotation nicht direkt abfragen. Man könnte alternativ heuristische Anfragen stellen, die zu viele Treffer erzielen, die man nachträglich manuell sichtet.

Personifizierung und Metaphorisierung

Weiterführend können Korpora dazu genutzt werden, die sprachliche Konstruktion von Bildlichkeit zu untersuchen. Im Folgenden soll am Beispiel der Personifikation und der Metaphorik exemplarisch gezeigt werden, wie Analysen vorgenommen werden können und in welcher Hinsicht Ergebnisse zu erwarten sind.

Personifikationen zeichnen sich dadurch aus, dass Nicht-Belebtes, Nicht-Menschliches, Abstrakta oder Kollektiva zu handelnden Instanzen werden. Auf der sprachlichen Ebene zeigt sich dies, indem bei diesen Entitäten z. B. Verben mit einer Semantik stehen, die Handlung ausdrücken. D. h. die Kombination aus Instanz und Handlung ist grundsätzlich nicht möglich und losgelöst vom Kontext ungewöhnlich (z. B. wenn das Herz spricht).

Die korpusbasierte Analyse von Personifikation erfordert ein hohes Maß an Abstraktion und Flexibilität, wenn man sich nicht von gängigen Personifikationen, wie z. B. (Frau) Welt, Herz, Tod, leiten lassen will. Die korpusbasierte Analyse möchte den Prozess der Personifizierung analysieren und stellt die Frage nach den hierfür genutzten sprachlichen Verfahren. Ein unvoreingenommener Ansatzpunkt könnten daher genuin menschliche Fähigkeiten sein, wie z. B. sprechen (verbale Kommunikation) oder lesen und schreiben (Kulturtechniken). Etwas weiter gefasst können auch Fähigkeiten einbezogen werden, die nur Lebewesen zu eigen sind, wie z. B. aus den Bereichen der Wahrnehmung (sehen, hören), der Bewegung (gehen, fahren), der verbalen (reden, raten, singen) und der nonverbalen Artikulation (zeigen). Entsprechende Verben können also den ersten Zugriff für eine Korpusabfrage bilden, in einem zweiten Schritt stellt sich die Frage nach den zugehörigen Substantiven, die die Agentiva sind.

Für die Suche nach vielversprechenden wie den oben genannten Verben nutzen wir das Merkmal »lemma«, z. B. in der Abfrage lemma=»sprëchen«. Neben dem Verb interessiert uns v. a. dessen Subjekt. Dazu nutzen wir den Näheoperator, den ANNIS anbietet, und suchen nach Nomen im Kontext von z. B. +/- 4 Wörtern (natürlich kann das Subjekt auch weiter entfernt stehen, allerdings steigt dann auch die Wahrscheinlichkeit, dass es sich um das Subjekt eines anderen Verbs handelt). Zusätzlich schränken wir die Nomen anhand der morphologischen Annotation auf Nomen im Nominativ ein.

Nun führt man wieder eine Frequenzanalyse durch, mit der die Lemmata der vorkommenden Subjekte gezählt werden. Die ersten 12 Lemmata bezeichnen, wie erwartet, Personen.[9] Die ersten unbelebten Lemmata sind

schrift (Platz 13 mit einer Frequenz von 59) und *stimme* (Platz 21), also Nomen des semantischen Felds »Kommunikation«. Auf späteren Plätzen kommen Nomen wie *gèist* (Platz 31), *rède* (Platz 33), *mund* (Platz 36), *wort* (Platz 39), *hërze* (Platz 42), *zît* (Platz 48) und *sêle* (Platz 52). Beispiel (3) zeigt einen typischen Beleg mit dem Nomen *schrift*.

(3)
Wan alſo div heilige ſchrift ſprichet.
Path: 12_2-bairalem-PV-G > M214-G1 (tok_dipl 2018 - 2029)
ANNIS: https://annis.linguistics.rub.de/?id=803dc1b6-2e69-45cf-8727-70644bafb8fe

Die Personifizierungen, die man über geeignete Verben gefunden hat, kann man wiederum als Ausgangspunkt für Abfragen nutzen, um weitere typische Verben zu finden: Welche Verben kommen z. B. mit dem Nomen *schrift* oder mit dem Nomen *hërze* in einem Kontext von +/- 4 Wörtern vor? Diese umgekehrte Sicht auf die Nomen kann zugleich zeigen, welche Verben als Marker für die Personifizierung dienen.

Mit dem Subjekt *schrift* kommen insgesamt 79 verschiedene Verben (im vorgegebenen Kontext) mit einer Gesamtfrequenz von 247 vor, mit dem Subjekt *hërze* sind es deutlich mehr: 372 mit einer Gesamtfrequenz von 865. (Allerdings ist das Nomen *hërze* generell auch viel häufiger als *schrift*). (4) zeigt die jeweils 10 häufigsten Verben für diese beiden Nomen, in Klammern ist jeweils die Frequenz der Verben angegeben.

(4)
a. Verben mit dem Subjekt *schrift*: *sprëchen* (59), *sagen* (38), *er-vüllen* (18), *gëben* (11), *nènnen* (8), *quëden* (7), *jëhen* (6), *râten* (6), *tuon* (5), *künden* (4)
ANNIS: https://annis.linguistics.rub.de/?id=17d9388d-0b94-4a97-86d5-791aea62e59a
b. Verben mit dem Subjekt *hërze*: *sprëchen* (24), *gër(e)n* (24), *tragen* (22), *tuon* (21), *vröuwen* (21), *stân* (19), *sëhen* (17), *be-ginnen* (13), *sagen* (12), *brëchen* (11)
ANNIS: https://annis.linguistics.rub.de/?id=859aa9ee-9d7d-42bc-9a76-4a1149dd0917

Die Auflistung in (4) zeigt, dass unter den hochfrequenten Verben ganz eindeutig diejenigen vorherrschen, die eine Personifizierung generieren. Bei beiden Nomen steht *sprëchen* ganz oben, es lassen sich aber unterschiedliche semantische Felder erkennen: Während beim Subjekt *schrift* Verben der Kommunikation eindeutig vorherrschen (7/10), sind es bei *hërze* vorwiegend Verben der nonverbalen Aktivität. Zudem ist die Verteilung über verschiedene Verben deutlich gleichmäßiger als bei *schrift*, wo alleine die beiden ersten Verben, *sprëchen* und *sagen*, schon knapp 40% der Belege ausmachen.

In einem ersten Schritt werden also aus allen bei *sprëchen* stehenden Substantiven (im Nominativ) diejenigen herausgefiltert, die üblicherweise nicht sprechen können (wie *schrift*, *sêle*, *rât*, *buoch*). Über die Verbsuche zeigt sich überdies, welche Substantive noch mit diesem Verb kombiniert werden – sowohl erwartbare als auch überraschende, d. h. möglicherweise weitere Personifikationen. Unterschiedliche Perspektivierungen können die vertiefende Analyse leiten:

- Einzelnes Lexem, z. B. *schrift*: Welche Verben stehen bei *schrift*? Gibt es neben *sprëchen* weitere Verben, die auf Personifikation hindeuten? Frequenz: Welchen Stellenwert haben diejenigen Verben, die Personifikation anzeigen, gegenüber anderen?
- Mehrere Lexeme: Vergleich der einzelnen Personifikationsprofile: Gibt es – neben *sprëchen* – gängige Verben, die Personifikationen anzeigen? Welche Funktionen lassen sich ablesen?
- Verallgemeinernd: Lassen sich sprachliche Verfahrensweisen definieren, die typisch für Personifikationen sind?

Dieses Verfahren bietet eine quantitative und sprachbasierte Grundlage, nach dem Schneeballprinzip können nun weitere Suchabfragen vorgenommen werden, die das Untersuchungsfeld untermauern und/oder neue Perspektiven eröffnen. Auf diese Weise gelangen Dinge in den Blick, die nicht primär oder gar nicht im Interesse der Ausgangsfrage standen. So lenkt die

Datenbasis die Analyse und ermöglicht auf empirischer Grundlage (= ausgewogenes Textverhältnis) valide Aussagen.

Metaphern sind innerhalb der Germanistik kontinuierlich Gegenstand der Forschung. Die Teildisziplinen setzen deutlich eigene Schwerpunkte: In der Linguistik stehen theoretische und methodische Zuwege im Zentrum (mit deutlichem Anschluss an die Forschungen aus dem anglo-amerikanischen Raum), in der Literaturwissenschaft steht zumeist eine Metapher im Gebrauch eines Werks oder Autors im Zentrum. Nur selten werden literaturwissenschaftliche Metaphernanalysen an die aktuellere linguistische Forschung angebunden (so Cöllen 2012).

Im Fokus der literaturwissenschaftlich ausgerichteten germanistischen Mediävistik stehen Metaphern in einzelnen Werken oder im Gebrauch einzelner Autoren.[10]

Eine korpusweite Analyse von Metaphern wird als Ansatzpunkt einen Bildbereich oder einen konkreten Begriff bzw. ein konkretes Lexem wählen müssen:

- Bildbereich: z. B. Naturelemente > Pflanzen > Blumen > konkrete Bezeichnungen; z. B. Liebe > mhd. Äquivalente *minne*, *liebe* und Ableitungen; der bildgebende Bereich muss eingeschränkt und auf (mehrere) Lexeme hin abstrahiert werden;
- Konkreter Begriff: Es steht von vornherein fest, dass z. B. *lieht*, *herz*, *wunde* das Kernlexem der Metapher ist, sodass von diesen ausgehend die Suche erfolgt.

Nach einem ersten Zugriff werden auf Wort- und Phrasenebene die lexikalischen Solidaritäten erfasst: Objektgebrauch (agentiv und patientiv), Präpositionalphrasen, direkte und indirekte Modifikation, Schlüsselwörter im weiteren Kontext (vgl. hierzu die Mehr-Ebenen-Analyse von Warnke/Spitzmüller 2008; Schultz-Balluff 2018; Schultz-Balluff 2020). Auf diese Weise können alle Elemente komplexer Metaphern erfasst werden. Methodisch werden zunächst – dem linearen Ansatz folgend (Lakoff/Johnson

1980) – Quell- und Zielbereiche erfasst, in weiteren Schritten können auch Metaphern-Netzwerke erschlossen werden (vgl. die Blending-Theorie von Fauconnier/Turner 2002). Eine korpusbasierte Analyse kann dem systematischen sprachlichen Aufbau von Metaphern nachgehen und damit den sprachlichen Prozess der Metaphorisierung zu erfassen versuchen. Am Beispiel des Konzepts von ›Wunde‹ und der metaphorischen Verwendung im religiösen Kontext soll im Folgenden die Vorgehensweise über gezielte Korpusabfragen verdeutlicht werden.

In einem ersten Suchlauf wird das Verbumfeld ermittelt und ausgewertet. Ähnlich wie bei der Suche nach Personifikationen können wir dazu wieder den Näheoperator einsetzen und nach Verben im Umfeld von +/- 4 Wörtern suchen. Diese Suchanfrage findet insgesamt 177 verschiedene Verben mit einer Gesamtfrequenz von 403. (5) zeigt die 15 häufigsten Treffer mit ihren Frequenzen. Im Gegensatz zu (3) und (4) oben ist hier das Lemma *wunte* in der Anfrage nicht auf einen bestimmten Kasus eingeschränkt, sodass es in unterschiedlichen syntaktischen Funktionen zu den Verben steht.

(5)
Kontext-Verben des Lemmas *wunte*: *hèilen* (21), *slahen* (20), *sprëchen* (12), *tuon* (12), *ent-vâhen* (12), *sëhen* (11), *salben* (10), *vlièzen* (10), *binten* (9), *machen* (9), *ge-nësen* (9), *ge-hèilen* (6), *vinden* (6), *zèigen* (6), *gëben* (6)
ANNIS: https://annis.linguistics.rub.de/?id=c25b40c0-7225-47df-b1b9-42f8bd015f2f

Abb. 4 zeigt das Ergebnis einer manuellen Klassifikation aller Verben im (größeren) Kontext von *wunte* nach Kategorien wie »aktiv« vs. »passiv« und »Art der Wahrnehmung«. Verwendet wird dem Ursprungsbereich der Medizin entsprechendes Vokabular, wie z. B. *heilen*, *erlîden*, *waschen*, oder es wird die Art und Weise des Zufügens bezeichnet (*snîden*, *stechen* usw.). Ungewöhnlich ist es, Wunden zu küssen oder zu offenbaren:

Abb 4: Verben im Kontext von *wunte* (entnommen aus Schultz-Balluff 2020, S. 42).

Über eine Zuordnung zu unterschiedlichen Lebensbereichen (Recht, Medizin, Kampf, Religion) und semantische Gruppierungen engt sich der Bereich ein, in dem Wunden eine transzendentale Dimension bekommen.

Jedoch führt allein die Untersuchung der Verbsemantik nicht zu einer Qualifikation als Metapher. Es müssen weitere Mosaiksteine gesammelt werden. Ein zusätzlicher Marker für metaphorischen Gebrauch kann die Attribuierung sein. Normalerweise wird der Zustand einer Wunde in medizinischer Sicht bezeichnet (*bluoten*), oder es wird die Schwere angezeigt (*grôz, tief, tôdlich*). Wenn eine Wunde allerdings *heilec* oder *edel* ist, kann davon ausgegangen werden, dass der physische Bereich verlassen und der transzendente beschritten worden ist. (6) zeigt die 15 häufigsten Adjektive einer Anfrage, die nach vorangestellten Adjektiven zum Lemma *wunte* im Kontext von 1-2 Wörtern sucht. Wie man sieht, sind die Belege nicht besonders häufig. Das liegt aber nicht daran, dass *wunte* eher selten modi-

fiziert würde: in 17% der Belege (72/421) steht ein vorangestelltes Adjektiv, also etwas häufiger als im Durchschnitt der Nomen appelativa (vgl. Tab. 2).

(6)
Adjektivische Modifikatoren zum Lemma *wunte*: *tièf* (13), *grôz* (9), *hèilig* (8), *tôd-lich* (5), *niuwe* (3), *vrèis-sam* (3), vrisch (2), *tûsent* (2), *bluoten* (2), *brëchen* (2), *rèin(e)* (2), *stëchen* (1), *sô-ge-tân* (1), *èdel(e)* (1), *bitter* (1)
ANNIS: https://annis.linguistics.rub.de/?id=76d404ac-6df7-4c75-b80e-19096e13b33d

Da Metaphern komplex sind und Elemente aus dem Quell- und dem Zielbereich enthalten, muss über mehrere markante Elemente ein metaphernspezifisches Cluster erarbeitet werden, über das es möglich wird, den Prozess der Metaphorisierung nachzuzeichnen. An einem konkreten Textbeispiel mit ›Wunde‹ soll verdeutlicht werden, dass die Qualifizierung einer Metapher nur über mehrere Suchabfragen im Umfeld eines zentralen Wortes erfolgen kann. Zunächst würde in der Trefferliste *wunde* + *salben* nicht ungewöhnlich erscheinen, erst die Präpositionalphrase *mit unser andacht* und das Substantiv *got* als Wunden salbender Akteur im weiteren Umfeld markieren Quellbereich (Medizin) und Zielbereich (Religion).

(7)
Got *wart wnt vmb vnſere ſunte. Die* ***wnden*** *ſuln wir im* ***ſalben*** *mit vnſer* ***andacht***.
Path: 13_1-bair-P-X > M168-N1 (tokens 11639 - 11652)
ANNIS: https://annis.linguistics.rub.de/?id=fffd8023-111f-4a4d-90ba-c49fa01f9c55

Zusammenfassung und Fazit: Worin liegen die Möglichkeiten und Grenzen der Auswertung strukturierter Korpora?

Mit der Suche in den Referenzkorpora ist sichergestellt, dass die Datenbasis handschriftennah ist und nicht auf – zum Teil veralteten – Editionen beruht. Die Transparenz bei der Auswahl der einzelnen Texte und Textzeugen ermöglicht die voneinander getrennte Betrachtung entlang der Datierung

des Textes und des jeweiligen Überlieferungsträgers, d. h. Sprachstand und sprachliche Form entsprechen immer der Zeit und dem Raum des konkreten Textzeugen.

Fazit aus korpuslinguistischer Sicht

Die gezeigten Fallbeispiele illustrieren, welche Möglichkeiten ein annotiertes, strukturiertes Korpus eröffnet. Auf Basis der Lemma- und Wortart-Annotationen lassen sich gezielt Belege suchen, die mit Hilfe der Frequenzanalyse und gegebenenfalls unter Einbeziehung der Metadaten (»time« und »genre«) weiter analysiert werden können. Umgekehrt gilt natürlich, dass nicht (oder nur erschwert) gesucht werden kann, was nicht explizit annotiert ist.

Typischerweise muss man bei Korpusanfragen einen Kompromiss finden zwischen einer allgemeineren Anfrage, die eine hohe Abdeckung hat (d. h. quasi alle oder die meisten Belege findet), die aber auf Kosten der Genauigkeit geht (d. h. es gibt auch viele ungewünschte Treffer), und einer spezifischen Anfrage, die den Schwerpunkt auf eine hohe Genauigkeit legt, bei der potenziell interessante Belege nicht gefunden werden. Oft lohnt es sich, zunächst allgemeinere Anfragen zu stellen, die man nach und nach weiter einschränkt und spezifischer macht.

ReM ermöglicht durch seine Größe und Strukturiertheit auch quantitative Analysen, d. h. man kann sich Muster und Tendenzen anschauen und zu (tentativen) Generalisierungen kommen. Dabei muss immer der Bezugspunkt klar sein: Womit lassen sich die Frequenzen der aktuellen Suchanfrage vergleichen? Einen solchen Bezugspunkt herzustellen, ist auch Zweck der »Gegenprobe«, die wir durchgeführt haben. Generell ist auch eine manuelle Sichtung wichtig, um einerseits sicherzugehen, dass die Anfrage auch wirklich die gewünschten Belege abdeckt, und um andererseits für eine Interpretation der Zahlen die eigentlichen Daten hinter den Zahlen wahrzunehmen.

Schließlich bietet ein Korpus die Möglichkeit, replizierbare Ergebnisse zu erzielen. Das Tool ANNIS unterstützt dies durch die Option, sich für eine Abfrage oder für einen Beleg einen Link generieren zu lassen, mit dem andere diese Abfrage nachvollziehen können — wie in diesem Beitrag gezeigt.

Fazit aus germanistisch-mediävistischer Sicht

Vielfach erfolgen Analysen auf Einzeltexte oder autor- bzw. gattungsbezogen, Ergebnisse und Aussagen basieren daher häufig auf einem – durchaus wohlbegründeten – Ausschnitt mhd. Literatur. Eine zusätzliche Korpusauswertung könnte entsprechende Ergebnisse innerhalb einer weitaus größeren Textlandschaft, die man vorsichtig als repräsentativ(er) bezeichnen könnte, verorten.

Auch bei einer spezifischen Fragestellung lohnt eine vergleichsweise unvoreingenommene Korpusauswertung, da die breite zeitliche, räumliche und inhaltliche Streuung in gewissem Maß verallgemeinernde Aussagen erlaubt.

Wenn man das Terrain der Befragung annotierter Korpora betritt, muss man allerdings den Weg ein gutes Stück weit als Ziel ansehen, da die Einarbeitung in die Abfragemodalitäten Zeit benötigt und nicht selten in Sackgassen führen kann. Weiterführend macht die philologische Perspektivierung der automatisch gewonnenen Ergebnisse den Großteil aus und erfordert ein hohes Maß manueller Nachbearbeitung, wenn z. B. eine Reihe von Belegstellen auf ihr kotextuelles Umfeld hin gelesen und möglicherweise zusätzlich ausgewertet werden muss. Die zwangsläufige Auseinandersetzung mit unterschiedlichen Textsorten erfordert eine Erarbeitung ihrer Genese und Zusatzkenntnisse einzelne Texte und Textzeugen betreffend – all dies muss zumeist (wieder) erarbeitet werden.

Der Mehrwert von korpusbasierten Analysen liegt zusammenfassend auf mehreren Ebenen:

- Korpusanalyse als Vergleichsgröße: Als Zusatz zu Einzeltextanalysen wird einerseits die Validität erhöht, andererseits können Ergebnisse im weiteren textkulturellen Umfeld verortet werden.
- Korpusanalyse als Überprüfung: Als Ausgangspunkt für Fragestellungen, denen bereits nachgegangen worden ist, bietet eine zusätzliche Korpusauswertung eine verlässliche Form für eine Verifizierung, Falsifizierung oder Modifikation von Ergebnissen und kann so möglicherweise bereits festgeschriebene Annahmen revidieren.
- Korpusanalyse als Methodenwechsel: Die Verfügbarkeit annotierter Korpora – sei es mit linguistischer oder inhaltlicher Schwerpunktsetzung – sollte zu neuen Fragestellungen, aber auch methodisch anderen Zugängen anregen, bspw. indem an einer allgemeinen, korpusweiten Analyse angesetzt wird und von dort aus das Markante zu Detailanalysen führt.
- Korpusanalyse als selbstverständlicher Einbezug sprachlicher Faktur: Die zwangsläufige Auseinandersetzung mit sprachlichen Entitäten führt deren Relevanz einmal mehr vor Augen und wird so zu einem festen Bestandteil literaturwissenschaftlicher Perspektivierungen.

Anmerkungen

1 Vgl. hierzu die auch korpuslinguistische Beiträge enthaltende Liste der Publikationen auf: www.linguistics.rub.de/rem/publications/index.html (letzter Aufruf: 19.06.2022).

2 Bei jedem Beispiel ist der »Korpuspfad« (»Path«) angegeben, der die genaue Fundstelle des Belegs angibt. Z. B. stammt der Beleg in (1a) aus dem Subkorpus »11-12_1-obd-PV-X« und entspricht den diplomatischen Tokens mit der Nummer 209-220 aus dem Text mit der Sigle »M010-N1«. Außerdem ist ein Link mit angegeben, der direkt zu ANNIS führt. Bei den Beispielen in (1)–(3) und (7) wird im Browser zunächst der Beleg im Kontext angezeigt, dazu gibt es einen Link »Show in ANNIS search interface«, mit dem man sich die betreffende Suchanfrage mit allen Treffern in ANNIS anzeigen lassen kann. Bei den

Beispielen in (4)-(6) gelangt man direkt zu den Anfragen in ANNIS. Die jeweiligen Suchanfragen können modifiziert und erneut ausgeführt werden.

3 Prädikative Adjektive (ADJD) stehen häufig nicht adjazent zum Bezugsnomen, so dass das Bezugsnomen nicht automatisch bestimmt werden kann. Dafür wäre eine syntaktische Annotation notwendig, die es in ReM aktuell nicht gibt.

4 Im Rahmen dieses Beitrags ist es nicht möglich, detaillierter auf die Funktionalitäten von ANNIS einzugehen. Eine ausführliche Dokumentation gibt es unter https://corpus-tools.org/annis/. In Dipper (2021, 2015) wird den Einsatz von ReM in ANNIS anhand von exemplarischen Fallstudien mit linguistischen Fragestellungen illustriert.

5 Die Form der Lemmata orientiert sich an Lexers Handwörterbuch, siehe die Dokumentation auf der ReM-Website: https://www.linguistics.ruhr-uni-bochum.de/rem/documentation/lemma.html.

6 Die betreffenden Token haben *sancte* und *sanctus* (322) als Lemma und sind hier zusammengefasst.

7 https://annis.linguistics.rub.de/?id=0673aaab-903e-4446-bd2f-b42e30557daa; Path: 13_1-obd-V-G > M321-G1 (tok_dipl 2950 - 2961).

8 https://annis.linguistics.rub.de/?id=0e1c0cea-8029-4a6e-a7a9-3db33b7fa862; Path: 12-13_1-mdnd-PV-X > M206-N1 (tok_dipl 25412 - 25423).

9 Diese Lemmata und ihre Frequenzen sind: *hêrre* (985), *got* (382), *mann* (267), *küni(n)g* (266), *mèister* (191), *vrouwe* (180), *vater* (104), *hèl(e)d* (93), sun (84), *küni(n)ginne* (79), *èngel* (75), *kind* (60).

10 So z. B. die taubenetzte Rose bei Wolfram von Eschenbach (in der Lyrik und Epik; Fuchs-Jolie 2004) bzw. weiterführend das Metaphernkonzept als Autornachweis für Wolfram (Fuchs-Jolie 2015) oder allgemein Metaphern und Metaphorik im ›Jüngeren Titurel‹ (Illibauer-Aichinger 2010). Gut im Blick sind auch bestimmte Metaphern, wie z. B. die Lichtmetaphorik (Cöllen 2012), die Metapher vom Wohnen im Herzen sowohl in der höfischen Epik als auch in religiösen Texten (Palmer 2005), die Metaphorik im Dreieck von Minne, Jagd und Tod in Wolframs von Eschenbach ›Titurel‹ (Kiening/Köbele 1998). Den Zusammenhang von Mythos und Metapher in Gottfrieds ›Tristan‹ beabeitet Köbele (2004). Die Metaphorik des Weges beleuchtet Friedrich (2014) ausgehend von der Metapherndefinition Aristoteles' in der Poetik.

Literaturverzeichnis

Sekundärliteratur

Cöllen, Sebastian: Minne und Metapher. Die Lichtmetaphorik Heinrichs von Morungen in kognitionslinguistischer Beleuchtung, in: Studia Neophilologica 84, 2 (2012), S. 201–220.

Dipper, Stefanie/Donhauser, Karin/Klein, Thomas/Linde, Sonja/Müller, Stefan/Wegera, Klaus-Peter: HiTS: ein Tagset für historische Sprachstufen des Deutschen, in: Journal for Language Technology and Computational Linguistics (JLCL) 28, 1 (2013), S. 85–137.

Dipper, Stefanie: Annotierte Korpora für die Historische Syntaxforschung: Anwendungsbeispiele anhand des Referenzkorpus Mittelhochdeutsch, in: Zeitschrift für Germanistische Linguistik 43, 3 (2015), S. 516–563.

Dipper, Stefanie: Das Referenzkorpus Mittelhochdeutsch: Nutzungsmöglichkeiten für morphologische Untersuchungen, in: Ganslmayer, Christine/Schwarz, Christian (Hrsg.): Historische Wortbildung. Theorie – Methoden – Perspektiven, Hildesheim/Zürich/New York 2021 (Germanistische Linguistik 252-254), S. 145–186.

Fauconnier, Gilles/Turner, Mark: The Way We Think. Conceptual Blending and the Mind's Hidden Complexities, New York 2002.

Friedrich, Udo: Erzähltes Leben - Zur Metaphorik und Diagrammatik des Weges, in: LiLi 176 (2014), S. 51–76.

Lakoff, George/Johnson, Mark: Metaphors We Live By. Chicago/London 1980.

Fuchs-Jolie, Stephan: *al naz von roete*. Visualisierung und Metapher in Wolframs Epik, in: Greenfield, John Thomas (Hrsg.): Wahrnehmung im ›Parzival‹ Wolframs von Eschenbach. Actas do Coloquio Internacional 2002, Porto 2004, S. 243–278.

Fuchs-Jolie, Stephan: Metapher und Metonymie bei Wolfram. Überlegungen zum ›Personalstil‹ im Mittelalter, in: Andersen, Elizabeth/Bauschke-Hartung, Ricarda/McLelland, Nicola/Reuvekamp, Silvia (Hrsg.): Literarischer Stil. Mittelalterliche Dichtung zwischen Konvention und Innovation. XXII. Anglo-German Colloquium Düsseldorf, Berlin/Boston 2015, S. 413–425.

Illibauer-Aichinger, Sandra: Metaphern und Metaphorik im ›Jüngeren Titurel‹, in: Baisch, Martin/Keller, Johannes/Kragl, Florian/Meyer, Matthias (Hrsg.): Der ›Jüngere Titurel‹ zwischen Didaxe und Verwilderung. Neue Beiträge zu einem schwierigen Werk, Göttingen 2010 (Aventiuren 6), S. 87–101.

Kiening, Christian/Köbele, Susanne: Wilde Minne. Metapher und Erzählwelt in Wolframs ›Titurel‹, in: PBB 120 (1998), S. 234–265.

Köbele, Susanne: Mythos und Metapher. Die Kunst der Anspielung in Gottfrieds ›Tristan‹, in: Friedrich, Udo/Quast, Bruno (Hrsg.): Präsenz des Mythos. Konfigurationen einer Denkform in Mittelalter und Früher Neuzeit, Berlin 2004, S. 218–246.

Krause, Thomas/Zeldes, Amir**:** ANNIS3: A new architecture for generic corpus query and visualization, in: Digital Scholarship in the Humanities 31 (2016), S. 118–139 (online).

Palmer, Nigel F.: Herzeliebe, weltlich und geistlich. Zur Metaphorik vom ›Einwohnen im Herzen‹ bei Wolfram von Eschenbach, Juliana von Cornillon, Hugo von Langenstein und Gertrud von Helfta, in: Hasebrink, Burkhard/Schiewer, Hans-Joachim/Suerbaum, Almut/Volfing, Annette (Hrsg.): Innenräume in der Literatur des deutschen Mittelalters. XIX. Anglo-German Colloquium Oxford 2005, Tübingen 2008, S. 197–224.

Schultz-Balluff, Simone: *triuwe* – Verwendungsweisen und semantischer Gehalt im Mittelhochdeutschen, in: Krieger, Gerhard (Hrsg.): Verwandtschaft, Freundschaft, Bruderschaft. Soziale Lebens- und Kommunikationsformen im Mittelalter. Akten des 12. Symposiums des Mediävistenverbandes 2007 Trier, Berlin 2009, S. 271–294.

Schultz-Balluff, Simone: Synergetisierung von Frame-Semantik und mediävistischer Literaturwissenschaft: Theoretische und methodische Überlegungen am Beispiel von Treue-Konzeptionen in mhd. Texten, in: PBB 136 (2014), S. 374–414.

Schultz-Balluff, Simone: Wissenswelt *triuwe*: Kollokationen – Semantisierung – Konzeptualisierung, Heidelberg 2018 (Germanistische Bibliothek 59).

Schultz-Balluff, Simone: Das Wissen über Wunden – Verwendungsweisen, Semantisierung und Konzeptualisierung von ahd. *wunti*/ as. *wunda*/ mhd. *wunde*, in: Bowden, Sarah/Miedema, Nine/Mossman, Stephen (Hrsg.): Verletzungen und Unversehrtheit in der deutschen Literatur des Mittelalters: XXIV. Anglo-German Colloquium 2015 Saarbrücken, Tübingen 2020, S. 37–66.

Warnke, Ingo H./Spitzmüller, Jürgen: Methoden und Methodologie der Diskurslinguistik – Grundlagen und Verfahren einer Sprachwissenschaft jenseits textueller Grenzen, in: Dies. (Hrsg.): Methoden der Diskurslinguistik. Sprachwissenschaftliche Zugänge zur transtextuellen Ebene, Berlin/New York 2008 (Linguistik – Impulse & Tendenzen 31), S. 3–54.

Online-Ressourcen

ANNIS (Annotation of Information Structure): https://corpus-tools.org/annis/.

ddd (Deutsch Diachron Digital. Referenzkorpora zur deutschen Sprachgeschichte): https://www.deutschdiachrondigital.de/.

ReM (Referenzkorpus Mittelhochdeutsch 1050–1350):
https://www.linguistics.rub.de/rem/.

Anschrift der Autorinnen:

Prof. Dr. Stefanie Dipper
Ruhr-Universität Bochum
Sprachwissenschaftliches Institut
Universitätsstraße 150
44801 Bochum
E-Mail: stefanie.dipper@rub.de

Prof. Dr. Simone Schultz-Balluff
Martin-Luther-Universität Halle-Wittenberg
Ludwig-Wucherer-Straße 2
06108 Halle (Saale)
E-Mail: simone.schultz-balluff@germanistik.uni-halle.de

Die Entstehung dieses Beitrags wurde teilweise gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – SFB 1475 – Projektnummer 441126958.

Katharina Zeppezauer-Wachauer

50 Jahre Mittelhochdeutsche Begriffsdatenbank (MHDBDB)

Eine Jubiläums-Zeitreise zwischen Lochkarten, Pixel-Drachen, relationaler Datenbank und Graphdaten

Abstract. Die seit 1972 betriebene Mittelhochdeutsche Begriffsdatenbank (MHDBDB) der Universität Salzburg ist ein umfassendes Recherchetool für das Mittelhochdeutsche. Kernelemente sind eine komplexe Suchmaschine und ein Wörterbuch, in dem mittels eines Begriffssystems Bedeutungen von korpusbasierten Wortartikeln erschlossen werden. Die einzelnen Vorkommensformen im Korpus werden auf diese Wortartikel und die im Kontext gültige Bedeutung sowie auf weitere Annotationsebenen wie beispielsweise grammatikalische Daten oder Metadaten bezogen. User können somit nicht nur nach Wörtern, Zeichenketten und Begriffen suchen, sondern auch wesentlich komplexere Fragestellungen auswerten. Dieser Beitrag liefert eine Rückschau über die letzten 50 Jahre sowie einen Ausblick auf zukünftige Entwicklungen.

Die Mittelhochdeutsche Begriffsdatenbank (MHDBDB) der Universität Salzburg ist seit den frühen 1970er-Jahren in Betrieb und inzwischen zu einem unerlässlichen Forschungswerkzeug der Fachcommunity geworden, wie erfreulicherweise auch dieser Sammelband zeigt. Mehrere im vorliegenden Band vorgestellten Digital-Humanities-Projekte arbeiten mit dem MHDBDB-Korpus: Dimpel [u. a.]; Brandes [u. a.]; Viehhauser. Der Dank ergeht an alle Kolleginnen und Kollegen, die das Feld um neue, lebendige Forschungsfragen und -methoden erweitern.

Als umfassendes Recherchetool für das Mittelhochdeutsche ist die MHDBDB auf der Grundlage digitaler Editionen und E-Texte aufgebaut. Ihr Korpus umfasst derzeit über 10,6 Millionen Tokens in fast 29.000 Wortartikeln und 666 Werken, wird jedoch beständig erweitert. Jährlich verzeichnet die Datenbank rund 28.000 individuelle User aus zahlreichen Ländern.

Kernelemente sind eine komplexe Suchmaschine und ein onomasiologisches Wörterbuch, in dem mittels eines Begriffssystems Bedeutungen von korpusbasierten Lemmata bzw. Wortartikeln erschlossen werden. Die einzelnen Vorkommensformen im Korpus (Tokens) werden auf diese Lemmata und die im Kontext gültige Bedeutung sowie auf weitere Annotationsebenen wie beispielsweise grammatikalische Daten oder Metadaten (s. u.) bezogen. User können somit nicht nur nach Wörtern, Zeichenketten und Begriffen suchen, sondern auch wesentlich komplexere Fragestellungen auswerten, etwa zur linguistischen, onomasiologischen, grammatikalischen, sprachhistorischen, Wortschatz- oder autorspezifischen Forschung.

1. MHDBDB 1.0: Lochkarten

Die Anfänge der MHDBDB liegen im Jahr 1972, als Klaus M. Schmidt mit seinen Vorarbeiten zum Begriffswörterbuch der mittelhochdeutschen Literatur, einer Dissertation zur computergestützten Analyse des Wortmaterials Ulrichs von Liechtenstein, an der Universität Michigan (Ann Arbor) promoviert (vgl. Schmidt 1972; Schmidt 1978). Daher stammt auch das Logo der MHDBDB: die Helmzier Ulrichs im Codex Manesse (cpg 848, fol. 237^{r}). Aus dieser Dissertation entwickelten sich die ersten elektronischen MHDBDB-Grundlagen an der Bowling Green State University (Ohio, USA) als so genanntes »maschinengestütztes Begriffswörterbuch«. Damals gab es noch keine Computer-Monitore, keine Tastaturen und Mäuse. Die Ein- und auch Ausgabe der Annotationen für das mhd. Begriffswörterbuch erfolgte mit Lochkartenstanzern über Fortran-Lochkar-

ten. Drucker bzw. Fernschreiber gaben die erfassten Daten, die auf den Karten gespeichert waren, aus:

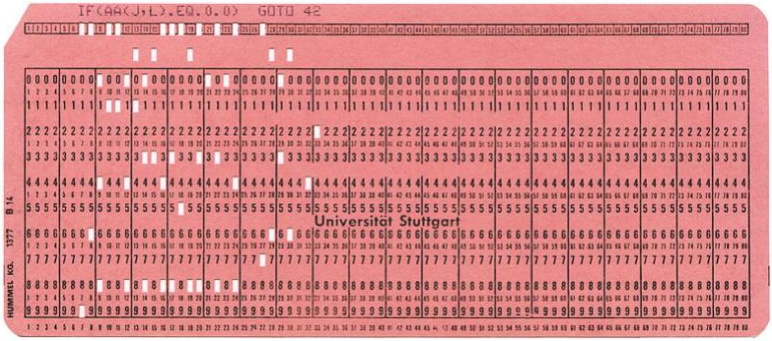


Abb. 1: Punch card Fortran Uni Stuttgart, Wikimedia Commons (CC BY-SA 3.0)

2. MHDBDB 2.0: MHDBDB goes online

Die nächste Großetappe für die MHDBDB lag in den 1990er Jahren: 1992 schlossen sich zwei Langzeitprojekte unter dem Namen »MittelHochDeutsche Begriffs-DatenBank« zusammen: Das Projekt »Namen in deutschen literarischen Texten des Mittelalters« von Horst P. Pütz (Christian-Albrechts-Universität zu Kiel) (vgl. Debus/Pütz 1987; darin insbes. Pütz, S. 287–299) und das erwähnte »Begriffswörterbuch der mittelhochdeutschen Literatur« von Schmidt. Pütz integrierte 1995 über 100 mittelhochdeutsche Texte und Namens-Annotationen und ermöglichte Schmidts Begriffswörterbuch dadurch einen gewaltigen Entwicklungsschub (vgl. Pütz/Schmidt 2001).

Auch das etwas sperrige, wenngleich mittlerweile weithin bekannte Akronym »MHDBDB« stammt aus dieser Zeit. Damals gab es allerdings auch noch eine englische Variante, die »MHGCDB« (»Middle High German Conceptual DataBase«).

Anlässlich des International Congress on Medieval Studies der Western Michigan University in Kalamazoo wurde die MHDBDB schließlich als menügesteuerte und in ihren Nutzungsmöglichkeiten noch limitierte Datenbank im Mai 1995 erstmalig über das Internet verfügbar gemacht (mhdbdb.bgsu.edu). Damit entstand ein weltweit zugängliches, einzigartiges Informationssystem zur mittelhochdeutschen Sprache und Literatur. Dank der Wayback Machine (archive.org) gibt es sogar noch Screenshots des typischen 90er-Jahre-Webdesigns, das auch eine text only-Version für den Fall anbot, dass die Nutzer*innen über keine leistungsfähige Grafikkarte verfügten:

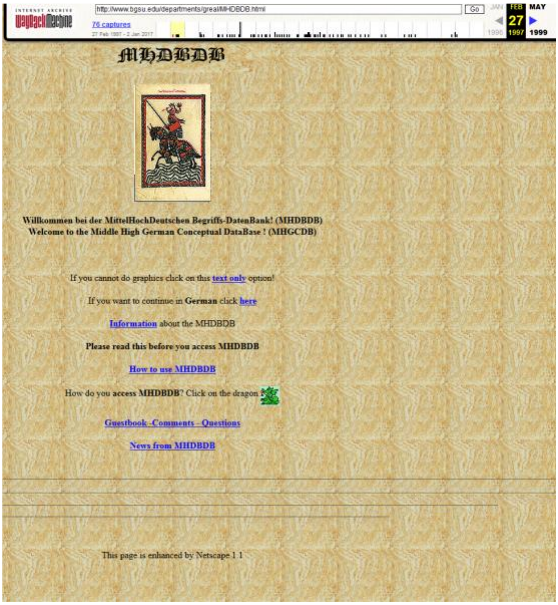


Abb. 2 + Abb. 3: Archivbild und Detail der MHDBDB-Website vom 27.02.1997, https://web.archive.org/web/19970227152339/http://www.bgsu.edu/departments/greal/MHDBDB.html (Link nicht mehr funktionsfähig, zuletzt abgerufen und gesichert am 07.02.2018)

Im Herbst 1998 wurde dieses System auf eine relationale Datenbank von ORACLE übertragen und eine neue Benutzeroberfläche auf der Basis von Websites im heutigen Verständnis erstellt.

3. Die MHDBDB im 21. Jahrhundert: Umzug von Ann Arbor nach Salzburg

Zwischen 2002 und 2004 fanden Verhandlungen mit der Paris Lodron-Universität Salzburg statt. 2004 wurde die MHDBDB schließlich in Salzburg installiert, ein kleiner Mitarbeiterstab (1,5 Stellen) zugesichert und von dort betrieben. Die Optik der Website änderte sich noch einmal geringfügig:

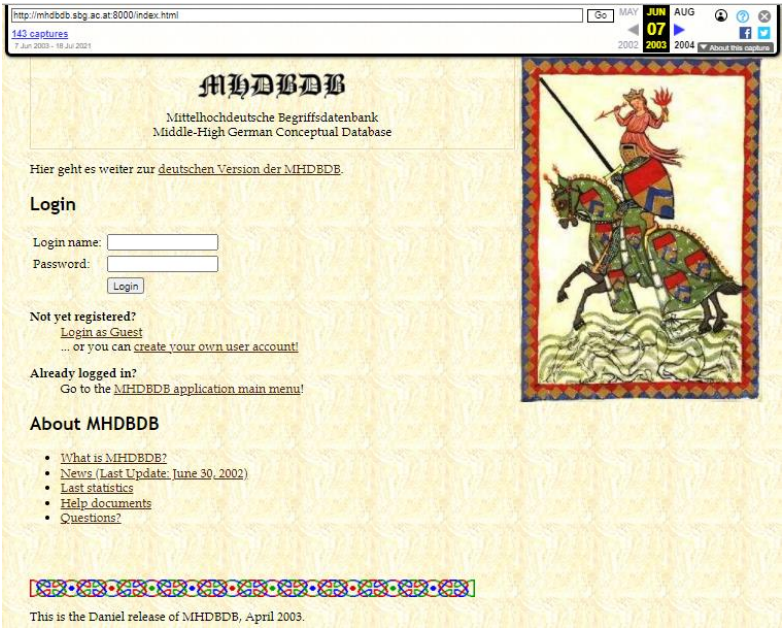


Abb. 4: Archivbild der MHDBDB-Website vom 07.06.2003, https://web.archive.org/web/20030607163507/http:/mhdbdb.sbg.ac.at:8000/index.html (zuletzt abgerufen und gesichert am 26.01.2022)

Das Rätsel um den »Daniel release« (unten am Bild) lässt sich über die alte Seitendoku aufklären, die für ein DH-Projekt aus dieser Zeit überraschend sorgfältig durchgeführt wurde. Die Releases waren die Versionennamen der Produktion, wobei »Daniel« noch vergleichsweise unspektakulär war, es gab etwa auch die Versionen »Craun«, »Erec«, »Wankelhart« oder »Fortimar«.

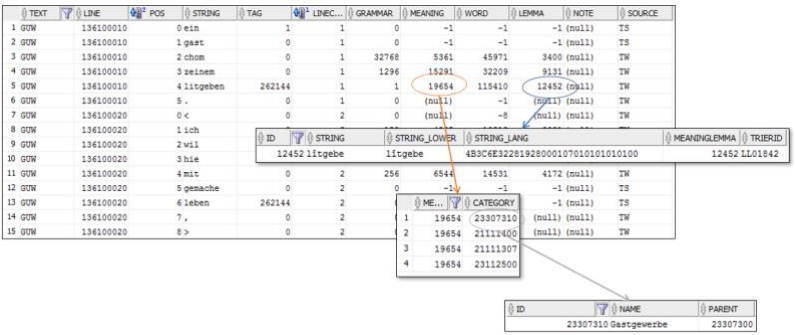


Abb. 5: Tabellenstruktur der relationalen Datenbank

Die folgende Zeit kann immerhin als direkter Augenzeugenbericht verfasst werden, denn meine erste Tätigkeit für die MHDBDB unter der ehemaligen Koordinatorin, Margarete Springeth, und dem sich bis heute als Berater und Mitarbeiter engagierenden Direktor, Klaus M. Schmidt, datiert ins Jahr 2010. 2012 bemühte ich mich zunächst – unter den gegebenen Bedingungen und mit de facto nicht vorhandenen finanziellen Mitteln – um ein kleineres optisches Update. Die Helmzier Ulrichs wurde von seinem Kopf gelöst und in ein Logo verwandelt, der Pixelhintergrund gegen ein *seamless* Pattern getauscht, die knallbunten, mit der Maus gezeichneten *divider* und andere visuelle Spuren der 1990er wurden peu à peu ausgetauscht. Auch inhaltlich kam es in den 2010ern neben der steten Erweiterung des Korpus zu zahlreichen Neuerungen wie beispielsweise in der Usability (Tooltip-Funktion bei der Textsuche, neue Filter, geschlechterspezifische Annotationen, breite Wortfeld-Annotationen, Neukonzeption des Begriffssystems, Ausgabemöglichkeit des text-/autorspezifischen Wortschatzes – Hapax-

legomena aka »Unique-Words-Tool«) sowie einem verstärkten Bemühen, die breiten Funktionalitäten der MHDBDB in der Fachcommunity zu verankern, etwa über implementierte Mini-Tutorials, themenspezifische Newsletter, Support über Social Media etc.

Darüber hinaus gab es zahlreiche Datenvernetzungen mit größeren Partnerprojekten: dem Trierer »Wörterbuchnetz«, dem Grazer »Portal der Pflanzen des Mittelalters«, dem italienisch-europäischen Poesie-Lexikon »Lirica Europea di TrobVers e MHDBDB« (vgl. Distilo 2013), der Kremser Bilddatenbank »REALonline«, der Oswald von Wolkenstein-Gesellschaft (Digitalisierung der Oswald-Ausgabe von K. K. Klein, siehe Literaturverzeichnis) sowie den diversen Editionsprojekten von Vlastimil Brom (Masaryk Universität Brünn), assoziiertem MHDBDB-Partner und Experte für deutschsprachige Texte in slawischer Umgebung (vgl. Brom 2019).[1]

Vom Institut für Literaturwissenschaft und dem Institut für Maschinelle Sprachverarbeitung der Universität Stuttgart wurde auf Basis der damals 10 Millionen MHDBDB-Tokens eine Trainingsdatei für einen Tree-Tagger des Mittelhochdeutschen erstellt (vgl. Ketschik [geb. Echelmeyer] [u. a.] 2017). Dabei handelt es sich um ein Werkzeug zur Annotation von Texten mit Part-of-Speech (PoS) und Lemma-Informationen, also der automatischen Zuordnung von Wörtern zu Wortarten.

Abb. 6: Logo der MHDBDB 2012–2022

4. MHDBDB 3.0: Die Reise geht weiter

Auf manche unserer Nutzer*innen mag es aufgrund des aktuellen Status quo der Website so wirken, als wäre die Entwicklung der MHDBDB seitdem zum Erliegen gekommen. Auf der dem Band zugrunde liegenden DH-Tagung prägte Joachim Hamm das treffende Bonmot der »Digitalen Editionsruine« (Gott sei Dank nicht auf die MHDBDB bezogen), das sich auf den ersten Blick auch für die Datenbank anbieten könnte. In Wahrheit ist jedoch das Gegenteil der Fall.

Die Verwaltung sämtlicher Daten (E-Editionen, Begriffssystem, Annotationen, Userdaten) findet seit 1998 in der relationalen ORACLE-Umgebung[2] statt; die Funktionen sind mit der objektorientierten Programmiersprache Java realisiert. War diese Technik 1998 noch auf dem neuesten Stand, so zeichnen sich inzwischen deutliche Mängel ab: Eine Vernetzung mit anderen Ressourcen, etwa in Form von Linked Open Data über das Semantic Web, ist, ebenso wie der unkomplizierte Import von Metadaten, kaum möglich. Metadaten sind im Grunde ›Daten über Daten‹ und beinhalten deskriptive, strukturelle, administrative oder technische Informationen zu einem Datensatz. Vor allem deskriptive Metadaten sind deshalb von großer Bedeutung, weil sie Daten auffindbar und nachnutzbar machen. Darüber hinaus können unter Metadaten auch Annotationen verstanden werden, die sich aber meist auf Teile eines Datensatzes beziehen, in Texten etwa auf einzelne Absätze oder Wörter, und deren Eigenschaften näher beschreiben (vgl. Schöch 2017, S. 223–233; Steiner 2021).

Aus den genannten Gründen des technischen Handicaps werden hinter den Kulissen bereits seit 2017 Vorkehrungen für einen Relaunch getroffen, für den bis heute weder die dafür benötigten finanziellen noch personellen Ressourcen zur Verfügung gestellt werden. Temporäre Hoffnung gab die Entwicklung eines Repositoriums der Universität Salzburg, ein Vorhaben, das aus Gründen leider wieder zum Erliegen gekommen ist und die avisierten Pilotprojekte ohne finales Ergebnis zurücklässt. Diese Umstände haben

dafür gesorgt, dass es leider nicht besonders wahrscheinlich ist, dass zum 50-jährigen Jubiläum der MHDBDB mit Ende des Jahres 2022 tatsächlich schon eine Beta-Version der komplett neu strukturierten, migrierten (Graph-)Datenbank[3], an der seit Jahren gearbeitet wird, erscheinen kann – doch die Chancen stehen gut für 2023. Im Folgenden soll ein Ausblick über die neue Datenlage gegeben werden, die bereits heute vorliegt und nur darauf wartet, in ein funktionales Frontend eingebettet zu werden.

Abb. 7: Neues Logo der MHDBDB

Die wichtigsten Neuerungen sind die Nutzung von Standards, genormte Datenformate (insbesondere XML-TEI, RDF, SKOS, OWL), eine Anbindung ans Semantic Web[4] und die Vernetzung mit Normdaten. Normdaten (engl. *authority files*), z. B. Wikidata und die GND, sind kontrollierte Vokabulare für bestimmte Domänen. In einer Normdatei hat eine Entität einen eindeutig referenzierbaren Eintrag, der mit weiterer spezifizierender Information ausgestattet sein kann. Solche Normdaten werden zukünftig auch von der MHDBDB nachgenutzt, damit User noch bessere Zugriffsmöglichkeiten auf das vorhandene Datenmaterial haben. Die Vernetzung mit Metadatenrepositorien ermöglicht das wechselseitige Anreichern der Daten (vgl. Steiner/Fritze 2021).

Beispiellos wird die neue Open-Access-Policy auf der Basis der FAIR-Prinzipien (*findable*, *accessible*, *interoperable* und *re-usable*)[5] sein: Das MHDBDB-Volltextkorpus kann in Zukunft nicht nur vollständig (bisher nur in Auszügen in der Textsuche) gelesen, sondern zu weiten Teilen auch

als reiner Lesetext (PDF) oder Arbeitsversion (in diversen Datenformaten) heruntergeladen werden. Ebenso werden sämtliche Annotationen wie Begriffssystem (Semantik, Onomasiologie), Namenssystem (Onomastik), PoS/Grammatik, Wort- und Satzgrenzen, Phrasen- und Satzstrukturen, Wortfelder, Metadaten, Verknüpfungen zu Normdaten u. v. m. unter einer Creative Commons-Lizenz (voraussichtlich CC BY-NC-SA 3.0 AT) zur Verfügung gestellt (vgl. Hinkelmanns/Zeppezauer-Wachauer 2020; Hinkelmanns/Zeppezauer-Wachauer 2021).

Die neue MHDBDB setzt auf ein Datenmodell basierend auf unterschiedlichen Semantic-Web-Technologien wie RDF (*Resource Description Framework*)-Vokabulare und Ontologien[6] sowie auf TEI (*Text Encoding Initiative*) zur Kodierung der Texte des Korpus (vgl. Pollin 2021b; Eibinger 2021). Sämtliche 666 E-Texte wurden bereits zur Gänze in das XML-TEI-Format konvertiert, das heute einen de facto-Standard in den Geisteswissenschaften darstellt. Die TEI-Texte wurden im Stand-off-Verfahren mit Linked Open Data verknüpft. Alle weiteren Forschungsdaten wie Annotationen sowie die bibliographischen Metadaten und die deskriptiven Metadaten zu Personen, Zeit, Orten und Ereignissen (basierend auf dem *Conceptual Reference Model* CIDOC-CRM für Kulturerbe-Daten, vgl. Pollin 2021a) liegen inzwischen als RDF-Daten vor. Sie werden direkt auf die Tokens der E-Texte bezogen. Die Vernetzung zwischen RDF- und TEI-Daten erfolgt mittels *Web Annotation Vocabulary* nach der Empfehlung des W3-Konsortiums.[7]

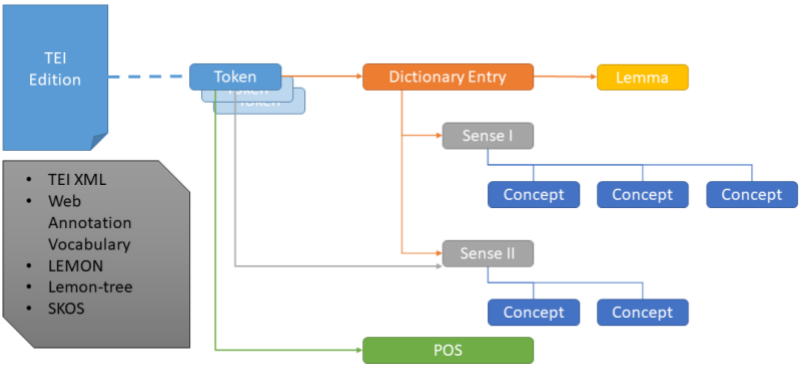


Abb. 8: Vereinfachtes Datenmodell der neuen MHDBDB

Die neue MHDBDB setzt durch die Verwendung kontrollierter Vokabularien und Ontologien zukünftig in sämtlichen Bereichen auf Nachhaltigkeit, um stabile Referenzen für Ressourcen sowie Interoperabilität zu gewährleisten. Kontrollierte Vokabulare und Normdaten helfen bei der Schematisierung von Inhalten geisteswissenschaftlicher Daten.

Die Wortartikel (Lemmata) wurden bereits nach den Vorgaben des *OntoLex-Lemon-Lexicography-Modules* kodiert. Weitere schon genutzte Ontologien bzw. Vokabulare sind *BibFrame* 2.0, die GND *Ontology* und eine im Rahmen unseres von der Österreichischen Akademie der Wissenschaften geförderten »Go! digital«-Projektes erstellte Semantic-Web-Ontologie für die Analyse narratologischer Muster in der Literatur und in Bildern des Mittelalters: ONAMA (*Ontology of Narratives of the Middle Ages*). Diese Ontologie bietet ein Modell für die medienübergreifende Beschreibung von Handlungen, Akteur*innen, Schauplätzen und zeitlichen Strukturen. Neben den konstituierenden Grundelementen transmedialer Erzählkerne erfassen diese Beschreibungen die jeweiligen textlichen und bildlichen Umsetzungen im Detail und gehen damit über die reine Handlung eines Erzähltexts oder Bilderzyklus hinaus. ONAMA erlaubt es, Muster und Besonderheiten in den Strukturen von Erzählungen zu identi-

fizieren, die dann auf ihre zugrundeliegenden Ursachen und Funktionen hin untersucht werden können. Mit ONAMA können Bild- und Textquellen so annotiert werden, dass die generierten Daten Aufschluss über die Genese und Tradierung von Erzählkernen, Figurenkonstellationen, Handlungsmustern etc. im jeweiligen Medium und in medienübergreifenden Zusammenhängen geben (vgl. Nicka [u. a] 2020; Zeppezauer-Wachauer [u. a.] 2021; Hinkelmanns [u. a] 2022; http://onama.sbg.ac.at/ontology/).

Das namensgebende Begriffssystem der MHDBDB war durch klassische, nach Sachgruppen strukturierte Thesauri inspiriert. Die Konzeption erfolgte in der Nachfolge von Rogets Thesaurus, als Weiterentwicklung der Dornseiff'schen Kategorisierung sowie vor allem in Anlehnung und Erweiterung des von Hallig und Wartburg entworfenen Begriffssystems (vgl. Roget 1852; Dornseiff 1970; Hallig/Wartburg 1963; Schmidt 1988, S. 40; Schmidt 1993, S. VII–X; Springeth 2009, S. 194).[8] Die korpusorientierte Anforderung an das System hat über die Jahre dazu geführt, dass es um zahlreiche Kategorien erweitert wurde und neben Autosemantika auch weitere Funktionalitäten integriert wurden, etwa Synsematika bzw. Funktionswörter, in den mhd. Texten vorkommende Fremdsprachen (Lateinisch, Französisch, Arabisch, Hebräisch, Griechisch, Slawisch, Tschechisch, Persisch, Italienisch, Spanisch, Ungarisch, Rätoromanisch, Englisch, Phönizisch-Punisch, Aramäisch, Jiddisch, Rotwelsch) oder eine Art wegbereitende Frühform von *Named-entity recognition* (NER). Das Design war konventionell hierarchisch und bei aller Nützlichkeit in seiner technischen Funktionalität sowie auch der Anwendungsfreundlichkeit limitiert. Usern, die das Begriffssystem zu schätzen und nutzen wussten, musste auch eine gewisse ›Leidensfähigkeit‹ attestiert werden, da die Arbeit mit Ziffern und Zahlencodes nicht besonders intuitiv war und eine längere Einarbeitung benötigte.

Aus diesem Grund wurde das Begriffssystem mittlerweile in einen polyhierarchischen SKOS-Thesaurus[9] überführt, der es erlaubt, einzelnen semantischen Begriffen mehrere *parents* und *children* gleichzeitig zuzu-

ordnen, um standardisierte Notizen zu ergänzen und zudem verschiedene Labels (Synonyme) für die Begriffe zu vergeben. Die Anwendung erfolgt zukünftig über eine visuelle Suchmaske und via SPARQL (*SPARQL Protocol And RDF Query Language,* einer Abfragesprache für Graphdatenbanken), man benötigt also keine Zifferncodes mehr. Der weitaus größte Teil des Begriffssystems blieb bei diesen Arbeiten erhalten und wurde um vielfache standardisierte Relationen untereinander (*skos:broader*; *skos:narrower*; *skos:broadMatch*; *skos:narrowMatch*; *skos:related Match*; *skos:closeMatch*; *skos:exactMatch*) bzw. weitere Informationen (die Sprachen/Language-Tags Deutsch, Englisch, Französisch und Latein; *skos:altLabel*; *skos:editorialNote*; *skos:definition*) angereichert. Konzepte, die keine semantische Annotationsebene i. e. S. darstellen (z. B. Fremdsprachen, neuhochdeutsche Kommentare, Abkürzungen) wurden zusätzlich mit standardisierten Beschreibungen ausgezeichnet, etwa in TEI oder als Verknüpfung mit einschlägigen kontrollierten Vokabularien.

Besonders viel Arbeit verursachte die Neukonzeption der Namen: Die bisherige Namenssystematik, die das Vermächtnis des 1992 integrierten Namensprojektes von Pütz darstellt und überaus umfassend ist, wurde aus dem onomasiologischen Begriffssystem herausgelöst und zu einem komplexen Namensverzeichnis (Onomastikon) auf SKOS-Basis, mit bidirektional gerichteten Relationen ins Begriffssystem, umgestaltet. Es wird somit sowohl optional integrativ über das Begriffssystem oder aber separat über ein reines Namenslexikon durchsuchbar sein. Eine weitere Baustelle wird es sein, die frühe Form von NER, die bisher über das Begriffssystem realisiert wurde, zu professionalisieren und in heutige technische Standards zu überführen. Bislang wurde beispielsweise jedes Token ›Walther‹ über den Wortindex mit dem Begriffskonzept ›Literatur/Namen‹ verknüpft, um Walther von der Vogelweide als Personenentität in Texten zu identifizieren. Dies war in den 1990ern revolutionär; heutzutage gibt es mit Semantic Web und Normdaten aber andere Möglichkeiten.

Vergleichbares kann über Metadaten wie Dichterbiografien, die in der alten MHDBDB aufwändig als *plain text*-Beschreibungstexte angelegt sind, über bibliografische Metadaten zu den zugrundeliegenden Editionen und Handschriften oder über Gattungen/Genres gesagt werden. Beispielsweise die bisherigen Texttypen werden als kollaboratives Joint Venture verschiedener Netzwerke und Arbeitsgruppen in eine vollkommen neue Gattungstypologie umgebaut, die ebenfalls in SKOS beschrieben wird. Die Arbeiten an einer ersten Beta-Version werden in den nächsten Monaten abgeschlossen, und eine Publikation der Typologie auf GitHub, einem netzbasierten Dienst für das Teilen von Software-Entwicklungsprojekten, wird angestrebt. Für diese neue MHDBDB-Gattungstypologie wird sowohl auf Knowhow vor Ort zurückgegriffen[10] als auch auf jenes internationaler Wissensverbünde und Expert*innen, beispielsweise der »Brevitas – Gesellschaft zur Erforschung vormoderner Kleinepik«[11] (insbesondere der »Brevitas Systematik Kleinepik«); des »Netzwerk Offenes Mittelalter« zu Linked Open Data in der deutschsprachigen Mediävistik (vgl. hierzu auch Borek/Zeppezauer-Wachauer [u. a.] 2022) sowie zuvorderst von Marco Heiles (RWTH Aachen), dessen Mitarbeit an der Typologie im Bereich der Kategorisierung von Wissens- und Gebrauchsliteratur zentral ist. Das »Netzwerk Historische Wissens- und Gebrauchsliteratur« (HWGL), in dem sowohl Heiles als auch ich Mitglieder sind, liefert auf gemeinsamen Tagungen und per Mailinglist ebenfalls wertvolle Impulse zu Anwenderfragen.

Die Forschungsdaten wurden und werden also auf den neuesten technischen (und damit einhergehend auch inhaltlichen) Stand gebracht. Inzwischen sind alle Texte voll lemmatisiert, die Lyrik und der größte Teil der Kurzprosa zusammen mit einigen Epen sogar disambiguiert. Diese Daten werden besonders hilfreich bei der Entwicklung eines lernfähigen Disambiguierungsmoduls sein. Freilich sind solche Arbeiten niemals zur Gänze abgeschlossen, jedoch darf der aktuelle Datenstand mittlerweile wieder als gut kuratiert, dokumentiert und nachhaltig konzipiert bezeichnet werden.

Anders gestaltet sich die Situation bei der Präsentation dieser Daten in einer Weboberfläche. Das völlig neue Webdesign wurde bereits mit eigenen Mitteln entworfen und weitgehend implementiert. Die ursprünglich vereinbarte Umsetzung des Frontend durch das IT-Center der Universität Salzburg ist aufgrund der Einstellung des universitären Repositoriums jedoch zum Erliegen gekommen und liegt nun – unerwartet – wieder in den Händen des MHDBDB-Teams, da keine personelle oder finanzielle Unterstützung vorgesehen ist. Aktuell nehmen sich dieser mühevollen Kleinarbeit die (ehem.) MHDBDB-Mitarbeiter Peter Hinkelmanns, Daniel Schlager und Peter Färberböck auf Basis von TypeScript/ Angular an. Die Aussichten sind vielversprechend, wenn auch aus den genannten Gründen noch etwas Geduld erforderlich sein muss: Nach Fertigstellung wird die rundumerneuerte Datenbank nicht nur optisch ansprechend und anwendungsfreundlicher, sondern auch responsiv sein. Sie kann dann auch auf mobilen Geräten wie Tablets oder Smartphones genutzt werden. Die graphenbasierte Abfragesprache SPARQL wird komplexe Recherchen ermöglichen; eine visuelle Suchmaske dient dazu, dass auch User, die keine SPARQL-Erfahrung haben, das mögliche Maximum für ihre Forschung herausholen können. Eine Nutzer*innen-spezifische Oberfläche wird nicht nur den lang gehegten User-Wunsch einer Speichermöglichkeit über die letzten Suchen bieten, sondern beispielsweise auch das Anlegen eigener Text-Sammlungen.

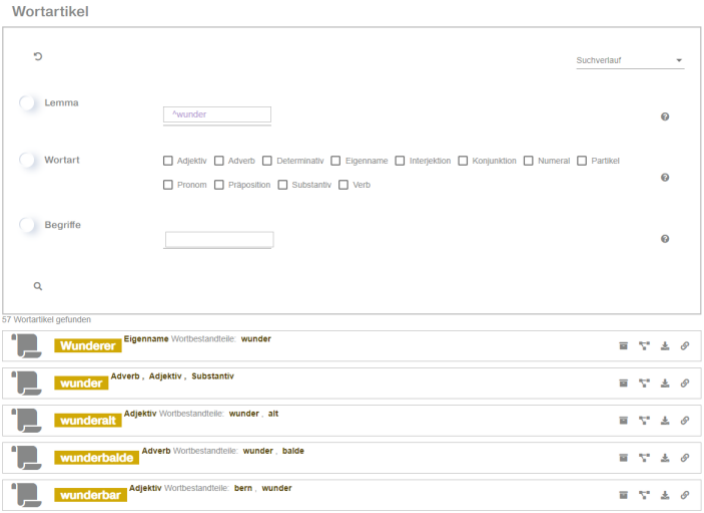


Abb. 9: *Sneak peek* Nr. 1 – Neues MHDBDB-Frontend (Wortartikel, die mit ›wunder-‹ beginnen)

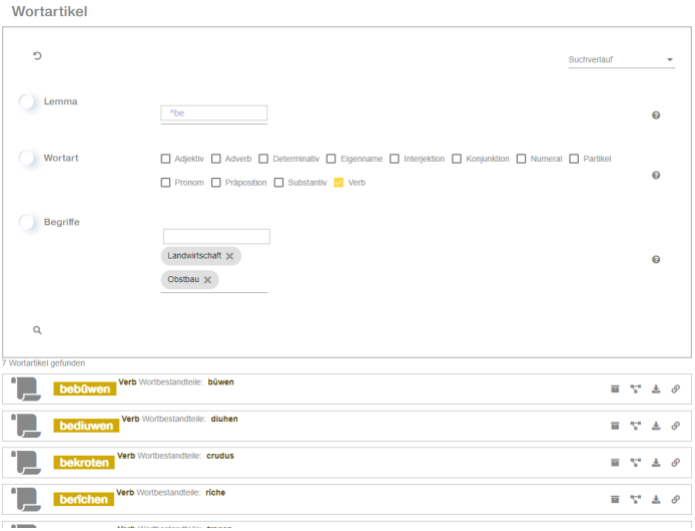


Abb. 10: *Sneak peek* Nr. 2 – Neues MHDBDB-Frontend (Wortsuche: Verba, die mit ›be-‹ beginnen und mit den Begriffen ›Landwirtschaft‹ sowie ›Obstbau‹ assoziiert sind)

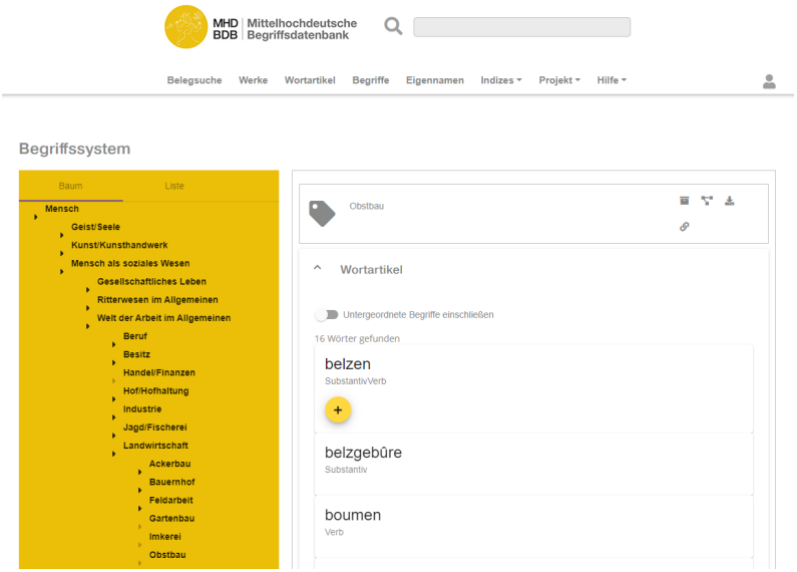
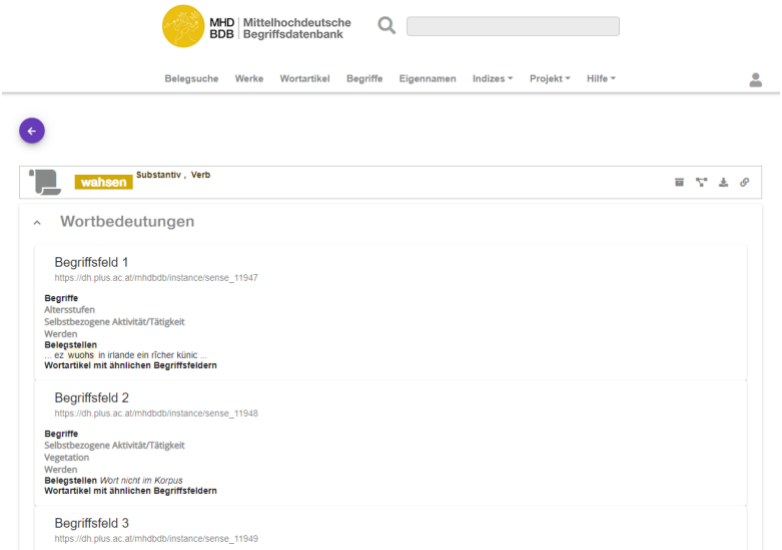


Abb. 11: *Sneak peek* Nr. 3 – Neues MHDBDB-Frontend (Begriffssystem, Beispiel ›Obstbau‹)



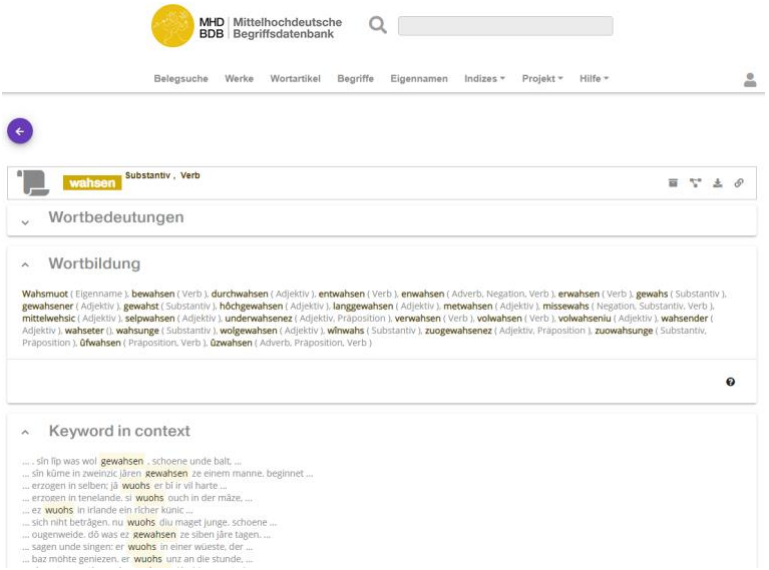


Abb. 12, Abb. 13 + Abb. 14: *Sneak peek* Nr. 4 – Neues MHDBDB-Frontend (Wortartikel ›*wahsen*‹)

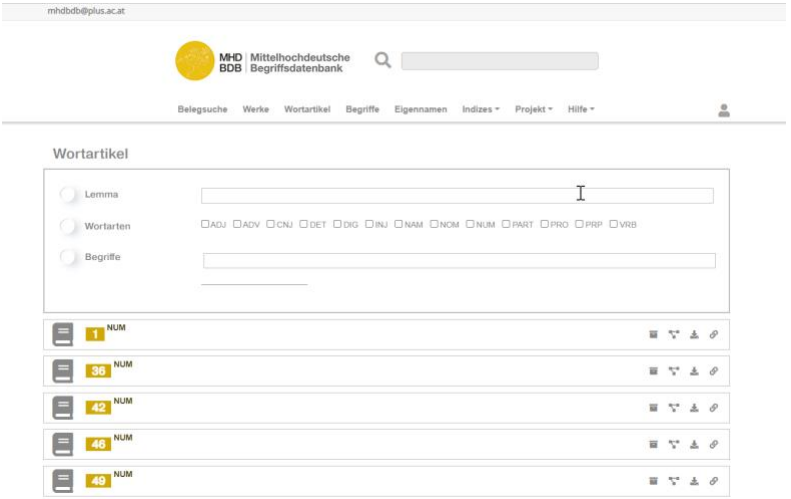


Abb. 15: *Sneak peek* Nr. 5 – Neues MHDBDB-Frontend (Wortsuche)

Weitere aktuelle Neuerungen finden sich auf der MHDBDB-GitHub-Seite, die beständig erweitert wird, so beispielsweise um eine Stoppwort-Liste für die quantitative Untersuchung mittelhochdeutscher Texte (vgl. Hinkelmanns 2021a), ein PoS-Tagger für das Mittelhochdeutsche für spaCy v3 (GitHub) (vgl. Hinkelmanns 2021b) und eine XQuery Module API für Transkribus PageXML (vgl. Hinkelmanns 2021d). Das XQuery-Modul ermöglicht den Zugriff auf Transkribus-PageXML-Dateien über XQuery-Funktionen.

Die kommende Zeit steht weiterhin im Zeichen von ›Nachhaltigkeit‹: Erste Priorität hat die Verwendung von maschinenlesbaren, standardisierten, offenen, dokumentierten und kuratierten Datenformaten und kontrollierten Vokabularien. Die Umsetzung von Back- und Frontend wird das kleine MHDBDB-Team noch eine Zeit lang gut beschäftigen. Die nächste große Herausforderung wird darin bestehen, endlich die technischen und auch organisatorischen Voraussetzungen zu schaffen, um die verfügbaren

Forschungsdaten dauerhaft in einem zertifizierten Repository zu speichern und verfügbar zu halten.

5. Epilog

Mit dem vorliegenden Jubiläumsbeitrag wollte ich eine umfassende Rückschau über 50 Jahre MHDBDB liefern, die zukünftigen Generationen wissenschaftstheoretische, methodisch-konzeptuelle, aber auch ganz persönliche Einblicke in die Entwicklung dieses Urgesteins der Digitalen Mediävistik geben sollte.

Wir Mediävist*innen sprechen gern in Jahrhunderten. 1972–2022: ein halbes Jahrhundert Mittelhochdeutsche Begriffsdatenbank. Eine glückliche Zeit der Entwicklung, des Fortschritts, Lernens, Erfolges und der Neugierde auf all die neuen Möglichkeiten, die das breite Digital-Humanities-Feld bietet. Es waren aber auch Jahre, die viel Flexibilität, organisatorische Kreativität, Resilienz, Geduld und Beharrlichkeit von uns allen forderten. Ohne eine so treue Fachgemeinschaft wäre vieles davon nicht möglich gewesen. Dieser Beitrag sei daher Ihnen und Euch allen gewidmet, die die MHDBDB regelmäßig nutzen, sie hervorheben und zitieren, sie konzeptuell befruchten und mit immer neuen Forschungsideen und -wünschen aufwarten.

Wir blicken den nächsten 50 aufregenden Jahren hoffnungsvoll entgegen.

Anmerkungen

1 Vgl. beispielsweise Brom 2019. Dort weiterführende Literatur von Vlastimil Brom zur Arbeit mit und an der MHDBDB. E-Texte in der MHDBDB ediert und ingestiert von Brom: ›Di tutsch kronik von Behem lant‹ und ›Pulkava Chronik‹.

2 Ein klassisches relationales Datenbankmodell basiert auf der Speicherung von Informationen in verschiedenen Tabellen, die untereinander über Relationen (Beziehungen) verknüpft sind, siehe Abb. 5. Das Entwicklungsunternehmen

ORACLE gehört gemeinsam mit Microsoft SQL und IBM DB zu den Marktführern bei Datenbankmanagementsystem-Software.

3 Eine Graphdatenbank nutzt statt tabellarischen Relationen Graphen, um vernetzte Informationen darzustellen und abzuspeichern. Die Beziehungen werden dabei mithilfe von so genannten Knoten (Objekte) und Kanten (Verbindungen) ausgedrückt.

4 »Mit dem *Semantic Web* wird das reguläre World Wide Web um semantische Informationen ergänzt, die einen globalen Informationsgraphen ergeben.« Zitat und genauere Erläuterung s. Hinkelmanns 2021c.

5 »Die FAIR *Guiding Principles for scientific data management and stewardship* wurden 2016 veröffentlicht und besagen, dass Forschungsdaten auffindbar (*Findable*), zugänglich (*Accessible*), interoperabel (*Interoperable*) und wiederverwendbar (*Re-usable*) sein sollen. Sie bilden die Grundlage für eine disziplinen- und länderübergreifende Nachnutzung von Forschungsdaten.« (Stigler 2021).

6 »In der Informationswissenschaft und den Digital Humanities handelt es sich bei Ontologien um eine Form der Wissensrepräsentation. Man versucht Wissen aus der Welt oder thematisch kleiner gefassten Bereichen zu modellieren, abzubilden und zu formalisieren. Durch die Formalisierung wird es möglich, Daten und Wissen weiterzuverwenden, auszutauschen und sogar logische Schlussfolgerungen zu ziehen. […] Ontologien stehen auch sehr stark in Zusammenhang mit dem *Semantic Web*, da sie dabei helfen, Informationen explizit und weiterverarbeitbar zu machen.« (Galka 2021).

7 Das *Web Annotation Vocabulary* spezifiziert das Set von RDF-Klassen, Prädikaten und Entitäten, die vom *Web Annotation Data Model*, einem strukturierten Modell für die Nutzung und Wiederverwendung von Annotationen, verwendet werden.

8 Onomasiologie wird seit Jahrzehnten innerhalb der deutschsprachigen historischen Lexikografie diskutiert. Ralf Plate schrieb beispielsweise schon 1992 in »Onomasiologische Umkehrlexikographie auf dem Prüfstand«: »Obgleich die Umkehrlexikographie […] keine Erfindung des Computer-Zeitalters ist, hat ihr doch die Möglichkeit, die außerordentlich aufwendigen Umsortierungsprozeduren der Maschine zu überlassen, gesteigerte Aufmerksamkeit zuteil werden lassen. Dabei rückte eine Eigenschaft von Umkehrwörterbüchern ins Blickfeld, die zuvor kaum beachtet worden war: diejenigen Lemmata, die im Bedeutungswörterbuch an verschiedenen Stellen mit demselben Interpretament erläutert werden, erscheinen im Umkehrwörterbuch zusammen hinter diesem

jetzt als Stichwort auftretenden Interpretament. Das Umkehrwörterbuch läßt also Synonymreihen, Wortfelder der Objektsprache sichtbar werden und erfüllt so teilweise die Aufgabe eines onomasiologischen Begriffswörterbuchs. Onomasiologische Wörterbücher aber, unentbehrliche Hilfsmittel für Untersuchungen z.B. zur Stilistik, zu systematischen inhaltlichen Wortschatzbeziehungen oder zum sprachgeschichtlichen Bezeichnungswandel, sind in der historischen Lexikographie des Deutschen eines der größten Desiderate: ein ›Dornseiff‹ oder ›Wehrle-Eggers‹ für das Mittelhochdeutsche etwa, der die großen semasiologischen Wörterbücher von G. F. Benecke/W. Müller/F. Zarncke (BMZ) und M. Lexer ergänzen müßte, ist nicht vorhanden und nicht in Sicht.« Plate 1992, S. 313.

9 Das »*Simple Knowledge Organization System* (SKOS) ist ein W3C-Metadatenstandard für die digitale Organisation von Wissen. […] SKOS versucht […] Strukturen abzubilden und für den Austausch und die Verlinkung im Semantischen Web aufzubereiten.« (Bleier 2021).

10 Besonderer Dank ergeht an Univ.-Prof. Dr. Manfred Kern sowie Mag. Manuel Schwembacher, beide Universität Salzburg.

11 Brevitas – Gesellschaft zur Erforschung vormoderner Kleinepik e. V., vertreten durch: Silvan Wagner (1. Vorsitzender), Patrizia Barton (2. Vorsitzende), Friedrich Michael Dimpel (Schatzmeister): http://brevitas.org/.

Literaturverzeichnis

Handschriften

Codex Manesse Heidelberg, Universitätsbibliothek, cpg 848 (online).

Primärliteratur

Die Lieder Oswalds von Wolkenstein. Digitalisierung der Ausgabe von K.K.Klein im Auftrag der Mittelhochdeutschen Begriffsdatenbank MHDBDB und der Oswald von Wolkenstein-Gesellschaft von Bettina Hatheyer (online).

Sekundärliteratur

Bleier, Roman: Metadaten-Schemata für LZA: SKOS, in: Klug (2021) (online).

Borek, Luise/Zeppezauer-Wachauer, Katharina/Ketschik, Nora: Eindeutig Uneindeutig. Zur Modellierung von Unschärfe in der Mediävistik, in: Mittelalter.

Interdisziplinäre Forschung und Rezeptionsgeschichte (16. Februar 2022) (online).

Brom, Vlastimil: Die Mittelhochdeutsche Begriffsdatenbank als ein vielseitiges Arbeitsinstrument zur Analyse älterer deutschsprachiger Texte. Middle High German Conceptual Database: a flexible tool for analysing older German texts, in: Brünner Beiträge zur Germanistik und Nordistik 33, iss. Supplementum (2019), S. 173–184 (online).

Debus, Friedhelm/Pütz, Horst P. (Hrsg.): Namen in deutschen literarischen Texten des Mittelalters. Vorträge. Symposion Kiel, 9.–12. 9. 1987, Neumünster 1989 (Kieler Beiträge zur deutschen Sprachgeschichte 12).

Distilo, Rocco: Per un portale del lessico poetico europeo (TrobVers/MHDBDB), in: Forme e la storia. Rivista di filologia moderna. VI, 1 (2013), S. 209–214 (online).

Dornseiff, Franz: Der deutsche Wortschatz nach Sachgruppen. 7. Aufl., Berlin 1970.

Eibinger, Julia: TEI (Text Encoding Initiative), in: Klug (2021) (online).

Galka, Selina: Ontologie, in: Klug (2021) (online).

Hallig, Rudolf/Wartburg, Walter von: Begriffssystem als Grundlage für die Lexikographie. Versuch eines Ordnungsschemas, Berlin 1963.

Hinkelmanns, Peter (Hrsg.): List of Middle High German stopwords. 2021a (online).

Hinkelmanns, Peter (Hrsg.): Middle High German pos tagger model for spacy. 2021b (online).

Hinkelmanns, Peter (2021c): Semantic Web, in: Klug (2021) (online).

Hinkelmanns, Peter (Hrsg.): XQuery Module API for Transkribus PageXML. 2021d (online).

Hinkelmanns, Peter/Landkammer, Miriam/Nicka, Isabella/Schwembacher, Manuel/Zeppezauer-Wachauer, Katharina: Beyond the Plot. Der Vergleich mittelalterlicher Narrative im Semantic Web mit ONAMA, in: Vienna Doctoral Academy – Medieval Academy (Hrsg.): Narrare – producere – ordinare. New approaches to the middle ages, Wien 2022, S. 143–158.

Hinkelmanns, Peter/Zeppezauer-Wachauer, Katharina: *ez ist ein wârheit, niht ein spel, daz netze was sinewel*. Die MHDBDB im Semantic Web, in: Fischer, Martin (Hrsg.): Digitale Methoden und Objekte in Forschung und Vermittlung der mediävistischen Disziplinen. Akten der Tagung Bamberg, 08.–10. November 2018, Bamberg 2020, S. 73–81 (online).

Hinkelmanns, Peter/Zeppezauer-Wachauer, Katharina: Mittelhochdeutsche Begriffsdatenbank (MHDBDB), in: Klug (2021) (online).

Ketschik [geb. Echelmeyer], Nora/Reiter, Nils/Schulz, Sarah: Ein PoS-Tagger für ›das‹ Mittelhochdeutsche, in: Dhd 2017 Konferenzabstracts. 2017, S. 141–147 (online).

Klug, Helmut W. (Hrsg.): KONDE Weißbuch. Unter Mitarbeit von Selina Galka/Elisabeth Steiner im HRSM Projekt »Kompetenznetzwerk Digitale Edition«, Graz 2021 (online).

Nicka, Isabella/Hinkelmanns, Peter/Landkammer, Miriam/Schwembacher, Manuel/Zeppezauer-Wachauer, Katharina: Erzählerische Spielräume. Medienübergreifende Erforschung von Narrativen im Mittelalter mit ONAMA, in: Schöch, Christof (Hrsg.): DHd 2020 Spielräume. Digital Humanities zwischen Modellierung und Interpretation, Paderborn 2020, S. 131–135 (online).

Plate, Ralf: Onomasiologische Umkehrlexikographie auf dem Prüfstand, in: Zeitschrift für Dialektologie und Linguistik 59 (1992), S. 312–329.

Pollin, Christopher (2021a): CIDOC-Conceptual Reference Model (CRM), in: Klug (2021) (online).

Pollin, Christopher (2021b): RDF, RDFS, OWL, in: Klug (2021) (online).

Pütz, Horst P.: Rechnergestützte Bearbeitung großer Datenmengen am Beispiel des entstehenden Lexikons, in: Debus, Friedhelm/Pütz, Horst P. (Hrsg.): Namen in deutschen literarischen Texten des Mittelalters. Vorträge. Symposion Kiel, 9.–12. 9. 1987, Neumünster 1989 (Kieler Beiträge zur deutschen Sprachgeschichte 12), S. 287–299.

Pütz, Horst P./Schmidt, Klaus M.: Die Mittelhochdeutsche Begriffsdatenbank, in: Mittelalter-Philologie im Internet. ZfdA 130 (2001), S. 493–495 (online).

Roget, Peter Mark: Thesaurus of English Words and Phrases. Classified and arranged so as to facilitate the Expression of Ideas and assist in Literary Composition [1. Aufl.], London 1852.

Schmidt, Klaus M.: Tendenzen zum Realismus in der ritterlichen Epik der nachklassischen Periode. Untersuchungen zu Ulrichs von Liechtenstein Frauendienst. Dissertation, University of Michigan, Ann Arbor MI 1972.

Schmidt, Klaus M.: Wege zu Begriffsglossaren und einem Begriffswörterbuch mittelhochdeutscher Epik, in: Lenders, Winfried/Moser, Hugo (Hrsg.): Maschinelle Verarbeitung Altdeutscher Texte. Beiträge zum Symposion Mannheim 11./12. Juni 1971, Berlin 1978, S. 127–146.

Schmidt, Klaus M.: Begriffsglossare und Indizes zu Ulrich von Liechtenstein, München 1980 (Indices zur deutschen Literatur 14/15).

Schmidt, Klaus M.: Der Beitrag der begriffsorientierten Lexikographie zur systematischen Erfassung von Sprachwandel und das Begriffswörterbuch zur mittelhochdeutschen Epik, in: Bachofer, Wolfgang (Hrsg.): Mittelhochdeutsches Wörterbuch in der Diskussion (Germanistische Linguistik 84), Tübingen 1988, S. 35–49.

Schmidt, Klaus M.: Begriffsglossar zu Ulrichs von Zatzikhoven ›Lanzelet‹, Tübingen 1993 (Indices zur deutschen Literatur 25).

Schmidt, Klaus M.: Begriffsglossar und Index zur ›Kudrun‹, Tübingen 1994 (Indices zur deutschen Literatur 29).

Schöch, Christof: Aufbau von Datensammlungen, in: Jannidis, Fotis/Kohle, Hubertus/Rehbein, Malte (Hrsg.): Digital Humanities. Eine Einführung, Stuttgart 2017, S. 223–233 (online).

Springeth, Margarete: Der analytische Weg ist das Ziel: Die Mittelhochdeutsche Begriffsdatenbank als Online-Textarchiv, in: Hofmeister, Wernfried/ Hofmeister-Winter, Andrea (Hrsg.): Wege zum Text. Über die Verfügbarkeit mediävistischer Editionen im 21. Jahrhundert. Grazer Kolloquium 17.–19. September 2008, Tübingen 2009 (Beihefte zu editio, Bd. 30), S. 185–202.

Steiner, Christian/Fritze, Christiane: Normdaten, in: Klug (2021) (online).

Steiner, Elisabeth: Metadaten (allgemein), in: Klug (2021) (online).

Stigler, Johannes: FAIR-Prinzipien, in: Klug (2021) (online).

Woesner, Katrin: Begriffsglossar und Index zu Albrechts ›Jüngerem Titurel‹, Tübingen 2008.

Zeppezauer-Wachauer, Katharina/Schwembacher, Manuel/Nicka, Isabella/Landkammer, Miriam/Hinkelmanns, Peter: Needful Things. Die Relationen der Dinge in einer Ontologie mittelalterlicher Narrative, in: MEMO. Medieval and Early Modern Material Culture Online 8 (2021) (online).

Online-Ressourcen

BibFrame 2.0 (Bibliographic Framework Initiative): http://www.loc.gov/bibframe/.

Brevitas – Gesellschaft zur Erforschung vormoderner Kleinepik (e. V): http://brevitas.org/.

GND (Gemeinsame Normdatei) Ontology: https://d-nb.info/standards/elementset/gnd.

List of Middle High German stopwords: https://github.com/Middle-High-German-Conceptual-Database/stopwords

Middle High German pos tagger model for spacy: https://github.com/Middle-High-German-Conceptual-Database/Spacy-Model-for-Middle-High-German.

MHDBDB (Mittelhochdeutsche Begriffsdatenbank). 2004–2022 (laufend): http://mhdbdb.sbg.ac.at/;

GitHub: https://github.com/Middle-High-German-Conceptual-Database; UniqueWords: http://mhdbdb.sbg.ac.at/mhdbdb/App?action=UniqueWords.

HWGL (e. V.) (Blog „Netzwerk Historische Wissens- und Gebrauchsliteratur“): https://hwgl.hypotheses.org/.

Netzwerk Offenes Mittelalter. DFG-gefördertes Projekt zu Linked Open Data in der deutschsprachigen Mediävistik: https://offenesmittelalter.org/.

ONAMA (Ontology of Narratives of the Middle Ages): http://onama.sbg.ac.at/;
GitHub: https://github.com/Middle-High-German-Conceptual-Database/onama;
Zenodo: 10.5281/zenodo.4285986.
OWL (Web Ontology Language): https://www.w3.org/TR/owl-features/.
MPS-Blog (Portal der Pflanzen des Mittelalters/Medieval Plant Survey) 2009–2022 (laufend): http://medieval-plants.org/.
REALonline. Bilddatenbank des Instituts für Realienkunde des Mittelalters und der frühen Neuzeit (IMAREAL) der Universität Salzburg. 2001–2022 (laufend): http://realonline.imareal.sbg.ac.at/.
SKOS (Simple Knowledge Organization System Reference): http://www.w3.org/TR/skos-reference.
spaCy - Industrial-Strength Natural Language Processing in Python: https://spacy.io/;
GitHub: https://github.com/explosion/spaCy.
Transkribus: https://readcoop.eu/transkribus/.
Web Annotation Vocabulary: https://www.w3.org/TR/annotation-vocab/.
Wörterbuchnetz: https://woerterbuchnetz.de/.
XQuery Module API for Transkribus PageXML: https://github.com/Middle-High-German-Conceptual-Database/xquery-pagexml-transkribus-module.

Anschrift der Autorin:

Dr. Katharina Zeppezauer-Wachauer
Paris-Lodron-Universität Salzburg
Mittelhochdeutsche Begriffsdatenbank MHDBDB
Erzabt-Klotz-Straße 1
A-5020 Salzburg
E-Mail: katharina.wachauer@plus.ac.at

Joachim Hamm

Repositorien und Datenbanken (Diskussionsbericht Sektion 4)

An die Projektvorstellungen schlossen sich in der Diskussion (Leitung: Jens Haustein) zunächst konkrete Sach- und Informationsfragen an. Die Verfügbarkeit der Daten des Handschriftencensus (HSC) einschließlich ihrer Webpräsentation werde, so Jürgen Wolf, durch das Marburger Rechenzentrum gewährleistet, doch sei über die auf 20 Jahre bemessene Laufzeit des Akademieprojekts hinaus eine institutionelle Absicherung wünschenswert. Auf die Nachnutzbarkeit der Daten zielte Gabriel Viehhausers Frage nach einer API, die eine maschinelle Abfrage der Daten erlaubt: Eine solche API sei, so Wolf, intern bereits verfügbar und solle nun allgemein zugänglich gemacht werden. Auch eine Recherche, welche die (paläographische bzw. kodikologische) Datierung der Handschriften abfragt (Brigitte Bulitta), sei bereits möglich, mit einer Datierungsgenauigkeit von etwa 25 Jahren bis in die Zeit von 1380/1400. Eine verbesserte Normierung der Signaturen, die Torsten Schaßan als wünschenswert erachtete, könnte ggf., so Wolf, mit Hilfe standardisierter Normdatenlösungen, die idealerweise das Handschriftenportal bereitstellt, erreicht werden. Darüber hinaus wurden Erweiterungen vorgeschlagen: Die bisher nicht aufgenommenen Urkunden (Kurt Gärtner) ließen sich, so Wolf, über künftige Urkundenprojekte einbinden. Dass auch die Übersetzungen genuin deutscher Texte in deren Datensätze aufgenommen werden, erscheine sinnvoll (Wieland Carls).

Mit der Frage nach der internationalen Vernetzung des Gesamtkatalogs der Wiegendrucke (GW) wurde ein zentrales Thema der Sektion ange-

sprochen (Henrike Lähnemann). Der Austausch etwa mit den Datenbanken Material Evidence in Incunabula (MEI) bzw. 15cBooktrade sei, so Oliver Duntze, vorhanden und lebendig, aber bisher nicht über technische Schnittstellen formalisiert.

Zum Bochumer Referenzkorpus Mittelhochdeutsch (ReM) wurden Fragen nach Suchfunktion und Textkorpus gestellt. Eine unscharfe Suche, die umständlich einzugebende Sonderzeichen ignoriere (Klaus Kipf), sei, so Stefanie Dipper, technisch nicht unproblematisch, eine Erweiterung des geschlossenen Korpus (Uta Goerlitz) nicht geplant. Simone Schultz-Balluff schlug vor, bisher nicht anderweitig edierte Texte des Korpus (»verborgene Schätze«) besser kenntlich zu machen.

An die Vorstellung der Mittelhochdeutschen Begriffsdatenbank (MHDBDB) schlossen sich Fragen zum Tagging der Texte an: Eine TEI-Strukturauszeichnung (Gabriel Viehhauser) sei, so Katharina Zeppezauer-Wachauer, bisher nur rudimentär erfolgt. Eine Erweiterung des Korpus um neu edierte Text (Friedrich Dimpel) sei wünschenswert, jedoch nur bei entsprechender Lizenzierung möglich. Damit war ein wichtiger Aspekt digitaler Repositorien berührt: Mehrere Diskutanten pflichteten der Empfehlung bei, künftige Editionsprojekte sollten Lizenzen für MHDBDB und Wörterbuchnetz von Beginn an mit den Verlagen vertraglich vereinbaren.

In der übergreifenden Diskussion traten vier Schwerpunkte hervor. Besondere Aufmerksamkeit fanden zunächst Fragen nach Dauerhaftigkeit und Langzeitverfügbarkeit der digitalen Daten. Mehrere Diskutanten betonten, dass gerade Datenbanken und Repositorien, die ja zu den ältesten Digital-Humanities-Projekten der Mediävistik zählten, von der Langlebigkeit digitaler Angebote zeugten, die zentralen Belangen und Bedürfnissen des Fachs entgegenkämen: Was benutzt werde, existiere auch weiter.

Ein zweiter Schwerpunkt der Diskussion lag auf der Forderung, die Zugänglichkeit und intuitive Nutzbarkeit von Datenbanken und Repositorien zu verbessern. Mehrfach angesprochen wurden die Suchfunktionen: Die Implementierung ›unscharfer‹ Suchfunktionen erschien zumal bei

historischen Datenbeständen einer *case*-sensitiven Recherche deutlich überlegen. Angesprochen wurde zudem, dass erweiterte, flexible Recherchemöglichkeiten erst dann in der Breite genutzt würden, wenn sie hinreichend und verständlich beschrieben seien. Insgesamt plädierten mehrere Teilnehmer dafür, die Frontends benutzerfreundlicher und intuitiver zu gestalten, was man, so Viehhauser und Zeppezauer-Wachauer, professionellen Programmierern mit Erfahrung in Usability und Web-Design überlassen könne.

Einen dritten Aspekt der Diskussion bildete die Forderung nach weiterer Standardisierung und Normierung. Hier seien nach allgemeiner Einschätzung der Austausch mit Normdatenbanken wie der GND, die zu intensivierende Kooperation der Wissenschaft mit normierenden Institutionen und Instanzen (insbesondere den großen Bibliotheken), die flächendeckende Einführung fester IDs (z. B. auf Handschriftenebene, vgl. die geplanten ›International Standard Manuscript Identifiers‹, ISMI) oder auch die weitere Nutzbarmachung der Normdaten (z. B. ihre Implementierung in die Suchfunktion) von hoher Bedeutung. Hierbei sollte der Austausch wechselseitig erfolgen. So seien die Autoren- und Werkdaten des HSC schon weitgehend in eine Normdatenebene überführt, auch wenn dieser Prozess noch nicht abgeschlossen sei. Ziel müsse die stärkere Vernetzung der einzelnen Datenbanken und Repositorien sein (Schulz-Balluff, Andrea Rapp), wie sie etwa in der Verknüpfung von GW und Typenrepertorium bereits erfolgt sei. Interessante Perspektiven böten eine Weiterentwicklung der Datenbanken zu Forschungsplattformen mit Portalfunktion oder die Erweiterung der öffentlichen Nachnutzung über technische Schnittstellen oder digitale Kommunikationstechnologien (Twitter usw.)

Ein abschließendes philologisches Caveat formulierte Gärtner: Er wies darauf hin, dass in digitalen Repositorien bei der Texterfassung Variantenapparate oft ausgespart blieben, so dass sich die Frage stelle, ob die Textkritik in diesen Gebrauchskontexten eine digitale Zukunft habe.

Literaturverzeichnis

Online-Ressourcen

15cBooktrade: https://15cbooktrade.ox.ac.uk/.
GW (Gesamtkatalog der Wiegendrucke):
https://www.gesamtkatalogderwiegendrucke.de/.
MEI (Material Evidence in Incunabula): https://data.cerl.org/mei/.
TW (Typenrepertorium der Wiegendrucke): https://tw.staatsbibliothek-berlin.de/.

Anschrift des Berichterstatters:

Prof. Dr. Joachim Hamm
Julius-Maximilians-Universität Würzburg
Institut für deutsche Philologie
Am Hubland
97074 Würzburg
E-Mail: joachim.hamm@uni-wuerzburg.de

Sektion 5: Online publizieren und digitale Wissenschaftskommunikation (Podiumsdiskussion)

Gabriel Viehhauser

Online publizieren und digitale Wissenschaftskommunikation (Bericht über die Podiumsdiskussion, Sektion 5)

Die digitale Transformation verändert nicht nur Formen des wissenschaftlichen Arbeitens, sie führt auch zu neuen Formen der Öffentlichkeit. Strategien des Online-Publizierens sowie die Potentiale, aber auch die Herausforderungen der digitalen Wissenschaftskommunikation waren auf unserer Tagung Gegenstand einer von Henrike Lähnemann geleiteten Podiumsdiskussion, der Albrecht Hausmann die folgenden vier Leitfragen voranstellte:

1. Wie verändert sich die mediävistische Publikationslandschaft mit ihren Spezifika (z. B. relativ ›kleines‹ Publikum, auch ›spezialistische‹ Formate, etwa bei Editionen) durch die Digitalisierung?
2. Welche Chancen, aber auch Risiken sind damit verbunden (z. B. Qualitätssicherung, Open Access, Probleme der Aufmerksamkeitslenkung, Fragen der Zugänglichkeit)?
3. Welche Rolle spielen dabei die Verlage, welche Bedeutung haben ›neue‹ Formate wie etwa das Archivum Medii Aevi Digitale (AMAD) oder die ›Beiträge zur mediävistischen Erzählforschung‹ (BmE)? Gibt es hier Gegensätze, Interessenkonflikte, ungleiche Startbedingungen, ökonomische Probleme?
4. Wie groß ist die Bereitschaft in der Wissenschaft selbst, auf digitale Formate zu setzen - oder wollen doch alle am Ende ihr ›Buch‹ in der Hand haben?

In seinem Eingangsstatement am Podium benannte Hausmann zunächst zwei Aspekte, die sich im Rahmen der digitalen Transformation stark verändert hätten: Zum einem stelle sich die Frage: Was ist eine digitale Publikation im digitalen Bereich? Sind etwa Tweets auf Twitter Publikationen, und wie steht es um die Zitierfähigkeit und das Prestige solcher Veröffentlichungen? Zum anderen werde die Frage des Open Access im Zuge der Digitalisierung besonders relevant, schließlich werde es durch die neuen medialen Formen einfacher und auch billiger, Texte zu publizieren. Insbesondere der zweite Punkt sei dabei kritisch, da sich die Frage stelle, wie man Transformation gestalten und Möglichkeiten nutzen könne, ohne dass sich vorhandene Institutionen hemmend auswirkten. Momentan, so zeigte sich Hausmann besorgt, sei eine Entwicklung weg von einem Nutzer*innenfinanzierten zu einem von den Produzent*innen finanzierten Modell zu beobachten. Früher seien hauptsächlich Bibliotheken und die öffentliche Hand Finanzierer von Büchern gewesen, diese orientierten sich jedoch neu, und nun bestehe die Gefahr, dass Forscher*innen unter die Räder kommen, die für ihre Publikationen mittlerweile Anträge schreiben müssten.

Hausmann berichtete zudem von seinem eigenen Projekt, der gemeinsam mit Anja Becker gegründeten Zeitschrift ›Beiträge zur mediävistischen Erzählforschung‹ (BmE). Diese verfolge durchaus gewollt ein konservatives Modell und nutze bei weitem nicht alle digitalen Möglichkeiten, sondern komme statt dessen konventionellen Bedarfen entgegen (etwa durch die Bereitstellung gut lesbarer PDFs), um die Akzeptanz der potentiellen Leser*innen nicht zu strapazieren. Der Erfolg gebe dem Modell recht, so verzeichne die Zeitschrift derzeit 1000 Dateizugriffe im Monat. Eine Förderung für die Beiträge gebe es nicht, da die DFG auf solche Publikationsformen nicht eingestellt sei.

Als zweite Sprecherin kam mit Karoline Döring die (Mit-)Gründerin des Mittelalter-Blogs und der Plattform Archivum medi aevi digitali (AMAD) zu Wort. Sie brach eine Lanze für den Blog als grundsätzlich mögliche Sphäre der Wissenschaftskommunikation. Blogs seien durchaus vielfältig,

und nicht alle von ihnen zeichneten sich durch Unfertigkeit und Vorläufigkeit aus. Beim Mittelalterblog etwa sorgten die vier Faktoren *editorial review* als Kooperation zwischen Herausgeber*in und Autor*in, änderungssicheres PDF/A-Format, zitierfähige DOI und Langzeitarchivierung für Stabilität. Die Reihe der Mittelalterblog-Beihefte reagiere mit einer Dynamisierung auf die Anforderungen des hohen Publikationsdrucks, die im Widerspruch zu konventionell zumeist langen Veröffentlichungszeiten und kurzen Vertragslaufzeiten von Forschenden in ihrer Qualifizierungsphase stünden. Döring sprach sich zudem für eine vielfältige Qualitätssicherung ohne deren Fetischisierung sowie für eine umfangreiche Vernetzung digitaler Angebote aus, wie sie etwa bei AMAD gegeben sei. Zu AMAD gebe es auch einen Twitterkanal, der nach *best practices* gestaltet ist. Die neue Form der Kommunikation verlange insbesondere nach neuen Formen der Kooperation, nach neuen Formen der Aufgaben- und Kompetenzenverteilung. Im Zentrum bei dieser Neuverteilung stünden dabei die Fragen »Wer braucht was?« und »Wer kann was?«, schlussendlich aber auch »Wer macht was?«.

Robert Forke vom Verlag de Gruyter betonte in seinem Podiumsbeitrag, dass die digitale Transformation insbesondere für geisteswissenschaftlich ausgerichtete Verlage große Herausforderungen mit sich bringe. Laut Forke habe sich das Verhältnis der Verkaufszahlen von gedruckten Büchern und *ebooks* in den letzten zwei Jahren (insbesondere in der Pandemie) gänzlich zugunsten des digitalen Formats umgekehrt. Aus diesen Befunden folge, dass sich das Geschäftsmodell der Verlage ändere und Finanzierung zum Problem werde. Allerdings ergäben sich auch neue Chancen des Austauschs, etwa mit Wissenschafter*innen und Bibliothekar*innen. Gefragt seien neue Formen der Qualitätssicherung und dass Verlage sich darauf ausrichteten, eine Partnerfunktion auszuüben.

Die anschließende Diskussion des Podiums drehte sich zunächst um die Frage der Qualitätssicherung: Hausmann wies darauf hin, dass gerade neue Unternehmungen stark auf Qualitätssicherung achten müssten, um den

Vorurteilen der digitalen Flüchtigkeit und der mangelnden Qualität zu entgehen. *Peer review* und Langzeitarchivierung seien unbedingt nötige Anforderungen. Döring berichtete, dass bei AMAD anfangs mit komplett offener *peer review* experimentiert wurde (alle dürfen alles schreiben), man sich aber von allzu offenen Formen mittlerweile verabschiedet habe; als Minimalanforderung müssten jedoch zumindest Autor*in und Reviewer*in bekannt sein. Henrike Lähnemann plädierte für kreative und neue Formen von Qualitätssicherung und für den Einbezug von oft ungenutztem Fachwissen in den Bibliotheken.

Weitere Aspekte der Diskussion auf dem Podium bezogen sich auf die auch internationale Sichtbarkeit der Publikationen (wie weitgehend soll in Englisch publiziert werden? Lähnemann wies darauf hin, dass ein Großteil der deutschsprachigen Publikationen von einem internationalen Publikum nicht wahrgenommen werde) und was eigentlich alles als Publikation gelten könne. Döring plädierte dafür, auch kleineren und ungewöhnlicheren Formaten Raum zu geben und den Begriff der Publikation grundlegend zu diskutieren. In Hinblick auf die Finanzierungsproblematik stellte Döring die Frage, wieso man sich eigentlich in finanzieller und rechtlicher Hinsicht so stark beschränke, wo es doch die vielfältigen digitalen Möglichkeiten gebe.

Schließlich war in der Podiumsdiskussion noch die Rolle der Verlage, insbesondere in Zusammenhang mit der Langzeitarchivierung, ein bestimmendes Thema. Hausmann stellte die Frage, worin der Vorteil einer an den Verlag delegierten Langzeitarchivierung liege (die als private Unternehmen ja auch wieder vom Markt verschwinden könnten). Forke antwortete hierauf, dass Verlage dieselben Langzeitsicherungsmechanismen nutzten wie Bibliotheken und dass auch DFG-Langzeitprojekte durchaus vom Verlag möglich gemacht würden (als Beispiel wurde etwa das ›Verfasserlexikon‹ genannt). Verlage stünden dem freien Zugriff und *open access* keineswegs negativ gegenüber, es gebe kein intrinsisches Interesse an Beschränkungen. Allerdings koste jede Publikation Geld, an die Stelle

des alten Modells (der Verlag schießt vor und finanziert sich durch die Zugangsbeschränkung) trete die Vorabpublikation. Im Einklang mit Döring betonte auch Forke die Notwendigkeit der Kooperation, um Ressourcen zu bündeln. So könnten Verlage etwa im Sinne der Frage nach dem »Wer macht was?« ohnedies belasteten Wissenschafter*innen die Publikationsarbeit abnehmen.

An diese Diskussionsstränge wurde danach auch in der für die Allgemeinheit geöffneten Diskussion angeknüpft. Elisabeth Lienert bestätigte, dass ihr auch aus traditioneller Sicht die Qualität der ›Beiträge zur mittelalterlichen Erzählforschung‹ sehr gut erscheine, die Qualitätssicherung aber de facto nur über eine ›Selbstausbeutung‹ der Herausgeber funktioniere. Qualitätssicherung sei eine Aufgabe für Profis, wo auch immer diese angesiedelt sind.

Torsten Schaßan lieferte auf der Grundlage des Publikationskonzepts des Handschriftenportals eine Antwort auf die Frage, was denn alles als Publikation zu verstehen sei, nämlich letztlich alles, was mit Namen zu einem definierten Zeitpunkt veröffentlicht wurde. Schon Annotationen oder Veränderungen von Zeitangaben (etwa im Handschriftenportal) stellten Mikropublikationen dar. Zur Frage der Finanzierung gab Schaßan zu bedenken, dass Publikationen im Digitalen zumindest heutzutage nicht mehr billiger sind und dass die Kosten für Publikationen immer schon zu tragen gewesen seien.

Stephan Müller strich die Notwendigkeit einer neuen Aufgabenverteilung für Bibliotheken und Verlage hervor, diese müssten sich neu erfinden. Für Verlage kämen etwa neue strukturelle Aufgaben hinzu (Bibliometrie, Rechteverwaltung), wodurch sich ihr Aufgabenspektrum verschiebe. Texte zur Verfügung zu stellen und zitierbar zu machen, liege ohnedies in der Kompetenz der Autor*innen. Demgegenüber seien weiterführende Aufgaben wie Qualitätssicherung und Lektorat (erst) die Rechtfertigung für die Rolle des Verlags.

Auch Jürgen Wolf betonte, dass Verlage nach wie vor wichtig seien, ihre Rolle aber neu definieren müssten. Zudem beklagte auch er den derzeitigen Ausbeutungscharakter des Online-Publikationssystems (Autor*innen müssten zusätzlich zum Erstellen der Artikel nun auch für deren Publikation sorgen) und unterstrich die Bedeutung von Mikropublikationen im digitalen Medium.

In Hinblick auf die Frage der Selbstausbeutung gab Döring zu bedenken, dass die *crowd* sich engagieren und zum Teil selbst ausbeuten wolle, da als Lohn der Arbeit zwar nicht Geld, aber Reputation und Sichtbarkeit zu erlangen seien. Klaus Kipf plädierte im Hinblick auf die Offenheit von *Reviews* dafür, bestimmte Teile der Wissenschaftskommunikation nicht öffentlich ablaufen zu lassen, um so auch Schutzräume zu schaffen. Hausmann führte schließlich die Diskussion um neue Aufgaben- bzw. Finanzierungsverteilungen und Selbstausbeutung zurück zu seinen eigenen Erfahrungen (und damit zu ihrem freilich nur vorläufigen Ende): Natürlich verursachten auch die BmE Kosten, letztlich sei aber nichts anderes zu leisten als das, was beispielsweise bei der Produktion eines Sammelbandes anfallen würde, nämlich die Abgabe eines fertig gelayouteten Produkts. Insofern erschienen die Kosten für die Wissenschaft heutzutage zwar verteilter, sie seien (zumindest in Ansätzen) aber immer schon da gewesen. Eine bessere Zusammenarbeit mit Verlagen sei allerdings durchaus vorstellbar.

Literaturverzeichnis

Online-Ressourcen

AMAD (Archivum medi aevi digitali): https://www.amad.org/;
Twitterkanal: https://twitter.com/amad_org.
BmE (Beiträge zur mediävistischen Erzählforschung): https://ojs.uni-oldenburg.de/ojs/index.php/bme.
De Gruyter: https://www.degruyter.com/.
Handschriftenportal: https://handschriftenportal.de/.

Anschrift des Berichterstatters:

Prof. Dr. Gabriel Viehhauser
Universität Stuttgart
Institut für Literaturwissenschaft
Herdweg 51
70174 Stuttgart
E-Mail: viehhauser@ilw.uni-stuttgart.de

Sektion 6: Stilometrie und Textanalyse

Gabriel Viehhauser

Digitale Methoden der Textanalyse für die Altgermanistik

Abstract. Der Beitrag gibt einen Einblick in einige Methoden der digitalen Textanalyse (Frequenzanalyse, *Principal Component Analysis*, *Lexical Diversity* und *Topic Modeling*), situiert diese in der Theoriediskussion der Digital Humanities und erprobt ihre Anwendung auf mittelhochdeutsche Literatur anhand eines Minnesangkorpus. Zum einem sollen dabei spezifische Herausforderungen herausgestellt werden, die sich bei der Anwendung digitaler Analysemethoden auf mittelhochdeutsche Texte bieten. Zum anderen plädiert der Beitrag dafür, die Methoden in einem multiperspektivischen Zugang zu nutzen, der das Spektrum von der digitalen Makro- bis zur qualitativen Detailanalyse umfasst.

1. Einleitung

Der folgende Beitrag will einen kurzen Einblick in einige Methoden geben, die sich zurzeit in den Digital Humanities im Rahmen der digitalen Textanalyse etabliert haben, und an Beispielen aus der mittelhochdeutschen Literatur überprüfen, ob solche Methoden auch in diesem Bereich zum Einsatz kommen können und welche besonderen Probleme damit verknüpft sind.

Letztlich basieren sämtliche Verfahren der digitalen Textanalyse, so vielgestaltig sie auch sein mögen, auf derselben Grundlage, nämlich auf der Auszählung von Features, die in einem Text auftreten, wobei in den meisten Fällen diese Features schlicht die einzelnen Wörter eines Textes darstellen. Die Methoden bewegen sich mithin notwendigerweise auf der Textober-

fläche. Der Computer kann im Rahmen der digitalen Textanalyse also eigentlich nichts anderes tun, als einfach nur Wörter zählen, aber dieses Wörterzählen kann in unterschiedlich elaborierten Formen erfolgen, die es zum Teil sogar erlauben, über die explizit gegebene Textoberfläche hinaus Schritte in Richtung auf komplexere Textbedeutungen hin zu gehen. Das Spektrum reicht dabei von der Erstellung von Konkordanzen und einfachen deskriptiven Statistiken über Verfahren, die in der Korpuslinguistik und im Information Retrieval gängig sind, bis hin zu komplexeren Modellen der distributionellen Semantik und Machine-Learning-Verfahren.

2. Methodologische Vorüberlegungen

Bevor ich ein paar dieser Verfahren vorstelle und auf konkrete Beispiele zu sprechen komme, erscheint es mir notwendig, einige methodologische Überlegungen voranzuschicken und die einschlägige Theoriediskussion der Digital Humanities und insbesondere der Digital Literary Studies zu rekapitulieren, da nur so deutlich werden kann, vor welchem Hintergrund sich solche digitalen Methoden bewegen und welcher Geltungsbereich für sie zu veranschlagen ist.

Computer sind als Rechenmaschinen immer dann besonders performant, wenn es darum geht, große Mengen an Daten auszuwerten und dieselben regelhaften Abläufe wiederholt auszuführen. Für die Untersuchung von Einzelfällen ist die Erstellung eines Algorithmus eigentlich überflüssig und wohl auch nicht nachhaltig, denn eine eigenständige Neuprogrammierung lohnt sich erst dann, wenn dem Algorithmus ein gewisser Grad der Generalisierbarkeit zu eigen ist. Softwarelösungen, die sich nur auf einen einzelnen Fall anwenden lassen und danach unbrauchbar werden, werden nur die wenigsten zur Benutzung eines Computers veranlassen.

Schon daraus ergibt sich, dass man bei der Arbeit mit dem Computer dazu tendiert, von Einzelfällen abzusehen und den Blick auf das große Ganze, die Makroperspektive, zu lenken. Mit dieser Fokussierung läuft die

Anwendung von Computern zunächst einer (zumindest unterstellten) grundsätzlichen Ausrichtung der Geisteswissenschaften auf das Individuelle und das Idiographische zuwider.[1] Auch wenn zu fragen ist, ob das Ansetzen einer solchen binären Klassifikation die individualisierenden Züge der Geisteswissenschaften nicht zu stark überbetont, so kann doch der Einsatz des Computers in den Humanities damit leicht zur Provokation werden.

Eine solche Provokation durch den Blickwechsel auf die Makroperspektive wird etwa durchaus bewusst von Franco Moretti in Kauf genommen, der das für die digitale Literaturwissenschaft wohl bislang wirkmächtigste Konzept entwickelt hat, nämlich den Ansatz des Distant Reading, anfangs sogar noch ohne den Blick auf das Digitale, nämlich im Kontext der Weltliteraturdebatte (Moretti 2000). Moretti geht dabei von dem Befund aus, dass die Gesamtmenge an vorhandenen Texten von einem einzelnen Menschen selbst bei größter Anstrengung schlicht nicht überschaut werden kann, da die dafür nötige Lesezeit die Lebenszeit einer Einzelperson bei weitem übersteigt. Daraus ergibt sich, dass bei jeglicher kulturwissenschaftlicher Analyse mit Selektivität und *bias* zu rechnen ist – und etwa Weltliteratur am Ende dann doch wieder vorwiegend als europäisches Konzept gedacht wird (Moretti 2000, S. 55). Moretti plädiert daher dafür, das genaue Lesen von Einzeltexten zugunsten eines kursorischen, sekundären Distant Reading zu verabschieden: Wenn dem Problem der unüberschaubaren Fülle mit immer mehr Lesen schlicht nicht beizukommen ist, dann, so die Überlegung Morettis, sollte man endlich aufhören zu lesen, und statt dessen nach neuen, alternativen Formen der Auswertung von Texten suchen (Moretti 2000, S. 57).

Distant Reading ist damit an das Versprechen eines ehrlicheren, weniger kolonialistisch und kanonistisch geprägten Zugangs geknüpft, der es nach Moretti nicht zuletzt erlauben soll, literaturgeschichtliche Zusammenhänge in ihrer Vollständigkeit zu erfassen und auf Muster gesellschaftlicher Machtstrukturen abzubilden. Wie in der Forschung bemerkt wurde (Underwood 2017), steht Morettis Konzeption damit in einer literatur-

soziologischen Tradition, auf deren Grundlage ästhetische Unterschiede nivelliert werden bzw. sekundär erscheinen und Texte in Hinblick auf ihre Strukturmuster abstrahiert werden, um so gesellschaftliche Machtmechanismen analysieren zu können: »Forms are the abstract of social relationships: so, formal analysis is in its own modest way an analysis of power« (Moretti 2000, S. 66) - neo-marxistische Strukturanalyse also an Stelle ästhetischer Ausdifferenzierung.

Dass mit dem Blick auf die Makroperspektive die Schärfe im Detail verloren geht, war freilich schon Moretti bewusst. Dieser Verlust muss seiner Ansicht nach jedoch als Erkenntnisbedingung für den Blick aufs ›große Ganze‹ in Kauf genommen werden (Moretti 2000, S. 57f.). Ein solcher Verlust an Detailschärfe ist aber nun eigentlich auch nicht spezifisch für den Einsatz digitaler Methoden, sondern ein erkenntnistheoretischer *trade-off*, der eigentlich immer einzukalkulieren ist, wenn es etwas zu erkennen gilt: Wenn man etwa von weit oben auf die Erde blickt, dann lässt sich gut der Zusammenhang der Kontinente erkennen; die Details der Erdoberfläche werden jedoch vernachlässigt. Fokussiert man hingegen auf die Details, bleiben die größeren Zusammenhänge aus dem Blick.

Letztlich, so ist in der Theoriediskussion der Digital Humanities der letzten Jahre immer wieder betont worden (vgl. in Auswahl: Ciula [u. a.] 2018; McCarty 2004; Flanders/Jannidis 2018; Piper 2017), beruhen Erkenntnisse auf Modellierungen, und Modelle (im Sinne von Stachowiak 1973) sind zwar immer Abbilder von etwas, von Sachverhalten oder der Wirklichkeit, aber nicht die Sache selbst. Sie sind daher notwendigerweise Verkürzungen, lassen bestimmte Details beiseite und heben bestimmte hervor. Modelle werden zudem zu einem bestimmten Zweck zum Einsatz gebracht - und dieser Zweck lässt sich überhaupt nur deswegen erfüllen, gerade weil Details beiseite gelassen und Ansichten akzentuiert werden.

Daraus folgt aber wiederum, dass kein Modell die Wirklichkeit in ihrer Komplexität vollständig abzubilden vermag; im Gegenteil, mit einem prominenten Aphorismus des Statistikers George Box ließe sich sogar umge-

kehrt sagen, dass alle Modelle letztlich falsch und defizitär sind: »all models are wrong«. Dass man sie dennoch zum Einsatz bringen möchte, liegt schließlich wieder an ihrer pragmatischen Komponente: »all models are wrong, but some are useful« (Box 1976, S. 201).

Es ist in den Digital Humanities schon öfter bemerkt worden, dass sich eine solche Vorstellung nun durchaus mit geisteswissenschaftlichen Idealen wie Komplexität und Multiperspektivität vereinbaren lässt (beispielsweise bei So 2017; Piper 2017; Pierazzo 2018; Underwood 2020; Viehhauser 2020): Es gibt eben nicht nur eine ›objektive‹ Sichtweise, sondern im Gegenteil lediglich unterschiedliche Blickpunkte. Gerade die besondere Bedeutung, die Modellierungen in den digitalen Geisteswissenschaften zugemessen wird, hat in besonderem Maße das Bewusstsein für diesen Sachverhalt geschärft. Mit dem Modellbegriff dürften sich wohl auch Erkenntnisprozesse in den ›traditionellen‹ Geisteswissenschaften beschreiben lassen, und es scheint paradox, dass dort das Bewusstsein für die Verkürzungen des eigenen Standpunkts mitunter sogar weit weniger ausgeprägt erscheint als in den technikgeprägten Digital Humanities (vgl. Piper 2016).

Dass der Rekurs auf den Modellierungsgedanken für die digitalen Geisteswissenschaften besonders naheliegend ist, ergibt sich wohl schon aus der Sache selbst: Schon per Definition besteht die Digitalisierung eines analogen Objekts (etwa eines Analogsignals) darin, dass dieses in diskrete Einheiten aufgelöst wird, die den analogen Zustand immer nur annähernd wiedergeben können, dafür aber zähl- und berechenbar sind. Ein Digitalsignal wäre so gesehen ein verkürzendes Modell eines Analogsignals, das zum Zweck der besseren Verarbeitung erstellt wird.

Zudem müssen Modelle in den Digital Humanities – und damit unterscheiden sich diese von den traditionellen Geisteswissenschaften – notwendigerweise immer auch formalisiert sein (Jannidis/Flanders 2018, S. 28; Jannidis 2018, S. 99). Erst durch die genaue Definition der einzelnen Komponenten und Relationen wird es möglich, Modelle zur Weiterverarbeitung an den Computer zu übergeben. Modellen in den digitalen

Geisteswissenschaften eignet daher ein höherer Grad an Explizitheit an, als dies in den traditionellen Geisteswissenschaften der Fall ist.[2]

Dies sollte jedoch keinesfalls zur Annahme verführen, dass es in Modellen der digitalen Geisteswissenschaften keine blinden Flecken gäbe. So tendieren Ansätze in den Digital Humanities oftmals dazu, tieferliegende Fragestellungen nicht direkt, sondern unter Zuhilfenahme von beobachtbaren Indikator- oder instrumentellen Variablen zu bearbeiten. (Moretti 2013, S. 2). Ein Beispiel wäre hier der Schluss von Worthäufigkeiten (auf der Textoberfläche) auf tieferliegende Bedeutungsschichten, etwa vom Auftreten von Begriffen auf die Wichtigkeit dieser Begriffe: Aufgrund der Ambiguität natürlicher Sprachen und deren stark impliziter Bedeutungserzeugung kann ein solcher Schluss immer nur eine gewisse Wahrscheinlichkeit beanspruchen. Tatsächlich beruhen die meisten Verfahren der Textanalyse auf Wahrscheinlichkeitsrechnungen, wodurch sich aber gerade wieder mögliche Verbindungen zu geisteswissenschaftlichen Denkweisen ergeben: Streng genommen treffen digitale Verfahren nämlich gerade keine binären Entscheidungen oder Kategorisierungen, sondern geben statt dessen lediglich die Wahrscheinlichkeiten für Zuordnungen an. Paradoxerweise bieten damit gerade digitale Methoden im Grunde durchaus differenziertere Instrumentarien als bloße Schwarz-Weiß-Entscheidungen (vgl. Craig/Greatly-Hirsch 2017, S. 3). Letztere ergeben sich erst dann, wenn man in die Skala der sich eigentlich erstaunlich analog ausnehmenden Wahrscheinlichkeitswerte nachträglich Schnitte einfügt (etwa einen Schwellwert, ab wann ein Ergebnis nicht als zufällig und daher als hypothesenbelegend anzusehen ist).

Schließlich muss es in Hinblick auf die Anschlussfähigkeit digitaler Methoden auch kein Nachteil sein, dass sich die starke Explizierung und Formalisierung der digitalen Modelle mit den oft komplexen und mehrdeutigen Objekten der Geisteswissenschaften reiben; denn gerade diese Reibung kann sich durchaus erkenntniserweiternd auswirken: Systematisch ließe sich in diesem Sinn etwa zwischen *data-driven* und *data-assisted*

Zugängen unterscheiden (Escobar Varela 2021, S. 7): Während erstere sich auf die strenge Methodik des formalen Modells einlassen und auf replizierbare Ergebnisse abzielen, fragen letztere danach, ob sich bestimmte Phänomene der Quantifizierung widersetzen, und vor allem, warum sie das tun. Der Umgang mit Daten wird hier also zum Ausgangspunkt für ein zwar methodisch geleitetes, aber interpretatives Vorgehen, das digitale Methoden sozusagen ›gegen den Strich‹ liest (vgl. Ramsey 2011).

3. Eine digitale Stilgeschichte des Minnesangs?

3.1 Minnesangs Meistererzählung

Aufgrund ihrer Detailvergessenheit laufen Untersuchungen aus der Makroperspektive nicht selten Gefahr, in teleologische und nivellierende Meistererzählungen abzugleiten. Dieser Befund gilt jedoch nicht bloß für die digitale Welt, sondern auch für die ›konventionelle‹ geisteswissenschaftliche Forschung. Eine der bekanntesten stilgeschichtlichen Großerzählungen der Altgermanistik ist etwa Hugo Kuhns Erzählung von ›Minnesangs Wende‹ (Kuhn 1952): Nach Kuhn wohnt dem späteren Minnesang eine Tendenz zur Objektivierung inne, die sich daraus ergibt, dass sich die Dichter nicht mehr an der Minne-Konzeption ›subjektiv‹ abarbeiten, sondern diese zur Konvention erstarrt, weshalb sich der Fokus auf die Form verschiebt (Kuhn 1952, S. 143–158; zur Kritik insbesondere Hübner 2013).

Auch in neueren Darstellungen wird dieser Gedanke aufgegriffen: Dass sich die Minnelyrik im 13. Jahrhundert transformiert, wird auch dort nur selten grundsätzlich bestritten (vgl. jedoch Hübner 2013, S. 387–390). Allerdings wird weit stärker die Vielgestaltigkeit und Ausdifferenzierung der Lieder betont, die es – als Einzelfälle – in den Blick zu bekommen gelte. Exemplarisch kann hierfür etwa die Einleitung zum Wolfram-Studien-Band ›Transformationen der Lyrik im 13. Jahrhundert‹ aus dem Jahr 2013

stehen. Dort wird – in deutlicher Anlehnung, aber zugleich Differenzierung der These von Kuhn – als Arbeitsprogramm formuliert:

> Wenn es richtig ist, daß der hochhöfische (›klassische‹) Minnesang innerhalb der Regeln spielt, dagegen der spätmittelalterliche Minnesang mit den Regeln selbst, muß es darauf ankommen, das Verhältnis von Konvention und Abweichung für möglichst viele Parameter zu klären, und zwar zunächst gesondert für jeden überlieferten Einzelfall. (Köbele 2013b, S. 9)

Konzeptionell bedingte Transformation ja, linearer teleologischer Entwicklungszusammenhang nein, so ließe sich also der Wechsel in der Einstellung zur Makroperspektive seit Kuhn resümieren. Für jeden Einzelfall ist gesondert zu eruieren, wie er sich zum Allgemeinen, zur Folie von Konvention und Abweichung, verhält.

3.2 Korpora und Wortfrequenzen

Ich möchte im Folgenden zeigen, dass sich dieses Wechselspiel von Verlauf und Einzelfall durchaus auch mit digitalen Mitteln nachzeichnen und zugleich auch als Perspektivenproblem kenntlich machen lässt. Ich nähere mich daher mit unterschiedlichen digitalen Zugängen dieser Frage an. Den Ausgangspunkt soll ein einziges, kurzes Wort und dessen Häufigkeit bilden, nämlich das Wort *ich*, dem eine zentrale Rolle im Minnesang zukommt. Es ist in der Forschung schon öfter bemerkt worden, dass sich die Frage nach der Objektivierung des Sanges über die Betrachtung der Rolle des *Ich* nachzeichnen lässt (vgl. etwa Schnell 2013). Dieser Befund bietet sich nun insbesondere für eine versuchsweise quantitative Auswertung an, da sich die Häufigkeit der Verwendung des Personalpronomens *ich* (und dessen abgeleiteter Formen) leicht auf der Textoberfläche nachzeichnen lässt. Die Frequenz von *ich* bietet also einen guten Indikator für Entwicklungen im Minnesang und ließe sich etwa als zeitliche Verlaufskurve gut darstellen (vgl. hierzu ausführlicher Viehhauser 2017).

Doch offenbart gerade ein solcher, vergleichsweise einfach erscheinender Versuch Probleme, die nun besonders für die spezifischen Situation der mittelalterlichen Literatur virulent werden, denn welches Korpus verwendet man zur Konstruktion einer solchen Verlaufskurve? Ich habe im Folgenden auf Texte zurückgegriffen, die leicht in digitaler Form zugänglich in der Mittelhochdeutschen Begriffsdatenbank (MHDBDB) vorliegen. Dort sind Minnesang-Texte überwiegend aus den klassischen Anthologie-Ausgaben abrufbar. Neben den Liedern aus Minnesangs Frühling (MF) habe ich jene aus Carl von Kraus' Liederdichtern (KLD) und den Schweizer Minnesängern (SM) berücksichtigt. Hinzu kommen die Lieder aus der Walther-Ausgabe (W) von Lachmann/Cormeau (1996) und die Lieder Konrads von Würzburg (KW) aus der Ausgabe Schröder (1924/59). Insgesamt umfasst das Korpus damit 103 Autoren.[3]

Wie leicht ersichtlich ist, werden überlieferungsgeschichtliche Feinheiten damit nicht erfasst. Es gehört zu den großen Leerstellen der digitalen Textanalyse, dass sie für Textvarianten, Mehrfachüberlieferungen oder ähnliches blind bleibt bzw. vielleicht sogar bis zu einem gewissen Grad blind bleiben muss, da sich solche Varianz schwer in die quantitative Analyse einrechnen lässt. Selbst wenn, wie angesichts der Errungenschaften der digitalen Editorik zu hoffen steht, überlieferungsgeschichtlich ausgerichtete Textausgaben in Zukunft auch stärker in digitaler Form zur Verfügung stehen werden, stellt sich weiterhin etwa bei einer Frequenzauszählung die Frage, auf welcher Textbasis diese zu erfolgen hat – aufgrund der Varianz kann der Frequenzbefund für unterschiedliche Textfassungen unterschiedlich ausfallen. Hier kommt also letztlich die oben angesprochene Tendenz der digitalen Textanalyse zur Vernachlässigung von Details zum Tragen: Auf die überlieferungsgeschichtlichen Einzelheiten kommt es nicht an, es zählt der Blick aufs große Ganze. Die gerade durch digitale Zugänge in der Editorik eröffnete Möglichkeit zur Darstellung von Komplexität wird in der Analyse also sogleich wieder nivelliert. Das Problem stellt sich grundsätzlich natürlich auch für neuere Texte, wird

bei mittelhochdeutscher Literatur mit ihrer unfesten Überlieferung aber besonders relevant.

Hinzu kommen Kategorisierungsprobleme: In den Ausgaben von Walther und Konrad etwa findet sich bekanntlich nicht nur Minnesang, sondern auch andere lyrische Formen wie Sangspruch, ohne dass immer eine klare Abgrenzung vorzunehmen ist, und generell sind die Übergänge zwischen den Gattungen fließend. Ich habe mich dafür entschieden, trotz dieser Unsicherheit für die vorliegenden Auswertungen möglichst nur Minnesang zu berücksichtigen. Für die Einschätzung, was genau zum Minnesang dazugehört und was nicht, habe ich mich an den im ›Verfasserlexikon‹ dokumentierten Forschungsstand gehalten und bin den dort verzeichneten Abgrenzungen gefolgt, wohl wissend, dass auch dies eine weitere Reduktion von Genauigkeit im Detail darstellt.

Schließlich stellt sich noch das Problem, wie eine zeitliche Verlaufskurve angesichts der Tatsache, dass sich die meisten Texte nur schwer und unsicher datieren lassen, überhaupt erstellt werden kann. Aus pragmatischen Gründen habe ich mich hier an Autorkorpora und deren gängige, etwa im ›Verfasserlexikon‹ oder bei Hübner 2008 angesetzten Datierungen gehalten. Erneut bringt also die digitale Methode fast notwendigerweise Komplexitätsreduktion mit sich.

Doch wie sieht nun ein solcher natürlich mit Unsicherheit behafteter Überblick über die Geschichte des Minnesangs aus? Für meine erste Analyse habe ich die einzelnen Autorenkorpora in sechs Zeitspannen eingeteilt, und zwar in Kategorie 1, früher Minnesang, 2, hoher Minnesang, sowie 3–6, später Minnesang, den ich wiederum in vier ungefähre zeitliche Phasen eingeteilt habe, die folgende Zeiträume betreffen: SpM 1: Anfang 13. Jahrhundert, SpM 2: Mitte 13. Jahrhundert, SpM 3: Ende 13. Jahrhundert, SpM 4: Ende 13./Anfang 14. Jahrhundert.

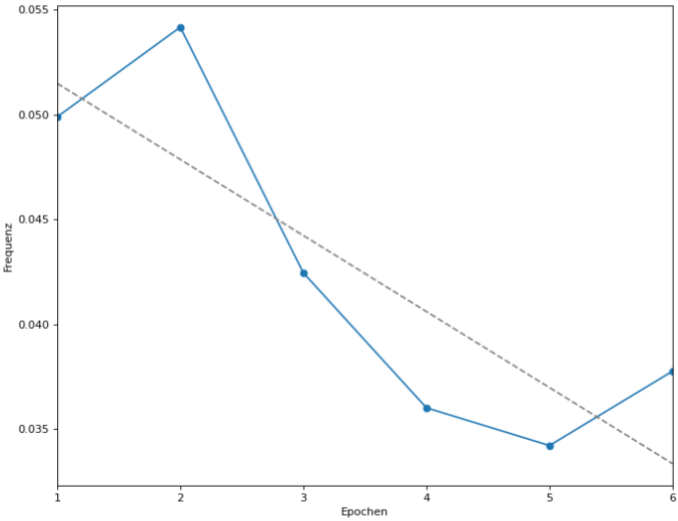


Abb. 1: *ich*-Frequenz im Minnesang nach Epochen

In Abb. 1 sind diese Phasen auf der x-Achse abgebildet, die y-Achse zeichnet die relative Wortfrequenz des Personalpronomens *ich* nach (also die Anzahl der *ich*-Belege geteilt durch die Gesamtanzahl der *tokens* des Teilkorpus). In der Tat zeigen nun die Verlaufskurve und die errechnete Trend-Gerade (gestrichelte Linie) eine Tendenz nach unten, scheinen also die Hypothese des abnehmenden *ich* -Bezugs zu bestätigen.[4]

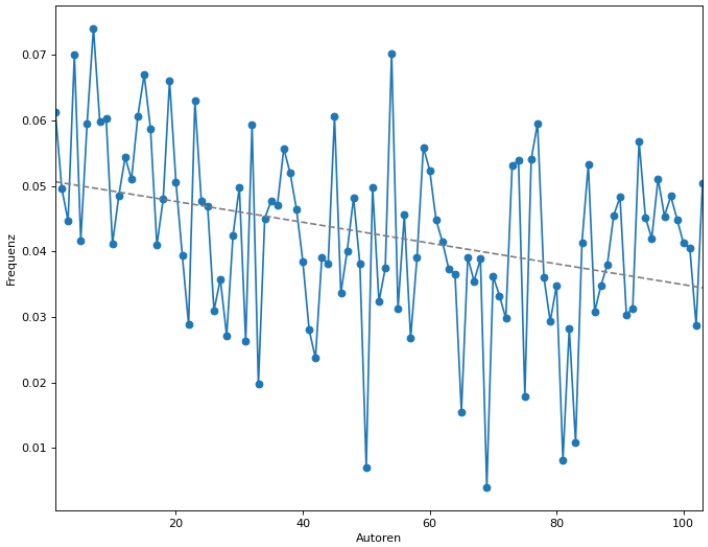


Abb. 2: *ich*-Frequenz im Minnesang nach Autorenkorpora

Ein differenzierteres Bild ergibt sich aber, wenn man, wie in Abb. 2, von der Makroperspektive etwas heranzoomt und nicht die zeitlichen Epochenkorpora, sondern die Autorenkorpora zur Grundlage macht. Hier zeigt sich nun recht augenfällig die oben angesprochene Diversität der Minnesangproduktion: Auch in der Spätphase begegnen durchaus noch Korpora mit durchschnittlichem oder sogar ausgesprochen hohem *ich* -Anteil. Die Frage, ob sich der Minnesang in seiner Ausrichtung grundsätzlich ändert, wird damit nicht zu einer binären Ja-Nein-Frage, sondern letztlich zu einer der Perspektive: Blickt man auf das große Ganze, dann zeigt sich ein Trend, zoomt man in die Details, dann sieht man Diversität, die dem Trend auch zuwiderlaufen kann.

Wie das Beispiel deutlich gemacht hat, lässt sich dieses Wechselspiel nun aber durchaus auch mit quantitativen Methoden fassen, ja sogar, dar-

auf hat So (2017, S. 670) hingewiesen, gerade besonders deutlich und gemäß dem Formalisierungsmodell explizit beschreiben. So könnten etwa die Abweichungen der einzelnen Datenpunkte von der errechneten Trendlinie in den Blick genommen werden: Je weiter sich die Einzelpunkte von der Linie entfernen, desto unsicherer ist es, von einem solchen Trend zu sprechen (So 2017, S. 670). Diese Unsicherheit ist nun aber gerade nicht eine Schwäche des Modells, sondern aus geisteswissenschaftlicher Hinsicht dessen Stärke:

> The advantage of statistical modeling is that it does not present cut- and- dried results that one accepts or rejects. Built into the modeling process is a self-reflexive account of what the model has sought to measure and the limitations of its ability to produce such a measurement. Again, as Box reminds us, »all models are wrong«. What's important is not to insist on how the model is right or nearly right but rather to understand how it is wrong. (So 2017, S. 671)

3.3 Hauptkomponentenanalyse (PCA)

Die Geschlossenheit bzw. Diversität der einzelnen Gruppierungen lässt sich auch mit einer etwas komplexeren statistischen Methode als dem bloßen Festhalten von Wortfrequenzen zur Darstellung bringen: Abb. 3 zeigt eine so genannte Hauptkomponentenanalyse bzw. *Principal Component Analysis* (PCA) der häufigsten Wörter (*Most Frequent Words,* abgekürzt MFW) in den zeitlichen Teilkorpora des Minnesangs.[5]

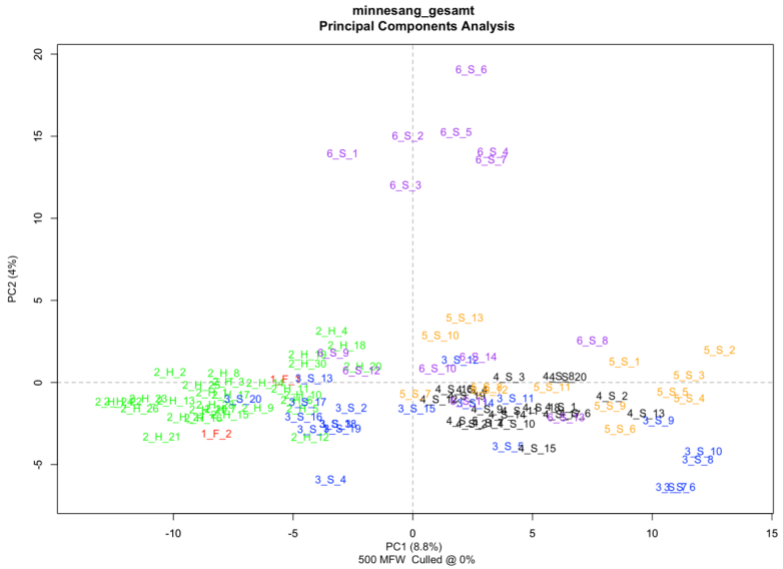


Abb. 3: PCA der Minnesangphasen (unnormalisiert)

In natürlichen Sprachen sind diese MFW üblicherweise Funktionswörter wie Artikel, Konjunktionen oder Präpositionen. In meinem Gesamtkorpus des Minnesangs tritt etwa *ich* mit 8933 Belegen am häufigsten auf; es folgen *daz* (5780), *ir* (4632), *der* (4100), *mir* (3905), *sô* (3485), *und* (2901), *mich* (2641), *mîn* (2624), *si* (2569) und *ist* (2524). Inhaltsbezogenere Wörter begegnen erst weiter hinten in der Liste der MFW; im Minnesang ist das erste dieser Wörter bezeichnenderweise *minne* (an 31. Stelle, mit 1113 Belegen).

Betrachtet man nun die 500 häufigsten Wörter für ein Textkorpus, dann ließe sich deren Frequenzverteilung als Vektor in einem 500-dimensionalen Vektorraum ansetzen, der auf einen Punkt in diesem Vektorraum zeigt. Platziert man in diesem Vektorraum die Wortfrequenz-Vektoren weiterer Korpora, dann lässt sich aus der Entfernung der Vektoren schließen, wie ähnlich die entsprechenden Korpora sind. Im 500-dimensionalen Raum bleiben solche Näheverhältnisse freilich höchst abstrakt. Abhilfe bietet hier

nun das Verfahren der PCA: Bei der PCA werden Variablen, die miteinander korrelieren, möglichst zusammengefasst und damit die Varianz in hochdimensionalen Vektorräumen auf einige wenige Dimensionen, eben die Hauptkomponenten, heruntergerechnet. So werden etwa in einem Datensatz, der verschiedene Schiffe nach deren Wasserverdrängung klassifiziert, die beiden Variablen Länge und Breite vermutlich so stark korrelieren, dass man sie zu einer Komponente (›Größe‹) zusammenfassen könnte.

Üblicherweise werden bei der PCA im Kontext von Textanalysen die ersten beiden Hauptkomponenten der Wortfrequenzverteilung identifiziert, da man diese übersichtlich in einem zweidimensionalen Koordinatensystem eintragen kann. In Abb. 3 habe ich die sechs zeitlichen Teilkorpora in Abschnitte zu jeweils 2000 Wörtern Länge aufgeteilt, die aufgrund ihres Frequenzprofils im zweidimensionalen Raum dargestellt werden können.[6] Abschnitte, die zum selben zeitlichen Teilkorpus gehören, sind durch gleiche Farbe gekennzeichnet (rot: früher Minnesang, grün: hoher Minnesang, blau: später Minnesang 1, schwarz: später Minnesang 2, orange: später Minnesang 3, violett: später Minnesang 4). Die erste Hauptkomponente ist in der Darstellung auf der x-Achse abgebildet, die zweite Hauptkomponente auf der y-Achse. Die Abbildung ist also so zu lesen, dass sich die links platzierten Textabschnitte hinsichtlich der ersten Hauptkomponente stark von jenen unterscheiden, die rechts positioniert sind. Die oben abgebildeten Textabschnitte unterscheiden sich in Hinblick auf die zweite Hauptkomponente von jenen, die unten platziert sind.[7]

Die Darstellung legt zunächst nahe, dass eigentlich nur der hohe Sang (gemeinsam mit den beiden Textabschnitten des frühen Sangs) und der zweite Abschnitt des späten Sangs eine einigermaßen homogene Gruppe ausbilden (in die sich aber auch manche Abschnitte anderer Gruppe einmischen), denn nur bei ihnen liegen die einzelnen Teilabschnitte beieinander, verhalten sich also stilistisch ähnlich. Für diesen Befund sind nun mehrere Deutungen möglich: Er könnte den Umstand reflektieren, dass die zeitliche Zuordnung der Autorenkorpora zu den vier Gruppen des späten

Sangs noch viel unsicherer ist als die der anderen Phasen, oder natürlich schlicht, dass sich eine solche zeitliche Unterscheidung eben nicht in einem klaren stilgeschichtlichen Verlauf abbildet. Vor diesem Hintergrund ist es vielleicht sogar überraschender, dass sich der hohe Sang dann doch vergleichsweise deutlich gruppiert und homogen bleibt, was den Verdacht nährt, dass sich hier ein weiterer Problemfaktor auswirken könnte, der gerade für mittelalterliche deutsche Texte von Belang ist: Nicht zu vergessen ist nämlich, dass die Texte des hohen und frühen Sang aus anderen Textausgaben bezogen sind als der späte Sang (Minnesangs Frühling bzw. die Walther-Ausgabe versus Carl von Kraus' Liederdichter oder die Schweizer Minnesänger). Die (relativ) starke Absonderung könnte also schlicht unterschiedliche Gepflogenheiten der Texteinrichtung in den Editionen reflektieren. Diese Editionsgepflogenheiten setzen wiederum, in mehr oder weniger starker Form, auf den handschriftlichen Quellen mit ihrer nicht regulierten Schreibung auf.

Um dem Zusammenhang eines möglichen Einflusses des Ausgabensignals nachzugehen, habe ich die einzelnen Teilkorpora automatisiert in normalisierte Texte umgewandelt, unterschiedliche Schreibungen also auf eine einheitliche Wortform reguliert. Bis vor kurzem war ein solcher *Preprocessing*-Schritt nur für moderne Texte denkbar, da entsprechende Programme für das Mittelhochdeutsche noch nicht greifbar waren. Mittlerweile liegt aber mit dem von Helmut Schmid entwickelten, auf *Deep-Learning*-Verfahren beruhenden RNNTagger (Recurrent Neural Network Tagger) (Schmid 2019) ein Tool vor, das auch für diese Sprachstufe durchaus brauchbare Ergebnisse liefert. Das Mittelhochdeutsch-Modell des RNNTaggers wurde auf dem Referenzkorpus Mittelhochdeutsch (Klein [u. a.] 2016) trainiert und bietet dementsprechend die Möglichkeit, Texte zu normalisieren, zu lemmatisieren und mit Wortarten-Labels zu versehen (*POS-Tagging*). Zwar können solche Tools niemals absolute Genauigkeit bieten, doch sind die ersten Ergebnisse des Taggers äußerst vielver-

sprechend - und einmal mehr gilt, dass von Unschärfen im Detail für die Makroperspektive zunächst einmal abgesehen werden soll.

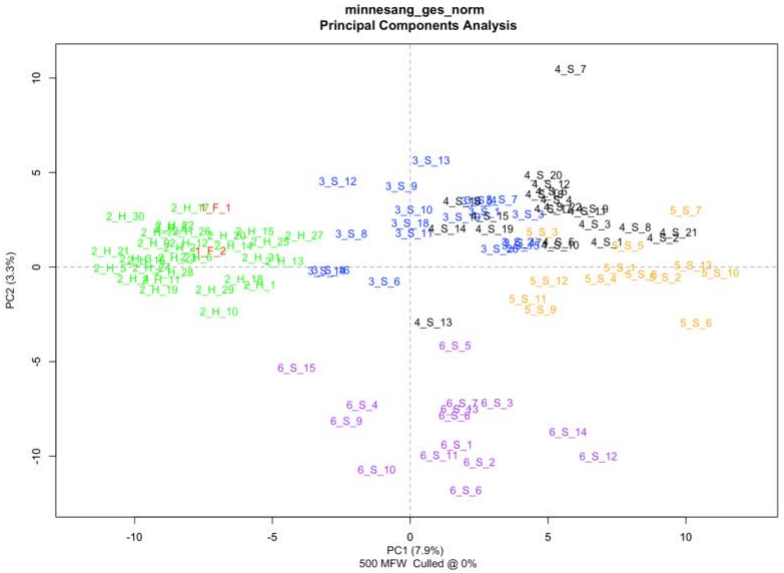


Abb. 4: PCA der Minnesangphasen (normalisiert)

Die PCA der normalisierten Texte (in Abb. 4) zeigt nun, dass sich nach dem *Preprocessing* die Gruppen auch im späten Sang deutlicher abzuzeichnen beginnen.

Nun könnte die Tatsache, dass die Texte als Grundlage für diese Abbildung mit demselben Modell normalisiert worden sind, zur Annahme verleiten, dass sich hier ein genaueres Bild der stilistischen Verhältnisse abzeichnet, da allfällige unterschiedliche Wortformen auf ein und dieselbe Form gebracht werden.

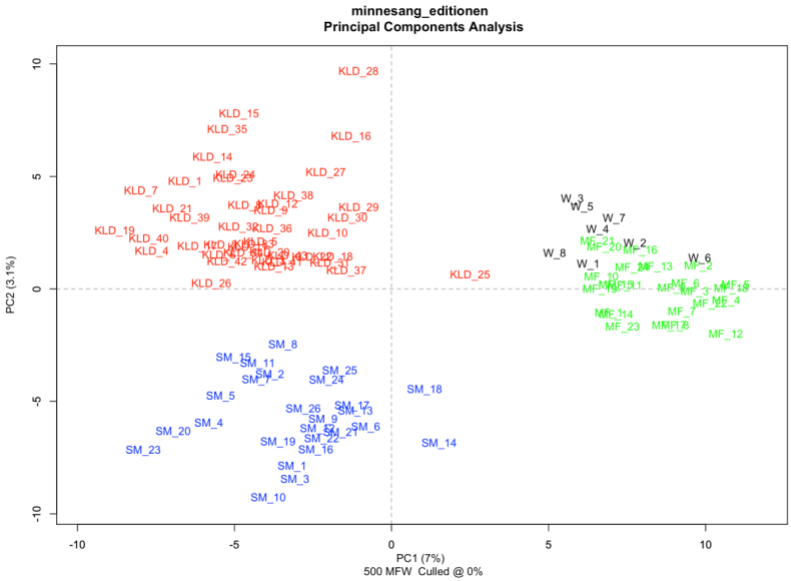


Abb. 5: PCA der Minnesang-Ausgaben (normalisiert)

Doch ist hier Vorsicht geboten, wie Abb. 5 zeigt: Hier habe ich das Gesamtkorpus nicht in zeitliche Unterkorpora aufgeteilt, sondern in solche nach Textausgaben.[8] Und hier wird nun ersichtlich, dass sich die Gruppen, die durch die Ausgaben vorgegeben sind, noch viel klarer abzeichnen (einzig Minnesangs Frühling und die Walther-Ausgabe haben Überschneidungen) und somit, dass das Ausgabensignal also doch eine erhebliche Rolle spielen dürfte.

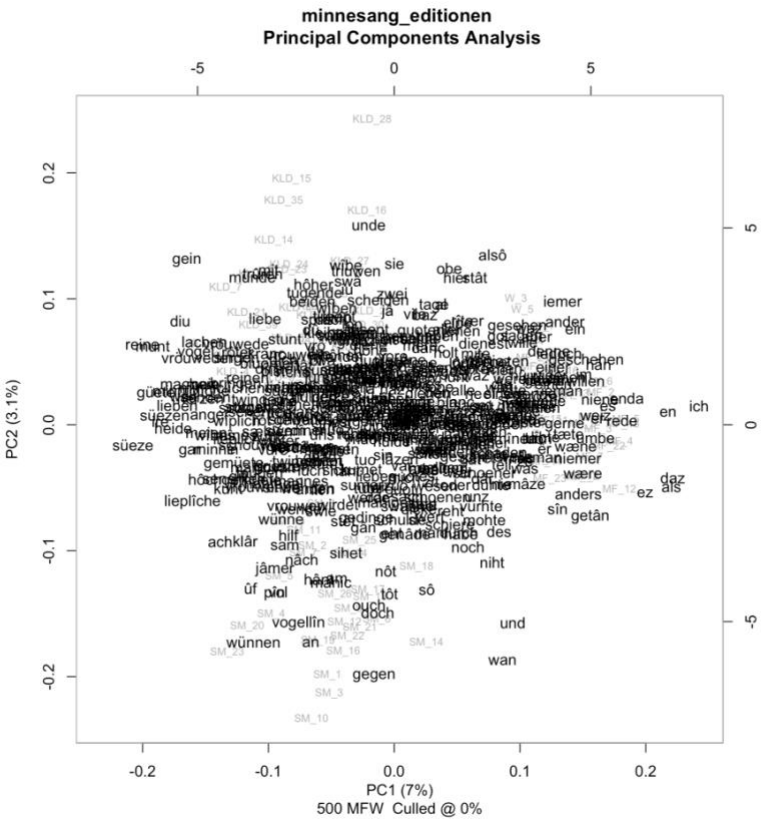


Abb. 6: PCA der Ausgabentexte mit *Loadings* der Wörter

Spätestens hier wäre es nun aufschlussreich zu wissen, welche Wörter für diese Sortierung ausschlaggebend sind. Dies lässt sich mit Hilfe eines sogenannten *Bi-Plots* eruieren, wie er in Abb. 6 ersichtlich ist. In dieser Darstellung wird nicht nur die Verteilung der Texte angezeigt (im Hintergrund in hellgrau), sondern auch die Verteilung der Wörter, die für die Ausprägung der Hauptkomponenten verantwortlich sind. So wird etwa rasch ersichtlich, dass sich die Ausgaben in Bezug auf die y-Achse (also die zweite

Hauptkomponente) z. B. durch die Oppositionen *unde* versus *und* oder *gegen* versus *gein* unterscheiden, die trotz der Normalisierung aufgrund ihrer metrischen Struktur als unterschiedliche Wortformen weitergeführt werden.

Allerdings zeigt die Abbildung auch, dass manche der Wortoppositionen doch auch über bloße Wortvarianten hinausgehen dürften. In Bezug auf die erste Hauptkomponente, die hohen vom späten Sang unterscheidet, nimmt das Wort *Ich* wieder eine prominente Stellung ein, als Gegenpol begegnen auf der rechten Seite Ausdrücke wie *lieplîche*, *süeze* und *süezen*, die sich wohl dem Frauenpreis zuordnen lassen. Dieser gerät also in Opposition zur Ich-Aussage. Ebenfalls auffällig ist die Extremposition von Ausdrücken, die sich dem Natureingangstopos zuordnen lassen (etwa *vogellîn* und *heide*). Wie später noch deutlich wird, stellen diese beiden Gruppen überhaupt ein starkes stilistisches Signal des späten Sangs dar.

3.4 Lexikalische Dichte: *Type Token Ratio* (TTR) und *Measure of Textual Lexical Diversity* (MTLD)

Ein weiteres, relativ einfaches Maß zur Beschreibung eines Korpus bietet die *Type-Token-Ratio* (TTR), die das Verhältnis der Gesamtworte (*tokens*) zu den einzelnen Wortformen (*types*) angibt.[9] Dabei kommt insbesondere die Variabilität im Wortschatz in den Blick: Werden in einem Text immer dieselben Wörter verwendet oder viele unterschiedliche Wörter? Der Satz ›ich bin ich‹ würde beispielsweise aus drei *tokens* ›ich‹, ›bin‹, ›ich‹), aber nur zwei *types* bestehen (›ich‹, ›bin‹). Die TTR beträgt demnach zwei Drittel und ist damit geringer als bei dem Satz ›Ich bin Stiller‹, der sich somit als lexikalisch reichhaltiger beschreiben lässt.

Eine unmittelbare Vergleichbarkeit der TTR wird jedoch dadurch erschwert, dass diese nicht unabhängig von der Länge eines Textes ist: Je länger ein Text ist, desto eher wird es vorkommen, dass sich Wörter wiederholen, wodurch die TTR bei längeren Texten automatisch sinkt. Zur Nor-

mierung dieses Textlängenproblems wurden daher auf Basis der TTR weitere, elaboriertere Maßzahlen entwickelt, wie etwa das *Measure of textual lexical diversity* (MTLD, McCarthy 2005). Mit dieser Methode wird die durchschnittliche Länge von Wortsequenzen berechnet, die über einem bestimmten Schwellenwert der TTR liegen. Die Funktionsweise der MTLD besteht also konkret darin, dass der Text Wort für Wort durchgegangen wird und bei jedem Wort die aktuelle TTR berechnet wird. Beim ersten Wort liegt die TTR notwendigerweise noch bei 100 Prozent (das erste *token* muss automatisch auch *type* sein); je weiter man im Text voranschreitet, desto öfter wird es aber vorkommen, dass sich Wörter wiederholen. Passiert diese Wiederholung so oft, dass die TTR unter einen Schwellwert fällt (der normalerweise bei 0,72 liegt), wird der Durchgang durch den Text gestoppt, notiert, wie viele Wörter man vorangekommen ist, und die Zählung von neuem gestartet. Ist der Text durchgearbeitet, kann man dann den Mittelwert der notierten Textsequenzen-Längen berechnen, der die MTLD angibt.

	Zeit	Autor	Zeichen	Tokens	Types	Mtld	Wortlänge
49	4	Konrad von Wuerzburg	21289	3742	1057	203	5.69
64	5	Schulmeister von Esslingen	2134	389	245	197	5.49
68	5	Der Kanzler	11574	2018	762	185	5.74
82	5	Goeli	6776	1213	595	177	5.59
94	6	Albrecht Marschall von Raprechtswil	2660	508	277	175	5.24
6	2	Engelhart von Adelnburg	869	172	123	173	5.05
101	6	Ulrich von Baumburg	8370	1575	654	173	5.31
78	5	Konrad von Kirchberg	6960	1279	539	169	5.44
80	5	Walther von Breisach	1340	243	166	168	5.51
66	5	Der Duering	6363	1167	521	164	5.45

Abb. 7: Statistiken für die zehn Autorenkorpora mit dem höchsten MTLD-Wert

Abb. 7 zeigt nun jene Autorenkorpora im Gesamtkorpus an, die den höchsten MTLD-Wert aufweisen.[10] Besonders hervorstechend ist Konrad von Würzburg mit einer MTLD-Score von 203, es folgen der Schulmeister von Esslingen (197) sowie der Kanzler (185). Dieser Befund ist nun kaum zufällig, sondern hängt vermutlich mit der Tatsache zusammen, dass die genannten Autoren allesamt auch als Verfasser von Sangspruchdichtung in Erscheinung getreten sind.[11] Es scheint sehr wahrscheinlich, dass sich die weniger monothematische Sangspruchproduktion dieser Dichter auch auf ihren Minnesang ausgewirkt hat.[12] Dass für diese eher als ›professionell‹ anzusehenden Dichter andere Maßstäbe gelten, hat auch Hübner (2013, S. 397) konzediert, der letztlich nur für diese Fälle die Kuhn'sche Objektivierungsthese in Ansätzen gelten lassen möchte.

Von den MTLD-Werten schließt sich zudem auffälligerweise wieder der Bogen zu den in Kapitel 3.2 dargestellten *ich*-Frequenzen, denn auch dort wird die Reihenfolge der Autoroeuvres mit der geringsten *ich*-Frequenz vom Kanzler (0.004) und von Konrad von Würzburg (0,007) angeführt, der Schulmeister von Esslingen nimmt Platz fünf dieses Rankings ein (0,015). Zudem finden sich auf den vorderen Plätzen weitere Sangspruch-erprobte Namen wie Walther von Breisach (0,008), aber auch der ausschließlich als Minnesänger bekannte Goeli (0,01), der ebenfalls einen der höchsten MTLD-Werte aufweist (177). Lexikalische Vielfalt und Abkehr vom Ich gehen also hier offensichtlich Hand in Hand und scheinen in vielen Fällen durch den Einfluss der Sangspruch-Form bedingt.

3.5 Wordclouds und Distinktivitätsmaße: TF/IDF, *Log Likelihood* und Burrows Delta

Wordclouds bieten eine mittlerweile weit verbreitete Möglichkeit, die MFW eines Korpus zu visualisieren.

Abb. 8: Wordcloud der MFW des Gesamtkorpus

Abb. 8 zeigt etwa eine solche Wordcloud der häufigsten Wörter im Gesamtkorpus des Minnesangs.[13] Die Schriftgröße der Wörter ist dabei in Relation zur Frequenz ausgeführt, je häufiger ein Wort im Korpus auftaucht, desto größer wird es dargestellt.[14] Es zeigt sich hier der oben angesprochene Befund, dass Funktionswörter die häufigsten in einem natürlichsprachlichen Text auftretenden Wörter sind. Im Minnesang dominieren insbesondere das Personalpronomen *ich* und dessen abgeleitete Formen.

Um einen besseren Einblick in den Inhalt der Texte zu haben, lassen sich nun, wie das bei vielen Verfahren der Textanalyse üblich ist, Funktionswörter mit Hilfe von sogenannten Stoppwortlisten aus der Untersuchung ausschließen.

Abb. 9: Wordcloud der MFW des Gesamtkorpus ohne Stoppwörter

Abb. 9 zeigt die Wordcloud der MFW des Gesamtkorpus nach Anwendung einer solchen Stoppwortliste. Übrig bleiben (durchaus erwartungsgemäß) die Leitbegriffe des Minnesangs, *minne*, *herze*, *vröude* und *vrouwe*. Auch ein wenig *leit* und *swaere* ist dabei (aber nicht so ausgeprägt wie die Freude-Komponente). Hinzu kommen schließlich Naturbegriffe (*bluomen, heide*) und Ausdrücke der Sinneswahrnehmung (*ougen*, *sinne*).

Abb. 10: Wordclouds der MFW der Einzelkorpora ohne Stoppwörter

Zoomt man von dieser Makroperspektive einen Schritt hinein und erstellt wie in Abb. 10 Wordclouds für die sechs zeitlichen Teilkorpora, dann ergibt sich ein differenziertes Bild, aus dem sich aber zunächst nur wenig Schlüsse ziehen lassen: Die Leitworte des Minnesangs (*minne, vrouwe, vröude)* dominieren auch hier und bleiben über die ganze Phase des Sangs relevant. Im frühen Sang ist *herze* das häufigste Wort, im hohen Sang *wîp* und in allen Phasen des späten Sangs *minne*. Mehr noch als Veränderung zeigen die *Wordclouds* damit also eine überraschende Konstanz an, Konvention statt Transformation und Wandel.

Eine Möglichkeit, die Unterscheidung der Teilkorpora genauer zu explorieren, bietet das aus dem Kontext des Information Retrieval stammende TF/IDF-Maß, mit dessen Hilfe distinktive Wörter der Subkorpora eruiert werden können (Spärck Jones 1972; zur Anwendung im altgermanistischen Kontext Braun/Reiter 2017). TF steht für *Term Frequency*, IDF für die

Inverse Document Frequency. TF/IDF blickt also ebenso wie eine normale Frequenzauszählung zunächst darauf, wie häufig ein Wort vorkommt, setzt dieses Vorkommen aber dazu in Relation, wie ungewöhnlich es ist, dass das Wort öfter auftritt. Dazu wird die TF durch die IDF geteilt, das heißt also, durch die Anzahl der anderen Dokumente aus dem Gesamtkorpus, in der das Wort auch vorkommt. In der TF/IDF-Darstellung (Abb. 11) zeichnen sich nun doch einige zeittypische Wörter ab, z. B. die *merkaere* für den frühen Sang oder der *reien* für die vorletzte Phase des späten Sangs.

Abb. 11: Wordclouds der Teilkorpora gewichtet nach TF/IDF

Schließlich können mit einem weiterem Schritt in das Material hinein die distinktiven Wörter für einzelne Autorenkorpora eruiert werden. Abb. 12–14 zeigen einige Beispiele, bei denen die Distinktivität nicht mit TF/IDF, sondern mit dem *log-likelihood*-Score berechnet wurde.[15] Dieser Score gibt an, ob ein Wort in einem Textkorpus signifikant häufiger erscheint als in den Vergleichskorpora.

Abb. 12: Wordcloud *log likelihood* Wolfram von Eschenbach

Abb. 13: Wordcloud *log likelihood* Reinmar

Abb. 14: Wordcloud *log likelihood* Bernger von Horheim

Bei Wolfram ragt der *urloub* hervor, bei Reinmar die *rede*, bei Bernger von Horheim die *liuge*.

Als nächste Annäherungsstufe könnte man nun in die Einzeltexte selbst hineingehen, und diese qualitativ lesen und interpretieren. Auch wenn der Computer die Makroperspektive betont, braucht man nämlich bei dieser nicht stehenzubleiben. Distant Reading erscheint so nicht als ausschließende Alternative zum Close Reading, sondern die beiden Formen bilden die äußersten Enden eines Kontinuums, zwischen denen man sich in fortwährender Perspektivenverschiebung hin und her bewegen kann. Der Anglist Martin Mueller (2014) hat einen solchen Zugang als Scalable Reading bezeichnet, der also die beiden Extrempositionen der überblicksmäßigen, quantitativen und der detailversessenen, qualitativen Lektüre vereint, und zwar vereint in der Dynamik der Bewegung zwischen den Perspektiven, mithin also in für die Geisteswissenschaften durchaus willkommener Multiperspektivität.

Systematisch gesehen ließen sich TF/IDF und *log likelihood* als Distinktivitätsmaße verstehen, die nicht wie etwa die bloße Auszählung von MFW allgemein auf Wortfrequenzen fokussieren, sondern insbesondere die Unterschiede zwischen zwei Texten bzw. Textgruppen in ihrem Wortgebrauch in den Blick nehmen. Ein weiteres, sehr intuitives und mathematisch relativ einfaches Verfahren stellt Burrows' Zeta dar (Burrows 2007), das anders als die anderen Methoden nicht im Kontext der Computerlinguistik oder des Information Retrieval entwickelt wurde, sondern aus den Digital Humanities selbst stammt (Schöch 2018, S. 81). Zeta zielt nicht auf die absolute Häufigkeit des Wortgebrauchs ab, sondern auf dessen Konsistenz, und zwar auf die Konsistenz des Wortgebrauchs eines Vergleichstextes im Verhältnis zu einem Zieltext (bzw. zwischen entsprechenden Korpora). Das Verfahren ist relativ einfach: Zuerst wird der Vergleichstext in einzelne Abschnitte (etwa zu 2000 *tokens*) zerlegt und dann für jedes Wort ausgezählt, ob und in wie vielen Abschnitten es vorkommt. Dann wird dieselbe Prozedur auf den Zieltext angewendet, und schließlich werden die Werte voneinander abgezogen. Daraus ergibt sich ein Zeta-Score, der zwischen 1 und -1 liegen kann: 1 würde (den in der Praxis natürlich kaum auftretenden Fall) bedeuten, dass das Wort in jedem einzelnen Abschnitt des Vergleichstextes vorkommt und in keinem Abschnitt des Zieltextes. -1 gibt den umgekehrten Fall an. Reiht man die Wörter anschließend nach ihrem Zeta-Score, ergibt sich eine Liste der vom Vergleichstext gegenüber dem Zieltext bevorzugten Wörter, die zugleich die Liste der vom Zieltext vermiedenen Wörter darstellt. Diese sind nur mehr selten Funktionswörter, da deren besonders häufiges Auftreten durch die Abschnittszählung (alle Vorkommnisse zählen pro Abschnitte genau nur wie ein Beleg) abgemildert wird. Damit geraten semantisch leichter zu deutende Wörter aus dem mittleren Frequenzspektrum in den Blick.

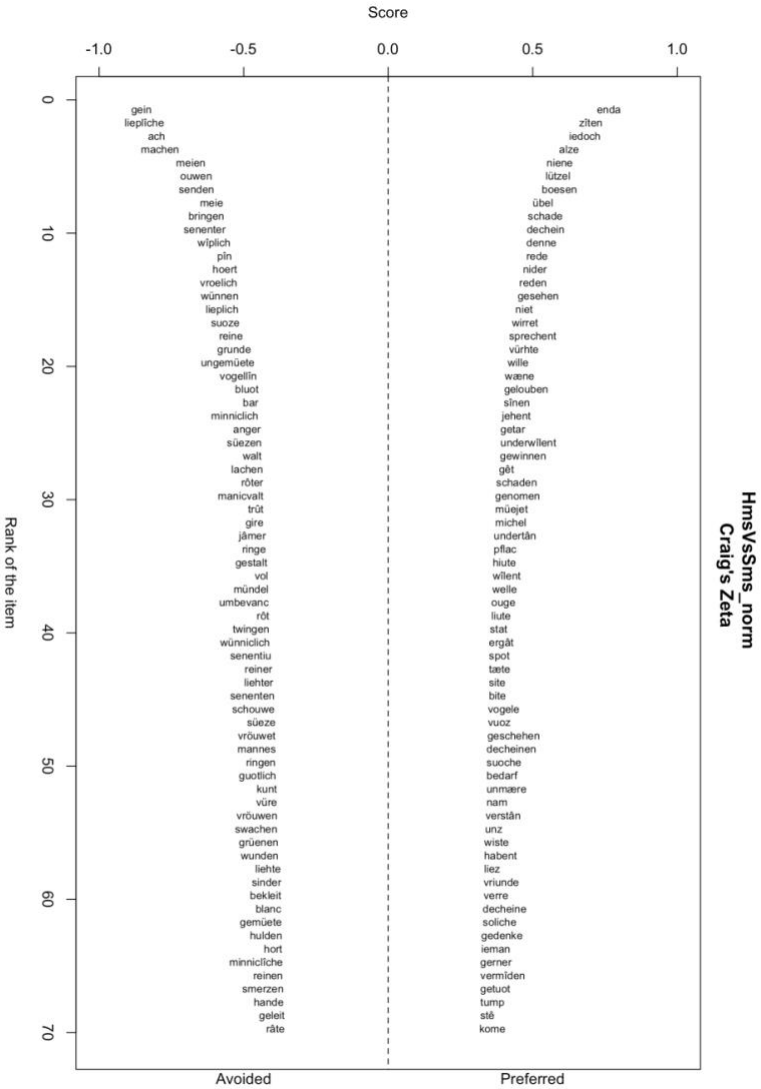


Abb. 15: Bevorzugte bzw. vermiedene Wörter zwischen hohem bzw. frühem Sang und spätem Sang

Zur Erstellung von Abb. 15 habe ich den frühen und hohen Minnesang als Vergleichstextkorpus den vier Teilkorpora des späten Sangs gegenübergestellt.[16] Rechts befinden sich die von der ersten Phase des Minnesangs bevorzugten Wörter, links die vermiedenen. Der Zeta-Score ist dabei auf der x-Achse abgetragen. Je weiter ein Wort also von der Mittellinie entfernt ist, desto deutlicher gehört es zu einer der beiden Gruppen. Die y-Achse gibt das Ranking der Wörter wieder, die einen Zeta-Score von 0,3 über- bzw -0,3 unterschreiten.

Die genauere Durchsicht bringt auf der Seite der in der Früh- und Hochphase vermiedenen (und damit in der Spätphase bevorzugten) Wörter den einheitlicheren Befund: Immer wieder begegnen hier Wörter, die mit dem Natureingangstopos in Verbindung zu bringen sind (*meie, meien, ouwen, vogellîn*), der sich ja schon in anderen Analysen als typisches Formmerkmal der Spätphase erwiesen hat.[17] Wörter wie *wîplich, lieplich, suoze, reine* deuten auf den Frauenpreis hin. Einen interessanten Einzelbefund stellt die Spitzenstellung der Interjektion *ach* dar, die offenkundig im späteren Sang häufiger auftritt. Als Wort mit dem ausgeprägtesten negativen Zeta-Score begegnet aber das schon bei der PCA der nach Editionen geordneten Texte zu Tage getretene einsilbige *gein*. Die vom Vergleichskorpus bevorzugten Wörter präsentieren sich deutlich uneinheitlicher. Auch hier steht an der Spitze mit *enda* vermutlich ein Ausgaben-Artefakt, danach begegnen vermehrt negative Wörter (*boesen, übel, schade*), auch das Wortfeld *rede*, *reden* und *sprechent* wird in den früheren Texten konsistent häufiger verwendet.

3.7 Distributionelle Semantik und Topic Models

Das wohl am ehesten geeignete Framework, um mit dem Computer über die reine Textoberfläche hinaus in tiefere Bedeutungsschichten vorzudringen, bietet das Theoriegebäude der distributionellen Semantik. Die distributionelle Semantik geht davon aus, dass sich die Bedeutung eines Wortes

nicht (oder nicht nur) aus sich selbst ergibt, sondern aus dem Kontext, in dem es erscheint. Der Linguist John Rupert Firth hat diese Grundannahme in dem berühmten Zitat »you shall know a word by the company it keeps« zusammengefasst (Firth 1957, S. 11). Dieser Umstand bietet nun auch dem Computer die Chance, komplexere und auch implizite Bedeutungen zu erkennen. So wird es etwa gemäß der distributionellen Hypothese für den Computer möglich, zwischen unterschiedlichen Bedeutungen des oberflächlich gleichen Wortes ›Bank‹ zu differenzieren: Begegnet dieses überwiegend im Kontext von anderen finanzbezogenen Ausdrücken, so ist der Bedeutungsaspekt von Bank als Geldinstitut wahrscheinlich, begegnen hingegen im Umfeld Naturausdrücke, so ist die Bedeutung als Parkbank eher anzunehmen. Dieses Beispiel zeigt bereits, dass auch die distributionelle Semantik nicht hundertprozentig sichere Ergebnisse liefern kann, sondern mit mehr oder minder großen Wahrscheinlichkeiten zu rechnen hat.

Eines der bekanntesten Verfahren, das auf den Grundsätzen der distributionellen Semantik beruht, ist das Topic Modeling, mit dessen Hilfe die Themenstruktur von größeren Korpora nachgezeichnet werden kann (Blei [u. a.] 2003; ausführlich zum Verfahren Horstmann 2018). Beim Topic Modeling wird von dem gemeinsamen Auftreten bestimmter Wörter auf zugrundeliegende Themen- oder besser gesagt *topic*-Cluster geschlossen.[18] So deutet etwa der Umstand, dass in einem Text sehr oft die Wörter ›Fisch‹, ›Boot‹, ›Netz‹ gemeinsam auftreten, darauf hin, dass es in diesem Text um ein gemeinsames Thema geht, das sich als ›Fischerei‹ benennen lässt. Ein Text kann und wird dabei durchaus mehrere solcher *topics* (ausgeprägt) aufweisen; ebenso kann ein Wort in unterschiedlichen *topics* (prominent) auftreten.[19]

Beim Topic Modeling handelt es sich um ein unüberwachtes Machine-Learning-Verfahren, das heißt, der Computer fügt in mehreren Trainingsdurchgängen, bei der die Kontextwahrscheinlichkeiten immer besser eingeschätzt werden, selbständig bestimmte Wörter zu Themenclustern zusammen. Wie viele solcher Themencluster angesetzt werden sollen, muss

jedoch von der menschlichen Benutzer*in vorgegeben werden; auch die Interpretation und letztendliche Benennung der *topics* bleibt dem Menschen überlassen. Der Computer stellt also nur fest, welche Wörter überzufällig oft gemeinsam auftreten, was auf eine latente Variable, nämlich einen semantischen Zusammenhang in einem *topic*, hinweisen könnte.

Abb. 16: Wordclouds der *topics* des Minnesangkorpus

Abb. 16 zeigt ein solches Topic Model des Minnesangs.[20] Ich habe vom Computer 15 *topics* berechnen lassen. Die einzelnen Topics sind in Wordcloud-Darstellungen visualisiert, wobei hier nun in proportionaler Größe jene Wörter angezeigt werden, die am meisten zur Formierung eines *topics* beitragen (und damit für das Thema am bedeutendsten sind). Neben erwartbaren Minne-spezifischen Themenclustern zeigen sich im Topic Model einige interessante Einzelcluster, etwa Topic Nr. 14, das mit Wörtern wie *sprach, liebe, tac, scheiden, naht, klage, wahtaer* und *ritaer* die Konstellation des Tageliedes reflektiert (vierte Wortcloud in der letzten Reihe) oder *topic* Nr. 5, das mit Wörtern wie *bluomen, winter, meien, vogel, heide, sumer* dem Natureingangstopos entspricht.

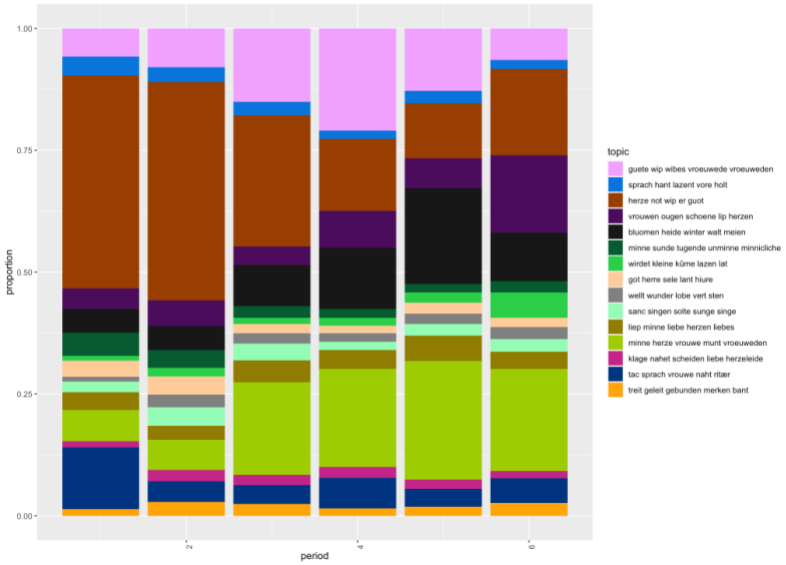


Abb. 17: Zeitliche Verteilung der *topics* im Minnesangkorpus

Auch im Kontext der Topic-Model-Analyse lässt sich demonstrieren, dass der Natureingang für den späten Sang kennzeichnend ist. Abb. 17 zeigt den zeitlichen Verlauf der Konjunktur der fünfzehn *topics*, die die Maschine berechnet hat. Die sechs Spalten auf der x-Achse stehen für die sechs zeitlichen Teilkorpora ein, auf der y-Achse repräsentieren die unterschiedlichen Farben den Anteil der 15 Topics an den Texten der Zeitstufe. Je breiter das Band eines *topics* ausfällt, desto mehr Textabschnitte lassen sich ihm mit höherer Wahrscheinlichkeit zuordnen. Und hier zeigt sich nun, dass das in Schwarz dargestellte *topic* 2, das den Natureingang abbildet, im Verlauf des späten Sangs immer mehr zunimmt und im fünften Abschnitt seinen Höhepunkt erreicht.

Andere, sich ebenfalls stark verändernde *topic*-Bänder sind weniger einfach zu deuten. So ist die Frühzeit etwa durch einen hohen Anteil von *topic* 3 geprägt, in dem die Wörter *herze, not, wip, êr* und *guot* dominieren. Daneben zeichnen sich jedoch Begriffsfelder der Klage (*leit*) ebenso wie der

vröuwede ab. Das *topic* scheint im weiteren Verlauf der Gattung durch *topic* 13 (*minne, herze, vrouwe, munt, vrouwe*) abgelöst zu werden.

4. Fazit

Mit meinem Durchgang durch einige Methoden der digitalen Textanalyse wollte ich demonstrieren, dass diese zwar notwendigerweise zur Darstellung größerer Zusammenhänge tendieren und damit Details nivellieren, dass man gerade mit digitalen Methoden aber auch die Perspektivenabhängigkeit dieser großen Zusammenhänge herausstellen kann. So lässt sich zwar die ›Geschichte‹ des späten Minnesangs als Transformation beschreiben, zugleich aber auf die immer noch gegebene Vielfältigkeit der stilistischen Muster verweisen.

Dabei haben sich deutlich spezifische Probleme gezeigt, die sich gerade bei der Anwendung digitaler Methoden auf mittelalterliche volkssprachige Literatur ergeben. Zwar ist es mittlerweile möglich, durch die Entwicklung des RNNTaggers Texte zu normalisieren und damit den ›Störfaktor‹ der uneinheitlichen Schreibung des Mittelhochdeutschen abzumildern. Doch wirkt die ungeregelte Orthographie, wie es scheint, trotzdem noch nach, etwa wenn unterschiedliche Textausgaben unterschiedliche Wortformen präferieren oder sich diese aus metrischen Gegebenheiten ergeben. Wie die PCA der nach Ausgaben gegliederten Texte gezeigt hat, dürften Störfaktoren im stilistischen Signal, die über eine ›normale‹ Stilmischung moderner Texte hinausreichen, nicht zu vernachlässigen sein.

Zudem stellt sich das Problem der Kategorienbildung: Zuweisungen von Texten zu Autorkorpora und deren Datierung sind auf unsicherem Boden gebaut. Dennoch bieten sie einen Ausgangspunkt, dessen Unzuverlässigkeit gerade durch digitale Methoden (wie etwa mit Hilfe der Herausstellung einer uneinheitlichen Verteilung der Textpunkte in der PCA) ausgestellt werden kann und unbedingt auch werden soll: Gerade die Abweichungen

von der Idealvorstellung des statistischen Modells sind besonders aufschlussreich.

Blickt man auf den inhaltlichen Ertrag, den die Analysen erbracht haben, dann zeigen sich immer wieder Einzelergebnisse, die vielleicht nicht grundsätzlich neu sind (das Korpus des Minnesangs, das hier exemplarisch herangezogen wurde, ist letztlich nicht so groß, dass es nicht auch qualitativ zu überschauen wäre), aber durch den stilistischen Befund bestehende Vermutungen bestärken können, neue Anregungen bieten und möglicherweise Offensichtliches, aber doch bislang Übersehenes zu Tage bringen. So konnte die Frequenzanalyse in Zusammenspiel mit der Untersuchung der lexikalischen Dichte den Verdacht erhärten, dass Transformationen des Minnesangs insbesondere bei Autoren auftreten, die auch als Sangspruchdichter in Erscheinung getreten sind und vermutlich von dort neue Formen in den Minnesang eingebracht haben. Es wurde zudem deutlich, wie stark der Natureingang als stilistisch-thematisches Spezifikum des späten Sanges in Erscheinung tritt. Und schließlich bieten die gezeigten Verfahren Anknüpfungspunkte für ein Scalable Reading der Texte, das von den Befunden auf der quantitativen Makro-Ebene seinen Ausgang nehmen und von dort auch wieder auf die qualitative Mikro-Ebene zurückführen kann.

Digitale Methoden sind so gesehen nicht der Endpunkt der Interpretation, sondern im Gegenteil erst der Ausgangspunkt; sie liefern keine endgültige Ergebnisse, die ohne Weiteres hinzunehmen sind, sondern Anregungen, die sich gerade aus der Friktion der digitalen Modelle mit den analogen Gegenständen ergeben können.

Anhang A: Verzeichnis der erfassten Autorenkorpora

ID	Abschnitt	Autor	Lieder	Zeichen	Tokens	Types	Ausgabe
1	1	Der von Kürenberg	2	2831	556	286	MF
2	1	Burggraf von Regensburg	2	876	171	109	MF
3	1	Dietmar von Eist	16	10785	2097	640	MF
4	1	Burggraf von Rietenburg	2	1844	361	194	MF
5	1	Meinloh von Sevelingen	3	4198	799	351	MF
6	1	Kaiser Heinrich	3	2071	409	218	MF
7	2	Engelhart von Adelnburg	2	869	172	123	MF
8	2	Ulrich von Gutenburg	4	12382	2468	695	MF
9	2	Friedrich von Hausen	17	15594	3122	747	MF
10	2	Heinrich von Veldeke	37	15068	2922	842	MF
11	2	Walther von der Vogelweide	76	91975	17857	2485	WL
12	2	Rudolf von Fenis	8	8193	1627	493	MF
13	2	Albrecht von Johansdorf	13	12528	2431	728	MF
14	2	Heinrich von Rugge	12	14158	2770	727	MF
15	2	Bernger von Horheim	6	5165	1023	376	MF
16	2	Hartwig von Rute	4	2020	400	207	MF
17	2	Bligger von Steinach	2	1486	299	178	MF
18	2	Heinrich von Morungen	35	31605	6231	1285	MF
19	2	Reinmar	70	79610	15866	1909	MF
20	2	Hartmann von Aue	18	18407	3631	882	MF
21	2	Gottfried von Straßburg	2	2893	544	300	MF
22	2	Wolfram von Eschenbach	9	10022	1898	699	MF
23	3	Hiltbolt von Schwangau	22	13014	2609	644	KLD
24	3	Otto von Botenlauben	12	6885	1357	500	KLD
25	3	Rubin	21	23699	4556	980	KLD
26	3	Der tugendhafte Schreiber	11	10951	2059	684	KLD
27	3	Gottfried von Neifen	51	56230	10449	1461	KLD
28	3	Burkhard von Hohenvels	18	21670	4005	1133	KLD

29	3	Der Markgraf von Hohenburg	6	4567	924	322	KLD
30	3	Wachsmuot von Künzingen	7	5673	1100	420	KLD
31	3	Christan von Hamle	6	5846	1075	477	KLD
32	3	Friedrich der Knecht	5	6128	1191	449	KLD
33	3	Friedrich von Leiningen	1	1575	308	189	KLD
34	3	Heinrich von Anhalt	2	1665	323	190	KLD
35	3	Rudolf von Rotenburg	11	11071	2127	618	KLD
36	3	Ulrich von Munegiur	3	2467	494	237	KLD
37	3	Walther von Mezze	10	11060	2152	679	KLD
38	3	Hesso von Rinach	2	1349	254	166	SM
39	3	Ulrich von Singenberg	31	33457	6445	1244	SM
40	4	Markgraf Heinrich von Meißen	6	4455	843	373	KLD
41	4	Hugo von Werbenwag	5	4149	763	368	KLD
42	4	Herrand von Wildonie	3	2200	431	231	KLD
43	4	Der Kol von Niunzen	4	1475	285	174	KLD
44	4	Reinmar von Brennenberg	5	12028	2243	717	KLD
45	4	Der Schenk von Limburg	6	6545	1235	462	KLD
46	4	Ulrich von Liechtenstein	59	74826	14224	1869	KLD
47	4	Ulrich von Winterstetten	40	54613	10236	1643	KLD
48	4	Bruno von Hornberg	4	3507	680	313	KLD
49	4	Burggraf von Lienz	2	2633	507	263	KLD
50	4	Konrad von Würzburg	23	21289	3742	1057	KW
51	4	Der von Sachsendorf	7	5965	1147	433	KLD
52	4	Wachsmuot von Mühlhausen	5	2830	535	271	KLD
53	4	Waltram von Gresten	3	1787	359	197	KLD
54	4	Willehelm von Heinzenburg	5	2931	569	286	KLD
55	4	Der von Stadegge	3	2057	392	218	KLD
56	4	Der von Suonegge	3	2145	401	199	KLD
57	4	Der von Wissenlo	4	2704	528	249	KLD
58	4	Günther von dem Forste	6	9966	1909	538	KLD

59	4	Heinrich von der Mure	3	2266	463	232	KLD
60	4	König Konrad der Junge	2	1323	250	147	KLD
61	4	Rudolf der Schreiber	3	3633	701	327	KLD
62	4	Heinrich von Sax	4	5272	1004	395	SM
63	4	Walther von Klingen	8	7402	1362	494	SM
64	5	Der wilde Alexander	2	2522	480	245	KLD
65	5	Schulmeister von Esslingen	2	2134	389	245	KLD
66	5	Brunwart von Augheim	5	3453	654	294	KLD
67	5	Der Düring	7	6363	1167	521	KLD
68	5	Der Dürner	1	1489	289	178	KLD
69	5	Der Kanzler	12	11574	2018	762	KLD
70	5	Der Püller	5	4196	777	346	KLD
71	5	Konrad von Landeck	22	32201	5942	1195	SM
72	5	Der von Buchein	3	1643	306	183	KLD
73	5	Der von Obernburg	7	5812	1104	410	KLD
74	5	Der von Scharpfenberg	2	2298	455	237	KLD
75	5	Der von Stammheim	1	3271	633	322	KLD
76	5	Hartmann von Starkenberg	3	1557	306	177	KLD
77	5	Herzog Heinrich von Breslau	2	2584	493	252	KLD
78	5	König Wenzel von Böhmen	3	4379	842	376	KLD
79	5	Konrad von Kirchberg	6	6960	1279	539	KLD
80	5	Markgraf Otto von Brandenburg	7	4088	752	371	KLD
81	5	Walther von Breisach	1	1340	243	166	KLD
82	5	Der Taler	3	3543	689	326	SM
83	5	Goeli	4	6776	1213	595	SM
84	5	Heinrich von Frauenberg	5	3863	743	322	SM
85	5	Heinrich von Stretelingen	3	2658	496	232	SM
86	5	Heinrich von Tettingen	2	1749	331	180	SM
87	5	Konrad von Altstetten	3	2845	553	280	SM
88	5	Kraft von Toggenburg	7	7487	1413	474	SM

89	5	Steinmar	14	14868	2818	801	SM
90	5	Winli	8	7500	1436	513	SM
91	6	Johannes Hadlaub	54	70350	13898	2022	SM
92	6	Christian von Lupin	7	5484	1072	439	KLD
93	6	Gösli von Ehenheim	2	2126	391	222	KLD
94	6	Heinrich Hetzbold von Weißensee	8	5697	1095	416	KLD
95	6	Albrecht Marschall von Raprechtswil	3	2660	508	277	SM
96	6	Der von Gliers	3	15153	2995	825	SM
97	6	Der von Trostberg	6	5217	967	432	SM
98	6	Heinrich Rost zu Sarnen	9	7639	1407	557	SM
99	6	Heinrich Teschler	13	15788	3033	851	SM
100	6	Jakob von Warte	6	7541	1403	479	SM
101	6	Otto zum Turm	5	5933	1104	455	SM
102	6	Ulrich von Baumburg	7	8370	1575	654	SM
103	6	Wernher von Hohenberg	8	5068	1010	396	SM

Abschnitte:

1 – Früher Minnesang

2 – Hoher Minnesang

3 – Später Minnesang 1 (Anfang 13. Jh.)

4 – Später Minnesang 2 (Mitte 13. Jh.)

5 – Später Minnesang 3 (Ende 13. Jh.)

6 – Später Minnesang 5 (Ende 13. / Anfang 14. Jh.)

Anmerkungen

1 Vgl. hierzu die klassische Beschreibung der Grenze zwischen nomothetischen und idiographischen Wissenschaften bei Windelband 1910.

2 Schon durch diese Explizierung können sich wissenschaftliche Mehrwerte ergeben, vgl. dazu Gius/Jacke 2015.

3 Es versteht sich von selbst, dass durch eine solche Auswahl keine Vollständigkeit zu erreichen ist, wie sie etwa Moretti vorgeschwebt haben mag (zur Frage, wa-

rum eine solche Vollständigkeit ohnedies Chimäre bleibt, vgl. Rosen 2011). Da die unterschiedliche Gestaltung von Textausgaben nicht unerheblichen Einfluss auf textanalytische Methoden hat (siehe dazu unten), habe ich versucht, die Zahl der Ausgaben einigermaßen klein zu halten, und dies bei größtmöglicher Abdeckung der Minnesangproduktion (eine größere Leerstelle stellt jedenfalls der überlieferungsgeschichtlich schwierige Neidhart dar, sonst ist ein Großteil der Minnesangproduktion vertreten). Eine genaue Aufstellung der berücksichtigten Autoren findet sich im Anhang.

4 Die Trendlinie wird mit der Methode der kleinsten Quadrate mithilfe der Funktion linregress des Python-Scipy-Packages berechnet. Ich habe dafür auf den Code zurückgegriffen, der in Karsdorp [u. a.] 2021, S. 21, beschrieben ist.

5 Die Darstellung wurde mit dem R-Package *stylo* von Eder [u. a] 2013 erstellt (Parameter: 500 MFW, Sample-Größe 2000 Wörter, correlation PCA). Zum Verfahren und dessen Anwendung in der Stilistik vgl. Craig/Greatley-Hirsch 2017.

6 Die Punkte werden dabei durch entsprechende Namenskürzel angezeigt: »3_S_13« gibt also z. B. den Punkt für den 13. Abschnitt von 2000 Wörtern im dritten der sechs Teilkorpora an; es handelt sich also um Lyrik vom Anfang des 13. Jahrhunderts. Das »S« zeigt an, dass sich das Teilkorpus dem späten Sang zuordnen lässt, 3_S wäre also das erste Teilkorpus des späten Sangs.

7 Zu beachten ist bei diesem und den folgenden Beispielen, dass die Hauptkomponenten im Fall des Minnesangkorpus recht schwach ausgeprägt sind. Die erste Hauptkomponente steht in Abb. 3 gerade einmal für 8,8% der Varianz ein, deckt also nur weniger als ein Zehntel der Gesamtvarianz ab. In den folgenden Beispielen ist der Wert noch geringer.

8 Die Konrad von Würzburg-Ausgabe habe ich dabei aufgrund ihrer Kürze (im Vergleich zu den anderen Anthologien) und der Besonderheiten des Konrad-Lied-Korpus beiseite gelassen.

9 Einen Überblick über die angeführten Methoden bieten Perkhun [u. a.] 2011

10 Die Berechnung erfolgte mit dem Python-Package *LexicalRichness* (Shen 2022) auf dem normalisierten Korpus.

11 Ich danke Sonja Glauch für entsprechende Hinweise. Bei Konrad ist allerdings wieder das Ausgaben-Problem zu berücksichtigen, da seine Lieder ja als einzige aus einer eigenen Edition bezogen sind.

12 Dass der TTR-Wert für Sangspruch höher liegt als der für den Minnesang bestätigen auch Braun/Reiter 2017, S. 15.

13 Die Wordclouds wurden mit dem Python-*wordcloud*-Package erstellt, unter Rückgriff auf den Code von https://towardsdatascience.com/how-to-make-word-clouds-in-python-that-dont-suck-86518cdcb61f.

14 Freilich können Wordcloud-Darstellungen auch trügerisch sein, da z. B. längere Wörter per se größer erscheinen als kürzere Wörter und dadurch überbetont werden. Doch gilt hier das oben für das Modell Gesagte: Daten-Visualisierungen sollten keinesfalls mit der Wahrheit ›an sich‹ verwechselt werden, sondern Ausgangspunkte für Interpretationen bieten.

15 Die Erstellung der Wordclouds erfolgte mit dem *wordclouds*-Package in R (Fellows 2018), unter Rückgriff auf Code von Wiedeman/Niekler 2017.

16 Die Grafik wurde mit der oppose()-Funktion es *stylo*-Package für R erstellt (Eder [u. a.] 2013), auf der Basis von Textabschnitten zu 3000 Wörtern.

17 Siehe dazu die Ergebnisse der PCA sowie schon Schnell 2013, S. 326, der so weit geht, dass er den Natureingang als neues Gattungssignal des Minnesangs im Spätmittelalter ansieht, das notwendig werde, da Sangspruch und Minnesang immer ununterscheidbarer werden.

18 Eine genaue Übersetzung der *topics* mit Themen wäre allerdings irreführend: *topics* können sich auch durch andere Bedeutungsformationen als Themen ergeben, etwa durch Einsprengsel von fremdsprachigen Ausdrücken, aber auch durch die Zugehörigkeit von Wörtern zu allgemeineren Gruppierungen wie etwa bei Ausdrücken der Zeit (vgl. Schöch 2017, S. 4)

19 Streng genommen sind sogar alle *topics* in einem Text enthalten und alle Wörter in allen *topics*, manche jedoch mit nur sehr geringer Wahrscheinlichkeit.

20 Erstellt in Anlehnung an Wiedeman/Niekler 2017 mit dem Package *topicmodels* in R (Grün/Hornik 2011). Siehe zum Topic Model des Minnesangs ausführlicher Viehhauser 2017. Im Gegensatz zur Darstellung dort wurde das Modell auf Grundlage der mit dem RNNTagger normalisierten Texte erstellt.

Literaturverzeichnis

Primärliteratur

KLD - Deutsche Liederdichter des 13. Jahrhunderts, hrsg. von Carl von Kraus, Tübingen 1952.

KW - Kleinere Dichtungen Konrads von Würzburg, hrsg. von Edward Schröder mit einem Nachwort von Ludwig Wolff, 3. Aufl, Berlin 1924/59.

MF - Des Minnesangs Frühling, unter Benutzung der Ausgaben von Karl Lachmann/Moriz Haupt/Friedrich Vogt/Carl von Kraus bearbeitet von Hugo Moser/Helmut Tervooren, 36. neugestaltete und erweiterte Aufl., Stuttgart 1977.

SM - Die Schweizer Minnesänger, nach der Ausg. von Karl Bartsch neu bearbeitet und hg. von Max Schiendorfer, Tübingen 1990.

W - Walther von der Vogelweide: Leich, Lieder, Sangsprüche, 14. völlig neubearbeitete Aufl. der Ausgabe von Karl Lachmann mit Beiträgen von Thomas Bein und Horst Brunner hrsg. von Christoph Cormeau, Berlin/New York 1996.

Sekundärliteratur

Blei, David M./Ng, Andrew Y./Jordan, Michael I.: Latent Dirichlet Allocation, in: Journal of Machine Learning Research 3 (2003), S. 993–1022 (online).

Box, George E. P. Robustness in the Strategy of Scientific Model Building, in: Launer, Robert L./Wilkinson, Graham N. (Hrsg.): Robustness in Statistics, New York [u. a.] 1976, S. 201–236.

Braun, Manuel/Reiter, Nils: Sangsprüche auf/in Wörterwolken oder: Vorläufige Versuche zur Verbindung quantitativer und qualitativer Methoden bei der Erforschung mittelhochdeutscher Lyrik, in: Brunner, Horst/Löser, Freimut/Franzke, Janina (Hrsg.): Sangspruchdichtung zwischen Reinmar von Zweter, Oswald von Wolkenstein und Michel Beheim, Wiesbaden 2017 (Jahrbuch der Oswald von Wolkenstein-Gesellschaft 21), S. 5–20 (online).

Burrows, John: All the Way Through. Testing for Authorship in Different Frequency Strata, in: Literary and Linguistic Computing 22 (2007), S. 27–47 (online).

Ciula, Arianna/Eide, Øyvind/Marras, Cristina/Sahle, Patrick (Hrsg.): Models and Modelling between Digital and Humanities: A Multidisciplinary Perspective. Historical Social Research, Supplement 31 (2018) (online).

Craig, Hugh/Greatley-Hirsch, Brett: Style, Computers, and Early Modern Drama Beyond Authorship, Cambridge 2017.

Eder, Maciej/Kestemont, Mike/Rybicki, Jan: Stylometry with R: a suite of tools, in: Digital Humanities 2013: Conference Abstracts. University of Nebraska-Lincoln, 16. –19. July 2013, NE, S. 487–489 (online).

Escobar Varela, Miguel: Theatre as Data. Computational Journeys into Theater Research, Ann Arbor 2021 (online).

Fellows, Ian: wordcloud: Word Clouds. R package 2018 (online).

Firth, John R.: A synopsis of linguistic theory 1930–1955, in: Firth, John R. (Hrsg.): Studies in Linguistic Analysis. Special volume of the Philological Society, Oxford 1957, S. 1–32.

Flanders, Julia/Jannidis, Fotis (Hrsg.): The Shape of Data in Digital Humanities. Modeling Texts and Text-based Resources, London 2018.

Gius, Evelyn/Jacke, Janina: Informatik und Hermeneutik. Zum Mehrwert interdisziplinärer Textanalyse, in: Zeitschrift für digitale Geisteswissenschaften 1 (2015) (online).

Grün, Bettina/Hornik, Kurt: topicmodels: An R Package for Fitting Topic Models 2011 (online).

Horstmann, Jan: Topic Modeling, in: *forTEXT. Literatur digital erforschen*, 2018 (online).

Hübner, Gert: Minnesang im 13. Jahrhundert. Eine Einführung, Tübingen 2008.

Hübner, Gert: Konzentration aufs Kerngeschäft. Späte Korpora der Manessischen Liederhandschrift und die Gattungsgeschichte des Minnesangs im 13. Jahrhundert, in: Köbele 2013a, S. 387–411.

Jannidis, Fotis/Flanders, Julia: A Gentle Introduction to Data Modeling, in: Flanders/Jannidis 2018, S. 26-96.

Jannidis, Fotis: Modeling in the Digital Humanities: a Research Program?, in: Ciula [u. a.] 2018, S. 96–100 (online).

Karsdorp, Folgert/Kestemont, Mike/Riddell, Allen: Humanities Data Analysis. Case Studies with Python, Princeton 2021.

Klein, Thomas/Wegera, Klaus-Peter/Dipper, Stefanie/Wich-Reif, Claudia (2016): Referenzkorpus Mittelhochdeutsch (1050–1350), Version 1.0 (online).

Köbele, Susanne (Hrsg., in Verbindung mit Eckart Conrad Lutz und Klaus Ridder): Transformationen der Lyrik im 13. Jahrhundert. Berlin 2013a (Wolfram-Studien 21).

Köbele, Susanne (2013b): Einleitung, in: Köbele 2013a, S. 9–17.

Kuhn, Hugo: Minnesangs Wende, Tübingen 1952.

McCarthy, Philip M.: An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity. Dissertation University of Memphis 2005.

McCarty, Willard: Modeling: A Study in Words and Meanings, in: Schreibman, Susan/Unsworth, John/Siemens, Ray (Hrsg.): A Companion to Digital Humanities, Oxford (2004) (online).

Moretti, Franco: Conjectures on World Literature, in: New Left Review 1 (2000), S. 54–68 (online).

Moretti, Franco: »Operationalizing«: or, the function of measurement in modern literary theory. Literary Lab Pamphlet 6 (2013) (online).

Mueller, Martin: Shakespeare His Contemporaries. Collaborative curation and exploration of Early Modern drama in a digital environment, in: Digital Humanities Quarterly 8 (2014) (online).

Perkhun, Rainer/Keibel, Holger/Kupietz, Marc: Ergänzungen zu Korpuslinguistik. 18. Juni 2012 (online).

Pierazzo, E[lena]: How Subjective Is Your Model?, in: Flanders/Jannidis 2018, S. 117–132.

Piper, Andrew: There Will Be Numbers, in: Cultural Analytics 1 (2016), S. 1–10 (online).

Piper, Andrew: Think Small: On Literary Modeling, in: PMLA 132 (2017), S. 651–658 (online).

Ramsay, Stephen: Reading Machines. Toward an Algorithmic Criticism, Champaign 2011 (online).

Rosen, Jeremy: Combining Close and Distant, or, the Utility of Genre Analysis: A Response to Matthew Wilkens's »Contemporary Fiction by the Numbers«, in: Post45 2011 (online).

Schmid, Helmut: Deep Learning-Based Morphological Taggers and Lemmatizers for Annotating Historical Texts, in: Proceedings DATeCH, May 2019 (online).

Schnell, Rüdiger: Minnesang und Sangspruch im 13. Jahrhundert. Gattungsdifferenzen und Gattungsinterferenzen, in: Köbele 2013a, S. 287–347.

Schöch, Christof: Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama, in: Digital Humanities Quarterly 11 (2017), H. 2 (online).

Schöch, Christof: Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie, in: Bernhart, Toni/Willand, Marcus/Richter, Sandra/Albrecht, Andrea: Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven, Berlin/Boston 2018, S. 77–94 (online).

Shen Y[an] S[hun], Lucas/Lesieur, David/Bedetti, Christophe: LexicalRichness: A small module to compute textual lexical richness 2022 (online).

So, Richard Jean: »All Models Are Wrong«, in: PMLA 132 (2017), S. 668–673 (online).

Spärck Jones, Karen: A Statistical Interpretation of Term Specificity and Its Application in Retrieval, in: Journal of Documentation 28 (1972), 11–21 (online).

Stachowiak, Herbert: Allgemeine Modelltheorie, Wien 1973.

Underwood, Ted: A Genealogy of Distant Reading, in: Digital Humanities Quarterly 11 (2017) (online).

Viehhauser, Gabriel: Digitale Gattungsgeschichten. Minnesang zwischen generischer Konstanz und Wende, in: Zeitschrift für digitale Geisteswissenschaften (2017) (online).

Viehhauser, Gabriel: Mittelalterliche Texte als Modellierungsaufgabe, in: Fischer, Martin (Hrsg.): Digitale Methoden und Objekte in Forschung und Vermittlung der mediävistischen Disziplinen. Akten der Tagung Bamberg, 08.–10. November 2018, Bamberg 2020, S. 15–50 (online).

Wiedemann, Gregor/Niekler, Andreas: Hands-on: A five day text mining course for humanists and social scientists in R. Proceedings of the 1st Workshop on Teaching NLP for Digital Humanities (Teach4DH@GSCL 2017), Berlin 2017 (online).

Windelband, Wilhelm: Geschichte und Naturwissenschaft. Rede zum Antritt des Rectorats der Kaiser-Wilhelms-Universität Strassburg, geh. am 1. Mai 1894. Strassburg 1894. Sitzungsberichte der Heidelberger Akademie der Wissenschaften, Philosophisch-Historische Klasse. Jg. 1910, Abh. 14 (online).

Online-Ressourcen

MHDBDB (Mittelhochdeutsche Begriffsdatenbank): http://mhdbdb.sbg.ac.at/.

Python-wordcloud: https://github.com/amueller/word_cloud.

RNNTagger (Recurrent Neural Network Tagger): https://www.cis.uni-muenchen.de/~schmid/tools/RNNTagger/.

scipy.stats.linregress: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.linregress.html.

TDS (Towards Data Science): https://towardsdatascience.com/how-to-make-word-clouds-in-python-that-dont-suck-86518cdcb61f.

Anschrift des Autors:

Prof. Dr. Gabriel Viehhauser
Universität Stuttgart
Institut für Literaturwissenschaft
Herdweg 51
70174 Stuttgart
E-Mail: viehhauser@ilw.uni-stuttgart.de

Phillip Brandes / Sophie Marshall / Felix Schneider

Stilfiguren aus der Distanz gelesen

Zur automatischen Detektion von Wortstellungsfiguren und deren Nutzen für die qualitative Analyse

Abstract. Im Projekt ›*Anomaly-based large-scale analysis of style and genre reflected in the use of stylistic devices in medieval literature*‹ befragen wir mittelhochdeutsche Texte der *Trois Matières* danach, welche Aussagen zu Textähnlichkeiten sich aufgrund des Gebrauchs von Stilmitteln der Wortstellung, wie etwa Parallelismus und Chiasmus, sowie der Tropen, wie etwa der Metapher, treffen lassen. Der hier vorliegende Artikel stellt die aktuellen Ergebnisse unserer bisherigen Parallelismus- und Chiasmusdetektion in 30 mittelhochdeutschen Texten vor. Es wird zu diskutieren sein, ob unsere Ergebnisse eine Korrelation des Stils mit der gemeinhin nach Stoffkreis definierten Gattung bezeugen, ob andere Faktoren wie Verfasserschaft oder Abfassungszeit eine übergeordnete Rolle spielen und ob darüberhinausgehende Aussagen zu den Ähnlichkeitsverhältnissen der Texte getroffen werden können.

1. Einleitung

»Was braucht das Fach?« war die Leitfrage der Tagung »Digitale Mediävistik«, in deren Rahmen der hier vorliegende Beitrag erstmals diskutiert wurde. Die Frage bezieht sich auf die Herausforderungen, denen sich die Germanistische Mediävistik aufgrund der wachsenden Bedeutung und Re-

levanz der Digital Humanities und angrenzender Nachbardisziplinen entgegengestellt sieht. Hier sollen aus der Praxis eines Projektes, das gleichermaßen aus Informatik und Mediävistik heraus betrieben wird, Bedürfnisse mitformuliert werden, die sich aus der computationellen Textanalyse ergeben.

Ziel des Projektes ›*Anomaly-based large-scale analysis of style and genre reflected in the use of stylistic devices in medieval literature*‹[1] ist es, computationelle Methoden zur Detektion von rhetorischen Stilmitteln zu entwickeln und diese Methoden auf ein Corpus, das aus mittelhochdeutschen Texten der *Trois Matières* besteht, anzuwenden und die daraus resultierenden Funde zu analysieren.

Das Projekt bewegt sich damit auch im Bereich der Stilometrie. Stilometrie ist, so kann man es kurzfassen, die computationelle Untersuchung von Stil.[2] Bisher ist die Attribution von Autorschaft das vorherrschende Anwendungsgebiet der Stilometrie (Büttner [u. a.] 2017).[3] Während Autorstil durch diverse *most-frequent-words*-Ansätze bereits hinreichend untersucht werden konnte, ist noch weitestgehend unklar, wie ein Gattungsstil untersuchbar wäre (oder überhaupt zu bestimmen ist). Gattung scheint nicht etwas zu sein, das man allein aufgrund von Worthäufigkeiten erkennen kann (Allison [u. a.]. 2017, insb. S. 31–33). Dies liegt wohl daran, dass ›Gattung‹ ein Begriff ist, der seine jeweilige Geltung aus den Konventionen zieht, die ihn hervorgebracht haben (Underwood 2019; Remele 2021) und Gattungseinteilungen nicht immer auf der systematisch gleichen Ebene vorgenommen werden (Viehhauser 2017).

Für Texte der *Trois Matières* erfolgt die Zuordnung nach Jean Bodel (vor 1200) zu einer ›Materie‹ aufgrund des Inhalts (einleitend dazu Herweg 2013). Was Artus und seine Tafelrunde zum Gegenstand hat, wird so der *Matière de Bretagne* zugeordnet. Formal sind zumindest die mittelhochdeutschen Vertreter dieser drei Stoffkreise über weite Teile identisch – alle Texte sind in kurzen Reimpaarversen verfasst (von kleinen Abweichungen wie dreiversigen Abschnittschlüssen, wie sie bspw. der ›Wigalois‹ kennt,

abgesehen). Die formale Gleichförmigkeit wird mit Wolframs ›Titurel‹ zunächst nur punktuell und erst später auch von Prosa-Texten herausgefordert. Möchte man sagen, dass es sich bei den *Trois Matières* also um jeweils eigenständige Gattungen handelt, hat man bisher nur inhaltliche Argumente. Wir fragen nun danach, ob die Quantität der Verwendung bestimmter Stilmittel – nämlich der Metapher sowie der Wortstellungsfiguren Chiasmus und Parallelismus – mit dem jeweils adaptierten Stoffkreis distinkt korreliert oder ob bezüglich der Stilmittel andere Parameter wie Verfasser, Entstehungszeit oder Szenentyp bestimmend sind. Den quantitativen Analysen folgen qualitative Anschlussuntersuchungen zu Form und Funktion der Stilmittel. Im vorliegenden Artikel wird nur auf bisherige Ergebnisse zu Chiasmen und Parallelismen eingegangen.

Wenn sich die jeweiligen Stoffkreise der *Trois Matières* – Antikenroman, Karlsepik und Artusroman (samt der Tristan-Texte) – hinsichtlich des Gebrauches von Wortstellungsfiguren in Form, Funktion und Häufigkeit unterscheiden und Einflüsse anderer Parameter wie Verfasserschaft und Entstehungszeit weniger zu beobachten sein sollten, könnte dies als ein Gattungssignal gewertet werden. Zu erwägen wäre dann also, ob für bestimmte Stoffkreise, möglicherweise aufgrund des Usus der altfranzösischen Vorlagen, etwa Parallelismen dem Chiasmus vorgezogen wurden. Das in mittelalterlicher Bildung zentrale Konzept des rhetorischen *aptum* legt auch nahe, dass bestimmte Themen innerhalb der Texte wie etwa Kämpfe rhetorisch elaborierter sind als andere (Ehrismann 1919). Dann würde nicht der Stoffkreis die Verwendung von Wortstellungsfiguren mitbestimmen, sondern einzelne Szenen oder Textabschnitte wie möglicherweise Kampf- oder Festszenen.[4] Dass die Ebene der Stilmittel nur einen Bruchteil von möglichen szenentypischen und/oder gattungsrelevanten Merkmalen bildet (neben beispielsweise narratologischen Kriterien von Zeit, Modus und vielen weiteren), ist selbstverständlich.

Das Corpus unseres Projektes umfasst folgende dreißig Texte[5]:

Matière de Bretagne	Matières de France	Matière de Rome
Crône (CRO)	Karl (KAR)	Ulrich v. Etzenbach: Alexander (AXU)
Daniel (DA)	Rennewart (REN)	Rudolf von Ems: Alexander (AXR)
Erec (ER)	Rolandslied (ROL)	Straßburger Alexander (AXS)
Garel (GAR)	Willehalm (WH)	Vorauer Alexander (AXV)
Gauriel (GL)	Alischanz (WUT)	Eneas (ENE)
Iwein (IW)		Trojanerkrieg (TRO)
Lanzelet (LZ)		Lied von Troja (TRY)
Mantel (MAN)		Göttweiger Trojanerkrieg (GWTK)
Meleranz (MEL)		
Parzival (PZ)		
Tandareis und Flordibel (TAN)		
Wigalois (WGL)		
Wigamur (WGM)		
Eilharts Tristant (EIL)		
Tristan (TR)		
Heinrich v. Freiberg Tristan (TRH)		
Ulrich von Türheim Tristan (TRU)		

Tabelle 1: Corpus

Im Folgenden sollen die quantitativen Befunde analysiert werden. Dabei werden wir zunächst auf die Herausforderungen, die die Modellierung literaturwissenschaftlicher Begriffe wie in unserem Fall Chiasmus und Parallelismus mit sich bringen, und auf unsere technische Umsetzung eingehen. Anschließend werten wir die gefundene Menge an Stilmitteln in den oben genannten dreißig Texten aus. Wir werden zeigen, dass Fragen, die auf Basis computationeller Methoden gestellt werden können, nicht zwangsläufig von traditionelleren Textuntersuchungsmethoden differieren. Dafür kann aber die Hinzuziehung computationeller Methoden den Blick für solche Fragen neu öffnen und scheinbar selbstverständliche Annahmen mit überraschenden Befunden zur Diskussion stellen.

2. Wortstellungsfiguren der Ähnlichkeit

Müssen computationelle Methoden für eine quantitative Untersuchung von Stilmitteln erst entwickelt werden – wie es in unserem Projekt der Fall ist –, liegt nahe, dass eine Eingrenzung auf bestimmte Stilmittel erfolgen

muss.[6] Dies hat Einschränkungen auf die Aussagekraft der Funde zur Folge. Denn es ist ja durchaus möglich, dass sich Texte bei der Verwendung etwa von Chiasmen ähneln, hinsichtlich der Verwendung von Apokoinu oder auch Paralipsen hingegen stark unterscheiden. Was Textähnlichkeit also ist, ab wann sich Texte auf Ebene der Stilmittelverwendung – darüber hinaus gäbe es ja, wie gesagt, noch die Ebene des Inhalts, des Metrums, der Zeit und viele weitere – ähneln (ab drei Stilmitteln? wenn die Menge an Stilmitteln ähnlich ist, oder doch eher die ›Art‹ der Verwendung?), ist diskutabel.

Was folgt, ist die ressourcenbedingte Einschränkung auf einige wenige Stilmittel: Neben Parallelismen und Chiasmen untersuchen wir derzeit Metaphern, wobei Letztere kein Gegenstand des hier vorliegenden Beitrags sind. Entsprechend ist die Aussagekraft der Befunde als womöglich bedingt einzustufen. Das soll nicht heißen, es könnten keine validen Beobachtungen angestellt werden. Es heißt nur, dass die hier getroffenen Aussagen vorläufig sein können (die Weiterentwicklung unserer Methoden zur Detektion von Stilmitteln bleibt ein Ziel). Denn grundsätzlich gilt, je mehr Variablen zur Untersuchung eines literarischen Phänomens – hier eben mögliche Abgrenzungskriterien der *Trois Matières* – herangezogen werden, desto wahrscheinlicher die Signifikanz und Plausibilität der Ergebnisse (Underwood 2019). Allerdings legen andere Studien nahe,[7] dass durchaus als wahrscheinlich gelten kann, dass unsere Ergebnisse ein Indiz für einen grundsätzlichen Trend in Bezug auf die Stilmittelhäufigkeit im Allgemeinen in den von uns untersuchten Textgruppen sein können – aber eben nicht zwingenderweise sein müssen. Unsere Untersuchung liefert also keineswegs keine Ergebnisse, nimmt aber eben – vorerst, das muss hier betont werden, – ›nur‹ Parallelismen und Chiasmen in den Blick.

Die Stilmittel, auf die wir uns beschränkt haben, erfüllen allerdings den zweifachen Anspruch, es wert zu sein, eine computationelle Methode zu entwickeln (Alliterationen etwa wären aus technischer Perspektive nicht lohnend gewesen), und gleichzeitig besonders ›wirkmächtig‹ zu sein.

Beide, Parallelismus und Chiasmus, sind jedenfalls prägende Merkmale mittelhochdeutscher Dichtung und sind auch heute noch Bestandteil von (literarischen) Texten. Andere Stilmittel wie etwa Hyperbaton lassen nur schwerlich textübergreifende Untersuchungen zu. Die Bedeutung, die Parallelismen und Chiasmen haben, gleicht die nur geringe Menge an untersuchten Stilmittelarten partiell aus; es kann wohl mit einigem Recht davon ausgegangen werden, dass eine Untersuchung der drei Stilmittel, die unser Projekt (die Metapher einbezogen) anstrebt, aussagekräftiger ist als eine Untersuchung von mehr Stilmitteln mit geringerer literargeschichtlicher Persistenz.

2.1 Das Verhältnis von Chiasmus und Parallelismus in mittelhochdeutscher Literatur

Chiasmus und Parallelismus sind zwar neuzeitliche Begriffe, dennoch ist das, was sie bezeichnen, bereits in antiken und mittelalterlichen Texten beobachtbar. Es handelt sich um Phänomene der syntaktischen, aber auch semantischen Parallelität und Gegensätzlichkeit – es fand keine begriffliche Trennung zwischen beiden Phänomenen statt, am ehesten treffen die Begriffe ›Isokolon‹ und ›Parison‹ zu. Als solche sind sie nach den im Mittelalter normsetzenden Rhetoriken Mittel der *compositio* (Scaglione/ Marvin 1994, Sp. 303).

Sowohl Parallelismus als auch Chiasmus haben eine a) syntaktische und b) semantische Ebene. Im Falle des Parallelismus ist mit syntaktischer Ebene die »syntaktische Äquivalenz zweier oder mehrerer aufeinanderfolgender Sätze oder Satzteile« (Ostrowicz 2003, Sp. 546) gemeint. ›Semantisch parallel‹ meint die Wiederholung »eines Gedankens in zwei oder mehr Satzteilen oder ganzen Sätzen« (Ostrowicz 2003, Sp. 546). Sonderformen des semantischen Parallelismus sind der synonyme und der synthetische Parallelismus. Bei Ersterem wird ein Gedanke mit variierendem Vokabular wiederholt, während sich ein synthetischer Parallelismus dadurch

auszeichnet, dass die im ersten Glied des Stilmittels getroffene Aussage nicht wiederholt, sondern fortgeführt und damit genauer bestimmt wird. Ein weiterer Sonderfall ist der antithetische Parallelismus, der syntaktisch parallele Strukturen aufweist, semantisch jedoch oppositionell aufgestellt ist.[8] Zuletzt kann bei weiter voneinander entfernten Textstellen und Strukturanalogien auch von einem Parallelismus im weiten Sinne gesprochen werden (Ackermann 2007).

Der Chiasmus wird als komplementär zum Parallelismus gesehen (Fauser 1994, Sp. 171). Die abhängigen Satzteile oder Sätze sind im Gegensatz zum Parallelismus syntaktisch überkreuzt (a–b — b–a), gleiches gilt für den semantischen Chiasmus. Ein Sonderfall des Chiasmus ist die Antimetabole: Die syntaktischen Einheiten sind in diesem Fall nicht nur gespiegelt, sondern erfordern eine exakte Wiederholung der Lexeme.

Die neuzeitliche Distinktion von Wortstellungsfiguren der Spiegelung und (kontrastierenden) Wiederholung in Chiasmus einerseits und Parallelismus andererseits galt für die Rhetoriken, vor deren Hintergrund die mittelalterlichen Texte verfasst wurden, nicht. Dass eine scharfe Trennung auch nur bedingt möglich ist, mag folgendes Beispiel verdeutlichen:

In Gottfrieds ›Tristan‹ heißt es im Prolog programmatisch: *ein man ein wîp, ein wîp ein man, / Tristan Îsolt, Îsolt Tristan* (V. 129f.). Diese zwei Verse decken bereits ein breites Spektrum der oben vorgestellten Stilmittel ab. Einerseits ist jeder der Verse für sich genommen eine Antimetabole – die Lexeme werden gespiegelt. Anderseits liegt auch – liest man die Nomina *man* und *wîp* als zu den Namen *Tristan* und *Îsolt* zugehörig – ein Parallelismus vor: *man* steht zu *Tristan* wie *wîp* zu *Îsolt*. Möchte man die Geschlechter Mann und Frau als ›gegensätzlich‹ lesen, liegt ein antithetischer Parallelismus – der mit einigem Recht auch semantischer Chiasmus heißen könnte – vor; liest man die Spiegelungen als Zeichen der Zusammengehörigkeit, eine Lesart, die angesichts der Handlung wohl plausibel ist, könnte auch von einem synthetischen Parallelismus gesprochen werden. Allein diese zwei Verse zeigen also, dass die Trennung in Chiasmus und Pa-

rallelismus für mittelalterliche Texte (und Dichtung allgemein) nicht immer und nur bedingt Geltung beanspruchen kann. Wir betrachten die mit computationellen Methoden erarbeiteten Ergebnisse daher nicht nur im Einzelnen für das jeweilige Stilmittel, sondern werten sie auch für Aussagen über (natürlich noch weiter zu ergänzende) Wortstellungsfiguren im Allgemeinen aus.

2.2 Modellierung: Was ›ist‹ ein Chiasmus?

Für computationelle Untersuchungen ist die Trennung in Chiasmus und Parallelismus durchaus sinnvoll. Denn um literarische Phänomene computationell untersuchbar zu machen, ist es nötig, sie und die zur Untersuchung herangezogenen Variablen mittels eines Modells genauestens zu profilieren. Das untersuchte ›Phänomen‹ ist hier im weiten Sinne Textähnlichkeit, konkret meint es für uns die Möglichkeit einer Gattungstrias; die Wortstellungsfiguren sind die Variable, anhand derer Aussagen über das Phänomen getroffen werden sollen; im Kontext der Modellierung sind die Wortstellungsfiguren bereits ›Phänomen‹. Verschiedene Modelle des gleichen Phänomens haben je nach intendiertem Zweck unterschiedliche Eigenschaften (Jannidis 2017, S. 100). Die Modellierung literarischer Sachverhalte bewegt sich dabei auch oftmals in einem Spannungsverhältnis aus technischer Machbarkeit und Aussagekraft für literaturwissenschaftliche Fragestellungen.

Exemplarisch sei hier auf eine Textstelle im ›Straßburger Alexander‹ verwiesen. Im Konflikt zwischen Alexander und Darius schicken sich beide Herrscher Gewürze in Form von Mohn bzw. Pfeffer – die Reaktionen auf die ›Geschenke‹ könnten unterschiedlicher nicht ausfallen. Während Alexander Darius' Mohn fast schon genussvoll zu sich nimmt, krampft Darius nach Verzehr weniger Pfefferkörner (vgl. ›StA‹, V. 2075, 2097). Die Passagen umspannen über 50 Verse (vgl. ›StA‹, V. 2046–2117). Dazu ist der Ablauf zwar nach dem Muster Gewürzerhalt-Freude – Gewürzerhalt-Wut

gespiegelt, so dass gewissermaßen von einem antithetischen Parallelismus im weiteren Sinne geredet werden könnte. Hinzu kommt, dass dies nicht der erste Gabenaustausch der beiden Kontrahenten war. Auch im ersten Gabenaustausch geht Alexander als ›Sieger‹ hervor – hier liegt also wiederum eine parallelistisch anmutende Wiederholung vor. Mehrere hundert Verse Umfang und das Erkennen-Können-Müssen, dass bei dieser Distanz Äquivalenzbeziehungen wie hier ›mehrmaliger Gabenaustausch‹ vorliegen, in das Stilmittelmodell zu integrieren, wäre aufgrund technischer Hürden nicht zielführend gewesen. Nicht zuletzt aufgrund eines fehlenden Syntax-Parsers für das Mittelhochdeutsche sind komplexere Konfigurationen der Wortstellungsfiguren, die über Wortarten hinausgehen, wohl (noch) nicht als Programm umsetzbar. Beide Textstellen sind jedoch auch auf der Mikroebene stark durch Wortstellungsfiguren überformt (Brandes 2022); diese automatisiert zu erkennen genügt, um die grundsätzliche Tendenz – dass, wie weiter unten genauer ausgeführt, der ›Straßburger Alexander‹ als Vertreter der Antikenromane wesentlich parallelistisch organisiert ist – zu erfassen, auch wenn nicht alle Strukturen, in denen der Text organisiert ist, erfasst werden können.

Da unsere Modellierungsarbeiten bei den Vorarbeiten der Stilmitteldetektion ansetzten – die sich primär mit der Antimetabole, wie oben beschrieben einem Spezialfall des Chiasmus, auseinandersetzen (Dubremetz/Nivre 2017) –, erläutern wir hier zunächst unser Chiasmus-Modell. In unserem Fall sollte also das Modell von ›Chiasmus‹ einerseits so ausgestaltet sein, dass eine spätere automatisierte Detektion tatsächlich machbar ist, und musste anderseits der literarischen Komplexität der Texte gerecht werden können. Relevant ist hier, dass unser Chiasmus-Modell nicht zwangsläufig auch für andere Arbeiten, die sich mit Chiasmen befassen, richtig sein muss. Ein Modell enthält immer nur die Eigenschaften des modellierten Gegenstandes, die für die Untersuchung maßgeblich sind (Jannidis 2017, S. 100).

Schließlich galten für unser Chiasmus-Modell folgende Merkmale als wesentlich: 1) ein Chiasmus ist eine Textstelle mit bis zu 30 Token Umfang,[9] 2) innerhalb dieser 30 Token gibt es vier ›stilmitteltragende‹ Wörter; stilmitteltragend meint, dass diese vier Wörter notwendig sind, um als betreffendes Stilmittel erkannt zu werden, 3) zwischen dem ersten und vierten sowie zweiten und dritten Wort herrscht eine syntaktische Ähnlichkeitsbeziehung und/oder eine semantische Ähnlichkeits- bzw. Kontrastbeziehung.

2.3 Technische Umsetzung

Die technische Umsetzung wird detailliert in Schneider ([u. a.] 2021) erläutert. Der bis dahin aktuelle *State-of-the-Art* der Chiasmusdetektion galt nur für den Spezialfall der Antimetabole. Für die Detektion von Chiasmen im Allgemeinen kann eine Wiederholung von Lexemen nützlich sein, reicht allein jedoch nicht aus. Unser Stilmittelmodell (s. o.) umfasst daher noch weitere syntaktische und semantische Informationen.

Auf syntaktischer Ebene war es möglich, mittels eines Part-of-Speech-Taggers für das Mittelhochdeutsche (Echelmeyer [u. a] 2017) nach invertierten Wortarten gemäß dem Muster a–b — b–a zu suchen. Um die semantische Ebene berücksichtigen zu können, trainierten wir *Word Embeddings* für die mittelhochdeutsche Sprache.[10] Um die Informationen, die die *Word Embeddings* bereitstellen, zu nutzen, haben wir Beispiele der gesuchten Stilmittel manuell in mittelhochdeutschen Texten annotiert. Zur Annotation gehörte auch die Markierung der ›stilmitteltragenden‹ Wörter, also jener Wörter, die konstitutiv für das jeweilige Stilmittel sind. Von diesen Wörtern haben wir den Cosinus-Abstand der Wortvektoren (s. Abb. 1) als vom Klassifikator, der die von der Maschine ausgegebenen Kandidaten mit einem Score versieht, zu berücksichtigendes Merkmal angegeben.

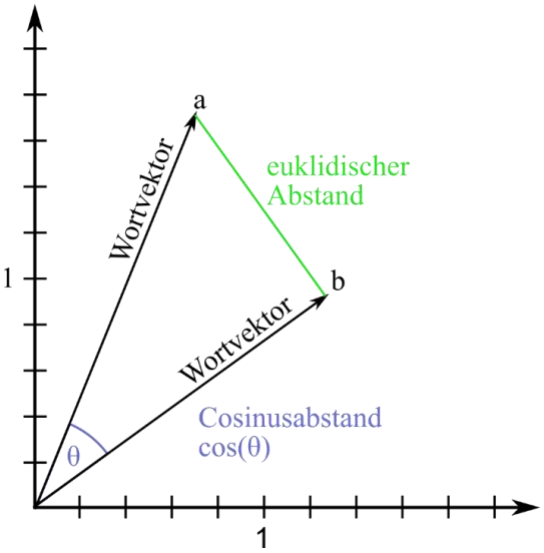


Abb. 1: Cosinusabstand und Wortvektoren

Die Idee war es, bei einer automatisierten Detektion von stilmitteltragenden Wörtern jenen Kandidaten ein höheres Ranking einzuräumen, bei denen der Cosinus-Abstand der Wortvektoren der stilmitteltragenden Wörter möglichst nah an dem unserer manuell annotierten Stilmittel war. Die These, die für dieses Vorgehen sprach, lautet also: Die Cosinus-Abstände der Wortvektoren von stilmitteltragenden Wörtern in Chiasmen sind bei allen Chiasmen tendenziell ähnlicher als die Cosinus-Abstände zufällig ausgewählter Wörter mit invertierter Wortart.

Es handelt sich bei der Nutzung unserer *Word Embeddings* also um ein *Machine-Learning*-Verfahren. Diese Verfahren haben sich in den *computational literary studies* (und auch in benachbarten Disziplinen wie der Computerlinguistik) als besonders leistungsfähig erwiesen (Kuhn 2020, S. 13f.). *Machine-Learning*-Verfahren setzen ›Trainingsmaterial‹ voraus, ein für die jeweiligen Zwecke geeignetes Set an Datensätzen, das Muster enthält, anhand derer die Maschine bessere Ergebnisse erzielen kann. In unserem Fall handelt es sich dabei um die manuell annotierten Daten.

Grundsätzlich gilt: je größer und korrekter der Datensatz, desto besser sind spätere Resultate. *Machine-Learning*-Verfahren sind also wesentlich mit von den Ressourcen, die zur Verfügung stehen, und den Datensätzen, die eingegeben werden, abhängig. Wären mehrere Millionen Texte vorab manuell annotiert worden – was eine *low-resource-language* wie das Mittelhochdeutsche nicht zulässt, selbst unannotierte Daten für unüberwachtes Lernen gibt es nicht in der dafür benötigten Größenordnung –, wären spätere Resultate wohl schneller und/oder besser erreicht worden. Auch hieraus folgt also die ressourcenbedingte Vorläufigkeit der Ergebnisse. Glücklicherweise gilt dies jedoch stärker für die Details der Ergebnisse als für die allgemeinen Trends, die aufgezeigt werden.

2.4 Optimierung des Verfahrens

Den *State-of-the-Art* der Antimetabole-Detektion konnten wir verbessern (Schneider [u. a.] 2021). Mit einer Quote von 35% innerhalb der besten einhundert Funde war die Aussagekraft über literarische Texte dennoch gering. Ziel war es in der Folge nicht mehr, den zugrundeliegenden technischen Hintergrund zu verändern, sondern die Erfolgsrate durch Entfernen von dem, was einen negativen Fund wahrscheinlicher macht, zu steigern.

Eine der ausgegebenen Textstellen ist folgende aus dem ›Tristan‹: *der gottinne Minne. / zer fossiure* [*oben*] *inne* / [*dâ*] *wâren cleiniu vensterlîn / durch daz lieht gehouwen în, / diu lûhten* [*dâ*] *unde hie.* / [*dâ*] *man ûz und in gie* (V. 16723–16728) – sie steht exemplarisch für eine Vielzahl von Funden, die keinem der gesuchten Stilmittel entsprechen. Die Wörter in eckigen Klammern wurden maschinell mit der Wortart ›Adverb‹ als stilmitteltragend markiert. Es fiel aufgrund zahlreicher ähnlicher Beispiele auf, dass sich die gesuchten Stilmittel überproportional häufig aus Wortarten wie Nomen, Adjektiv und Verb zusammensetzten und dass hingegen Wortarten wie Pronomina, Hilfsverben und eben auch Adverbien

nahezu immer in ausgegebenen Textstellen vorkamen, die den gesuchten Stilmitteln nicht entsprachen. Ein Problem mit Funktionswörtern wie etwa Hilfsverben war die häufige Ausgabe von Textstellen nach folgendem Muster: [*Dâ*] [*wâren*] *gesazt under / starkir radere viere. / Starker elfentiere sehs unde drîzich / (daz* [*was*] [*vil*] *hêrlîch)* (›StA‹, V. 6108–6112) – die Spiegelung aus Adverb–Hilfsverb—Hilfsverb–Adverb veranlasste, einen relativ hohen Score anzusetzen, obwohl keines der gesuchten Stilmittel vorlag.

Um diesem Problem zu begegnen, haben wir in einem iterativen Verfahren eine Liste von POS-Tags und auch einzelner Lemmata zusammengetragen, die überproportional häufig in falschen Ergebnissen mündeten (entweder weil sie häufig falsch getaggt wurden oder schlicht keine Beispiele bekannt wurden, in denen sie den gesuchten Stilmitteln entsprachen, dazu gehört beispielsweise das Lemma ›*ër*‹) und die wir daher nicht mehr berücksichtigt haben. Dieses Vorgehen birgt freilich die Gefahr, Texten, die bevorzugt Stilmittel mit der Zusammensetzung ›Adverb–Pronomen — Pronomen–Adverb‹ enthalten, nicht gerecht zu werden. Denn Stilmittel mit dieser Zusammenstellung würden selten bis gar nicht in den Top 100-Ergebnissen angezeigt – so hätten wir, ohne es zu wissen, unseren Ansatz systematisch an Texten ausgerichtet, die ohnehin bereits weiter oben im Ranking platziert waren. Um derartige Fehler zu vermeiden, haben wir die Ergebnisse von Einzeltexten, die niedrige Scores auswiesen, analysiert, ohne bestimmte POS-Tags und Lemmata auszuschließen, und konnten feststellen, dass keiner der untersuchten Texte durch unsere gefilterte POS-Tag-Liste benachteiligt wurde.

Parallelismen und Chiasmen, das kann als – wenn auch wenig überraschendes – Zwischenergebnis festgehalten werden, scheinen sich in überwiegender Mehrheit durch bedeutungstragende Wortarten zu konstituieren, während Wortarten wie Pronomina tendenziell kein konstitutiver Bestandteil der Stilmittel sind. Dabei bleibt natürlich zu betonen, dass es durchaus Stilmittel mit diesen Wortarten geben kann, sie bei einer Trend-

analyse auf Basis einer großen Textmenge allerdings keinen oder einen zu vernachlässigenden Einfluss auf die zu beobachtenden Trends haben.

Die Parallelismusdetektion wurde erst im Anschluss an die funktionierende Chiasmusdetektion begonnen. Einziger Unterschied ist hier, dass nicht auf eine invertierte Reihenfolge der Wortarten, sondern eine sich wiederholende Reihenfolge nach dem Muster A-B–A'-B' geachtet wurde. Für die Top-100-Ergebnisse erzielten wir so eine Erfolgsrate von über 90%.

3. Quantitative Analyse

Die folgenden Tabellen und Abbildungen stellen aktuelle Ergebnisse der Stilmitteldetektion dar. Ausgegeben wurden Textstellen, die gewichtet nach den oben genannten Features – syntaktische und semantische Relation, Wortumfang, Lemma-Wiederholung – mittels eines Klassifikators mit einem bestimmten Score gerankt wurden. Je höher der Score, desto eher handelt es sich bei der ausgegebenen Textstelle um eines der gesuchten Stilmittel. Bis zu einem Score von 11 können wir davon ausgehen, dass es sich bei den ausgegebenen Stilmitteln mit einer Wahrscheinlichkeit von mindestens 50% um ein korrekt erkanntes Stilmittel handelt. Um die Ergebnisse aussagekräftig zu halten, haben wir also nur solche berücksichtigt, die eben einen Score von ≥ 11 aufweisen.

Rang	Text	Vers/Stilmittel
1	MAN	497
2	TRH	861
3	TR	1222
4	TRU	1250
5	ER	1274
6	TRY	1318
7	KAR	1356
8	AXR	1546

Score Relativ	Text	Rang
25,64	WGM	1
22,55	GAR	2
18,4	MAN	3
8,1	TR	4
7,17	TRU	5
7,06	TRY	6
6,78	TRH	7
5,27	DA	8

9	TRO	1608	4,66	ER	9
10	DA	1696	4,63	KAR	10
11	GAR	1938	3,84	ROL	11
12	GL	2083	3,01	AXR	12
13	IW	2722	2,97	TRO	13
14	PZ	2757	2,94	REN	14
15	WGL	2928	2,79	AXS	15
16	AXU	3500	2,74	IW	16
17	WUT	3535	2,58	ENE	17
18	TAN	3610	2,44	GL	18
19	MEL	4281	1,62	WUT	19
20	WGM	6115	1,56	MEL	20
21	ENE	6784	1,51	CRO	21
22	WH	7001	1,38	TAN	22
23	AXS	7098	1,34	PZ	23
24	REN	7300	1,28	AXU	24
25	ROL	9094	1,27	WGL	25
26	CRO	30077	1,06	WH	26
27	GWTK	–	0	GWTK	27
28	LZ	–	0	LZ	28
29	AXV	–	0	AXV	29
30	EIL	–	0	EIL	30

Tabelle 2: Chiasmusdetektion

Die linke Hälfte der Tabelle ist nach Versanzahl je Stilmittel sortiert. Erkennbar ist, dass weder Entstehungszeit noch Autorschaft (auch wenn es mit Hartmann, Wolfram und dem Pleier nur drei Autoren gibt, die mehr als einen der hier aufgeführten Texte verfasst haben) oder Stoffkreis einen Einfluss darauf zu haben scheinen, wie viele Chiasmen im Text vorkommen. Allenfalls sind leichte Tendenzen, was den Stoffkreis anbelangt, erkennbar. So sind in drei der vier Tristan-Texte vergleichsweise viele Stil-

mittel gefunden worden, lediglich der Text Eilharts weicht hier stark ab. Auch dass die karlsepischen Texte (mit Ausnahme des ›Karl‹) weiter unten in der Tabelle platziert sind, ist erkennbar.

Die rechte Hälfte der Tabelle ist nach dem relativen Score sortiert. Dieser sagt aus, wie viele aller gefundenen Stilmittel mit einem Score ≥ 11 bewertet wurden. Die Texte mit einem hohen relativen Score weisen also im Verhältnis viele mit erhöhter Wahrscheinlichkeit richtige Funde auf. Auch hier ist erkennbar, dass Texte des Bretagne-Stoffkreises das obere Drittel dominieren. Solche Resultate sind gemeint, wenn in Abschnitt 2 von der Notwendigkeit möglichst vieler Variablen, die zur Beantwortung einer literaturwissenschaftlichen Fragestellung herangezogen werden, die Rede war. Obwohl die Reihenfolge der Texte in beiden Tabellenhälften im Detail unterschiedlich ist, ist der grundsätzliche Trend – Texte aus dem Bretagne-Stoffkreis weisen mehr Chiasmen auf – in beiden Tabellen erkennbar.

Auffällig ist auch, dass in einigen Texten keine der ausgegebenen Textstellen einen Score ≥ 11 aufweist. Zu diesen gehört auch der ›Vorauer Alexander‹. Bedenkt man, dass dieser in weiten Teilen den ersten rund 1500 Versen des ›Straßburger Alexander‹ entspricht, ließe sich für Letzteren folgern, dass die ›Chiastizität‹ des Textes im zweiten Textabschnitt begründet liegt. Tatsächlich zeigt eine unserer qualitativen Arbeiten (Brandes 2022), dass der Beginn des ›Straßburger Alexander‹ eher von (auch antithetischen) Parallelismen geprägt ist.

Rang	Text	Verse/ Stilmittel
1	TRH	492
2	DA	652
3	AXV	767
4	AXR	1082
5	AXS	1183
6	KAR	1220
7	TRY	1230

Score Relativ	Text	Rang
26,9	AXV	1
15,5	AXS	2
12,3	DA	3
11,2	TRH	4
10,9	ROL	5
8,3	ENE	6
7,6	GAR	7

8	ER	1274	7,3	TRY	8
9	IW	1361	6,3	TR	9
10	TR	1504	5,9	REN	10
11	TRO	1511	5	IW	11
12	ENE	1938	4,8	KAR	12
13	PZ	2481	4,4	ER	13
14	WGL	2928	4	AXR	14
15	TAN	3008	3	TRO	15
16	ROL	3031	2,3	TRU	16
17	AXU	3111	2	WH	17
18	REN	3318	1,5	TAN	18
19	WH	3501	1,4	PZ	19
20	TRU	3751	1,4	CRO	20
21	GAR	5330	1,3	AXU	21
22	MEL	6421	1,2	WGL	22
23	WUT	10604	1	MEL	23
24	CRO	30077	0,5	WUT	24
25	GL	–	0	GL	25
26	MAN	–	0	MAN	26
27	GWTK	–	0	GWTK	27
28	WGM	–	0	WGM	28
29	LZ	–	0	LZ	29
30	EIL	–	0	EIL	30

Tabelle 3: Parallelismusdetektion

Betrachtet man die Parallelismen-Ergebnisse, ergibt sich ein etwas anderes Bild. Antikenromane sind nun mehrheitlich in der oberen Hälfte zu finden – drei Alexanderromane sogar innerhalb der ersten fünf Plätze. Die Platzierungen der Artusromane fallen nun – konträr zu den Antikenromanen – insgesamt niedriger aus. In Bezug auf karlsepische Texte bleibt das Bild ähnlich. Der ›Karl‹ des Strickers weist erneut die meisten Funde auf; in den

übrigen Texten dieses Stoffkreises ließen sich vergleichsweise wenige der gesuchten Stilmittel finden.

Sortiert nach relativem Score, zeichnet sich der gleiche Trend – bei Unterschieden in den Details – ab: Texte der *Matière de Rome* rangieren deutlich weiter oben als die übrigen Texte. Die quantitativen Funde bestätigen damit unsere qualitative Analyse (Brandes 2022). Das ist insofern gut, als es zeigt, dass unser doch ›reduziertes‹ Modell der Stilmittel – so entfallen etwa die Möglichkeiten eines Syntax-Parsers – uns nicht daran hindert, die grundsätzliche Tendenz eines Textes bezüglich seiner Stilmittelhäufigkeit erkennen zu können.

Die hier dargelegten Beobachtungen sind noch leichter erkennbar, wenn man die Funde nach Stoffkreisen sortiert:

Stoffkreis	rel. Score Parallelismen
Antikenroman	8,3
Karlsepik	4,8
Bretagne	3,3

rel. Score Chiasmen	Stoffkreis
6,5	Bretagne
2,8	Karlsepik
2,5	Antikenroman

Tabelle 4: Relativer Score der Stoffkreise

Hier ist deutlich erkennbar, dass die Antikenromane zum Parallelismus zu tendieren scheinen, während die Artustexte vorsichtig als ›Stoffkreis des Chiasmus‹ bezeichnet werden könnten. Die Karlsepischen Texte nehmen – wohl dank Strickers ›Karl‹ – stets einen Mittelplatz ein. Generell wurden mehr Parallelismen als Chiasmen detektiert. Das liegt – so unsere Vermutung – in einem Detail unserer Detektion begründet: Geachtet wird, wie oben erläutert, darauf, ob das erste und dritte sowie das zweite und vierte (oder im Fall des Chiasmus das erste und vierte bzw. zweite und dritte, aber das ist hier nicht relevant) stilmitteltragende Wort von gleicher Wortart sind. Es gibt jedoch keinen Mechanismus, der ausschließt, dass alle Wörter

die gleiche Wortart haben können – einfach, weil Beispiele wie *ein man ein wîp, ein wîp ein man* (›TR‹, V. 129) dann ausgeschlossen würden. Dies hat zur Folge, dass bei der Parallelismus-Detektion einige synonyme und synthetische Parallelismen detektiert wurden, obwohl diese Sonderform des Parallelismus nicht zwangsläufig im Fokus unserer Bemühungen lag. Solche ›Reihungen‹ (von der Aufzählung unterscheiden sie sich nur durch das Merkmal der syntaktischen Abhängigkeit) sind natürlich ›einfach‹ umzusetzen und kommen in den Texten vielfach vor.

Wie in Abschnitt 2.1 erwähnt, wird man den mittelhochdeutschen Texten jedoch nur bedingt gerecht, wenn man die untersuchten Wortstellungsfiguren streng in Parallelismus und Chiasmus einteilt. Dies zeigen auch zahlreiche Funde, die – je nach Auslegung der Textstelle – sowohl als Chiasmus als auch als Parallelismus bezeichnet werden könnten. Zusätzlich zum oben genannten Beispiel aus dem ›Tristan‹ sei folgende Passage genannt, die maschinell detektiert wurde: *der* [*arme*] *und der* [*riche*] *der* [*junge*] *und der* [*alte*]. Diese für das Mittelhochdeutsche fast schon topische Formel, die schlicht ›alle‹ meint, stammt in diesem Fall aus dem ›Rennewart‹ (V. 25806f.). Natürlich liegt hier syntaktisch gesehen ein Parallelismus vor. Semantisch ist der Fall nicht ganz so eindeutig. Für sich genommen stellen *arme* und *riche* sowie *junge* und *alte* jeweils eine Antithese dar. Den gesamten Textauszug semantisch chiastisch zu deuten – nämlich so, dass ›arm und alt‹ also ›reich und jung‹ umklammern, wäre plausibel, wenn man die Setzung, dass ›arm‹ und ›alt‹ hier pejorativ konnotiert sind, wohingegen ›reich‹ und ›jung‹ meliorativ verwendet werden, mitträgt.

Dass die Trennung in Chiasmus und Parallelismus nur bedingt ihren Sinn hat, macht auch der Sonderfall des antithetischen Parallelismus deutlich. Während die Syntax parallelistisch organisiert ist, stellt sich die Semantik – in einigen Fällen auch chiastisch – entgegen. Wir haben uns daher – wegen der im Mittelalter nicht vorhandenen Trennung in Parallelismus und Chiasmus und den diese Dichotomie ohnehin unterlaufenden Sonder-

fällen wie dem antithetischen Parallelismus – dafür entschieden, unsere Ergebnisse auch zu Wortstellungsfiguren insgesamt darzustellen:

Rang	Text	Vers/Stilmittel
1	TRH	405
2	MAN	497
3	DA	565
4	AXR	722
5	AXV	767
6	TRY	769
7	TR	782
8	ER	849
9	KAR	872
10	TRU	938
11	TRO	1061
12	IW	1167
13	AXS	1183
14	PZ	1460
15	GAR	1640
16	WGL	1673
17	ENE	1696
18	AXU	2000
19	TAN	2005
20	GL	2083
21	WH	2334

Score Relativ	Text	Rang
18,8	AXV	1
17,4	GAR	2
16,9	WGM	3
12,2	MAN	4
10,6	AXS	5
9,6	TRH	6
9,5	DA	7
8,7	TRY	8
8,6	TR	9
7,9	ROL	10
6,4	ENE	11
5,8	TRU	12
5,6	REN	13
4,6	KAR	14
4,4	ER	15
4,3	AXR	16
3,9	IW	17
3,1	TRO	18
2,2	WH	19
1,9	CRO	20
1,7	PZ	21

22	REN	2433	1,6	GL	22
23	WUT	2651	1,5	WGL	23
24	ROL	3031	1,5	TAN	24
25	MEL	3211	1,4	AXU	25
26	WGM	6115	1,4	WUT	26
27	CRO	15039	1,3	MEL	27
28	GWTK	–	0	GWTK	28
29	LZ	–	0	LZ	29
30	EIL	–	0	EIL	30

Tabelle 5: Wortstellungsfiguren

Doppelte Funde, also solche, die sowohl bei der Suche nach Chiasmen als auch nach Parallelismen detektiert wurden, wurden nur einfach – mit dem je höchsten Score – gezählt. Das bisherige Bild bestätigt sich. Es gibt kein eindeutiges Gattungssignal, allerdings ansatzweise eine Tendenz. Im oberen Drittel gibt es mit dem ›Karl‹ nur einen karlsepischen Text, die übrigen neun Plätze werden von Antikenromanen und Artusromanen besetzt. Auch wenn karlsepische Texte nicht weit oben platziert sind, so gibt es doch zumindest für jeden der karlsepischen Texte Resultate – keine Funde kommen bei karlsepischen Texten nicht vor. Mit Ausnahme des ›Karl‹ folgen diese Texte in der linken Tabellenhälfte sogar direkt aufeinander (Rang 21–24), sind sich also ähnlich; die drei Texte, die zum ›Willehalm‹-Umkreis gehören (Rang 21–23), weisen besonders geringe Abstände zueinander auf. – Hier kann sehr vorsichtig vermutet werden, dass die gemeinsame *materia* der Texte – denn Autorschaft und Entstehungszeit können es nicht sein – Auswirkungen auf die Stilmittelhäufigkeit hat. Dieses Bild wird allerdings von den weiter auseinander platzierten Alexandertexten unterlaufen. Doch nur weil es kein oder nur ein schwaches Gattungs- oder Stoffsignal in den Antiken-

romanen gibt, heißt das nicht, dass dies auch für karlsepische Texte gelten muss. Zu bedenken ist in diesem Zusammenhang auch die formal von den anderen Gattungen abweichende Gruppe der altfranzösischen Karlsepik (zunächst in Laissen gedichtete *chansons de geste*). Möglicherweise hat die intensive Auseinandersetzung der deutschen Verfasser mit diesen Texten auf die eigene Wahl von Stilmitteln eingewirkt.[11]

In Bezug auf den relativen Score fällt auf, dass zwei der vier oben platzierten Texte, der ›Vorauer Alexander‹ und das ›Mantel‹-Fragment, zu den kürzesten Texten des Corpus gehören. Nur sehr wenige Funde mit einem Score ≥ 11 genügen also, um einen hohen relativen Score aufzuweisen. Hier kann (zumindest teilweise) Verzerrung aufgrund der Textlänge vermutet werden. Auch hier gilt: Die Reihenfolge der Texte mag im Detail unterschiedlich sein, die grundsätzliche Verteilung der Stoffkreise bleibt ähnlich.

Insgesamt, so kann vorerst konstatiert werden, ist allein die Häufigkeit von Stilmitteln der Wortstellungsfiguren kein eindeutiger Indikator für die Zugehörigkeit zu einem der Stoffkreise. So haben Texte der Karlsepik etwa vergleichsweise wenig Funde, aber nicht alle Texte mit verhältnismäßig wenigen Funden gehören der Karlsepik an. Hinzu kommt der geringe Anteil karlsepischer Texte am Corpus sowie der mit ›Karl‹ selbst betitelte Text, der innerhalb seines Stoffkreises immer die Ausnahme bildet. Und doch, so kann das Ergebnis zumindest mit Blick auf die Texte Wolframs und des Pleiers auch vorsichtig formuliert werden, gruppieren sich die einander ähnlichen Texte eher nach dem Stoffkreis als nach der Autorschaft.

Wenn auch die bloße Quantität der untersuchten Stilmittel wohl kein Gattungssignal erkennen lässt, so lassen die Funde möglicherweise doch noch weitere Aussagen über die Texte zu. Hierzu folgende Abbildung:

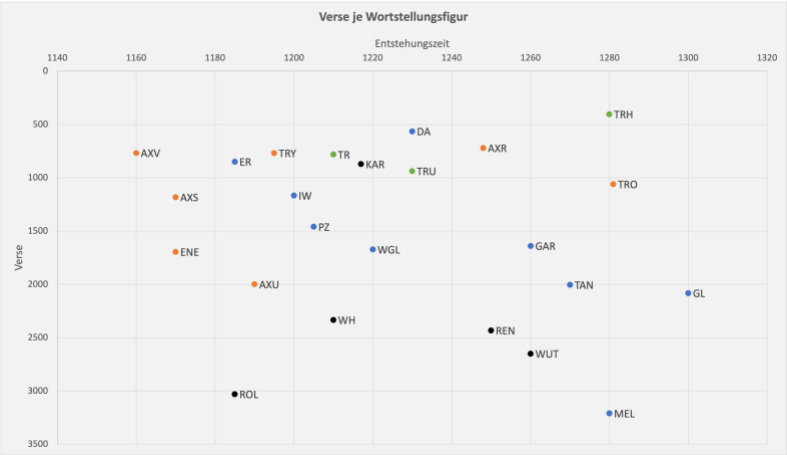


Abb. 2: Verse je Wortstellungsfigur

In dieser Abbildung fehlen die Texte ohne Funde (›Göttweiger Trojanerkrieg‹, ›Lanzelet‹ und Eilharts ›Tristrant‹) sowie ›Wigamur‹ und ›Crône‹, da beide Texte die übrigen in einer linearen Darstellung zu nah aneinanderrücken würden. Die vollständige Abbildung kann im hier eingesehen werden.

Abgebildet ist, mit welcher Anzahl von Versen Abstand (y-Achse) ein Stilmittel detektiert wurde und wann[12] (x-Achse) der Text entstanden ist. Erkennbar ist, dass früheste Texte ein gewisses Mindestmaß an Stilmitteln zu enthalten scheinen. Erst im ›Rolandslied‹ und in nur wenigen Texten nach ihm wurde ≥ alle 2000 Verse ein Stilmittel entdeckt. In der überwiegenden Mehrheit der Texte konnte alle 500 bis 2000 Verse ein Stilmittel detektiert werden. Davon weichen nur neun Texte ab. In acht dieser Texte ist die Zahl ›benötigter‹ Verse größer, davon sind vier Texte der Karlsepik zuzuordnen, die übrigen der Artusepik, wobei hier keiner der ›Klassiker‹ und mit ›Meleranz‹ und ›Gauriel‹ die zwei spätesten Artusromane betroffen sind.

Die Dimension ›Entstehungszeit‹ prädeterminiert zwar nicht, ob ein Text viele oder wenige Stilmittel aufweist, dennoch ist ein grundsätzlicher

Trend erkennbar: Später geschriebene Texte enthalten weniger Stilmittel,[13] wobei auch hier die Ausnahme in Form von Heinrichs ›Tristan‹ die Regel bestätigt. Es ist jedoch nicht so, dass die späten Texte mit wenigen Stilmitteln, etwa der ›Meleranz‹, erheblich weniger Stilmittel hätten als frühere Texte mit wenigen Funden, etwa das ›Rolandslied‹; beide liegen auf der y-Achse nah beieinander – es fehlt in der zweiten Hälfte des 13. Jh. also schlicht eine gewisse Anzahl an Texten, in denen alle 1000 oder 1500 Verse ein Stilmittel gefunden wurde. Separiert man die Abbildung nach Stoffkreis, wird klar, dass der ›abnehmende‹ Trend nicht für alle Stoffkreise gilt:

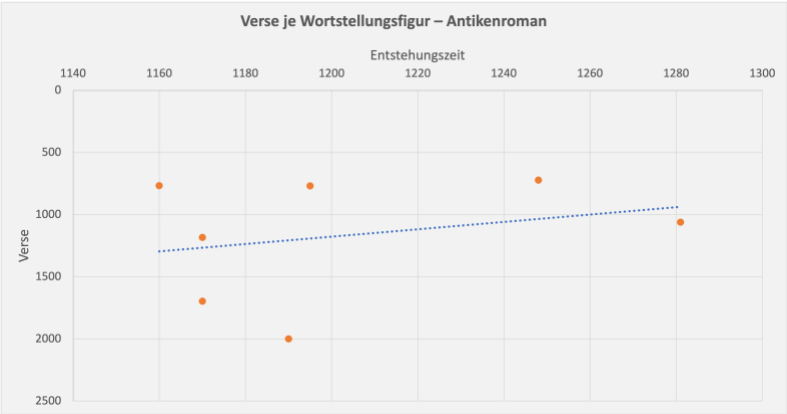


Abb. 3: Verse je Wortstellungsfigur – Antikenroman (lineare Trendlinie)

Im Gegenteil ist in dieser Abbildung, die nur noch die Datenpunkte der Antikenromane enthält, zu sehen, dass sogar ein leichter Aufwärtstrend vorhanden ist. Aufgrund der geringen Textmenge von nur sieben sollte diesem leichten Trend allerdings nicht allzu viel Gewicht beigemessen werden. Noch weniger Datenpunkte weist die Abbildung zur Karlsepik auf:

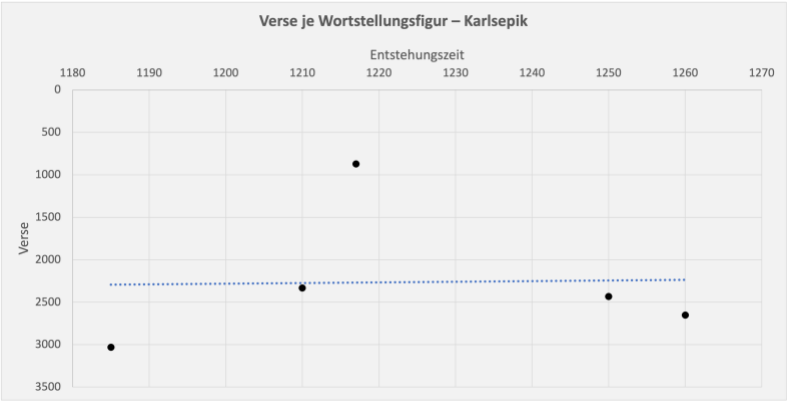


Abb. 4: Verse je Wortstellungsfigur – Karlsepik (lineare Trendlinie)

Hier ist nun kein Trend, allenfalls ein minimaler, so man ihn so nennen möchte, erkennbar. Dass in vier der fünf Texte vergleichsweise wenig Stilmittel gefunden werden konnten, zeigten auch die bisherigen Ausführungen. Die größte Textmenge weist die *Matière de Bretagne* auf, der die folgende Abbildung – hier mit den Ausreißern ›Crône‹ und ›Wigamur‹ – gilt.

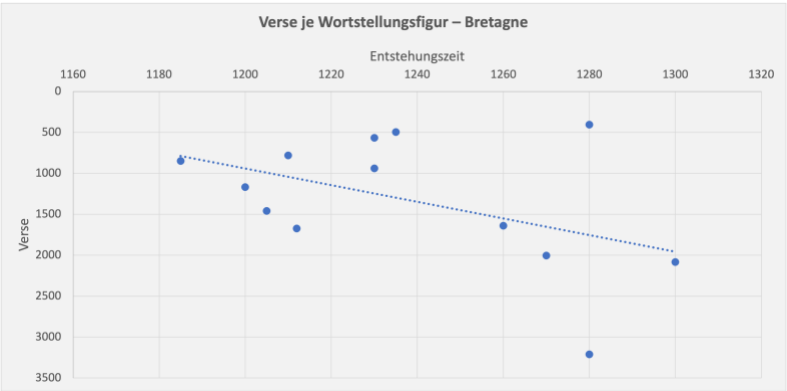


Abb. 5: Verse je Wortstellungsfigur – Bretagne (lineare Trendlinie)

Der in Abbildung 1 sichtbare Abwärtstrend über die Zeit hinweg basiert also im Wesentlichen auf den Texten der Artusromane. Hier ist auch der ›stärkste‹ Trend erkennbar.

Zu fragen wäre nun, wie mit den hier erzielten Ergebnissen weiter verfahren werden soll. Einerseits muss auf ihnen aufgebaut werden, indem die Genauigkeit des Detektors weiter verbessert wird – dazu ist bereits ein iteratives Verfahren geplant, bei dem die Ergebnisse von Durchgang zu Durchgang besser werden, was auch zur Folge hätte, dass alle POS-Tags berücksichtigt werden können und der Fokus auf die Zusammensetzung der Stilmittel erweitert wird. Denn wenn die Häufigkeit nicht oder nur geringfügig an die Stoffkreiszugehörigkeit geknüpft ist, ist dies möglicherweise bei einer quantitativen Auswertung der Wortarten der Fall, aus denen die jeweiligen Stilmittel bestehen. Auch eine Ergänzung um unsere Metaphern-Detektion steht aus.[14]

Anderseits bieten die jetzigen Ergebnisse bereits einen explorativen Zugang zu den Texten. Intuitiv würde man Texten wie dem ›Parzival‹, dem ›Willehalm‹, aber auch dem ›Eneas‹ und dem ›Tristan‹ wohl einen rhetorisch ›artifiziellen‹ Stil attestieren – gerade Wolfram und Gottfried gelten als Dichter, deren Sprachvirtuosität heraussticht. Auch sind späte oder ›postklassische‹ Texte wie der ›Garel‹, der ›Gauriel‹ oder auch der ›Lanzelet‹ (auf den das Attribut ›spät‹ gar nicht zutrifft) in Bezug auf ihre literarische Qualität – wozu eben auch Stil gehört – als ›schlicht‹ abgewertet worden (exemplarisch zum ›Gauriel‹ die Zusammenfassung von Achnitz 1997). Die hier gezeigten Tabellen decken diese Positionen der Forschung jedoch nur bedingt ab – wieso?

Zunächst einmal – das ist eingangs bereits gesagt worden – besteht Stil aus mehr als der bloßen Menge der verwendeten Stilmittel. Darüber hinaus beziehen sich die hier vorliegenden quantitativen Ergebnisse auf zwei Stilmittel, nicht auf alle. Nicht zuletzt umfasst unser Modell der Stilmittel eben nicht alle möglichen Ausprägungen der Stilmittel. So könnte etwa für Wolframs Texte vermutet werden, dass die in ihnen verwendeten Wort-

stellungsfiguren deutlich komplexer sind, als es unsere Stilmitteldetektion bisher erfassen kann.

Autor, Entstehungszeit, Stoffkreis und ›Szenen‹ stehen, wie oben erläutert, im Verdacht, den Stilmittelgebrauch zu beeinflussen. Für die Texte von ein und demselben Autor – Hartmann, Wolfram und Pleier – war erkennbar, dass durchaus Unterschiede in der Menge der Stilmittelverwendung vorhanden sind. Ein klares Autorsignal ist in dieser Stichprobe nicht bemerkbar. Dies gilt auch für die Entstehungszeit. Auch wenn es – von Stoffkreis zu Stoffkreis unterschiedlich stark ausgeprägt – leichte Tendenzen zu geben scheint, gibt es zu jeder Zeit Texte, in denen besonders viele, aber auch besonders wenige Stilmittelfunde vorliegen. Auch die Stoffkreise selbst lassen keine eindeutigen Cluster erkennen. Karlsepische Texte und zum Tristan-Stoffkreis gehörende Texte scheinen allerdings eine Tendenz zu wenigen bzw. zu vielen Stilmitteln zu haben – eine Tendenz, die dann von den jeweiligen Ausnahmen der Textgruppen (›Karl‹ respektive ›Tristrant‹) unterlaufen wird. Szenen, etwa Figurenrede oder Kämpfe, könnten den Stilmittelgebrauch wohl beeinflussen. Zu einzelnen Szenentypen und deren Verteilung in den Einzeltexten können hier (noch) keine quantitativen Aussagen getroffen werden. Allgemeine Beobachtungen könnten aber erste Vermutungen zulassen. So beinhalten Texte der *chansons de geste* vergleichsweise viele Schlachten. Möglich ist etwa, dass Schlachten weniger Wortstellungsfiguren, aber umso mehr Metaphern enthalten. Dies zu überprüfen, gehört zu den nächsten geplanten Schritten unseres Projektes.

Dieses Resümee sollte neben allen Potenzialen der neuen Methoden die besondere Herausforderung deutlich gemacht haben: Auszuloten, welche Bedeutung den computationell gewonnenen Befunden beigemessen werden kann, ist aus literaturwissenschaftlicher Sicht die wichtigste und schwierigste Aufgabe. Und auch wenn in der Forschung teils stark gemacht wird, dass es nicht nur auf die Menge von Stichproben und Variablen ankommt, da schon einzelne bereits sehr plausibel auf große Trends hin-

weisen könnten, ist klar, dass bei einem literaturwissenschaftlichen Thema wie ›Stil und Gattung‹ mit der quantitativen Untersuchung von Wortstellungsfiguren allenfalls ein wichtiger Anfang gemacht sein kann. Literarische Texte sind zu komplex, als dass eventuell genrebedingte Unterschiede zwischen ihnen an derart wenigen Merkmalen festzumachen wären. Differenzen zwischen den Stoffkreis-Bearbeitungen wird quantitativ wohl nur über Merkmalsbündel beizukommen sein – diese in kleinen Schritten zu erarbeiten, verstehen wir aber als lohnenden Weg.

4. Literaturwissenschaft als Experiment?

Literaturwissenschaft erhält – davon konnte hier nur ein flüchtiger Eindruck vermittelt werden – in Verbindung mit den Möglichkeiten der Statistik und der Informatik (samt benachbarter Disziplinen wie der Computerlinguistik) Experimentcharakter. Bis in die 2010er-Jahre hinein schien das Paradigma der Exploration vorherrschend zu sein. Es galt umzusetzen, was technisch umsetzbar war, um erst dann zu überprüfen, ob und was mit den Ergebnissen über Literatur ausgesagt werden kann (als technische Machbarkeitsstudie kann etwa auch Jockers 2013 gelesen werden).[15]

Spätestens mit Ted Underwoods ›Distant Horizons‹ (2019) galten rein explorative Vorgehensweisen, bei denen vor allem auch technische Potenz im Vordergrund stand, als obsolet. Ausgangspunkt für ein *Distant Reading* oder generell für computergestützte Literaturwissenschaft sollte stets literaturwissenschaftliches Hintergrundwissen sein (Underwood 2019, S. 17). Aus diesem bilden sich Interessen und Fragestellungen, die dann experimentell untersucht werden. Es gilt danach zu fragen, welche Variable Aussagekraft für welches Phänomen hat (also hier die Variable ›Stilmittelhäufigkeit‹ für das Phänomen ›Stoffkreiszugehörigkeit‹). Je mehr verschiedene Variablen in Bezug auf ein bestimmtes Phänomen untersucht werden, umso aussagekräftiger sind die zu beobachtenden Trends.

Zum Abschluss soll kurz der Versuch unternommen werden, eine weitere Variable zur Untersuchung von Textähnlichkeiten heranzuziehen. Braun und Ketschik (2019) unternehmen es, anhand einer Netzwerkanalyse, die Figurenbeziehungen untersucht, Aussagen über die Märchenhaftigkeit von Artusromanen treffen zu können. Dabei wird gezeigt, dass der ›Iwein‹ weniger komplex in der Interaktion der Figuren ist als der ›Erec‹ (Braun/Ketschik 2019, S. 59); dies fällt vor allem auf, wenn man beide Figurennetzwerke betrachtet.[16] Der ›Parzival‹ ist wiederum deutlich komplexer als die vorigen Texte. Die Unterschiede fallen so deutlich aus, dass »zunächst einmal die beiden Artusromane Hartmanns dem ›Parzival‹ Wolframs gegenüber [gestellt werden]« (Braun/Ketschik 2019, S. 70). Diese Gegenüberstellung trifft partiell auch auf die von uns gemachten Stilmittelfunde zu, wenn auch auf den ersten Blick in umgekehrter Weise. Denn während ›Erec‹ und ›Iwein‹ einen relativen Score von ~4 aufweisen, kommt der ›Parzival‹ nur auf einen relativen Score von 1,7. Es könnte also sein, dass ›Komplexität‹ gemessen an der Figureninteraktion mit einer spezifischen Verwendung der Stilmittel korreliert, so dass diese bei höherem ›Schwierigkeitsgrad‹ entweder seltener vorhanden sind oder – was bei Wolfram wahrscheinlicher scheint – mit unseren bisherigen Ansätzen schwieriger detektiert werden können, weil sie ›komplexer‹ sind.

Wird unsere Variable also um weitere ergänzt, wie es eben nur angedeutet wurde und wie wir es mit Metaphernhäufigkeiten, Zusammensetzung der Wortstellungsfiguren, inhaltlichem Kontext der Metaphern und Wortstellungsfiguren, Szenentypen-Analysen sowie dem Vergleich mit Texten außerhalb der drei Stoffkreise planen,[17] kann eine für das Mittelhochdeutsche umfassende Studie zu Textgruppenkonstitutionen erwartet werden. Eine solche Studie wäre kein Experimentieren um des Experimentierens willen. Sie wäre den Anforderungen an eine Literaturwissenschaft, die nun immer mehr von computationellen und statistischen Methoden Gebrauch machen kann und sollte, angemessen – denn komplexe Phänomene wie ›Gattung‹ und ›Stil‹ werden nur dann hinreichend erforscht

werden können, wenn von den ebenso komplexen Untersuchungsmethoden, die zur Verfügung stehen (und zukünftig zur Verfügung stehen werden), Gebrauch gemacht wird.

Anmerkungen

1 Das Projekt unter der Leitung von Prof. Dr.-Ing. Joachim Denzler und Jun.-Prof. Dr. Sophie Marshall ist Teil des von der DFG finanzierten Schwerpunktprogramms 2207 ›Computational Literary Studies‹. Informationen zu allen Projekten und unseren publizierten Ergebnissen sind hier einsehbar: https://dfg-spp-cls.github.io/.

2 Was Stil genau ist, variiert in entsprechenden Untersuchungen je nach zugrundeliegender Theorie. Stil wird hier als den Ausdrucksformen funktionszuweisendes Konzept (Hübner 2015) verstanden. Stilfiguren sind demnach Teil dieser Ausdrucksformen. Zur weiteren Diskussion von Stil und Stilkonzepten s. Andersen [u. a.] (2015) und Gumbrecht (1986). Für die nachfolgende Untersuchung soll betont sein, dass Stil nicht allein mit der Verwendung bestimmter Stilmittel gleichgesetzt wird.

3 Weitere Schwerpunkte der Stilometrie sind u. a. – wenn auch bei weitem nicht so etabliert wie Autorschaftsattribution – *topic modelling* (Viehhauser 2018) und das noch neue Feld der automatisierten Szeneneinteilung (Zehe [u. a.] 2021). Für die Mediävistik und unser Projekt von Interesse ist vor allem die Beobachtung, dass es nicht den einen Indikator gibt, anhand dessen alle Gattungen untersucht werden können (Viehhauser 2015).

4 Eine ›klassisch-analoge‹ Untersuchung aus dem Projekt beschäftigt sich mit Szenen entsprechender Texte, die Gabenlogisches verhandeln (Brandes 2022).

5 Bereitgestellt werden die maschinenlesbaren Dateien der Texte von der Mittelhochdeutschen Begriffsdatenbank (MHDBDB), der dafür an dieser Stelle herzlich gedankt wird.

6 Dass das Detektieren der drei Stilmittel Chiasmus, Parallelismus und Metapher bereits durchaus ambitioniert ist, zeigen allein bisherige Arbeiten in der Chiasmusdetektion. Diese befassten sich nahezu ausschließlich mit dem Sonderfall der Antimetabole (Dubremetz/Nivre 2017) und waren aus computerlinguistischer Sicht zweifellos ein wichtiger Entwicklungsschritt; für literaturwissenschaftliche Untersuchungen sind sie aufgrund der geringen Genauigkeit aber irrelevant.

7 Underwood macht klar, dass Korrelation noch längst keine Kausalität darstellt. Ebenso auffällig scheint aber zu sein, dass beobachtbare Trends auf sehr spezieller Ebene (hier die Ebene zweier Wortstellungsfiguren) sich in der Regel auf größerer Ebene wiederfinden. In Underwoods Untersuchung gilt dies etwa für die spezielle Ebene der Farbadjektive, deren häufigeres Vorkommen auf grundsätzlich häufigeres Vorkommen von beschreibenden Adjektiven hinwies (Underwood 2019, insb. S. 11–14). Für unser Experiment würde das bedeuten, der Spezialfall weniger Stilmittel der Wortstellung lässt durchaus Rückschlüsse auf den allgemeinen Gebrauch von Stilmitteln im jeweiligen Werk zu.

8 Als Antithese kann hingegen bereits eine bloße Zusammenführung gegensätzlicher Inhalte gelten (Villwock 1992, Sp. 722–750). Hier liegt also die Struktur A–B vor, die Antithese ist in diesem Sinne damit keine der hier untersuchten Wortstellungsfiguren.

9 30 Token sind in der Computerlinguistik ein etablierter Umfang, wenn es um die Untersuchung mikrostruktureller Textphänomene geht, und solch ein Umfang kommt auch bei der bisherigen Antimetabole-Detektion zum Einsatz (Dubremetz/Nivre 2017).

10 *Word Embeddings* ermöglichen es, Wörter im Hinblick auf ihre Semantik repräsentieren zu können. Die semantische Relation, die bei *Word Embeddings* beachtet wird, ist das Umfeld von Wörtern. *Word Embeddings* liegt also die Annahme zugrunde, die Bedeutung eines Wortes ergäbe sich aus dessen Kontext. Um maschinenlesbar zu sein, erfolgt die Abbildung von Wörtern in einem mehrdimensionalen Vektorraum (Pilehvar/Camacho-Collados 2020, S. 5f.). Jedes Wort erhält aufgrund seiner Relation zu anderen Wörtern seinen eigenen Wert; in einem 4-dimensionalen Vektorraum beispielsweise (2,7,5,9), üblich sind deutlich höherdimensionale Abbildungen über 750 Dimensionen. Zur Entstehung eines Vektors s. Pilehvar/Camacho-Collados 2020). Unsere *Word Embeddings* wurden mit fastText erzeugt.

11 Dies möchte unser geplantes Anschlussprojekt näher in den Blick nehmen.

12 Die Entstehungszeiten der Texte sind Herweg (2013) entnommen. Ist eine Zeitspanne angegeben, wurde der frühere Zeitpunkt ausgewählt. Ausnahmen bilden das ›Rolandslied‹, für das angenommen werden kann, dass es nach dem ›Straßburger Alexander‹ verfasst wurde, sowie die Texte des Pleiers, die alle um 1260/80 datiert werden (Herweg 2013, S. XXII). Es kann bei der Entstehung der Texte von der Reihenfolge ›Garel‹, ›Tandareis und Flordibel‹, ›Meleranz‹ ausgegangen werden, weshalb sich diese in der Abbildung widerspiegelt. Wichtig ist

dann nicht die exakte Jahreszahl, sondern die Beobachtung, dass in späteren Texten weniger Stilmittel gefunden worden sind.

13 Dies ist eine Tendenz, die wohl auch die ›analoge‹ Forschung stützen würde. Lange Zeit – und für einige Texte gilt dies bis heute – galten Texte, die nicht zu den ›Klassikern‹ gehören, als weniger ›kunstvoll‹, als weniger ›gut gemacht‹. Häufigkeit von Stilmitteln und Qualität sind nun freilich verschiedene Dinge, doch auf beiden Ebenen kann von ›wenig‹ Stil (›Stil‹ gewichtet dann das bloße Vorhandensein von Stilmitteln stark) gesprochen werden. Die Ausnahme dieser Tendenz bildet dann der ›Tristan‹ Heinrichs von Freiberg, der als einer der letzten Texte die meisten Funde aufweist.

14 Erste Metaphern-Ergebnisse werden auf der DH 2022 vorgestellt. Thema ist vor allem die Metaphernverwendung in mittelhochdeutschen Texten mit thematischem Bezug zu ›Asien‹.

15 Ein Paradebeispiel für eine explorative Tool-Sammlung ist Voyant – wie sehr etwa Wordclouds die Literaturwissenschaft bereichern, ist dort nicht die Fragestellung; lediglich ein ›anderes‹ Kennenlernen der Texte steht im Vordergrund. Statistisch reichhaltiger ist etwa das Tool ›Correlations‹.

16 Die Figurennetzwerke können unter folgendem Link eingesehen werden (Braun/ Ketschik 2019, S. 61 Anm. 23): 10.17879/55189456328.

17 Denn denkbar wäre ja auch, dass die deutschsprachigen Varianten der französischen Heldenepik, die *chansons de geste*, in Bezug auf die Stilmittelverwendung näher an der ›deutschen‹ Heldenepik sind als an den Stoffkreisen *de Rome* und *de Bretagne*. Ebenso gut möglich ist aber auch, dass die *Trois Matières* ein gemeinsames Cluster gegenüber den anderen Gattungen bilden.

Literaturverzeichnis

Primärliteratur

Gottfried von Straßburg: Tristan und Isold, hrsg. von Walter Haug und Manfred Günter Scholz, Berlin 2012.

Pfaffe Lambrecht: Alexanderroman. Mittelhochdeutsch/Neuhochdeutsch, hrsg., übers. und komm. von Elisabeth Lienert, Stuttgart 2007.

Ulrich von Türheim: Rennewart. Aus der Berliner und Heidelberger Handschrift hrsg. von Alfred Hübner, Berlin/Zürich 1968.

Sekundärliteratur

Achnitz, Wolfgang: Interpretationsansätze, in: Ders. (Hrsg.): Der Ritter mit dem Bock. Konrads von Stoffeln ›Gauriel von Muntabel‹. Neu hrsg., eingeleitet und komm. von dems., Tübingen 1997 (Texte und Textgeschichte 46), S. 195–232.

Ackermann, Irmgard: Art: Parallelismus, in: Metzler Lexikon Literatur. Begriffe und Definitionen, (2007), S. 570 Sp. 2.

Allison, Sarah/Heuser, Ryan/Jockers, Matthew/Moretti, Franco/Witmore, Michael: Quantitativer Formalismus. Ein Experiment, in: Algee-Hewitt, Mark/Allison, Sarah/Gemma, Marissa/Heuser, Ryan/Jockers, Matthew/Pestre, Dominique/Steiner, Erik/Tevel, Amir/Walser, Hannah/Witmore, Michael/Yamboliev, Irena (Hrsg.), unter der Leitung von Franco Moretti: Literatur im Labor, Paderborn 2017, S. 17–48.

Andersen, Elizabeth/Bauschke-Hartung, Ricarda/McLelland, Nicola/Reuvekamp, Silvia (Hrsg.): Literarischer Stil. Mittelalterliche Dichtung zwischen Konvention und Innovation. XXII. Anglo-German Colloquium Düsseldorf, Berlin/Boston 2015.

Brandes, Phillip: Von Gaben mit Stil erzählen. Zur Korrelation von reziproken Logiken in Inhalt und Form im ›Straßburger Alexander‹, ›Lanzelet‹ und ›Willehalm‹, in: PBB 144 (2022), S. 368–395.

Braun, Manuel/Ketschik, Nora: Soziale Netzwerkanalysen zum mittelhochdeutschen Artusroman oder: Vorgreiflicher Versuch, Märchenhaftigkeit des Erzählens zu messen, in: Das Mittelalter 24 Nr. 1 (2019), S. 54–70 (online).

Büttner, Andreas/Dimpel, Friedrich Michael/Evert, Stefan/Jannidis, Fotis/Pielström, Steffen/Proisl, Thomas/Reger, Isabella/Schöche, Christof/Vitt, Thorsten: ›Delta‹ in der stilometrischen Autorschaftsattribution, in: Zeitschrift für digitale Geisteswissenschaften (2017) (online).

Dubremetz, Marie/Nivre, Joakim: Machine Learning for Rhetorical Figure Detection: More Chiasmus with Less Annotation, in: Proceedings of the 21st Nordic Conference on Computational Linguistics (2017), S. 37–45 (online).

Ehrismann, Gustav: Studien über Rudolf von Ems. Beiträge zur Geschichte der Rhetorik und Ethik im Mittelalter, Heidelberg 1919 (Sitzungsberichte der Heidelberger Akademie der Wissenschaften 8).

Fauser, Markus: Art. Chiasmus, in: Historisches Wörterbuch der Rhetorik, Bd. 2 (1994), Sp. 171–173.

Gumbrecht, Hans Ulrich/Pfeiffer, K. Ludwig (Hrsg.): Stil. Geschichten und Funktionen eines kulturwissenschaftlichen Diskurselements, Frankfurt a. M. 1986 (Suhrkamp Taschenbuch Wissenschaft 633).

Herweg, Mathias: Volkssprachige Großepik im deutschen Mittelalter. Stoffe, poetologische Konzepte, diskursive Profile im Überblick, in: Achnitz, Wolfgang

(Hrsg.): Deutsches Literatur-Lexikon. Das Mittelalter. Bd. 5: Epik (Vers – Strophe – Prosa) und Kleinformen, Berlin/Boston 2013, S. VII–XXVI.

Hübner, Gert: Historische Stildiskurse und historische Poetologie, in: Andersen [u. a.] 2015, S. 17–38.

Jannidis, Fotis: Grundlagen der Datenmodellierung, in: Jannidis, Fotis/Kohle, Hubertus/Rehbein, Malte (Hrsg.): Digital Humanities. Eine Einführung. Stuttgart 2017, S. 99–108.

Jockers, Matthew L.: Macroanalysis: Digital Methods and Literary History, Urbana [u. a.] 2013 (Topics in the Digital Humanities).

Kuhn, Jonas: Einleitung, in: Reiter, Nils/Pichler, Axel/Kuhn, Jonas (Hrsg.): Reflektierte Algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt, Berlin/Boston 2020, S. 9–40.

Ostrowicz, Philipp: Art. Parallelismus, in: Historisches Wörterbuch der Rhetorik, Bd. 6 (2003), Sp. 546–552.

Pilehvar, Mohammad Taher/Camacho-Collados, Jose: Embeddings in Natural Language Processing. Theory and Advances in Vector Representations of Meaning, Toronto 2020 (Synthesis Lectures on Human Language Technologies 47).

Remele, Florian: Theorie und Methode der Gattungsgeschichtsschreibung. Mediävistische Perspektiven, in: Journal of Literary Theory 15 (2021), S. 53–80 (online).

Scaglione, Aldo/Marvin, William P: Art. Compositio, in: Historisches Wörterbuch der Rhetorik, Bd. 2 (1994), Sp. 300–305.

Schneider, Felix/Barz, Björn/Brandes, Phillip/Marshall, Sophie/Denzler, Joachim: Data-Driven Detection of General Chiasmi Using Lexical and Semantic Features, in: SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (2021), S. 96–100 (online).

Underwood, Ted: Distant Horizons. Digital Evidence and Literary Change, London 2019.

Viehhauser, Gabriel: Historische Stilometrie? Methodische Vorschläge für eine Annäherung textanalytischer Zugänge an die mediävistische Textualitätsdebatte, in: Baum, Constanze/Stäcker, Thomas (Hrsg.): Grenzen und Möglichkeiten der Digital Humanities, Sonderband der Zeitschrift für digitale Geisteswissenschaften 1 (2015) (online).

Viehhauser, Gabriel: Digitale Gattungsgeschichten. Minnesang zwischen generischer Konstanz und Wende, in: Zeitschrift für digitale Geisteswissenschaften (2017) (online).

Viehhauser, Gabriel: Digital Humanities ohne Computer? Alte und neue quantifizierende Zugänge zum mittelhochdeutschen Tagelied, in: Bernhart, Toni/Willand, Marcus/Richter, Sandra/Albrecht, Andrea (Hrsg.): Quantitative An-

sätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven, Berlin/Boston 2018, S. 173–203.

Villwock, Jörg: Art. Antithese, in: Historisches Wörterbuch der Rhetorik, Bd. 1 (1992), Sp. 722–750.

Zehe, Albin/Konle, Leonard/Dümpelmann, Lea Katharina/Gius, Evelyn/Hotho, Andreas/Jannidis, Fotis/Kaufmann, Lucas/Krug, Markus/Puppe, Frank/Reiter, Nils/Schreiber, Annekea/Wiedmer, Nathalie: Detecting Scenes in Fiction: A new Segmentation Task, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics Nr. 1 (2021), S. 3167–3177 (online).

Online-Ressourcen

CLS (Computational Literary Studies): https://dfg-spp-cls.github.io/.

DH 2022 (Digital Humanities 2022): https://dh2022.adho.org/.

Digitale Mediävistik. Perspektiven der Digital Humanities für die Altgermanistik. Internationale Tagung Bremen, vom 9. bis 11. Februar 2022: https://www.uni-bremen.de/fb-10/forschung/perspektiven-der-digital-humanities-fuer-die-altgermanistik.

fastText. Library for efficient text classification and representation learning: https://fasttext.cc/.

MHDBDB (Mittelhochdeutsche Begriffsdatenbank): http://www.mhdbdb.sbg.ac.at.

Voyant Tools: https://voyant-tools.org/; https://voyant-tools.org/docs/#!/guide/correlations.

Anschrift der Autorin und Autoren:

Phillip Brandes M.A.
Friedrich-Schiller-Universität Jena
Institut für Germanistische Literaturwissenschaft
Leutragraben 1, Raum 21N08
07743 Jena
E-Mail: phillip.brandes@uni-jena.de

Jun.-Prof. Dr. Sophie Marshall
Friedrich-Schiller-Universität Jena
Institut für Germanistische Literaturwissenschaft
Fürstengraben 18
07743 Jena
E-Mail: sophie.marshall@uni-jena.de

Felix Schneider M.A.
Friedrich-Schiller-Universität Jena
Fakultät für Mathematik und Informatik
Ernst-Abbe-Platz 2
07743 Jena
E-Mail: felix.schneider@uni-jena.de

Friedrich Michael Dimpel / Andre Blessing
Peter Hinkelmanns / Nora Ketschik
Katharina Zeppezauer-Wachauer

Figuren und ihr Handeln

Eine computergestützte Untersuchung von Figurenaktivitäten im Kontext von Figurenreferenzen mit Hilfe des Begriffssystems der MHDBDB

Abstract. Im Kontext von Figurenreferenzen, die mit Hilfe der Annotationsumgebung CRETAnno teilautomatisch annotiert wurden, werden bestimmte Kategorien aus dem Begriffssystem der MHDBDB erfasst, die Aktivitäten von Figuren repräsentieren können. Diese Kategorien werden für vier Figurentypen (weibliche und männliche Hauptfigur, Zofen, Opponenten) ermittelt und mit einer Studie verglichen, die mit manueller Annotation Aktivitäten von Figurentypen erfasst hat – jeweils für ›Iwein‹, ›Tristan‹, ›Partonopier‹ und ›Mauritius von Craun‹. Es zeigt sich unter anderem, dass die automatische Erfassung von Aktivitäten eine Unterscheidung der Figurentypen in diesem Korpus ermöglicht. Mit Ausnahme der Aktivität ›Kopulieren‹, die meist nicht unmittelbar lexikalisch, sondern eher verhüllend oder metaphorisch adressiert wird, werden plausible Ergebnisse erzielt.

1. Rückblick: Die Paidia-Studie

In dem Paidia-Beitrag »Versuch einer quantitativen Analyse von Figurenaktivitäten in ›Iwein‹, ›Tristan‹, ›Partonopier‹ und ›Mauritius von Craun‹ in Analogie zu Computerspielen« (Dimpel 2018) wurde untersucht, inwieweit Aktivitäten wie ›Suchen‹, ›Töten‹ oder ›Argumentieren‹ bestimmten

Figurentypen zugeschrieben wurden. Die Studie konzentrierte sich dabei auf die weibliche und männliche Hauptfigur, ein bis drei Opponenten und je einen Adjuvanten in Gestalt einer Zofe. Auch wenn das Ziel eine quantitative Auswertung war, wurden die Aktivitäten doch zunächst manuell registriert: Gezählt wurde im Rahmen eines Lektürevorgangs, ob die jeweiligen Aktivitäten in einem bestimmten Textsegment den jeweiligen Figuren eingeschrieben sind, und zwar dann, wenn in den Texten explizit von den Aktivitäten berichtet wird. Darüber hinaus erfolgte eine Zählung mitunter auch dann, wenn der Kontext es erlaubt, mit einem hohen Grad an Offensichtlichkeit auf die fragliche Aktivität zu inferieren (vgl. zur Inferenzbildung Jannidis 2004, S. 206).

Auch wenn eine solche manuelle Erfassung in der Regel präziser sein wird als eine automatische, war und ist doch zu betonen, dass zwar viele, aber nicht alle Textbefunde einen eindeutigen Schluss darauf zulassen, ob etwa eine kommunikative Aktivität bereits unter ›Beklagen‹ fällt oder ob eine leichte Auseinandersetzung bereits das Etikett ›Kämpfen‹ verdient. Die mitunter vorhandenen Probleme, eine Textpassage einer Aktivität zuzuordnen, beruhen nicht nur auf der häufig konstatierten Polyvalenz literarischer Texte. Vielmehr hat man mit einem grundlegenden Problem der digitalen Literaturwissenschaft zu tun: Wenn man nicht nur alltagssprachlich umschreiben will, ob eine Figur kämpft, sondern digital über Sein oder Nichtsein einer Aktivität entscheiden soll, ist es nötig, die Aktivität möglichst eindeutig zu modellieren. Willard McCarty hat die Modellbildung und Modellweiterentwicklung als die zentrale Tätigkeit von digitalen Literaturwissenschaftler*innen beschrieben, die nicht etwa nur einen Mangel wie unvollständige Explizitheit kompensiere, sondern die vielmehr häufig unmittelbar zu tiefergehenden Erkenntnissen über den Gegenstand des Modellierten führe (vgl. McCarty 2004, S. 254–270, insbes. S. 258, sowie McCarty 2005, S. 23–72.).

Die Ergebnisse der quantitativen Auswertung sind durchaus kompatibel zu einer konventionellen literaturwissenschaftlichen Einschätzung: Die

männlichen Hauptfiguren haben erhöhte Werte bei ›Sich-Bewegen‹; die weiblichen Hauptfiguren haben mit Ausnahme von Isolde im ›Tristan‹ verhältnismäßig niedrige Bewegungswerte und wären daher eher als statische Figuren einzustufen; Zofen helfen und beraten oft. Männliche Hauptfiguren essen häufig, während Opponenten und weibliche Hauptfiguren viel seltener bei der Nahrungsaufnahme gezeigt wurden. Bei den Opponenten wird lediglich Marke aus dem ›Tristan‹ bei einem Beilager gezeigt. Bei der Gegenüberstellung von sprachlichen und nichtsprachlichen Aktivitäten ergab sich ein überraschender Befund: Dass weibliche Hauptfiguren eher mittels Sprache handeln, war erwartbar, nicht hingegen, dass es sich bei den Opponenten ebenso verhält.

2. Automatisierungsversuch

Während in dem Paidia-Beitrag die Aktivitäten manuell gezählt wurden, soll hier erprobt werden, inwieweit man mit einer automatischen Zählung der Aktivitäten zu ähnlichen oder anderen Ergebnissen gelangen kann.

Bei diesem Projekt handelt es sich um eine Vorstudie, in der erstens für ein kleines Testkorpus ein Procedere etabliert wird, bei dem Informationen auf Basis des Begriffssystems der Mittelhochdeutschen Begriffsdatenbank (MHDBDB) und Figurenreferenzen mithilfe der Annotationsumgebung CRETAnno in ausgewählten Textausschnitten annotiert werden. Zweitens wird ein Auswertungssystem entwickelt, das eine Plausibilitätsprüfung der Daten erlaubt. Drittens erfolgen eine Prüfung der Ergebnisse und ein Abgleich der Ergebnisdaten mit den Daten der Paidia-Studie. Diese zeichnet sich zwar gegenüber automatischen Verfahren durch die größere Verlässlichkeit einer manuellen Annotation aus, sie hat aber den Nachteil, dass sie nicht ohne enormen Aufwand auf größere Textmengen und auf weitere Kategorien hochskaliert werden kann. Bei dem Vergleich mit den Paidia-Daten muss allerdings in Kauf genommen werden, Äpfel mit Birnen zu vergleichen – schon deshalb, weil im Begriffssystem der MHDBDB keine

exakten Entsprechungen zu den Aktivitäten vorliegen, die in der Paidia-Studie untersucht wurden. Mit dieser Vorstudie wird auch das Anliegen verfolgt, Probleme und Chancen auszuloten, wenn künftig in größerem Umfang Begriffe aus der MHDBDB im Umkreis von Figurenreferenzen in einem größeren Korpus und bei weiteren Figurentypen untersucht werden.

Bei einem solchen automatischen Unternehmen stellt sich erstens die Problematik, Referenzen auf Figuren zu identifizieren, zweitens, im Text erwähnte Aktivitäten einer Figur zuzuordnen, und drittens, eine Aktivität automatisch zu erkennen und zu klassifizieren.

Bei unserem Versuchsaufbau und bei der zugrundeliegenden Modellbildung gehen wir davon aus, dass die lexikalische Zuordnung von MHDBDB-Kategorien zu Aktivitäten, die wiederum auf einer Verknüpfung der Kategorien mit Lemmata beruht, zumindest eine Annäherung an die gesuchten Aktivitäten realisieren kann. Selbstverständlich kann man nicht davon ausgehen, dass die jeweilige Aktivität – wie auch immer man sie in traditionellen literaturwissenschaftlichen Studien konzeptualisieren wollte – tatsächlich vollständig und exakt durch diesen lexikalischen Ansatz abgebildet werden kann.

3. Figurenreferenzen erfassen mit CRETAnno

Das Erkennen von Figuren fällt dem menschlichen Leser meist relativ leicht. Figuren können mit ihrem Eigennamen (z. B. ›Parzival‹) bezeichnet werden, mit einem Gattungsnamen (z. B. *der knappe*) oder mit Pronomina (›er‹ oder ›dieser‹). All diese Figurenreferenzen stehen stellvertretend für eine bestimmte Figureninstanz und werden dieser bei der Lektüre eines Textes durch einen menschlichen Leser ›einfach‹ zugeordnet – Unsicherheiten bei der Zuordnung gibt es in seltenen Fällen bei unscharfen Referenzen auf Figurengruppen oder bei komplexen Gemengelagen. Eine automatische Erkennung von Figurenreferenzen sieht sich im Gegensatz zur menschlichen Lektüre mit vielfältigen Herausforderungen konfrontiert.

Systematisch zerfällt die Erkennung von Figurenreferenzen in zwei Teilaufgaben: Erstens die sog. Named Entity Recognition (NER), die zur Aufgabe hat, Eigennamen in einem Text zu finden und diese einer bestimmten Kategorie (die gängigen Kategorien sind Person, Ort, Organisation) zuzuordnen. Zweitens die Koreferenzresolution, die sowohl die Erkennung von Figurenerwähnungen (etwa Gattungsnamen wie ›Frau‹) als auch ihre Auflösung auf die entsprechende Entität (z. B.: ›Frau‹ referiert auf die Instanz ›Angela Merkel‹) umfasst. Für die vollautomatische NER gibt es zwar eine Reihe etablierter Werkzeuge (bekannt sind etwa die für verschiedene Sprachen verfügbaren Stanford NER-Modelle), für das Mittelhochdeutsche fehlen solche Werkzeuge jedoch. Zusätzlich funktionieren sie primär für Daten, die den Trainingsdaten ähneln (i. d. R. Zeitungskorpora), wohingegen ihr Einsatz auf literarische Texte mit einer erheblichen Verschlechterung der Erkennungsquote einhergeht. Für das hier angestrebte Untersuchungsvorhaben lag es daher nahe, eine teilautomatische Erkennung der Figurenreferenzen mit Unterstützung der Annotationsumgebung CRETAnno zu verfolgen. Der zugrundeliegende Workflow wurde im Rahmen des Stuttgarter Zentrums für reflektierte Textanalyse (CRETA) entwickelt und erprobt.[1] Indem die Figurenreferenzen zunächst in einem kleinen Textabschnitt manuell annotiert werden und anschließend auf Basis dieser Annotationen ein Vorhersagemodell trainiert wird, wird das Verfahren auf sprachliche und textliche Spezifika zugeschnitten. Der teilautomatische Ansatz hat den Vorteil, dass er einerseits die Annotationsarbeit gegenüber einer rein manuellen Vorgehensweise beschleunigt, andererseits aber eine Qualitätskontrolle – nämlich eine Korrektur der automatischen Erzeugnisse über die Benutzeroberfläche – erlaubt. Der Workflow zur Annotation der Figurenreferenzen für die hier vorgelegte Studie wird im Abschnitt ›Workflow 2‹ genauer erläutert.

Nachdem die Figurenreferenzen erfasst sind, gilt es, die Aktivitäten den jeweiligen Figuren zuzuordnen. Das naheliegende Mittel ist eine Zuordnung über Kookkurrenzen – wenn also eine Aktivität im Umkreis einer

bestimmten Anzahl an Wörtern von einer Figurenreferenz erwähnt wird. Allerdings bringt dieses Vorgehen gewisse Unschärfen mit sich, beispielsweise bei Formulierungen, die den Bezug von Begriffen auf Figuren einschränken, etwa durch Negationen oder durch Konjunktiv II. Während zudem ein menschlicher Leser bei einer Formulierung wie ›Isolde belügt Marke‹ keine Probleme haben dürfte, würde durch das Kookkurrenz-Verfahren auch Marke mit der Aktivität ›Lügen‹ in Verbindung gebracht. Um eine Zuordnung einer Textpassage zu einer Figur mit größerer Sicherheit automatisch prozessieren zu können, wäre es nötig, zuvor die Kategorie ›Figurenbezug‹ manuell zu annotieren, was wiederum mit einem erheblichen Aufwand verbunden wäre (vgl. Dimpel 2017, S. 94).

4. Wortschatzbasierte Begriffe der MHDBDB als Aktivitäten

Bei wortschatzbasierten Zugriffen auf tatsächliche Phänomene trifft man ebenfalls auf Probleme – etwa bei ambigem Vokabular. Doch bis bessere Optionen realisierbar sind, scheint mit einem wortschatzbasierten Vorgehen immerhin eine gewisse Annäherung an die gesuchten Phänomene möglich zu sein.

Das Begriffssystem der MHDBDB geht zurück auf den Entwurf eines »Begriffssystems als Grundlage für die Lexikographie« Rudolf Halligs und Walther von Wartburgs und ergänzt diesen um Begriffe, die eine genauere Erfassung in Bezug auf die mittelalterliche Lebenswelt ermöglichen (vgl. Hallig/Wartburg 1963). Damit reiht es sich ein in die onomasiologischen Thesaurus-Ansätze von Roget 1864 bis hin zum ›Historical Thesaurus of the OED‹. Allen gemein ist, dass eine hierarchische Gliederung von Begriffen oder Kategorien dazu dient, den Wortschatz inhaltsseitig zu erschließen. In der MHDBDB werden die Bedeutungen eines Lemmas mittels Zusammensetzung aus Begriffen ausgedrückt. Diese Bedeutungen können von den Tokens der Korpustexte referenziert werden, um die Bedeutung eines Wortes im Kontext zu kennzeichnen. Somit werden etwa auch Wort-

schatzanalysen ermöglicht, indem alle über Bedeutungspositionen mit einem Begriff verbundenen Wortartikel angezeigt werden können.

5. Aktivitäten und Begriffe – Tentatives Mapping

Den Aktivitäten aus dem Paidia-Beitrag wurden Begriffe aus der MHDBDB zugeordnet. Teilweise war es notwendig, diesen Aktivitäten mehrere Begriffe zuzuweisen, so dass insgesamt 62 Begriffe berücksichtigt werden.

Zu den Aktivitäten ›Suchen‹, ›Listhandeln‹, ›Verstecken‹, ›Warten‹, ›Gefangennahme‹, ›Erpressen‹, ›Stehlen‹, ›Zuhören‹, ›Unterweisen‹, ›Beraten‹, ›Drohen‹, ›Streiten‹, ›Verurteilen‹ und ›Abweisen‹ wurden keine plausiblen Entsprechungen im Begriffssystem gefunden. Bei den sprachlichen Aktivitäten konnte damit leider nur bei einem kleinen Teil ein Mapping zum Begriffssystem der MHDBDB erstellt werden. Da angestrebt wurde, jeden Begriff nur einer Aktivität zuzuordnen, wurde etwa darauf verzichtet, 22826000 (›Willensausübung auf andere‹) und 22642000 (›Wahrheit/Irrtum/Lüge/Täuschung ‹) auch dem nichtsprachlichen ›Listhandeln‹ zuzuordnen. Ein Blick in die Daten ergab, dass diese Begriffe eher zu den sprachlichen Aktivitäten ›Veranlassen‹ bzw. ›Lügen‹ passen.

a_Essen	Mahlzeiten	21111200
a_Essen	Ernährung	21111000
a_Essen	Gerichte/Speisen	21111306
a_Essen	Nahrungsmittel/Lebensmittel	21111300
a_Feiern	Festlichkeiten/Feiern	23135100
a_Feiern	Feste/Spiel/Unterhaltung	23135000
a_Feiern	Spiel/Zerstreuung	23135200
a_Feiern	Spielzeug	23135220
a_Feiern	Schachspiel	23135210
a_Feiern	Andere Spiele	23135230
a_Feiern	Tanz	25132000
a_Feiern	Tanz/Namen	25132100
a_Feiern	Instrumentalmusik	25153000
a_Feiern	Instrumentalmusik/Namen	25153100
a_Feiern	Gesang	25152000
a_Feiern	Gesang/Namen	25152100
a_Feiern	Musik/Allgemeines	25150000
a_Schlafen	Schlaf/Ausruhen	21080000
a_Helfen	Hilfeleistung/Dienst/Widerstand	23133000

a_Hilfsmittel	Voraussetzungen zur Aktion - Mittel	22831500
a_Hilfsmittel	Objektbezogene Aktivität/Tätigkeit	21072000
a_Hilfsmittel	Werken/Werkzeuge/Utensilien/Ausrüstung	23304100
a_Jagen	Jagd/Fischerei	23302000
a_Jagen	Jagd/Fischerei/Waffen/Geräte	23302200
a_Jagen	Jagd/Fischerei/Waffen/Geräte/Namen	23302210
a_Jagen	Jagd/Fischerei/Gebräuche/Rituale	23302300
a_Jagen	Jagdtiere/Namen	23302400
a_Jagen	Fischzucht	23302500
a_Jagen	Raubvögel	14024000
a_Kämpfen	Kriegswesen/Kampf/Gewalt	23240000
a_Kämpfen	Kriegswesen/Kampf/Namen	23241000
a_Kämpfen	Heerfahrt	23242000
a_Kämpfen	Schlacht	23243000
a_Kämpfen	Aventiure	23244000
a_Kämpfen	Belagerung	23245000
a_Kämpfen	Zweikampf	23246000
a_Kämpfen	Rangordnung	23247000
a_Kämpfen	Ritterwesen/Allgemeines	23200000
a_Kämpfen	Waffen	23231000
a_Kämpfen	Waffen/Namen	23231100
a_Kämpfen	Kampf-/Waffentechnik	23231200
a_Kopulieren	Sexualleben/Erotik	21112000
a_Kopulieren	Geschlechtsmerkmale/Genitalien/Zeugung/ Geburt	21045000
a_Bewegen	Bewegung	31800000
a_Turnieren	Turnier/Organisation/Aufzug	23261000
a_Turnieren	Turnier/Schauplatz	23262000
a_Turnieren	Turnier/Kampfformen	23263000
a_Turnieren	Turnier/Rituale	23264000
a_Turnieren	Turnierwesen	23260000
a_Töten	Tod	21104000
a_Peinigen	Ärger/Zorn/Wut/Aggression	22705200
b_Reden	Mündliche Kommunikation	23121000
b_Reden	Eigenschaften/Mängel der stimml. Äußerung	23121100
b_Reden	Stimmlicher Ausdruck	23121200
b_Reden	Kommunikation und Sprache	23120000
b_Argumentieren	Beweis/Zustimmung/Widerspruch	22641000
b_Veranlassen	Autorität/Anordnung/Befehl	22826100
b_Veranlassen	Willensausübung auf andere	22826000
b_Beklagen	Gefallen/Missfallen	22702010
b_Beklagen	Glückseligkeit/Unglück	22702020
b_Beklagen	Freude/Leid	22702030
b_Lügen	Wahrheit/Irrtum/Lüge/Täuschung	22642000

Tabelle 1: Mapping Aktivitäten-Begriffe. Ein vorgestelltes »a_« zeigt nichtsprachliche, ein vorgestelltes »b_« zeigt sprachliche Aktivitäten an.

6. Figurenstudien in der digitalen Literaturanalyse

Es gibt eine Reihe von Studien zu Figuren in den DH, die jedoch nur geringes Anknüpfungspotential zum hier vorgelegten Vorhaben bieten. Die von Peer Trilcke und Frank Fischer vorgelegten Analysen zu Dramen bedienen sich der Methode der Sozialen Netzwerkanalyse, um u. a. offene und geschlossene Dramen miteinander zu vergleichen (vgl. grundlegend: Trilcke 2013, S. 201–247, sowie das Forschungsprojekt dlina, Digital Literary Network Analysis). Damit basieren die Untersuchungen zwar auch auf einer Erfassung von Figurenvorkommen, diese bleibt aber auf Eigennamen beschränkt, die im Drama aus dem Nebentext extrahiert werden können. Auch die Forschungen des QuaDramA-Projekts (Quantitative Drama Analytics) beschäftigen sich mit Dramentexten, widmen sich hierbei aber auch Problemen der Koreferenz und ihrer Resolution (vgl. z. B. Pagel/Reiter 2020; Wiedmer [u. a.] 2020). Figurenbezogene Studien zu Erzähltexten wurden etwa hinsichtlich ihrer sozialen Interaktion (vgl. etwa: Elson [u. a.] 2010; Agarwal/Rambow 2010; Agarwal [u. a.] 2012; Dimpel 2020) oder ihres Raum-Bezugs (Viehhauser/Barth 2017) durchgeführt; spannend ist die Studie zu Figurentypen im modernen Drama von Krautter ([u. a.] 2020). Für mittelhochdeutsche Erzähltexte liegen erst wenige Studien vor; diese sind im Rahmen des DH-Zentrums CRETA entstanden und beschäftigen sich insbesondere mit sozialen Relationen (vgl. Braun/Ketschik 2019; Ketschik [u. a.] 2020, hier insb. Kap. 3.2). Der Zusammenhang von Figuren und Aktivitäten wurde bisher nicht in den Blick genommen. Gegenüber Topic Modeling-Ansätzen bietet das hier vorgestellte Verfahren den Vorteil, dass es gezielt ausgewählte Begriffe (hier: Aktivitäten) und ihre Kookkurrenz mit ausgewählten Figureninstanzen untersucht. Über ein Topic Modeling-Verfahren werden hingegen dominierende Themenfelder für einen bestimmten Textabschnitt extrahiert. Zusätzlich bleibt hier – im Gegensatz zu vollautomatischen Verfahren – die Transparenz gewahrt, wie die

ausgewählten Begriffe zustande kommen, was eine kritische Reflexion der resultierenden Ergebnisse ermöglicht.

7. Workflow 1: sql-Abfrageergebnis nach xml konvertieren

In einem ersten Schritt wurden Daten zu den Texten und den gewählten Begriffen aus der MHDBDB erfasst. Exportiert worden sind die Begriffsrelationen und das POS-Tagging der Tokens, deren Bedeutungspositionen mindestens einen der zuvor ausgewählten Begriffe umfasst. Für Gottfrieds ›Tristan‹ ergab die Abfrage beispielsweise:

1,0,65537,6748,"gedaehte",32768,1992,"gedenken","TR",21072000.

Diese Daten wurden wieder in Texte konvertiert. Dazu wurden Perl-Programme entwickelt, die die Zeilennummer in ein xml-Tag umsetzen. Danach folgen die Wortformen des Textes und, falls vorhanden, das Kategorien-Tag <kat>, das kommasepariert die Kategorien bzw. Begriffe zur vorstehenden Wortform auflistet. ›Tristan‹, V. 9, erhielt damit folgende Form:

<l>9</l> ich hoere es velschen <kat>21072000,22642000,23120000, 23304100</kat> harte vil.

Bereits dieser Schritt, der zunächst trivial anmutet, war mit einigem Aufwand verbunden, da die ebenfalls in Perl realisierte Konsistenzprüfung, die auch einen Abgleich mit vorhandenen Texten umfasste, eine Reihe an Zuordnungsproblemen und Abweichungen ergab, die teilweise aus verschiedenen Textgrundlagen resultierten und teilweise aus einer abweichenden Nummerierung.[2] Die mit den Kategorien-Tags versehenen Texte wurden sodann in CRETAnno eingespeist.

8. Workflow 2: Erkennung von Figurenreferenzen

Das Annotationstool CRETAnno war ursprünglich für eine vollständige Annotation bestimmter Entitätenklassen (etwa Personen, aber auch Orte oder Organisationen) gemäß den zugrundeliegenden Richtlinien (online) entwickelt. Für die neue Studie, die lediglich einige zuvor festgelegte Figuren in den Texten untersuchen möchte, wurde die Benutzeroberfläche so angepasst, dass das Korrekturmenü nur die Liste mit den zu untersuchenden Figuren enthielt.[3] Um den halbautomatischen Annotationsprozess nicht bei null beginnen zu müssen, wurde ein bereits entwickeltes Vorhersagemodell auf die neuen Daten angewandt. In einer ersten Sichtprüfung zeigte sich, dass das System, das an anderen Texten trainiert worden war, die auf diesen Daten erzielte Erkennungsquote (Recall von bis zu 76%, Precision von bis zu 91%; Blessing [u. a.] 2017) auf den neuen Texten nicht erreichen konnte. Zusätzlich galt es, das Modell für den neuen Anwendungsfall, nämlich die Beschränkung auf bestimmte Figuren, anzupassen. Deshalb wurden für den ›Tristan‹ und den ›Partonopier‹ jeweils ca. 1000 Verse manuell vorannotiert,[4] um das Erkennungsmodell neu zu trainieren. Der ›Iwein‹ lag in Stuttgart bereits annotiert vor, für den kurzen ›Mauritius‹ hätte sich der Aufwand dafür nicht amortisiert.

Nach diesem Schritt wurden die automatisch erzeugten Annotationsvorschläge vollständig kontrolliert und nachannotiert – für diese Arbeit danken wir Anne Faust (WHK, Darmstadt). Zu unklaren Textbefunden wurde eine Protokolldatei angelegt, die durch Friedrich Michael Dimpel überprüft wurde.[5]

Zusätzlich zu Figurenreferenzen über Eigennamen (etwa ›Isolde‹) oder Appellative (wie *mîn muoter* oder *der getriuwe*) wurden in bestimmten Fällen pronominale Referenzen auf eine Figur erfasst – und zwar dann, wenn eine Referenz auf diese Figur in den zehn Versen zuvor nicht bereits annotiert worden ist. Dies war notwendig, um auch Kategorien, die nur im

Kontext pronominaler Figurenreferenzen vorkommen, der entsprechenden Figur zuordnen zu können.

Nach Abschluss der Annotation wurden die Texte aus CRETAnno exportiert – einschließlich der xml-Informationen zu den Kategorien und zu den Figurenreferenzen. Mit einem Perl-Skript wurde die Konsistenz der Daten überprüft. Für Hartmanns ›Iwein‹, der im Rahmen der CRETA-Arbeiten bereits mit Figurenreferenzen annotiert wurde, wurden die <kat>-Tags integriert sowie abschließend noch einige pronominale Referenzen nachannotiert.

9. Workflow 3: Bereinigung der Begriffskategorien aus der MHDBDB

Bei ersten Auswertungsversuchen fiel auf, dass insgesamt überraschend viele Wortformen mit Begriffen annotiert wurden. Ein Blick in die Daten zeigt, dass die Lemma-Begriff-Zuordnung in der MHDBDB eher großzügig vorgenommen wurde, da die verwendeten Korpustexte noch nicht bedeutungsdisambiguiert sind. Das Lemma *ritter* ist beispielsweise in einer der drei differenzierten Bedeutungen dem Begriff 21111306 ›Gerichte/Speisen‹ zugeordnet. Diese Zuordnung hat durchaus ihre Berechtigung, listet doch der große Lexer unter *ritter* auch ›*arme ritter, ein backwerk*‹ als Bedeutungsangabe (Lexer 1992, Bd. 2, Sp. 466). Die Zuordnung ist für den primären Zweck der MHDBDB sinnvoll: Viele Anwender verwenden die MHDBDB, um herauszufinden, welche Begriffe oder Konzepte in welchen Texten anzutreffen sind, um nach diesem Rechercheschritt manuell die Fundstellen auszuwerten, die dann wiederum etwa in eine Textanalyse einfließen können. Da die Texte der Begriffsdatenbank zwar lemmatisiert, aber bislang nicht vollständig nach verschiedenen Bedeutungspositionen disambiguiert worden sind, müssen bei nicht bedeutungsannotierten Texten zunächst ›falsche‹ Begriffe herausgefiltert werden.

Für die vorliegende Analyse ist es jedoch durchaus relevant, dass die Begriffe, die im Umkreis von Figurenreferenzen gefunden werden, nicht nur im Einzelfall tatsächlich für den gesuchten Begriff einschlägig sind, sondern idealerweise ausschließlich, zumindest aber möglichst überwiegend: Ritter werden im Artusroman eher selten verspeist, auch wenn es in ›Dietrichs Flucht‹ um Ortnit anders bestellt sein mag. Das Konzept 21111000 ›Ernährung‹ ist auch Bedeutungspositionen der Lemmata *starken*, *twingen, vaste* zugeordnet, die in diesen Texten ebenfalls eher selten im Ernährungskontext verwendet werden.

Um die Begriff-Zuordnung um eher periphere Bedeutungen bereinigen zu können, wurden 2020 aus der MHDBDB alle hier untersuchten Begriffe, die einer Wortform in den vier hier untersuchten Texten zugeordnet sind, exportiert.[6] Ein Perl-Skript wurde entwickelt, das nach Wortform-Begriff-Zuordnungen sucht, die in den vier hier untersuchten Texten fünf Mal oder häufiger vorkommen. Wenn diese Zuordnung auf einer Wortbedeutung beruht, die im Korpus als eher selten eingestuft wurde, wurde die Zuordnung des zugehörigen Lemmas[7] in eine Löschliste aufgenommen – also dann, wenn ein deutlicher Verdacht bestand, dass das Lemma im Korpus überwiegend nicht in dieser Bedeutung vorkommt. Den Begriffen sind auch Lemmata zugeordnet, die nur ex negativo zur Kategorie passen. Auch solche Zuordnungen wurden bereinigt, beispielsweise *swîgen* bei ›Stimmlicher Ausdruck‹. Zwar hat *swîgen* durchaus mit einem stimmlichen Ausdruck zu tun, allerdings ist der Begriff 23121200 der Aktivität ›Reden‹ zugeordnet, so dass ein Vergleich mit den Paidia-Daten ins Leere laufen würde, würde man *swîgen* beibehalten.

Die Löschliste wurde anschließend erstens nach Lemmata und zweitens nach Begriffen sortiert und in beiden Fassungen jeweils manuell auf eine konsistente Entscheidungsstrategie hin überprüft. Es ist wohl unnötig zu betonen, dass trotzdem kaum objektivierbare Kriterien für diesen Vorgang, bei dem der Zeitfaktor doch auch eine kritische Größe war, festgelegt wer-

den konnten – ähnlich wie bei der älteren Editionsphilologie basiert die Entscheidung auf dem Iudicium des Anwenders.

Ein weiteres Skript hat die Zuordnungen, die in der Löschliste notiert sind, nach den jeweiligen Wortformen in den Texten eliminiert. Zur Wortform-Lemma-Zuordnung wurde dabei eine Liste verwendet, die aus dem Export aus der MHDBDB von 2020 erzeugt werden konnte. Im Rahmen dieser Bereinigung wurde bei 2.514 Lemmata eine manuelle Entscheidung getroffen. Davon wurden bei 1.569 Lemma-Begriff-Zuordnungen Inkonsistenzen bereinigt.[8] Beim Abarbeiten der Löschliste wurden zudem auch solche Begriff-Lemma-Zuordnungen gelöscht, die 2018 noch in der MHDBDB eingetragen waren, die aber 2020 in der MHDBDB nicht mehr vorhanden waren. Bislang nicht zugeordnete Lemmata wurden auch im Rahmen dieser Bereinigung nicht zugeordnet.

Schwierig bleibt vor allem die Erfassung von Begriffen, die in den Texten meist nicht direkt lexikalisch abgebildet werden, sondern die oft metaphorisch oder verhüllend umschrieben werden, etwa beim Begriff ›Sexualleben/Erotik‹ (21112000). Mit einigen Bedenken wurden u. a. die Verben *umbevâhen*, *lieben*, *triuten* und *minnen* beibehalten, auf die Löschliste wurden u. a. übernommen: *amîs*, *amor*, *brennen*, *gebern*, *gelieben*, *geluste*, *herzeliep*, *kiusche*, *küssen*, *lieplich*, *maget*, *minneclîch*, *munt*, *reinegen*, *sanfte*, *senender*, *süeze*, *tougen*, *tragen*, *triuwe*, *trôst*, *trûtgeselle*, *vriunt*, *vrouwe*.

10. Das Scoring-System zur Auswertung

In dem Paidia-Aufsatz ging es darum, festzustellen, ob eine Aktivität in einem bestimmten Textsegment überhaupt vorlag oder nicht. Anschließend wurde betrachtet, welche Aktivitäten in wie vielen Segmenten vorhanden waren; mehrfache Okkurrenzen im gleichen Segment wurden nicht berücksichtigt. Ein Vorteil dieses Verfahrens war, dass die Werte zu verschiedenen Aktivitäten eher vergleichbar sind. Erwartbar ist etwa, dass

›Reden‹ deutlich häufiger vorkommt als ›Töten‹, die Unterschiede konnten durch den Segmentbezug jedoch zueinander in einem aussagekräftigen Verhältnis verglichen werden.

Anders als in der Paidia-Studie wurde hier versucht, auch quantitativ zu berücksichtigen, wie häufig welche Begriffe im Figurenkontext vorkommen. Bei der Auswertung sind mehrere Problemfelder vorhanden: Einerseits ist die Häufigkeit der Aktivitäten/Kategorien sehr unterschiedlich verteilt, andererseits sind die Texte unterschiedlich lang. Außerdem muss überlegt werden, bei welchem Abstand eine Kategorie noch dem Figurenkontext zugerechnet wird und ab welchem Abstand nicht mehr.

Für die Zuordnung zum Figurenkontext wurde ein ternäres Abstands-Scoring-System etabliert. Bei einem geringen Abstand wurde ein Score von 1,3 festgelegt, bei einem mittleren Abstand ein Score von 1,0 und bei einem größeren Abstand ein Score von 0,66. Die Scores für mittlere und große Abstände wurden v. a. für die Reihenfolge ›Figurenreferenz‹ > ›Kategorie‹ angedacht; für die umgekehrte Reihenfolge ›Kategorie‹ > ›Figurenreferenz‹ wurde ein zusätzlicher ›Rückwärts‹-Grenzwert eingeführt, der dafür sorgt, dass Kategorien nur dann einer Figur zugeordnet werden, wenn die Figur kurz nach der Kategorie genannt wird, weil davon ausgegangen wurde, dass die Zuordnung einer Kategorie zu einer Figur lange vor deren Nennung größere Anforderungen an das Hörverstehen stellt und daher wohl eher die Ausnahme ist, weil der Rezipient hier rätseln muss, auf wen sich eine Aussage bezieht.

Geeignete Parameter für einen geringen, mittleren und großen Abstand wurden in einem Abgleich mit den Ergebnisdaten aus dem Paidia-Aufsatz ermittelt, der unten beschrieben wird. Überraschenderweise haben sich ausgesprochen niedrige Parameter als geeignet erwiesen: fünf, zehn und 15 Worte Abstand; als ›Rückwärts‹-Grenzwert für Kategorien vor der Figurenreferenz wurde ein Abstand von fünf Worten verwendet. Wenn eine Kategorie innerhalb dieser Abstände im Umkreis einer Figurenreferenz auftritt, wird dies im Weiteren kurz als Kookkurrenz bezeichnet.

Jede gefundene Kookkurrenz wird zunächst mit dem jeweiligen Abstandsparameter multipliziert, also maßvoll schwächer oder stärker gewichtet: Wenn auf eine Figurenreferenz eine Kategorie im Abstand von bis zu fünf Wörtern folgt, wird diese Kookkurrenz mit Faktor 1,3 ungefähr doppelt so stark gewichtet wie bei einer Kategorie, die erst nach bis zu 15 Wörtern folgt.

Zusätzlich zu diesem Abstands-Scoring-System wird noch ein System benötigt, das die Anzahl der Kookkurrenzen berücksichtigt. Wenn man jedoch lediglich diese gewichteten Kookkurrenzen anschließend je Figur und Kategorie aufaddiert, erhält man ein stark textlängenabhängiges Maß. Ein Vergleich der Anzahl der Kookkurrenzen zwischen ›Mauritius‹ (1.784 Verse) und ›Partonopier‹ (21.784 Verse) wäre auf diesem Weg sinnlos. Wenn man andererseits diesen Wert durch die Anzahl der Wörter im jeweiligen Text oder auch durch die Anzahl der Figurenreferenzen im jeweiligen Text bzw. die Referenzen der jeweiligen Figur im jeweiligen Text teilt, erhält man Werte, die die kurzen Texte und seltene Figuren ganz erheblich begünstigen. Auch diese Quotienten würden also längenabhängig bleiben.

Ebenfalls experimentell erprobt und verworfen wurde eine Z-Wert-Normalisierung,[9] bei der für jede Figur der Mittelwert zu allen gewichteten Kookkurrenzen und für alle Kategorien und auf dieser Grundlage Standardabweichung und Z-Werte berechnet wurden. Grundlage für diesen Weg war die Überlegung, dass hier eine Aussage möglich sein könnte, ob innerhalb der Werte zu den Kategorien bei den jeweils gleichen Figuren einer Kategorie erhöhte oder niedrigere Werte erreicht werden. Auch hier haben sich keine überzeugenden Daten ergeben: Im ›Mauritius‹ etwa erreicht der Ehemann-Graf bei Begriff 22642000 ›Wahrheit/Irrtum‹ auf der Grundlage von nur einer Kookkurrenz bei 15 Figurenreferenzen einen höheren Wert als die Zofe, die auf zehn Kookkurrenzen bei 29 Figurenreferenzen kommt. Vermutlich stellen dabei Nullwerte bei den Kookkurrenzen einen Störfaktor dar: Wenn eine Nebenfigur viele Nullwerte aufweist, sind dort, wo zumindest nur wenige Kookkurrenzen überhaupt vor-

handen sind, die Z-Werte recht hoch und wohl auch höher als bei häufigen Figuren, bei denen bei mehr Kategorien Kookkurrenzen überhaupt gezählt werden können. Das führt dazu, dass niedrige Distanzwerte bei Nebenfiguren zu höheren Z-Werten führen können als bei den Hauptfiguren in der gleichen Kategorie.

Bei der Anzahl der Kookkurrenzen ist also ein Scoring-System notwendig, das sowohl die Problematik der unterschiedlichen Textlänge als auch der unterschiedlichen Anzahl der Figurentags (also Haupt- versus Nebenfigur) berücksichtigt. Gleichzeitig sollen aber auch einzelne Kookkurrenzen, die trotz ihrer geringen Häufigkeit recht signifikant sein können (etwa ›Tod‹ oder ›Sexualleben/Erotik‹), statistisch nicht allzu stark nivelliert werden. Das Kookkurrenzen-Anzahl-Scoring-System erhält in sieben Stufen jeweils Scores von 5–15; weiterhin bleibt der Score bei null Punkten, wenn keine einschlägige Kookkurrenz gefunden wurde. Eine einzelne Kookkurrenz erhält die niedrigste Score-Stufe von fünf Punkten; wenn zwei oder drei Kookkurrenzen vorliegen, werden zunächst sieben Punkte vergeben. Für alle häufigeren Kookkurrenzen werden Scores vergeben, die davon abhängen, ob die relative Häufigkeit der Kookkurrenzen zu einer Figur und zu einer Kategorie eher als höher oder eher als niedriger betrachtet werden kann. Um zu ermitteln, ob die relative Häufigkeit eher als höher oder als niedriger betrachtet werden kann, werden zwei Vergleichswerte gebildet: einerseits ein normalisierter Kategorien-Mittelwert MW-K(Kat), der nicht figurenspezifisch ist, und andererseits ein figurenspezifischer Wert zur jeweiligen Kategorie F-K(Kat).

$$\mathbf{MW\text{-}K(Kat)} = \frac{\mathrm{AK(Kat)}}{\mathrm{AF}}$$ > normalisierter Kategorien-Mittelwert, nicht figurenspezifisch

AK(Kat): Anzahl aller Kookkurrenzen zur jeweiligen Kategorie in allen Texten zu allen Figuren
AF: Anzahl aller Figurenreferenzen in allen Texten

$$\mathbf{F\text{-}K(Kat, Fig)} = \frac{\mathrm{AK(Kat,Fig)}}{\mathrm{AF(Fig)}}$$ > figurenspezifischer Wert zur jeweiligen Kategorie

AK(Kat,Fig): Anzahl der Kookkurrenzen je einer Figur zu je einer Kategorie
AF(Fig): Anzahl der Figurenreferenzen zu dieser Figur

Formel 1: MW-K(Kat) und F-K(Kat)

Der Wert MW-K(Kat) wird für alle Kategorien separat gebildet: Die Anzahl aller Kookkurrenzen zur jeweiligen Kategorie in allen Texten zu allen Figuren wird geteilt durch die Anzahl aller Figurenreferenzen in allen Texten. Dieser Wert besagt, wie häufig bei einer Kategorie Kookkurrenzen im Durchschnitt bei allen Texten in Relation zu allen Figurenreferenzen vorliegen. Es handelt sich also um einen Kategorien-Mittelwert, der jedoch nicht in Bezug auf die Textlänge, sondern in Bezug auf die Gesamtzahl der Figurenreferenzen normalisiert wird.

Gebildet wird weiterhin der Wert F-K(Kat). Dieser Wert wird für alle Kategorien und alle Figuren separat gebildet. Er wird berechnet aus der Anzahl der Kookkurrenzen je einer Figur zu je einer Kategorie, die geteilt wird durch die Anzahl der Figurenreferenzen zu dieser Figur. Dieser Wert besagt, wie häufig bei einer Kategorie Kookkurrenzen zu einer Figur in Relation zu der Häufigkeit der Figurenreferenzen dieser Figur im jeweiligen Text vorliegen.

Berücksichtigt werden auf diesem Weg auch relativ höhere oder relativ niedrige Werte, unabhängig davon, ob es sich um eine häufige Kategorie wie »Objektbezogene Aktivität« (hoher Wert MW-K(Kat): 290,7) oder um eine Kategorie handelt, die in allen Texten eher selten ist wie etwa ›Tod‹ (niedriger Wert MW-K(Kat): 68,7). Damit es zu einem Score von zehn Punkten kommt, muss der figurenspezifische Wert F-K(Kat) mindestens den allgemeinen kategorienspezifischen Mittelwert erreichen. Über die Anzahl der Figurenreferenzen wird weiterhin berücksichtigt, ob eine Figur häufig oder selten vorkommt. Verwendet werden folgende Scores:

Score=5, wenn eine Kookkurrenz vorliegt (zur jeweiligen Figur und Kategorie)
Score=7, wenn 2-3 Kookkurrenzen vorliegen
Score=9, wenn F-K(Kat) > MW-K(Kat) x 0,9
Score=10, wenn F-K(Kat) >= MW-K(Kat)
Score=11, wenn F-K(Kat) > MW-K(Kat) x 1,1
Score=13, wenn F-K(Kat) > MW-K(Kat) x 1,3
Score=15, wenn F-K(Kat) > MW-K(Kat) x 1,66

Gewählt wird jeweils der höchste Score: Wenn eine Figur bspw. nur drei Kookkurrenzen aufweist, aber nur sehr selten vorkommt, kann bei einer seltenen Kategorie der relative Figurenwert F-K(Kat) dennoch über dem allgemeinen Mittelwert MW-K(Kat) liegen, so dass ein höherer Score als sieben Punkte möglich ist.

Abschließend wird der Anzahl-Score mit dem Abstands-Score kombiniert: Zu den einzelnen Abstand-Scores, die zu niedrigen Abständen (fünf Worte), mittleren (zehn) oder größeren Abständen (15) vergeben wurden, wird pro Figur und Kategorie ein Abstandsmittelwert gebildet, der zwischen 0,66 und 1,33 liegen kann. Mit diesem Abstands-Score wird der Anzahl-Score (Werte von 0–15) multipliziert, so dass der endgültige Score in einem Bereich von 0–19,95 liegen kann.

Ein Beispiel: Der allgemeine kategorienspezifische Mittelwert für die Kategorie ›Tod‹ MW-K(›Tod‹) liegt für alle Figuren und alle Texte bei 68,7. Bei Lunete wurden elf Kookkurrenzen zu ›Tod‹ erfasst. Der figurenspezifische Wert F-K(›Tod‹) für Lunete liegt bei 71,9. Dieser Wert ist größer als MW-K(›Tod‹), jedoch nicht größer als MW-K(›Tod‹) x 1,1. Der Score beträgt also zehn Punkte – noch vor der Berücksichtigung der Abstand-Scores. Die Abstand-Scores betragen für jede der elf Kookkurrenzen entweder 0,66, 1,0 oder 1,3; aus diesen elf Abstand-Scores wird ein Mittelwert gebildet. Mit diesem Abstand-Score-Mittelwert (hier 0,87) wird der Kookkurrenzen-Anzahl-Score (hier zehn Punkte) multipliziert, so dass sich als kombinierter Score für Lunete und ›Tod‹ 8,7 Punkte ergeben.

Dieses Scoring-Verfahren, das auf der Anzahl der Kookkurrenzen beruht, stellt sicher, dass die Scores eher aussagekräftig und weniger textlängenabhängig sind; sie liegen zudem in einer vergleichbaren Größenordnung.

Während bei häufigen Figuren eher zu erwarten ist, dass sie im Kontext einer breiten Palette von Kategorien gezeigt werden, ist zu erwarten, dass selten vorkommende Figuren im Kontext einer weniger breiten Kategorien-Palette gezeigt werden; mitunter können an seltene Figuren spezifische

narrative Funktionen gebunden sein. Da in die Gewichtung auch die Häufigkeit der Referenzen auf die jeweilige Figur eingeht, ist das Scoring-Verfahren nicht vollständig längenunabhängig: Es weist eine leichte Tendenz dazu auf, seltenen Figuren bei den vorhandenen Kategorien höhere Werte zuzuerkennen als häufig vorkommenden Figuren. Diese Plausibilitätsüberlegung stimmt damit überein, dass etwa der Graf und die Gräfin im ›Mauritius‹ mitunter etwas höhere Scores aufweisen als vergleichbare Figuren in den Romanen.

11. Auswertung: Einzelne Kategorien und Figuren

Folgende Figuren wurden untersucht:

Text	**Zofe**	**Hauptfigur ♂**	**Hauptfigur ♀**	**Opponenten**
›Partonopier‹	Irekel	Partonopier	Meliur	Mutter, Sornagiur, Mareis
›Iwein‹	Lunete	Iwein	Laudine	Ascalon, Harpin, Truchsess
›Tristan‹	Brangaene	Tristan	Isolde	Marke, Irischer Truchsess, Marjodo
›Mauritius v.C.‹	Zofe-MvC	Mauritius	Graefin-MvC	Graf-MvC

Tabelle 2: Texte und Figuren

Die ausführlichen Ergebnis-Daten sind im Dariah-Repository unter doi:10.20375/0000-000F-322E-6 zu finden. Dort stehen zwei Tabellen: Erstens eine Tabelle für alle Figuren, zweitens eine Tabelle, die eine nach Aktanten gruppierte Ausgabe bietet.

Die Scores pro Kategorie und Figur sind dann mit einem (+) bzw. (-) markiert, wenn der Figurenscore bei der jeweiligen Kategorie oberhalb oder unterhalb des Konfidenzintervalls (bei einem Konfidenzniveau von 80%) liegt – hervorgehoben werden also bereits mäßig über- oder unter-

durchschnittliche Werte. Auf Spitzenwerte kommt es hier zunächst nicht an; im Paidia-Aufsatz ging es nur darum, ob eine Aktivität in einem Segment überhaupt vorhanden war oder nicht. Bei Aktivitäten, die aus mehreren Kategorien bestehen, wird hier angegeben, bei wie vielen dieser einzelnen Kategorien sich solche erhöhten Werte für eine Figur ergeben – beispielsweise also, ob bei fünf von zehn Kategorien, die zu einer Aktivität gehören, bei einer Figur Werte oberhalb des Konfidenzintervalls vorhanden sind. Bei diesem Analyseschritt geht jedoch nicht ein, ob erhöhte Werte für eine Figur nur marginal oder ganz erheblich über dem Konfidenzintervall liegen.

Ein zweiter Analyseschritt bezieht die Höhe der Scores mit ein. Wenn eine Aktivität aus mehreren Kategorien besteht, wird der Mittelwert der Scores zu all diesen Kategorien gebildet. Dieses Verfahren ist nicht ganz unproblematisch: Wenn etwa aus den Mittelwerten der zwölf Kategorien, die der Aktivität ›Kämpfen‹ zugeordnet sind, wiederum ein Mittelwert gebildet wird, könnte damit theoretisch nivelliert werden, wenn eine Figur zwar ein enormes Kampfpensum absolviert, das sich etwa bei ›Zweikampf‹ dokumentiert, diesen hohen Wert aber durch niedrige Werte etwa bei ›Rangordnung‹ einbüßt. In der Praxis haben sich jedoch meist plausible Werte ergeben, die mehr oder weniger ähnliche Tendenzen dokumentieren wie die Betrachtung der Anzahl der Überschreitungen des Konfidenzintervalls.

Aktivität	alle[10]	Zofen	Hauptfigur ♂	Hauptfigur ♀	Opponenten
Essen	5,7	**6,4**	**10,2**	4,7	4,1
Feiern	2,3	1,5	**4,3**	**3,2**	1,5
Jagen	3,1	1,5	**5,6**	**3,8**	2,5
Schlafen	9,1	**10,2**	**10,1**	**10,2**	7,8
Helfen	11,4	**13,6**	9,9	**11,4**	11,0

Tabelle 3: Höhe der Scores je Aktantengruppe, Teil 1 (Summenwerte aller Kategorien, die einer Aktivität zugeordnet sind)

Essen: Der Aktivität ›Essen‹ sind vier Kategorien zugeordnet. Bei jeder Figur können sich also bis zu vier Werte ergeben, die oberhalb des Konfidenzintervalls liegen. Bei den folgenden Figuren ist mehr als eine Kategorie mit solcherart erhöhtem Wert vorhanden: Irekel (2), Partonopier (3), Meliur (2), Partonopiers Mutter (2), Mareis (2), Lunete (2), Iwein (3), Harpin (3), Brangäne (2), Tristan (3), Marke (2), Mauritius (3) – also vorwiegend männliche Hauptfiguren, Zofen und einzelne Opponenten.

In den Mittelwerten (Tabelle 3) würden demnach die Opponenten noch seltener als die weiblichen Hauptfiguren im Essens-Kontext genannt, obwohl einzelne Opponenten wie Marke und Partonopiers Mutter wohl aufgrund ihrer zentralen Stellung am Hof überdurchschnittliche Werte aufweisen.

Feiern: Der Aktivität ›Feiern‹ sind 13 Kategorien zugeordnet. Bei jeder Figur können sich also bis zu 13 Werte ergeben, die oberhalb des Konfidenzintervalls liegen. Bei den folgenden Figuren sind bei mehr als einer Kategorie erhöhte Werte vorhanden: Partonopier (6), Meliur (3), Sornagiur (5), Iwein (2), Brangäne (3), Tristan (7), Isolde (6), Marke (5), Irischer Truchsess (2), Mauritius (2). Feiern bzw. kulturelle Aktivitäten werden also insbesondere im Umkreis von männlichen Hauptfiguren vorgefunden; zudem sind zwei Könige (Sornagiur, Marke) beteiligt, deren Hofhaltung ebenfalls höfische Züge trägt. Die weiblichen Hauptfiguren sind nur teilweise in der Nähe solcher Kategorien angesiedelt – bei Isoldes hohem Wert darf man sofort an das gemeinsame Musizieren mit Tristan denken. Der leicht erhöhte Wert beim irischen Truchsess könnte mit seiner Präsenz am Königshof zusammenhängen.

In den Mittelwerten der 13 Kategorien (Tabelle 3) haben männliche und weibliche Hauptfiguren erhöhte Werte. Dass die weiblichen Hauptfiguren bei höfischen Festivitäten präsent sind, ist plausibel; Zofen werden bei solchen Gelegenheiten im Durchschnitt ebenso selten erwähnt wie die Opponenten – dies könnte gut mit der Fokusführung korrespondieren, die eher

Hauptfiguren als ihre Gegner bei diesen Aktivitäten zeigt. Insgesamt liegen damit ähnliche Tendenzen vor wie bei ›Essen‹ – mit der Ausnahme allerdings, dass Zofen weniger bei Festen als im Ernährungskontext genannt werden. Dies würde gut zu ihrer dienenden Funktion passen – man denke an Lunete, die den im Torraum gefangenen Iwein erst einmal mit Nahrung versorgt. – Eine ähnliche Tendenz wie bei ›Feiern‹ findet sich bei ›Jagen‹.

Schlafen: Hier gibt es bei den Aktanten-Mittelwerten nur geringe Unterschiede bei den Zofen, den männlichen und weiblichen Hauptfiguren; die Werte der Opponenten sind deutlich niedriger.

MW_alle	MW_Zofen	MW_mHF	MW_wHF	MW_Opp
9,1	10,2	10,1	10,2	7,8

Bei 23133000 (›Hilfeleistung‹) finden sich Werte oberhalb des Konfidenzintervalls bei Partonopiers Mutter, die immerhin meint, ihrem Sohn Gutes zu tun, sowie bei Sornagiur, der für die Freilassung von Partonopier sorgt, bei Lunete, Laudine, Brangäne sowie bei allen vier Figuren im ›Mauritius‹. Wie im Paidia-Aufsatz fallen in Tabelle 4 erwartungsgemäß die höheren Werte für die Zofen ins Auge. Die männlichen Hauptfiguren schneiden, denkt man an Iweins *helfe*-Aventiuren, überraschend schlecht ab, die weiblichen Hauptfiguren liegen nah am Mittelwert, die Opponenten knapp darunter.

Aktivität	alle	Zofen	Hauptfigur ♂	Hauptfigur ♀	Opponenten
Helfen	11,4	**13,6**	9,9	**11,4**	11,0
Hilfsmittel gebrauchen	7,8	7,1	**9,7**	**8,4**	7,1
Kämpfen	5,5	4,1	**7,2**	5,0	**5,5**
Turnieren	4,4	2,8	**6,9**	3,4	**4,4**

Tabelle 4: Höhe der Scores je Aktantengruppe, Teil 2 (Summenwerte aller Kategorien, die einer Aktivität zugeordnet sind)

Die drei Kategorien zur Aktivität ›Hilfsmittel gebrauchen‹ beziehen sich recht allgemein auf den Gebrauch von Gegenständen. Im Mittelwert haben die männlichen und weiblichen Hauptfiguren höhere Werte als die Opponenten und Zofen (Tabelle 4).

Der Aktivität ›Kämpfen‹ sind zwölf Kategorien zugeordnet. Mehr als ein Wert oberhalb des Konfidenzintervalls findet sich bei Irekel (3), Partonopier (8), Meliur (2), Sornagiur (8), Mareis (6), Iwein (5), Ascalon (4), Harpin (4), Laudines Truchsess (4), Brangäne (2), Isolde (2), Irischer Truchsess (7), Mauritius (3), Gräfin und Graf MvC (2). Erwartungsgemäß sind die Zofen und die weiblichen Hauptfiguren hier schwächer vertreten als die männlichen Hauptfiguren und die Opponenten; überraschend hoch ist der Wert des irischen Truchsessen, dessen Drachenkampf parodistische Züge trägt, und der Umstand, dass Tristan nur einen erhöhten Score aufweist. Der der mittlere Opponenten-Score (Tabelle 4) würde höher liegen, würde man Partonopiers Mutter (2,0) als weibliche Figur herausrechnen; auch Marjodo (2,4) zeichnet sich bekanntlich mehr durch Intrigen aus denn durch ritterliche Bewährung.

Ähnlich verhält es sich bei ›Turnieren‹. Dieser Aktivität sind fünf Kategorien zugeordnet. Mehr als ein erhöhter Wert findet sich bei Irekel (2), Partonopier (4), Sornagiur (3), Mareis (2), Iwein (2), Harpin (2), Laudines Truchsess (2), Irischer Truchsess (3), Mauritius (2), Gräfin MvC (1). Der niedrige Wert bei Tristan (1) könnte damit zusammenhängen, dass Tristan häufiger in der Nähe von anderen Aktivitäten erwähnt wird.

Die Rangfolge der Mittelwerte in Tabelle 4 ist analog zum Befund bei ›Kämpfen‹: Zofen haben deutlich unterdurchschnittliche Scores, weibliche Hauptfiguren leicht unterdurchschnittliche. Bei Opponenten finden sich durchschnittliche und bei männlichen Hauptfiguren deutlich überdurchschnittliche Werte.

Bei 31800000 (›Bewegung‹) haben erwartungsgemäß die weiblichen Hauptfiguren als eher statische[11] Figuren das Nachsehen gegenüber den übrigen Figuren; auffällig (Tabelle 5) ist, dass die Zofen in diesen Texten

noch häufiger im Bewegungskontext gezeigt werden als die männlichen Hauptfiguren. Während im Paidia-Aufsatz höhere Werte bei Isolde beobachtet wurden, weist hier die Gräfin im ›Mauritius‹ höhere Scores auf – womöglich spielt hier das Berechnungsmodell eine Rolle, das hier eher selten auftretende Figuren leicht bevorzugt.

Aktivität	alle	Zofen	Hauptfigur ♂	Hauptfigur ♀	Opponenten
Bewegung	10,6	**12,4**	**11,0**	8,9	10,5
Töten	10,1	8,3	9,4	9,1	**11,5**
Peinigen	7,4	**8,1**	**10,9**	7,1	5,8
Lügen	9,5	**10,8**	8,4	**11,1**	8,8
Reden etc.	9,8	**11,5**	9,7	**10,4**	8,9
Kopulieren	5,9	3,5	4,9	**9,5**	5,8

Tabelle 5: Höhe der Scores je Aktantengruppe, Teil 3 (Summenwerte aller Kategorien, die einer Aktivität zugeordnet sind)

›Töten‹ (21104000, ›Tod‹) ist erwartungsgemäß weniger Frauen- als Männersache, wobei auch die männlichen Hauptfiguren in Tabelle 5 unterhalb des Mittelwerts bleiben: Im Wesentlichen ist der Begriff an Opponenten gekoppelt.

›Peinigen‹ hat sich nur notdürftig zu der Kategorie 22705200 (›Ärger/Zorn/Wut‹) zuordnen lassen. Hier ist der niedrige Wert der Opponenten auf den ersten Blick kontraintuitiv. Denkbar ist jedoch, dass ›Ärger/Zorn/Wut‹ ausführlich bei den männlichen Hauptfiguren zur Darstellung kommt, auch wenn Ursache dafür andere Figuren waren, deren Referenz weiter zurückliegt. Dass die Darstellung von Leiderfahrungen bei den Opponenten nicht allzu ausgiebig erfolgt, ist wiederum erwartbar, wenn man eine Sympathiesteuerung voraussetzen darf, die eher die Hauptfiguren als die Opponenten begünstigt.

›Lügen‹ bzw. 22642000 (›Wahrheit/Irrtum‹) findet trotz des höheren Scores von Tristan im Kontext von männlichen Hauptfiguren unterdurch-

schnittlich statt. Die Daten in Tabelle 5 deuten eine genderspezifische Verteilung an – da Zofen und weibliche Hauptfiguren häufiger sprachlich handeln (s. u.), greifen sie womöglich häufiger zur Lüge. Allerdings kann bei den Opponenten ein breites Spektrum beobachtet werden: Während starke oder brutale Opponenten wie Ascalon oder Harpin sehr offen und direkt ihre Vorhaben ankündigen, findet man beim irischen Truchsessen und bei Marjodo im ›Tristan‹ besonders hohe Werte.

Gruppiert man die zehn Kategorien, die ›Reden‹, ›Argumentieren‹, ›Veranlassen‹ oder ›Beklagen‹ zugeordnet sind, findet sich mit Ausnahme von Ascalon und Laudines Truchsess bei allen Figuren mehr als ein Wert oberhalb des Konfidenzintervalls. Der Mittelwert für alle Figurenwerte beträgt 9,8. Überdurchschnittliche Werte weisen die Zofen auf – gefolgt von den weiblichen Hauptfiguren (Tabelle 5). Minimal unterdurchschnittlich sind die männlichen Hauptfiguren, deutlich unterdurchschnittlich die Opponenten.

Im Paidia-Aufsatz wurde der überraschende Befund beobachtet, dass die Opponenten beim sprachlichen Handeln sehr hohe Werte aufweisen. Dieser Befund könnte auf dem Berechnungsverfahren beruhen, das dort nur berücksichtigt, ob eine Figur in einem Segment die Aktivität überhaupt ausübt, während hier die Häufigkeit der Aktivität stärker berücksichtigt wird. Hier bleiben Opponenten nunmehr unter dem Mittelwert.

Während die bislang vorstellten Daten zumindest im Großen und Ganzen nicht in Widerspruch zum Lektüreeindruck menschlicher Rezipienten stehen, zeigen die Mittelwerte der beiden Kategorien 21112000 und 21045000 (›Sexualleben/Erotik‹ und ›Geschlechtsmerkmale/Genitalien/Zeugung/Geburt‹) teilweise unplausible Befunde in Tabelle 5. Unter den Zofen erfährt man nur bei Brangäne explizit von einem Beilager in Markes Hochzeitsnacht. Dass die Opponenten höhere Werte aufweisen als die männlichen Hauptfiguren, wird nicht einleuchten; der Befund wird auch nicht allein etwa durch Harpins Drohung verständlich, der die

Tochter von Gaweins Schwager zur Frau nehmen oder sie durch seine Knechte vergewaltigen lassen will.

Abschließend sei deshalb die Problematik der Wortform-basierten Analyse mit einem detaillierteren Blick auf die Kategorie 21112000 (›Sexualleben/Erotik‹) vorgeführt. Der Mittelwert für alle Figuren und alle Texte beträgt 7,27. Werte oberhalb des Konfidenzintervalls weisen auf: Meliur (13,1), Lunete (9,1), Laudine (9,1), Harpin (13,0), Brangäne (12,2), Tristan (11,6), Isolde (14,3), Marke (16,0), der irische Truchsess (13,1) und Marjodo (11,0). Bei Meliur, Laudine, Brangäne, Tristan, Isolde und Marke handelt es sich um erwartbare Befunde. In Widerspruch zu einer manuellen Erfassung der Aktivität ›Kopulieren‹ stehen die Werte von Lunete, Harpin, dem irischen Truchsess und Marjodo.

Zu einem erhöhten Score führen bei Lunete vier Nennungen im Kontext der Wortformen *entwunge* (8089), *kuster* (7976), *beruorte* (5384) und *nacket* (3238); bei Harpin die Wortform *ruote* (5058), beim irischen Truchsess die Wortformen *damoysele* (9165), *gewaltesaere* (11027), *schranken* (11254) und *understan* (11050), bei Marjodo *geminne* (13467). Zwar unterdurchschnittliche, dennoch aber fragwürdige Werte größer null haben Irekel, Partonopiers Mutter und Sornagiur, die auf den Wortformen *umbevangen* (8827), *minnedieben* (11430), *triute* (6647), *minnetranc* (6957), *geminnet* (6905), *muotgelust* (5893) und *grans* (5489) beruhen. Während zu *nacket* und *umbevâhen* bei der Kategorienbereinigung die zwar auch diskussionswürdige Zuordnung zu Kategorie 21111200 beibehalten wurde, sind die anderen genannten Wortformen so selten im Korpus, dass sie bei der Kategorienbereinigung nicht überprüft worden sind.

Diese Zuordnungen sind ein gutes Beispiel für das Verfahren der Kategorienzuordnung in der MHDBDB: ein Verfahren, das suchende Benutzer im Zweifelsfall lieber zu einem Suchtreffer führt, als potentielle Suchtreffer bei unsicherer Relevanz zu übergehen. Die Berührung im ›Iwein‹ (*beruorte* V. 5384) gilt den Panzerringen des Truchsessen durch den Löwen; Lunete wird im nächsten Vers erwähnt. Zur unsicheren

Zuordnung im Einzelfall kommt auch hier das Problem hinzu, dass Figuren nur zufällig im Kontext der Wortform stehen können, dass die Aktivität aber von anderen vollführt wird.

Unerwartet niedrige Werte finden sich bei Partonopier, bei Mauritius und der Gräfin im ›Mauritius‹. Hier zeigen sich die Grenzen von lexikalischen Zugriffen auf Phänomene, die oft umschreibend, verhüllend oder gar mit Unsagbarkeitstopos auf der Textoberfläche realisiert werden. Im ›Mauritius‹ heißt es: *nû begunde ouch er erwarmen. / unde tete der vrouwen ichn weiz waz. / waz hulfe ez, sagete ich iu daz? ez ist ungesaget alsô guot / ir wizzet wol waz man dâ tuot.* (V. 1614–1618. ›Nun wurde auch ihm immer wärmer. Ich weiß nicht, was er mit der Dame anfing. Was würde es helfen, wenn ich Euch das sagen würde? Wenn das ungesagt bleibt, ist es genauso gut. Ihr wisst ja gut, was man da so macht.‹) Das erste Beilager im ›Tristan‹ ist metaphorisch umschrieben: *ietwederez schancte unde tranc / die süeze, diu von herzen gie. / sô sî die state gewunnen ie, / sô gie der wehsel under in / slîchende her unde hin* (V. 12042–12046. ›Jeder von ihnen schenkte und trank die Süße, die vom Herzen kam. Wann immer sie die Möglichkeit dazu finden konnten, so ging dieser Warentausch zwischen ihnen leise hin und her.‹)

Je weniger unmittelbar ein Phänomen auf der Textoberfläche adressiert wird, umso schwieriger wird es auch künftig bleiben, dieses Phänomen mit automatischen Verfahren zu erfassen. Bei der Beilager-Thematik ist dieses Problem besonders virulent, weil es sich erstens um ein Phänomen handelt, das im höfischen Roman nicht offen adressiert wird, und zweitens um eines, das etwa im Artusroman im Vergleich zu Aktivitäten wie ›Kämpfen‹ oder ›Reden‹ seltener vorkommt. Bei häufigeren Phänomen ist eher zu erwarten, dass direkte Benennungen zumindest auch neben verhüllenden Umschreibungen stehen, so dass computerbasierte Verfahren weniger stark von spezifischen Einzelformulierungen abhängig sind, die manche Autoren gerade bei dieser Aktivität besonders elaboriert ausgestalten – man denke etwa an Wolframs *hirzwurz*-Metapher (›Parzival‹, 643,28).

12. Figurentypen mit Delta unterscheiden

In einem weiteren Analyseschritt werden alle Scores zu den 62 Kategorien in ein Delta-Verfahren eingespeist. Das Verfahren wird typischerweise für Studien zur Autorschaftsattribution verwendet und gruppiert hierbei basierend auf Wortfrequenzen Texte nach Autoren. Für die vorliegende Untersuchung werden analog dazu die Figuren in Bezug auf ihre Kategorien in Figurencluster gruppiert bzw. voneinander unterschieden.[12]
Erwartungsgemäß weist die Zofe zur weiblichen Hauptfigur in ›Iwein‹, ›Partonopier‹ und ›Mauritius‹ die größte Nähe auf (blau-rote Paare). Im ›Tristan‹ sind die Ähnlichkeiten von Tristan und Isolde derart groß, dass sie sogar stärker ins Gewicht fallen als die Gender-Differenz etwa zu Brangäne; eventuell könnte dieser Befund damit korrespondieren, dass Tristan und Isolde in vielen Szenen gleichzeitig im Erzählfokus stehen.

Die Opponenten Laudines Truchsess, Ascalon, Harpin, irischer Truchsess, Marjodo, Mareis und Sornagiur (schwarz in Abb. 1) weisen erhebliche Distanzen zu weiblicher Hauptfigur und Zofe auf – diese eher typischen Opponenten lassen sich also gruppieren.

Opponenten wie Marke und der Graf im ›Mauritius‹, die zugleich auch Ehemänner der weiblichen Hauptfigur sind, clustern relativ nahe zur weiblichen Hauptfigur; im grafischen Befund spiegelt sich damit, dass der Opponententyp ›Ehemänner‹ von anderen Opponenten wie Harpin oder Mareis unterschieden werden kann.

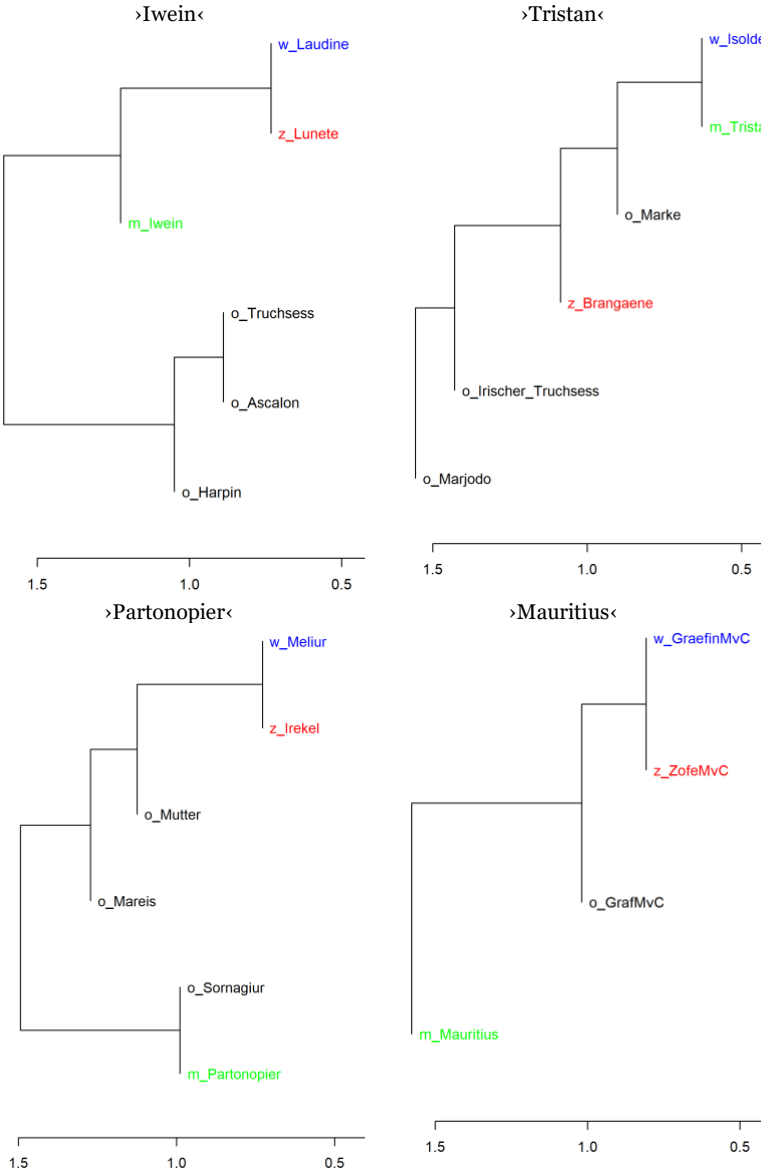


Abb. 1: Delta-Cluster zu den Aktanten

Bei den männlichen Hauptfiguren (grün) clustern Iwein und Tristan erwartungsgemäß nahe bzw. sehr nahe zur weiblichen Hauptfigur. Bei Partonopier überrascht auf den ersten Blick seine Nähe zum Opponenten Sornagiur. Allerdings ist Sornagiur kein klassischer Antagonist – nach dem Verrat von Mareis bewirkt er die Freilassung von Partonopier; die Gegnerschaft am Beginn wird schließlich in ein freundschaftliches Verhältnis überführt. In Bezug auf die Motive Tabubruch und Verrat kann Sornagiur auch als Komplementärfigur zu Partonopier aufgefasst werden (vgl. Dimpel 2015a, S. 63–67). Ein Delta-Plot ohne Sornagiur lokalisiert Partonopier wie etwa im ›Iwein‹ an den gleichen Ast, an dem auch die Zofe und weibliche Hauptfigur hängen; die beiden verbleibenden Opponenten sind deutlich von den übrigen Figuren getrennt:

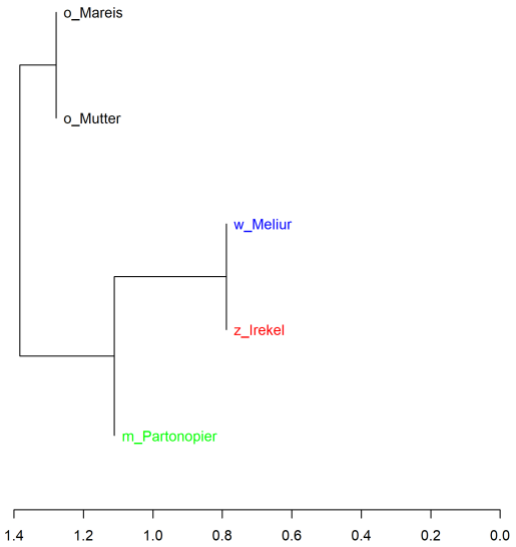


Abb. 2: Delta-Cluster zu den Aktanten im ›Partonopier‹ ohne Sornagiur

Leider sind im ›Mauritius‹ die Turniergegner derart marginal dargestellt, dass nur der Graf als singulärer Opponent, der noch dazu Ehemann der

Gräfin ist, in die Analyse eingehen kann; der Delta-Graph ist deshalb nur eingeschränkt mit den anderen Texten vergleichbar.

Jenseits der Figurenarmut des ›Mauritius‹ und der Einordnung von Tristan und Partonopier bei Isolde bzw. Sornagiur, die durch textspezifische Besonderheiten erklärbar ist, gelingt eine Erfassung der verschiedenen Figurentypen angesichts der niedrigen Zahl von nur 62 Textmerkmalen, die in das Delta-Verfahren eingehen, überraschend gut.

13. Abgleich mit der manuellen Annotation – Äpfel mit Birnen vergleichen?

Nachdem die automatischen Verfahren erfreulich plausible Ergebnisse hervorbringen, sollte abschließend überprüft werden, wie hoch der Grad der Übereinstimmung mit den manuell annotierten Daten aus dem Paidia-Aufsatz ist. Immer dann, wenn automatische Verfahren einer manuell gefundenen Lösung nahekommen sollen, träumt man von einem F1-Wert von 1,0 – also von einem perfekten Recall bei perfekter Precision. Ein solches Ziel wäre bei dem vorliegenden Versuch von vornherein utopisch – die zahlreichen Gründe dafür seien hier kurz skizziert:

- Bereits die manuelle Annotation ist nicht vollständig objektiv, weil nicht alle Aktivitäten tatsächlich explizit in den Texten vorliegen.
- Die Zuordnung der Aktivitäten zu Kategorien der MHDBDB müsste perfekt passen; tatsächlich wurden die Beschreibung der Aktivitäten und die Konzeption der Kategorien der MHDBDB für ganz unterschiedliche Zwecke erstellt, so dass eine Zuordnung von Kategorien zu Aktivitäten nur tentativ möglich ist. So zielte etwa die Paidia-Aktivität ›Hilfsmittel‹ auf zentrale Gegenstände wie Ringe, Zaubergürtel oder Salben. In den zugeordneten MHDBDB-Kategorien sind Lemmata zu Gegenständen im weiteren Sinn (Werkzeuge bzw. gegenstandsgebundene Tätigkeiten jeder Art) enthalten. Auch die Unterschiede etwa

zwischen ›Reden‹, ›Argumentieren‹ und ›Veranlassen‹ sind eine erhebliche Herausforderung für automatische Verfahren.

- Die Lemmata, die einer Kategorie zugeordnet sind, müssen diese Kategorie auch tatsächlich beschreiben; andere Bedeutungen bei Polysemie dürfen nicht überwiegen. Im Idealfall dürften Wortformen nur dann einer Kategorie zugerechnet werden, wenn sichergestellt ist, dass die im Text verwendete Bedeutung tatsächlich der Kategorie entspricht. Polysemie kann vereinzelt den Recall etwas verbessern, sie verursacht jedoch sehr häufig eine deutlich schlechtere Precision.
- Die Wortform sollte tatsächlich auf die Figur bezogen sein. Text im Umkreis von Figurenreferenzen kann jedoch auch auf andere Figuren, auf Gegenstände oder abstrakte Entitäten bezogen sein. Systematische Fehler sind zu erwarten im Zusammenhang mit Aussagen in Negationen oder Konjunktiv II.
- Im Paidia-Aufsatz wurden Aktivitäten-Zuordnungen zu Figuren in Analepsen nicht im Slot der eigentlichen erzählten Zeit erfasst, sondern sie wurden dem Segment zugeordnet, dem die nachgetragene Information eigentlich angehört. Für wichtige Analepsen wurden teilweise eigene Segmente gebildet. Hier werden Kookkurrenzen an der Textposition gezählt, an der sie vorgefunden werden.
- Einige Kategorien involvieren mehrere Figuren. Während in einem Beispielsatz wie ›Iwein kämpft mit Harpin‹ die Zuordnung von ›Kämpfen‹ zu Iwein und Harpin gleichermaßen korrekt ist, wäre in einem Beispielsatz wie ›Lunete hilft Iwein‹ zwar bei manueller Annotation evident, dass die Aktivität ›Helfen‹ nicht Iwein, sondern Lunete eingeschrieben ist. Bei einer automatischen Erfassung von Kategorien, die die fünf Worte vor und die 15 Worte nach einer Figurenreferenz einbezieht, führen solche Fälle jedoch zwangsläufig dazu, dass es sich bei der Hälfte der Treffer um False-Positives handelt; entsprechend ist mit einer niedrigen Precision zu rechnen.

- In der Paidia-Studie wurden Häufigkeitswerte nicht in gleicher Weise wie hier gebildet. Annotiert wurde, ob in einem Textsegment eine Aktivität überhaupt vorhanden ist oder nicht – ob eine Aktivität in einem Segment einmal oder häufig vorkommt, wurde nicht berücksichtigt. Die Anzahl der Segmente, in der die jeweilige Aktivität vorlag, wurde dort durch die Zahl der Segmente geteilt. Dieses Verfahren führt dazu, dass Figuren, die nur in wenigen Segmenten überhaupt vorkommen, tendenziell schlechtere Werte erzielen. Hier wurde ein komplexes Scoring-System entwickelt, das von solchen Effekten weniger stark betroffen ist, aber leicht dazu neigt, selten vorkommende Figuren zu bevorzugen.

13.1 Segment-Vergleich

Im ersten Analyseversuch hat das Phänomen, das im letzten Spiegelstrich beschrieben wurde, keine Rolle gespielt: Ein Skript liest zunächst die Daten aus der Paidia-Studie dazu ein, in welchem Textsegment zu welcher Figur die jeweilige Aktivität vorhanden ist oder nicht. In einem zweiten Abschnitt werden die mit Figurenreferenzen und Kategorien annotierten Dateien eingelesen – sowie die Mapping-Tabelle, die die Zuordnungen von Kategorien zu Aktivitäten enthält. Überprüft wird nun, ob in den Segmenten, in dem zu einer Figur mindestens eine Aktivität vorhanden ist, auch mindestens eine zugeordnete Kategorie im Umkreis von fünf, zehn, 15 oder 20 Worten vor und nach einer Figurenreferenz vorhanden ist (hier ohne Rückwärtsschwelle). Hier die Ergebnisse:

Spannweite	Recall	Precision	F1
5	0,688	0,337	0,452
10	0,813	0,300	0,438
15	0,858	0,283	0,425
20	0,875	0,270	0,412

Tabelle 6: Vergleich mit manueller Annotation

Diese Werte sind zwar erheblich von den idealen Werten entfernt; gemessen an den hier skizzierten Problemen handelt es sich jedoch um ein Ergebnis, das in Relation zu dem Aufwand, der in dieser Studie möglich war, noch akzeptabel ist. Mit steigender Spannweite nimmt die Precision ab; der Recall steigt zwar etwas, aber nicht in einem so großen Ausmaß, dass dadurch bei größeren Intervallen das Absinken der Precision kompensiert werden könnte.

Auch mit Blick auf das Zeitbudget wurden in der Annotationsphase pronominale Referenzen auf eine Figur nur dann erfasst, wenn eine Figurenreferenz in den zehn Versen zuvor nicht bereits annotiert wurde. Retrospektiv hat sich diese Entscheidung als recht unglücklich herausgestellt, da sich Spannweiten, die 15 Worte übersteigen, als ungeeignet erwiesen haben; am Beginn dieser Studie stand die Annahme, dass man mit deutlich größeren Spannweiten operieren könnte. Hier die Einzelwerte bei Spannweite = 10:

	Recall	Precision	F1
a_Hilfsmittel	0,970	0,204	0,337
a_bewegen	0,879	0,548	**0,675**
a_essen	0,600	0,200	0,300
a_feiern	0,143	0,022	0,038
a_helfen	0,711	0,274	0,395
a_jagen	0,500	0,075	0,131
a_kopulieren	0,577	0,313	0,405
a_kämpfen	0,900	0,135	0,235
a_schlafen	0,636	0,103	0,177
a_turnieren	0,875	0,137	0,237
a_töten	0,700	0,080	0,144
b_argumentieren	0,500	0,327	0,396
b_beklagen	0,925	0,343	**0,500**
b_lügen	0,714	0,215	0,331
b_reden	0,906	0,783	**0,840**
b_veranlassen	0,935	0,274	0,424

Tabelle 7: Vergleich mit manueller Annotation; Einzelwerte

Während ›Bewegen‹, ›Beklagen‹ und ›Reden‹ offenbar recht zuverlässig erkannt werden können, ist die automatische Erfassung von ›Feiern‹, ›Ja-

gen‹, ›Schlafen‹ und ›Töten‹ mit dem hier verwendeten Verfahren offenbar besonders fehleranfällig.

13.2 Vergleich von erhöhten Werten

Während im vorstehenden Abschnitt das gemeinsame Vorhandensein oder Fehlen einer Aktivität bzw. Kategorie im gleichen Segment untersucht wurde, wird hier verglichen, ob die Werte bzw. Scores für die Aktivitäten und die ihnen zugeordneten Kategorien jeweils erhöhte Werte aufweisen.

Bei den Aktivitäten geht dabei ein, ob die Aktivität einer Figur oberhalb des Konfidenzintervalls bei einem Konfidenzniveau von 80% liegt.[13] Da im Paidia-Aufsatz nur das Vorhandensein einer Aktivität in einem Segment berücksichtigt wurde, bleiben Unterschreitungen des Konfidenzintervalls hier unberücksichtigt. Wenn mehrere Kategorien einer Aktivität zugeordnet sind, genügt es für einen positiven Vergleich, wenn eine der Kategorien der jeweiligen Figur einen erhöhten Score oberhalb des Konfidenzintervalls aufweist: So, wie bei Erfassung der Aktivität ›Kämpfen‹ nicht spezifiziert wurde, ob es sich um ›Zweikampf‹ oder ›Heerfahrt‹ handelt, genügt für den positiven Abgleich mit der Aktivität ›Kämpfen‹, wenn eine dieser Kategorien einen erhöhten Score aufweist.

Auch auf diesem Weg wurden letztlich Äpfel mit Birnen verglichen, so dass sich ähnliche Vergleichswerte ergeben:

Recall	Precision	F1
0,681	0,325	0,440

14. Bilanz und Ausblick

Wichtiger als die Analyseergebnisse war es, mit dieser Studie zunächst Konzepte zu erarbeiten, um semantische Kategorien im Kontext von Figurenreferenzen untersuchen zu können. Neben der Annotation von Figurenreferenzen und der Übertragung der Daten aus dem Begriffssystem der

MHDBDB in die digitalen Texte wurde ein Scoring-System zur Auswertung solcher Kookkurrenzen entwickelt.

Unter den Analyse-Ergebnissen waren viele Befunde, die sich gut mit intuitiven interpretativen Annahmen parallelisieren lassen – etwa Befunde zu mittelalterlichen Gender-Konzepten: Kämpfen oder Turnieren erfolgt eher im Kontext von männlichen Figuren, mündliche Kommunikation eher bei weiblichen Figuren. Kulturelle Aspekte wie Feiern oder Jagen sind eher an Hauptfiguren gekoppelt – eventuell eine Konsequenz der Fokusführung. Weibliche Hauptfiguren, die mitunter als statische Figuren betrachtet werden, kommen selten im Kontext von Bewegungsvokabular vor. Zunächst überraschende Befunde werden meist plausibel, wenn man betrachtet, welche einzelnen Figuren durch ihre Spezifika zu den überraschenden Befunden beitragen. Zugleich werden aber auch die Grenzen des Verfahrens offenkundig – etwa bei verhüllender Redeweise bei Sexualität.

Die Delta-Plots zu allen 62 Kategorien weisen meist eine Nähe von Zofe und weiblicher Hauptfigur aus, viele typische Opponenten lassen sich gut gruppieren, während spezielle Opponenten (wie Ehemänner der weiblichen Hauptfigur) sich von den übrigen Opponenten abgrenzen lassen.

Eine interessante Herausforderung bestand darin, Figuren und Texte zu analysieren, die im Rahmen der Paidia-Studie bereits mittels manueller Annotation unter zumindest ähnlichen Gesichtspunkten untersucht worden sind. Angesichts der erheblichen konzeptionellen Differenzen zwischen beiden Studien ist es nicht möglich, die Paidia-Daten als Messlatte für die Qualität der hier automatisiert erhobenen Daten zu verwenden – es ist eher erstaunlich, dass der Vergleich überhaupt einen F1-Wert von 0,45 erreichen kann. Vergleicht man allerdings die zentralen Tendenzen wie höhere oder niedrigere Werte von bestimmten Aktantengruppen (etwa: weniger ›Bewegung‹ bei weiblichen Hauptfiguren als bei den anderen Aktanten; vgl. oben, Abschnitte 1 und 11), so zeigt sich, dass die Ergebnisse der Paidia-Studie mit Ausnahme der Aktivität ›Sexualleben/Erotik‹ in hohem Maße mit den aktuellen Ergebnissen übereinstimmen.

Folgestudien können sich von der ursprünglichen Konzeption der Paidia-Studie freimachen und weitere Kategorien, weitere Figurentypen und weitere Texte einbeziehen. Zu fragen bleibt, ob eventuell nicht nur Figurentypen, sondern auch Gattungs- oder Epochenspezifika auszumachen wären. Dafür wäre es künftig wichtig, alle Figurenreferenzen auch bei pronominalen Referenzen zu annotieren, und die Kategorien-Bereinigung auch für Wortform-Häufigkeiten bereits ab einer niedrigeren Okkurrenz-Schwelle vorzunehmen. Zudem wäre eine Auflösung von Polysemien erfreulich. Zumindest bei höherfrequenten Wortformen sollte es machbar sein, typische von seltenen Bedeutungen im Korpus unterscheiden zu können.

15. Was braucht das Fach?

Die digitale Literaturanalyse hat in den letzten 20 Jahren enorme Fortschritte gemacht. In der Mediävistik sind die Fortschritte allerdings viel kleiner als in den neueren Philologien. Ein möglicher Grund dafür ist, dass sich die computeraffinen Mediävisten überwiegend um Editionen und Handschriften verdient machen. Ein anderer Grund ist unser Ressourcenproblem: Immerhin sind allmählich etwas mehr Texte digital frei verfügbar, doch sind wir noch weit davon entfernt, dass zumindest alle zentralen Texte in aktuellen Ausgaben digital zur Verfügung stehen. Dass bei der DFG die Pflicht zur digitalen Publikation von Editionen nur eine Soll- und keine Muss-Bestimmung ist, ist angesichts der öffentlichen Finanzierung dieser Editionen ein unhaltbarer Zustand.

Mit den vorhandenen digitalen Texten sind Studien zur Autorschaftsattribution, zu Gattungen und wortschatzbezogenen Fragen möglich. Wir können heute eine hohe Wahrscheinlichkeit dafür vorrechnen, dass die ›Halbe Birne‹ von Konrad von Würzburg ist und dass die Nürnberger Weingrüße von Rosenplüt sind (Dimpel [u. a.] 2019; Dimpel/Wagner 2022).

Die hier vorgestellte Studie ist ein Beispiel dafür, welche Möglichkeiten bereits bestehen, wenn man auch die Figurenebene in den Blick nehmen will. Es werden aber vor allem auch unsere Ressourcenprobleme sichtbar. Gerade die digitale Literaturanalyse kommt nicht mit reinen Textdateien aus: Man braucht Anreicherungen, etwa eine vollständige Annotation von Figurenreferenzen, die auch Pronomina erfasst. Solange keine Texte mit entsprechenden Anreicherungen verfügbar sind, bleibt die Automatisierbarkeit von figurenbezogenen Studien schwierig, was bedauerlich ist, da die Figur eine doch recht zentrale Kategorie darstellt.

Hätte man Syntax-Parser, könnte man Aussagen in Negation und Konjunktiv II identifizieren.[14] In der vorliegenden Studie könnte man damit etwa erfassen, ob Figurenprofile plausibler werden, wenn man negierte oder unsichere Kategorien im Figurenkontext übergeht. Für andere Studien bräuchte man die Annotation von Orten oder weitere narratologische Annotationen.

Das ist umso wichtiger, als wir gegenüber der Neugermanistik und der Anglistik massiv in Nachteil sind: Dort gibt es gute Parser und andere Tools, von denen Mediävist*innen nur träumen können. Andererseits sind unsere zentralen Texte längst nicht so zahlreich, sodass es höchste Zeit ist, großangelegte Annotationsprojekte anzugehen: Wir brauchen eine breite Grundlage, damit man auch in der Mediävistik vernünftig digital arbeiten kann.

Weiterhin wäre es erfreulich, wenn gesprochene und gedachte Figurenrede einbezogen werden könnte. Dann wären auch Auswertungen denkbar, die einen höheren Grad an Sicherheit haben könnten: Welche Figur erwähnt welche Kategorie? Solche Annotationen wären auch für viele andere Zwecke hilfreich – SNA-Studien arbeiten beispielsweise häufig mit Figurenreferenzen und Figurenrede.

Für manche Zwecke benötigt man auch handschriftennahe Texte. Aber für quantifizierende und vergleichende literaturwissenschaftliche Studien sind normalisierte Texte notwendig. Und ideal wäre es, wenn alle digitalen

Texte konsequent nach den gleichen Regeln normalisiert wären; auch für die Parser-Entwicklung wäre das ein Vorteil.

Diese Ressourcen-Frage ist auch eine Frage der Nachwuchsförderung. W3-Professores können es sich eher leisten, Hilfskraftstunden in ein Projekt zu investieren, wenn man eine angereicherte Ausgangsbasis braucht. Im DH-Bereich ist aber überwiegend der wissenschaftliche Nachwuchs am Werk, der nicht über solche Ressourcen verfügt. Daher ist es essenziell, die Verfügbarmachung von Textkorpora sowie die Entwicklung geeigneter Tools in der digitalen Mediävistik voranzutreiben. Wenn hier keine wesentlichen Fortschritte stattfinden, wird sich der Rückstand der Mediävistik auf die neueren Philologien bei der digitalen Literaturanalyse noch weiter vergrößern.

Im Rahmen der Tagungsdiskussion hat ein Kollege bedauert, dass digitale Studien vielfach erwartbare Befunde vorstellen würden. Es wurde die Ansicht vertreten, dass Literaturanalyse wohl weiterhin Handarbeit bleiben würde. Eine solche Skepsis hat bereits Jannidis (1999) beobachtet:

> Quantitative Verfahren werden in der Literaturwissenschaft nicht immer gern gesehen. Bestätigen sie gängige Einsichten, stehen sie im Verdacht, überflüssig zu sein. Widersprechen sie aber den üblichen Ansichten, schafft man sie sich mit dem leisen Hinweis vom Hals, daß man Statistiken ohnehin nicht trauen könne.

Mit Blick auf manche traditionellen Literaturwissenschaftler mag es auch 2022 noch hilfreich sein, wenn man Bekanntes bestätigen kann, um demonstrieren zu können, dass man mit digitalen Methoden verlässliche Ergebnisse hervorbringen kann, so dass man wenigstens nicht dem Verdikt verfällt, man würde fragwürdige Statistiken erzeugen. Doch eigentlich hat die digitale Literaturanalyse einige Fortschritte erlebt: Das Bestätigen von Bekanntem kann man mit Eibl (2013, S. 37), auch positiv als »Kontrollpeilung« einstufen. Vielfach wird (wie in der vorliegenden Studie) eine Methode zuerst anhand von bekanntem Material evaluiert, bevor man sie

in einem größeren Rahmen (hier: Korpus und weitere Kategorien) einsetzt. Andrea Rapp hat in der Diskussion darauf hingewiesen, dass man in den Naturwissenschaften geradezu froh ist, wenn man mit verschiedenen Methoden zu den gleichen Befunden kommt und so eine unabhängige Bestätigung findet.

Wenn man konventionelle Fragestellungen so modelliert, dass sie digital implementierbar werden, wird offensichtlich, dass man viel kleinschrittiger und präziser vorgehen muss als bei konventionellen oder narratologischen Beschreibungen: »Die kleinste Lücke, Ambiguität oder gar Widersprüchlichkeit im narratologischen Modell führt, sobald sie sich in der Software niederschlägt, zum Desaster.« (Meister 2013, S. 294; vgl. auch McCarty 2005, S. 46.) Insofern sind die Modellerzeugung und die Modellweiterentwicklung eine wesentliche Leistung der digitalen Literaturanalyse (zum »Modelling« vgl. McCarty 2005, S. 20–72); oft werden dabei Voraussetzungen und Vorannahmen präziser benannt als in konventionellen Studien. So wird in der vorliegenden Studie nicht die Aktivität ›Kämpfen‹ erfasst, sondern es werden ausgewählte lexikalische Referenzen erfasst. Die Lemmata, die den einschlägigen MHDBDB-Begriffen zugeordnet sind, realisieren dabei eine graduelle Annäherung zur Aktivität ›Kämpfen‹ (zum Begriff »proxy« als Stellvertretung für die gesuchten Phänomene bei digitalen Operationalisierungen vgl. etwa Moretti 2013, S. 2–5). Dabei ist anzumerken, dass viele traditionelle Studien keine Rechenschaft darüber ablegen, welches Konzept eigentlich gemeint ist, wenn sie über die Aktivität ›Kämpfen‹ schreiben. Dennoch bleibt es ein Vorzug einer digitalen Studie, dass nicht nur an ihrem Beginn interessegeleitete Fragestellungen und Modellierungen klar benannt werden und dass an ihrem Ende transparente Interpretationen von Ergebnisdaten stehen, sondern dass es dazwischen auch empirische Teilstrecken gibt (zum Computerphilologen als »Teilzeitempiriker« vgl. Dimpel 2015b, S. 349–354).

Anmerkungen

1 Vgl. hierzu z. B. Ketschik [u. a.] 2020; Blessing [u. a.] 2017. Zu CRETA vgl. https://www.creta.uni-stuttgart.de/.

2 Beim ›Mauritius von Craun‹ folgt der Datenbankabzug der Pretzel-Ausgabe, abgeglichen wurde er mit einem digitalen Text, der auf der älteren Ausgabe von Edward Schröder beruht. Gegenüber dem Schröder-Text fehlten in dem Datenbankabzug knapp 20 Verse.

3 Vier Figuren im Falle des ›Mauritius‹.

4 Für das Vorannotieren wurden Passagen gewählt, in denen alle Figuren einigermaßen häufig auftreten: Partonopier: V. 5623-6134, 7435-7535, 7755-7908, 8595-9114; Tristan: V. 9365-9896, 12507-12674, 13451-13672.

5 Für die Annotationsphase wurden bei 14h pro Monat neun Monate benötigt.

6 Bei einem Plausibilitätscheck der exportierten Daten hat sich ergeben, dass seit dem Export der Texte im Jahr 2018 durch die fortlaufende Arbeit an der Datenbank einige Begriff-Lemma-Zuordnungen in der MHDBDB entfernt worden sind.

7 Das Verfahren ist nicht vollständig zuverlässig, da Wortformen zu verschiedenen Lemmata gehören können.

8 Die MHDBDB strebt eine Disambiguierung der Bedeutungen im Kontext an; wenn künftig disambiguierte Texte vorliegen, werden solche Zuordnungen weniger Probleme bereiten. Ebenfalls bislang nicht gelöst ist das Problem, dass eine Wortform zu mehreren Lemmata zugeordnet sein kann. Da die Bereinigung auf Lemma-Ebene und nicht auf Wortform-Ebene erfolgt, wurden Zuordnungen bei allen Wortformen entfernt, wenn ein zugehöriges Lemma in der Löschliste enthalten ist.

9 Vgl. etwa zur Z-Wert-Begrenzung bei Autorschaftsattributionsstudien Evert [u. a.] 2016, S. 62–64.

10 Im Gesamtmittelwert »MW-alle« für alle Figuren sind die Opponenten stärker repräsentiert als die anderen Aktantengruppen, da mit Ausnahme des ›Mauritius‹ drei Opponenten pro Text untersucht wurden, jedoch nur eine Zofe sowie eine männliche und weibliche Hauptfigur. »MW-alle« entspricht also nicht einem Mittelwert der Mittelwerte der Aktantengruppen (MW_Zofen, MW_mHF, …), da es sich um den Mittelwert aller Einzelwerte handelt.

11 Zu dynamischen vs. statischen Figuren vgl. Lotman 1981.

12 Für die Delta-Analyse und die Cluster-Grafik wurde das stylo-R-Paket von Eder [u. a.] 2017 verwendet. Übergeben wurden die hier ermittelten Scores als »table_with_frequencies.txt«. Zu Delta bei Autorschaftsfragen vgl. etwa

Burrows 2003, Büttner [u. a.] 2017. Damit stylo-R den Figuren, die zum gleichen Aktantentyp gehören, die gleiche Farbe zuordnet, wurde den Figurenbezeichnungen hier ein z_, m_, w_ oder o_ vorangestellt.

13 Diese Schwelle wurde etwas niedriger gewählt als die Schwelle im Paidia-Aufsatz, damit eine breitere Vergleichsbasis vorliegt, in die auch leicht erhöhte Werte eingehen. Die im Paidia-Aufsatz mit + bzw. – markierten Werte wurden nur innerhalb der jeweiligen Aktantengruppe berechnet; hier werden die Abweichungen in Bezug auf alle Figuren berechnet.

14 Der Stuttgarter POS-Tagger für das Mittelhochdeutsche erreicht bei der Wortartenbestimmung eine Erkennungsgenauigkeit von F1=0,82 sowie eine Accuracy von bis zu 0,91. Vgl. Echelmeyer [u. a.] 2017; Schulz/Ketschik 2019. Für eine Anwendung im Kontext quantitativer literaturwissenschaftlicher Studien wäre eine Steigerung der Erkennungsgenauigkeit wünschenswert. Ein erstes Experiment zu einem Tagger, der auch Flexionsformen bestimmt, hat Helmut Schmid (LMU) im Workshop »Automatische Annotation digitaler Editionen« (Stuttgart, 16./17.3.2022) vorgestellt; eine Evaluation auf einer ausreichenden Datenmenge steht noch aus.

Literaturverzeichnis

Primärliteratur

Gottfried von Strassburg: Tristan und Isold, hrsg. von Friedrich Ranke, Berlin 1968.

Hartmann von Aue: Iwein. Eine Erzählung von Hartmann von Aue, hrsg. von G.F. Benecke und Karl Lachmann, neu bearb. von Ludwig Wolff, siebente Ausgabe, Berlin 1968.

Konrad von Würzburg: Partonopier und Meliur, hrsg. von Karl Bartsch, Berlin 1970.

Moritz von Craûn, hrsg. von Ulrich Pretzel, 4. Aufl., Tübingen 1973 (ATB 45).

Wolfram von Eschenbach: Parzival. Studienausgabe, 2. Aufl. Mittelhochdeutscher Text nach der sechsten Ausgabe von Karl Lachmann, Übersetzung von Peter Knecht. Mit Einführung zum Text der Lachmannschen Ausgabe und in Probleme der ›Parzival‹-Interpretation von Bernd Schirok, Berlin/New York 2003.

Sekundärliteratur

Agarwal, Apoorv/Corvalan, Augusto/Jensen, Jacob/Rambow, Owen: Social Network Analysis of ›Alice in Wonderland‹, in: Workshop on Computational Linguistics for Literature. Montréal, Canada, June 8, 2012, Montréal 2012, S. 88–96 (online).

Agarwal, Apoorv/Rambow, Owen: Automatic Detection and Classification of Social Events, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Massachusetts 2010, S. 1024–1034 (online).

Blessing, André/Echelmeyer, Nora/John, Markus/Reiter, Nils: An End-to-end Environment for Research Question-Driven Entity Extraction and Network Analysis, in: Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Vancouver, Canada 2017 (online).

Braun, Manuel/Ketschik, Nora: Soziale Netzwerkanalysen zum mittelhochdeutschen Artusroman – oder: Vorgreiflicher Versuch, Märchenhaftigkeit des Erzählens zu messen, in: Das Mittelalter 24 (2019), S. 54–70 (online).

Büttner, Andreas/Dimpel, Friedrich Michael/Evert, Stefan/Jannidis, Fotis/Pielström, Steffen/Proisl, Thomas/Reger, Isabella: »Delta« in der stilometrischen Autorschaftsattribution, in: Zeitschrift für digitale Geisteswissenschaften (2017) (online).

Burrows, John F.: Questions of Authorship: Attribution and Beyond. A Lecture Delivered on the Occasion of the Roberto Busa Award, ACH-ALLC 2001, New York, in: Computers and the Humanities 37 (2003), S. 5–32 (online).

Dimpel, Friedrich Michael: Wertungsübertragungen und korrelative Sinnstiftung im ›Herzog Ernst B‹ und im ›Partonopier‹, in: DVjs 89 (2015a), S. 41–69.

Dimpel, Friedrich Michael: Der Computerphilologe als Interpret – ein Teilzeit-Empiriker? in: Borkowski, Jan/Descher, Stefan/Ferder, Felicitas/Heine, Philipp David (Hrsg.): Literatur interpretieren: Interdisziplinäre Beiträge zur Theorie und Praxis, Münster 2015b, S. 339–359.

Dimpel, Friedrich Michael: Novellenschätze narratologisch auszeichnen und analysieren am Beispiel Victor von Scheffels ›Hugideo‹ und der sozialen Netzwerkanalyse, in: Weitin, Thomas/Werber, Niels (Hrsg.): Scalable Reading, Siegen 2017 (LiLi 47), S. 87–108 (online).

Dimpel, Friedrich Michael: Versuch einer quantitativen Analyse von Figurenaktivitäten in ›Iwein‹, ›Tristan‹, ›Partonopier‹ und ›Mauritius von Craun‹ in Analogie zu Computerspielen, in: Ascher, Franziska/Müller, Thomas (Hrsg.): Paidia Sonderausgabe: Vom ›Wigalois‹ zum ›Witcher‹ – Mediävistische Zugänge zum Computerspiel, 2018 (online).

Dimpel, Friedrich Michael/Schlager, Daniel/Zeppezauer-Wachauer, Katharina: Der Streit um die Birne. Autorschafts-Attributionstest mit Burrows' Delta und dessen Optimierung für Kurztexte am Beispiel der ›Halben Birne‹ des Konrad von Würzburg, in: Bleier, Roman/Fischer, Franz/Hiltmann, Torsten/Viehhauser, Gabriel/Vogeler, Georg (Hrsg.): Digitale Mediävistik, 2019 (Das Mittelalter. Perspektiven mediävistischer Forschung. Zeitschrift des Mediävistenverbandes, Band 24), S. 71–90.

Dimpel, Friedrich Michael: Soziale Netzwerkanalyse und Erzählschemata. Eine explorative Vorstudie, in: Ernst, Marlene/Hinkelmanns, Peter/Zangerl, Lina Maria/Zeppezauer-Wachauer, Katharina (Hrsg.): digital humanities austria 2018. empowering researchers, Wien 2020, S. 95–111 (online).

Dimpel, Friedrich Michael/Wagner, Silvan: Rosenplüt als Autor der Nürnberger Weingrüße. Philologische und computerphilologische Analysen, in: Jurchen, Sylvia/Wagner, Silvan (Hrsg.): Schlechtes Wetter und Grenzüberschreitungen, Oldenburg 2022 (Zeitschrift Brevitas 2, BmE Sonderheft) (online).

Echelmeyer, Nora/Reiter, Nils/Schulz, Sarah: Ein PoS-Tagger für »das« Mittelhochdeutsche, in: Konferenzabstracts DHd 2017: Digitale Nachhaltigkeit, Bern 2017, S. 141–147 (online).

Eder, Maciej/Rybicki, Jan/Kestemont, Mike: stylo R package, in: 2017 (online).

Eibl, Karl: Ist Literaturwissenschaft als Erfahrungswissenschaft möglich? Mit einigen Anmerkungen zur Wissenschaftsphilosophie des Wiener Kreises, in: Ajouri, Philip/Mellmann, Katja/Rauen, Christoph (Hrsg.): Empirie in der Literaturwissenschaft, Münster 2013 (Poetogenesis. Studien zur empirischen Anthropologie der Literatur 8), S. 19–45.

Elson, David K./Dames, Nicholas/McKeown, Kathleen R.: Extraction Social Networks from Literary Fiction, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala 2010, S. 138–147.

Evert, Stefan [u. a.]: Burrows' Delta verstehen, in: Burr, Elisabeth (Hrsg.): Konferenzabstracts DHd 2016. Modellierung – Vernetzung – Visualisierung. Die Digital Humanities als fächerübergreifendes Forschungsparadigma, Leipzig 2016, S. 62–65 (online).

Hallig, Rudolf/Wartburg, Walther von: Begriffssystem als Grundlage für die Lexikographie. Versuch eines Ordnungsschemas, Berlin 1963.

Jannidis, Fotis: Was ist Computerphilologie? in: Jahrbuch für Computerphilologie 1 (1999), S. 39-60 (online).

Jannidis, Fotis: Figur und Person. Beitrag zu einer historischen Narratologie, Berlin/New York 2004 (Narratologia 3).

Ketschik, Nora/Blessing, André/Murr, Sandra/Overbeck, Maximilian/Pichler, Axel: Interdisziplinäre Annotation von Entitätenreferenzen, in: Reiter, Nils/Pichler, Axel/Kuhn, Jonas (Hrsg.): Reflektierte algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt, Berlin/Boston 2020, S. 203–263 (online).

Krautter, Benjamin/Pagel, Janis/Reiter, Nils/Willand, Marcus: »[E]in Vater, dächte ich, ist doch immer ein Vater«. Figurentypen im Drama und ihre Operationalisierung, in: ZfdG (2020) (online).

Lexer, Matthias: Mittelhochdeutsches Handwörterbuch. 3 Bde., Leipzig 1872–1878. Nachdr. Stuttgart 1992 (online).

Lotman, Jurij M.: Die Entstehung des Sujets – typologisch gesehen, in: Ders. (Hrsg.): Kunst als Sprache. Untersuchungen zum Zeichencharakter von Literatur und Kunst, Stuttgart 1981, S. 175–204.

McCarty, Willard: Modeling: A Study in Words and Meanings, in: Schreibman, Susan/Siemens, Ray/Unsworth, John (Hrsg.): A Companion to Digital Humanities, Oxford 2004, S. 254–270 (online).

McCarty, Willard: Humanities Computing, London; New York 2005.

Meister, Jan Christoph: Computerphilologie vs. Digital Text Studies. Von der pragmatischen zur methodologischen Perspektive auf die Digitalisierung der Literaturwissenschaften, in: Grond-Rigler, Christine/Straub, Wolfgang (Hrsg.): Literatur und Digitalisierung, Berlin/Boston 2013, S. 267–296.

Moretti, Franco: »Operationalizing«: or, the function of measurement in modern literary theory, Stanford 2013 (Pamphlets of the Stanford Literary Lab 6).

Pagel, Janis/Reiter, Nils: GerDraCor-Coref: A Coreference Corpus for Dramatic Texts in German, in: Proceedings of the 12th Language Resources and Evaluation Conference (LREC), Marseille 2020.

Roget, Peter Mark: Thesaurus Of English Words And Phrases. Classified and arranged so as to facilitate the expression of ideas and assist in literary composition. Unter Mitarbeit von Barnas Sears. Revised and Edited, Boston 1864.

Schulz, Sarah/Ketschik, Nora: From 0 to 10 million annotated words: part-of-speech tagging for Middle High German, in: Language Resources and Evaluation 53 (2019), S. 837–863 (online).

Trilcke, Peer: Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft, in: Ajouri, Philip/Mellmann, Katja/Rauen, Christoph (Hrsg.): Empirie in der Literaturwissenschaft, Münster 2013 (Poetogenesis. Studien zur empirischen Anthropologie der Literatur 8), S. 201–247.

Viehhauser, Gabriel/Barth, Florian: Towards a Digital Narratology of Space, in: Digital Humanities 2017. Conference Abstracts. McGill University & Université de Montréal. Montréal, Canada 2017.

Wiedmer, Nathalie/Pagel, Janis/Reiter, Nils: Semi-Automatische Extraktion von Beziehungen zwischen dramatischen Figuren, in: DHd 2020. Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts, Paderborn 2020, S. 194–200 (online).

Online-Ressourcen

Ausführliche Ergebnis-Daten zu diesem Beitrag im Dariah-Repositorium: https://dx.doi.org/10.20375/0000-000F-322E-6.

CRETA (Center For Reflected Text Analystics): https://www.creta.uni-stuttgart.de/, https://www.creta.uni-stuttgart.de/tools/index.html.
CRETAnno [web-basierte Website zur Annotation, in Arbeit]: http://hdl.handle.net/11022/1007-0000-0007-E1BE-5. CRETAnno Annotationsrichtlinien: https://www.creta.uni-stuttgart.de/wp-content/uploads/2016/09/Annotationsrichtlinienv1_1.pdf.
dlina (Digital Literary Network Analysis): https://dlina.github.io/.
MHDBDB (Mittelhochdeutsche Begriffsdatenbank): http://mhdbdb.sbg.ac.at/.
NER (Named Entity Recognition): https://fortext.net/routinen/methoden/named-entity-recognition-ner;
Stanford NER-Modelle: https://nlp.stanford.edu/software/CRF-NER.html.
OED (Historical Thesaurus of the Oxford English Dictonary): https://www.oed.com/thesaurus/.
QuaDramA-Projekts (QuaDramA: Quantitative Drama Analytics): https://www.ims.uni-stuttgart.de/forschung/projekte/quadrama/.

Anschrift der Autorinnen und Autoren:

Prof. Dr. Friedrich Michael Dimpel
Friedrich-Alexander-Universität Erlangen-Nürnberg
Bismarckstraße 1
91054 Erlangen
E-Mail: mail@dimpel.de
ORCID: 0000-0003-4833-4897
GND: 1057584525

Andre Blessing
Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung
Pfaffenwaldring 5b
70197 Stuttgart
E-Mail: andre.blessing@ims.uni-stuttgart.de
ORCID: https://orcid.org/0000-0001-7573-578X
GND: https://d-nb.info/gnd/1058601865

Peter Hinkelmanns
Paris-Lodron-Universität Salzburg
Mittelhochdeutsche Begriffsdatenbank MHDBDB

Erzabt-Klotz-Straße 1
A-5020 Salzburg
E-Mail: peter.hinkelmanns@plus.ac.at
ORCID: https://orcid.org/0000-0001-8618-0185

Nora Ketschik
Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung
Pfaffenwaldring 5b
70569 Stuttgart
E-Mail: nora.ketschik@ilw.uni-stuttgart.de
ORCID-ID: 0000-0001-8758-5432

Dr. Katharina Zeppezauer-Wachauer
Paris-Lodron-Universität Salzburg
Mittelhochdeutsche Begriffsdatenbank MHDBDB
Erzabt-Klotz-Straße 1
A-5020 Salzburg
E-Mail: Katharina.Wachauer@plus.ac.at
ORCID: 0000-0001-9310-9029
GND: 1140611232

Elisabeth Lienert

Stilometrie und Textanalyse (Diskussionsbericht Sektion 6)

Stilometrie und Textanalyse sind die Bereiche, wo sich am deutlichsten zeigen wird, wie die digitalen Methoden die Mediävistik verändern werden (Manuel Braun). Die Diskussion (Leitung: Manuel Braun) aller drei Vorträge fokussierte vor allem auf die Materialgrundlage in Form von Editionen; Probleme der (quantitativen) Methodik, besonders der Annotation und der Kontextualisierung; das Verhältnis von Input und Output, Vorannahmen/Fragestellungen und Ergebnis, damit verbunden einerseits die Fragen von Objektivität, Scheinobjektivität, Modellbildung und andererseits die Frage nach der literaturwissenschaftlich-interpretatorischen Relevanz; das Problem der Komplexitätsreduktion; weitere Voraussetzungen für Arbeiten in diesem Bereich.

Als Grundlage für quantitative digitale Analysen seien normalisierte und zwar möglichst einheitlich normalisierte Ausgaben nötig. Eine Gefahr bestehe darin, dass unterschiedlich verfahrende Editionen im Korpus die Ergebnisse beeinflussen (etwa durch Unschärfen bei der Erfassung von Begriffen aufgrund unterschiedlicher Normalisierung: Klaus Kipf; hierzu fragte Sonja Glauch, ob der Einfluss der Editionsqualität ›herausrechenbar‹ sei); eine weitere darin, dass Konzentration auf die ›Höhenkammliteratur‹ drohe (Freimut Löser). In die Arbeiten sollten möglichst immer die Ergebnisse der MHDBDB einbezogen werden, weitere Vernetzung mit (annotierten) digitalen Wörterbüchern sei wünschenswert. Die in den vorgestellten Projekten entwickelten Tools selbst sollten (Stephan Müller) und können (Phillip Brandes) nachnutzbar sein.

Das Methodenarsenal sei begrenzt, rein quantitativ (›Zählen‹ von Wörtern und allenfalls ihren – auf wenige Wörter oder Verse begrenzten – Kontexten) und damit aussagekräftig für große Mengen, nicht für Einzelfälle – schon gar nicht für Interpretationen konzeptionell und ästhetisch außergewöhnlicher Einzelwerke (Gabriel Viehhauser, Müller). Quantitative Ergebnisse wie Wordclouds bedürften der Auswertung (Brigitte Bulitta). Quantative Analysen und ›qualitative‹ Interpretation, *distant* und *close reading*, Statistik und Interpretation könnten sich jedoch sinnvoll ergänzen (Müller, Simone Schultz-Balluff). Probleme bei der (Nicht-)Berücksichtigung von Negiertem, bei der Identifikation von Ironie, Polysemie, Metapher (Michael Stolz) könnten über Word Embeddings angegangen werden (Viehhauser); insbesondere Disambiguierung bleibe jedoch schwierig (Friedrich Dimpel), manuelle Kontrolle ggf. nötig (Tina Terrahe, Dimpel). Auch Probleme der Kontextualisierung minderten die Präzision (Dimpel); pronominale Referenzen müssten praktisch immer manuell präzisiert werden (Dimpel).

Ausgegangen werde immer von Vorannahmen, häufig literaturgeschichtlich längst Etabliertem (etwa Epochen- und Gattungsbegriffen), womöglich gar Veraltetem (Müller). Es besteht die Gefahr der Zirkularität (Albrecht Hausmann, Löser) und Bestätigung des Altbekannten (Löser, Müller). Der Vorteil digitaler Verfahren liege jedoch in der Transparenz und Explizitheit hinsichtlich der Vorannahmen (Viehhauser, Dimpel), ohne die auch traditionelle Verfahren nicht auskämen. Die Gefahr der Scheinobjektivität quantitativer Verfahren sei nicht auszuschließen (Hausmann). Visualisierung etwa durch Wortwolken wolle jedoch nur graduell objektivieren (Schultz-Balluff); erkenntnisgeleitete Fragestellungen und Auswertung der Befunde seien immer nötig (Löser).

Erfasst werde bei digitalen Stilanalysen vielfach nur die Wortebene; damit stelle sich die Frage, wie man zu den Texten gelange (Hausmann): über POS (Part of Speech-Tagger) für die Syntax, Word Embeddings für die Semantik (Viehhauser, Brandes, Dimpel). Moniert wurde das Missver-

hältnis von Aufwand und Ergebnis, Komplexität der Verfahren und (bislang) Schlichtheit der Fragestellungen und Ergebnisse (Elisabeth Lienert). Hier sei auf die Weiterentwicklung der Verfahren und die Entwicklung neuer Fragestellungen zu hoffen.

Für die digitale Analyse mittelalterlicher Literatur würden digitale Texte (TEI) benötigt, insbesondere (einheitlich) normalisierte (annotierte) Editionen für quantitative Analysen. Gezielte öffentliche Förderung sei umso notwendiger, als DH-affine, nicht-etablierte Forscher*innen in besonderem Maß mit Ressourcen-Problemen konfrontiert seien (Dimpel). Kontrovers diskutiert wurden allerdings die Forderung (Dimpel) nach einem großen Korpus mit basalen Entity- sowie narratologischen Annotationen und die Durchführbarkeit dieser Forderung. Besonders wichtig seien auch in diesem Kontext Vernetzung (insbesondere die Anknüpfung an MHD-BDB: Dimpel; disambiguierte Verlinkungen mit den Wörterbüchern: Torsten Schaßan; Kooperation zwischen Projekten: Brandes) und der Aufbau zentraler Ressourcen (Dimpel). Die Praktikabilität eines annotierten Korpus sei jedoch fraglich (Braun), da nicht klar sei, wer entscheide, welche Texte annotiert werden, und kein Konsens bestehe, was aus der Überfülle der Möglichkeiten annotiert werden solle (Löser, Braun), zumal Bedarfslagen interessegeleitet seien (Dimpel). Hier votierte Viehhauser dafür, die Annotationen auf Grundsätzliches zu beschränken (etwa Entitäten, Figuren) und zunächst auf zu Spezielles und zu Komplexes zu verzichten.

Anschrift der Berichterstatterin:

Prof. Dr. Elisabeth Lienert
Universität Bremen
Fachbereich 10
Universitäts-Boulevard 13
28359 Bremen
E-Mail: elienert@uni-bremen.de

Elisabeth Lienert

Bericht über die Abschlussdiskussion

Martin Schubert[1] als Diskussionsleiter stellte die Diskussion in die Spannungsfelder von retrospektiven Feststellungen vs. prospektiven Erwartungen an neue Fragen und Methoden, von Deklarationen und Appellen. In der Retroperspektive lasse sich feststellen, dass mittels Digitalisierung mindestens das im Fach Erreichte weiter verbreitet werden könne. Das Fach brauche auch digitale Publikationsformen, deren Renommee und Relevanz derzeit überhaupt erst etabliert würden (Schubert). Für Arbeiten mit Verfahren der Digital Humanities (DH) schienen Wahl und Art des Korpus, begriffliche Voraussetzungen und insbesondere geeignete Fragestellungen entscheidend (Schubert).

In der Diskussion wurden vor allem der Status digitaler Verfahren (hilfreiche Methodiken/eigenständige Forschung), die Akzeptanz von Publikationsformen sowie die bisherigen Erfolge von DH-Forschungen (Schubert) verhandelt. Unstrittig sind Nutzen und Mehrwert digitaler Tools: Als Hilfsmittel für die Philologie erleichtern sie die Arbeit (Albrecht Hausmann, Stephan Müller u. a.). Digitale Editionen erschienen einigen Diskutanten problematischer. Die größten Schwierigkeiten stellten sich beim Einsatz von DH-Methoden für genuin literaturwissenschaftliche Fragen (Hausmann). Es gebe Fragestellungen, Methoden und Erkenntnisse, die Forscher*innen gar nicht ans Digitale delegieren wollten (Schubert). Hier plädierte Andrea Rapp für eine Skala von Digitalität; wichtig sei Offenheit für die digitalen Methoden; es werde noch zu viel vom einfachen menschlichen Lesen ausgegangen; es gehe darum, digitale Verfahren in die Philologie einzuordnen; einfache Dichotomien von digitalem Hilfsmittel und forschen-

dem menschlichem Hirn, von digitaler Hilfswissenschaft und eigentlicher Wissenschaft griffen zu kurz (Schubert, Klaus Kipf); Digital Humanities seien nicht Hilfswissenschaft, sondern Grundlagenforschung (Tina Terrahe).

Freimut Löser argumentierte gegen die Dichotomie von Philologie und Digitaler Edition: Auch digitale Editionen müssten philologisch aufbereitet sein; auch hybride Modelle seien eine Option. Stets geklärt werden müssten die Fragen, was mit welchen Methoden wie, warum und wozu erreicht werden solle und könne. Dabei müsse jedes Projekt seine digitalen Methoden selbst entwickeln – eine einfache Übertragbarkeit sei nicht automatisch gegeben. Gabriel Viehhauser hielt Wert und Nutzen digitaler Editionen mittlerweile für unstrittig. Für die ›höhere Kritik‹ dagegen sei es nötig, Analysemethoden zu finden und zu verfeinern. Müller betonte die durch die DH angefachte Dynamik, etwa bei der Handschriftenkatalogisierung. Mit den DH änderten sich auch die Forschungsgegenstände und Methoden: *distant* und *close reading* ergänzten sich. Neue, komplexere Fragestellungen seien zu entwickeln (Terrahe); dabei seien Experimente sinnvoll und auch die Publikation von Zwischenschritten und Erprobungen (Uta Goerlitz).

Das Problem der (Nicht-)Akzeptanz digitaler Publikationen wurde angesprochen (Alan van Beek); in diesem Bereich sei vieles im Fluss, seitens der Institutionen und der Fördergeber (Schubert, Müller); Schubert appellierte, auch digital publizierte Dissertationen in den Rezensionsturnus aufzunehmen.

Als ein Hauptergebnis der DH für die Altgermanistik wurden Verknüpfungen und Vernetzungen festgestellt (Michael Stolz, Hausmann, Löser), die auch neue Fragen ermöglichten. Gleichwohl wurde die vielfach mit den DH verbundene Komplexitätsreduktion bei Fragestellungen und Ergebnissen als Herausforderung wahrgenommen. Gewinn liege freilich gerade auch im Quantitativen: etwa einem möglichen Projekt der Komplett-OCR aller mittelalterlichen Handschriften (Schubert, einen Vorschlag von Torsten Schaßan aufgreifend); der stärkeren Berücksichtigung unbe-

kannter/noch unedierter Texte (Christine Glaßner); der Schnelligkeit, mit der Texte digital zur Verfügung gestellt werden können; der Möglichkeit, ›niedere‹ Arbeiten zu delegieren (Kurt Gärtner).

Phillip Brandes verwies darüber hinaus auf die Notwendigkeit, die DH stärker auch in der Lehre zu verankern; Jürgen Wolf nannte beispielhaft den Bachelorstudiengang Digital Philology der TU Darmstadt und den Masterstudiengang Cultural Data Studies (Universität Marburg).[2]

Abschließend betonte Schubert das Problem des Exitus von Daten, das Bedürfnis nach einer Clearing-Stelle für die Zukunft von Daten und die Notwendigkeit abgestimmter Normierung und Datenstandards. Hier sei vor allem das Engagement der Nationalbibliotheken gefordert.

Anmerkungen

1 Martin Schubert danke ich für Unterstützung.
2 Weitere DH-Studiengänge siehe http://www.dh-curricula.de/.

Anschrift der Berichterstatterin:

Prof. Dr. Elisabeth Lienert
Universität Bremen
Fachbereich 10
Universitäts-Boulevard 13
28359 Bremen
E-Mail: elienert@uni-bremen.de