

## **Korpusliteraturwissenschaft. Zur Konzeption und Praxis am Beispiel eines Korpus zur literarischen Moderne**

Ein Fach oder eine Teildisziplin ‚Korpusliteraturwissenschaft‘ gibt es nicht, nicht so jedenfalls, wie es Korpuslinguistik als einen Bereich der Sprachwissenschaft gibt, der in systematisierender wie historischer Absicht natürlchsprachliche Äußerungen sammelt (Bubenhofer 2009; Biber/Conrad/Reppen 2006; Lemnitzer/Zinsmeister 2015). Auch spricht man in der Literaturwissenschaft und selbst in der Literaturgeschichte nur selten von einem Quellenkorpus, wie das in historischen Fächern üblich ist, wenn unter Quellen die je nach Fragestellung systematische Zusammenstellung möglichst breit angelegter Sammlungen von Zeugnissen verstanden wird (Maurer 2002). Außerdem fehlt in der Literaturwissenschaft weitgehend die Verbindung von qualitativen und quantitativen Methoden der Textanalyse, wie sie in der Linguistik zu finden ist (Kim 2012; Meindl 2011).

Das Fehlen einer Korpusliteraturwissenschaft kann auf drei Entwicklungen in der Literaturwissenschaft zurückgeführt werden. Erstens ist die Literaturwissenschaft stark am Kanon der besonderen Werke ausgerichtet (Rosenberg 2000), was auch literaturwissenschaftliche Editionen betrifft. Weil das Besondere traditionell mehr als das Typische zählt, geht es in der Literaturwissenschaft nicht um eine evidenzbasierte Sichtung umfassender Werkcorpora. Besondere Texte werden hier vor allem als exemplarische Belege eingeschaltet oder auch in genauer Einzelanalyse untersucht, nicht aber in größerem Umfang zur Beschreibung genereller Trends oder allgemeiner Phänomene kompiliert. Hinzu kommt zweitens der Bedeutungsverlust der Literaturgeschichtsschreibung seit den 1990er Jahren, als mit dem Ende der Sozialgeschichte als leitendes Forschungsparadigma eine integrierende Fragestellung für die Literaturgeschichte an Geltung verloren hat (Huber/Lauer 2000). Die Literaturwissenschaft ist damit von den historischen Wissenschaften und ihrem Quellenverständnis abgerückt. Die dominierenden theoretischen und methodologischen Konzepte der Literaturwissenschaft brauchen kaum noch verbreiterte Text-Sammlungen, um ihre Fragen bearbeiten zu können. Die kulturwissenschaftliche Erweiterung des Fachs durch eine generellere Theorie

von Medien und Kommunikation hat zwar das Gegenstandsfeld entgrenzt. Aber diese Entgrenzung wurde nicht von einer vergleichbaren Erweiterung der Quellen abgesichert. Das liegt an der Forschungspraxis einer kulturwissenschaftlich angeleiteten Literaturwissenschaft, die sich eher mit der Entwicklung kulturphilosophischer Modelle und der eher kleinteiligen Untersuchung von Fallbeispielen befasst (Graevenitz 1999). Eine systematische Zusammenstellung von literarischen Texten über die Breite der publizierten, gelesenen und besprochenen Literatur einer Region, einer Sprache oder einer Zeit hat auch aus diesem Grund seit Längerem nicht im Zentrum der Literaturwissenschaft gestanden, wie es in älteren Konzepten der Literaturgeschichte, wie etwa in August Boeckhs epigraphischem Projekt zu den griechischen Inschriften, dem *Corpus Inscriptionum Graecarum*, geordnet nach Landschaften, noch durchaus üblich war (Vogt 1979). Schließlich ist drittens die Literaturwissenschaft fachsystematisch weit von der Linguistik weggerückt. Deren empirische Ansätze qualitativen wie quantitativen Typs spielen in der Literaturwissenschaft kaum eine Rolle. Von der strukturalistischen Einheit des Fachs, die in Zeitschriften wie *LILI. Zeitschrift für Literaturwissenschaft und Linguistik* institutionalisiert worden war, ist wenig geblieben. Nicht zuletzt hat der bestimmende antisientifische Impuls die Literaturwissenschaft weit von der stärker methodenorientierten Linguistik entfernt.

Eine an Traditionen der Philologie nach August Boeckh anknüpfende Korpusliteraturwissenschaft hätte jedoch gegenüber bloß quantitativen und IT-gesteuerten Ansätzen mehrere Vorteile: Sie stünde zunächst in einer Tradition, die qualitative, hermeneutische Vorgehensweisen pflegt, also kritische und kontextualisierende Verfahren. Die Korpusliteraturwissenschaft fände damit besonders fruchtbare Bedingungen für die Verschränkung von quantitativen und qualitativen Verfahren (Herrmann 2017). Zudem würde eine Korpusliteraturwissenschaft, die das philologische Sammeln, Katalogisieren und Verfügbarmachen von textuellen Zeugnissen zu ihrer Praxis mache, die stichhaltige Modellierung der in Anschlag gebrachten Theorien und Begriffe zentral stellen. Sie würde mit Giambattista Vico danach trachten, nicht nur eine beliebige Zahl von Einzelbefunden anzuhäufen, sondern diese im Sinne einer historischen Anthropologie zu synthetisieren, „die mannigfaltigen Aspekte der menschlichen Kultur in ihrem synchronen Nebeneinander ebenso wie in ihrer diachronen Entwicklung auf einige wenige fundamentale Grundsätze zurückzuführen“, wie Vittorio Hösle Vicos Anspruch an die Geisteswissenschaften zusammengefasst hat (Hösle 1990 CIX). Schließlich ist für die Korpusliteraturwissenschaft die Qualität der Daten ein zentrales

Anliegen. Gegenwärtig aber kann Literaturwissenschaft in ihrer allgemeinen Ausprägung nicht mit Philologie gleichgesetzt werden und ist sie – zumindest zurzeit – nicht wie die Philologie im besten Sinne peinlich genau. Die genannten Gründe erschweren die Etablierung einer Korpusliteraturwissenschaft.

In der Summe ist Korpusliteraturwissenschaft trotz möglicher Rückbezüge auf die Fachtradition (noch) kein Teil der Disziplin Literaturwissenschaft. Im Folgenden argumentieren wir dennoch, dass trotz mancher Widrigkeiten eine Korpusliteraturwissenschaft für die Disziplin eine wesentliche Erweiterung ihrer Methoden und Verbesserung ihrer Erkenntnismöglichkeiten darstellt. Wir referieren zunächst die wenigen bisherigen korpusorientierten Ansätze zur Untersuchung von Literatur und diskutieren anschließend die Anforderungen an ein repräsentatives Korpus als abstrahierendes Modell diskursiver (literarischer) Realität (Biber 1993; Rieger 1979). Am Beispiel des Korpus der Literarischen Moderne (KOLIMO) zeigen wir exemplarisch, wie ein valides Korpus aus einer Fragestellung heraus konzipiert wird.

## Korpusliteraturwissenschaft, Vorarbeiten und Ansätze

Wenn etwas in den letzten Jahren Korpusliteraturwissenschaft ermöglicht hat, dann die Digitalisierung sehr großer Bestände der Literaturgeschichte. Die Sammlung Gutenberg, die Quellen bei Wikisource, ECCO, Evans, Gallica oder die HathiTrust-Bibliothek und das derzeit wohl größte Digitalisierungsprojekt, Google Books, mit seinen mehr als 30 Millionen Bänden, sind Ressourcen für korpusliteraturwissenschaftliche Fragestellungen in einem Maßstab, der noch vor kurzem nicht vorstellbar war. Diese Art von Sammlungen wird ergänzt und erweitert durch historische oder gattungsspezifische Kollektionen, etwa zum Theater der Französischen Klassik (Fièvre 2007), zur antiken Überlieferung im Portal Perseus (Crane 2016; Almas/Beaulie 2016), das Verzeichnis Deutscher Drucke des 18. Jahrhunderts (Bürger 2008) oder die 19th Century Collection Online. Historisch breit aufgestellte Ressourcen, wie sie etwa im Deutschen Textarchiv (DTA) zusammengestellt und sprach-historisch erschlossen werden (Geyken/Gloning 2015), in der Digitalen Bibliothek bei TextGrid zu finden sind (Neuroth/Rapp/Söring 2015) oder für das 20. Jahrhundert auch literarische Texte einbinden wie im Deutschen Referenzkorpus DeReKo des Instituts für Deutsche Sprache (IDS) oder im C4-Projekt, bestehend aus österreichischen, bundesdeutschen, Südtiroler und schweizerischen Texten (Dittmann/Đurčo et al. 2012), erweitern noch einmal das Spektrum.

So eindrucksvoll die Zahlen der in solchen Kollektionen zu findenden Texte sind, nur wenige sind nach literaturwissenschaftlichen Kriterien erstellt, dabei gattungshistorisch ausbalanciert und offen zugänglich. Viele Sammlungen unterliegen vielmehr verschiedenen urheberrechtlichen und anderen Einschränkungen, so dass die Erstellung eines eigenen Korpus aus diesen Ressourcen nur beschränkt möglich ist. Im fachwissenschaftlichen Sinne repräsentative, weil kriteriengeleitet und mit Bezug auf eine angenommene Grundgesamtheit gesampelte, Korpora gibt es daher derzeit in der Literaturwissenschaft nur ansatzweise.

So wie die Literaturwissenschaft zu fragen gewohnt ist, nämlich nach möglichst originellen Erkenntnissen über oftmals kanonische Texte und nicht nach allgemeinen Regularitäten von Grundgesamtheiten, hat sie kaum Bedarf an Textkorpora. Es sei denn, man stellt Fragen quer zur disziplinären Logik der Literaturwissenschaft. Besonders prominent hat das Franco Moretti (2013) getan, ursprünglich motiviert durch die komparatistische Absicht, eine Geschichte des europäischen Romans zu schreiben. In verschiedenen Aufsätzen hat er Wege gesucht, die Evolution und Variation der Gattung des Romans zu fassen. Schon dass er nach einer formalen Regularität fragt, nämlich nach der Evolution der Form des Romans, unterscheidet seinen Ansatz von herkömmlichen Modellen der Literaturgeschichtsschreibung. Zum anderen geht es Moretti um die Definition einer Grundgesamtheit der literarischen Gattung des Romans, also um die Zielgrundgesamtheit (*'target population'*), die jeder Stichprobenwahl oder jeden Maßnahmen zu Steigerung der Abbildungsgenauigkeit (*'balancing'*) eines Korpus vorgeordnet sein muss (Biber 1993). Diese beiden Fragen haben eine Forschungsarbeit angestoßen, die für die Schätzung der Anzahl und Art der Romantexte in der europäischen Geschichte hinreichend verlässliche bibliographische Daten braucht. Zusammen mit Matthew Jockers hat Moretti am Stanford Literary Lab bald gesehen, dass die Untersuchung einer historischen Morphologie des Romans eine quantitative und abstrahierende Aufgabenstellung ist, die mit Methoden wie geographischen Abbildungen der imaginären Orte der Romanhandlung, Vergleichen von Romananfängen und Romanenden, Netzwerken von Figurenkonstellationen, Bibliographien von Übersetzungen und auch Diagrammen von bibliothekarischen Sammelpolitiken vorgehen kann und auch muss. Morettis Vorschlag betrifft so gesehen eine deskriptive, quantitativ-verallgemeinernde Literaturgeschichtsschreibung. Ihre Voraussetzung ist das Korpus, das als Untersuchungsgegenstand eine annähernd repräsentative Auswahl dessen ist, was als Grundgesamtheit angenommen werden kann.

Unabhängig von Moretti haben wir schon vor mehr als zehn Jahren versucht, ein Korpus von 350 Romanen in deutscher Sprache zu erstellen, das Teil eines historischen Referenzkorpus des Deutschen werden sollte (Jannidis/Lauer/Rapp 2005). Wesentlich für das Korpus war, dass dessen Erstellung nur dann sinnvoll ist, wenn die Fragestellungen an dieses Roman-Korpus andere als die fachlich etablierten sind. Uns ging es damals um eine Geschichte des Erzählens, um eine historische Narratologie, ein Projekt, das aus einer Reihe von fachpolitischen Gründen erst jetzt mehr als eine Dekade später wiederaufgenommen werden kann. Wie Morettis Gruppe halten wir an sozialgeschichtlichen Fragestellungen fest, die ansonsten im Fach weitgehend aufgegeben worden sind. Weil wir anders fragen, rückt die Notwendigkeit der Erstellung von Korpora in den Vordergrund. Eine aktuelle Initiative der komparatistischen Korpusliteraturwissenschaft ist die COST-Action “Distant Reading for European Literary History” ([http://www.cost.eu/COST\\_Actions/ca/CA16204](http://www.cost.eu/COST_Actions/ca/CA16204)). Auch sie zielt auf die Erstellung eines Referenzkorpus für die Literaturgeschichte des europäischen Romans. In der Motivation ist sie vergleichbar mit einem weiteren Romankorpusprojekt, dem Projekt “Text Mining the Novel” ([https://novel-tm.ca/?page\\_id=22](https://novel-tm.ca/?page_id=22)).

Die genannten Romankorpora sind differenziert nach Untergattungen und Genres und so umfangreich, dass ein traditionell vorgehender Literaturhistoriker ein solches Korpus nicht mehr alleine bearbeiten könnte. Die Datenfülle begründet allein schon zum guten Teil, warum Korpusliteraturwissenschaft ein eigener Bereich in der Literaturwissenschaft sein dürfte. Aber es geht eben auch um ein verändertes Forschungsinteresse. Moretti (2000, 57, Kursivierungen im Original) schreibt:

[L]iterary history will quickly become very different from what it is now: it will become ‘second hand’: a patchwork of other people’s research, *without a single direct textual reading*. Still ambitious, and actually even more so than before (world literature!); but the ambition is now directly proportional to the distance from the text: the more ambitious the project, the greater must the distance be.

Diesen im Kern deskriptiven Ansatz hat Moretti zunächst fast eher beiläufig als ‚distant reading‘ bezeichnet. Damit hat er, wenn auch etwas ungenau, einen Gegensatz zum ‚close reading‘ aufgemacht, dem intensiven (Wieder-)Lesen der kanonischen Texte der Literaturgeschichte. ‚Distant reading‘ beschreibt bei genauerer Betrachtung eine andere Skalierung des literaturwissenschaftlichen Gegenstandsbereichs. Nicht der Kanon, sondern die Grundgesamtheit dessen, was zu einem bestimmten Zeitpunkt und Blickwinkel als ‚Literatur‘, als

,Gothic Novel‘, ,Detektivroman‘ usw. verstanden wurde, bestimmt in diesem Feld die literaturwissenschaftliche Forschung. Die wachsenden digitalen Korpora, die diese Grundgesamtheiten abbilden, zusammen mit der Entwicklung computergestützter Methoden der Auswertung, haben distant reading zu mehr als einem Schlagwort in der Literaturwissenschaft gemacht. Festzuhalten ist: Erst wenn man die Fragestellung ändert, haben Korpora ihren Sinn. Dann erst gibt es im Fach Literaturwissenschaft auch eine Korpusliteraturwissenschaft.

In den letzten beiden Dekaden ist aus der Verbindung von Digitalisierung großer Textbestände und der Entwicklung neuer Fragestellungen eine noch überschaubare Zahl unterschiedlicher Beiträge zu einer korpusbasierten Literaturwissenschaft erwachsen (Underwood 2017; Weitin 2015). Die Gruppe in Stanford und hier besonders Matthew Jockers haben durch die Publikation von Tools und Lehrbüchern (Jockers 2013 und 2014) die Methodenentwicklung wesentlich vorangebracht und gezeigt, wie man sich mit formalen Methoden in die Lage versetzt, Gattungen automatisiert zu erkennen und in ihrer historischen Entwicklung zu unterscheiden. Die *Pamphlets* des Stanford Literary Labs gehören inzwischen zu den Referenzpapieren der Modellierung und Methodenentwicklung im Fach. Für die Methodenentwicklung werden überwiegend Testkorpora einzelner Werke wie *Sense and Sensibility* oder *Moby Dick* genutzt, die vor allem dazu dienen, die Analyseergebnisse kritisch und von Hand einzuschätzen. Die Bildung von größeren und aufwändiger stratifizierten Korpora ist dabei nachgeordnet (vgl. Algee-Hewitt/ Allison et al. 2016).

Andere Literaturhistoriker wie Andrew Piper oder Ted Underwood haben an umfangreichen Korpora, die mehrere tausend Publikationen der deutschen und englischen Literatur des 18. und 19. Jahrhunderts umfassen, untersucht, wie sich Themen innerhalb von Nationalliteraturen ausdifferenzieren, wie sich das Prestige von Gattungen wandelt oder sich Figurenkonstellationen ändern (Underwood 2016; Underwood/Sellers 2016). An mehr als tausend Bestsellern in unterschiedlichen Genres haben Piper und Portlance (2016) Regeln der kulturellen Kapitalbildung untersucht, während Archer und Jockers (2016) Regularitäten von Bestsellern algorithmisch abbilden. Auch systematische Fragestellungen, ob etwa Fiktionalität eine Eigenschaft von Texten oder nur eine Zuschreibung durch Leser ist, werden an vergleichenden Korpora von fiktionalen und nicht-fiktionalen Texten analysiert. Piper (2016) konnte zeigen, dass sich fiktionale Texte stilistisch von nicht-fiktionalen Texten unterscheiden, Fiktionalität daher eine sprachliche Eigenschaft von Texten sei. Fragen zu Nationalliteraturen werden anhand von Korpora der irisch-amerikanischen Literatur untersucht, die quantitative Auswertungen von literaturhistorischen

Prozessen der Amerikanisierung der irischen Literatur erlauben (Jockers 2013). Die Korpusgröße und -zusammensetzung sind dabei von der jeweiligen Fragestellung abhängig und variieren in den Untersuchungen entsprechend. So hat Thomas Weitin eine historisch prominente Novellensammlung ausgewählt, die von Paul Heyse und Hermann Kurz zwischen 1871 und 1876 herausgegebene Sammlung *Deutscher Novellenschatz*. Das Korpus ist eines der eher spezialisierten Korpora, die an einem historisch überschaubaren Umfang die Exploration korpusliteraturwissenschaftlicher Methoden erlauben (Weitin 2017). Mit kleineren Korpora arbeitet traditionell auch die Korpusstilistik, zum Beispiel autororientiert wie etwa Michaela Mahlbergs Dickens-Studien (Mahlberg 2013), werkorientiert wie Michael Stubbs Analysen von Joseph Conrads *Heart of Darkness* (Stubbs 2005) oder schon breiter angelegt wie die narratologische Studie zu Redewiedergabe von Elena Semino und Mick Short (Semino/Short 2004).

Obwohl es literaturwissenschaftliche Fragestellungen gibt, die globale literarische Referenzkorpora nach Art der linguistischen Referenzkorpora wie des British National Corpus oder DeReKo benötigen, hat deren Entwicklung gerade erst begonnen. Dies hat auch mit der genuin literaturwissenschaftlichen Dimension der Korpusliteraturwissenschaft zu tun, die sich im Gegenstand, etwa mit Blick auf Form und Entwicklung bestimmter literarischer Gattungen, aber auch im Zugang, hier besonders im Kontextualisieren, zeigt. In der Stilometrie etwa ist die Untersuchung von exemplarischen, auf die jeweilige Problemstellung der Methode zugeschnittenen Korpora das gängige Verfahren. Auf den Spuren von John Burrows (2002, 2003) bahnbrechenden Arbeiten zur literarischen Stilometrie und Autorschaftsbestimmung sind verschiedene kleinere Modellkorpora zusammengestellt worden. Zu diesen gehören etwa Hugh Craigs Korpora zur Autorschaftsfrage der Dramen Shakespeares oder zur Dialogführung in verschiedenen Theaterepochen (Burrows/Craig 1994; Craig 2012; Craig/Greatley-Hirsch 2017), und darüber hinaus die Korpora, die Fotis Jannidis und Gerhard Lauer zur Unterscheidung von literarischen Epochen und von Autorinnen und Autoren erstellt haben (Jannidis/Lauer 2014), ähnlich dann auch Christof Schöch zur Gattungsbestimmung des Dramas in der Französischen Klassik (Schöch 2014). Der Erfolg von Burrows' statistischer Methodik der Delta-Maße hat wesentlich dazu beigetragen, dass sich die Forderung nach einer Vergleichbarkeit der zugrundeliegenden Korpusdaten aufdrängt. Die auf der freien Statistiksoftware R basierenden Stilometrie-Skripte von Maciej Eder, Jan Rybicki und Mike Kestemont (2016) haben die Methodenentwicklung in der stilometrischen Korpusliteratur-

wissenschaft noch einmal verdichtet und inzwischen kritische Bewertungen statistischer Stilometrie-Maße ermöglicht (Evert/Proisl et al. 2017). Dies gilt auch in Bezug auf die Frage, wie umfangreich Untersuchungskorpora (und enthaltene Einzeltexte) sein müssen, um valide untersucht werden zu können (Eder 2010). Die hierbei verwendeten Korpora sind ihrem Umfang nach in der Regel klein und dienten ursprünglich der Autorschaftsbestimmung, auch in Übersetzung (Rybicki 2012), oder zur Identifizierung von Kopistenhänden in mittelalterlichen Manuskripten (Kestemont/van Dalen-Oskam 2009). Ein Referenzkorpus für diese und ähnliche stilometrischen Untersuchungen bildet sich gerade erst heraus.

Ein weiteres Feld der Korpusliteraturwissenschaft sind die Ansätze, die Topic Modeling zur Identifizierung von thematischen Strukturen, Sentimentanalyse zur Ermittlung von emotionalen Verläufen in literarischen Texten und Netzwerkanalysen zur Bestimmung von Mustern der Figurenkonstellationen nutzen, um Literaturgeschichte anders als bisher zu modellieren. Auch hier sind die zugrundeliegenden Textsammlungen eher Samples als vollständige Korpora. Ausnahmen sind etwa Matt Erlins Literaturgeschichte der Erzählungen von 1731-1864, einer Verbindung von Topic Modeling und Netzwerkanalyse (Erlin 2014), oder Mariona Coll Ardanuys und Caroline Sporleders Untersuchung der Figuren-Netzwerke in mehreren hundert englischen Romanen des 19. Jahrhunderts (Ardanuy/Sporleder 2015). Zu nennen ist hier auch Christof Schöchs Untersuchung zur thematischen Entwicklung von Untergattungen in der Geschichte des französischen Dramas zwischen 1610 und 1810, der ein historisch repräsentatives Korpus von mehr als 800 Dramen zugrunde liegt (Schöch 2017). Eine andere Repräsentativität haben Andrew Reagan und das Computational Story Lab für ihre Sentimentanalysen angestrebt, die untersuchten, welche fiktionalen Texte derzeit bei Gutenberg am häufigsten heruntergeladen werden. Das resultierende Korpus von mehr als tausend Büchern typisierten sie auf Emotionsmuster und folgerten, dass sechs Plotline-Haupttypen das Interesse der heutigen Leser bestimmten (Reagan/Mitchell et al. 2016; Archer/Jockers 2016). Insgesamt ist die Forschungslandschaft korpusbasierter Literaturwissenschaft ebenso klein wie methodisch homogen. Bezogen auf die genutzten Korpora ist sie aber höchst divers. Dies schränkt die Vergleichbarkeit von Ergebnissen ein, denn schon allein die unterschiedliche Morphologie der Sprachen Englisch, Französisch oder Polnisch erfordert den Einsatz differenzierter Analyseparameter.

Das alles ist Korpusliteraturwissenschaft, aber mit einer Reihe von Einschränkungen. Diese betreffen die Forschungspraxis in Bezug auf Korpus-

sampling, die editorische Qualität der Daten, aber auch die theoretische Modellierung und den interpretatorischen Zugriff. Das Sampling der Texte orientiert sich vielfach vor allem am pragmatischen Kriterium der Zugänglichkeit bereits digitalisierter Texte (eine anschauliche Beschreibung des vergeblichen Versuchs eine repräsentative Zufallsstichprobe für anglophone Literatur für die Periode 1750-1880 zu erlangen geben Algee-Hewitt/Allison et al. 2016). In den Hintergrund geraten in vielen Studien dabei theoretisch relevante Kriterien wie die historische Repräsentativität und die Ausgewogenheit zwischen verschiedenen Gattungen und Untergattungen, Autorengruppen oder auch Rezipientengruppen. Dies führt paradoxe Weise dazu, dass viele computerlinguistische Beiträge ihre Korpora anhand von Kanon-Listen (wie etwa der *ZEIT-Bibliothek der 100 Bücher*, Raddatz 2009) kompilieren, die eine normativ-wertende Perspektive auf Literatur oft unkommentiert reproduzieren. Weil eine historisch-philologische Reflexion und Praxis der Korpusbildung oftmals nur ein Randthema ist, laufen daher viele Studien Gefahr, ihren Gegenstand als extern gegeben zu reifizieren. Insbesondere wenn zusätzlich auf eine gründliche literaturhistorische Interpretation der Ergebnisse verzichtet wird, kritisiert die historisch orientierte Literaturwissenschaft mit guten Gründen manche der Korpora und die mit ihnen verknüpften Ansätze als philologisch unzuverlässig (Bode 2017). Der deutlich höhere forschungspraktische Aufwand einer evidenzbasierten Literaturwissenschaft kann in solchen Fällen nur zu oft nicht hinreichend gerechtfertigt werden.

Neben den entstehenden Korpora hat sich in den letzten Jahrzehnten in der Literaturwissenschaft ein noch überschaubares Panorama von korpusliteraturwissenschaftlichen Ansätzen entfaltet. Zu nennen sind hier vor allem Beiträge aus den Digital Humanities (Jockers/Underwood 2015; Meister 2013), Beiträge der gebrauchsorientierten Korpuslinguistik mit ihrem registerübergreifenden Ansatz (Biber/Johansson et al. 1999) und Korpusstilistik (Mahlberg 2015) sowie Beiträge der strukturalistischen Stilistik (Herrmann/van Dalen-Oskam/Schöch 2015), neben Ausgriffen anderer Fächer wie der Sozialpsychologie und ihrem Interesse an der Psychologie des Schreibens (Pennebaker/Ireland 2011).

Korpusliteraturwissenschaft ist eng an die Entwicklung neuer Methoden des Fachs gebunden. Im Vordergrund steht bislang nicht das Korpus, sondern die Methodenentwicklung. Weil aber Korpus und Methode eng zusammenhängen, lohnt es sich, die Kriterien für die Erstellung eines literaturwissenschaftlichen Korpus nicht nur von der Methodenseite her zu entwickeln, sondern die theoretische Modellierung des Korpus nach vorne zu stellen.

## Korpusliteraturwissenschaft: Praktische Kriterien

Die Anforderungen an Korpora für Zwecke der Literaturwissenschaft sind nicht deckungsgleich mit denen an Korpora für Zwecke der Linguistik. Aber viele Kriterien, die an wissenschaftliche Korpora angelegt sein müssen, gelten für beide Bereiche (siehe z. B. Liu 2016). Die Unterschiede wie die Gemeinsamkeiten festzuhalten, ermöglicht es, eine kritische Prüfliste für ein literaturwissenschaftlich valides Korpus erstellen. Eine erste Gruppe von Kriterien ist mit dem Begriff der Repräsentativität eines Korpus überschrieben, wobei proportional so viele Vertreter bestimmter Texte gesammelt werden, dass die gesamte Bandbreite der Variabilität der Grundgesamtheit umfasst ist: „Representativeness refers to the extent to which a sample includes the full range of variability in a population“ (Biber 1993, 243). Zunächst scheint es für literaturwissenschaftliche Korpora vergleichsweise leicht, die jeweilige Grundgesamtheit zu bestimmen. Im Unterschied zum Gegenstand ‚Sprache‘ mit seinen historischen und gegenwärtigen Varianten mündlicher und geschriebener Art ist die literarische Sprache erst einmal eine überschaubare Teilmenge. Für viele Literaturen und historische Abschnitte der Literaturgeschichte haben wir Zugriff auf die jeweiligen Publikationszahlen - zumal es sich bei Literatur in überwältigender Mehrheit um schriftlich fixierte und in der Regel dann auch publizierte Werke handelt. In das Korpus kann dann auch der Teil des Archivs aufgenommen werden, der zur Beantwortung einer spezifischen Fragestellung benötigt wird. Zur Beziehung zwischen Grundgesamtheit und Korpus prognostiziert das Literary Lab (Algee-Hewitt/Allison et al. 2016) recht zuversichtlich eine Konvergenz der Ebenen des Publizierten, des davon Erhaltenen und des Korpus: Das Korpus ist dann keine indirekte Modellierung, sondern die direkte Repräsentation der tatsächlich publizierten Literatur.

Befragt man Bibliographien und Katalogdaten, Archiv- und Buchhandelskataloge, Auktions- und Messkataloge, so geben sie vergleichsweise verlässlich Auskunft über die Menge des Publizierten und Erhaltenen. Natürlich gibt es auch Forschungsfelder wie etwa die mündliche Überlieferung von Märchen, die eine Ermittlung der Grundgesamtheit kaum erlauben. Wenn man sich aber auf das Erkenntnisinteresse der Literaturwissenschaft bezüglich Textsammlungen bezieht, mit den Leitkategorien der Gattung und des Autors/der Autorin, so lassen sich die Grundgesamtheiten für viele Epochen vergleichsweise genau bestimmen. Zu fordern, dass ein Korpus nicht nur eine repräsentative Stichprobe, sondern eine möglichst vollständige Repräsentation

einer Population von Texten sein müsse, scheint unter dieser Perspektive eine begründete Anforderung. Im Feld der Literaturwissenschaft können Korpora – so gesehen – eine repräsentative Zusammenstellung aus verschiedenen Gattungen, Genres und Registern sein, ein Remix der Literatur (Liu 2016).

In der Praxis ist die Bestimmung der Grundgesamtheit für die (deutschsprachige) Literatur aber schwieriger, und zwar in mindestens zweifacher Hinsicht: Zum einen ist zur statistischen Bestimmung eine Abschätzung der Gesamtheit der publizierten Titel samt weiterer Angaben zu Autorschaft, Publikationsdatum und Ort sowie Ausgabe und Gattungsinformationen notwendig. Für den deutschsprachigen Bereich nach 1913 ist die Deutsche Nationalbibliothek ein Vorzeigeprojekt, deren Katalogdaten per API-Schnittstelle oder auch direkt durchsuchbar sind (Fischer/Jäschke 2018). Allerdings offenbart sie bei automatisiertem Zugriff Inkonsistenzen in den Metadaten (etwa Tippfehler, nichtstandardisierte Datumsangaben und Ortsangaben). Das größte Manko ist jedoch, dass Daten vor 1913 nicht valide aufgenommen wurden, was am späten Einsatz des ‚Unternehmens Nationalbibliothek‘ liegt und der ganzen anderen Sammelpolitik im Vergleich etwa mit der Bibliothèque Nationale de France (1666) oder der Library of Congress (1800). Für andere Nationalsprachen als das Deutsche ist die Situation anders, je nach Gründungsgeschichte der jeweiligen Nationalbibliothek. Für die deutschsprachige Literatur vor 1913 gibt es damit derzeit keine eindeutige Quelle zu Publikationszahlen, sondern ein Sammelsurium verschiedener Kataloge, Bibliographien etc. Die Grundgesamtheit der literarischen Texte, in ihren Gattungen, Autorschaften und nationalen Zuordnungen für die Epochen vor 1913 zu bestimmen, ist deshalb ein Desiderat der quantitativen Literaturwissenschaft.

Die Grundgesamtheit fordert ihre genaue Bestimmung auch aus einem zweiten Grund heraus: Man muss entscheiden, ob hier die tatsächliche historische Rezeption, wie sie die Geschichte des Lesens zu rekonstruieren versucht, oder die druckgeschichtlichen Daten gemeint sind. Nicht jeder Erstdruck ist auch der tatsächlich gelesene, was unter anderem daran zu erkennen ist, dass Goethes *Wahlverwandtschaften* im Erstdruck noch im 20. Jahrhundert vielfach in Antiquariaten unaufgeschnitten gekauft werden konnten. Soll also die Rezeption – und eben nicht die Publikation – abgebildet werden, muss das Korpusampling gegebenenfalls auf Zahlen aus Leihbibliotheken oder die Sichtung von Lesezeugnissen (etwa in Tagebüchern und Briefen) zurückgreifen.

Offen ist letztlich auch die Frage, wie groß eine Stichprobe bezogen auf die Grundgesamtheit sein muss, um als repräsentativ gelten zu können. Der

gängige stratifikatorische Lösungsansatz der Korpuslinguistik liefert hier einen Ansatz, indem er die aus inferenzstatistischen Gründen geforderte randomisierte Zufallsstichprobe erst innerhalb vordefinierter Gruppen (‘strata’) wie etwa ‚weibliche Autoren, nach 1860 geboren‘ ansetzt. So kann man sicher gehen, dass das durch das Korpus dargestellte Modell der Population keine zu starke statistische Schieflage (‘skewness’) zeigt. Wie allerdings stratifiziert wird und wie groß die Stichproben sind, muss aus literaturwissenschaftlichen Fragestellungen heraus begründet werden

Anstelle des gängig gewordenen Gegensatzes von close vs. distant reading wird daher auch mit einer Formulierung von Martin Mueller von ‚scalable reading‘ gesprochen (Mueller 2012). Gemeint ist damit, dass nur in Abhängigkeit von der jeweiligen Forschungsfrage die Größe des einer Untersuchung zugrunde gelegten Korpus und der Grad der Abstraktion, das Forschungsdesign und die Forschungsmethodik bestimmt werden können. Eine Untersuchung zur überschaubaren Geschichte des Barockromans beispielsweise hat mit einem anderen Mengengerüst zu arbeiten als eine Untersuchung zum Erfolgsmodell ‚Novelle im 19. Jahrhundert‘. Was aber die Wahl der Methodik angeht, beinhaltet eine Skalierung vor allem, den Grad der Abstraktion und der Kontextualisierung zu modulieren, wie etwa für eine pragmatische Korpuslinguistik vorgeschlagen von Marcus Müller (2011). Manche Fragen an die Novelle, etwa solche zur Verwendung von Modalwörtern in Strategien der Inferenzsteuerung, sollten demnach am besten in explorativen Analysen mit hohem textuellen Kontextualisierungsgrad gestellt werden (Herrmann 2016).

Repräsentativität ist dann noch aus einem anderen Grund schwieriger zu bestimmen, als es zunächst scheint. Eine Grundgesamtheit ist selbst auch ein Konstrukt, denn Daten haben eine bestimmte Granularität, je nach Beobachter und Grund der Beobachtung (Giere 1999). Repräsentativität kann daher bezogen auf die Grundgesamtheit immer nur in einer bestimmten Perspektive behauptet werden. Und schließlich ist die Schaffung von Repräsentativität noch ganz profan beeinflusst vom Urheberrecht, das sich gerade für die Untersuchung der neueren Literaturgeschichte einschränkend auswirkt.

Ein weiteres Kriterium der Wohlgeformtheit von Korpora betrifft die Abbildgenauigkeit der Korpusdaten, die üblicherweise unter dem Begriff des balancierten Korpus verhandelt wird (Lemnitzer/Zinsmeister 2015, vgl. Atkins/Clear/Ostler 1992). Genauigkeit ist abhängig von der jeweiligen Fragestellung, historische Genauigkeit kann sich auf so unterschiedliche Dimensionen wie die Wirkungsgeschichte, auf die produktive Rezeption durch Autoren, den Einfluss oder die Konstellationen von Autoren bezie-

hen. Forschungsvorhaben, die die Wirkungsgeschichte von Goethes Lyrik oder die besondere Rolle Kleists untersuchen, brauchen jeweils Daten, die in unterschiedlicher Richtung genau sind, etwa in der Erfassung von Metaphern der Naturerfahrung oder in der Vergleichbarkeit von Dramenformen. Sind akribisch erstellte Editionen notwendig - oder genügen Texte ohne Überlieferungs- und Entstehungsvarianten? Genaugigkeit kann sich auf die kulturelle, soziale und historische Variabilität von Gattungen beziehen oder auf systematische Fragen etwa des Fiktionalitätsstatus von Texten. Was etwa einen Roman definiert, ist um 1700 anders klassifiziert worden als um 1850. Wie kann die Entstehung neuer Gattungen wie etwa des Kriminalromans überhaupt textsortengenau erfasst werden, wenn diese Gattung erst im Entstehen begriffen ist? Ein balanciertes Korpus kann sich auch auf die klassifikatorische Bestimmung von Autorschaft beziehen oder auf die deskriptive Unterscheidung zwischen Autorinnen und Autoren, auf die Verschiedenheit der Bildungszügänge oder der Konfession. Ein Korpus kann als balanciert gelten, wenn es die Besonderheiten von Epochen, Gruppen oder Bewegungen, Zuordnung zu politischen oder kulturellen Räumen einbezieht, immer vorausgesetzt, dass dies Forschungsfragen sind, die an ein Korpus gerichtet sind. Je nach Fragestellung können auch die Unterschiede der Sprache, Dialekte und Soziolækte und andere Register der Literatur in einem Korpus abgebildet sein.

Eine andere Gruppe von Kriterien für den Bau eines literaturwissenschaftlichen Korpus hat mit der jeweils vorausgesetzten Text-, wohl auch Zeichen- und Literaturtheorie zu tun, mit den Vorannahmen über Textualität und Intertextualität, über Historizität, Literarizität und Systematik. Ob ein Korpus aus Fragmenten in Textfiles bestehen soll oder aus Werken als ästhetischen Einheiten, ist ausschlaggebend (Winko 2008). Die theoretische Klärung schließt auch die Prüfung ein, ob das Korpus und die für die Fragestellung gewählten Indikatoren (oder ‚features‘) zusammenpassen. Sind die Indikatoren beispielsweise auf der Satzebene angesiedelt, muss das Korpus so strukturiert sein, dass Sätze als Einheiten in der Analyse überhaupt erkannt werden können. Metadaten können Angaben über die Lebenszeit von Autoren, den Publikationsort oder die Gattung enthalten. Wie viele solcher und ähnlicher Kontextinformationen in die Strukturierung der Korpusdaten eingehen, hängt an einer Reihe theoretischer Vorannahmen. Ob eine Untersuchung mit Volltexten arbeitet oder etwa mit Anfangs- oder Endpassagen von Werken, ob die Wörter ungeordnet als sogenannte ‚bag of words‘ analysiert werden oder ihrer Wortstellung nach, ob diplomatische Umschriften und textgenaue Editionen zugrunde gelegt sein sollen oder auch modernisierte Textausgaben

genutzt werden können, fällt ebenfalls unter die Vorannahmen, die die Korpusbildung anleiten. Eine Klärung der eigenen Vorannahmen ist daher für das systematische und formal genaue Vorgehen in der Korpusbildung auch in der Literaturwissenschaft notwendig.

Ein weiteres Kriterium betrifft Übersetzungen, denn ein literarischer Diskurs profitiert stark von Übersetzungen und Nachahmung anderer Literaturen. Der Vorbildcharakter der französischen Literatur für die deutsche ist dabei nur eines der vielen Beispiele. Obwohl Übersetzungen selten eine Rolle bei der Erstellung von Korpora spielen, sind sie für viele literaturwissenschaftliche Fragestellungen jedoch wesentlich. Gezielt Übersetzungen in ein Korpus aufzunehmen ist daher für viele, aber nicht alle, Fragestellungen des Fachs erkenntnisfördernd.

Schließlich kann ein Korpus nur dann wissenschaftlich genannt werden, wenn es transparent dokumentiert und für weitere Forschung zugänglich ist. Dazu müssen Korpora die unvermeidlichen Kompromisse bei der Datenerstellung und ihre technischen Standards ausweisen. Das Korpus muss zudem eindeutig zitierbar und sollte nachhaltig und zugänglich archiviert sein, um die Nachnutzung zu fördern. Ein Korpus, das schon in der Web-Präsenz auch Anreicherungen, Qualitätskontrolle und Kommentierungen ermöglicht, verbessert mit einiger Gewissheit seine kritische Evaluierung und auch seine Validität (siehe die „FAIR“-principles, <https://www.force11.org/group/fairgroup/fairprinciples>). Nicht zuletzt kostet die Erstellung von Korpora Zeit und Geld. Die gemeinsame Nutzung von Daten hat in der Literaturwissenschaft wahrscheinlich nirgends mehr Sinn als im Umgang mit Korpora. Aus verschiedenen Korpora neue Daten gewinnen zu können, wäre daher eine kleine Revolutionierung der literaturwissenschaftlichen Forschung.

## Korpusliteraturwissenschaft in Anwendung

Ein Beispiel für ein literaturwissenschaftliches Korpus ist KOLIMO, das Korpus der Literarischen MODerne (<https://www.kolimo.uni-goettingen.de>). Es ist neben wenigen anderen wohl das erste literaturwissenschaftliche Korpus, das im Paradigma des distant reading philologischen Ansprüchen sowohl an Details wie an Synthese gerecht wird. Der substantielle Aufwand, der auf technischer und auf theoretisch-modellierender Ebene für den Korpusbau betrieben werden muss, hat ein doppeltes Ziel. Es liegt erstens darin, Schwächen der werk- und kanonorientierten Literaturwissenschaft zu überwinden, zweitens aber eben auch den Problemen der rein quantitativen Ansätze zu begegnen.

So werden einerseits Befunde über textuell belegbare Merkmale von Gattung, Epoche und Autor im Sinne des distant reading auf einer breiten Datenbasis überprüfbar und bislang unbeobachtete literaturhistorische Muster können im Sinne der Data Science (Underwood 2017) oder auch der explorierenden Korpusstylistik (Mahlberg 2015) aufgedeckt werden. Andererseits bleiben aggregierte Ergebnisse aber nicht dekontextualisiert, denn schon die Bauweise der Ressourcebettet sie in ein sorgfältig erarbeitetes literaturhistorisches wie literaturtheoretisches Modell ein. Bei der Konzeption des Korpus ist neben der theoretischen Modellierung der Grundsatz handlungsleitend, eine Ressource für transparente und überprüfbare Analysen zu schaffen. Praktisch bedeutet dies eine systematische Dokumentation der technischen wie der philologischen Dimensionen des Projekts und die möglichst freie Publikation der Ressource im Sinne der Open Data Science.

KOLIMO ist ein Korpus von narrativen fiktionalen Erzähltexten, das die literarische Epoche der Moderne repräsentativ abbilden soll. Ausgehend von einer Operationalisierung von Stil als quantitativ beschreibbarer Eigenschaft von Texten (Herrmann/van Dalen-Oskam/Schöch 2015) dient KOLIMO zur vergleichenden textbasierten Untersuchung der Moderne als Epoche der Neueren deutschen Literatur. Von besonderem Interesse sind der programmatische Anspruch der Literatur um 1900 auf Modernität und damit gerade auch Übergangsphänomene zwischen Realismus und Moderne (Baßler 2015; Wünsch 2007). Aussagekräftige sprachliche Merkmale sollen auf ihre Verteilung über Autoren, narrative Sub-Gattungen, aber auch Schulen und geographische und nationalterritoriale Topographien untersucht werden. Eine solche Analyse der Moderne in systematischer wie historischer Hinsicht ist das leitende korpusliteraturwissenschaftliche Interesse an KOLIMO.

Sein Grundstock besteht aus frei verfügbaren deutschsprachigen Texten, die einer narrativen Gattung zugeordnet und die in der Regel klar entweder literarischer oder nichtliterarischer Kommunikation zugewiesen werden können. Es sind dies Texte aus dem Deutschen Textarchiv (DTA), aus TextGrid und dem Projekt Gutenberg sowie prospektiv auch eine wachsende Anzahl von Retrodigitalisaten. Insgesamt wurden in das Korpus bislang 42.694 Texte von 2.029 Autorinnen und Autoren mit insgesamt über 558 Mio. Tokens (Wörtern) importiert (s. Abbildung 1). Das Korpus umfasst ein möglichst breites Spektrum der deutschen Literatur der zweiten Hälfte des 19. und des frühen 20. Jahrhunderts, verteilt über kanonische und nichtkanonische Texte. Ungefähr ein Fünftel der Texte fällt derzeit in die Kategorie ‚Gebrauchstexte‘,

um Vergleiche von literarischen mit nichtliterarischen Texten zu ermöglichen. Ungefähr 34.000 Texte sind als ‚literarisch‘ zu klassifizieren.

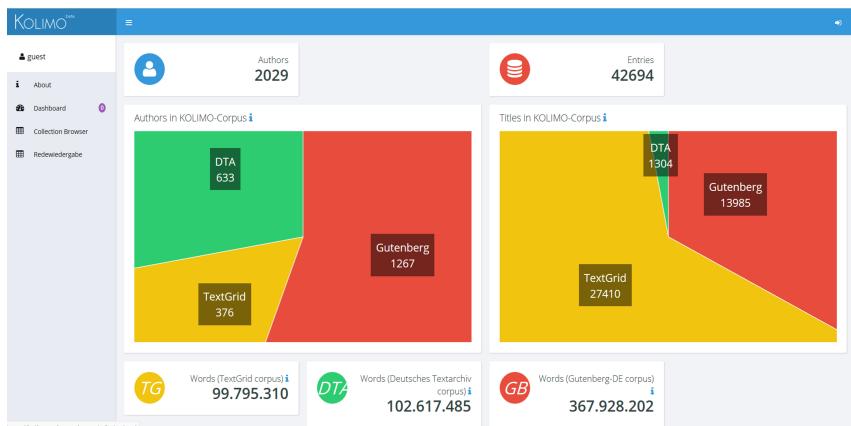


Abbildung 1: Screenshot KOLIMO Dashboard (Stand Januar 2018)

Wie Abbildung 1 zeigt, entfällt der größte Anteil auf das Projekt Gutenberg mit fast 368 Mio. Tokens, während TextGrid und das DTA jeweils mit etwa 100 Mio. Tokens zum Korpus beitragen. Diese Zahlen bilden einen maximalen Zugriff ab, enthalten also multiple Versionen von Texten. Die geschätzte Gesamtzahl von expliziten und impliziten ‚Dubletten‘ liegt derzeit bei ca. 5.000 Texten, was eine geschätzte Gesamtzahl von etwa 38.000 unikalen Texten ergibt. Die Indizierung der Versionen ist einer der nächsten Schritte in der Korpusaufbereitung. Auf ihn folgt eine abschließende Überprüfung des Gattungsschemas.

## Sampling

Die Epoche der ‚Moderne‘ wird zunächst operationalisiert als Texte mit einer Erstpublikation zwischen 1880 und 1930. Die Epoche des ‚Realismus‘, die zu Vergleichszwecken kontrastiv mitgesampelt wird, umfasst Texte aus der Zeit zwischen 1850 und 1880. Dass dieses Verfahren nicht *a priori* allen denkbaren oder auch richtigen Zuschreibungen gerecht werden kann, ist deutlich – die Problematik wird jedoch verringert durch die transparente Dokumentation und die grundsätzliche Flexibilität der Einteilung. Die Variabilität und Vielfalt der Texte und Schreibweisen wird repräsentativ abgebildet durch ein strati-

fiziertes Sampling, das eine große Stichprobe über Autoren und Autorinnen, kanonische/populäre Texte, aber auch narrative Untergattungen versammelt.

Das Sampling wurde durch zwei externe Ressourcen validiert, die Nennung der Autorinnen und Autoren in einschlägigen Literaturgeschichten sowie durch Experten-Ratings für Titel (Herrmann/Lauer 2016 a, b).

## Datenbank und Format

Um philologischen Ansprüchen an den editorischen Status literarischer Texte und die Abbildung von Epochen sowie Gattungskonzepten zu genügen, ist eine hohe Genauigkeit und Konsistenz bei der informatischen Vorverarbeitung besonders wichtig. Nicht minder wichtig ist ein hoher Grad an Flexibilität, den KOLIMO als *Stand-Off* Korpus in Form einer TEI-XML Datenbank mit Web-Applikation (*eXist*; Siegel/Retter 2014) umgesetzt hat. XML ist das gängige Datenbankformat in den Korpuswissenschaften, mit dem offenen Standard der Text Encoding Initiative (TEI), der den Vergleich und Austausch von Daten auf der Ebene von Headern, Volltexten (*text bodies*) und auch Annotationen garantiert – eben auch für spezifisch literaturwissenschaftliche Metadaten und Annotationen. Um Austausch und Kompatibilität zu gewährleisten, nutzt KOLIMO eine Adaption des DTA-Basisformats (DTABf, <http://www.deutschestextarchiv.de/doku/basisformat/einfuehrung.html>), einer Spezifikation der P5-Richtlinien der TEI. Die flexible und zugleich anschlussfähige Architektur der Datenbank erlaubt die Versionierung und den Austausch der Daten. Auf der Ebene der Volltexte ist zum Beispiel die Aufnahme eines Gebrauchstext-Samples mit Kafkas *Amtlichen Schriften* ein nächster Schritt für einen sinnvollen stilistischen Vergleich („Ähnelt Kafkas literarische Prosa im Vergleich mit anderen Autoren und Autorinnen stärker einem formalen Gebrauchstext? Wenn ja, inwiefern?“).

Die XML-Datenbank erlaubt auch, dass auf der Ebene der bestehenden Metadaten-Schemata bei bereits aufgenommenen Werken *missing entries*, also bislang leere Felder bei Erscheinungsdatum oder Gattung, ergänzt werden. Auf der Ebene der Annotationen können bestehende Auszeichnungen wie die Wortarten fortlaufend überarbeitet werden, und auch neue Ebenen wie Formen der Redewiedergabe lassen sich zusätzlich auftragen. Schließlich kann hier flexibel mit der Veränderung von Metadaten-Ontologien und Annotationstaxonomien umgegangen werden. Solche Veränderungen und Ergänzungen der Datenstruktur wären in SQL-Datenbanken, die wiederum in der Regel schneller sind als ihre XML-Pendants, nicht so flexibel handhabbar.

Die Hauptaufgabe der informatischen Dimension des Projektes besteht in der Implementierung eines praktikablen und anschlussfähigen Workflows mit Volltext-Einspeisung (*ingest*) und Transformation der gegebenen Volltexte und Metadaten in ein einheitliches TEI XML-Format. Einige Schritte der Korpusauszeichnung wurden durch ein internes eXist-Webinterface umgesetzt. Das Korpus wird auf einem eigenen Server mit standardisierten Datenschnittstellen sowie ein Datenbankabbild (*nonpublic*) mit regelmäßigen Backups (auch der Korpus-Versionen) zur Langzeitarchivierung veröffentlicht.

### **Versionen und Dokumentation**

Dass sich die von uns einbezogenen Online-Repositorien in der editionsphilologischen Textqualität deutlich unterscheiden, ist aus philologischer Perspektive problematisch. Die Qualität der Editorik, aber auch die der informatischen Vorverarbeitung der Ausgangsdaten mit Tokenisierung, Lemmatisierung sowie Normalisierung, ist besonders hoch beim DTA, das eine fortlaufende Qualitätskontrolle (s. DTAQ <http://www.deutschestextarchiv.de/doku/dtaq>) einsetzt. Dagegen legen Repositorien wie das Projekt Gutenberg mit offenen Crowd-Sourcing-Verfahren weniger klare Standards an. Der qualitativen Heterogenität der Ausgangsdaten begegnen wir pragmatisch durch die Wahl der jeweils besten verfügbaren Ausgabe und dem Ziel, die Fehlermarge bei Transkriptions- und Tokenisierungsfehlern unter dem statistisch vertretbaren Wert von zwei Prozent zu halten. Zudem erlaubt unser Korpus mit seiner flexiblen Architektur fortlaufende Verbesserungen und die weitere Anreicherung mit qualitativ höherwertigen oder auch einfach anderen Versionen von Werken. Durch die nahtlose Dokumentation wird dabei die nötige Transparenz zur Versionenkontrolle gewährleistet (momentan auf <https://gitlab.gwdg.de/kolimo>). Korpusanalysen können (und sollten) so jeweils genau angeben, auf welchen Daten innerhalb des Korpus sie beruhen. Durch die Entscheidung, im Zweifel mehrere Versionen zugleich im Korpus abzubilden, werden zudem Fragestellungen zum Vergleich von Editionen wie etwa der Kritischen Ausgabe Franz Kafkas mit Brods Edition möglich.

### **Metadaten**

Die Metadatenmodellierung ist eine wichtige Schnittstelle zwischen der informatischen und philologischen Dimension einer Korpusliteraturwissenschaft. Durch Metadaten modellieren wir die Variablen unserer Analyse in das Korpus hinein (vgl. zur angewandten literaturwissenschaftlichen Modellierung auch

Piper 2017). Die Arbeit an Metadaten ist neben dem Textsampling der Ort, an dem durch Kontextualisierung das Studienobjekt ‚Korpus‘ geformt wird, und zwar im besten Falle theoriegeleitet. In der Forschungspraxis involviert dies zudem wichtige Arbeitsschritte zu Formaten, Schemata, aber auch an fehlenden oder fehlerhaften Einträgen. In KOLIMO haben wir etwa das Problem fehlender Publikationsdaten nach Aufnahme der Quellrepositoryn durch eine heuristische Arithmetik auf der Grundlage der Lebensdaten der Autoren zu lösen versucht. Lebensdaten stehen durch die Einbindung der *Gemeinsamen Normdateien* (GNDs) der Bibliothekskataloge zur Verfügung, die wir manuell am Einzelfall validiert haben.

Die typischen Metadaten-Elemente der Korpuslinguistik können auch für literaturwissenschaftliche Fragestellungen fruchtbar gemacht werden. Im KOLIMO-TEI-XML-Header sind zum Ziele der Autorstilistik zunächst die Elemente ‚Autor‘, ‚Name‘ und ‚Gender‘ enthalten. Unabdingbar für unsere Forschungsfrage ist die Kontextualisierung temporaler Art durch das ‚Erscheinungsdatum‘ des Einzeltextes und das ‚Geburtsdatum‘ der Autorin bzw. des Autors. In KOLIMO werden diese durch das literaturwissenschaftliche Metadatum ‚Gattung‘ ergänzt, wobei die ursprünglichen Gattungsetiketten aus den verwendeten Repositoryn abgebildet und durch eigene Operationalisierungen ergänzt werden. Dazu gehören eine dreistufige Einteilung der Textlänge („kurz“, „mittel“, „lang“) und die binäre Kategorie ‚Literatur‘/„Nichtliteratur“.

Geplant ist ein valides Gattungsmetadatenschema, das die deduktive und induktive Gattungszuordnung korreliert. Induktiv werden zwei Arten der Selbstbeschreibung, Kookkurrenzprofile von Features (durch Topic Modeling, Wortartenprofile, Lesbarkeitsmaße) und paratextuelle Angaben („Roman“, „Erzählung“ etc.) berücksichtigt, auf die die deduktive ‚Fremdbeschreibung‘ aus den Repositoryn („Belletristik“, „Roman“ etc.) bezogen wird. Das Ziel ist hier ein deskriptives Gattungsschema einer datengetriebenen Gattungsstilistik, die induktive und deduktive Perspektiven miteinander verzahnt.

## Annotation

Bei der Anreicherung des Korpus stand bislang eine linguistische Annotation nach Wortarten (Part of Speech, POS) im Vordergrund. Zudem annotierten wir ein KOLIMO-Subsample, das Erzählanfangskorpus (EAK; Herrmann i. Dr.), auf Metaphern und experimentierten mit der Annotation von Redewiedergabe. Wortarten, die in der Korpuslinguistik im deutschen Sprachraum meist mittels des Stuttgart-Tübingen-TagSet (STTS; vgl. Schiller/Teufel et al.

1995) repräsentiert werden, bieten einen Zugriff auf die Beschreibung von Register- und Genrevariation (z. B. Biber/Conrad 2009). Auch gelten sie im Vergleich mit anderen Features wie Syntax oder auch Metaphorizität durch eine relativ genaue automatische Annotation als praktikabel. Um für KOLIMO die Qualität der POS-Annotation einschätzen zu können, untersuchten wir die Genauigkeit einer Reihe von POS-Taggern (Herrmann 2018). Im Zentrum unserer ersten semiautomatischen Evaluation stand das POS-Tagging des DTA, das durch den Einsatz einer morphologischen Segmentierung (Jurish 2011) als besonders genau gilt. Eine Zufallsstichprobe von narrativen Texten mit Erscheinungsdatum 1800-1930 wurde händisch nachkorrigiert (DTA-Version 01.09.2017, verfügbar unter [http://media.dwds.de/dta/download/dta\\_kernkorpus\\_2017-09-01\\_tcf.zip](http://media.dwds.de/dta/download/dta_kernkorpus_2017-09-01_tcf.zip)). Unsere Analyse zeigte eine durchschnittliche 90% Genauigkeit des DTA-Taggings, mit starker Varianz der unterschiedlichen Tags. Dieser Wert ist zwar größer als die Genauigkeit der verglichenen Tagger für das Sample, liegt aber zunächst unter der allgemein angenommenen Benchmark-Genauigkeit für das Gegenwartsdeutsch (98%). Dies zeigt zunächst vor allem, dass die erhältlichen Ressourcen nicht unbesehen auf literarische und historische Texte angewandt werden können. Zurzeit prüfen wir, inwiefern die Gesamtgenauigkeit von der Tagging-Genauigkeit einzelner Tags abhängig ist. Zudem erarbeiten wir auf der Grundlage der nachkorrigierten Daten eine Verbesserung des automatischen Tagging.

KOLIMO ist in der Beta-Version bereits veröffentlicht und steht so der Forschungsgemeinschaft zur Verfügung. Das Webinterface liefert einen variablen Zugriff auf die annotierten Daten, u. a. eine Volltextansicht (siehe Abbildung 2).

KOLIMO wurde zum Zwecke hypothesengetriebener, aber auch explorativer quantitativer Stilistik erstellt (Herrmann 2017). Die Publikation erster Ergebnisse zur stilistischen Variation in der literarischen Moderne ist in Vorbereitung, ebenso wie weitere Analysen von Teilkorpora, für die Befunde etwa zu Metaphorik in Erzählanfängen (Herrmann i. Dr.) und Modalpartikeln in Kafkas *Das Urteil* (Herrmann 2016) bereits vorliegen.

Unser Projekt dokumentiert Entscheidungen auf verschiedenen konzeptionellen, analytischen und prozeduralen Ebenen. Es zeigt, dass der Aufbau eines digitalen literarischen Korpus für den synchronen und diachronen quantitativen Vergleich einer Schwerpunktepoche eine komplexe Aufgabe ist, die sich kaum auf etablierte Lösungen stützen kann. Hypothesen zur Konstitution von Epochen, Autorschaft und Gattungen steuern die Korpuskompilation. Greifbar sind sie als Metadaten (u. a. Autor, Titel, Publikationsdatum, Publika-

tionsort, Gattung) und linguistische Merkmale wie POS oder Metaphern. Sie sind für KOLIMO die Ansatzpunkte, an denen philologische Fragestellungen in präzise und praktikable Kategorien der Korpusliteraturwissenschaft umgewandelt werden können. Durch die flexible Architektur des Korpus können zusätzliche Annotationsebenen einbezogen werden können. Damit will KOLIMO die Offenheit des hermeneutischen Erkenntnisprozesses aufnehmen.

The screenshot shows the KOLIMO Textview interface. On the left is a sidebar with navigation links: guest, About (selected), Dashboard (with a purple notification dot), Collection Browser, and Redewiedergabe. The main area displays a text in German. At the top of the text area is a blue header bar with the KOLIMO logo and a search icon. The text itself is a narrative from a 19th-century German book. It includes several lines of text and some annotations in the margin, such as "Weg nach Winkelteg." and "Hier am Scheidewege stand ein hohes hölzer-nes Kreuz mit drei Querbalken und den bildlich dargestellten Martyrerwerkzeugen der heiligen Leidens-gelichtete, als drei Nägeln. Auf einem Felsen stand der Pfahl, wettergrau und bemoost. Eng daneben stand der Balken mit dem Arme und der Inschrift; „Weg nach Winkelteg.“". The text continues with a description of a path through a forest, mentioning a cross, a nail, and a path sign.

*Abbildung 2: Screenshot KOLIMO Textview (Peter Rosegger  
Die Schriften des Waldschulmeisters, 1875)*

## Erstes Fazit und Ausblick

Die Arbeit am Korpus ist in der Literaturwissenschaft noch Neuland. Neu für das Fach ist vor allem, wie sehr die Konzeption des Korpus die Explikation des angelegten theoretischen Modells samt Fragestellung herausfordert. Korpusliteraturwissenschaftliches Arbeiten kommt nicht ohne Formalismus aus, und weicht dadurch spürbar von den etablierten Ansätzen des Fachs ab, findet gleichwohl in formorientierten Ansätzen wie Jürgen Petersens (2014) *Formengeschichte der deutschen Erzählkunst* und Moritz Baßlers (2015) *Geschichte literarischer Verfahren* auch literaturhistorische Bezüge. Annahmen und Fragestellungen müssen in klar definierte Schritte der Korpuserstellung und Indikatorendefinition übersetzt werden und leiten die Modellbildung, die jedem Korpus vorausgeht. Doch jede Formalisierung stößt an Grenzen,

sei es die der Repräsentativität einer Textauswahl für die Epoche der Literarischen Moderne oder die einer Grundkategorie wie Stil, oder aber die tiefeliegender, semantischer und pragmatischer Dimensionen. An dieser Stelle kommt die hermeneutische Tradition in Spiel, die ein probates Korrektiv der Schwächen formalistischer und explanativ-generalisierender Modelle offeriert, indem situierende und explorative Fallstudien reichhaltige Bezüge aufwerfen. Generell verspricht die Verknüpfung quantitativ-formaler und qualitativ-hermeneutischer Zugängen eine Verbesserung der methodischen Validität der Literaturwissenschaften. Doch auch Mixed Methods-Verknüpfungen sind kein Königsweg zur Überwindung der bestehenden Grenzen der Forschung: Fragestellungen und Begriffe sind immer nur begrenzt formalisierbar, Zugeständnisse an Subjektivität wie Reduktionismus bleiben – und das ideale Korpus bleibt eine regulative Idee.

Freilich sind diese Einwände ihrerseits nur begrenzt gültig. Eine Korpusliteraturwissenschaft muss ebenso wie eine Korpuslinguistik pragmatisch vorgehen, und Kompromisslösungen akzeptieren auf dem Weg einer sich inkrementell verbessernden Datenlage, Methodik und theoretischen Modellierung. Dabei kann nicht abschließend vorausgesagt werden, wie sich Fragestellungen, Methodik und Gegenstandsbereich zueinander verhalten. Dass sich die Sementanalyse dazu eignet, Generalisierungen über den Handlungsverlauf von literarischen Texten zu erstellen, war nicht abzusehen, noch, dass Wahrscheinlichkeitsmodelle der Wortverteilung dafür genutzt werden können, Topics in der Literatur zu ermitteln. Aufbauend auf neuen Schemata der Annotation zeichnet sich die Rolle des maschinellen Lernens für die Literaturwissenschaft derzeit bestenfalls in Umrissen ab. Korpora sind daher eine Variable im Dreieck von Fragestellung, Methodik und Gegenstandsbereich und damit eine der Möglichkeiten, neue Erkenntnisse über Literatur zu gewinnen. Unter den Bedingungen der korpusorientierten Philologie dürfte sich dann auch das Verhältnis der Fächer Literaturwissenschaft und Linguistik zueinander verändern. Sie könnten wieder voneinander lernen. Erkenntnisfortschritt hängt daher auch in der Literaturwissenschaft daran, ob es gelingt, systematischer als bisher im Fach üblich Fragestellungen zu verbessern, Methoden zu kalibrieren und eben repräsentative Korpora zu erstellen.

## Literaturverzeichnis

- Algee-Hewitt, Mark/Sarah Allison/Gemma, Marissa/Heuser, Ryan/Moretti, Franco/Walser, Hannah (2016): Canon/Archive. Large-Scale Dynamics in the Literary Field. In: Stanford Literary Lab Pamphlet 11, URL: <<https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf>>
- Almas, Bridget/ Beaulieu, Marie-Claire (2016): The Perseids Platform: Scholarship for all!. In: Bodard, Gabriel/ Romanello, Matteo (Hg.): Digital Classics out of the Echo-Chamber: Teaching, Knowledge Exchange & Public Engagement, London: Ubiquity Press, S. 171–187
- Archer, Jodie/Jockers, Matthew L. (2016): The Bestseller Code. Anatomy of the Blockbuster Novel. New York: Penguin
- Ardanuy, Mariona Coll/Sporleder, Caroline (2015): Clustering of Novels Represented as Social Networks. In: Linguistic Issues and Language Technology 12,4, URL: <<http://csli-lilt.stanford.edu/ojs/index.php/LILT/article/view/60>>
- Atkins, Sue/Clear, Jeremy/Ostler, Nicholas (1992): Corpus Design Criteria. In: Literary and Linguistic Computing 7,1, S. 1–16, URL: <<https://doi.org/10.1093/lrc/7.1.1>>
- Baßler, Moritz (2015): Deutsche Erzählprosa 1850–1950. Eine Geschichte literarischer Verfahren. Berlin: Erich Schmidt
- Biber, Douglas (1993): Representativeness in Corpus Design. In: Literary and Linguistic Computing 8,4, S. 243–257, URL: <<https://doi.org/10.1093/lrc/8.4.243>>
- Biber, Douglas/Conrad, Susan (2009): Register, Genre, and Style. Cambridge: Cambridge University Press
- Biber, Douglas/Conrad, Susan/Reppen, Randi (2006): Corpus Linguistics. Investigating Language Structure and Use. Cambridge: Cambridge University Press
- Biber, Douglas /Johansson, Stig/Leech, Geoffrey/Conrad, Susan/Finegan, Edward (1999): The Longman Grammar of Spoken and Written English. London: Longman
- Bode, Katherine (2017): The Equivalence of “Close” and “Distant” Reading; or, Toward a New Object for Data-Rich Literary History. In: Modern Language Quarterly 78,1, S. 77–106
- Bubenhofer, Noah (2009): Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse. Berlin: de Gruyter
- Burrows, John (2002): Delta: A Measure of Stylistic Differences and a Guide to Likely Authorship. In: Literary and Linguistic Computing 17,3, S. 267–283
- Burrows, John (2003): Questioning of Authorship: Attribution and Beyond. In: Computer and the Humanities 37,1, S. 5–32

- Burrows, John/Craig, Hugh (1994): Lyrical Drama and the “Turbid Mountebanks”. Styles of Dialog in Romantic and Renaissance Tragedy. In: Computers and the Humanities 28,2, S. 63–86
- Bürger, Thomas et al. (2008): Das VD 18. Eine Einladung ins 18. Jahrhundert. In: Bibliothek. Forschung und Praxis 32, S. 195–202
- Crane, Gregory (2016): Greco-Roman Studies in a Digital Age. In: Daedalus 145, S. 127–133
- Craig, Hugh (2012): George Chapman, John Davies of Hereford, William Shakespeare, and A Lover’s Complaint. In: Shakespeare Quarterly 63, S. 147–174
- Craig, Hugh/Greatley-Hirsch, Brett (2017): Style, Computers, and Early Modern Drama. Beyond Authorship. Cambridge: Cambridge University Press
- Dittmann, Henrik/Ďurčo, Matej/Geyken, Alexander/Roth, Tobias/Zimmer, Kai (2012): Korpus C4 – A Distributed Corpus of German Varieties. In: Schmidt, Thomas/Wörner, Kai (Hg.): Multilingual Corpora and Multilingual Corpus Analysis. Amsterdam: Benjamins, S. 339–346
- Eder, Maciej (2010): Does size matter? Authorship Attribution, Small Samples, Big Problem. In: Digital Humanities 2010: Conference Abstracts. King’s College London, S. 132–135
- Eder, Maciej/Rybicki, Jan/Kestemont, Mike (2016): Stylometry with R. A Package for Computational Text Analysis. In: R Journal 8,1, S. 107–121
- Erlin, Matt (2014): The Location of Literary History. Topic Modeling, Network Analysis, and the German Novel 1731–1864. In: Erlin, Matt/Tatlock, Lynne (Hg.): Distant Readings. Topologies of German Culture in the Long Nineteenth Century. Rochester: Camden, S. 55–90
- Evert, Stefan/Proisl, Thomas/Jannidis, Fotis/Reger, Isabella/Pielström, Steffen/Schöch, Christof/Vitt, Thorsten (2017): Understanding and Explaining Distance Measures for Authorship Attribution. In: Digital Scholarship in the Humanities 32, S. 4–16
- Fièvre, Paul (2007): Théâtre classique, URL: <<http://www.theatre-classique.fr/>>
- Fischer, Frank/Jäschke, Robert (2018): Liebe und Tod in der Deutschen Nationalbibliothek. In: Kritik der digitalen Vernunft. DHd Jahrestagung, Köln. Konferenzabstracts, S. 261–266, URL: <http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf>
- Geyken, Alexander/Gloning, Thomas (2015): A Living Text Archive of 15th-19th-Century German. Corpus Strategies, Technology, Organization. In: Gippert, Jost/Gehrke, Ralf (Hg.): Historical Corpora. Challenges and Perspectives. Tübingen: Narr, S. 165–180

- Giere, Ronald N. (1999): Using Models to Represent Reality. In: Magnani, Lorenzo/Nersessian, Nancy/Thagard, Paul (Hg.): *Model-Based Reasoning in Scientific Discovery*. Boston: Springer, S. 41–57, URL: [https://doi.org/10.1007/978-1-4615-4813-3\\_3](https://doi.org/10.1007/978-1-4615-4813-3_3)
- Graevenitz, Gerhart (1999): Literaturwissenschaft als Kulturwissenschaft? In: Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte 73, S. 69–93
- Herrmann, J. Berenike (2016): „Läuse im Pelz der Sprache?“ Zu Funktionen von Modalpartikeln in narrativen (De-)Motivierungsstrategien bei Franz Kafka. In: Horváth, Márta/Mellmann, Katja (Hg.): *Die biologisch-kognitiven Grundlagen narrativer Motivierung*. Münster: Mentis, S. 169–192
- Herrmann, J. Berenike (2017): In a Test Bed with Kafka. Introducing a Mixed-Method Approach to Digital Stylistics. In: Chambers, Sally/Jones, Catherine/Kestemont, Mike/Koolen, Marijn/van Zundert, Joris (Eds.). Special Issue DHBenelux 2015, In: *Digital Humanities Quarterly* 11,4, URL: <http://www.digitalhumanities.org/dhq/vol/11/4/000341/000341.html>
- Herrmann, J. Berenike (2018): Praktische Tagger-Kritik. Zur Evaluation des POS-Tagging des Deutschen Textarchivs. In: *Kritik der digitalen Vernunft*. Jahrestagung der DHd, Köln. Konferenzabstracts, S. 287–290, URL: <http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf>
- Herrmann, J. Berenike (i. Dr.): *Anschaulichkeit messen. Eine quantitative Meta-phernanalyse deutschsprachiger Erzählanfänge zwischen 1880 und 1926*. In: Köppé, Tilmann/Singer, Rüdiger (Hg.). *Show, don't tell: Konzepte und Strategien narrativer Anschaulichkeit*. Bielefeld: Aisthesis
- Herrmann, J. Berenike / Lauer, Gerhard (2016a): KAREK: Building and Annotating a Kafka/Reference Corpus. International Conference Digital Humanities 2016, July, Krakow, Poland. Konferenzabstracts, S. 552–553, URL: <http://dh2016.adho.org/abstracts/427>
- Herrmann, J. Berenike / Lauer, Gerhard (2016b): Aufbau und Annotation des Kafka/Referenzkorpus. In: Modellierung, Vernetzung, Visualisierung. Die Digital Humanities als fächerübergreifendes Forschungsparadigma. Jahrestagung der DHd, Leipzig. Konferenzabstracts, S. 158–160, URL: <http://dhd2016.de/boa.pdf>
- Herrmann, J. Berenike/van Dalen-Oskam, Karina/Schöch, Christof (2015): Revisiting Style, a Key Concept in Literary Studies. In: *Journal of Literary Theory* 9,1, S. 25–52
- Hösle, Vittorio (1990): Einleitung. Vico und die Idee der Kulturwissenschaft. In: Giovanni Battista Vico [1725]: Prinzipien einer neuen Wissenschaft über die gemeinsame Natur der Völker. Übers. von Vittorio Hösle, Vittorio/Jermann, Christoph. Bd. 1, Hamburg: Meiner

- Huber, Martin/Lauer, Gerhard (2000): Neue Sozialgeschichte? Poetik, Kultur und Gesellschaft. Zum Forschungsprogramm der Literaturwissenschaft. In: Huber, Martin/Lauer, Gerhard (Hg.): Nach der Sozialgeschichte. Konzepte für eine Literaturwissenschaft zwischen Historischer Anthropologie, Kulturgeschichte und Medientheorie. Tübingen: Niemeyer, S. 1–11
- Jannidis, Fotis/Lauer, Gerhard/Rapp, Andrea (2005): Alte Romane und neue Bibliotheken. Zum Projekt eines digitalen historischen Referenzkorpus des Deutschen. In: Nielsen, Erland Kolding/Saur, Klaus G./Ceynowa, Klaus (Hg.): Die innovative Bibliothek. Elmar Mittler zum 65. Geburtstag. München: Saur, S. 139–150
- Jannidis, Fotis/Lauer, Gerhard (2014): Burrows's Delta and Its Use in German Literary History. In: Erlin, Matt/Tatlock, Lynne (Hg.): Distant Readings. Topologies of German Culture in the Long Nineteenth Century. Rochester: Camden House, S. 29–54
- Jockers, Matthew L. (2013): Macroanalysis. Digital Methods and Literary History. Illinois: University of Illinois Press
- Jockers, Matthew L. (2014): Text Analysis with R for Students of Literature. Heidelberg: Springer
- Jockers, Matthew L./Underwood, Ted (2015): Text-Mining the Humanities. In: Schreibman, Susan/Siemens, Ray/Unsworth, John (Hg.): A New Companion to Digital Humanities. Chichester: John Wiley & Sons, S. 291–306. URL: <https://doi.org/10.1002/9781118680605.ch20>
- Jurish, Bryan (2011): Fine-State Canonicalization Techniques for Historical German. Dissertation Universität Potsdam. URL: <<https://publishup.uni-potsdam.de/opus4-ubp/frontdoor/index/index/docId/5562>>
- Kestemont, Mike/van Dalen-Oskam, Karina (2009): Predicting the Past. Memory-Based Copyist and Author Discrimination in Medieval Epics. In: Proceedings of the Twenty-First Benelux Conference on Artificial Intelligence 21, S. 121–128
- Kim, Dan (2012): Mixed Methods. The Encyclopedia of Applied Linguistics. Wiley Online Library. URL: <http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0766/abstract>
- Lemnitzer, Lothar/Zinsmeister, Heike (2015): Korpuslinguistik. Eine Einführung. 3. Auflage. Tübingen: Narr
- Liu, Alan (2016): N+1. Plea for Cross-Domain Data in the Digital Humanities. In: Debates in the Digital Humanities. URL: <http://dhdebates.gc.cuny.edu/debates/text/101>
- Mahlberg, Michaela (2013): Corpus Stylistics and Dickens's Fiction. New York: Routledge

- Mahlberg, Michaela (2015): Literary Style and Literary Texts. In: Biber, Douglas/Repken, Randi (Hg.): *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press, S. 346–361
- Maurer, Michael (Hg.) (2002): *Aufriß der Historischen Wissenschaften*. Band 4: Quellen. Ditzingen: Reclam
- Meindl, Claudia (2011): *Methodik für Linguisten*. Tübingen: Narr
- Meister, Jan-Christoph (2013): Computerphilologie vs. Digital Text Studies. Von der pragmatischen zur methodologischen Perspektive auf die Digitalisierung der Literaturwissenschaften. In: Grond-Rigler, Christine/Straub, Wolfgang (Hg.): *Literatur und Digitalisierung*. Berlin, Boston: de Gruyter, S. 267–296
- Moretti, Franco (2000): Conjectures on World Literature. In: *New Left Review* 1, S. 54–68
- Moretti, Franco (2013): *Distant Reading*. London: Verso
- Mueller, Martin (2012): Scalable Reading, URL: <[https://scalablerreading.northwestern.edu/?page\\_id=22](https://scalablerreading.northwestern.edu/?page_id=22)>
- Müller, Marcus (2011): Vom Wort zur Gesellschaft: Kontexte in Korpora. Ein Beitrag zur Methodologie der Korpuspragmatik. In: Felder, Ekkhard/Müller, Marcus/Vogel, Friedemann (Hg.): *Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen*. Berlin, Boston: De Gruyter, S. 33–82
- Neuroth, Heike/Rapp, Andrea/Söring, Sibylle (2015): Text Grid: Von der Community – für die Community. Glückstadt: Werner Hülsbusch
- Pennebaker, James W./Ireland, Molly E. (2011): Using Literature to Understand Authors. The Case for Computerized Text Analysis. In: *Scientific Study of Literature* 1,1, S. 34–48
- Petersen, Jürgen H. (2014): *Formgeschichte der deutschen Erzählkunst: von 1500 bis zur Gegenwart*. Berlin: Erich Schmidt
- Piper, Andrew (2016): Fictionality. In: Cultural Analytics, DOI: 10.22148/16.011, URL: <<http://culturalanalytics.org/2016/12/fictionality/>>
- Piper, Andrew (2017): Think Small: On Literary Modeling. In: Publications of the Modern Language Association 132,3, S. 651–658, URL: <<https://doi.org/10.1632/pmla.2017.132.3.651>>
- Piper, Andrew/Portelance, Eva (2016): How Cultural Capital Works: Prizewinning Novels, Bestsellers, and the Time of Reading. In: Post45, URL: <<http://post45.research.yale.edu/2016/05/how-cultural-capital-works-prizewinning-novels-bestellers-and-the-time-of-reading/>>
- Raddatz, Fritz J. (2009): *Die ZEIT-Bibliothek der 100 Bücher* (13. Aufl.). Frankfurt/M: Suhrkamp

- Reagan, Andrew J./ Mitchell, Lewis/Kiley, Dilan/ Danforth, Christopher M./ Dodds, Peter Sheridan (2016): The Emotional Arcs of Stories Are Dominated by Six Basic Shapes. In: EPJ Data Science 5, 1, URL: <<https://doi.org/10.1140/epjds/s13688-016-0093-1>>
- Rieger, Burghard (1979): Repräsentativität: von der Unangemessenheit eines Begriffs zur Kennzeichnung eines Problems linguistischer Korpusbildung. In: Bergen-holtz, Henning/Schaeder, Burghard (Hg.): Empirische Textwissenschaft. Aufbau und Auswertung von Text-Corpora. Königstein: Scriptor, S. 52–70
- Rosenberg, Rainer (2000): Art. Kanon. In: Fricke, Harald/Weimar, Klaus/Müller, Jan-Dirk (Hg.): Reallexikon der deutschen Literaturwissenschaft. Berlin, Bd. II, S. 224–227
- Rybicki, Jan (2012): The Great Mystery of the (Almost) Invisible Translator. In: Oakes, Michael P./Meng, Ji (Hg.): Quantitative Methods in Corpus-Based Translation Studies. A Practical Guide to Descriptive Translation Research. Amsterdam: Benjamins, S. 231–248
- Schiller, Anne/Teufel, Simone/Stöckert, Christine/Thielen, Christine (1995): Guidelines für das Tagging deutscher Textcorpora mit STTS. Manuscript, Universities of Stuttgart and Tübingen. URL <<http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf>>
- Schöch, Christof (2014): Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik. In: Phin Beiheft 7, S. 130–157
- Schöch, Christof (2017): Topic Modeling Genre. An Exploration of French Classical and Enlightenment Drama. In: Digital Humanities Quarterly 17,2, URL: <<http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html>>
- Semino, Elena/Short, Mick (2004): Corpus stylistics. Speech, Writing and Thought Presentation in a Corpus of English Writing. London: Routledge.
- Siegel, Erik/Retter, Adam (2014): eXist: A NoSQL Document Database and Application Platform. Sebastopol: O'Reilly
- Stubbs, Michael (2005). Conrad in the Computer: Examples of Quantitative Stylistic Methods. In: Language and Literature 5,1, S. 5–24
- Underwood, Ted (2016): The Life Cycles of Genres. In: Cultural Analytics, DOI: 10.22148/16.005, URL: <<http://culturalanalytics.org/2016/05/the-life-cycles-of-genres/>>
- Underwood, Ted (2017): A Genealogy of Distant Reading. In: Digital Humanities Quarterly 11,2, URL: <<http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html>>

- Underwood, Ted/Sellers, Jordan (2016): The Longue Durée of Literary Prestige. In: *Modern Language Quarterly* 77,3, S. 321–344
- Vogt, Ernst (1979): Der Methodenstreit zwischen Hermann und Böckh und seine Bedeutung für die Geschichte der Philologie. In: Flashar, Hellmut/Gründer, Karlfried/Horstmann, Axel (Hg.): *Philologie und Hermeneutik im 19. Jahrhundert. Zur Geschichte und Methodologie der Geisteswissenschaften*. Göttingen: Vandenhoeck & Ruprecht, S. 103–121
- Weitin, Thomas (2015): Digitale Literaturwissenschaft. In: *Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte* 89,4, S. 651–656
- Weitin, Thomas (2017): Scalable Reading. In: *Zeitschrift für Literaturwissenschaft und Linguistik* 47, S. 1–6
- Winko, Simone (2008): Textualitätsannahmen und die Analyse literarischer Texte. In: *Zeitschrift für Germanistische Linguistik* 36,3, S. 427–443
- Wünsch, Marianne (2007): Realismus (1850–1890). Zugänge zu einer literarischen Epoche. Kiel: Ludwig



# Korpusbasiertes Arbeiten und epigraphische Datenbanken: Möglichkeiten und Herausforderungen am Beispiel von EPIDAT und DIO

## 1. Die Daten

Ausgangspunkt der folgenden Überlegungen sind die epigraphischen Portale EPIDAT, *Forschungsplattform zur jüdischen Grabsteineigraphik* (EPIDAT) und *Deutsche Inschriften Online* (DIO). Beide Sammlungen verdanken ihr Entstehen primär sozial-, kultur- und kunsthistorischen sowie philologischen Fragestellungen.<sup>1</sup> Korpuslinguistische Herangehensweisen sind bislang nicht oder nur im kleinen Rahmen zum Tragen gekommen. Allerdings ist mit dem Referenzkorpus *Deutsche Inschriften* (ReDI) ein bedeutender Anfang gemacht.<sup>2</sup>

### EPIDAT

EPIDAT wird seit 2002 am Essener Steinheim-Institut für deutsch-jüdische Geschichte entwickelt. Die Datenbank ging 2006 online und konnte ihren Umfang seitdem kontinuierlich erweitern. Die Bestände sind über die Webseite des Forschungsportals frei zugänglich: [www.steinheim-institut.de/cgi-bin/epidat](http://www.steinheim-institut.de/cgi-bin/epidat). Jeder Datensatz wird unter einer Creative Commons-Lizenz als Open Access-Publikation veröffentlicht.<sup>3</sup> Derzeit dokumentiert die Datenbank 180 historische jüdische Friedhöfe mit mehr als 33.000 Inschriften und etwa 65.000 digitalen Abbildungen. EPIDAT enthält nahezu ausschließlich Grabinschriften.<sup>4</sup>

- 
- 1 Zum Entstehungskontext der Deutschen Inschriften (DI) s. Brandi 1937, Panzer 1938, Nikitsch 2008. – Zu DIO Schrade 2012. – Zu EPIDAT s. Kollatz 2004 und 2015.
  - 2 S. Herbers 2016, besonders S. 27–29; Desiderate, S. 31–32; Herausforderungen, S. 40–41, Fn. 64; <http://www.ruhr-uni-bochum.de/wegeRa/ReDI/index.htm> (Zugriff: 15.11.2017).
  - 3 Zu Offenen Lizenzen und Nachnutzung s. Kreutzer 2016.
  - 4 EPIDAT dokumentiert auch einige auf jüdischen Friedhöfen errichtete Gedenksteine, s. Index der Gedenksteine, <http://steinheim-institut.de/cgi-bin/epidat?info=index&anzeige=memorials> (Zugriff: 15.11.2017).

Topographisch sind vor allem Bestände aus heutigen und historischen Regionen Deutschlands, aber auch einige der Niederlande, Tschechiens und Litauens vertreten. Zeitlich umspannt EPIDAT einen Zeitraum, der vom 11. bis zum 20. Jahrhundert reicht. Die Hauptsprache der Inschriften ist bis Ende des 18. Jahrhunderts das Hebräische. Die Eulogien belegen einen ebenso kreativen wie souveränen Umgang mit dem Wortschatz der hebräischen Bibel, des rabbinischen Schrifttums und der synagogalen Liturgie. Jiddisch ist zu keiner Zeit Sprache der Inschriften.<sup>5</sup> Deutsch lässt ab Mitte des 19. Jahrhunderts von regionalen Ausnahmen abgesehen das Hebräische in den Hintergrund treten,<sup>6</sup> das im 20. Jahrhundert zumeist nur noch in traditionellen Einleitungs- und Schlussformeln (פָנָן חֲנַצְבָּה) auftritt. Insgesamt werden die Inschriftentexte in der Moderne stereotyper und kürzer. Diesem Textschwund wird mit einer Formenvielfalt in der Gestaltung der Inschrifenträger begegnet.<sup>7</sup>

Zu allen Zeiten finden sich Ornamente, bildhafte und symbolische Elemente auf den Grabmalen (Abb. 1). Letztere stehen oft in direktem Bezug zu Namen oder Amt der Verstorbenen. David wird durch den Davidschild (hebr. Magen David) symbolisiert (Abb. 3), Frau Röschens Grabmal mit einer Rose geschmückt (Abb. 2).<sup>8</sup> Auch Familien- und Zunamen werden visualisiert, so etwa der Name des Verstorbenen David Hammerschlag durch ein doppeltes Symbol, nämlich einen Davidschild, in den zwei Hämmer integriert sind (Abb. 3).<sup>9</sup> Zuweilen finden sich bildliche Verweise auf den Beruf, etwa eine Harfe für eine Musiklehrerin<sup>10</sup> oder ein Messer für das religiöse Amt des Beschneiders.<sup>11</sup> Die am häufigsten verwendeten Symbole sind die Zeichen,

5 Ausnahme von der Regel ist die teilweise in jiddischer Sprache gehaltene Grabinschrift eines durch einen Arbeitsunfall ums Leben gekommenen Industriearbeiters: לא ל'גט רעד' (da liegt der auf schreckliche Weise ums Leben gekommene ...), s. du6-11 <http://steinheim-institut.de/cgi-bin/epidat?id=du6-11> (Zugriff: 15.11.2017).

6 Lediglich die sogenannten Austrittsgemeinden halten in der Auseinandersetzung mit der innerjüdischen Reformbewegung bewusst am Hebräischen fest.

7 S. Arera-Rütenik/Kollatz 2016.

8 S. hha-3938 <http://www.steinheim-institut.de/cgi-bin/epidat?id=hha-3938> (Zugriff: 15.11.2017) und weitere Beispiele.

9 S. hha-2338 <http://www.steinheim-institut.de/cgi-bin/epidat?id=hha-2338> (Zugriff: 15.11.2017). Die Eulogie nimmt das Spiel mit dem Namen auf und verwendet Zitate aus den Erzählungen um den biblischen David (2. Sam 15,30, 1. Kön 2,1).

10 S. bay-591 <http://www.steinheim-institut.de/cgi-bin/epidat?id=bay-591> (Zugriff: 15.11.2017).

11 S. hgl-275 <http://steinheim-institut.de/cgi-bin/epidat?id=hgl-275> (Zugriff: 15.11.2017). Die Textaussage „[er] eilte zum Gebot der Beschneidung morgens und abends“ wird durch das Symbol des Messers unterstützt.

die die beiden biblischen Geschlechter Kohen und Levi symbolisieren, erste durch die segnenden Hände der Priester, letztere durch die Kannen zur Reinigung der Leviten.



Abb. 1



Abb. 2 und 3

Im Aufbau sind die Inschriften im Allgemeinen formularisch gehalten. Vereinfacht gesagt variieren sie das Grundmuster: Einleitung – Formel – Eulogie – Name – Sterbetag – Schlussformel:<sup>12</sup>

פ"ט ילדה חכמה פרadicי צפורה בת כי מאיר ב"א ב"מ נפטר ונקברה יוי א' דה"ה של  
פסח תקיה לא פ"ק הנצבר"ה

Konstitutive Elemente sind die Einleitungsformel (פ"ט – Hier ist geborgen), das eigentliche Totenlob („ein kluges Kind“), der Name der Verstorbenen („Fradche Zipora, Tochter des Meir, Sohn des Awraham, Sohn des Meir“), das Todes- bzw. Begräbnisdatum („begraben am 1. Tag der Zwischenfeiertage von Pessach im Jahr 511 kleiner Zählung“) und die Schlussformel (הנצבר"ה meist

<sup>12</sup> S. hha-309 <http://www.steinheim-institut.de/cgi-bin/epidat?id=hha-309>