Korpuserstellung als literaturwissenschaftliche Aufgabe

Gius, Evelyn
evelyn.gius@uni-hamburg.de
Universität Hamburg, Deutschland

Katharina, Krüger
katharina.krueger@uni-hamburg.de
Universität Hamburg, Deutschland

Carla, Sökefeld
carla.soekefeld@studium.uni-hamburg.de
Universität Hamburg, Deutschland

Korpuserstellung in der (digitalen) Literaturwissenschaft

Die Praxis der Zusammenstellung von Primärtexten zu einem Korpus ist gewissermaßen literaturwissenschaftliches Alltagsgeschäft, trotzdem wird sie selten problematisiert. Die beiden Standardfälle der nicht-digitalen Korpusanalyse erscheinen auch bezüglich der Korpuszusammenstellung unproblematisch: (1) Die Forschungsfrage erfordert eine bestimmte Textbasis (etwa bei einer Untersuchung zu Krankheitsdarstellungen bei Thomas Mann), (2) Die Korpuserstellung basiert auf der Kanonizität der Texte. Zumindest im zweiten Fall ist das geeignete Vorgehen allerdings nicht selbstverständlich, denn es müsste begründet werden, nach welchen Kriterien Werke tatsächlich exemplarisch und repräsentativ sind. Dies wird jedoch kaum thematisiert (vgl. Gius, in Vorbereitung).

Bei der konkreten Erstellung eines *digitalen* Korpus ist wiederum aufgrund der Menge der verfügbaren Texte häufig nur eine balancierte Sammlung nach vordefinierten Kriterien realisierbar, die allerdings die Gefahr von unerwünschten Korrelationen birgt (Schöch 2017).[1] Trotz dieser Problematik wird die methodologische Bedeutung der Korpuszusammenstellung auch in den *Digital Humanities* kaum wahrgenommen . So ergibt eine Stichwortsuche nach „Korpus“ bzw. „corpus“ in allen Tagungsbänden der *Digital Humanities*- und *Digital Humanities im deutschsprachigen Raum*-Konferenzen von 2014-2017 jeweils hunderte Beiträge. Allerdings gehen nur acht der 49 Abstracts mit literaturwissenschaftlicher Thematik explizit auf die Korpus *erstellung* ein, obwohl in den übrigen durchaus auch von repräsentativen Korpora die Rede ist. Lediglich fünf[2] dieser Beiträge beschäftigen sich dezidiert mit der Problematik eines literaturwissenschaftlichen Korpus.

Ein Grund für die geringe Auseinandersetzung mit dem Thema kann sein, dass die Praxeologie der Zusammenstellung literarischer digitaler Korpora verhältnismäßig neu ist und sich bislang keine literaturwissenschaftlichen Routinen etablieren konnten, die das Zusammenstellen des Untersuchungsobjektes betreffen.[3] Ein anderer Erklärungsansatz ist, dass schon in der traditionellen Literaturwissenschaft der “literaturwissenschaftliche Objektumgang vor allem durch implizite Normen strukturiert” ist (Schruhl 2018) und entsprechend auch eine Übersetzung dieses nicht-digitalen Zugangs in einen digitalen Zugang – also eine Art Operationalisierung der Korpuserstellung – nicht möglich ist.

Die digitale literaturwissenschaftliche Korpuserstellung stellt also ein ungelöstes methodologisches Grundproblem dar.

Wir erläutern im Folgenden die Problematik der literaturwissenschaftlichen Korpuserstellung exemplarisch, um eine Diskussion darüber anzustoßen sowie mögliche Lösungsansätze vorzustellen.

Erstellung eines Korpus zur Untersuchung von genderspezifischer Darstellung von Krankheit

Im Rahmen unseres Forschungsprojektes zu genderspezifischer Darstellung von Krankheit in literarischen Texten im Forschungsverbund hermA (“Automatische Modellierung hermeneutischer Prozesse”)[4] geht es um die Frage nach der genderspezifischen Darstellung von Krankheit bei literarischen Figuren um 1900. In einem korpusbasierten Ansatz werden Methoden für die Analyse der Darstellung von Krankheit entwickelt. Da wir einen konstruktivistischen Ansatz in Bezug auf Genderkonzepte verfolgen und außerdem grundsätzlich davon ausgehen, dass sich semantische Aspekte in literarischen Texten nur bedingt an der Textoberfläche materialisieren,[5]mussten wir Strategien entwickeln, um ein möglichst großes Ausgangskorpus zu etablieren. Die Grundlage zur Erstellung dieses Korpus war das Kolimo-Korpus nach Herrmann und Lauer (2017). Es enthält über 42.000 Texte mit einem Fokus auf der Zeitspanne von 1880 bis 1930 aus dem Deutschen Textarchiv, dem TextGrid Repository und dem Projekt Gutenberg-DE (vgl. Tabelle 2).

Da die Entwicklung von Strategien zur Bildung themenspezifischer Subkorpora ebenfalls Ziel des Forschungsvorhabens ist, gab es für die Textauswahl zunächst keine inhaltlichen Einschränkungen. Kriterien zur Textauswahl waren deshalb nur das Datum

der Erstveröffentlichung, Textsprache, das Genre und Textlänge; Grundlage für ihre Ermittlung bildeten die Kolimo-Metadaten.

Kriterium	Relevanter Bereich	Zweifelsfälle	Festlegung durch:
Erscheinungsdatum	1870-1920	Zeitspannen, widersprüchliche Angaben, fehlende Angaben	Metadaten + Recherche zu Ersterscheinungsdaten bzw. Lebensdaten
Sprache	Deutschsprachig (Standardsprache), keine Übersetzungen		Automatisierte Überprüfung: >90% deutschsprachig
Gattung/Genre	Prosa und dazugehörige Genres	Kinder- und Jugendbuch (im Kernkorpus) Autobiographie, Biographie, Fabel, Märchen, Memoire, Tagebuch (nicht im Kernkorpus) Ohne spezifisches Genre	Metadaten Kategorisierung der Angaben in den Metadaten Recherche
Textlänge	> 1.000 Wörter	Kapitelweise vorliegende Texte u.ä.	automatische Wortzählung

Tabelle 1: Übersicht der Kriterien für die Korpuserstellung

Ins Korpus aufgenommen wurden Texte mit Erstveröffentlichung zwischen 1870 und 1920. Die Auswahl wurde anhand der vorliegenden Metadaten getroffen und bei Bedarf weiter ergänzt. Je nach Repositorium unterschied sich die zusätzlich nötige Recherchearbeit stark: elf DTA-Texte, 7.602 Gutenberg-Texte, und 27.300 TextGrid-Texte hatten keine Datumsangabe.

Da wir uns auf Phänomene konzentrieren, die in der deutschsprachigen zeitgenössischen Literatur verhandelt wurden, lag der Fokus auf standarddeutschen Texten. Übersetzungen wurden in ein Sonderkorpus aufgenommen, zu weniger als 90% deutschsprachige Texte aussortiert.

Außerdem wurden ausschließlich Prosatexte berücksichtigt, wobei für die (Sub-)Genres Autobiographie, Biographie, Fabel, Märchen, Memoire und Tagebuch eigene Sonderkorpora erstellt wurden. Texte der Kategorie Kinder- beziehungsweise Jugendbuch wurden hingegen in das Kernkorpus miteinbezogen.

Um Kürzestprosa auszuschließen, die sich meist deutlich von der narrativen Struktur anderer Formen unterscheidet, wurden nur Texte mit über 1.000 Wörtern aufgenommen.

Das nach diesen Kriterien manuell erstellte Kernkorpus umfasst nur noch etwa 2.700 Werke.

Korpus	Texte (DTA/Textgrid/ Gutenberg)
Ausgangskorpus Kolimo	42.710 (1.317 / 27.412 / 13.981)
Kernkorpus	2.726 (50 / 1.137 / 1.539)
Sonderkorpus Märchen	2.601
Sonderkorpus Übersetzungen	465
Weitere Sonderkorpora	69

Tabelle 2: Übersicht Korpusgrößen

Korpusbereinigung

Die dargestellten Schwierigkeiten bei der Entscheidung über die Aufnahme eines Textes in das Kernkorpus basierten auf unvollständigen, widersprüchlichen oder nicht ohne weiteres aufeinander abbildbaren Metadaten und konnten anhand der dargelegten Setzungen vergleichsweise einfach gelöst werden. Weitaus schwieriger gestaltete sich hingegen die Identifikation von Dubletten und die Konzipierung einer geeigneten Problembehandlung – ein leider typisches Problem bei der Aggregation eines Korpus aus verschiedenen Quellen.

So zeigte sich bei einer ersten manuellen Durchsicht der Liste und der stichprobenartigen Überprüfung der Volltexte, dass Dubletten nicht eindeutig anhand von Metadaten identifizierbar sind. Die naheliegende Lösung, ein Volltextvergleich, müsste jedoch bei 2.700 Texten über sieben Millionen mögliche Paare vergleichen und würde auch bei der Nutzung von Cluster Computing Jahrzehnte dauern. Deshalb mussten Heuristiken entwickelt werden, um das Verfahren abzukürzen.

Entsprechend haben wir einen Workflow entworfen, um mit möglichst hoher Treffgenauigkeit Dubletten zu identifizieren. Ziel war, alle echten Dubletten zu finden und das Korpus entsprechend zu bereinigen, ohne Texte vorschnell zu streichen.

Dafür wurden folgende, größtenteils automatische Prüfungen durchgeführt:[6]

1. Identifikation eindeutiger Dubletten:
   - Werke, denen derselbe Volltext zugeordnet wurde
   - Werke mit einer Edit-Distanz (Levensthein 1966) # 2 bei Autor und Titel
2. Identifikation Dublettenkandidaten:
   - Werke mit einer Edit-Distanz # 2 bei Autor (und mehr beim Titel)
3. Volltextvergleich Dublettenkandidaten:
   - gemessene einseitige Edit-Distanz (in beiden Richtungen):
     - #15%: Dublette
     - # 80%: keine Dublette
   - alle anderen Fälle: manuelle Prüfung (vgl. Tabelle 3)

4. Entfernung verbleibender Texte mit
   - # 1.000 Wörter und/oder
   - # 10% nichtdeutscher Textanteil

Bei den verbleibenden Textpaaren gehen wir von echten Dubletten aus. Für diese erfolgt eine Repositorien-Priorisierung nach der Qualität der Repositorien (1. DTA, 2. TextGrid, 3. Gutenberg).

Problem	Bsp.	Lösungsansatz
Text enthält zusätzlichen Paratext	“Hymnen” (Ferdinand von Saar) mit Vorwort des Herausgebers	Variante ohne Paratext
Eindeutig stark veränderte Ausgabe	“Die Pilger der Wildnis” (Johannes Scherr) als “für die Jugend bearbeitete Ausgabe“	“Ursprungsvariante” wählen
Mehrbändiges Werk (# Dubletten)	“Auch Einer” (Friedrich Theodor von Vischer, 1879): Band 1 und 2 mit zwei nahezu gleichlautenden Dateinamen	zusammenführen
Teile eines Sammelbandes (# Dubletten)	“Der Schmied seines Glückes” von Gottfried Keller, erschienen in “Die Leute von Seldwyla”	aufteilen

Tabelle 3: Dubletten: Lösungsansätze bei manueller Überprüfung

Ansätze für die literaturwissenschaftliche Korpuserstellung?

Die Digitalisierung fördert die literaturwissenschaftliche Korpuserstellung in neuem Umfang und macht dadurch die Textzusammenstellung als literaturwissenschaftliches Problem offensichtlich, das methodologisch kaum beleuchtet ist. Oft dominiert die Frage, welche Texte überhaupt ins Korpus können, also aus welchen bereits digitalisierten oder noch digitalisierbaren Texten ausgewählt werden kann, die literaturwissenschaftlichen Überlegungen zur Textauswahl.

Da die Menge digitalisiert vorliegender Texte stetig steigt, haben Korpora außerdem immer häufiger eine Größe, die von einzelnen Forscher/innen nicht mehr überschaubar ist. Deshalb muss sich die literaturwissenschaftliche Begründung der Relevanz der Texte in einem Korpus einer gewissen Größe im Zweifelsfall auf Aspekte, die als Informationen in den Metadaten der Texte vorliegen – wie Zeit, Gattung/Genre oder Autor/innen – beschränken.

Sowohl die Menge der zur Verfügung stehenden Texte als auch die Qualität der Primär- und Metadaten sind noch steigerungsfähig. Hier ist die Forschungsgemeinschaft genauso gefordert wie Bibliotheken und Archive.

Im Sinne einer wissenschaftlichen Qualitätssicherung ist es daher umso wichtiger, im Hinblick auf das *zur Verfügung* stehende Datenmaterial Qualitätskriterien für die Zusammenstellung der Korpora zu entwickeln und umzusetzen, wie das vorgestellte Verfahren zeigen soll. So können aus Texten mit schlechten oder fehlenden Metadaten vergleichsweise homogene Korpora erstellt und anschließend mit weiteren Informationen angereichert werden.

Da das literaturwissenschaftliche Wissen über ein Korpus mit steigender Korpusgröße notwendigerweise ab- und damit die Gefahr unerwünschter Korrelationen zunimmt, sollte das Korpus außerdem mit Informationen angereichert werden, die für die Interpretation von Analyseergebnissen genutzt werden können (Epochenzugehörigkeit, thematische Zuordnung etc). Dadurch können entdeckte Korrelationen auf einer größeren Basis an Texteigenschaften – also: besser – interpretiert werden. Da das manuelle Ergänzen von Informationen sehr arbeitsaufwändig ist, sollten dafür (halb-)automatische Verfahren entwickelt bzw. genutzt werden.[7]

Die Anforderungen an ein Korpus variieren je nach Kontext und Forschungsfrage des Projekts, deshalb müssen allgemeine Qualitätskriterien für die Korpuserstellung eine gewisse Offenheit aufweisen. Grundsätzlich gilt aber: Für die Korpuserstellung muss frühzeitig eine Strategie zur Priorisierung von Problemen entwickelt – und dokumentiert! – werden. Dabei geht es darum, sich wie in diesem Beitrag skizziert eine Übersicht über konzeptuelle und technische Aspekte zu verschaffen und anschließend einfach zu lösende Probleme und konzeptuell wichtige Entscheidungen in eine geeignete Reihenfolge zu bringen – den Workflow für die Korpuserstellung.

Fußnoten

1. Zu den verschiedenen Typen von Datensammlungen bzw. Korpora vgl. ebenfalls Schöch (2017).
2. Farrar (2016): Corpus of Revenge Tragedy; Herrmann und Lauer (2016a/b): Kafka/Referenzkorpus; Herrmann und Lauer (2017): Kolimo; Pernes et al. (2017): historisch-literarisches Metaphernkorpus.
3. Trilcke & Fischer (2018) sprechen hier davon, dass das “epistemische Ding” der Literaturwissenschaft sich vom (einzelnen) Primärtext hin zu durch Weiterverarbeitung entstandene Zwischenformate und Korpora entwickelt.
4. Vgl. https://www.herma.uni-hamburg.de und Gaidys et al. (2017).

5. So wird häufig eine für den Texte zentrale Krankheit nur andeutungsweise erzählt, etwa in Schnitzlers Novelle *Sterben* (1894), in der die Krankheit des Protagonisten über den Großteil des Textes ausschließlich über die Symptom- und Behandlungsbeschreibungen als Tuberkulose erkennbar ist.
6. Für die Mitarbeit bei der Planung und die Implementierung der Verfahren danken wir Benedikt Adelmann.
7. Zum Beispiel durch die Einbindung von fachlichen Wissensbasen wie etwa den in der Deutschen Nationalbibliothek verfügbaren bibliographischen Metadaten.

Bibliographie

Farrar, Danielle Marie (2016): *The Corpus of Revenge Tragedy (CoRT): Toward Interdisciplining Early Modern Digital Humanities and Genre Analysis*. In: DH 2016 - Conference Abstracts. S. 789–790.

Gaidys, Uta/Gius, Evelyn/Jarchow, Margarete/ Koch, Gertraud/Menzel, Wolfgang/Orth, Dominik/ Zinsmeister, Heike (2017): *Project Description. HermA: Automated Modelling of Hermeneutic Processes*. In: Hamburger Journal für Kulturanthropologie 7 (2017), S. 119–123.

Gius, Evelyn (in Vorbereitung): *Digitale Hermeneutik: Computergestütztes close reading als literaturwissenschaftliches Forschungsparadigma?* In: **Jannidis, Fotis (Hg.)**: *Digitale Literaturwissenschaft*. Metzler.

Herrmann, Berenike/Lauer, Gerhard (2017): *Das „Was-bisher-geschah" von KOLIMO. Ein Update zum Korpus der literarischen Moderne*. In: DHd 2017 Digitale Nachhaltigkeit Konferenzabstracts. S. 107–111.

Herrmann, Berenike/Lauer, Gerhard (2016a): *Aufbau und Annotation des Kafka/Referenzkorpus*. In: DHd 2016 Modellierung - Vernetzung - Visualisierung Konferenzabstracts. S. 158–159.

Herrmann, Berenike/Lauer, Gerhard (2016b): *KARREK: Building and Annotating a Kafka/Reference Corpus*. In: DH 2016 - Conference Abstracts. S. 552–553.

Levenshtein, Vladimir (1966): *Binary codes capable of correcting deletions, insertions, and reversals*. In: Soviet Physics Doklady. Vol. 10, No. 8, S. 707–710.

Pernes, Stefan/Keller, Lennart/Peterek, Christoph (2017): *Aufbau eines historisch-literarischen Metaphernkorpus für das Deutsche*. In: DHd 2017 Digitale Nachhaltigkeit Konferenzabstracts. S. 92–94.

Trilcke, Peer/Fischer, Frank (2018): *Literaturwissenschaft als Hackathon. Zur Praxeologie der Digital Literary Studies und ihren epistemischen Dingen*. In: Zeitschrift für digitale Geisteswissenschaften, Sonderband 3 "Wie Digitalität die Geisteswissenschaften verändert. Neue Forschungsgegenstände und Methoden".

Schöch, Christof (2017): *Aufbau von Datensammlungen*. In: **Jannidis, Fotis/Kohle, Hubertus/Rehbein, Malte (Hg.)**: Digital Humanities: eine Einführung. Stuttgart: J.B. Metzler Verlag. S. 223–233.

Schruhl, Friederike (2018): *Objektumgangsnormen in der Literaturwissenschaft*. In: Zeitschrift für digitale Geisteswissenschaften, Sonderband 3 "Wie Digitalität die Geisteswissenschaften verändert. Neue Forschungsgegenstände und Methoden".