Algorithmische Korpusarchäologie: Eine Git-basierte Analyse von Korpora als dynamische epistemische Objekte in den Computational Literary Studies

Börner, Ingo
ingo.boerner@uni-potsdam.de
Universität Potsdam, Deutschland
ORCID: 0000-0001-8294-2541

In den Computational Literary Studies (CLS) hat sich das ‚Korpus' als eines der zentralen epistemischen Objekte herauskristallisiert (vgl. Gavin 2023) Trotz der großen Bedeutung, die dem Korpus naturgemäß für die Forschung zukommt, findet jedoch, wie auch Piotrowski (2022) anmerkt, in den CLS kaum eine Diskussion über Korpora und ihr Zustandekommen statt. Dabei sind insbesondere die technischen Rahmenbedingungen ausgesprochen günstig, um evidenzbasiert die Genese von Korpora beleuchten zu können. Viele der in den CLS verwendeten Korpora werden auf den Plattformen *GitHub* oder *GitLab* entwickelt[1], auf denen der Entstehungsprozess in Form von ‚Commits' mehr oder weniger vollständig transparent und nachvollziehbar ist. Dennoch werden die reichhaltigen technischen Metadaten über diese Entstehungsprozesse der Korpusdaten bisher kaum für die Forschung fruchtbar gemacht.

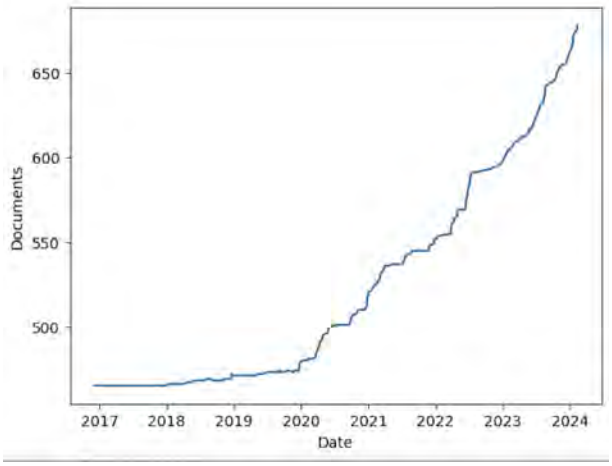
Das mag auch daran liegen, dass die Funktion von Code-Versionierungssystemen, wie *Git*, in den Informationswissenschaften eine rein pragmatische ist: Versionierungssysteme unterstützen das kollaborative Arbeiten am Quellcode und werden allenfalls zur Nachverfolgung von Fehlern verwendet. Für die Digital Humanities bieten sie darüber hinaus aber eine weitere vielversprechende Möglichkeit, nämlich konkret für einen Forschungsansatz, der auf die Historisierung und das Herausarbeiten der ‚Gemachtheit' seiner Forschungsgegenstände, der Daten, abzielt. So ermöglicht eine Analyse der *Git Commit History* beispielsweise die detaillierte Nachverfolgung der Entstehungsprozesse von Korpora.

In dem im Rahmen des *CLS INFRA* Projektes (https://cls-infra.io) entstandenen Report "On Versioning Living and Programmable Corpora" (Börner und Trilcke 2024) konnte gezeigt werden, dass Git-Commits als stabile Referenzen auf Versionen lebender Korpora fungieren können und somit einen Mechanismus für eine transparente Versionierung darstellen. Im oben genannten Report werden DraCor Korpora als „lebende Korpora" beschrieben, die über den Zeitraum ihres Bestehens ‚gewachsen' und sich auch in Bezug auf das in ihnen verwendete Inventar an TEI-XML Tags und den verwendeten Kodierungsstrategien gewandelt haben.

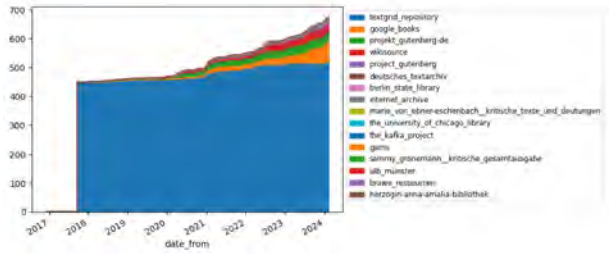
Die angewandte Methode, die hier „algorithmische Korpus-Archäologie" genannt wird, kann diese Veränderungsprozesse durch eine detaillierte Analyse der Git Commit History herausarbeiten. Für die Analyse der *DraCor* Corpora (Fischer et al. 2019) wurde eine Reihe von Python-Skripten entwickelt, mit deren Hilfe die *RESTful API* der Plattform *GitHub* abgefragt und dadurch eine Vielzahl zusätzlicher technischer Metadaten über den Arbeitsprozess an den Korpora bezogen werden kann. Eine Analyse dieser Daten ermöglicht ein tiefergehendes Verständnis der Genese und der strukturellen Eigenschaften der Korpora und liefert insbesondere Einblicke in die Prozesse der Integration von Daten aus unterschiedlichen Quellen.

Im Rahmen der vorgeschlagenen Poster-Präsentation soll diese Methode am Beispiel des *Deutschen Dramenkorpus* (*GerDraCor*) erläutert und in Form eines interaktiven Showcases in einem „Corpus Archeology Dashboard" vorgeführt werden. Grundlage ist die umfangreiche Git Commit Historie des GitHub-Repositories (https://github.com/dracor-org/gerdracor) die ca. 1.500 commits umfasst. Die über die GitHub API abgerufenen und prozessierten Daten werden genutzt, um die Entwicklung des Korpus detailliert nachzuvollziehen und analysieren zu können. Dabei werden verschiedene Aspekte der Korpusgenese betrachtet: Die Integration neuer Texte, die Korrektur von Fehlern, Umstellungen im Markup, die Anpassung der Metadaten zu Texten, usw.

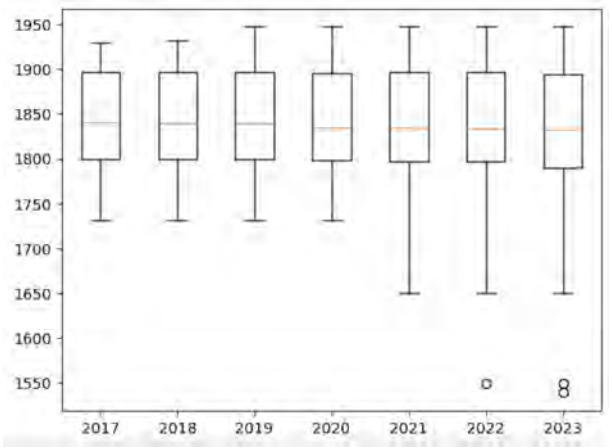
Beispielsweise lassen sich das ‚Korpuswachstum' und die Diversifizierung der digitalen Quellen in verschiedenen Grafiken visualisieren: Abbildung 1 zeigt die Entwicklung der Anzahl der Stücke, die in den Korpusversionen von *GerDraCor* enthalten sind. Aus der Grafik ist ersichtlich, dass das Korpus ab 2018 nur langsam wächst und erst ab 2020 die Anzahl der pro Jahr hinzugefügten Stücke deutlich zunimmt. Dieses ‚Wachstum' resultiert – bis zu einem gewissen Grad – auch aus der Diversifizierung der digitalen Quellen, aus denen Stücke in das Korpus aufgenommen werden. Abbildung 2 zeigt die Verteilung der verwendeten Quellen im Laufe der Zeit. Dabei hat sich das *GerDraCor*-Korpus auch hinsichtlich des durch die Stücke abgedeckten Zeitraums erweitert (siehe Abbildung 3).



Entwicklung der Anzahl an Korpusdokumenten in allen Versionen in *GerDraCor*

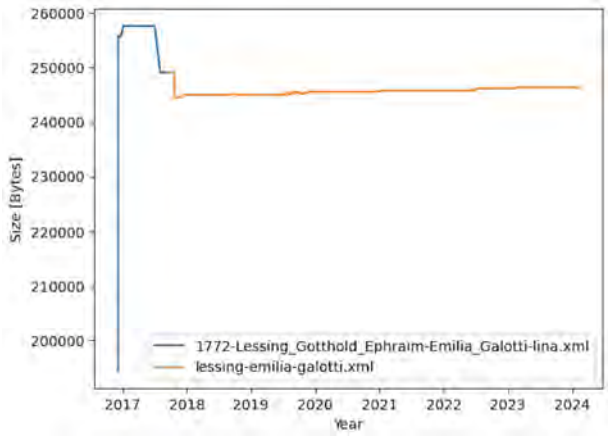


Anzahl der Stücke nach Quelle über den Zeitverlauf



Entwicklung des in *GerDraCor* abgedeckten Zeitraums (property "YearNormalized")

Die grafische Darstellung der Dateigröße einer einzelnen im Korpus enthaltenen Datei über einen bestimmten Zeitraum hinweg ist ferner nützlich, um zu verstehen, ob und wann eine einzelne Datei geändert wurde. Als Beispiel wird in Abbildung 4 die Dateigröße der einzelnen Versionen der TEI-XML-Datei des Stücks „Emilia Galotti“ von Gotthold Ephraim Lessing dargestellt.



Entwicklung der Dateigröße der Datei zu „Emilia Galotti“

Diese Beispiele zeigen im Ansatz die Möglichkeiten auf, die eine Korpusarchäologie auf der Grundlage von Informationen aus Versionierungssystemen bieten kann. Ein Wissen um die Dynamik und Veränderlichkeit von Korpora spielt insbesondere auch für Fragen der Reproduzierbarkeit von datenbasierten Studien eine entscheidende Rolle. Das entwickelte Tool, das im Rahmen der Posterpräsentation vorgestellt werden soll, kann die Gewinnung und Analyse der GitHub Commit History und der in einem Repositorium enthaltenen Daten einfacher als bisher zugänglich machen.

Fußnoten

1. Dies betrifft neben den DraCor Korpora beispielsweise die *ELTeC* -Korpora (Odebrecht et al. 2021; https://github.com/COST-ELTeC), das *Korpus Redewiedergabe* (Brunner et al. 2020; https://github.com/redewiedergabe/corpus), *Modes of Narration and Attribution Corpus* (*MONACO*) (Barth et al. 2021; https://gitlab.gwdg.de/mona/korpus-public) und das *DiBiLit*-Korpus (Boenig/Hug 2021), um nur einige zu nennen. Stabile Korpus-Versionen werden in der Regel in Repositories, wie beispielsweise auf *Zenodo* veröffentlicht, dort sind dann die Bearbeitungsspuren nicht mehr ersichtlich.

Bibliographie

Barth, Florian et al. 2021. *MONACO: Modes of Narration and Attribution Corpus*. https://gitlab.gwdg.de/mona/korpus-public .

Boenig, Matthias und Hug, Marius. (2021). *DiBiLit-Korpus (v3.0)*. https://doi.org/10.5281/zenodo.5786725 .

Börner, Ingo und Peer Trilcke. 2024. *CLS INFRA D7.3 On Versioning Living and Programmable Corpora: (Executable) Report and Prototypes for Reproducible Research.* https://doi.org/10.5281/ZENODO.11081934 .

Brunner, Annelen et al. 2020. „Corpus REDEWIEDERGABE“. In *Proceedings of The 12th*

Language Resources and Evaluation Conference, Marseille, pp. 803 -812.

Fischer, Frank et al. 2019. „Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama“. In *DH2019:»Complexities«. 9–12 July 2019. Book of Abstracts*. Utrecht: Utrecht University. https://doi.org/10.5281/ZENODO.4284002 .

Gavin, Michael. 2023. *Literary mathematics: quantitative theory for textual studies*. Stanford text technologies. Stanford, California: Stanford University Press.

Odebrecht, Carolin et al. 2021. *European Literary Text Collection (ELTeC), version 1.1.0, April 2021, COST Action Distant Reading for European Literary History (CA16204)*. doi.org/10.5281/zenodo.4662444 .

Piotrowski, Michael. 2022. *Epistemological Issues in Digital Humanities*. https://doi.org/10.5281/ZENODO.6498979 .