ALGORITHMISCHE KORPUSARCHÄOLOGIE

EINE GIT-BASIERTE ANALYSE VON KORPORA ALS DYNAMISCHE EPISTEMISCHE OBJEKTE IN DEN COMPUTATIONAL LITERARY STUDIES

DYNAMISCHE “LEBENDE” KORPORA

In den Computational Literary Studies (CLS) wird das Korpus als grundlegendes epistemisches Objekt genutzt. GitHub-basierte Versionierungssysteme bieten eine bislang kaum genutzte Möglichkeit, die Entwicklung von Korpora transparent nachzuverfolgen und für wissenschaftliche Analysen nutzbar zu machen.

Der im Rahmen des EU-Horizon 2020 Projekts **CLS INFRA** (https://clsinfra.io) erstellte Bericht “On Versioning Living and Programmable Corpora” beschreibt die auf der Plattform DraCor (https://dracor.org) veröffentlichten Korpora als “lebende Korpora” – dynamische, offene Textsammlungen, die sich in einem fortlaufenden Prozess der Digitalisierung, Erweiterung und Veränderung befinden. Sie sind nicht statisch abgeschlossen, sondern entwickeln sich kontinuierlich weiter und werden um neue Texte ergänzt.

HERAUSFORDERUNGEN FÜR DIE FORSCHUNG

Dieser dynamische Charakter von DraCor-Korpora stellt eine besondere Herausforderungen für die digitale Dramenforschung dar, insbesondere im Hinblick auf die Reproduzierbarkeit von Studien.
Eine Analyse der mittlerweile über 80 Studien, die Korpora von DraCor nutzen, zeigt, dass zwar meist die Anzahl der Stücke dokumentiert ist, jedoch unklar bleibt, welche spezifische Version des Korpus verwendet wurde.
Für eine exakte Nachvollziehbarkeit und Reproduzierbarkeit der Studien der Datengrundlage wäre es daher essenziell, präzise Informationen zur verwendeten Version eines Korpus zur Verfügung zu haben. Im Falle von DraCor lassen sich Git-Commits nutzen, um Versionen zu identifizieren.

GIT-BASIERTE KORPUSARCHÄOLOGIE

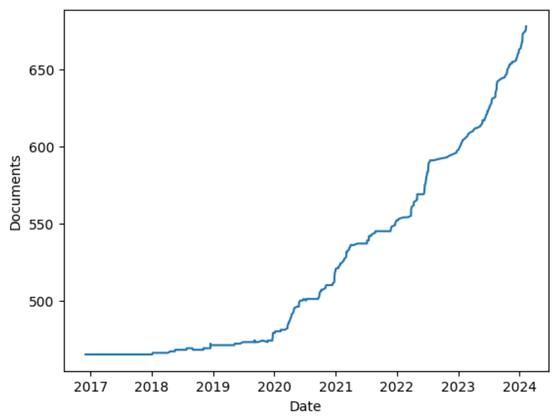
Die Methode der algorithmischen Korpusarchäologie analysiert die Versionierungshistorie von Korpora mittels der GitHub-API. Dies ermöglicht beispielsweise:

- Detaillierte Nachverfolgung von Änderungen (z. B. Korrekturen, Metadatenanpassungen, Markup-Umstellungen)
- Untersuchung des “Korpuswachstums” durch Integration neuer Texte
- Ermittlung der Datenquellen und deren Diversifizierung über den Zeitverlauf
- Visualisierung der Evolution einzelner Dokumente innerhalb eines Korpus

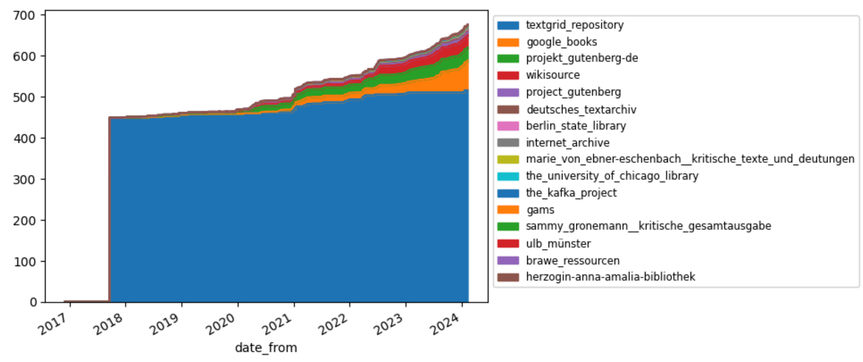
FALLSTUDIE

Das Deutsche Dramenkorpus (GerDraCor) (https://github.com/dracor-org/gerdracor) umfasst ca. 1500 Commits, die eine detaillierte Analyse seiner Entwicklung ermöglichen.

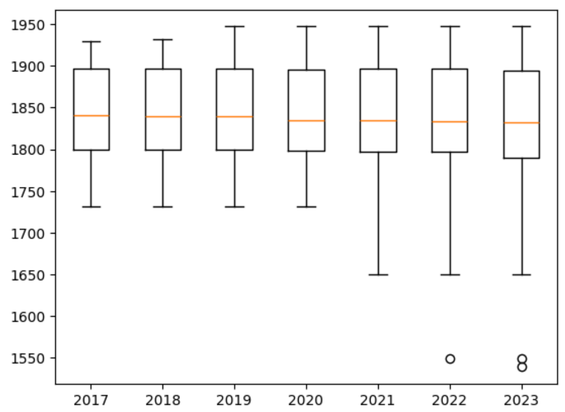
- **Korpuswachstum**: Abb. 1 zeigt ein allmähliches Wachstum ab 2018. Besonders ab 2020 nimmt die Anzahl der jährlich hinzugefügten Dramen deutlich zu. Dies ist vor allem zurückzuführen auf eine
- **Diversifizierung der Quellen**: Frühe Versionen basierten auf den Dramen aus dem TextGrid-Repository, später wurden weitere digitale Quellen integriert, wie Abb. 2 zeigt.
- **Abgedeckter Zeitraum:** Im Laufe seiner Entwicklung erweitert sich GerDraCor auch in Bezug auf den abgedeckten Zeitraum (Abb. 3).
- Die Visualisierung der Dateigröße über einen bestimmten Zeitraum hilft dabei, **Änderungen an einer einzelnen Datei nachzuvollziehen**. Als Beispiel zeigt Abb. 4 die Entwicklung der Dateigröße der verschiedenen Versionen der XML-Datei des Dramas Emilia Galotti von Gotthold Ephraim Lessing.



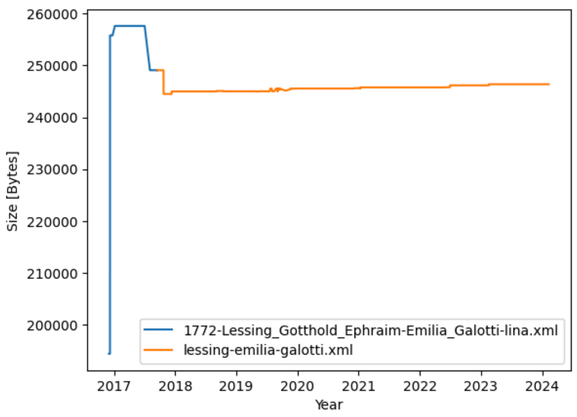
[1] Entwicklung der Anzahl der Dokumente in allen Versionen in GerDraCor



[2] Verteilung der verwendeten Quellen im Laufe der Zeit



[3] Entwicklung des von GerDraCor abgedeckten Zeitraums („YearNormalized“)



[4] Entwicklung der Dateigröße des Stücks Emila Galotti über alle Versionen in GerDraCor

LITERATUR

Börner, Ingo, und Peer Trilcke. „CLS INFRA D7.3 On Versioning Living and Programmable Corpora: (Executable) Report and Prototypes for Reproducible Research“, 27. Februar 2024. https://doi.org/10.5281/ZENODO.11081934.

Ingo Börner
Universität Potsdam
ingo.boerner@uni-potsdam.de

CLSINFRA COMPUTATIONAL LITERARY STUDIES INFRASTRUCTURE