Einführung in DraCor
Programmable Corpora für die digitale Dramenanalyse

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

Fischer, Frank
frank.fischer@dariah.eu
National Research University Higher School of Economics Moscow; DARIAH-EU

Milling, Carsten
cmil@hashtable.de
Universität Potsdam

Sluyter-Gäthje, Henny
sluytergaeth@uni-potsdam.de
Universität Potsdam

Zielstellung des Workshops

In dem ganztägigen Workshop wird DraCor (https://dracor.org), eine offene Plattform zur Erforschung von Dramen in verschiedenen Sprachen, vorgestellt und anhand von praktischen Beispielen aus der digitalen Dramenanalyse erprobt. Im Zentrum von DraCor stehen so genannte ‚Programmable Corpora'. Hierunter verstehen wir infrastrukturell-forschungsorientierte, offene, erweiterbare, Linked-Open-Data-freundliche Volltextkorpora, die es ermöglichen sollen, auf niederschwellige Weise diverse Forschungsfragen aus dem Bereich der digitalen Literaturwissenschaft anhand von Korpora datenbasiert, nachvollziehbar und reproduzierbar zu bearbeiten (Fischer u. a. 2019).

Der Workshop richtet sich an Personen, die

- mit literarischen Texten und insbesondere mit Dramen arbeiten oder arbeiten möchten und dazu eigene Korpora erstellen oder bereits vorhandene Korpora nachnutzen möchten;
- Methoden der digitalen Dramenanalyse (Netzwerkanalyse, Stilometrie) erlernen oder auf Basis des Programmable Corpora-Ansatzes erproben wollen;
- Interesse an den Möglichkeiten zur Erforschung von literarischen Texten mithilfe von Linked Open Data (LOD) haben.

Es erfolgt eine Vorstellung des Konzepts der ‚Programmable Corpora' sowie einer Demonstration der exemplarischen Umsetzung in der Plattform DraCor inklusive einer Vorstellung aller Komponenten. In Form von Hands-on-Tutorials wird den Teilnehmer*innen eine praktische Einführung in das Erstellen und Kuratieren eigener Dramenkorpora zur Analyse mit DraCor geben. Ein weiterer Teil führt anhand praktischer Beispiele zu den Methoden Stilometrie und Netzwerkanalyse in die Verwendung der DraCor-API sowie der Python-Bibliothek PyDraCor ein. Die API-Schnittstelle (Application Programming Interface) ermöglicht den maßgeschneiderten direkten Zugriff auf bestimmte Teile der Korpora. Die Möglichkeiten zu korpusübergreifenden Abfragen und Einbeziehung von Informationen aus der Linked-Open-Data-Cloud mit SPARQL werden ebenso erprobt.

Das Konzept der ‚Programmable Corpora'

Den Kern von DraCor bilden Korpora von Dramen in elf Sprachen (Deutsch, Russisch, Französisch, Italienisch, Schwedisch, Spanisch, Altgriechisch, Elsässisch, Lateinisch, Baschkirisch und Tatarisch) sowie zwei weitere Autoren-Korpora (Shakespeare, Calderón), zu denen die Plattform eine Vielzahl an möglichen Forschungszugängen bietet: Die Dramen sind als XML-Dateien entsprechend der TEI-Guidelines kodiert und unter einer offenen Lizenz frei über GitHub unter https://github.com/dracor-org verfügbar. Sie können von dort geladen, gegebenenfalls selbst transformiert oder angereichert und zur weiteren Beforschung in beliebigen Tools weiterverwendet werden.

Neben diesem „klassischen" modus operandi der korpusbasierten Forschung bietet DraCor als offenes digitales Ökosystem jedoch noch weitere Schnittstellen und angeschlossene Tools (Netzwerkvisualisierungen, Shiny App, Easy Linavis). Grundlegend hierfür ist die DraCor REST API (https://dracor.org/doc/api), die sowohl Funktionen zum Abrufen der Daten in unterschiedlichen Formaten (TEI, JSON, Plaintext, RDF, GEXF, GraphML) als auch einige eingebaute Analysefunktionalitäten (bspw. zu Netzwerkmetriken) bereitstellt. Über die API können neben Struktur- und Metadaten auch die Volltexte ohne weiteres Markup abgerufen werden, um so ohne weiteren Zwischenschritt zur Entfernung von Markup Methoden wie stilometrische Analysen oder Topic Modeling anzuwenden. Die DraCor API ist im OpenAPI-Standard dokumentiert und kann in einer mittels Swagger UI implementierten interaktiven Dokumentation (https://dracor.org/documentation/api) direkt aus dem Webbrowser heraus verwendet werden.

Für die Programmiersprachen Python (PyDraCor: https://github.com/dracor-org/pydracor) und R (rdracor: https://github.com/dracor-org/rdracor) sind API-Bibliotheken verfügbar, die eine schnelle und auf die jeweilige Programmiersprache angepasste Einbindung der API-Funktionalitäten ermöglichen. Für komplexe Abfragen steht auf der Plattform ein SPARQL-Endpoint (https://dracor.org/sparql) zur Verfügung. Hierüber sind sowohl korpusübergreifende als auch kombinierte Abfragen (federated queries) möglich, bei denen DraCor gleichzeitig mit anderen als LOD verfügbaren Ressourcen, wie beispielsweise Wikidata, abgefragt werden kann.

Digitale Dramenanalyse mit DraCor

Korpusbasierte, in der Regel quantitative Methoden verwendende Analysen von Dramen haben sich in den vergangenen Jahren zu einem eigenen Subfeld der Computational Literary Studies (CLS) entwickelt (vgl. Willand et al. 2017; Reiter 2021). Dabei hat sich die Bereitstellung gemeinsam kuratierter und offener Ressourcen wie DraCor als produktiv auch für angrenzende Disziplinen wie die Computerlinguistik erwiesen (vgl. beispielsweise Pagel, Reiter 2020).

Auf Wortebene operierende Verfahren haben sich dabei etwa auf die Autorschaftsattribution (Schöch 2014) oder Genreklassifikation mit Topic Modeling (Schöch 2017) fokussiert. Aktuell werden vielversprechende Neukonzeptualisierungen stilometri-

scher Maße wie das Kontrastmaß Zeta entwickelt und angewendet (Schöch 2018). Auf der Grundlage von strukturell ausgezeichneten Korpora lassen sich darüber hinaus gezielte Analysen etwa von Bühnenanweisungen durchführen, die mit POS-Informationen oder semantischen Feldern operieren (Trilcke et al. 2020).

Im Bereich der strukturellen Analyse wurden Dramenkorpora früh schon, beginnend mit den Arbeiten von Stiller, Nettle, Dunbar (2003) und fortgesetzt etwa bei Moretti (2011), mit netzwerkanalytischen Ansätzen untersucht. Typologische Arbeiten beispielsweise zum Konzept der Small Worlds (Trilcke et al. 2016) stehen hier u.a. neben Ansätzen zur quantitativen Klassifizierung von Figurentypen (Fischer et al. 2018).

Wenngleich semantische Technologien mittlerweile zum festen Bestandteil des Methodenspektrums der Digitalen Geisteswissenschaften zählen, gelangen sie in den korpusbasierten CLS bisher selten Anwendung (zu Prosa bspw. Frank und Ivanovic 2018; Dittrich 2017). Die Erfassung von Metadaten als Linked Data und die Anbindung an externe Referenzressourcen, insbesondere Wikidata, ermöglichen jedoch weitreichende Abfragemöglichkeiten und lassen sich zur Analyse von literarischen Korpora gewinnbringend nutzen. Beispielsweise sind in den DraCor-Korpusdaten keine detaillierten Informationen zu Autor*innen und Aufführungsorten enthalten. Da aber zu den einzelnen Stücken die eindeutigen Wikidata-Identifikatoren hinterlegt sind, können diese Informationen per federated queries in SPARQL abgerufen und in unterschiedlichen Visualisierungsformen, wie zum Beispiel als Karte, dargestellt werden.

Lernziele und Ablauf des Workshops

Im ersten Teil des Workshops wird zunächst das Konzept der ‚Programmable Corpora' eingeführt und diskutiert. Daran anschließend werden die Plattform DraCor und die einzelnen Komponenten vorgestellt, wobei auch immer wieder kürzere Übungsphasen vorgesehen sind, in denen die Teilnehmer*innen die vorgestellten Komponenten und Tools unmittelbar ausprobieren können. Insbesondere werden die unterschiedlichen Möglichkeiten zum Bezug und zur Analyse der Korpusdaten erprobt. Ein Fokus liegt dabei auf der Verwendung der API. Anhand der interaktiven Dokumentation werden die API-Funktionalitäten erläutert und können von den Teilnehmer*innen ausgiebig getestet werden. Im Anschluss daran wird ein kurzer Überblick zur Korpuserstellung und zu den Besonderheiten der TEI-Kodierung geben, wie sie in DraCor zum Einsatz kommen.

Den zweiten Teil des Workshops bilden Gruppenarbeitsphasen, in denen drei Themenbereiche vertieft werden können:

(1) Korpuserstellung und -kuratierung mit DraCor: Die Teilnehmenden vertiefen die TEI-Kodierung von Dramen anhand von praktischen Übungen und lernen, wie eine lokale Instanz der Plattform mittels Docker aufgesetzt, gegebenenfalls angepasst und mit eigenen Korpora bestückt werden kann.

(2) Dramenanalyse mit DraCor-API und Python: Mittels Jupyter Notebooks mit ausführlich dokumentiertem Python-Programmcode werden die Teilnehmer*innen an Methoden der digitalen Dramenanalyse unter Verwendung der DraCor-API herangeführt. Die Notebooks sollen es auch Teilnehmer*innen, die bisher noch keine Erfahrungen im Programmieren mit Python gemacht haben, im Sinne eines Literate-Programming-Ansatzes ermöglichen, die einzelnen Analyseschritte nachzuvollziehen und auch selbst adaptieren zu können. Die Notebooks setzen konkrete Forschungsfragen zur Dramenanalyse um, etwa zur literaturhistorischen Entwicklung netzwerkanalytischer Maße oder zur quantitativen Dominanz von Figuren.

(3) Dramenanalyse mit Linked Data: Den Schwerpunkt bilden praktische Analysen, die aus der Anbindung von DraCor an die Linked Open Data Cloud möglich werden. Im Workshop wird ein kurzer Crashkurs in die Abfragesprache SPARQL gegeben, um dann im Anschluss gemeinsame Abfragen von DraCor und Wikidata vorzunehmen und die Ergebnisse zu visualisieren.

Die Ergebnisse der Arbeitsgruppen werden anschließend im Plenum präsentiert und diskutiert.

Organisatorisches

Anzahl der möglichen Teilnehmer*innen: 25

Teilnehmer*innen benötigen einen eigenen Laptop mit Internetzugang; Hinweise zu vorab zu installierender Software (Oxygen XML-Editor, Docker, ...) werden im Vorfeld bekanntgegeben. Die Materialien werden auf GitHub bereitgestellt; die Jupyter Notebooks werden unter (https://github.com/dracor-org/dracor-notebooks) veröffentlicht.

Weitere benötigte technische Ausstattung am Veranstaltungsort: Beamer, WLAN

Beitragende / Kontaktdaten

Ingo Börner (ingo.boerner@uni-potsdam.de) arbeitet als wissenschaftlicher Mitarbeiter im Projekt „CLSInfra" an der Universität Potsdam an der Weiterentwicklung von DraCor. Seine Arbeitsschwerpunkte umfassen Datenmodellierung und Linked Open Data.

Frank Fischer (frank.fischer@dariah.eu) ist Associate Professor an der Higher School of Economics in Moskau und einer der Direktoren von DARIAH. Seine Beschäftigung mit digitaler Dramenanalyse geht zurück auf das Projekt zur Digitalen Literaturwissenschaftlichen Netzwerkanalyse DLINA (https://dlina.github.io), aus dem DraCor hervorgegangen ist.

Carsten Milling (cmil@hashtable.de **)** ist Webdeveloper und ist im Projekt „CLSInfra" an der Universität Potsdam für die Entwicklung der DraCor-Plattform zuständig.

Henny Sluyter-Gäthje (sluytergaeth@uni-potsdam.de) ist wissenschaftliche Mitarbeiterin an der Professur für deutsche Literatur des 19. Jahrhunderts an der Universität Potsdam. Sie hat ein Masterstudium of Science in Cognitive Systems mit Schwerpunkt Computerlinguistik abgeschlossen und arbeitet an der algorithmischen Verarbeitung literarischer Texte.

Fördernachweis

DraCor wird gegenwärtig im Rahmen des von der EU Horizon 2020 geförderten Projekts „CLSInfra" (Fördernummer: 101004984, https://cordis.europa.eu/project/id/101004984) weiterentwickelt.

Bibliographie

Dittrich, Andreas (2017): "Intra-Connecting an Exemplary Literary Corpus with Semantic Web Technologies for Exploratory Literary Studies" in: Bański, Piotr et al. (Hg.): *Proceedings of the*

Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017. Mannheim: Institut für Deutsche Sprache. https://nbn-resolving.org/urn:nbn:de:bsz:mh39-62441.

Fischer, Frank / Trilcke, Peer / Kittel, Christopher / Milling, Carsten / Skorinkin, Daniil (2018): "To catch a protagonist: Quantitative dominance relations in german-language drama (1730–1930)" in: Digital Humanities 2018. Conference Abstracts. Mexico City: El Colegio de México / Universidad Nacional Autonoma de México / Red de Humanidades Digitales 193–201.

Fischer, Frank / Börner, Ingo / Göbel, Mathias / Hechtl, Angelika / Kittel, Christopher / Milling, Carsten / Trilcke, Peer (2019): "Programmable Corpora: Die digitale Literaturwissenschaft zwischen Forschung und Infrastruktur am Beispiel von DraCor " in: *DHd2019: »Digital Humanities: multimedial & multimodal«. Book of Abstracts*. Mainz/Frankfurt a. M.: Johannes Gutenberg Universität Mainz / Goethe Universität Frankfurt, 194–197.

Frank, Andrew / Ivanovic, Christine (2018): "Building Literary Corpora for Computational Literary Analysis – A Prototype to Bridge the Gap between CL and DH" in: Calzolari, Nicoletta et al. (Hg.): *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association.

Moretti, Franco (2011): "Network Theory, Plot Analysis " in: *Stanford Literary Lab Pamphlets* 2. http://litlab.stanford.edu/LiteraryLabPamphlet2.pdf [letzter Zugriff 13.7.2021].

Pagel, Janis / Reiter, Nils (2020): "GerDraCor-Coref: A Coreference Corpus for Dramatic Texts in German" in: *Proceedings of the Language Resources and Evaluation Conference (LREC).* Marseille 55-64 http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.7.pdf [Letzter Zugriff: 15.7.2021].

Reiter, Nils (2021): "Möglichkeiten Quantitativer Dramenanalyse" in: *Comparatio. Zeitschrift für Vergleichende Literaturwissenschaft* 12(2): 39–52.

Schöch, Christof (2017): "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama" in: *Digital Humanities Quarterly* 11, Nr. 2 http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html [Letzter Zugriff: 15.7.2021].

Schöch, Christof (2018): "Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie" in: Bernhart, Toni et al. (Hg): *Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven*. Berlin: de Gruyter 77–94 doi: 10.1515/9783110523300-004.

Schöch, Christof (2014): "Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik" in: Schneider, Lars / Schöch, Christof (Hg.): *Literaturwissenschaft im digitalen Medienwandel*. Beihefte zu Phin 7 http://web.fu-berlin.de/phin/beiheft7/b7t08.pdf [Letzter Zugriff: 15.07.2021].

Stiller, James / Nettle, Daniel / Dunbar, Robin I. M. (2003): "The Small World of Shakespeare's Plays" in: *Human Nature* 14: 397–408.

Trilcke, Peer / Fischer, Frank / Göbel, Mathias / Kampkaspar, Dario / Kittel, Christopher (2016): "Theatre Plays as ›Small Worlds‹? Network Data on the History and Typology of German Drama, 1730-1930" in: *Digital Humanities 2016. Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków 385-387 https://dh2016.adho.org/abstracts/static/dh2016_abstracts.pdf [Letzter Zugriff: 15.07.2021].

Trilcke, Peer / Kittel, Christopher / Reiter, Nils / Maximova, Daria / Fischer, Frank (2020): "Opening the Stage. A Quantitative Look at Stage Directions in German Drama" in: *Digital Humanities 2020. Conference Abstracts*. Ottawa: University of Ottawa https://dh2020.adho.org/wp-content/uploads/2020/07/337_OpeningtheStageAQuantitativeLookatStageDirectionsinGermanDrama.html [Letzter Zugriff: 15.07.2021].

Willand, Marcus / Trilcke, Peer / Schöch, Christof / Rißler-Pipka, Nanette / Reiter, Nils / Fischer, Frank (2017): "Aktuelle Herausforderungen der Digitalen Dramenanalyse" in: *DHd 2017. Digitale Nachhaltigkeit. Konferenzabstracts*. Bern: Universität Bern 175–180 doi: 10.5281/zenodo.3684825.