Ressourcenbeitrag: DROC: Deutsches Romankorpus

Jan Horstmann [1]

1. Universität Münster

Thema:	Korpusbildung	DOI:	10.48694/fortext.3812
Jahrgang:	1	Ausgabe:	2
Erscheinungsdatum:	12-06-2024	Erstveröffentlichung:	2019-08-19 auf fortext.net
Lizenz:			open ∂ access

Allgemeiner Hinweis: Rot dargestellte Begriffe werden im Glossar am Ende des Beitrags erläutert. Alle externen Links sind auch am Ende des Beitrags aufgeführt.

1. Kurzbeschreibung

Das Deutsche Romankorpus (DROC) (vgl. Korpus) versammelt 90 annotierte (vgl. Annotation) Fragmente deutschsprachiger Romane (jeweils ca. 200 Sätze) vom 17. bis 20. Jahrhundert. Es enthält neben automatisch generiertem Markup (Textauszeichung) zu Kapiteln, Segmenten, Dependenz- und Morphologieinformationen, Wortarten (POS), Sätzen und Absätzen auch über 50.000 manuell erstellte Annotationen zu benannten Entitäten (vgl. Named Entity Recognition (Schumacher 2024)), Koreferenzen, direkter Rede, sowie Sprechern und Adressaten dieser direkten Rede. Die dichte Annotation macht das DROC zu einer guten Ressource für Machine-Learning-Routinen (vgl. Machine Learning) oder die Kombination mit anderen qualitativen Annotationen. Das DROC stellt keine grafische Benutzeroberfläche (vgl. GUI) zur Verfügung, zur Exploration der Daten ist daher etwas technisches Know-How (z. B. über die Formate TEI-XML oder Apache-UIMA-XMI) vonnöten.

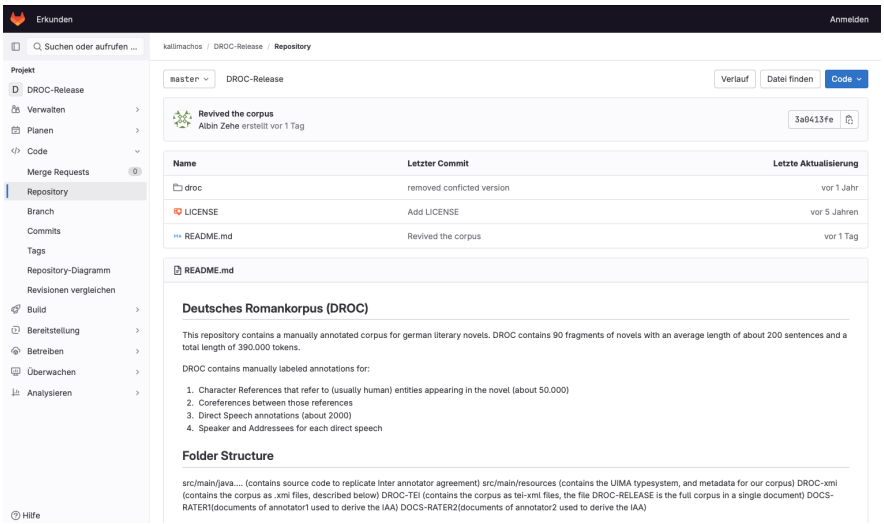

Abb. 1: Übersichtsseite des DROC

Steckbrief

- https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/DROC-Release
- 90 zufällig ausgewählte Fragmente verschiedener deutschsprachiger Romane (auch Übersetzungen); jeweils ca. 200 Sätze; insgesamt ca. 393.000 Tokens (vgl. Type/Token)
- im Projekt Kallimachos (gefördert vom BMBF) an der Universität Würzburg hergestellt
- die Sammlung soll insbesondere bereits vorhandene automatisierte Tools in den Literaturwissenschaften unterstützen und bereichern (Machine Learning)
- Schwerpunkt: Annotation von Figurenreferenzen; enthält manuell erstellte Annotationen für knapp über 50.000 annotierte Figurenreferenzen und ihre Koreferenzen, ca. 2000 Annotationen von direkter Rede und deren jeweiligen Sprechern und Adressaten
- in zwei unterschiedlichen Dokumentformaten erhältlich: TEI-XML und Apache-UIMA-XMI; in den Metadaten werden aufgeführt: Titel, Autor, Jahr, Geschlecht der Autor*innen, Gattung, Erzählerposition, Happyend, Epoche (Jahr und Fachwissenschaft), Strömung, Originalsprache

2. Anwendungsbeispiel

Sie wollen den Einsatz direkter Rede in deutschsprachigen Romanen vergleichend untersuchen. DROC bietet Ihnen für diesen Anwendungsfall ein gründlich annotiertes Korpus aus 90 Romanfragmenten mit Figuren-, Koreferenz-, direkter-Rede- inkl. Sprecher- und Adressatenannotationen.

3. Diskussion

3.1 Kann ich das DROC für wissenschaftliche Arbeiten nutzen?

Ja. Die Texte entstammen dem TextGrid Repository (Horstmann 2024), das Preprocessing sowie Annotation- und Auswertungsregeln für das DROC werden transparent gemacht. Die Textauswahl erfolgte zufällig aus 450 kanonisierten Texten als auch aus der Sammlung „Deutsche Literatur von Frauen“. Die Fragmente aus diesen beiden Textgruppen wurden ebenfalls zufällig ausgewählt. Dieses Preprocessing wird in Krug u. a. (2018, Abschn. 4) dokumentiert und begründet.

Die annotierten Textfragmente entstammen Romanen aus der Zeit zwischen dem 17. und dem 20. Jahrhundert, die orthographisch - bis auf neun Texte aus der Zeit von 1650-1800 - zum Großteil nicht standardisiert sind. Die Texte können anhand ihrer Metadaten gefiltert werden. Vertreten sind zu 60% männliche und zu 40% weibliche Autor*innen sowie kanonisierte und unbekannte Texte. Die Annotationen wurden mithilfe eines vorab erstellten Annotator-Agreements in der vom Kallimachos-Projekt selbst programmierten Desktop-Applikation ATHEN manuell und semi-automatisch erstellt.

Da es beim DROC um die Qualität der Metadaten und nicht um die Primärtexte geht, besteht bei den ausgewählten Texten kein Anspruch auf Vollständigkeit. Die Volltexte der Romane können bei Bedarf im Textgrid Repository eingesehen und heruntergeladen werden. Die Texte unterliegen der Creative Commons License CC-BY und können mit entsprechender Zitation als Quellenangabe genutzt werden.

Die hochwertigen Metadaten können zukünftig durch weitere qualitative Annotationen ergänzt werden. So wurde beispielsweise bereits ein Subset von 30 Romanfragmenten des DROC für ein kollaboratives Annotationsprojekt (Jacke 2024) zur Klassifikation von Textsorten (deskriptiv, argumentativ oder narrativ) genutzt (Schlör, Schöch und Hotho 2019). Die Kombination unterschiedlicher qualitativer Annotationen in einer Ressource eröffnet die Möglichkeit, neue Fragestellungen digital zu erforschen.

3.2 Wie benutzerfreundlich ist die Arbeit mit dem DROC?

Die Arbeit mit den Texten und Annotationen des DROC setzt technische Kenntnisse voraus. Die Strukturen der beiden Datenformate TEI-XML und Apache-UIMA-XMI machen eine Einarbeitung erforderlich, die für die Arbeit mit dem DROC derzeit unumgänglich ist. Das Korpus können Sie in seinen beiden Formaten von der Plattform GitHub herunterladen, deren Nutzung ebenfalls einer gewissen Einarbeitung bedarf.

4. Wie funktioniert die Textsuche im DROC?

Um das Korpus beispielsweise im TEI-XML-Format herunterzuladen, klicken Sie dem Link zur GitHub-Seite20 und wählen dort dann - wie in Abbildung 2 gezeigt - den ZIP-Download oben rechts unter „Download this directory“ aus.

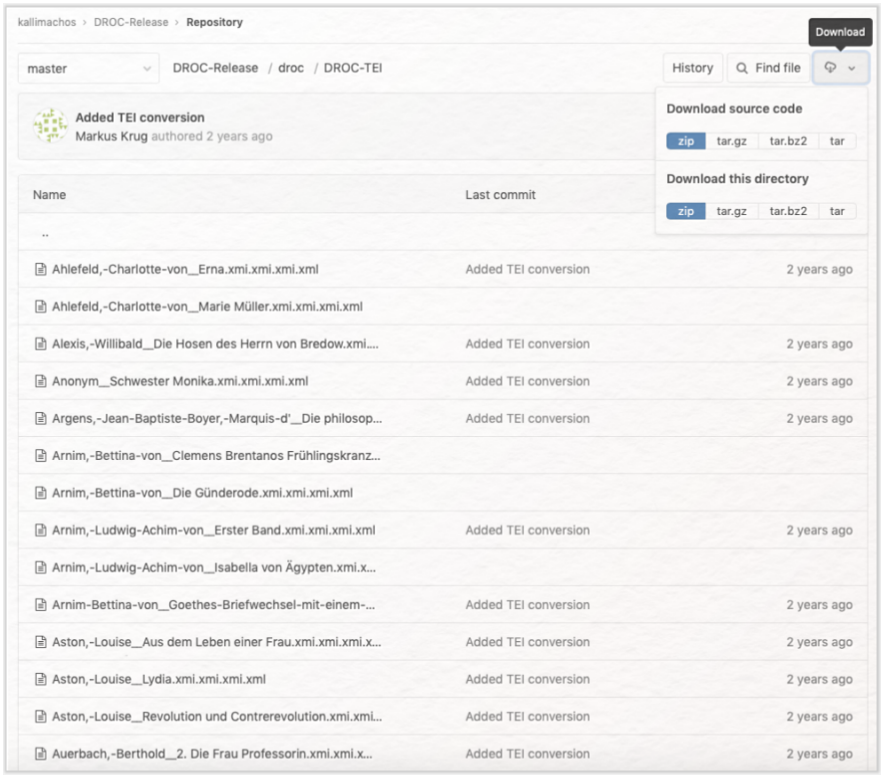

Abb. 2: Download des DROC im TEI-XML-Format

In der oberen Zeile können Sie zudem durch die einzelnen Seiten des GitHub-Repositorys navigieren. Unter DROC-Release finden Sie beispielsweise eine README-Datei, die (auf Englisch) grundlegende Informationen über das DROC versammelt. Wenn Sie an einem bestimmten Romanfragment interessiert sind, können Sie auch einfach den „Find file"-Button (vgl. Query) oben rechts bedienen und die Freitextsuchzeile ausfüllen.

Externe und weiterführende Links

- ATHEN: https://web.archive.org/save/https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/Athen (Letzter Zugriff: 04.06.2024)
- DROC Korpus: https://web.archive.org/save/https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/DROC-Release/tree/master/droc/DROC-TEI (Letzter Zugriff: 04.06.2024)
- Kallimachos: https://web.archive.org/save/http://www.camerarius.uni-wuerzburg.de/kallimachos/index.php/Hauptseite (Letzter Zugriff: 04.06.2024)
- Zitation: https://web.archive.org/save/https://creativecommons.org/licenses/by/3.0/de/ (Letzter Zugriff: 04.06.2024)

Bibliographie

Horstmann, Jan. 2024. Ressourcenbeitrag: TextGrid Repository. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 11. Bibliografie (29. November). doi: 10.48694/fortext.3794, https://fortext.net/ressourcen/textsammlungen/textgrid-repository.

Jacke, Janina. 2024. Methodenbeitrag: Kollaboratives literaturwissenschaftliches Annotieren. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 4. Manuelle Annotation (7. August). doi: 10.48694/fortext.3749, https://fortext.net/routinen/methoden/kollaboratives-literaturwissenschaftliches-annotieren.

Krug, Markus, Frank Puppe, Isabella Reger, Lukas Weimer, Luisa Macharowsky, Stephan Feldhaus und Fotis Jannidis. 2018. Description of a Corpus of Character References in German Novels - DROC [Deutsches ROman Corpus]. DARIAH-DE working papers. http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2018-2-9 (zugegriffen: 5. August 2019).

Schlör, Daniel, Christof Schöch und Andreas Hotho. 2019. Classification of Text-Types in German Novels. In: *Digital Humanities 2019 Conference Papers*. doi: 10.34894/OMLKRN, https://doi.org/10.34894/OMLKRN (zugegriffen: 5. August 2019).

Schumacher, Mareike. 2024. Methodenbeitrag: Named Entity Recognition (NER). Hg. von Evelyn Gius. *forTEXT* 1,

Nr. 9. Named Entity Recognition (30. Oktober). doi: 10.48694/fortext.3765, https://fortext.net/routinen/methoden/named-entity-recognition-ner.

Glossar

Annotation Annotation beschreibt die manuelle oder automatische Hinzufügung von Zusatzinformationen zu einem Text. Die manuelle Annotation wird händisch durchgeführt, während die (teil-)automatisierte Annotation durch Machine-Learning-Verfahren durchgeführt wird. Ein klassisches Beispiel ist das automatisierte PoS-Tagging (Part-of-Speech-Tagging), welches oftmals als Grundlage (Preprocessing) für weitere Analysen wie Named Entity Recognition (NER) nötig ist. Annotationen können zudem deskriptiv oder analytisch sein.

Browser Mit Browser ist in der Regel ein Webbrowser gemeint, also ein Computerprogramm, mit dem das Anschauen, Navigieren auf, und Interagieren mit Webseiten möglich wird. Am häufigsten genutzt werden dafür Chrome, Firefox, Safari oder der Internet Explorer.

Commandline Die Commandline (engl. *command line interface* (CLI)), auch Kommandozeile, Konsole, Terminal oder Eingabeaufforderung genannt, ist die direkteste Methode zur Interaktion eines Menschen mit einem Computer. Programme ohne eine grafische Benutzeroberfläche (GUI) werden i. d. R. durch Texteingabe in die Commandline gesteuert. Um die Commandline zu öffnen, klicken Sie auf Ihrem Mac „cmd“ + „space“, geben „Terminal“ ein und doppelklicken auf das Suchergebnis. Bei Windows klicken Sie die Windowstaste + „R“, geben „cmd.exe“ ein und klicken Enter.

GUI GUI steht für *Graphical User Interface* und bezeichnet eine grafische Benutzeroberfläche. Ein GUI ermöglicht es, Tools mithilfe von grafischen Schaltflächen zu bedienen, um somit beispielsweise den Umgang mit der Commandline zu umgehen.

HTML HTML steht für *Hypertext Markup Language* und ist eine textbasierte Auszeichnungssprache zur Strukturierung elektronischer Dokumente. HTML-Dokumente werden von Webbrowsern dargestellt und geben die Struktur und Online-Darstellung eines Textes vor. HTML-Dateien können außerdem zusätzliche Metainformationen enthalten, die auf einer Webseite selbst nicht ersichtlich sind.

Korpus Ein Textkorpus ist eine Sammlung von Texten. Korpora (Plural für „das Korpus“) sind typischerweise nach Textsorte, Epoche, Sprache oder Autor*in zusammengestellt.

Lemmatisieren Die Lemmatisierung von Textdaten gehört zu den wichtigen Preprocessing-Schritten in der Textverarbeitung. Dabei werden alle Wörter (Token) eines Textes auf ihre Grundform zurückgeführt. So werden beispielsweise Flexionsformen wie „schneller“ und „schnelle“ dem Lemma „schnell“ zugeordnet.

Machine Learning Machine Learning, bzw. maschinelles Lernen im Deutschen, ist ein Teilbereich der künstlichen Intelligenz. Auf Grundlage möglichst vieler (Text-)Daten erkennt und erlernt ein Computer die häufig sehr komplexen Muster und Gesetzmäßigkeiten bestimmter Phänomene. Daraufhin können die aus den Daten gewonnen Erkenntnisse verallgemeinert werden und für neue Problemlösungen oder für die Analyse von bisher unbekannten Daten verwendet werden.

Markup (Textauszeichung) Die Textauszeichnung (eng. *Markup*) fällt in den Bereich der Daten- bzw. Textverarbeitung, genauer in das Gebiet der Textformatierung, welche durch Auszeichnungssprachen wie XML implementiert wird. Dabei geht es um die Beschreibung, wie einzelne Elemente eines Textes beispielsweise auf Webseiten grafisch dargestellt werden sollen.

Markup Language Markup Language bezeichnet eine maschinenlesbare Auszeichnungssprache, wie z.B. HTML, zur Formatierung und Gliederung von Texten und anderen Daten. So werden beispielsweise auch Annotationen durch ihre Digitalisierung oder ihre digitale Erstellung zu Markup, indem sie den Inhalt eines Dokumentes strukturieren.

Metadaten Metadaten oder Metainformationen sind strukturierte Daten, die andere Daten beschreiben. Dabei kann zwischen administrativen (z. B. Zugriffsrechte, Lizenzierung), deskriptiven (z. B. Textsorte), strukturellen (z. B. Absätze oder Kapitel eines Textes) und technischen (z. B. digitale Auflösung, Material) Metadaten unterschieden werden. Auch Annotationen bzw. Markup sind Metadaten, da sie Daten/Informationen sind, die den eigentlichen Textdaten hinzugefügt werden und Informationen über die Merkmale der beschriebenen Daten liefern.

Named Entities Eine Named Entity (NE) ist eine Entität, oft ein Eigenname, die meist in Form einer Nominalphrase zu identifizieren ist. Named Entities können beispielsweise Personen wie „Nils Holgerson“, Organisationen wie „WHO“ oder Orte wie „New York“ sein. Named Entities können durch das Verfahren der Named Entity Recognition (NER) automatisiert ermittelt werden.

POS PoS steht für *Part of Speech* , oder „Wortart“ auf Deutsch. Das PoS- Tagging beschreibt die (automatische) Erfassung und Kennzeichnung von Wortarten in einem Text und ist of ein wichtiger Preprocessing-Schritt, beispielsweise für die Analyse von Named Entities.

Preprocessing Für viele digitale Methoden müssen die zu analysierenden Texte vorab „bereinigt“ oder „vorbereitet“ werden. Für statistische Zwecke werden Texte bspw. häufig in gleich große Segmente unterteilt (*chunking*), Großbuchstaben werden in Kleinbuchstaben verwandelt oder Wörter werden lemmatisiert.

Query *Query* bedeutet „Abfrage“ oder „Frage“ und bezeichnet eine computergestützte Abfrage zur Analyse eines Textes. Um Datenbestände zu durchsuchen, werden Abfragesprachen eingesetzt, die *Queries* (Anfragen) an den Datenbestand senden. So bilden alle möglichen Queries zusammen die *Query Language* eines Tools.

TEI Die *Text Encoding Initiative* (TEI) ist ein Konsortium, das gemeinsam einen Standard für die Darstellung von Texten in digitaler Form entwickelt. Die TEI bietet beispielsweise Standards zur Kodierung von gedruckten Werken und zur Auszeichnung von sprachlichen Informationen in maschinenlesbaren Texten (siehe auch XML und Markup).

Type/Token Das Begriffspaar „Type/Token“ wird grundsätzlich zur Unterscheidung von einzelnen Vorkommnissen (Token) und Typen (Types) von Wörtern oder Äußerungen in Texten genutzt. Ein Token ist also ein konkretes Exemplar eines bestimmten Typs, während ein Typ eine im Prinzip unbegrenzte Menge von Exemplaren (Token) umfasst.
Es gibt allerdings etwas divergierende Definitionen zur Type-Token-Unterscheidung. Eine präzise Definition ist daher immer erstrebenswert. Der Satz „Ein Bär ist ein Bär.“ beinhaltet beispielsweise fünf Worttoken („Ein“, „Bär“, „ist“, „ein“, „Bär“) und drei Types, nämlich: „ein“, „Bär“, „ist“. Allerdings könnten auch vier Types, „Ein“, „ein“, „Bär“ und „ist“, als solche identifiziert werden, wenn Großbuchstaben beachtet werden.

XML XML steht für *Extensible Markup Language* und ist eine Form von Markup Language, die sowohl computer- als auch menschenlesbar und hochgradig anpassbar ist. Dabei werden Textdateien hierarchisch strukturiert dargestellt und Zusatzinformationen i. d. R. in einer anderen Farbe als der eigentliche (schwarz gedruckte) Text dargestellt. Eine standardisierte Form von XML ist das TEI-XML.

ZIP ZIP steht für ein Dateiformat (zip = engl. Reißverschluss), in welchem mehrere Einzeldateien verlustfrei, komprimiert zusammengefasst werden. ZIP-Dateien werden beim Öffnen entweder automatisch entpackt oder lassen sich per Rechtsklick extrahieren.