Ressourcenbeitrag: KOLIMO: Korpus der literarischen Moderne

Jan Horstmann [1]

1. Universität Münster

forTEXT

Thema:	Korpusbildung	DOI:	10.48694/fortext.3813
Jahrgang:	1	Ausgabe:	2
Erscheinungsdatum:	12-06-2024	Erstveröffentlichung:	2019-02-04 auf fortext.net
Lizenz:	CC BY SA		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 KOLIMO (Korpus der literarischen Moderne) versammelt deutschsprachige narrative, fiktionale Erzähltexte der literarischen Moderne aus den Textsammlungen Deutsches Textarchiv (Horstmann und Kern 2024), TextGrid Repository (Horstmann 2024) und Gutenberg, vereinheitlicht die bestehenden Metadaten und fügt weitere hinzu, um epochenspezifische und aufgrund einheitlicher Daten verlässliche Abfrageergebnisse erhalten zu können.

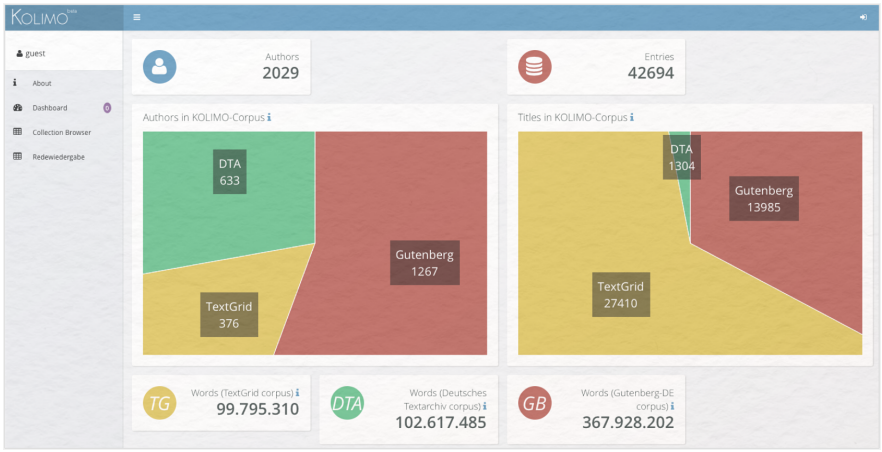

Abb. 1: KOLIMO-Benutzeroberfläche

Steckbrief

- Das Korups steht Ihnen auf Gitlab und Zenodo zur Verfügung
- großes Spektrum der literarischen Moderne: literarische und nicht-literarische Texte sowie Vergleichstexte früherer Epochen
- Teil des laufenden Projekts Q-LIMO (Quantitative Analyse der literarischen Moderne)
- seit Herbst 2016 ist die Beta-Version veröffentlicht
- Quellen: TextGrid Repository, Gutenberg-DE, Deutsches Textarchiv (DTA), Kafka-Referenzkorpus (vgl. Korpus)
- konsistente und manuell erweiterte Metadaten zu bspw. Autor*innen, Publikationsdatum und Gattung und grundlegende linguistische Annotation
- Textsorten: verschiedenen Genres narrativer/fiktionaler Texte
- Ziel: Ermöglichung des synchronen und diachronen Vergleichs einer literarischen Epoche
- Downloadformate: XML/ TXT (vgl. Reintext-Version); das KOLIMO kann auch heruntergeladen und unabhängig von der grafischen Benutzeroberfläche (vgl. GUI) verwendet werden

2. Anwendungsbeispiel

Sie wollen untersuchen, wie sich der Stil narrativer Texte der literarischen Moderne gegenüber ihren Vorgängerepochen unterscheidet.

Für eine derartige Fragestellung bietet sich die Arbeit im KOLIMO an. Durch die Konzentration auf diese spezifische Epoche ist es dem KOLIMO möglich, relevante Texte aus diversen anderen Textsammlungen in sich zu vereinen und vergleichbare Metadaten zur Verfügung zu stellen. So ist es Ihnen möglich, vergleichende quantitative Abfragen (vgl. Query) durchzuführen, die sich auf eine repräsentative Textmenge beziehen. Zudem bietet Ihnen KOLIMO die Möglichkeit, literarische mit nicht-literarischen narrativen Texten der Epoche oder mit Texten aus der Zeit vor der literarischen Moderne zu vergleichen.

3. Diskussion

3.1 Kann ich das KOLIMO für wissenschaftliche Arbeiten nutzen?

Ja, aber mit etwas Vorsicht.

Die Texte im KOLIMO werden aus unterschiedlichen Quellen bezogen: TextGrid, Gutenberg-DE, DTA und dem Kafka-Referenzkorpus. Das Problem dabei ist, dass die importierten Ressourcen in qualitativer Hinsicht stark variieren – die Texte aus Gutenberg sind generell nicht wissenschaftlich zitierfähig. Das KOLIMO eignet sich daher besonders für quantitative Vergleichsanalysen in Form eines *Distant Reading*. Sollten Sie einen bestimmten digitalisierten Text für ein zitierfähiges *Close Reading* suchen, schauen Sie entweder genau in den Metadaten des jeweiligen Textes nach (im sog. TEI-Header (vgl. TEI)), aus welcher Quelle das Digitalisat stammt, oder suchen Sie direkt in einer der enthaltenen zitierfähigen Textsammlungen.

Das KOLIMO erhebt jedoch mit großer Mühe einheitliche und vergleichbare Metadaten für alle enthaltenen Texte und hat dafür einige verbindliche Richtlinien festgelegt: Die Metadaten eines jeden Dokumentes werden aus der ursprünglichen Textquelle übernommen und unter Einhaltung des DTA-Basisformats TEI ergänzt. Dazu werden beispielsweise fehlende Erscheinungsdaten recherchiert oder unterschiedliche Gattungsangaben vereinheitlicht. Neben der öffentlichen Zugriffsmöglichkeit durch die Speicherung auf einem eigenen Server wird zur nachhaltigen Langzeitarchivierung ein Datenbankabbild gespeichert.

Die Texte selbst sollen vor allem dem stilistischen Vergleich dienen und wurden dafür automatisierten linguistischen Annotationen unterzogen. Um eine höhere Genauigkeit der Wortart-Annotationen (vgl. Annotation) zu gewährleisten, hat das KOLIMO mehrere POS-Tagger eingesetzt und pro Register nur den Tagger mit den jeweils besten Ergebnissen verwendet. So wurde eine epochensensitive POS-Annotation erzeugt. Als Epochen wurden unter Zuhilfenahme von Literaturgeschichten Moderne, Barock, Aufklärung, Romantik und Realismus festgelegt.

3.2 Wie benutzerfreundlich ist die Arbeit mit dem KOLIMO?

Die Startseite („Dashboard“) des KOLIMO zeigt eine Übersicht über die Anzahl der Autor*innen, Titel und Wörter pro Quelle und außerdem die zeitliche Verteilung der vertretenen Primärtexte, deren Schwerpunkt auf der Zeit um 1900 liegt.

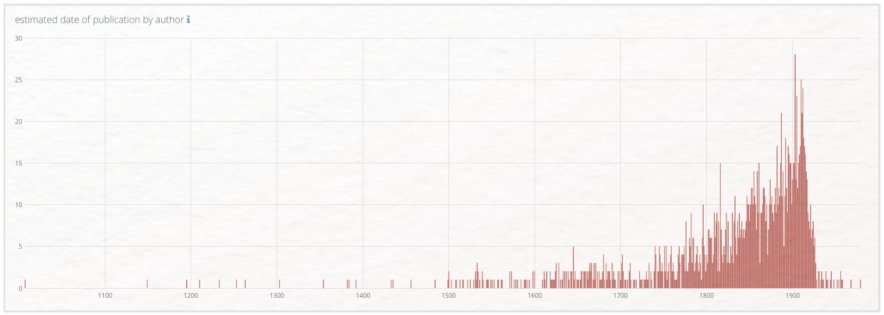

Abb. 2: KOLIMO - Chronologie

Unter dem Navigationspunkt „About“ werden das Projekt, seine Ziele und Konzeption, die Schritte der Textaufbereitung und die Lizenzbedingungen englischsprachig erklärt. Hinweise, wie die Textsammlung schrittweise verwendet werden kann, oder gar Tutorials fehlen jedoch bislang.

Die Webseite ist übersichtlich und klar strukturiert und man findet sich schnell zurecht. Es wird jedoch deutlich, dass es sich um eine Beta-Version handelt, die Textsammlung sich also noch in der Entwicklung befindet:

Die Text- und Metadaten können zwar sämtlich und in unterschiedlichen Formaten heruntergeladen werden, es besteht in den Daten aber noch viel Rauschen („noise“), von dem sie weiterhin bereinigt werden sollen. Das KOLIMO-Team setzt zudem stark auf die Mitarbeit der Nutzer*innen: Es bittet bspw. um die Zusendung von Fehlermeldungen, Anregungen und digitalisierter Volltexte aus der Zeit vor 1800 (die den beschriebenen Qualitätskriterien entsprechen) an litre@gwdg.de.

4. Wie funktioniert die Textsuche im KOLIMO?

Unter dem Menüpunkt „Collection Browser“ gelangen Sie auf die intuitive Suchmaske für die Volltexte. Hier können Sie bspw. nach Autor*innen, Titeln, Veröffentlichungsdaten, Genres (externe Kategorien aus dem DTA, TextGrid und Gutenberg), Epochen (Realismus, Moderne) und Textlängen (von KOLIMO implementiert) suchen. Einige Suchanfragen bedürfen hier momentan noch weiterer Programmierung seitens des KOLIMO-Teams. Mit der Plustaste lassen sich auch kombinierte Abfragen zu diesen Kategorien starten. Das Seitensymbol jeweils rechts neben den einzelnen Einträgen der Ergebnisliste bringt Sie dann zu den Volltexten.

Externe und weiterführende Links

- KOLIMO auf Gitlab: https://web.archive.org/save/https://gitlab.gwdg.de/kolimo (Letzter Zugriff: 04.06.2024)
- KOLIMO auf Zenodo: https://web.archive.org/save/https://zenodo.org/records/10246193 (Letzter Zugriff: 04.06.2024)

Bibliographie

Herrmann, J. Berenike und Gerhard Lauer. 2016a. Aufbau und Annotation des Kafka/Referenzkorpus. In: *DHd 2016. Modellierung – Vernetzung – Visualisierung. Die Digital Humanities als fächerübergreifendes Forschungsparadigma. Konferenzabstracts*, 158–160. Universität Leipzig. http://www.dhd2016.de/boa.pdf (zugegriffen: 20. Dezember 2018).

———. 2016b. KAREK. Building and Annotating a Kafka/Reference Corpus. In: *Digital Humanities 2016: Conference Abstracts*, 552–553. Kraków. http://dh2016.adho.org/abstracts/427 (zugegriffen: 20. Dezember 2018).

———. 2017. Das ‚Was-bisher-geschah' von KOLIMO. Ein Update zum Korpus der literarischen Moderne. In: *DHd 2017: Digitale Nachhaltigkeit. Konferenzabstracts*, 107–110. Universität Bern. http://www.dhd2017.ch/wp-content/uploads/2017/02/Abstractband_ergaenzt.pdf (zugegriffen: 20. Dezember 2018).

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.

Horstmann, Jan und Alexandra Kern. 2024. Ressourcenbeitrag: Deutsches Textarchiv (DTA). Hg. von Evelyn Gius. *forTEXT* 1, Nr. 11. Bibliografie (29. November). doi: 10.48694/fortext.3791, https://fortext.net/ressourcen/textsammlungen/deutsches-textarchiv-dta.

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.

Close Reading Close Reading bezeichnet die sorgfältige Lektüre und Interpretation eines einzelnen oder weniger Texte. Close Reading ist in der digitalen Literaturwissenschaft außerdem mit der manuellen Annotation textueller Phänomene verbunden (vgl. auch Distant Reading als Gegenbegriff).

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.

CSV CSV ist die englische Abkürzung für *Comma Separated Values*. Es handelt sich um ein Dateiformat zur einheitlichen Darstellung und Speicherung von einfach strukturierten Daten mit dem Kürzel `.csv` , sodass diese problemlos zwischen IT-Systemen ausgetauscht werden können. Dabei sind alle Daten zeilenweise angeordnet. Alle Zeilen wiederum sind in einzelne Datenfelder aufgeteilt, welche durch Trennzeichen wie Semikola oder Kommata getrennt werden können. In Programmen wie Excel können solche Textdateien als Tabelle angezeigt werden.

Distant Reading Distant Reading ist ein Ansatz aus den digitalen Literaturwissenschaften, bei dem computationelle Verfahren auf häufig große Mengen an Textdaten angewandt werden, ohne dass die Texte selber gelesen werden. Meist stehen hier quantitative Analysen im Vordergrund, es lassen sich jedoch auch qualitative Metadaten quantitativ vergleichen. Als Gegenbegriff zu *Close Reading* wurde der Begriff insbesondere von Franco Moretti (2000) geprägt.

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.

Reintext-Version Die Reintext-Version ist die Version eines digitalen Textes oder einer Tabelle, in der keinerlei Formatierungen (Kursivierung, Metadatenauszeichnung etc.) enthalten sind. Reintext-Formate sind beispielsweise TXT, RTF und CSV.

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.