Toolbeitrag: CorpusExplorer

Mareike Schumacher [1]

1. Universität Regensburg

Thema:	Korpusbildung	DOI:	10.48694/fortext.3810
Jahrgang:	1	Ausgabe:	2
Erscheinungsdatum:	12-06-2024	Erstveröffentlichung:	2020-12-14 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.

Workflow: Upload von Textdaten in reiner oder vorannotierter Form, Aufbereitung des Korpus und Durchführen von Abfragen, Download der aufbereiteten Daten.

- **Systemanforderungen:** CorpusExplorer ist eine Desktopapplikation (vgl. Webanwendung) für Windows. Technich versierte Nutzer*innen können auf Mac oder Linux eine Konsolen-Version (vgl. Commandline) verwenden.
- **Stand der Entwicklung:** Die jetzige Version des CorpusExplorer wurde 2013 herausgebracht und kontinuierlich weiter entwickelt.
- **Herausgeber:** Jan Oliver Rüdiger
- **Lizenz:** kostenfrei
- **Weblink:** www.corpusexplorer.de
- **Im- und Export:** Der CorpusExplorer unterstützt über 100 unterschiedliche Datei- und Textformate für Im- und Export, darunter gängige Formate wie CSV oder XML
- **Sprachen:** Sprachunabhängig (unterstützt UTF-8 (vgl. Unicode/UTF-8))

1. Für welche Fragestellungen kann der CorpusExplorer eingesetzt werden?

Der CorpusExplorer eignet sich vor allem für explorative Zugänge zu großen Textkorpora (vgl. Korpus). Diese können automatisch in Teilkorpora unterteilt und so immer wieder neu betrachtet werden. Verwendung bestimmter Wortarten oder häufig in ähnlichen Zusammenhängen auftretende Wörter können mit Hilfe automatischer Routinen untersucht werden. Durch die Verknüpfung mit literaturwissenschaftlich relevanten Ressourcen wie DTA (Horstmann und Kern 2024), TextGrid (Horstmann 2024b) und DraCor (Horstmann 2024a) können bereits mit literaturwissenschaftlich relevanten Annotationen (vgl. Annotation) ausgezeichnete Texte automatisch importiert und vergleichend betrachtet werden. So wäre es z.B. möglich, folgende Fragestellung zu verfolgen: Wie ist das Verhältnis von Sprecher- zu Sprecherinnen-Text in 500 deutschsprachigen Dramen des 18. - 20. Jahrhunderts?

2. Welche Funktionalitäten bietet der CorpusExplorer und wie zuverlässig ist das Tool?

Funktionen:

- Auswertung kleiner und großer Textsammlungen (vgl. Korpus)
- Automatisierte Text-/Metadatenextraktion, Bereinigung und Annotieren von Korpora
- Bereits über 50 zum Teil experimentelle Auswertungen und Visualisierungen, u. a. Frequenzanalyse, Kookkurrenzen, Heatmaps oder Geovisualisierung
- Analyse unterschiedlichster Quellen (z. B. Transkripte (vgl. Transkription), Tweets, Dramen oder Romane)
- Die Abfrageroutinen (vgl. Query) zielen auf Reproduzierbarkeit der Datenaufbereitung

- Export der Analyseergebnisse und Korpora in verschiedene offene Formate
- Einbindung in andere Programmiersprachen wie Python, R, C# oder Java durch Konsolen-Schnittstelle möglich.

Zuverlässigkeit: CorpusExplorer wird kontinuierlich weiterentwickelt, kann auf dem eigenen Rechner installiert werden und läuft zuverlässig.

3. Ist der CorpusExplorer für DH-Einsteiger*innen geeignet?

Checkliste	∨ / teilweise / -
Methodische Nähe zur traditionellen Literaturwissenschaft	-
Grafische Benutzeroberfläche	∨
Intuitive Bedienbarkeit	teilweise
Leichter Einstieg	teilweise
Handbuch vorhanden	∨
Handbuch aktuell	∨
Tutorials vorhanden	∨
Erklärung von Fachbegriffen	teilweise
Gibt es eine gute Nutzerbetreuung?	∨

Die grafische Benutzeroberfläche (vgl. GUI) ermöglicht eine weitgehend intuitive Bedienung, die Vielzahl der angebotenen Features (vgl. Feature), die nur zum Teil selbsterklärend sind, erschwert aber den Einstieg für weniger technikaffine Nutzer*innen. Die relevanten Funktionen lassen sich ohne technisches Vorwissen nicht sofort gewinnbringend ausführen. Allerdings werden ein aktualisiertes Handbuch sowie hilfreiche Tutorials bereitgestellt, um den Einstieg zu erleichtern und verschiedene Funktionen des Tools aufzuzeigen. Ein E-Mail-Support zur Unterstützung sowie Klärung von Fragen und Problemen steht zur Verfügung.

4. Wie etabliert ist der CorpusExplorer in den (Literatur-)Wissenschaften?

Der CorpusExplorer wird bereits in einigen, überwiegend korpuslinguistischen, Studien zitiert. Auch für Diskursanalysen wird das Tool verwendet. In den (digitalen) Literaturwissenschaften ist der CorpusExplorer noch wenig etabliert.

5. Unterstützt der CorpusExplorer kollaboratives Arbeiten?

Nein, der CorpusExplorer hat keine Funktionalitäten, die kollaborativ genutzt werden können.

6. Sind meine Daten beim CorpusExplorer sicher?

Ja. Für die Nutzung des CorpusExplorers ist keine Angabe persönlicher Daten notwendig. Die verarbeiteten Textdaten bleiben auf dem eigenen PC. Seit einem Update im Jahr 2019 fragt der CorpusExplorer, ob die auf einer eigenen OpenSource basierte Infrastruktur zur Telemetrieerhebung genutzt werden darf. Stimmen Nutzende dem zu, werden anonymisierte Ereignisse wie Programmfehler oder genutzte Funktionen erhoben. Daten werden nicht an Dritte übermittelt. Wird der Nutzung der Telemetrie widersprochen, werden keinerlei Daten erhoben oder übermittelt. Die Nutzung des CorpusExplorers ist also unter datenschutzrechtlichen und auch unter urheberrechtlichen Gesichtspunkten unproblematisch.

Externe und weiterführende Links

- www.corpusexplorer.de: https://web.archive.org/save/http://www.corpusexplorer.de (Letzter Zugriff: 04.06.2024)

Bibliographie

Horstmann, Jan. 2024a. Ressourcenbeitrag: DraCor - Drama Corpora Project. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 6. Netzwerkanalyse (30. August). doi: 10.48694/fortext.3785, https://fortext.net/ressourcen/textsammlungen/dracor-drama-corpora-project.

———. 2024b. 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.
Rüdiger, Jan Oliver. 2018. CorpusExplorer. Universität Kassel, Universität Siegen. http://corpusexplorer.de.

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.

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.

Feature Unter Features können Einzelfunktionen eines Tools verstanden werden, die beispielsweise komplexe Funktionen wie die Visualisierung eines Textes als Wordcloud ermöglichen, oder auch kleinere Funktionseinheiten wie den Abgleich einzelner Spracheigenschaften (Properties) mit annotierten Beispieltexten darstellen.

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.

Property Property steht für „Eigenschaft", „Komponente" oder „Attribut". In der automatischen Annotation dienen konkrete Worteigenschaften wie Groß- und Kleinschreibung zur Klassifizierung von Wörtern oder Phrasen. Durch die Berücksichtigung solcher Eigenschaften in den Features eines Tools kann maschinelles Lernen bestimmter Phänomene umgesetzt werden. In der manuellen Annotation können als Properties auch Eigenschaften von Annotationen benannt werden.

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).

Transkription Die Definition des Begriffs „Transkription" ist weit gefasst und stark abhängig vom wissenschaftlichen Bereich. Grundsätzlich bezieht sich die Transkription auf das Umschreiben, Übertragen oder Umformen einer Entität. In den Geisteswissenschaften kann sie grundsätzlich als Verschriftlichung von Medien wie Audio-, Videodateien aber auch Texten verstanden werden. In der Editionswissenschaft handelt es sich beispielsweise um die buchstabengenaue Abschrift oder Kopie eines Textes.

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.

Unicode/UTF-8 Unicode ist ein internationaler Standard, der für jedes Schriftzeichen oder Textelement einen digitalen Code festlegt. Dabei ist UTF-8 die am weitesten verbreitete Kodierung für Unicode-Zeichen. UTF-8 ist die international standardisierte Kodierungsform elektronischer Zeichen und kann von den meisten Digital-Humanities-Tools verarbeitet werden.

Webanwendung Eine webbasierte Anwendung ist ein Anwendungsprogramm, welches eine Webseite als Schnittstelle oder Front-End verwendet. Im Gegensatz zu klassischen Desktopanwendungen werden diese nicht lokal auf dem Rechner der Nutzer*innen installiert, sondern können von jedem Computer über einen Webbrowser „online" genutzt werden. Webanwendungen erfordern daher kein spezielles Betriebssystem.

Wordcloud Eine *Wordcloud* , oder auch Schlagwortwolke, ist eine Form der Informationsvisualisierung, beispielsweise von Worthäufigkeiten in einem Text oder einer Textsammlung. Dabei werden unterschiedlich gewichtete Wörter, wie die häufigsten Wörter, i.d.R. größer oder auf andere Weise hervorgehoben dargestellt. Die horizontale/vertikale Ausrichtung und die Farbe der dargestellten Wörter hat meistens allerdings keinen semantischen Mehrwert.

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.