Ressourcenbeitrag: DraCor: Drama Corpora Project

Jan Horstmann [1]

1. Universität Münster

forTEXT

Thema:	Netzwerkanalyse	DOI:	10.48694/fortext.3785
Jahrgang:	1	Ausgabe:	6
Erscheinungsdatum:	2024-08-30	Erstveröffentlichung:	2020-12-03 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

DraCor bietet für eine Vielzahl von deutschsprachigen, russischen, italienischen, schwedischen, altgriechischen, spanischen, tatarischen, elsässer, antik römischen oder auch für Shakespeare- und Caldéron-Dramen den zielgenauen Zugriff auf bestimmte Textuntermengen (wie etwa der gesprochene Text pro Figur; nur der Nebentext, nur Texte weiblicher Figuren etc.). DraCor generiert zudem automatisch Netzwerke (vgl. Netzwerkanalyse (Schumacher 2024a)), die figürliche Kopräsenzen anzeigen. Die Netzwerkdaten können Sie im CSV oder Gephi-kompatiblen GEXF-Format herunterladen und weiterverarbeiten (Mehr zu Gephi siehe Schumacher (2024b)). Eine API bietet zudem zahlreiche Möglichkeiten (vgl. Feature) und zusätzliche Formate zur Weiterverarbeitung der vereinheitlichten Metadaten.

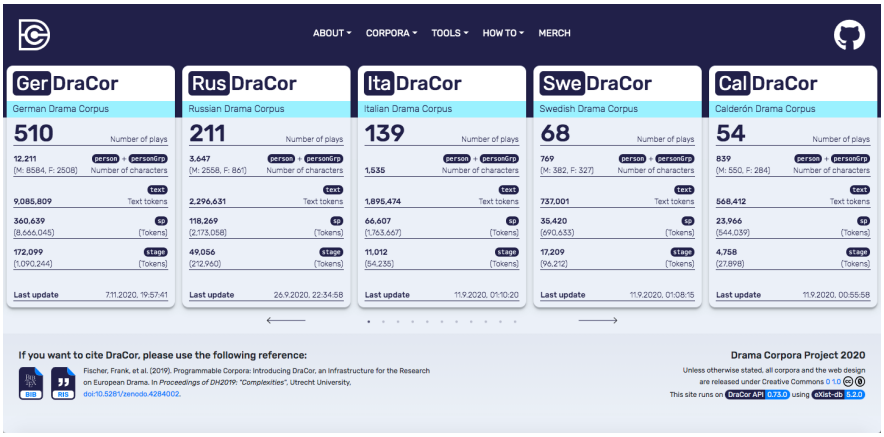

Abb. 1: Startseite von DraCor mit fünf von elf Korpora

Steckbrief

- https://dracor.org
- herausgegeben durch das Centre for Digital Humanities an der HSE Moskau und die Universität Potsdam; entstanden aus dem mittlerweile eingestellten DLINA-Projekt und Ezlinavis (Flüh 2024)
- derzeit 510 deutschsprachige, 211 russische, 139 italienische, 68 schwedische, 39 griechische, 36 antik römische , 25 spanische, 7 elsässer und 3 tartarische sowie 54 Caldéron- und 37 Shakespeare-Dramen
- strukturelle (Akte, Szenen, Auftritte, Szenenbeschreibungen etc.), semantische (z. B. Gender der jeweiligen Figuren) und weitere Metadaten: Autor, Entstehungsdatum, Veröffentlichung und Premiere des Stücks, Figurennamen, Link (vgl. URI) zur Textquelle
- zwischen 1730–1930 erschienene vollständige Dramentexte (keine Fragmente), insb. übernommen vom TextGrid Repository (Horstmann 2024)

2. Anwendungsbeispiel

Sie untersuchen in einem Forschungsprojekt Figurenkonstellationen in Dramen des Sturm und Drang und interessieren sich für das Verhältnis von weiblichen und männlichen Figuren. Mit DraCor erhalten Sie mit wenigen Klicks automatisch (vgl. Text Mining) Figurennetzwerke zu den Dramen und können diese miteinander

vergleichen. Sie sehen, welche Figuren in einzelnen Akten gemeinsam auftreten und wie groß der Anteil einer jeden Figur am gesprochenen Text ist. Die Texte können Sie per API in vielerlei Formen (bspw. Text pro Figur inkl. Genderangabe, nur Nebentexte/Regieanweisungen, nur gesprochene Texte etc.) als TXT- (vgl. Reintext-Version), RDF-, CSV-, GEXF- oder JSON-Datei herunterladen und weiterverwenden.

3. Diskussion

3.1 Kann ich DraCor für wissenschaftliche Arbeiten nutzen?

Ja. Den deutschen Texten liegt insbesondere das TextGrid Repository (Horstmann 2024) mit dem dort erarbeiteten TEI-XML (vgl. TEI) zugrunde, das korrigiert und angereichert wurde. Zudem werden Texte von Gutenberg-DE und Wikisource integriert und formal angeglichen. Außerdem werden zukünftig weitere Dramen aus dem Deutschen Textarchiv (DTA) (Horstmann und Kern 2024) integriert. Für die zusätzlichen Dramen aus Gutenberg-DE und Wikisource ist das Aufnahmekriterium, dass es eine bekannte Druckvorlage gibt, die als Vergleichsgrundlage herangezogen wird. Die russischen Dramen stammen aus der Wikisource, der Russian Virtual Library, der Online Library of Alexei Komarov und der Maxim Moshkov's Library, die Shakespeare-Dramen aus der Folger Shakespeare Library, die spanischen Dramen aus der Biblioteca Electrónica Textual del Teatro en Español (BETTE), die altgriechischen aus der Perseus Digital Library. Eigene wissenschaftliche Anwendungen von DraCor sind auf dieser Webseite dokumentiert. Es werden nur Texte in DraCor aufgenommen, die erstens keine Fragmente und zweitens in den Jahren 1730–1930 erschienen sind. Dabei möchte man eher ein repräsentatives als ein riesiges Korpus aufbauen, das Ziel sind etwa 1000 deutschsprachige und mehr als 500 russische Stücke. Zusätzlich zu Caldéron- und Shakespearekorpus sind derzeit zwei weitere autorspezifische Korpora zu Henrik Ibsen und Ludvig Holberg geplant.
Das Projekt ist nicht primär als Repositorium für Volltexte gedacht, auch wenn Sie die gesamte Textsammlung mit einer Zeile in der Commandline downloaden können; eine Anleitung dazu finden Sie über diesen Link auf Github am Ende von „Corpus Description". Ein Gesamtdownload der jeweiligen XML-Datei (vgl. XML) ist mit Rechtsklick auf „TEI version" zwar möglich, in Textsammlungen wie TextGrid Repository (Horstmann 2024) (das in DraCor zu jedem Drama verlinkt ist) oder dem Deutschen Textarchiv (DTA)(Horstmann und Kern 2024) jedoch etwas benutzerfreundlicher gestaltet. Das Ziel ist vielmehr, einen Ausgangspunkt für verschiedene digitale Projekte bereitzustellen, für die man sich die Daten in der jeweils präferierten Form über die API-Abfrage herunterladen und weiterverarbeiten kann. Gemäß dieses Grundsatzes sind alle Daten auf DraCor für jeden zugänglich. Bei Bedarf ist es zudem möglich, die gesamte Plattform lokal zu installieren (Folgen Sie dafür diesem Link). Mit dem von DraCor bereitgestelltenTool Shiny DraCor haben Sie weitere Möglichkeiten, die Korpora zu explorieren bzw. die Visualisierungen zu beeinflussen (zu finden im Menü unter „Tools").
Strukturelle Metadaten wie auch in das TEI übernommene Seitenumbrüche ermöglichen die Orientierung in den Texten. Die Metadaten sind auf vereinheitlichtem Niveau, im Shakespeare-Korpus fehlen bislang lediglich die Genderannotationen (vgl. Annotation) für Figuren. Geplant ist hierbei die Möglichkeit, dass Benutzer*innen weitere Annotationen über GitHub beisteuern können.

3.2 Wie benutzerfreundlich ist die Arbeit mit DraCor?

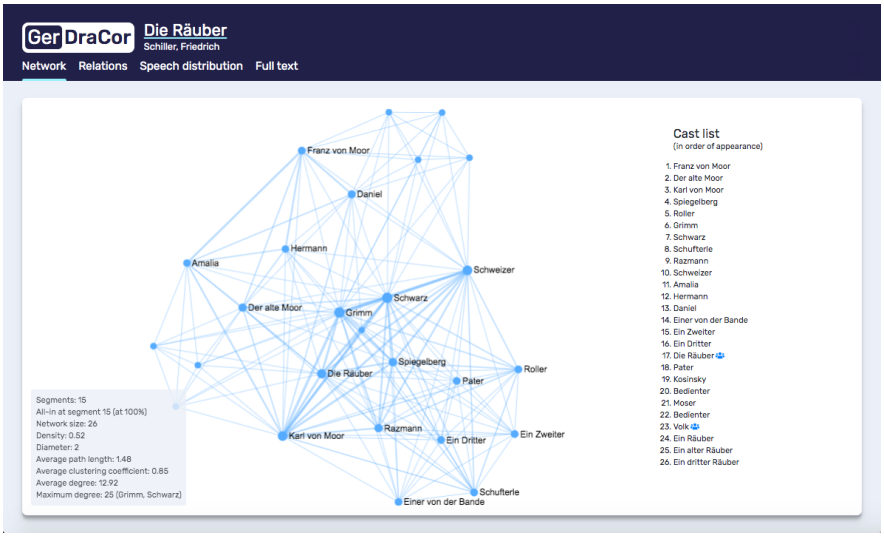

Abb. 2: Netzwerk für Schillers Drama Die Räuber *(1781) in DraCor*

Die gesuchten Texte finden Sie in DraCor sehr schnell (s. Abschnitt 4 dieses Beitrags). Ein Klick auf den Titel des jeweiligen Dramas öffnet das entsprechende Netzwerk (vgl. Abb. 2), in das Sie beliebig hineinzoomen können. Um zu verdeutlichen, dass eine Netzwerkvisualisierung immer eine Interpretation von Daten ist und z. B. die Position der Figurennamen auf der Fläche keine Rolle spielt (im Gegensatz zu ihren Verbindungen), sieht das jeweilige Netzwerk zu einem Stück bei jedem Aufruf etwas anders aus (vgl. Textvisualisierung (Horstmann und Stange 2024)).
Klicken Sie im Menü unter „Tools“ auf „API“, gelangen Sie zu den zahlreichen Möglichkeiten, mehrere oder einzelne Texte (bzw. auch Textteile) in unterschiedlichen Formen und Formaten zu exportieren (verschiedene Metadaten, TEI, Figuren, verschiedene Listen). Diese Funktionen können eine erhebliche Zeitersparnis bedeuten, wenn Sie an bestimmten Aspekten der Dramentexte interessiert sind.
In den Grundfunktionen (Textsuche, Netzwerkgenerierung, Download von Netzwerkdaten als CSV-Datei oder im Gephi (Schumacher 2024b)-kompatiblen GEXF) funktioniert DraCor sehr intuitiv. Die Arbeit mit der API-Seite (vgl. Abb. 3) kann zu Beginn eine Herausforderung darstellen, wird mit der Zeit aber handhabbar.

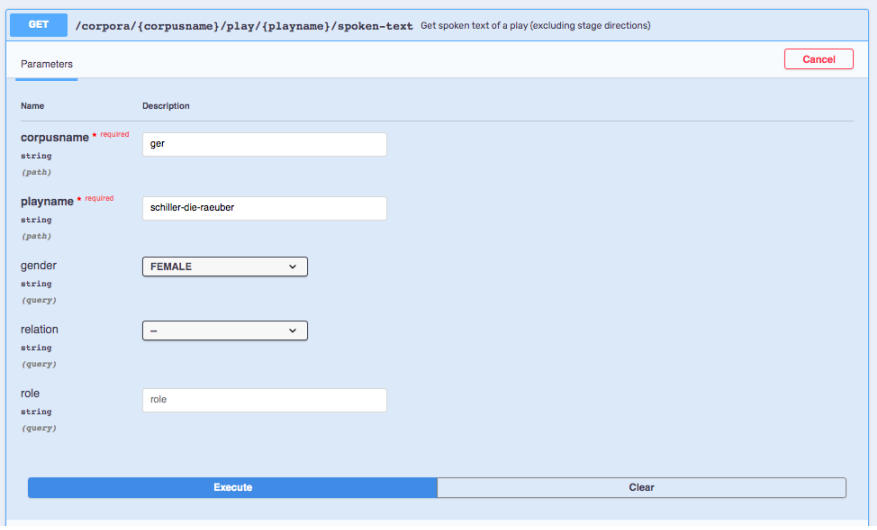

Abb. 3: Export der gesprochenen Texte weiblicher Figuren (Amalia) in Schillers Die Räuber *(1781)*

Auf der GitHub-Seite von DraCor gibt es für das deutschsprachige Dramenkorpus ein Wiki, das Notizen zu einzelnen Dramen enthält (z. B.: „Zwischen Franz und Karl wird wegen IV/2 eine Relation hergestellt. Aber eigentlich begegnen sie sich gar nicht“ für *Die Räuber*) sowie eine Dokumentation der Veränderungen im Zuge des Imports von TextGrid bereitstellt. Über GitHub bzw. die API-Seite („contact the developer“) können Sie die Herausgeber von DraCor zudem per Email kontaktieren.
Für Texte, die Sie nicht in DraCor finden, können Sie sehr leicht selbst ein entsprechendes Netzwerk aufbauen. Das Tool Ezlinavis (bzw. Easy Linavis; mehr zu Ezlinavis siehe Flüh (2024)) existierte bereits vor DraCor und ist auf der DraCor-Seite im Menü unter „Tools“ verlinkt. Es bietet die Möglichkeit, Netzwerke zügig und gänzlich ohne technische Vorkenntnisse zu erstellen.

4. Wie funktioniert die Textsuche in DraCor?

Auf der Startseite entscheiden Sie sich für eine der angebotenen Textsammlungen, die Sie mit einem Klick öffnen können. Dort haben Sie die Möglichkeit, die aufgeführten Dramen per Eingabe im Suchfeld oben zu suchen (vgl. Query) oder nach Autor*in, Titel, Netzwerkgröße, Erscheinungsjahr (generiert aus Entstehungsjahr, Premierenjahr und Druckjahr) oder Textquelle neu zu sortieren, indem Sie auf den entsprechenden Reiter klicken (siehe Abbildung 4). Die Suche einzelner Texte über die Suchleiste funktioniert sehr schnell. Ein Klick auf den Titel des jeweiligen Dramas öffnet das entsprechende Netzwerk.

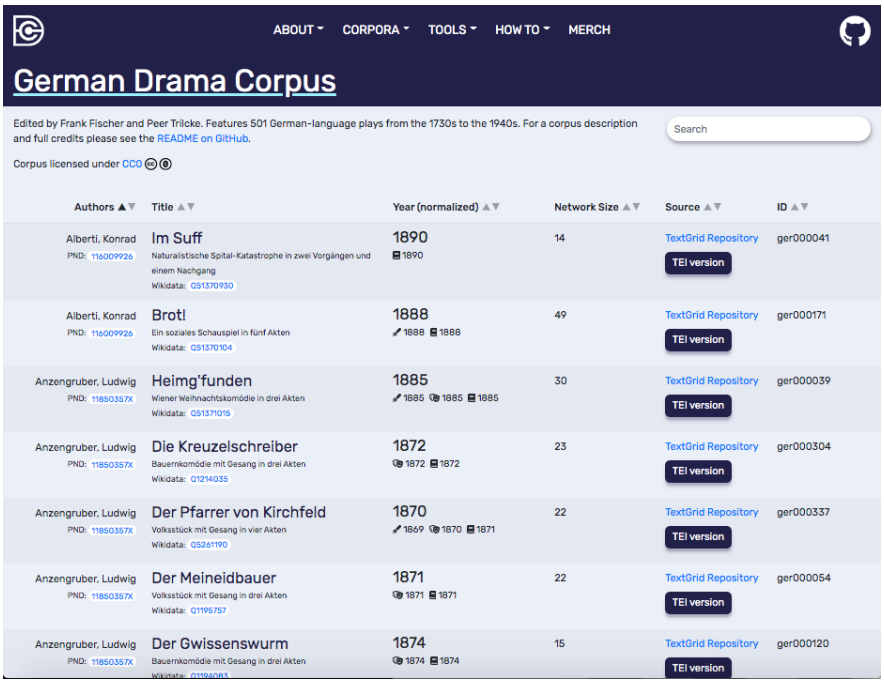

Abb. 4: Textsuche im German Drama Corpus von DraCor

Um sich die von Ihnen präferierte Form des Textes (bspw. die weibliche Figurenrede) downloaden zu können, gehen Sie auf die API-Seite, klicken (in diesem Fall bei „Get spoken text of a play (excluding stage directions)“) auf „GET“ und dann auf „Try it out“ und füllen die entsprechende Suchmaske aus. Die für eine erfolgreiche API-Abfrage benötigte Form des „corpusname“ sowie des „playname“ folgt einem Schema, das Sie in jeder URL der Dramennetzwerke wiederfinden können. Das German Drama Corpus hat bspw. das Kürzel „ger“, Schillers *Räuber* das Kürzel „schiller-die-raeuber“, woraus sich die URL dracor.org/ger/schiller-die-raeuber ergibt. Nach einem Klick auf „Execute“ erhalten Sie die angeforderten Daten zum Download.

Externe und weiterführende Links

- DLINA: https://web.archive.org/save/https://dlina.github.io (Letzter Zugriff: 01.08.2024)
- DraCor: https://web.archive.org/save/https://dracor.org/ (Letzter Zugriff: 01.08.2024)
- DraCor API: https://web.archive.org/save/https://github.com/dracor-org/dracor-api (Letzter Zugriff: 01.08.2024)
- GitHub-Seite von DraCor: https://web.archive.org/save/https://github.com/dracor-org (Letzter Zugriff: 01.08.2024)
- GerDraCor: https://web.archive.org/save/https://github.com/dracor-org/gerdracor/ (Letzter Zugriff: 01.08.2024)
- Shiny DraCor: https://web.archive.org/save/https://shiny.dracor.org/ (Letzter Zugriff: 01.08.2024)
- Wissenschaftliche Anwendungen von DraCor: https://web.archive.org/save/https://dracor.org/doc/research (Letzter Zugriff: 01.08.2024)

Bibliographie

Fischer, Frank, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christopher Kittel, Carsten Milling und Peer Trilcke. 2019. Programmable Corpora - Die digitale Literaturwissenschaft zwischen Forschung und Infrastruktur am Beispiel von DraCor. In: *DHd 2019. Digital Humanities: multimedial & multimodal.Konferenzabstracts*, 194–197. Frankfurt am Main. doi: 10.5281/zenodo.2596095,.

Fischer, Frank, Tatyana Orlova, German Palchikov, Irina Pavlova, Daniil Skorinkin und Natasha Tyshkevich. 2017. Introducing RusDraCor. A TEI-encoded Russian Drama Corpus for the Digital Literary Studies. https://dlina.github.io/presentations/2017-spb/#/ (zugegriffen: 15. Mai 2019).

Flüh, Marie. 2024. Toolbeitrag: Ezlinavis. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 6. Netzwerkanalyse (30. August). doi: 10.48694/fortext.3784, https://fortext.net/tools/tools/ezlinavis.

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.

Horstmann, Jan und Jan-Erik Stange. 2024. Methodenbeitrag: Textvisualisierung. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 5. Textvisualisierung (7. August). doi: 10.48694/fortext.3772, https://fortext.net/routinen/methoden/textvisualisierung.

Schumacher, Mareike. 2024a. Methodenbeitrag: Netzwerkanalyse. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 6. Netzwerkanalyse (30. August). doi: 10.48694/fortext.3759, https://fortext.net/routinen/methoden/netzwerkanalyse.

———. 2024b. Toolbeitrag: Gephi. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 6. Netzwerkanalyse (30. August). doi: 10.48694/fortext.3783, https://fortext.net/tools/tools/gephi.

Skorinkin, Daniil, Frank Fischer und German Palchikov. 2018. Building a Corpus for the Quantitative Research of Russian Drama: Composition, Structure, Case Studies. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018"*. Moscow. http://www.dialog-21.ru/media/4332/skorinkind.pdf (zugegriffen: 15. Mai 2019).

Trilcke, Peer, Frank Fischer, Mathias Göbel und Dario Kampkaspar. 2016. Theatre Plays as „Small Worlds"? Network Data on the History and Typology of German Drama, 1730-1930. In: *Digital Humanities 2016. Conference Abstracts*, 385–387. Kraków. https://dh2016.adho.org/abstracts/360 (zugegriffen: 15. Mai 2019).

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.

API API steht für *Application Programming Interface* und bezeichnet eine Programmierschnittstelle, die Soft- und Hardwarekomponenten wie Anwendungen, Festplatten oder Benutzeroberflächen verbindet. Sie vereinheitlicht die Datenübergabe zwischen Programmteilen, etwa Modulen, und Programmen.

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.

Data Mining Data Mining gehört zum Fachbereich Information Retrieval und bezieht sich auf die systematische Anwendung computergestützter Methoden, die darauf abzielt, in vorhandenen Datenbeständen Muster, Trends oder Zusammenhänge zu erkennen. Textbasierte Formen des Data Minings sind u. a. Text Mining, Web Mining und Opinion Mining.

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.

Information Retrieval Die Teildisziplin der Informatik, das Information Retrieval, beschäftigt sich mit der computergestützten Suche und Erschließung komplexer Informationen in meist unstrukturierten Datensammlungen.

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.

Opinion Mininig Unter Opinion Mining, oder Sentiment Analysis, versteht man die Analyse von Stimmungen oder Haltungen gegenüber einem Thema, durch die Analyse natürlicher Sprache. Das Opinion Mining gehört zu den Verfahren des Text Minings.

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.

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).

Text Mining Das Text Mining ist eine textbasierte Form des Data Minings. Prozesse & Methoden, computergestützt und automatisch Informationen bzw. Wissen aus unstrukturierten Textdaten zu extrahieren, werden als Text Mining zusammengefasst.

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.

URI *Uniform Resource Identifier* (URI) ist ein Identifikator zur eindeutigen Erkennung von Online-Ressourcen wie Webseiten. Im „Raum“ des Internets können so alle Inhalte eindeutig identifiziert werden, unabhängig davon, ob es sich dabei beispielsweise um eine Seite mit Text oder Video handelt. Die am häufigsten verwendete Form eines URI ist die Webseitenadresse, die URL.

Web Mining Unter Web Mining versteht man die Anwendung von Techniken des Data Mining zur Extraktion von Informationen aus dem World Wide Web. Das Web Mining ist ein Teilbereich des Data Minings und zählt zu einem der wichtigsten Anwendungsgebiete für das Text Mining.

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.