Lerneinheit: Korpusanalyse mit CATMA

Mareike Schumacher [1]

1. Universität Regensburg

forTEXT

Thema:	Manuelle Annotation	DOI:	10.48694/fortext.3751
Jahrgang:	1	Ausgabe:	4
Erscheinungsdatum:	2024-08-07	Erstveröffentlichung:	2020-11-16 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.

Eckdaten der Lerneinheit

- Anwendungsbezug: Genderverteilung in Romanen des 18. Jahrhunderts
- Methoden: Analyse, Visualisierung
- Angewendetes Tool: CATMA
- Lernziele: quantitative Analyse von Text- und Annotationsdaten; Erstellen von Queries und Visualisierungen
- Dauer der Lerneinheit: ca. 90 Minuten
- Schwierigkeitsgrad des Tools: leicht

Bausteine

- Anwendungsbeispiel: Welchen Text und welche Annotationen werden Sie erforschen? Hier erfahren Sie, wie Sie ein kleines Korpus aus Romanen des 18. Jahrhunderts digital erforschen können.
- Vorarbeiten: Was müssen Sie vor der Analyse erledigen? Hier bekommen Sie Informationen über notwendige Vorarbeiten.
- Funktionen: Welche Funktionen können Sie in CATMAs *Analyze*-Modul verwenden? Lernen Sie die einzelnen Komponenten des Moduls kennen und lösen Sie Beispielaufgaben.
- Lösungen zu den Beispielaufgaben: Haben Sie die Beispielaufgaben richtig gelöst? Hier finden Sie Antworten.

1. Anwendungsbeispiel

Mit dieser Lerneinheit können Sie beispielhaft ein kleines Korpus aus Romanen des 18. Jahrhunderts mit CATMA analysieren und visualisieren. CATMA (Computer Assisted Text Markup (Textauszeichnung) and Analysis) ist ein frei verfügbares, webbasiertes (vgl. Webanwendung) Tool, das Ihnen ermöglicht, digitale bzw. digitalisierte Texte manuell zu annotieren, analysieren und visualisieren (Horstmann 2024a) – alleine oder auch kollaborativ im Team. Dabei sind Sie völlig frei in der Wahl Ihrer Annotationskategorien und können darum ganz undogmatisch vorgehen. Diese Lerneinheit baut auf der Lerneinheit Manuelle Annotation mit CATMA (Horstmann 2024b) auf, d.h. Sie sollten bereits einen Account und ein Projekt in CATMA anlegen. Als Testmaterial stellen wir hier vorannotierte Texte zur Verfügung, in denen mit einem CRF-Modell und dem StanfordNER (Schumacher 2024) automatisch Genderzuschreibungen markiert wurden. Dabei greifen wir auf ein Modell zurück, das im Projekt m*w entwickelt wurde (Flüh und Schuhmacher 2020). Sie können die hier vorgestellten Funktionen entweder an Ihrem eigenen, bereits annotierten Datenmaterial erproben oder die hier bereitgestellten Testdaten verwenden.

2. Vorarbeiten

Loggen Sie sich bei CATMA ein und erstellen Sie ein neues Projekt. Wenn Sie dazu eine Schritt-für-Schritt-Anleitung erhalten möchten, schauen Sie in der Lerneinheit Manuelle Annotation (Jacke 2024) nach und folgen Sie den dortigen Erläuterungen bis einschließlich Absatz 15. Laden Sie sich außerdem die Testdaten auf Zenodo herunter (forTEXT 2020).

Mit einem Klick auf das „+“-Symbol in der Kachel „Documents and Annotations“ und einem anschließenden

Klick auf „Add Documents“ öffnen Sie einen Upload-wizard. Klicken Sie auf das Symbol mit dem kleinen Pfeil nach oben (siehe Abb. 1), so gelangen Sie in Ihre lokale Ordner-Struktur.

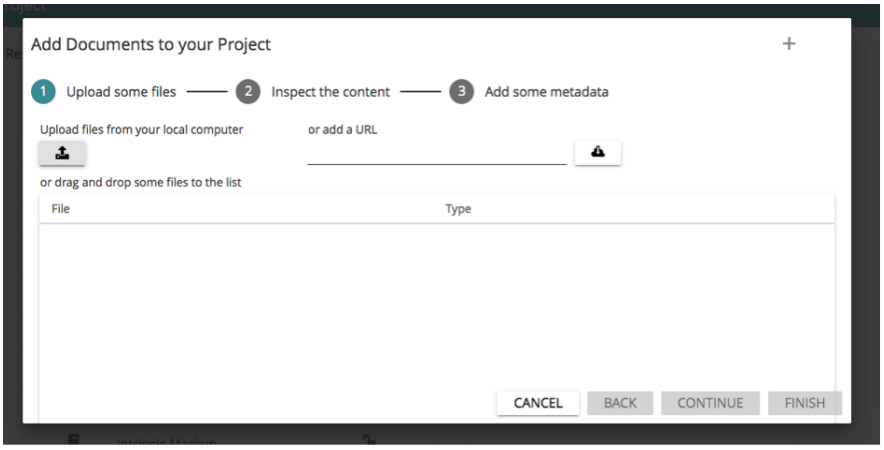

Abb. 1: Upload-Screen

Wählen Sie nun eines der vorannotierten Dokumente, die Sie heruntergeladen haben. Laden Sie die Dokumente bitte nach und nach einzeln hoch. CATMA kann zwar auch mehrere Text-Dokumente gleichzeitig hochladen, doch die vorannotierten XML-Dateien erfordern komplexere Verarbeitung, da CATMA beim Upload das interne Markup (Textauszeichnung) erkennt und es in ein Tagset umwandelt oder es mit einem bestehenden Tagset kombiniert.
Wenn Sie ein Dokument ausgewählt haben, klicken Sie auf „Continue“ und gelangen zu einer Vorschau. Kontrollieren Sie, ob der Text richtig angezeigt wird. Klicken Sie dann wieder auf „Continue“. In der nun folgenden Ansicht haben Sie die Möglichkeit, die Metadaten Ihres Textes, also Titel, Autor, Verlag und Beschreibung anzupassen. Klicken Sie dazu doppelt auf eines der Textfelder. Nun können sie alle Metadaten-Felder so bearbeiten, wie Sie es für sinnvoll halten. Klicken Sie auf „Continue“, so sehen Sie Informationen zum Tagset, das CATMA aus dem internen Markup erstellt oder mit bestehenden Tagset-Informationen zusammengeführt hat. Ein weiterer Klick auf „Continue“ führt Sie zu einer Ansicht, die Sie über die „Annotation Collection“ informiert, in der CATMA die Annotationen ablegt. Diese „Annotation Collection“ ist wie eine zweite Schicht, die über den Text gelegt wird, sodass das eigentliche Textdokument beim Annotieren unangetastet bleibt.
Laden Sie auf diese Weise alle 10 Texte des Test-Korpus in Ihr Projekt.

3. Funktionen

Wenn Sie, wie jetzt gerade, CATMA zur Korpusanalyse nutzen, so werden Sie entweder Ihre eigenen, im Close Reading-Verfahren erstellten, Annotationen auswerten oder automatisch vorannotierte Texte oder Reintextdaten mit Abfragen (vgl. Query) analysieren und visualisieren. Ist Letzteres der Fall, so nutzen Sie die Methode des Distant Reading (vgl. Distant Reading). Da wir genau das hier tun wollen, gehen Sie nun direkt ins Analyze-Modul, indem Sie links auf die Schaltfläche „Analyze“ klicken. Es öffnet sich eine Ansicht, in der an der linken Seite eine Lasche die Abfrage-Leiste verdeckt (siehe Abb. 2).

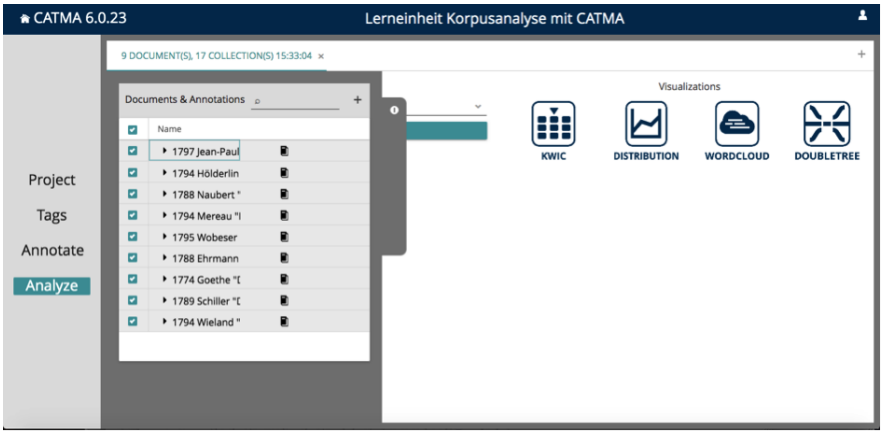

Abb. 2: Startbildschirm im Analyze-Modul

In dieser Lasche müssen Sie auswählen, welche der Texte in der Analyse berücksichtigt werden sollen. Klicken Sie auf das Kästchen ganz oben links neben „Name“. Damit wählen Sie alle Texte aus. Durch einen Klick auf die graue Schaltfläche am rechten Rand der Lasche schließen Sie diese wieder und gelangen zur Abfrage-Leiste. Tippen Sie in die Abfrage ein t ein, so wird Ihnen eine der CATMA-Standard-Abfragen vorgeschlagen, die Tagabfrage. Wählen Sie diese aus dem Drop-Down-Menü aus, so durchsucht CATMA das Korpus nach allen Annotationen. Ist die Anfrage beendet, sehen Sie eine Tabelle, in der alle Wörter aufgelistet sind, die mit einem der Tags aus Ihrem Tagset belegt sind (siehe Abb. 3). Wir wollen nun sehen, wie die Analysekategorien Ihres Tagsets über das Korpus verteilt sind.

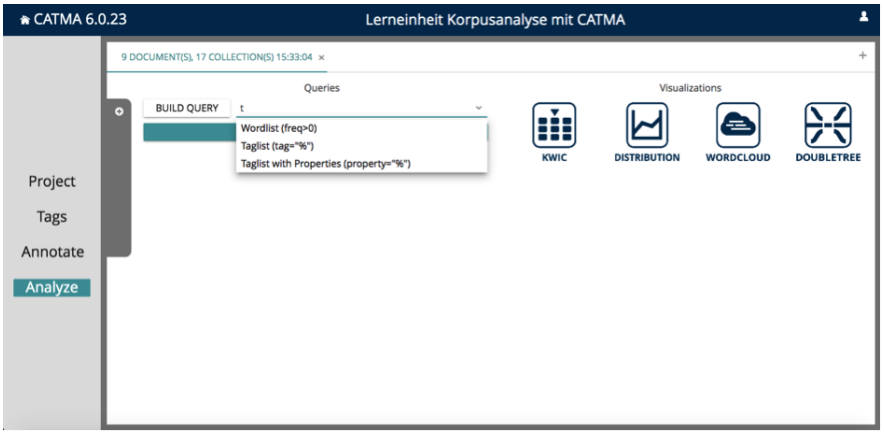

Abb. 3: Standard-Abfragen in CATMA

Klicken Sie dazu auf das Drei-Punkte-Menü und wählen Sie „group by tag path“ aus (siehe Abb. 4). Jetzt sehen Sie in der Tabelle die Anzahlen der Wörter, die mit den einzelnen Tag-Kategorien belegt sind.

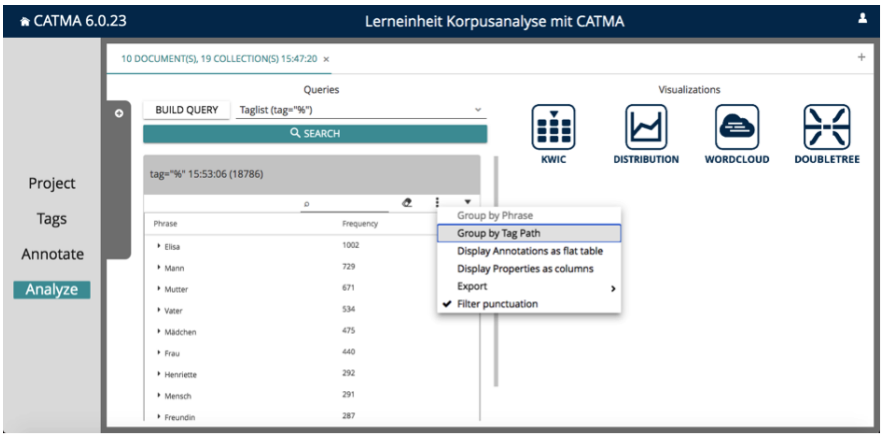

Abb. 4: Suchergebnisse nach Tags gruppieren

Aufgabe 1: Wie ist das Verhältnis der einzelnen Tag-Kategorien zueinander? Welche Kategorie verzeichnet die meisten Annotationen? Und in welchem Verhältnis stehen die Zahlenwerte der einzelnen Kategorien zu denen der anderen?

Die Tag-Abfrage zeigt in der Tabelle die Zahlen für das gesamte Korpus. Nun wollen wir uns die Verteilung der Annotationskategorien in den einzelnen Texten vergleichend anschauen. Dazu erstellen wir eine Distribution-Graph-Visualisierung. Klicken Sie auf das Icon, unter dem „Distribution“ steht.
Klicken Sie im sich neu öffnenden Fenster auf das Drei-Punkte-Menü oben mittig und wählen Sie dann „Select all“ (siehe Abb. 5). Es wird nun einen Moment dauern, bis CATMA die Visualisierungen in Form von Small Multiples (eine Visualisierung pro Text) erstellt hat.

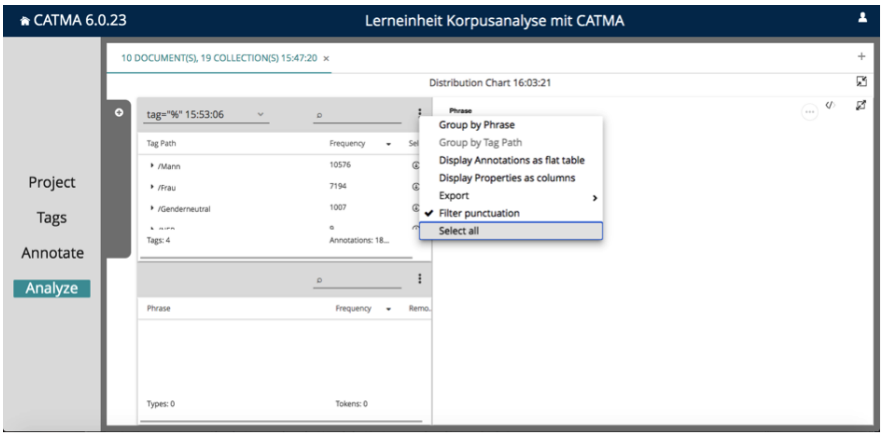

Abb. 5: Alle Abfrageergebnisse für eine Visualisierung auswählen

Aufgabe 2: In wie vielen Texten sind männliche Genderzuschreibungen dominant? In wie vielen gibt es deutlich mehr weibliche Genderzuschreibungen? Wie viele Texte zeigen ein in etwa ausgeglichenes Verhältnis zwischen weiblichen, männlichen und neutralen Genderzuschreibungen?

Im nächsten Schritt unserer Korpusanalyse schauen wir uns die einzelnen Kategorien genauer an. Dazu schließen wir zunächst die Small Multiples der Distributionsgraphen. Klicken Sie dazu auf das (obere) Icon mit den zwei kleinen Pfeilen (siehe Abb. 6).

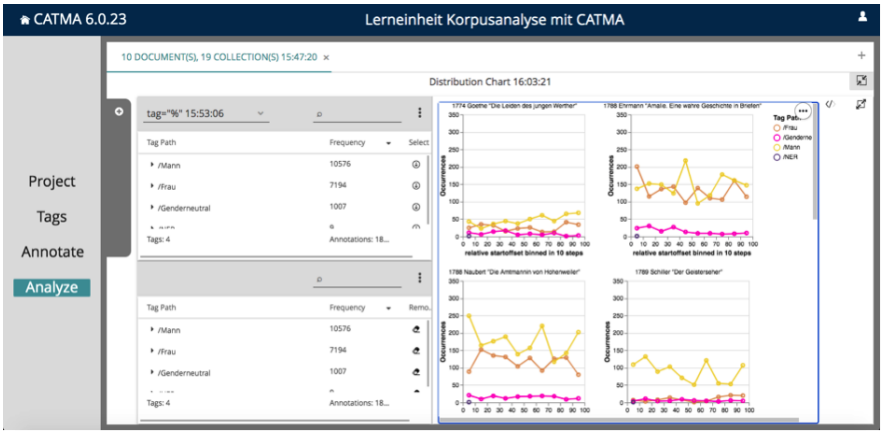

Abb. 6: Visualisierungsansicht schließen

Die Visualisierung geht dabei nicht verloren, denn Sie können Sie jederzeit über die graue Schaltfläche mit dem Titel „Distribution Chart“ und dem Zeitstempel wieder aufrufen (siehe Abb. 7).

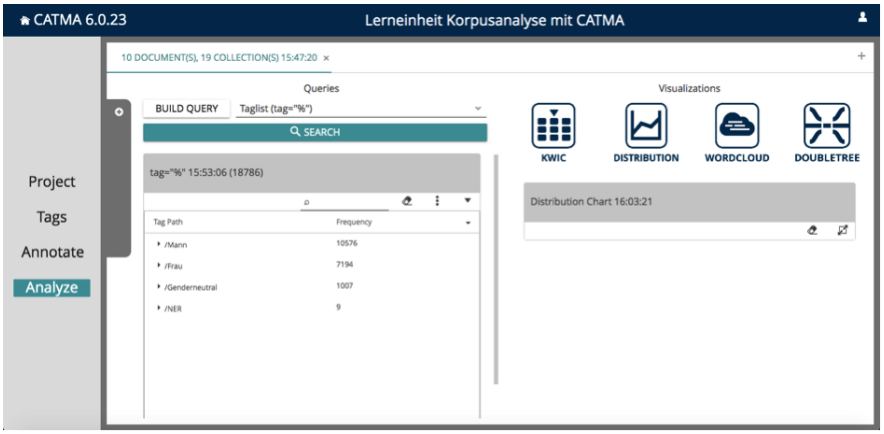

Abb. 7: Liste erstellter Visualisierungen im rechten Bildschirmbereich

Tippen Sie nun in die Abfrage-Leiste tag=„Mann“ ein und gehen Sie auf „Search“. Wenn das Abfrage-Ergebnis da ist, klicken Sie auf das Icon, unter dem „Wordcloud“ steht. Wählen Sie dann im Drei-Punkte-Menü bei der Abfrage-Tabelle „Select all“ aus. Es kann wieder einen Moment dauern, bis CATMA die Small Multiples Wordclouds erstellt hat. Scrollen Sie dann im Visualisierungsbereich ganz nach unten. Stellen Sie den Regler „No. of Types“ ganz nach rechts, bis er „500“ anzeigt. Ziehen Sie dann mit Ihrer Maus an der Mittelleiste das Feld der Visualisierungen so groß, dass Sie alle Wordclouds ganz sehen können (siehe Abb. 8).

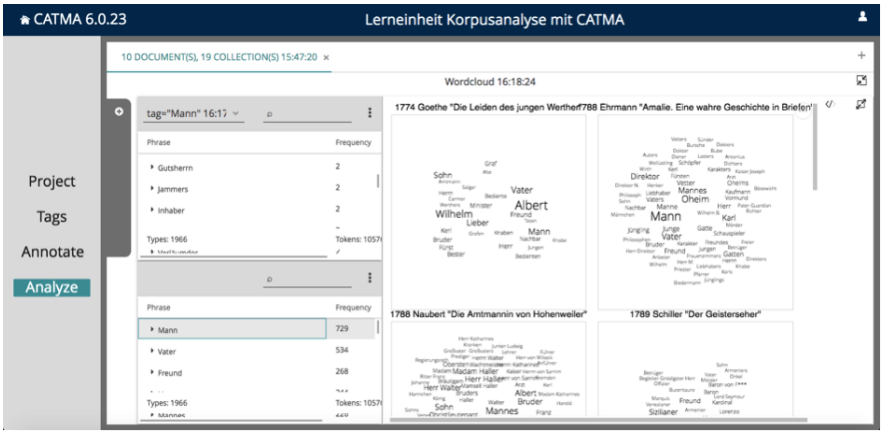

Abb. 8: Anpassbare Ansicht der Small-Multiples-Wordclouds-Visualisierung

Speichern Sie Ihre Visualisierung über das Drei-Punkte-Menü oben rechts als Bild ab (siehe Abb. 9).

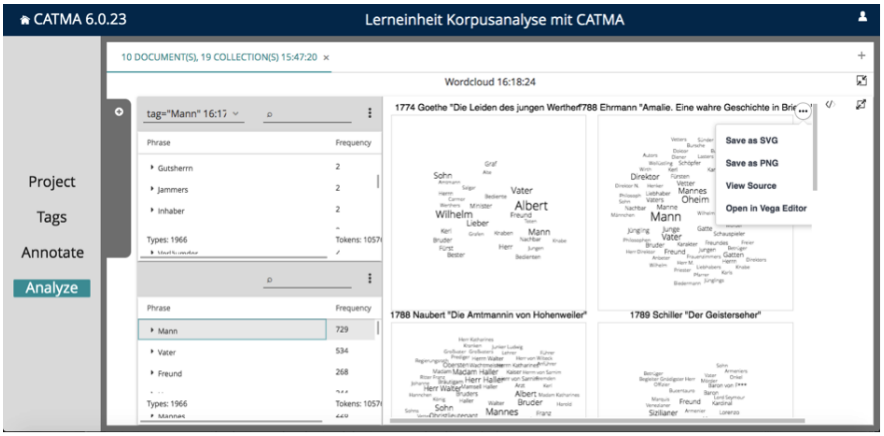

Abb. 8: Downloadansicht der Small Multiples Wordclouds Visualisierung

Erstellen Sie auf gleiche Weise Small Multiples Wordclouds des Tags „Frau“ und speichern Sie die Visualisierung ab.

Aufgabe 3: Vergleichen Sie die beiden Small-Multiples-Wordcloud-Visualisierungen miteinander! Was fällt Ihnen auf?

Verkleinern Sie in CATMA die Small Multiples Wordclouds wieder. Machen Sie eine Tag-Abfrage für die Kategorie „Genderneutral“. Gehen Sie dann rechts auf das Icon, unter dem „Double Tree“ steht, und wählen Sie das Wort „Mensch“ aus den Abfrage-Ergebnissen aus, indem Sie auf den kleinen Pfeil dahinter klicken. Es öffnet sich eine neue Form der Visualisierung. In diesem Double Tree steht in der Mitte das Wort „Mensch“ und rechts und links davon sind Wörter zu sehen, die häufig davor oder dahinter stehen (siehe Abb. 10). Je größer das Wort dargestellt ist, desto häufiger steht es vor oder nach dem Wort in der Mitte. Klicken Sie auf ein Wort vor oder hinter dem Wort in der Mitte, so öffnet sich ein Pfad, der häufige Satzstrukturen andeutet.

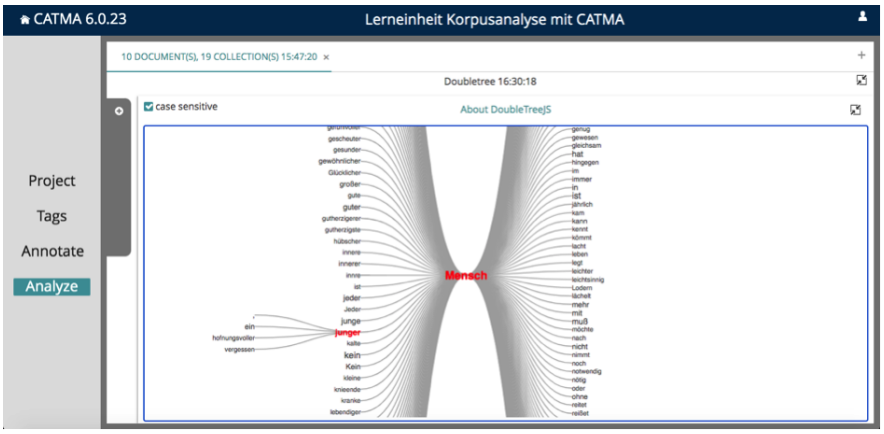

Abb. 10: Double Tree Visualisierung

Aufgabe 4: Welche Wortarten stehen am häufigsten vor oder nach dem Wort „*Mensch*“? Was schließen Sie aus den häufig nach dem Wort „*Mensch*“ stehenden Wörtern für den genderneutralen Wortgebrauch im Korpus?

Die Vorannotationen im Testkorpus wurden mit einem Modell erstellt, das Genderzuschreibungen automatisch erkennen kann. Die Kategorien „weiblich“, „männlich“ und „neutral“ wurden so trainiert, dass Substantive zwar erfasst werden, Pronomen allerdings nicht. Um die Daten der Vorannotationen mit einer wortbasierten Abfrage der Pronomen zu vergleichen, werden nun noch Distributionsgraphen erstellt, die die Pronomen „er“, „sie“ und „es“ und deren Verteilung im Korpus zeigen.
Schließen Sie den Double Tree und erstellen Sie eine Abfrage der drei oben genannten Pronomen. Dazu nutzen Sie die Wildcard-Abfrage, mit der einzelne oder mehrere Wörter im Korpus gesucht werden können. Sie können die Abfrage entweder über den Query Builder starten oder die CATMA-Abfragesprache nutzen. In dieser Lerneinheit werden wir Letzteres tun. Wenn Sie den Query-Builder ebenfalls austesten möchten, finden Sie eine Anleitung dazu in der Lerneinheit Analyse und Visualisierung (Horstmann 2024a). Eine einfache Wort-Abfrage hat in CATMA die Form wild=„Wort“. Die Abfrage aller drei Personalpronomen lautet ((wild=„er“) , (wild=„sie“)) , (wild=„es“). Erläuterungen zur Abfrage-Sprache von CATMA finden Sie [hier][30]. Erstellen Sie nun wie oben beschrieben Small Multiples Distributionsgraphen für alle drei Personalpronomen.

Aufgabe 5: Vergleichen Sie die Visualisierung mit denen der Tag-Verteilungen. Was fällt Ihnen auf? Warum muss man bei der Analyse der Personalpronomen vorsichtig sein?

Um nun etwas tiefer in die Analyse einsteigen zu können, können Sie sich die von Ihnen zuvor abgefragten Wörter in einer Keyword-in-Context (vgl. KWIC) Tabelle anschauen. Gehen Sie dazu zurück zum Abfrage-Bildschirm, indem Sie Ihre Small-Multiples-Distributiongraph-Visualisierung schließen. Klicken Sie dann auf das Icon, unter dem „KWIC“ steht (siehe Abb. 11).

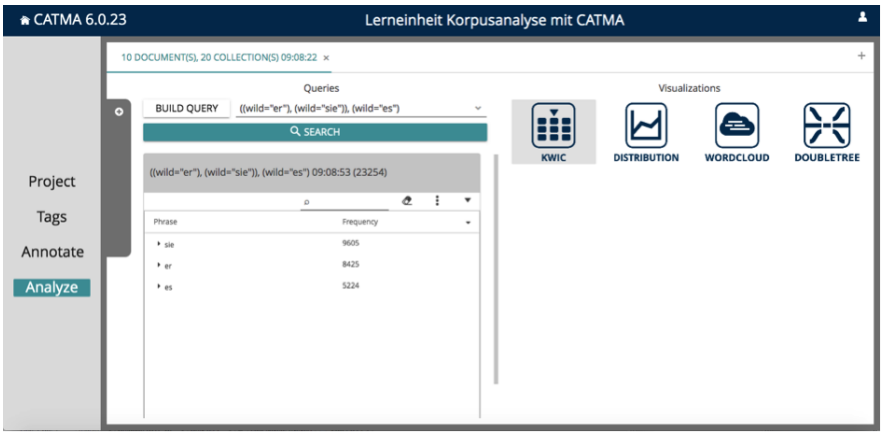

Abb. 11: Auswahl der KWIC-Visualisierung

Klicken Sie im sich dann öffnenden Fenster auf den kleinen Pfeil in der Tabelle links hinter dem Wort „sie“. Schauen Sie sich nun die einzelnen Zusammenhänge, in denen das Wort vorkommt, genauer an (Tipp: klicken Sie doppelt auf eines der Keyword-Vorkommnisse, so gelangen Sie zum ursprünglichen Text im Annotate-Modul).

Aufgabe 6: Worauf bezieht sich das Wort „*sie*“ hier meistens? Was bedeutet das für Ihre Einschätzung der Personalpronomen-Abfrage in Aufgabe 4?

Sie haben nun die Grundfunktionen zur Korpusanalyse in CATMA kennengelernt. Wenn Sie von hier aus mit dem halbautomatischen Annotieren beginnen wollen, um die Ergebnisse der Vorannotation zu verbessern, so finden Sie Hinweise dazu in der Lerneinheit „Analysieren und Visualisieren in CATMA“ (Horstmann 2024a).

4. Lösungen zu den Beispielaufgaben

***Aufgabe 1**:* Wie ist das Verhältnis der einzelnen Tag-Kategorien zueinander? Welche Kategorie verzeichnet die meisten Annotationen? Und in welchem Verhältnis stehen die Zahlenwerte der einzelnen Kategorien zu denen der anderen?

10.576 Annotationen fallen auf die Annotationskategorie „Mann“, 7194 auf die Kategorie „Frau“ und 1007 auf „Genderneutral“. Neutrale Genderzuschreibungen sind also vergleichsweise selten, kommen nur 1/10 bzw. 1/7 Mal so häufig vor wie die Zuschreibungen der anderen beiden Kategorien. Männliche Genderzuschreibungen sind in diesem Korpus eindeutig am häufigsten – und das, obwohl die Anzahl der Schriftstellerinnen im Korpus genau so hoch ist wie die der männlichen Autoren.

Aufgabe 2: In wie vielen Texten sind männliche Genderzuschreibungen dominant? In wie vielen gibt es deutlich mehr weibliche Genderzuschreibungen? Wie viele Texte zeigen ein in etwa ausgeglichenes Verhältnis zwischen weiblichen, männlichen und neutralen Genderzuschreibungen?

In der Hälfte der Texte sind männliche Genderzuschreibungen im gesamten Verlauf der Erzählung deutlich zahlreicher als weibliche. In nur einem Text sind die Nennungen weiblicher Genderzuschreibungen im gesamten Textverlauf deutlich häufiger als männliche. In vier Texten ist das quantitative Verhältnis der Kategorien „Frau“ und „Mann“ relativ ausgeglichen. In 80% der Texte ist die Anzahl genderneutraler Figurenbezeichnungen geradezu verschwindend gering gegenüber den anderen beiden Kategorien.

***Aufgabe 3**:* Vergleichen Sie die beiden Small-Multiples-Wordcloud-Visualisierungen miteinander! Was fällt Ihnen auf?

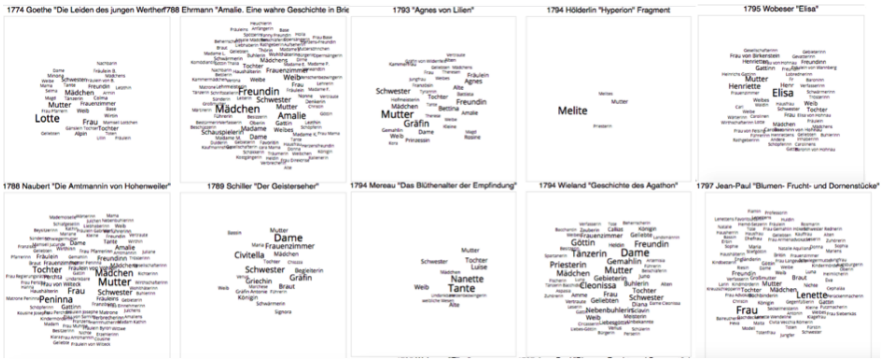

Abb. 12: Small Multiples Wordclouds der mit dem Tag „Mann“ belegten Wörter

In beiden Visualisierungen sind Rollenbenennungen besonders zahlreich. Dabei stechen klassische Familienrollen wie Mutter, Tante, Schwester, Tochter und Vater, Sohn besonders hervor. Außerdem fällt in fast allen Texten jeweils ein weiblicher und ein männlicher Vorname besonders ins Auge. Dies ist sogar der Fall, wenn der Protagonist ein Ich-Erzähler ist wie bei Goethes „Die Leiden des jungen Werther“. Die Anzahl der unterschiedlichen Wörter, die mit einer Kategorie belegt sind, variiert bei beiden Tagkategorien stark. Die Texte, in denen besonders viele unterschiedliche Wörter mit Tags belegt sind, sind für beide Kategorien meist die gleichen, und zwar „Blumen-, Furcht- und Dornenstücke“, „Geschichte des Agathon“, „Die Amtmannin von Hohenweiler“ und „Amalie“. Es fällt auf, dass in „Hyperion“ insgesamt besonders wenig weibliche und männliche Genderzuschreibungen vorkommen. „Das Blüthenalter der Empfindung“ ist ebenfalls ein Text, in dem relativ wenig Genderzuschreibungen gefunden wurden. Allerdings sind die Nennungen weiblicher Genderzuschreibungen in diesem Text im Vergleich sehr viel zahlreicher als die männlichen, die nur fünf unterschiedliche Wörter umfassen.

***Aufgabe 4**:* Welche Wortarten stehen am häufigsten vor oder nach dem Wort „Mensch“? Was schließen Sie aus den häufig nach dem Wort „Mensch“ stehenden Wörtern für den genderneutralen Wortgebrauch im Korpus?

Vor dem Wort „Mensch“ stehen besonders häufig Artikel („Der“, „der“ oder „Ein“, „ein“). Danach stehen besonders häufig Verben, am häufigsten „ist“ und „muss“. Auffallend ist auch das häufig danach stehende Wort

„von“, das, nachdem man darauf geklickt hat, Pfade zu Wörtern wie „Ansehn“, „Verstand“, „Kopf“, aber auch „Sinnen“ eröffnet. Dies lässt vermuten, dass die genderneutrale Beschreibung häufig besonders angesehene oder besonders derangierte Zustände indiziert („von Ansehn sein“ bzw. „von Sinnen sein“).

***Aufgabe 5**:* Vergleichen Sie die Small-Multiples-Distrubutiongraph-Visualisierung der Personalpronomen mit denen der Tag-Verteilungen. Was fällt Ihnen auf? Warum muss man bei der Analyse der Personalpronomen vorsichtig sein?

Tendenziell sind sich die Verlaufskurven nicht unähnlich. Allerdings zeigt die Personalpronomen-Abfrage sehr viel häufiger die genderneutrale potentielle Figuren-Referenz „es“. Einerseits kann dies darauf hindeuten, dass tatsächlich häufiger genderneutrale Figuren erwähnt werden, als die Referenzierung mit genderneutralen Substantiven wie „Mensch“, „Person“ oder „Kind“ nahe legt. Auf der anderen Seite kann das Wort „es“ auch in ganz anderen Zusammenhängen vorkommen, die nichts mit der Genderthematik zu tun haben, wie z.B. in dem Satz „es war ein schöner Tag“. Auch die Verlaufskurven, die für die Verteilung des Personalpronomens „sie“ in den Texten stehen, sind zumeist höher als die Verlaufskurven für weibliche Genderzuschreibungen. Besonders eklatant ist dies in dem Roman „Die Amtmannin von Hohenweiler“. Es liegt darum nahe, dass speziell in diesem Roman die Hauptfigur häufiger mit dem Personalpronomen „sie“ referenziert wird als mit genderstereotypen Rollenbeschreibungen. Eine solche These kann im reinen Distant-Reading-Verfahren aber lediglich aufgestellt und nicht verifiziert werden. Dafür müsste nun der Schritt zurück in den Text erfolgen, um die Kontexte besser einordnen zu können. Die generell höhere Verwendung des Personalpronomens „sie“ im gesamten Korpus legt zwar nahe, dass es insgesamt mehr Referenzen auf weibliche Figuren gibt, als die automatische Vorannotation mit einem Modell zeigt, das am besten typisierte Genderzuschreibungen erkennt. Aber auch das Wort „sie“ kann mehrere Bedeutungen annehmen. So kann damit z.B. auch eine Gruppe von Figuren bezeichnet werden. Generell sind die Analysen einzelner Wortvorkommnisse auf Korpusebene in Distant-Reading-Verfahren also eher dazu geeignet, erste Mutmaßungen aufzustellen, die dann in tiefergehenden Analysen, die mit automatischer und / oder manueller Annotation einher gehen, genauer betrachtet werden müssen. Da Mehrdeutigkeiten ausgesprochen viele Wörter betreffen, besteht hier beim reinen Distant Reading die Gefahr, fehlgeleitet zu werden.

Aufgabe 6: Worauf bezieht sich das Wort „sie“ hier meistens? Was bedeutet das für Ihre Einschätzung der Personalpronomen-Abfrage in Aufgabe 5?

Das Wort „sie“ bezieht sich hier in der Tat häufig auf weibliche Figuren. Figurengruppen oder andere Referenzen sind deutlich seltener. Die in Aufgabe 5 analysierten Verteilungsgraphen zeigen also tatsächlich ein leicht verzerrtes Bild. Dennoch kann es als Hinweis auf eine Tendenz gedeutet werden, dass Referenzen auf weibliche Figuren sehr häufig über Personalpronomen umgesetzt sind. Es wäre für eine Anschlussuntersuchung nun möglich, die tatsächlichen Referenzen auf weibliche Figuren über das Personalpronomen „sie“ manuell zu den automatisch vorannotierten Referenzen über substantivische Genderzuschreibungen hinzuzufügen. Der Aufwand einer manuelle Annotation für 10 Romane ist allerdings relativ hoch.

Externe und weiterführende Links

- CATMA: https://web.archive.org/save/https://catma.de/ (Letzter Zugriff: 03.07.2024)
- CATMA Query Language: https://web.archive.org/save/https://catma.de/how-to/query-language/ (Letzter Zugriff: 03.07.2024)
- Testdaten: https://web.archive.org/save/https://doi.org/10.5281/zenodo.10592623 (Letzter Zugriff: 03.07.2024)

Bibliographie

Flüh, Marie und Mareike Schuhmacher. 2020. m*w Figurengender zwischen Stereotypisierung und literarischen und theoretischen Spielräumen Genderstereotype und -bewertungen in der Literatur des 19. Jahrhundert. In: *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts*, hg. von Christof Schöch, 162–166. Paderborn. doi: 10.5281/zenodo.3666690,.

forTEXT. 2020. Korpusanalyse mit CATMA. Zenodo, 16. November. doi: 10.5281/zenodo.10592622, https://doi.org/10.5281/zenodo.10592623.

Horstmann, Jan. 2024a. Lerneiheit: Analyse und Visualisierung mit CATMA. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 4. Manuelle Annotation (7. August). doi: 10.48694/fortext.3752, https://fortext.net/routinen/lerneinheiten/analyse-und-visualisierung-mit-catma.

———. 2024b. Lerneinheit: Manuelle Annotation mit CATMA. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 4. Manuelle Annotation (7. August). doi: 10.48694/fortext.3750, https://fortext.net/routinen/lerneinheiten/manuelle-annotation-mit-catma.

Jacke, Janina. 2024. Methodenbeitrag: Manuelle Annotation. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 4. Manuelle

Annotation (7. August). doi: 10.48694/fortext.3748, https://fortext.net/routinen/methoden/manuelle-annotation.

Schumacher, Mareike. 2024. Toolbeitrag: Stanford Named Entity Recognizer. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 9. Named Entity Recognition (30. Oktober). doi: 10.48694/fortext.3767, https://fortext.net/tools/tools/stanford-named-entity-recognizer.

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).

CRF-Modell CRFs (*Conditional Random Fields*) sind eine Klasse statistischer Modellierungsmethoden, die häufig in der Mustererkennung und im maschinellen Lernen eingesetzt werden. CRF-Algorithmen sind der Kern kontextsensitiver Programme. Ein CRF-Modell ist daher das Ergebnis eines Trainingsprozesses, bei dem ein Modell auf Grundlage manuell anmontierter Beispiele trainiert wird, welches dabei lernt bestimmte Muster zu erkennen, um diese dann auf neue, unbekannte Texte anzuwenden. In diesen unbekannten Texten werden die erlernten Phänomene dann automatisch erkannt.

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.

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.

KWIC KWIC steht für *Keyword in Context* . Dabei handelt es sich um eine Darstellungsform, bei welcher die Treffer eines bestimmten Suchbegriffs in ihrem Kontext zeilenweise aufgelistet werden. Die Größe der Kontexte, also die Anzahl der angezeigten Umgebungswörter, kann meist individuell festgelegt werden.

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.

Tagset Ein Tagset definiert die Taxonomie, anhand derer Annotationen in einem Projekt erstellt werden. Ein Tagset beinhaltet immer mehrere Tags und ggf. auch Subtags. Ähnlich der Type/Token -Differenz in der Linguistik sind Tags deskriptive Kategorien, wohingegen Annotationen die einzelnen Vorkommnisse dieser Kategorien im Text sind.

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.

Upload-wizard Der Upload-Wizard ist ein Assistent zum Hochladen von Dateien in eine Webanwendung, der Nutzer*innen Schritt für Schritt durch den Prozess begleitet.

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.

Wildcard Als Wildcard bezeichnet man in der Informatik Platzhalter für beliebige Zeichen oder Zeichenketten.

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.