Lerneinheit: Textvisualisierung mit Voyant

Marie Flüh [1]

1. Universität Hamburg

Thema:	Textvisualisierung	DOI:	10.48694/fortext.3773
Jahrgang:	1	Ausgabe:	5
Erscheinungsdatum:		Erstveröffentlichung:	2019-06-17 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: Gotthold Ephraim Lessings bürgerliches Trauerspiel *Emilia Galotti*
- Methode: Distant Reading (vgl. Distant Reading) und Textvisualisierung
- Angewendetes Tool: Voyant
- Lernziele: Textauswahl und Nutzung elementarer Voyant-Funktionalitäten: Erstellen einer Stoppwortliste, Arbeit mit dem Voyant-Toolkit, Export der erstellten Visualisierungen und deren Interpretation
- Dauer der Lerneinheit: ca. 60 Minuten
- Schwierigkeitsgrad des Tools: einfach

Bausteine

- Anwendungsbeispiel: Welche Texte werden analysiert? Untersuchen Sie Lessings *Emilia Galotti* mittels Distant-Reading-Verfahren und interpretieren Sie die unterschiedlichen Visualisierungen der Textanalyse.
- Vorarbeiten: Welche Arbeitsschritte müssen vor der Textanalyse erledigt werden? Laden Sie sich den Primärtext herunter und speisen diesen in die Voyant-Tools ein.
- Funktionen: Welche Funktionen bietet Voyant für die digitale Textanalyse? Lernen Sie ausgewählte Tools, deren Analysefunktionen und Visualisierungsformen kennen.
- Lösungen zu den Beispielaufgaben: Haben Sie die Beispielaufgaben richtig gelöst? Hier finden Sie die Antworten.

1. Anwendungsbeispiel

In dieser Lerneinheit werden Sie Lessings bürgerliches Trauerspiel *Emilia Galotti* mit dem Textanalysetool Voyant (Flüh 2024) untersuchen. Voyant vereint Distant-Reading-Verfahren und Formen der Textvisualisierung (Horstmann und Stange 2024). Im Hintergrund der Voyant-Tools steht die Methode des Distant Readings. Im Rahmen einer quantitativen Analyse werden Texte statistisch ausgewertet. Die einzelnen und individuell auswählbaren Voyant-Tools (vgl. Feature) visualisieren die Ergebnisse dieser quantitativen Textanalyse auf ganz unterschiedliche Art und Weise. Distant-Reading-Verfahren eignen sich sowohl für die Exploration großer Textmengen (z. B. das Œuvre von Autor*innen) als auch für die Analyse vergleichsweise kleiner Textmengen (wie einzelner Werke). Für die erste Auseinandersetzung mit digitalen Methoden der Textanalyse und -visualisierung werden wir einen Einzeltext mithilfe von Voyant untersuchen.

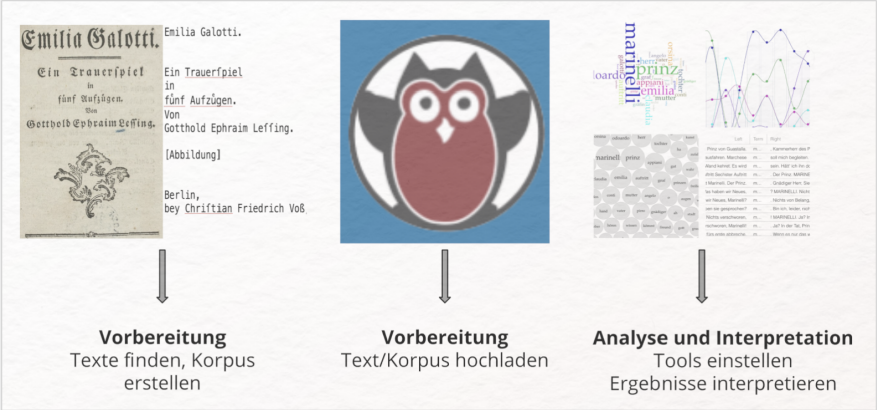

Abb. 1: Die drei Hauptbestandteile dieser Lerneinheit nehmen circa 60 Min. in Anspruch

2. Vorarbeiten

Zu Beginn beinahe jeder Form der digitalen Textanalyse steht die Frage nach der **Textauswahl**: Woher bekomme ich genau die Texte, die ich untersuchen möchte, oder sogar ein bereits zusammengestelltes thematisch passendes Textkorpus (vgl. Korpus), das in brauchbarer Form digital vorliegt? Das Deutsche Textarchiv (DTA) (Horstmann und Kern 2024) stellt eine etablierte Anlaufstelle für die Beantwortung dieser Frage dar. Hier werden Sie nun Ihren Untersuchungsgegenstand *Emilia Galotti* für die Analyse mit Voyant herunterladen. Suchen Sie die Homepage des DTA auf, indem Sie diesem Link folgen. Nun befinden Sie sich auf der Startseite des DTAs und geben den passenden Suchbegriff in das Suchfeld ein (siehe Abb. 2).

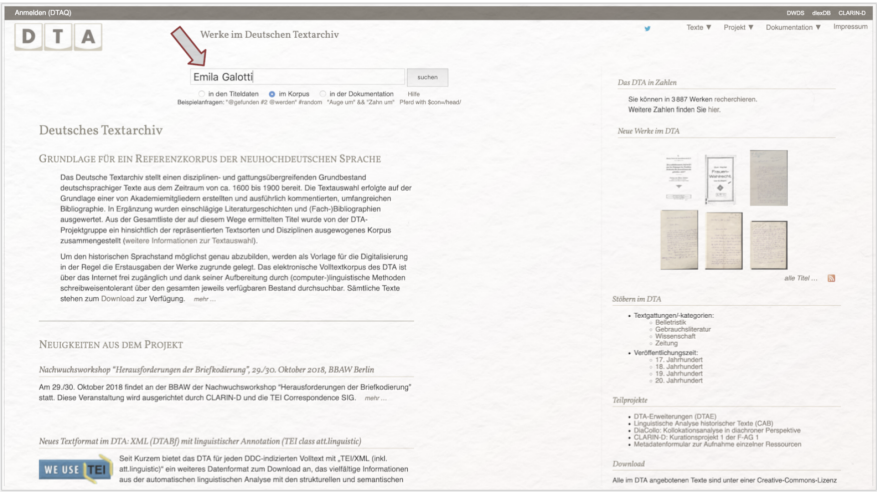

Abb. 2: Auswahl eines digitalen Textes auf der Seite des DTA durch eine Suchanfrage nach „Emilia Galotti“

Unter den Suchergebnissen befindet sich neben unterschiedlichen Werken, in denen der Suchbegriff vorkommt, auch unsere Primärquelle *Emilia Galotti* von Lessing. Per Mausklick auf den mit „#5“ gekennzeichneten Primärtext gelangen Sie in einem nächsten Schritt zu einer detaillierten dreiteiligen Werkansicht.

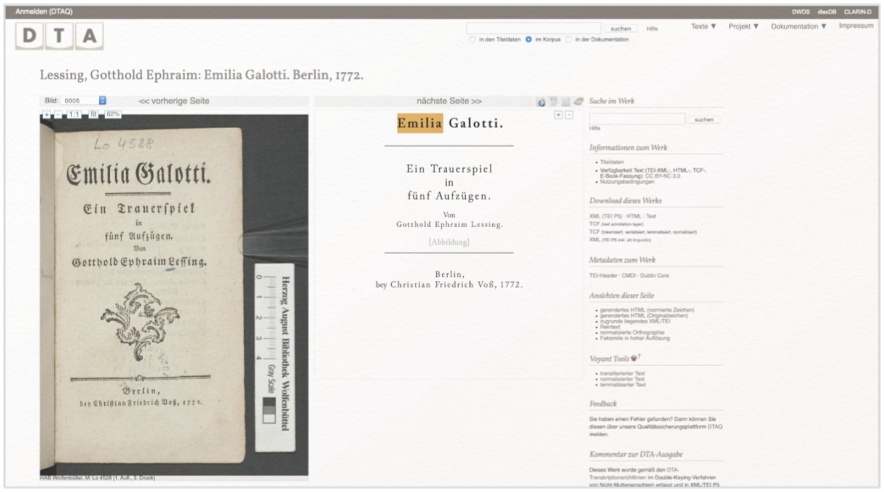

Abb. 3: Emila Galotti im Deutschen Textarchiv

Hinweis: Durch eine Kooperation zwischen dem DTA und Voyant-Tools besteht bei einigen der im Repositorium enthaltenen Texte die Möglichkeit, die gewünschte Quelle direkt über die Seite des DTAs in Voyant zu importieren. Da Download und Import von Datensätzen als Teil der Korpuserstellung einen essentiellen Bestandteil der digitalen Textanalyse darstellen, werden wir in dieser Lerneinheit den „Umweg“ über das eigenhändige Herunterladen des Textes einschlagen. Voyant kann mit Texten in unterschiedlichen Formaten wie XML, HTML, RDF, RTF, MS-Word-Dateien oder Reintexten (vgl. Reintext-Version) (TXT) arbeiten. Für die Arbeit mit Voyant bietet sich dieses Textformat an, da wir ausschließlich den Reintext von *Emila Galotti* untersuchen möchten. Wählen Sie per Mausklick das Format „Text“ aus und beginnen den Downloadprozess (siehe Abb. 4).

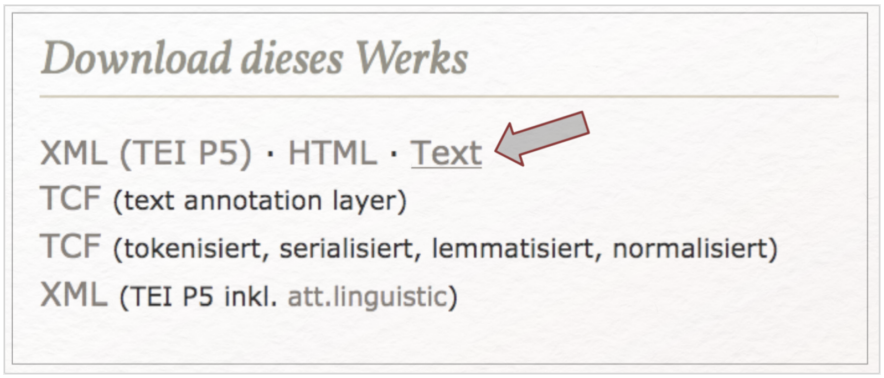

Abb. 4: Download von Emilia Galotti im Text-Format

Klicken Sie auf den „Text“-Button und speichern Sie die Datei „lessing_emilia_1772.txt“ z. B. auf Ihrem Desktop ab. Der Downloadprozess des Textes kann je nach Betriebssystem variieren. I. d. R. werden Sie dazu aufgefordert, den Text direkt zu öffnen und an einem beliebigen Ort auf Ihrem Rechner zu speichern (siehe Abb. 5).

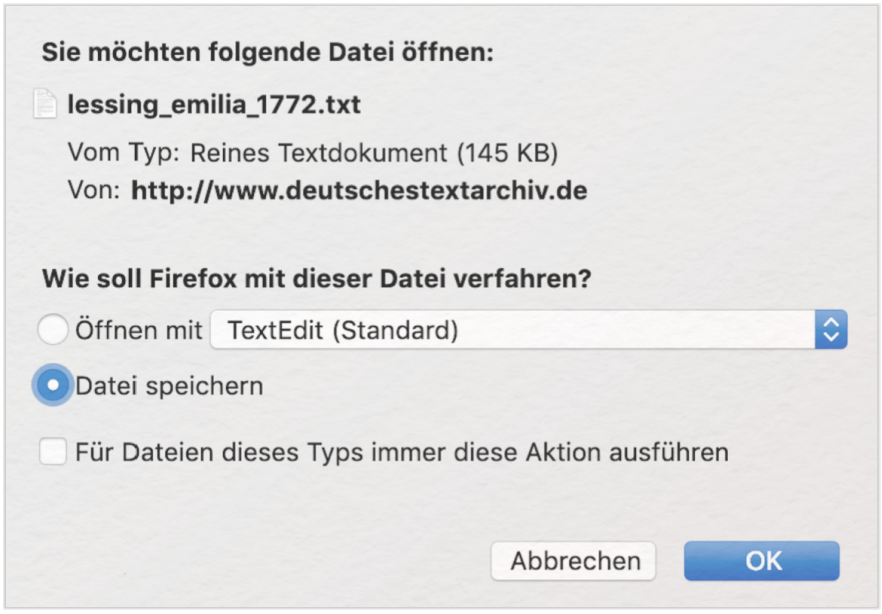

Abb. 5: Der Downloadprozess des Textes hier in der Apple-Variante

Nun haben Sie die ersten Schritte auf dem Weg zur quantitativen digitalen Textanalyse von *Emilia Galotti* erfolgreich absolviert: Ihr Untersuchungsgegenstand liegt in einem mit Voyant kompatiblen Format auf Ihrem Desktop vor und ist bereit für die Implementierung in Voyant. Das Tool selbst verwenden wir in diesem Fall als webbasierte Variante (vgl. Browser). Um *Emilia Galotti* bei Voyant **hochzuladen**, suchen Sie bitte die Startseite von Voyant auf, indem Sie diesem Link folgen. Die sich nun öffnende Seite beinhaltet ein Eingabefeld. Um Texte hochzuladen, stehen Ihnen drei unterschiedliche Möglichkeiten zur Verfügung: Die direkte Eingabe in das zentrale Eingabefeld, das Öffnen bereits erstellter Textkorpora oder das Hochladen eigener Dokumente. Im Eingabefeld können Sie URLs (vgl. URI) eingeben (mehrere URLs werden pro Zeile eingegeben, durch ein „Enter“ voneinander getrennt und per Klick auf den blauen „Reveal“-Button hochgeladen) oder Text per Copy & Paste (ebenfalls durch „Reveal“ bestätigen) einfügen. Über den Open-Button lassen sich übrigens vorbereitete Textkorpora – bestehend aus acht Werken Jane Austens oder aus 27 Werken Shakespeares – hochladen und mit Voyant explorieren. Da Sie aber einen eigenen Text untersuchen möchten, muss der Text über den „Upload“-Button hochgeladen werden (siehe Abb. 6).

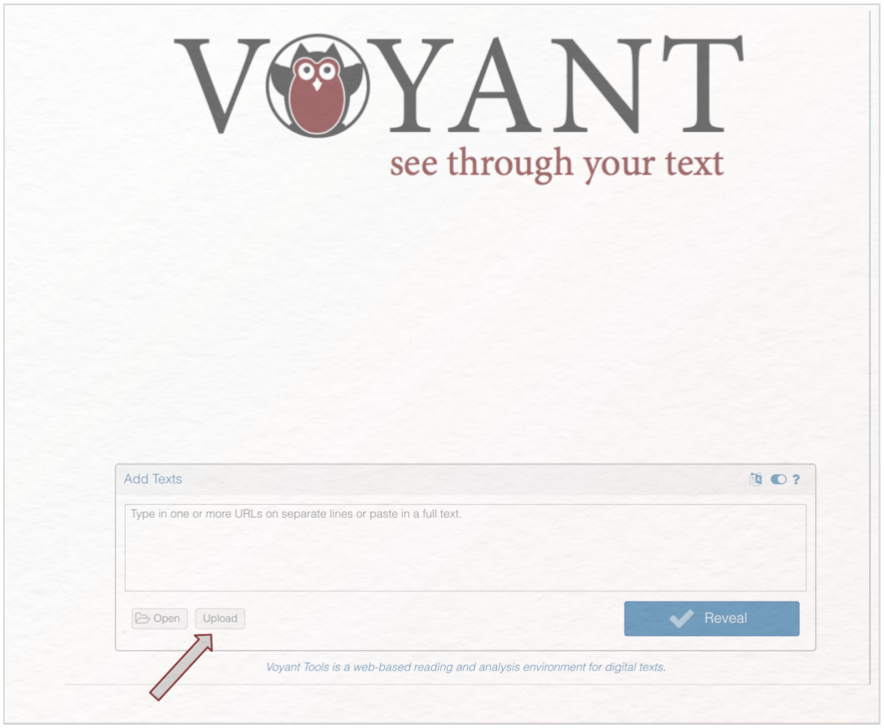

Abb. 6: Die Startseite von Voyant – Hochladen eigener Texte

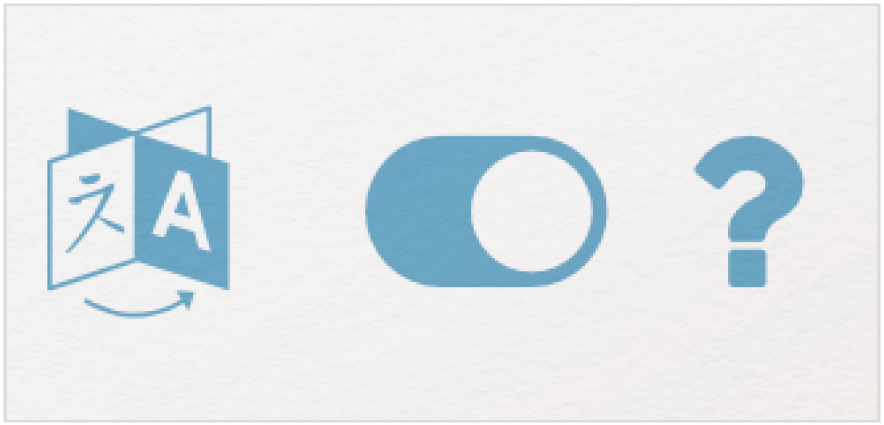

Abb. 7: Die Menüleiste des Eingabefeldes: Links: Sprachauswahl; Mitte: Festlegen von Voreinstellungen; Rechts: Erklärung der unterschiedlichen Optionen, Texte oder Textsammlungen hochzuladen

Hinweis: Bereits vor dem Hochladen Ihres Textes können Sie Voreinstellungen festlegen. Hierfür nutzen Sie die Menüleiste in der oberen rechten Ecke des Eingabefeldes (siehe Abb. 7).
Das Interface von Voyant ist auf Englisch voreingestellt, darüber hinaus aber über „Language Interface Options“ in zehn weiteren Sprachen abrufbar. Sofern Sie ein größeres Textkorpus hochladen möchten, können Sie via „Options“ bereits an dieser Stelle bestimmte Voreinstellungen vornehmen, die während des Hochladens auf das Textkorpus angewendet werden. Hier können Sie bspw. Titel für mehrere Dokumente in einem Korpus festlegen, ein Zugangspasswort zu einer Voyant-Session vergeben, festlegen, dass nur Teile einer HTML-Datei in die Voyant-Analyse einbezogen werden oder nur JSON-Dateien bearbeitet werden dürfen. Die meisten dieser Einstellungen richten sich an erfahrene Nutzer*innen und spielen bei dieser Lerneinheit keine Rolle. Tipp: Eine Beschreibung sämtlicher Tools finden Sie im Voyant-Guide. Hier werden alle Visualisierungsmöglichkeiten detailliert vorgestellt, häufig gestellte Fragen beantwortet und weitere hilfreiche Hinweise für die Arbeit mit Voyant bereitgestellt. Das Hochladen eigener Texte erfolgt in drei Schritten. Klicken Sie im Eingabefeld zunächst den Upload-Button und navigieren Sie zu der auf Ihrem Desktop abgespeicherten Datei „lessing_emilia_1772.txt“. Diese laden Sie hoch, indem sie die Datei markieren und die Auswahl bestätigen, indem Sie den „Öffnen“-Button betätigen.

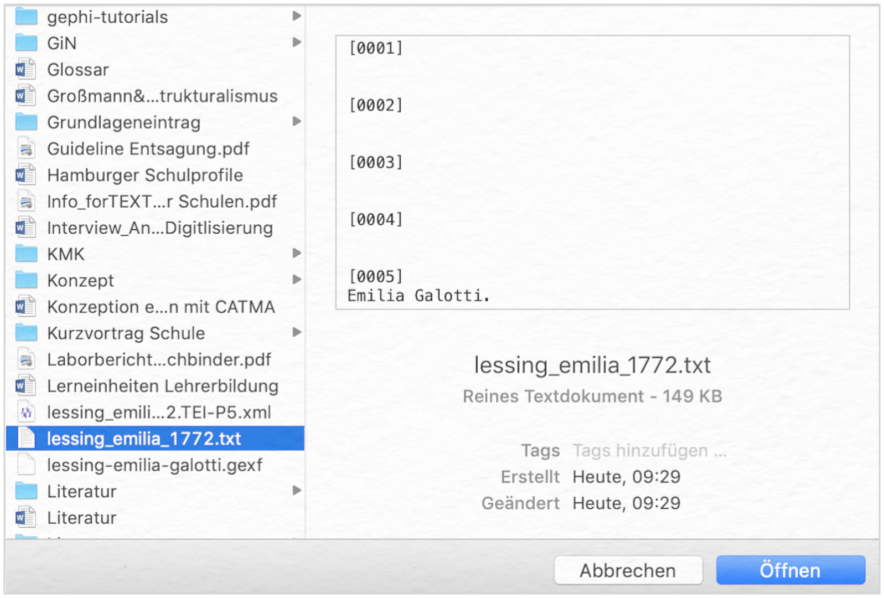

Abb. 8: Das Hochladen des Textes in Voyants Eingabefeld: Klick auf „Upload“, Auswahl der Datei, Bestätigung der Auswahl per Klick auf „Öffnen“

Nachdem Sie das Hochladen der Datei bestätigt haben, lädt Voyant den Text automatisch hoch, worüber das bewegte „Uploading-Corpus“– bzw. „Fetching your Corpus“– Symbol informiert.

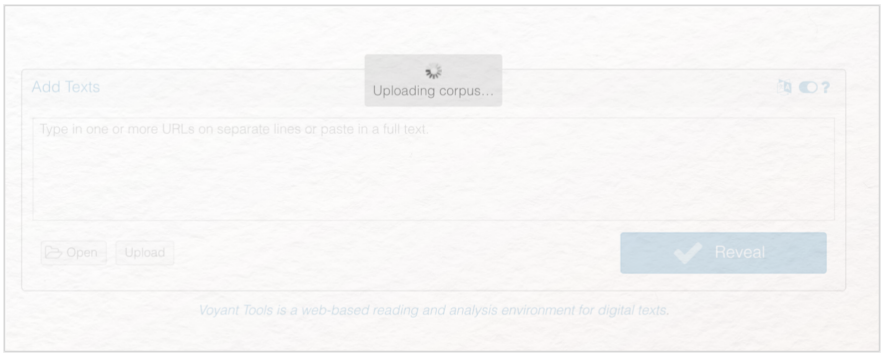

Abb. 9: Hochladen des Textes

Dieser Prozess kann je nach Größe der Datei einige Zeit in Anspruch nehmen, in diesem Fall sollte er aber in wenigen Sekunden abgeschlossen sein. In dieser Zeit wird der von Ihnen ausgewählte Text statistisch ausgewertet (vgl. Text Mining). Nach dem Upload-Prozess öffnet sich automatisch das Voyant-Interface (vgl. GUI) (siehe Abb. 10). Nun öffnet sich die Benutzeroberfläche von Voyant. Nehmen Sie sich einen Augenblick Zeit und verschaffen sich einen ersten Überblick über die Benutzeroberfläche: Sie sehen unterschiedliche Visualisierungen der statistischen Auswertung von *Emilia Galotti*.

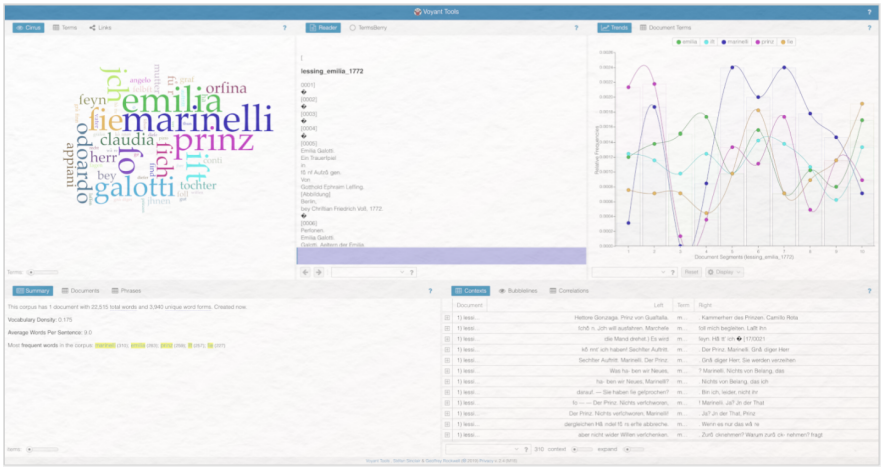

Abb. 10: Das Interface von Voyant mit fünf Panels: Oben v.l.n.r.: Cirrus, Reader, Trends*; Unten v.l.n.r.:* Summary und Contexts

3. Funktionen

Im Rahmen dieser Lerneinheit werden Sie zunächst die voreingestellte Benutzeroberfläche von Voyant kennenlernen. Diese beinhaltet die fünf Tools *Cirrus, Reader, Trends, Contexts* und *Summary*. Jedes dieser Tools ist in einem Panel beheimatet. Die fünf voreingestellten Tools heben unterschiedliche Ergebnisse der Textanalyse auf ganz unterschiedliche Art und Weise hervor. Doch zunächst zu den Gemeinsamkeiten: Jedes Panel verfügt über eine Symbolleiste, über die z. B. Tooleinstellungen angepasst und die entworfenen Grafiken heruntergeladen werden können. Diese Leiste befindet sich stets in der oberen rechten Ecke eines jeden Panels. Die Symbolleiste erscheint erst, wenn Sie mit der Maus über den Bereich links neben dem in jedem Panel sichtbaren Fragezeichen-Symbol hovern. Die Exportfunktionen und die Möglichkeit, das ausgewählte Tool durch ein anderes zu ersetzen, besteht für jedes Panel. Die manuell anpassbaren Tooleinstellungen unterscheiden sich von Tool zu Tool bzw. sind nicht für jedes Tool vorhanden.

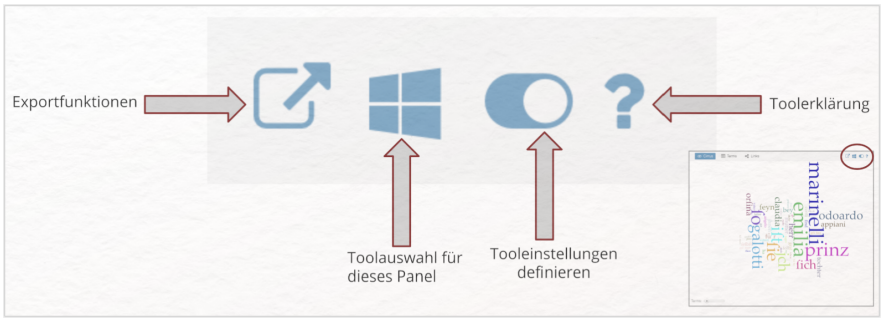

Abb. 11: Die Symbolleiste bei Voyant

Zu einer wichtigen und manuell festlegbaren Tooleinstellung zählt v. a. das Generieren einer **Stoppwortliste**. Stoppwörter sind diejenigen Wörter, die in Ihrer Textanalyse unberücksichtigt bleiben sollen. Sämtliche durch Voyant automatisch durchgeführten Analyseschritte sollen folglich unter Ausschluss dieser Wörter durchgeführt werden. In den meisten Fällen handelt es sich dabei um Funktionswörter (Personalpronomen, Präpositionen, Konjunktionen, Interrogativpronomen, Possessivpronomen), die in Texten – grammatisch bedingt – besonders häufig vorkommen und dadurch den Blick auf die Inhaltswörter (Substantive, Verben, Adjektive) versperren. Da es in diesem Fall vor allem die Inhaltswörter sind, die dabei helfen, aus den Ergebnissen der quantitativen Textanalyse eine Interpretation abzuleiten, möchten wir möglichst viele Funktionswörter auf die Stoppwortliste setzen und diese dadurch aus den Visualisierungen herausrechnen lassen. Erkunden Sie hierzu zunächst die Symbolleiste der unterschiedlichen Tools. Über den „Define options for this tool"-Button gelangen Sie zu der Stoppwortliste.

Aufgabe 1:
Für welche der fünf Tools lässt sich eine Stoppwortliste festlegen und für welche nicht? Welche Grundeinstellungen finden Sie hier vor?

Wie Sie wahrscheinlich festgestellt haben, muss nicht für jedes einzelne Tool eine Stoppwortliste erstellt werden. Durch das Häkchen vor „apply globally“ wird die einmal entworfene Stoppwortliste automatisch in die Auswertung sämtlicher Tools einbezogen. Generell gilt: Die fünf Tools agieren miteinander. Voyant bietet für unterschiedliche Sprachen bereits definierte Stoppwortlisten an. Das ist hilfreich und zeitsparend, da sprachspezifische Funktionswörter bereits in einer Liste gesammelt wurden, die Sie nun per Mausklick von Ihrer Textanalyse ausschließen können. In den folgenden Schritten werden wir nun die deutschsprachige Stoppwortliste aktivieren. Klicken Sie zunächst in einem der drei möglichen Tools auf den „Define options for this tool“-Button. Nun öffnet sich ein Fenster, in dem Sie ihre Stoppwortliste bearbeiten können.

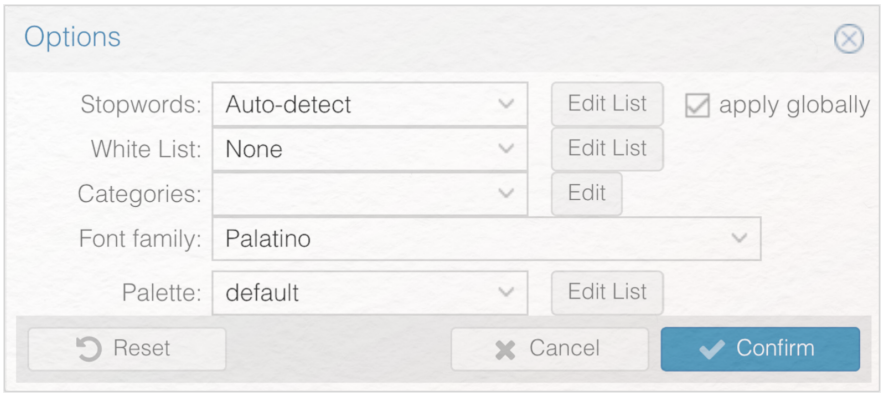

Abb. 12: Bearbeiten der Stoppwortliste in Voyant

Das Erstellen einer guten Stoppwortliste stellt einen zentralen Arbeitsschritt der digitalen Textanalyse mit Voyant dar. Sobald Sie die Stoppwortliste aufgerufen haben, stehen unterschiedliche Möglichkeiten zur Verfügung, um die Liste auf Ihre Anliegen anzupassen. Unter „Stopwords“ finden Sie eine Auswahl aus vordefinierten Stoppwortlisten. Unter „White List“ können Sie Wörter festlegen, die in jedem Fall in die Analysen einbezogen werden sollen, während unter „Categories“ gleiche Wörter in Kategorien zusammengefasst sind, die Sie im Ganzen auf die Stoppwortliste setzen können. Unter „Font family“ können Sie die Schriftart und unter „Palette“ die Farben der Visualisierungen festlegen. Wählen Sie hinter „Stopwords“ nun „German“ aus und bestätigen Sie diese Auswahl mit einem Klick auf den „Confirm“-Button. Damit haben Sie bereits einige für die Interpretation Ihrer Ergebnisse störende Funktionswörter ausgeschlossen. Doch welche sind das überhaupt und wie lassen sich weitere individuell ausgewählte Stoppwörter in die Liste aufnehmen? Um einsehen zu können, welche Wörter auf der Liste enthalten sind, gehen Sie erneut via „Define Options for this Tool“ in das Menü, in dem Sie die Stoppwortliste bearbeiten können (siehe Abb. 12). Hinter der zuvor ausgewählten deutschsprachigen Stoppwortliste gelangen Sie per Klick auf den „Edit-List“-Button auf die detaillierte Ansicht der Stoppwortliste. Nun können Sie zum einen sehen, welche Wörter bereits auf der Liste stehen und zum anderen manuell Wörter ergänzen, indem Sie diese am Ende der Liste eintippen. Pro Zeile, die durch das Betätigen der Enter-Taste voneinander getrennt werden, kann ein Wort eingegeben werden. Durch einen Klick auf den „Save“-Button bestätigen Sie die Aktualisierung Ihrer Stoppwortliste.

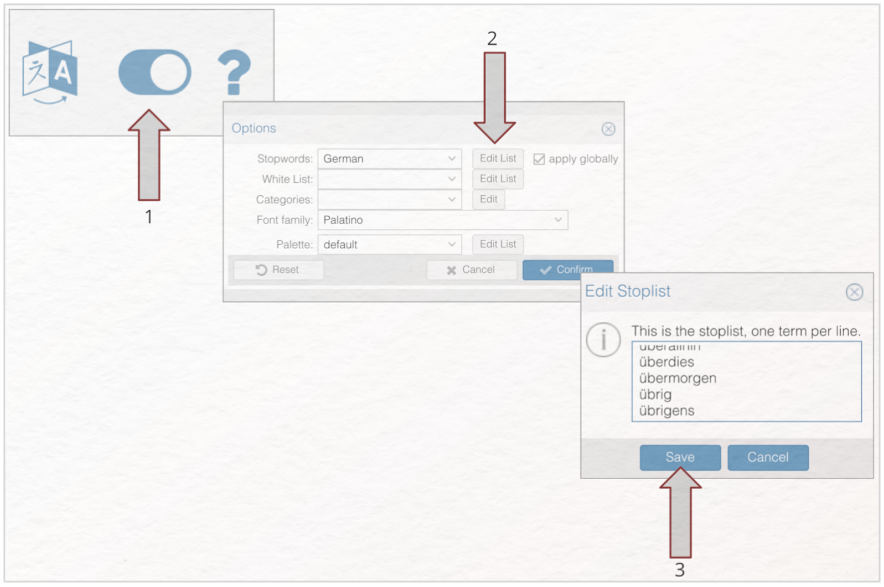

Abb. 13: Die Stoppwortliste ansehen und ergänzen: 1: „Define options for this tool“; 2: „Edit List“; 3: Die Liste ansehen und per Eingabe an das Ende der Liste ergänzen

Aufgabe 2: Welche Wörter sollten auf einer „guten“ Stoppwortliste stehen?

Nun haben Sie einen wichtigen Bestandteil von Voyant kennengelernt, der darüber hinaus in vielen Methoden der digitalen Textanalyse eine tragende Rolle spielt. Eine weitere Funktion, die sämtlich Panels gemeinsam haben, ist die Exportfunktion. Grundsätzlich bestehen zwei Exportmöglichkeiten: Sie können entweder die einzelne Visualisierung eines Panels oder die gesamte Voyant-Sitzung exportieren. Einzelne Abbildungen lassen sich als SVG oder PNG-Datei exportieren. In diesem Fall exportieren Sie quasi eine Momentaufnahme Ihres Analyseprozesses. Darüber hinaus besteht die Möglichkeit, die Visualisierung als URL zu exportieren und sich eine Zitierhilfe anzeigen zu lassen. Wählen Sie die URL-Exportmöglichkeit, exportieren Sie die dynamische Visualisierung in Form der URL, die zu Ihrer Visualisierung zurückführt. Die gesamte Voyant-Sitzung lässt sich nur als URL exportieren und bleibt damit – genau wie der Export der dynamischen Visualisierung – webbasiert. Der Export der Bilddatei ist vor allem dann sinnvoll, wenn Sie die Visualisierungen bspw. in eine Printpublikation einbinden möchten.

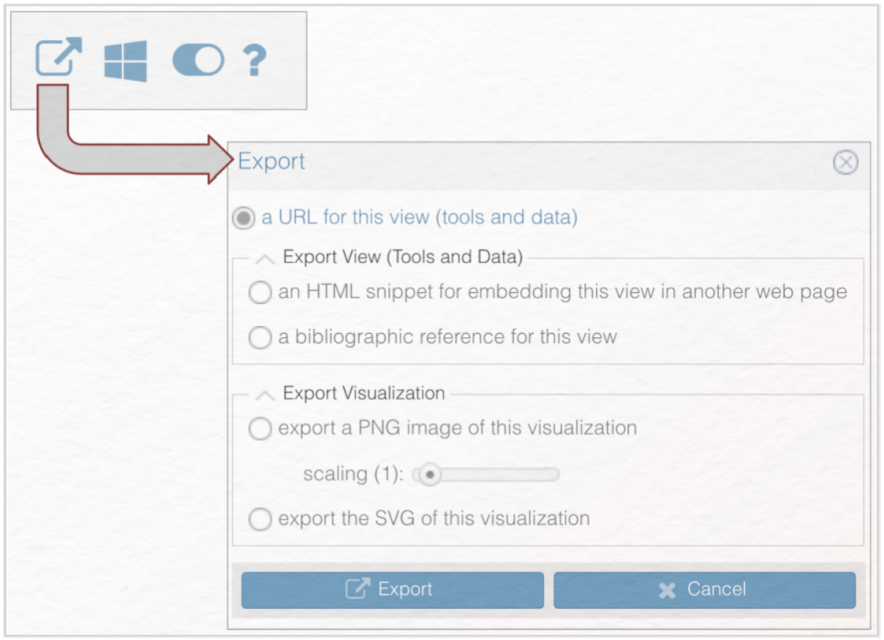

Abb. 14: Aufrufen der unterschiedlichen Exportfunktionen

Nun haben Sie elementare Voyant-Bestandteile kennengelernt, die jegliche Arbeit mit Voyant begleiten werden: die Stoppwortliste, die Funktionen der Menüleiste der einzelnen Panels und die unterschiedlichen Exportfunktionen. Jetzt wenden wir uns den einzelnen Tools und den hier dargestellten Formen der Textvisualisierungen zu. In den unterschiedlichen Panels werden die Ergebnisse der statistischen Auswertung von *Emilia Galotti* unterschiedlich dargestellt. Doch welches Panel verbildlicht welche Textdaten und welche Interpretation lässt sich daraus ableiten? Um die schrittweise erfolgende Beantwortung dieser Fragen wird es in dem folgenden Teil der Lerneinheit gehen. Kernelement der digitalen Textanalyse mit Voyant sind die Auswertung und Interpretation der erhobenen Daten und deren Visualisierung.

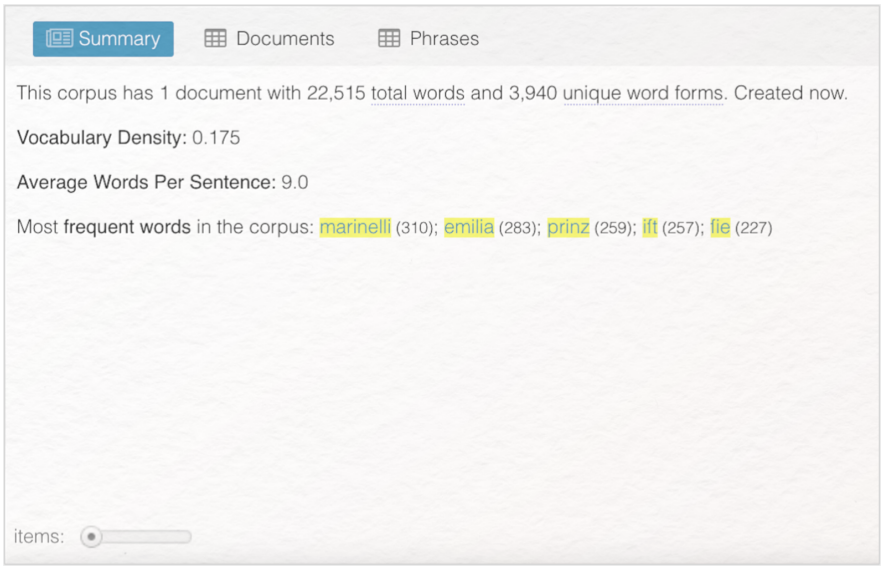

Abb. 15: Das Summary*-Panel in Voyant*

Summary listet die wichtigsten Ergebnisse der quantitativen Textanalyse auf und hat vor allem eine informierende Funktion. Angezeigt wird die gesamte Wortzahl, die Wortdichte (Relation von Type und Token (vgl. Type/Token) bzw. Wortschatz: je näher die Token- der Typezahl, desto größer der jeweilige Wortschatz), die durchschnittliche Anzahl von Wörtern pro Satz sowie die am häufigsten vorkommenden Wörter (*most frequent words, MFW*). Durch einen Blick auf *Summary* entsteht ein erstes Gefühl für die quantitative Beschaffenheit der Textgrundlage: Hier finden Sie die konkreten Zahlen, auf denen z. B. die Wortwolke (vgl. Wordcloud) basiert. *Summary* ist vor allem dann hilfreich, wenn Sie mit einem aus mehreren Texten bestehenden Textkorpus arbeiten. In diesem Fall werden für jedes Textdokument die Ergebnisse der quantitativen Auswertung aufgelistet. Anhand der Ergebnisse über Wortvorkommen, Textlänge etc. lassen sich die Texte rein quantitativ miteinander vergleichen.

Abb 16: Das Cirrus*-Panel in Voyant*

In ***Cirrus*** wird eine Wortwolke dargestellt. Hier werden Wörter der Häufigkeit ihres Vorkommens entsprechend visualisiert. Häufig vorkommende Wörter werden größer dargestellt als selten vorkommende Wörter.

Aufgabe 3: Werten Sie die Wortwolke aus: Welche Wörter kommen besonders häufig, welche weniger häufig vor? Leiten Sie basierend auf den Worthäufigkeiten eine erste Interpretationshypothese ab, indem Sie Vermutungen über Inhalte, Handlungen, Figuren, Orte oder epochencharakteristische Merkmale anstellen. Spielen Farbe und die topografische Anordnung der Wörter eine Rolle und wenn ja, welche? Exportieren Sie Ihre Grafik als PNG-Dokument.

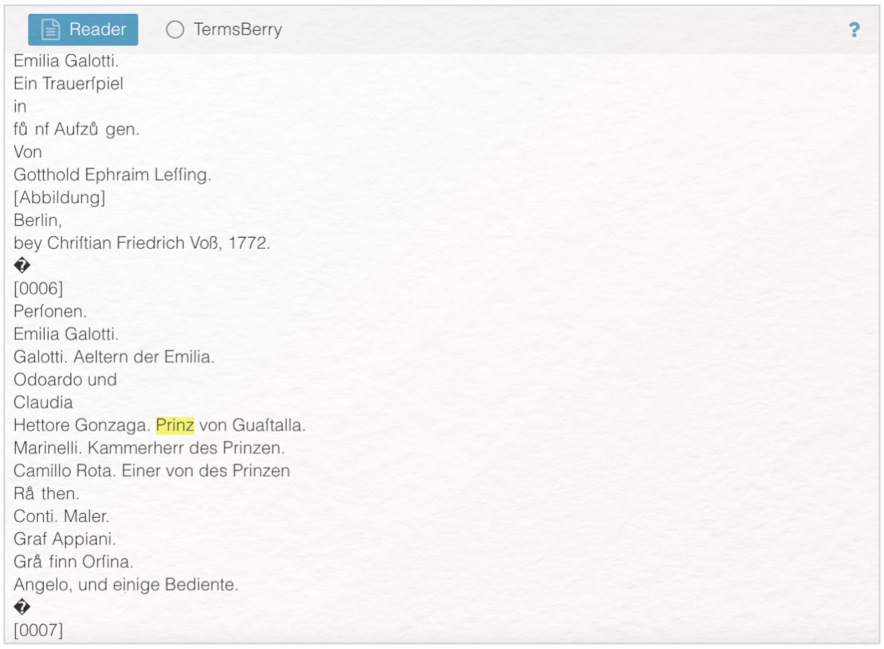

Abb. 17: Das Reader*-Panel in Voyant*

Der ***Reader*** bildet den gesamten hochgeladenen Text ab. Wenn Sie in diesem Panel mit der Maus über einzelne Wörter hovern, wird Ihnen die Vorkommenshäufigkeit dieses Wortes angezeigt. Über eine unter dem Text

platzierte Suchleiste können Sie nach jedem beliebigen Wort innerhalb des Textes suchen. Das Ergebnis wird im Text gelb markiert.

Aufgabe 4: Welche Wörter, die mit dem Wort „schuld“ verwandt sind, kommen in *Emilia Galotti* vor?
Wie bereits erwähnt, sind die einzelnen Tools in Voyant miteinander vernetzt: Das gilt auch für die Tools *Reader*, *Trends* und *Contexts*.

Aufgabe 5: Klicken Sie im *Reader* auf „Marinelli“. Wie reagieren die Einstellungen der anderen Tools darauf?

Trends visualisiert das Vorkommen der fünf am häufigsten im gesamten Text/Textkorpus vorkommenden Wörter. Eine Legende am oberen Rand des Graphen teilt jedem Wort eine Farbe zu. Da Sie vorher im Reader „Marinelli“ ausgewählt haben, müssen Sie Trends zunächst in seine Grundeinstellung zurückversetzen, um nicht mehr ausschließlich Marinellis Verlaufskurve, sondern diejenige sämtlicher MFWs angezeigt zu bekommen. Hierfür wird die „Reset“-Funktion verwendet, die fester Bestandteil eines jeden Panels ist und manuell vorgenommene Tooleinstellungen deaktiviert.

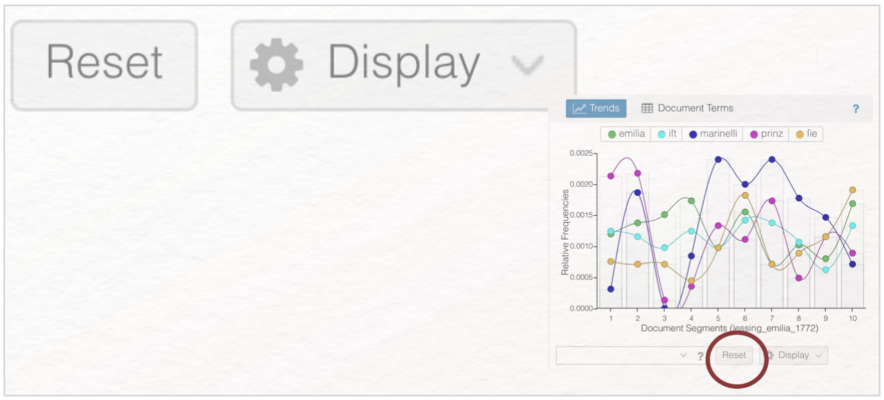

Abb. 18: Die „Reset“-Taste bei Voyant

Klicken Sie nun auf die „Reset“-Taste und setzen das Tool in seine Grundeinstellung zurück. Wenn Sie auf ein Wort in der Legende klicken, wird es im Graphen nicht mehr angezeigt. Voyant teilt den gesamten Text in *Trends* automatisch in gleichgroße Segmente auf. Diese werden im Graphen als farbige Säulen dargestellt. Für jedes Segment wird angegeben, wie häufig jedes der fünf Wörter innerhalb dieses Textsegments vorkommt. Wenn Sie über die Knotenpunkte hovern, werden Ihnen Informationen zu dem jeweiligen Wort (Wort, Häufigkeit, Titel des Textes, Segmentnummer) dargeboten. Sobald Sie auf den Knotenpunkt klicken, wird eine Vernetzung zu zwei weiteren Tools hergestellt: *Contexts* und *Reader*. Im *Reader* wird das angeklickte Wort innerhalb des gesamten Textes gelb markiert (siehe Abb. 17). Unter *Contexts* (s. u.) werden Ihnen Textstellen angezeigt, die vor und nach dem angeklickten Wort auftauchen.

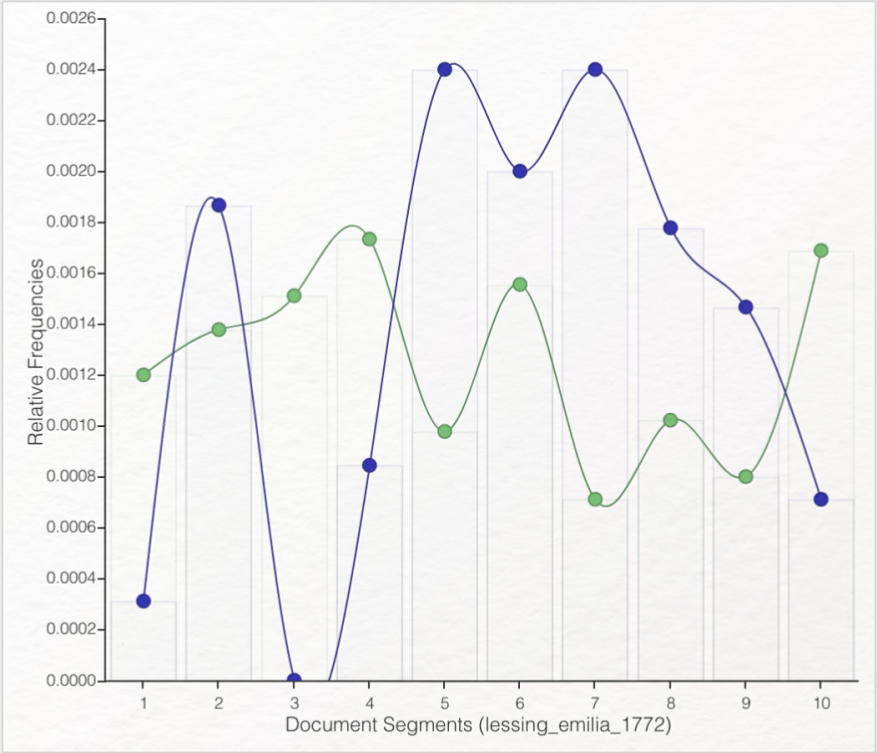

Abb. 19: Das Trends*-Panel in Voyant*

Aufgabe 6: Erstellen Sie einen Graphen, in dem nur die beiden häufigsten Wörter vorkommen und exportieren diese Grafik. In welchen beiden Segmenten kommen diese Wörter am häufigsten vor und in welchen Segmenten kommen beide Wörter zusammen (vgl. Kollokation) vor? Welche Rückschlüsse lassen sich hieraus auf die Figurenkonstellation des Stücks ziehen?
Tipp: Nutzen Sie den „Display"-Button rechts neben der „Reset"-Funktion, um die aktuelle Verlaufskurve durch einen anderen Graphentypen zu ersetzen.

Contexts zeigt Ihnen die direkte Umgebung eines ausgewählten Wortes an und steht in Verbindung mit den Tools *Trends* und *Reader*. *Contexts* ermöglicht also – genau wie *Reader* –, aus der statistischen Erhebung heraus wieder in den Text zurückzukehren und sich das Umfeld eines ausgewählten Wortes – die sog. Keywords in Context (vgl. KWIC) – genauer anzuschauen.

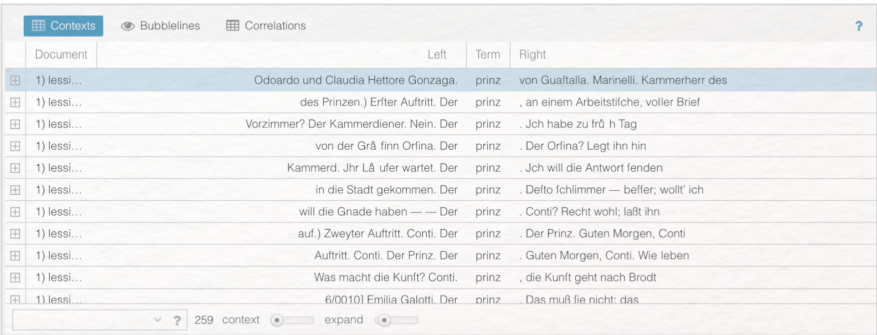

Abb. 20: Das Contexts*-Panel in Voyant*

Nun haben Sie die fünf klassischen und zum Teil miteinander vernetzten Voyant-Tools kennengelernt. Während *Reader* und *Contexts* eher als Close-Reading-Tools (vgl. Close Reading) bezeichnet werden können, haben Sie mit *Cirrus*, *Trends* und *Summary* drei Tools der quantitativen Textanalyse kennengelernt. Darüber hinaus haben Sie einige der Grafiken interpretiert und damit den Brückenschlag zwischen statistischer Auswertung von Texten und literaturwissenschaftlicher Interpretation gemacht. Die Grundlage Ihrer Interpretation – die Visualisierungen – haben Sie exportiert. Neben der Arbeit mit den fünf voreingestellten Tools besteht in jedem Panel die Möglichkeit, das voreingestellte Tool durch ein anderes Werkzeug aus dem Voyant-Toolkit zu ersetzen. Anders als die standardmäßig aktivierten Tools sind die inaktiven Tools nicht blau hinterlegt. Im *Cirrus*-Panel können

Sie per Mausklick auf die Tools *Terms* oder *Links* umschalten, während *Reader* sich durch *TermsBerry* ersetzen lässt. *Trends* können Sie mit *Document Terms* auswechseln, *Contexts* mit *Bubblelines* oder *Correlations* austauschen. *Summary* kann per Mausklick auf die Tools *Documents* oder *Phrases* umgeschaltet werden.

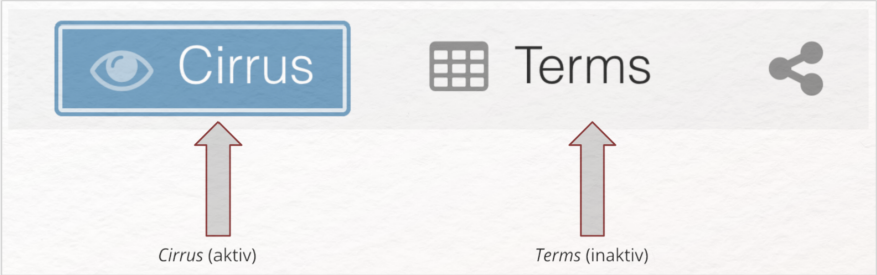

Abb. 21: Umschaltmöglichkeit innerhalb eines Panels: Cirrus *ist hier aktiviert, per Mausklick lässt sich von* Cirrus *auf* Terms *umschalten*

Nehmen Sie sich etwas Zeit, um die unterschiedlichen Tools auszuprobieren. Keine Angst: Sie können ohne Bedenken zwischen unterschiedlichen Tools hin- und herspringen. *TermsBerry* visualisiert durch farbliche Kennzeichnungen, im Zusammenhang mit welchen anderen Wörtern die MFWs besonders häufig vorkommen.

Aufgabe 7: Ersetzen Sie *Reader* durch *TermsBerry*. In Verbindung mit welchen vier Wörtern kommt „marinelli“ am häufigsten vor und wie verhält es sich mit „emilia“? Welche Rückschlüsse lassen sich hieraus ziehen?

Bei dem Tool *Bubblelines*, welches in dem *Contexts*-Panel ausgewählt werden kann, repräsentiert die horizontale Linie das Textdokument. Diese Linie, d. h. das Textdokument, ist in gleich lange Segmente unterteilt (Grundeinstellung: 50 Segmente, auch dies lässt sich in den Tooleinstellungen verändern). Jedes von Ihnen ausgewählte Wort wird als kreisförmige Blase (eben als Bubble) angezeigt. Die Größe der Blase zeigt die Häufigkeit, in der dieses Wort innerhalb des Segments vorkommt: Je größer die Blase, desto häufiger kommt das Wort in dem jeweiligen Textsegment vor. Wenn Sie *Bubblelines* auswählen, werden zunächst sämtliche MFWs in einer Linie angezeigt. Das Durcheinander von bunten Bubbles kann schnell etwas unübersichtlich anmuten. Um einzelne Wörter aus der Bubbleline zu entfernen, klicken Sie einfach auf das entsprechende Wort oberhalb der Bubbleline und blenden es per „Hide Term“ aus oder entfernen es endgültig via „Remove Term“. Durch die Interdependenz der Panels kann es sein, dass in Ihrer Bubbleline zunächst nur ein bestimmtes Wort angezeigt wird. Sollten Sie in anderen Panels ein bestimmtes Wort angeklickt haben, wird Ihnen bspw. nur die Bubbleline dieses Wortes angezeigt. Um sämtliche MFWs zu integrieren, nutzen Sie das Eingabefeld in der unteren linken Ecke des Panels (siehe Abb. 22). Begriffe können über die Suchbox hinzugefügt werden, indem Sie den Suchbegriff eingeben und mit „Enter“ bestätigen.

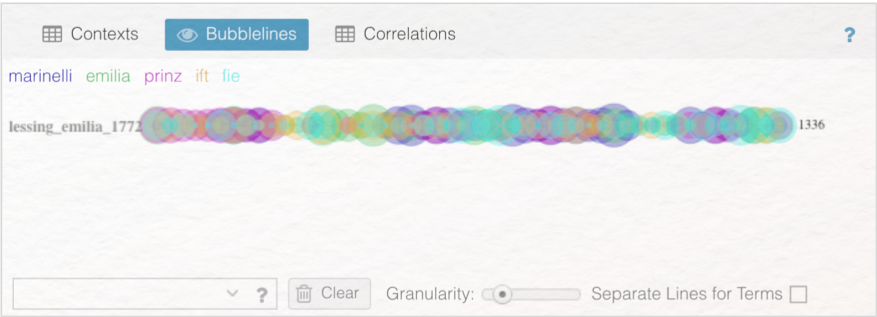

Abb 22: Das Bubblelines*-Panel in Voyant*

Aufgabe 8: Ersetzen Sie *Contexts* durch *Bubblelines* und integrieren Sie die fünf MFWs in Ihre Bubbleline. Klicken Sie auf „Seperate Lines from Terms“, um sich pro Segment nur das Vorkommen eines der MFWs anzeigen zu lassen. Blenden Sie die Wörter „ist“ und „sie“ aus. Exportieren Sie auch diese Grafik. Was bildet dieses Tool ab und wie lassen sich die Ergebnisse interpretieren?

Wie anfangs erwähnt, handelt es sich bei Voyant um den Zusammenschluss von mittlerweile 29 unterschiedlichen Tools, die jeweils unterschiedliche Formen der Textvisualisierung ermöglichen. In den vorangegangenen Schritten haben Sie unterschiedliche interne Tools angewendet. Sie sind Bestandteil der standardmäßig festgelegten Benutzeroberfläche. Den Einbezug eines weder in den Voreinstellungen noch in der Menüleiste der Panels enthaltenen Tools – d. h. eines externen Voyant–Tools – setzen Sie in wenigen Schritten in die Tat um:

Klicken Sie in der Menüleiste des gewünschten Panels auf „Click to choose another tool for this panel location“. Nun können Sie ein beliebiges Tool für das jeweilige Panel auswählen.

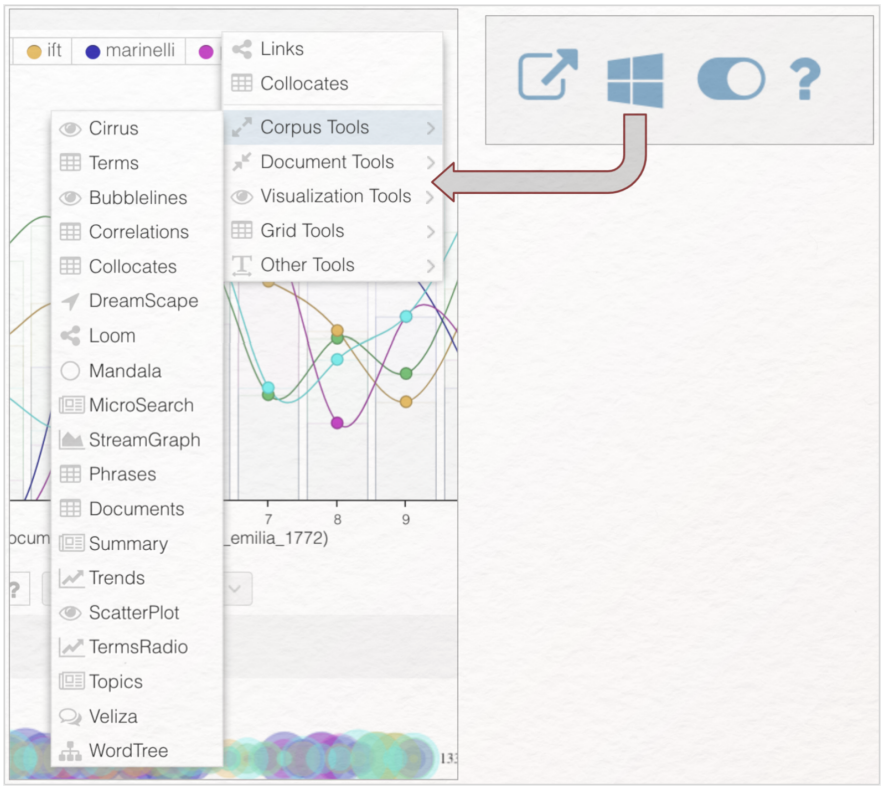

Abb. 23: Ersetzen der voreingestellten Tools durch die Integration eines externen Voyant-Tools

Auf diese Weise können Sie sich eine individuelle Toollandschaft erstellen. Die 29 unterschiedlichen Tools lassen sich in unterschiedliche Kategorien einteilen. Einige eignen sich besonders gut, um einzelne Dokumente zu untersuchen, so wie Sie es in dieser Lerneinheit gemacht haben. Andere sind dafür prädestiniert, ein aus mehreren Texten bestehendes Textkorpus zu analysieren, so wie das vorbereitete Shakespeare- oder Austen-Textkorpus. Wieder andere Tools erstellen ausschließlich Rastergrafiken. An den folgenden Symbolen erkennen Sie die Kategorien.

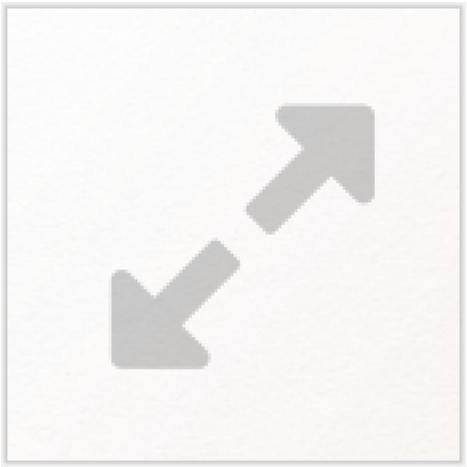

Abb. 24: Symbol für Corpus Tools

Corpus Tools: Tools, die sich für die Analyse eines Textkorpus eignen: *Cirrus, Terms, Bubblelines, Correlation, Collocates, DreamScape, DreamScape, Loom, Mandala, MicroSearch, StreamGraph, Phrases, Documents, Summary, Trends, ScatterPlot, TermsRadio, Topics, Veliza, WordTree*.

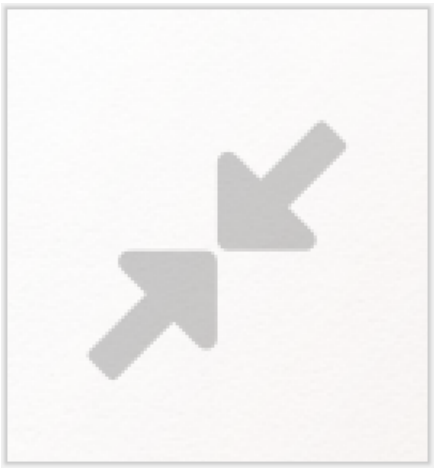

Abb. 25: Symbol für Document Tools

Document Tools: Tools, die sich für die Analyse einzelner Dokumente eignen: *Bubbles*, *Cirrus*, *Document Terms*, *Reader*, *TextualArc*, *Trends*, *Knots*, *Topics.*

Abb. 26: Symbol für Visualizationtools

Visualizationtools: *Cirrus*, *Bubbles*, *Bubblelines*, *Links*, *DreamScape*, *Loom*, *Knots*, *Mandala*, *MicroSearch*, *StreamGraph*, *ScatterPlot*, *TextualArc*, *Trends*, *Termsberry*, *TermsRadio*, *WordTree.*

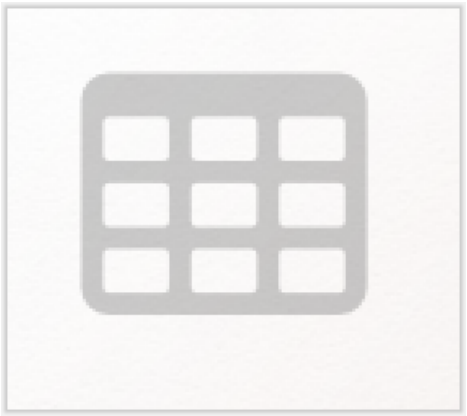

Abb. 27: Symbol für Grid Tools

Grid Tools: Tools, die Rastergrafiken/Tabellen erstellen: *Terms*, *Collocate, Correlations*, *Phrases*, *Contexts*, *Document Terms*, *Documents*,*Topics*.

Aufgabe 9: Erstellen Sie eine Toollandschaft, die aus jeder der vier Hauptkategorien mindestens ein Tool enthält

und exportieren Sie diese Voyant-Sitzung als URL.

Es ist ratsam, sich vorab im Voyant-Guide über die Funktionen der unterschiedlichen Tools zu informieren, damit Sie an dieser Stelle bewusst eine bestimmte Visualisierung auswählen können, deren Ergebnisse Sie interpretieren können. Ohne Kenntnis über die Konzeption und Funktion der unterschiedlichen Visualisierungen kann deren Auswertung und Interpretation kaum gelingen.
Nun sind Sie am Ende der Lerneinheit angelangt und haben wichtige Schritte der digitalen Textanalyse mit Voyant ausprobiert: Korpuserstellung, Hochladen der Daten, Erstellen einer Stoppwortliste, Analyse und Interpretation der Visualisierungen (*Cirrus*, *Trends*, *Contexts*, *Summary*, *Bubblelines*, *Reader*, *TermsBerry*) und Export der Visualisierungen. Ihre Interpretationsansätze beruhen auf der statistischen Auswertung von *Emilia Galotti*, dem Sie sich mit Distant-Reading-Verfahren angenähert haben. Um Ihre Hypothesen zu überprüfen, empfiehlt sich spätestens jetzt die Methode des Close Reading: Kommt die Titelfigur wirklich weniger häufig vor als Marinelli? In welchem Verhältnis stehen die *most frequent words* „marinelli“, „prinz“ und „emilia“ zueinander? Adel vs. Bürgertum: Trifft diese Konstellation zu und wie wirken Gefühlswörter wie „herz“, „hand“ oder „auge“ auf Sie?

4. Lösungen zu den Beispielaufgaben

Aufgabe 1: Für welche der fünf Tools lässt sich eine Stoppwortliste festlegen und für welche nicht? Welche Grundeinstellungen finden Sie hier vor?

Eine Stoppwortliste lässt sich über die Symbolleiste der Tools *Cirrus*, *Trends* und *Summary* erstellen. Über die Tools *Reader* und *Contexts* kann keine Stoppwortliste erstellt werden. Die Voreinstellungen hinter „Stopwords“ sind auf „Auto-detect“ eingestellt. Bestimmte Stoppwörter werden also automatisch bereits erfasst und in den Visualisierungen nicht abgebildet. Das angekreuzte „apply globally“ verweist darauf, dass eine Stoppwortliste für sämtliche Tools gilt – unabhängig davon, über welches der drei Tools die Stoppwortliste bearbeitet wird.

Aufgabe 2: Welche Wörter sollten auf einer „guten“ Stoppwortliste stehen?

Die ersten drei Wörter der deutschsprachigen Stoppwortliste lauten „ab“, „aber“ und „abgerufen“. Dabei handelt es sich um Funktionswörter, die für eine Interpretation oder Analyse des Textes nicht essentiell sind und deshalb auch in der Visualisierung nicht zwingend abgebildet werden müssen. Das Erstellen einer fundierten und gut durchdachten Stoppwortliste stellt einen wichtigen vorbereitenden Schritt der digitalen Textanalyse dar, weil sie sich in den generierten Visualisierungen niederschlägt, die in der digitalen Textanalyse häufig die Grundlage der Interpretation darstellen. Auf einer „guten“ Stoppwortliste stehen folglich die Wörter, die keine tragende Rolle bei der Textinterpretation spielen und die Sie auch in einer völlig analog durchgeführten Textanalyse nicht berücksichtigen würden. Wörter wie „zu“, „in“, „aber“, „ab“, „welcher“, „welchem“ oder „welche“ können relativ bedenkenlos auf die Liste gesetzt werden und stehen berechtigterweise auch bei Voyant auf der umfangreichen deutschsprachigen Stoppwortliste. Dennoch gilt grundlegend: Die Stoppwortliste stellt einen Filter dar, der die Ergebnisse der digitalen Textanalyse beeinflusst. Transparenz und Nachvollziehbarkeit sind auch hier zu berücksichtigen.

Aufgabe 3: Werten Sie die Wortwolke aus: Welche Wörter kommen besonders häufig vor und welche Wörter kommen weniger oft vor? Leiten Sie basierend auf den Worthäufigkeiten eine erste Interpretationshypothese ab, indem Sie Vermutungen über den Inhalt, Handlungen, Figuren, Orte oder epochencharakteristische Merkmale anstellen. Spielen Farbe und die topografische Ausrichtung der Wörter eine Rolle und wenn ja, welche?

Abb. 28: „Cirrus: Emilia Galotti", Bildquelle: Sinclair und Rockwell (2019): https://voyant-tools.org/docs/#!/guide/cirrus

In der Wordcloud sind die Wörter „marinelli" (325), „prinz" (259), „emilia" (131), „odoardo" (120), „claudia" (92), „orsina" (85) und „appiani" (74) besonders zentral und groß abgebildet. Die Visualisierung liefert einen Überblick über die handelnden Personen, die im Mittelpunkt der Handlung stehen. Die titelgebende Figur „emilia" ist in der Wortwolke allerdings kleiner abgebildet als „marinelli" oder „prinz". Schließt man von der Worthäufigkeit auf die Zentralität der in dem Stück beteiligten Figuren – also auf die Hauptfigur – müssten „marinelli" oder der „prinz" die Hauptfiguren sein. Der Schluss von der Worthäufigkeit auf die Zentralität einer Figur bzw. auf die Hauptfigur stellt dabei nur eine Möglichkeit von vielen dar, um Zentralität zu definieren und lädt angesichts des Titels dazu ein, die Figurenkonstellation genauer zu untersuchen. Die ebenfalls in der Wortwolke abgebildeten Interjektionen „o" und „ah" verweisen auf eine ausdrucks- und gefühlsbetonte Sprache. „grafen" (51), „gräfin" (43) und „gnädiger" (31) zeigen, dass der untersuchte Text die Rolle des Adels thematisiert und einige der handelnden Figuren dem Adel zugerechnet werden können. Dem adeligen Setting und einer möglichen politischen Dimension kann ein familiäres, bürgerliches Setting gegenübergestellt werden. Die „tochter"- (72), „mutter"- (61), „vater"- (47) Konstellation wird in der Wortwolke deutlich abgebildet und verweist auf eine bürgerliche Opposition. Zentrale Themenkomplexe des Textes könnten die Themenfelder Recht („recht": 35), Tod („tot": 22), Wahrheit („wahr": 20), Leben („leben": 15) und Ehre („ehre": 14) berühren. Die Lexeme „herz" (14) „hand" (25) und „augen" (25) werden zwar nicht primär dem Emotionswortschatz zugeordnet, gelten jedoch als bildlicher Ausdruck für Emotionen. Sie werden als Signalwörter behandelt, die inhaltlich über ein besonders hohes Assoziationspotenzial verfügen, um Gefühle darzustellen. Hinter diesen Gefühlswörtern verbergen sich häufig Konzeptualisierungsstrukturen von emotionalem Erleben. Sie werden – z. B. in Literaturkritiken – eingesetzt, um ein intrasubjektives Erleben zu gestalten. Die Lexeme ermöglichen der Leserschaft also ein Sich-Einfühlen in die erzählte Welt. Werden diese Wörter besonders häufig verwendet, könnte man dem Autor eine „Sinngebung der Empfindung durch Ritualisierung" (Stoeva-Holm 2015, 37f.) unterstellen. Eine mögliche Hypothese könnte es also sein, dass in dem untersuchten Text Gefühlswörter eingesetzt wurden, um auf die Leserschaft einzuwirken und sie emotional zu involvieren. Weder die Ausrichtung der Wörter in der Wortwolke noch die Farbgebung haben eine semantische Bedeutung. Ob die Wörter horizontal oder vertikal ausgerichtet sind oder welche Position innerhalb der Visualisierung sie einnehmen, spielt keine Rolle.

Aufgabe 4: Welche Wörter, die mit dem Wort „schuld" verwandt sind, kommen in *Emilia Galotti* vor?

Die Wörter „schuld" (5), „schuldigkeit" (4) und „schulden" (1) kommen in Lessings Trauerspiel wortwörtlich vor. Es handelt sich aber um eher selten vorkommende Wörter.

Aufgabe 5: Klicken Sie im *Reader* auf „Marinelli“. Wie reagieren die Einstellungen der anderen Tools darauf?

Wenn Sie im *Reader* auf „Marinelli“ klicken, wird der Name im gesamten im *Reader* abgebildeten Text gelb markiert. *Trends* generiert eine Einzelansicht der Wortverteilung im gesamten Text. Im dritten Segment kommt „marinelli“ am wenigsten vor.

Aufgabe 6: Erstellen Sie einen Graphen, in dem nur die beiden häufigsten Wörter vorkommen und exportieren diese Grafik. In welchen beiden Segmenten kommen diese Wörter am häufigsten vor und in welchen Segmenten kommen beide Wörter zusammen vor? Welche Rückschlüsse lassen sich hieraus auf die Figurenkonstellation des Stücks ziehen?

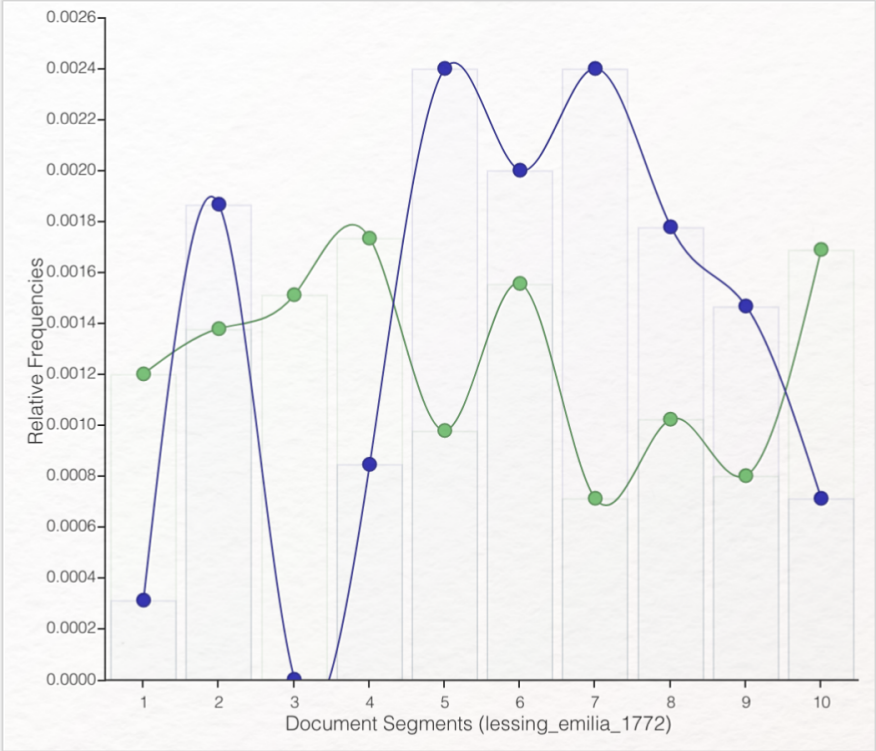

Abb. 29: „Trends: emilia und marinelli, Bildquelle: Sinclair und Rockwell (2019): https://voyant-tools.org/docs/#!/guide/trends

Um in *Trends* wieder alle MFWs anzeigen zu lassen (nicht mehr nur „marinelli“), müssen Sie den Graphen zunächst per Klick auf den „Reset“-Button in die Grundeinstellung zurückversetzen. Nun können Sie „claudia“, „odoardo“ und „prinz“ aus dem Graphen entfernen, indem Sie diese Wörter anklicken. Bei „marinelli“ und „emilia“ handelt es sich um die *most frequent words*. „marinelli“ kommt in den Segmenten fünf und sieben am häufigsten vor, während „emilia“ in den Segmenten vier und zehn am häufigsten genannt wird. Die Verlaufskurve beider Wortvorkommen verdeutlicht, dass die titelgebende „emilia“ deutlich weniger häufig vorkommt, als „marinelli“. Die bei der Auswertung der Wordcloud aufgestellte Annahme verhärtet sich: Die Titelfigur ist hier nicht mit der Hauptfigur gleichzusetzen. Schließt man von der Häufigkeit des Vorkommens eines Wortes im gesamten Text auf den Rang einer Figur, kann in diesem Fall Marinelli als Hauptfigur bestimmt werden. Die Beantwortung der Frage, warum die weibliche titelgebende Figur nicht die Hauptfigur des Stücks ist, könnte den Ausgangspunkt für eine tiefergehende genderspezifische Erforschung des Textes darstellen.

Aufgabe 7: Ersetzen Sie *Reader* durch *TermsBerry*. In Verbindung mit welchen vier Wörtern kommt „marinelli“ am häufigsten vor und wie verhält es sich mit „emilia“? Welche Rückschlüsse lassen sich hieraus ziehen?

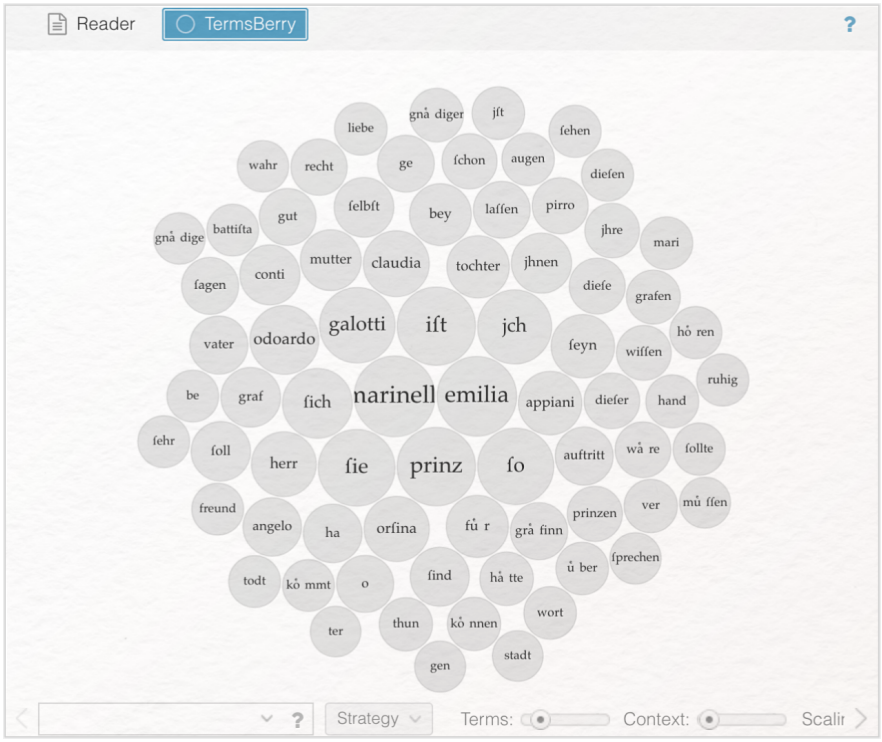

Abb. 30: „TermsBerry: Emilia Galotti", Bildquelle: Sinclair, Stéphan und Geoffrey Rockwell (2019): https://voyant-tools.org/docs/#!/guide/termsberry

TermsBerry visualisiert – genau wie die Wortwolke – besonders häufig vorkommende Wörter, indem es diese in zentraler Position und innerhalb einer Bubble mit größerem Radius abbildet. Hovert man über die Wörter, wird durch die Farbgebung angezeigt, mit welchen anderen Wörtern das ausgewählte Wort besonders häufig gemeinsam vorkommt. „marinelli" kommt am häufigsten in Verbindung mit „prinz" (23) vor, gefolgt von „emilia" (28), „ich" (14) und „ist" (9). „emilia" taucht am häufigsten mit „galotti" (183), „marinelli" (27), „appiani (8) und"orfina" (5) auf. Das Tool eignet sich, um erste Vermutungen bspw. über Figuren und deren Beziehungen zueinander anzustellen. Marinelli scheint in einer sehr engen Beziehung zu dem Prinzen zu stehen. Der Prinz könnte also eine für Marinelli wichtige Bezugsperson darstellen. Anders als für Emilia, die zwar für Marinelli eine wichtige Rolle spielt, mit dem Prinzen aber vergleichsweise wenig interagiert. Marinelli und Emilia können als zentrale Figuren aufgefasst werden. Dass „galotti" überdurchschnittlich oft in Verbindung mit „emilia" vorkommt, ist nicht verwunderlich. Schließlich handelt es sich hierbei um den Nachnamen der Figur. Über Marinelli könnte man mutmaßen, dass er ichbezogene, egozentrische oder besonders selbstbewusste Charaktereigenschaften aufweist, da er den eigenen Standpunkt vertritt und/oder häufiger über sich selbst spricht als Emilia es tut.

Aufgabe 8: Ersetzen Sie *Contexts* durch *Bubblelines* und blenden die Wörter „ist" und „sie" aus. Was bildet dieses Tool ab und wie lassen sich die Ergebnisse interpretieren?

Das Tool zeigt, wie häufig die Wörter „marinelli", „emilia" und „prinz" in dem gesamten Text, der in 50 gleich große Segmente unterteilt wurde, vorkommen. Durch den Vergleich der drei *Bubblelines* wird deutlich, dass „marinelli" im Vergleich zu „prinz" und „emilia" am häufigsten vorkommt. „emilia" taucht stetig auf, während „marinelli" und „prinz" phasenweise gar nicht vorkommen. Darüber hinaus lässt diese Visualisierung Rückschlüsse auf die Interaktion der Figuren untereinander zu: „prinz" und „marinelli" kommen häufiger zusammen vor als „marinelli" und „emilia". Der längste zusammenhängende Abschnitt, in dem „prinz" und „emilia" gemeinsam vorkommen, findet sich in der Mitte der Bubbleline. Geht man davon aus, dass ein gleichzeitiges Vorkommen beider Wörter auf eine tatsächliche Begegnung und Interaktion beider Figuren miteinander im Text schließen lässt, stehen „marinelli" und der „prinz" sich besonders nahe.

Externe und weiterführende Links

- „Cirrus: Emilia Galotti", Bildquelle: Sinclair und Rockwell (2019): https://web.archive.org/save/https://voyant-tools.org/docs/#!/guide/cirrus (Letzer Zugriff: 17.06.2024)
- „Trends: emilia und marinelli, Bildquelle: Sinclair und Rockwell (2019): https://web.archive.org/save/https://voyant-tools.org/docs/#!/guide/trends (Letzer Zugriff: 17.06.2024)
- „TermsBerry: Emilia Galotti", Bildquelle: Sinclair, Stéphan und Geoffrey Rockwell (2019): https://web.archive.org/save/https://voyant-tools.org/docs/#!/guide/termsberry (Letzer Zugriff: 17.06.2024)

- Voyant Startseite: https://web.archive.org/save/https://voyant-tools.org/ (Letzer Zugriff: 18.06.2024)
- Voyant-Guide: https://web.archive.org/save/https://voyant-tools.org/docs/#!/guide/about (Letzer Zugriff: 18.06.2024)

Bibliographie

Flüh, Marie. 2024. Toolbeitrag: Voyant. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 5. Textvisualisierung (7. August). doi: 10.48694/fortext.3775, https://fortext.net/tools/tools/voyant.

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.

Stoeva-Holm, Dessislava. 2015. Gefühle worten. Zum Emotionalisieren in zeitgenössischer Literaturkritik. In: *Literaturkritik heute: Tendenzen – Traditionen – Vermittlung*, hg. von Heinrich Kaulen und Christina Gansel, 27–42. Göttingen: V & R Unipress.

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.

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.

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.

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.

Kollokation Als Kollokation bezeichnet man das häufige, gemeinsame Auftreten von Wörtern oder Wortpaaren in einem vordefinierten Textabschnitt.

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.

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.

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.

Stoppwortliste Stoppwörter sind hochfrequente Wörter, meist Funktionswörter, die, aufgrund ihrer grammatisch bedingten Häufigkeit, beispielsweise die Ergebnisse von inhaltlichen oder thematischen Analysen verzerren können. Deshalb werden diese Wörter, gesammelt in einer Stoppwortliste, bei digitalen Textanalysen meist nicht berücksichtigt.

SVG SVG steht für *Scalable Vector Graphics* und ist ein freies, standardisiertes Dateiformat, das Bilddateien bezeichnet, die als 2D-Vektorgrafiken größenunabhängig reproduziert werden können. Bei SVG-Dateien wird im Gegensatz zu anderen Bildgrafiken somit die Auflösung der Abbildung beim Vergrößern nicht schlechter. Es basiert auf den Strukturen von XML und wird dazu verwendet, Bilddaten zu repräsentieren.

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.