Lerneinheit: Analyse und Visualisierung mit CATMA

Jan Horstmann [1]

1. Universität Münster

forTEXT

Thema:	Manuelle Annotation	DOI:	10.48694/fortext.3752
Jahrgang:	1	Ausgabe:	4
Erscheinungsdatum:	2024-08-07	Erstveröffentlichung:	2019-11-29 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: Franz Kafkas *Erstes Leid* (1924)
- Methoden: Analyse, Visualisierung, halb-automatische Annotation
- Angewendetes Tool: CATMA
- Lernziele: quantitative Analyse von Text- und Annotationsdaten; Erstellen von Queries (vgl. Query) und interaktiven Visualisierungen; Anpassen der Visualisierungen; automatische Annotation ausgewählter Keywords
- 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 Kafkas Erzählung *Erstes Leid* analysieren und Analyseergebnisse visualisieren 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 die Analyse- und Visualisierungsfunktionen von CATMA (Schumacher 2024) erlernen. CATMA (Computer Assisted Text Markup (vgl. Markup Language) and Analysis) ist ein frei verfügbares, webbasiertes (vgl. Browser) Tool, das Ihnen ermöglicht, digitale bzw. digitalisierte Texte manuell zu annotieren (Jacke 2024a), analysieren und visualisieren (Horstmann und Stange 2024) - alleine oder auch kollaborativ (Jacke 2024b) im Team. Die Annotationskategorien können dabei frei (d. h. „undogmatisch“) festgelegt werden. Aus diesem Grund eignet sich CATMA insbesondere für literaturwissenschaftliche Projekte, in denen häufig der Untersuchungsgegenstand die Untersuchungskategorien bestimmt und vorbestimmte Analysekategorien nicht den nötigen Forschungsspielraum böten. Teilweise werden wir auf Ergebnissen der Lerneinheit Manuelle Annotation mit CATMA (Horstmann 2024) aufbauen und die dort erstellten Annotationen zur Distanz der Erzählinstanz zum Erzählten neben den Textdaten aus Kafkas Erzählung *Erstes Leid* (1924) analysieren und visualisieren. Die meisten der hier beschriebenen Funktionen (vgl. Feature) und Aufgaben können Sie jedoch auch ohne manuell gesetzte Annotationen ausführen.

2. Vorarbeiten

Um diese Lerneinheit erfolgreich bearbeiten zu können, müssen Sie aus der vorangegangenen Lerneinheit Manuelle Annotation mit CATMA (Horstmann 2024) mindestens den Abschnitt „2. Vorarbeiten“ (bis einschließlich Absatz 24) durchlaufen haben. Dort lernen Sie, wie Sie sich einen CATMA-Account anlegen, ein Projekt erstellen und den zu erforschenden Text hochladen. Diese Vorarbeiten sind auch für die vorliegende Lerneinheit obligatorisch. Der Abschnitt „3. Funktionen“ aus der Lerneinheit zur manuellen Annotation ist hingegen fakultativ. Haben Sie ihn nicht vorab bearbeitet, können Sie die vorliegende Lerneinheit im Großen und Ganzen dennoch durchlaufen.

Um die Funktionen des *Analyze*-Moduls in CATMA bedienen zu können, gehen Sie auf https://catma.de12, klicken auf die Schaltfläche „Work with CATMA 6“ und loggen sich mit Ihrem Account ein. Auf der als nächstes angezeigten Home-Seite (vgl. Abb. 1) sollten Sie das in der vorherigen Lerneinheit angelegte Projekt wiederfinden (dort haben wir es „forTEXT-Lerneinheit“ genannt). Betreten Sie das Projekt durch einen Klick auf die entsprechende Kachel.

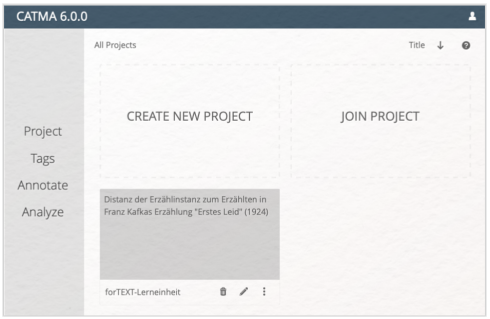


Abb. 1: Projekte im Home-Bereich von CATMA

3. Funktionen

CATMA unterstützt hermeneutische Textarbeit, indem es Textanalyse und Textannotation in einen zyklischen, iterativen Prozess integriert. In der Regel werden Sie Ihre Arbeit im *Annotate*-Modul mit der Nutzung des *Analyze*-Moduls kombinieren. Während Sie beim Annotieren ein Close Reading (vgl. Close Reading) praktizieren, unterstützt das *Analyze*-Modul auch einen Distant Reading-Ansatz: Hier können Sie Ihre Texte durchsuchen, sogar ohne sie zu lesen.

Vom *Projects*-Modul aus gibt es zwei Wege zum *Analyze*-Modul:

- Sie können zunächst den Text per Doppelklick öffnen und anschließend im *Annotate*-Modul auf die Schaltfläche „ANALYZE“ unterhalb des Textes klicken (so analysieren Sie den Text zusammen mit den dazugehörigen Annotation Collections).
- Alternativ können Sie im linken Navigationsbereich direkt auf „Analyze“ klicken. Dort müssen Sie dann in der sich öffnenden Lasche (vgl. Abb. 2) die **Texte und Annotation Collections auswählen**, die Sie analysieren möchten. In dieser Lerneinheit wählen Sie in der Lasche sowohl den Text als auch Ihre Annotation Collection aus (sämtliche Inhalte können insgesamt auch per Klick in das obere Kästchen neben „Name“ ausgewählt werden).

Abb. 2: Auswahl von Texten und Annotation Collections im Analyze*-Modul von CATMA*

Was bedeutet es, Texte und Annotation Collections für die Analyse auszuwählen? Im *Analyze*-Modul werden quantitative Operationen vorgenommen. Diese können sich entweder ausschließlich auf Textdaten (wie etwa Worthäufigkeiten oder -verteilungen) oder auf die von Ihnen erstellten Annotationsdaten (wie etwa Taghäufigkeiten (vgl. Tagset) oder -verteilungen) beziehen. CATMA bietet außerdem die Möglichkeit, komplexere Abfragen zu gestalten, die etwa Text- und Annotationsdaten in Kombination betreffen können.
Wann immer Sie im *Annotate*-Modul eine oder mehrere Annotationen hinzufügen oder löschen, verändern Sie die Annotationsdaten, sodass Sie im *Analyze*-Modul potentiell andere Ergebnisse erhalten. Ein erneuter Klick auf die „ANALYZE“-Schaltfläche im *Annotate*-Modul wird daher einen neuen Tab im *Analyze*-Modul öffnen, um die unterschiedlichen Datengrundlagen voneinander getrennt zu halten. Die einzelnen Analysen sind mit Zeitmarkierungen versehen, sodass Sie ihre chronologische Abfolge im Auge behalten können (vgl. Abb. 3). Ähnlich können Sie über das kleine Plus rechts oben im *Analyze*-Modul einen neuen Query-Tab öffnen und bspw. weitere Annotation Collections in die Analyse integrieren.

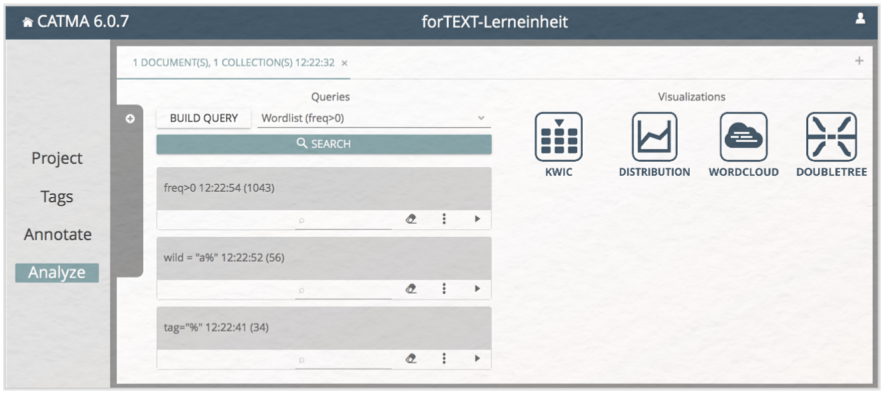


Abb. 3: Das Analyze*-Modul in CATMA mit mehreren geöffneten Analysen*

Wenn Sie sich die Gliederung des *Analyze*-Moduls anschauen, sehen Sie, dass es generell in zwei Bereiche aufgeteilt ist: „**Queries**“ auf der linken und „**Visualizations**“ auf der rechten Seite. Eine Query bestimmt, welche Informationen aus den Text- und/oder Annotationsdaten abgefragt werden soll. Die Ergebnisse der Abfrage können Sie in verschiedenen interaktiven Visualisierungen betrachten, explorieren und weiterverarbeiten. Die Textvisualisierung (Horstmann und Stange 2024) wird damit als integraler Bestandteil der Textanalyse verstanden.
Für die einfachsten Abfragen bietet CATMA Ihnen ein paar voreingestellte Queries an. Klicken Sie dafür auf den kleinen nach unten zeigenden Pfeil neben „Select or enter a free query“ und wählen die obere Option „Wordlist (freq>0)“ aus (vgl. Abb. 4). Durch diese Aktion erhalten Sie eine nach Häufigkeit geordnete Liste aller im Text vorkommenden Wörter (vgl. Abb. 5). Übrigens: Der in Klammern hinter „Wordlist“ angegebene technische Ausdruck „freq>0“ ist die Query-Language-Variante und besagt, dass alle Wörter mit einer Frequenz über null angezeigt werden sollen.

Abb. 5: Die Wortliste zeigt alle nach Häufigkeit geordneten Wörter in Kafkas Erstes Leid

Aufgabe 1: Schauen Sie sich die erstellte Wortliste genauer an. Wie viele Wörter enthält Kafkas Erzählung? Welches sind die drei am häufigsten vorkommenden Inhaltswörter (d. h. Wörter mit „mehr“ Bedeutung als Funktionswörter wie Artikel, Pronomen etc.)? Was fällt Ihnen außerdem an den angegebenen „Wörtern“ auf? Und wie müsste eine Abfrage lauten, die alle Wörter des Textes ausgibt, die mehr als fünfmal vorkommen?

Wenn Sie eine weitere Query eingegeben haben, um nach allen Wörtern zu suchen, die mehr als fünfmal vorkommen, wird Ihnen aufgefallen sein, dass das Ergebnis zu dieser Abfrage oberhalb der zuvor erstellten Wortliste angezeigt wird, die ihrerseits nicht verschwindet. Es ist generell möglich, mehrere Queries nacheinander zu stellen und die **Ergebnislisten** in dieser Form zu **sammeln**. Jedes Query-Ergebnis hat ein kleines Suchfeld in der Kopfzeile, um die entsprechende Liste schnell zu durchsuchen (vgl. Abb. 6). Durch einen Klick auf das kleine Radiergummi-Symbol können Sie eine Ergebnisliste wieder löschen. Das Dreipunktemenü rechts daneben enthält u. a. Exportmöglichkeiten für Ihre Abfragen in den Formaten CSV, XLSX oder JSON. Das Pfeilsymbol ermöglicht, die Ergebnislisten ein- und auszuklappen, um bei mehreren Abfragen einen besseren Überblick zu behalten.

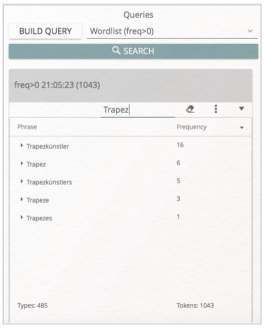


Abb. 6: Suchzeile und Symbole in CATMAs Query-Ergebnislisten

Worthäufigkeiten lassen sich ebenfalls gut visuell in Form der Wordcloud explorieren. Dies ist eine der derzeit in CATMA implementierten Möglichkeiten zur **Visualisierung** von Query-Ergebnissen. Um eine Visualisierung zu erstellen, führen Sie eine Query aus (wie etwa freq>0) und klicken anschließend auf die gewünschte Schaltfläche auf der rechten Seite; in diesem Fall „Wordcloud".
Die **Wordcloud**-Ansicht öffnet sich und auf der linken Seite sehen Sie wiederum die Query-Ergebnisliste. Aus dieser Liste können Sie nun diejenigen Wörter auswählen, die in der Wordcloud angezeigt werden sollen, indem Sie in der Spalte „Select" die nach unten zeigenden Pfeile bedienen. Auf diese Art stellen Sie sich unten manuell eine Liste derjenigen Wörter zusammen, die rechts in der erscheinenden Wordcloud angezeigt werden. Tipp: Wählen Sie im Dreipunktemenü der Query-Ergebnisliste die Option „Select all", um sämtliche Wörter in die Wordcloud zu integrieren. Sie erhalten eine Wordcloud, in der die Größe der Wörter ihre Häufigkeit im Text repräsentiert (vgl. Abb. 7). Auch in der zweiten so erzeugten Liste können Sie einzelne Werte (wie etwa Satzzeichen, Zahlen, oder Wörter, die klar dem Paratext zugeordnet werden können) mithilfe des Radiergummi-Symbols wieder löschen, sodass sie auch aus der Visualisierung verschwinden.

Abb. 7: CATMA-Wordcloud für die Häufigkeit aller Wörter in Kafkas Erstes Leid

Unter der Wordcloud haben Sie mehrere Möglichkeiten, Ihre Visualisierung per Slider zu manipulieren. So können Sie die Anzahl der Wörter, die Größe der Wordcloud wie der enthaltenen Wörter und auch den Abstand der Wörter („Word padding") zueinander verändern.
Exkurs: Alle Visualisierungen in CATMA weisen ein blasses Dreipunktemenü in der oberen rechten Ecke auf, das Ihnen ermöglicht, Ihre Visualisierung als Bild- (PNG) oder Vektordatei (SVG) herunterzuladen. Sie können sich hier außerdem den sog. Source Code der jeweiligen Visualisierung anschauen. CATMA-Visualisierungen basieren auf der **Visualisierungssprache Vega**. Mit etwas Einarbeitung ermöglicht Vega Ihnen, jede beliebige (auch interaktive) Visualisierung passend zu Ihren Daten zu erstellen. Sie öffnen dazu den Vega-Editor, erstellen potentiell eine ganz neue Visualisierung auf Basis der ausgewählten Query-Ergebnisse und laden die neue Visualisierung zurück in Ihr CATMA-Projekt. Diese Expertenfunktionen werden wir in der vorliegenden einführenden Lerneinheit jedoch überspringen. Etwas einfacher ist es, auf die kleine Schaltfläche rechts neben dem Dreipunktemenü zu klicken. Der Visualisierungs-Code erscheint rechts in einer Spalte: Manipulationen in diesem Code verändern (nach einem Klick auf die Kreispfeile oben) die angezeigte Visualisierung. Verändern Sie doch einfach einmal die Schriftart (engl. „font") wie wir es in Abb. 8 beispielhaft getan haben.

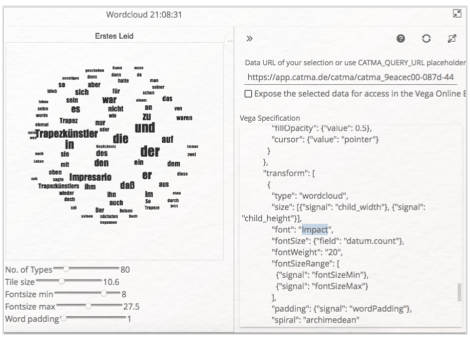


Abb. 8: Ausschnitt aus dem Vega-Code für die CATMA-Wordcloud mit veränderter Schriftart

Um eine Visualisierung zu schließen und zum *Analyze*-Modul zurückzukehren, klicken Sie auf die Schaltfläche mit den beiden zueinander zeigenden Pfeilen ganz rechts oben. Die Visualisierung wird minimiert und ähnlich wie die einzelnen Query-Ergebnisse im Visualisierungsbereich des *Analyze*-Moduls gesammelt (vgl. Abb. 9). Von hier können Sie einzelne Visualisierungen stets wieder aufrufen.

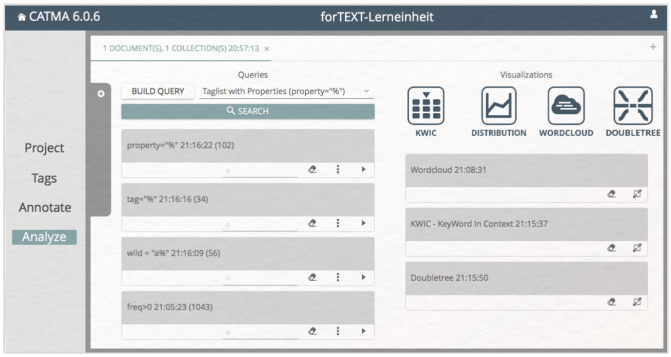


Abb. 9: Das Analyze*-Modul mit gesammelten Queries und mehreren Visualisierungen*

Wenden wir uns nun wieder den Queries zu. Unter den vorgeschlagenen Queries finden Sie neben der Wortliste auch eine Tagliste (optional inkl. Properties) und die Option „Wildcard“. Mit der Wildcard-Option können Sie bspw. Wortanfänge suchen. Wildcard steht für Platzhalter (das sind in CATMAs Query Language (vgl. Reguläre Ausdrücke) die „%“-Symbole (vgl. Reguläre Ausdrücke)). Die vorgeschlagene Wildcard-Abfrage „a%“ wird Ihnen folglich alle Wörter anzeigen, die mit einem kleinen „a“ beginnen.

Aufgabe 2: Wie viele Wörter im Kafka-Text beginnen mit dem Buchstaben „a“, wie viele mit „b“?

Um weitere Abfragen zu erstellen, bietet CATMA die „**BUILD-QUERY**“-Funktion, die Sie mithilfe normalsprachlicher Fragen dabei unterstützt, Ihren Abfragebefehl zusammenzustellen. Nach jedem Durchgang durch „BUILD QUERY“ erscheint die erstellte Query oben im *Analyze*-Modul. Die Form dieser Queries zu beachten, ermöglicht Ihnen, im Handumdrehen CATMAs **Query Language** zu erlernen. Einen systematischen Überblick über CATMAs Query Language finden Sie zudem auf dieser Seite.
In dieser Lerneinheit werden wir nun die BUILD-QUERY-Funktion benutzen. Klicken Sie auf die entsprechende Schaltfläche, werden Sie zunächst gefragt, wonach Sie suchen wollen (vgl. Abb. 10): nach Wörtern oder Formulierungen, nach Ähnlichkeiten, nach Tags, nach Kollokationen (vgl. Kollokation) oder nach Häufigkeiten. Wählen Sie zunächst „by frequency“ aus und klicken anschließend auf „CONTINUE“.

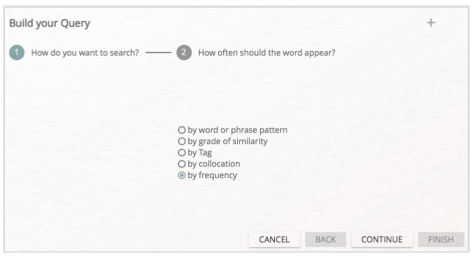


Abb. 10: Die BUILD-QUERY-Funktion in CATMA: Wonach soll gesucht werden?

Im nächsten Fenster können Sie festlegen, welche Wörter angezeigt werden sollen: alle Wörter, die genau x-mal im Text erscheinen („exactly"), alle Wörter, die mehr oder weniger als x-mal („more than" und „less than"), mehr bzw. genau x-mal („more or equal than") oder weniger bzw. genau x-mal („less or equal than"), oder auch zwischen x- und y-mal („between") vorkommen (vgl. Abb. 11). Die Zahlenwerte bestimmen Sie selbst. Unten im Fenster sehen Sie bereits, wie die jeweilige Query aussehen müsste.

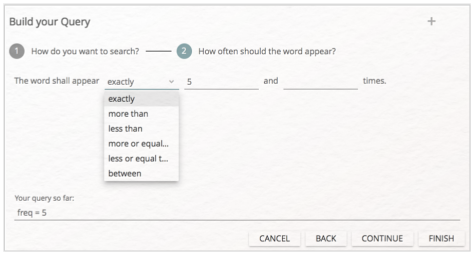


Abb. 11: Die BUILD-QUERY-Funktion in CATMA: Suche nach Häufigkeiten

Aufgabe 3: Wie viele Wörter erscheinen in Kafkas Erzählung fünfmal oder häufiger, aber weniger als 16 Mal? Wie sieht die entsprechende Query aus?

Auf diese Art und Weise können Sie jetzt eine Vielzahl an Queries bauen, ohne der Query Language mächtig zu sein (möchten Sie allerdings einen Text oder ein Textkorpus (vgl. Korpus) mit vielen Abfragen analysieren, empfiehlt es sich, diese recht einfache Sprache zu erlernen - Sie sparen dann bei den Abfragen schnell viel Zeit. Nicht alle der komplexeren Queries können außerdem mithilfe von BUILD QUERY erstellt werden). Klicken Sie sich nun selbstständig durch die anderen BUILD-QUERY-Funktionen (außer „by Tag") und versuchen Sie, die folgenden Fragen zu beantworten.

Aufgabe 4:Welche Wörter in Kafkas Erzählung haben eine 60-prozentige Ähnlichkeit mit dem Wort „Kunst"? Wie viele sind es bei 65 % und 70 %? Wie häufig kommt das Wort „Kunst" in der Nachbarschaft von zehn Wörtern des Wortes „oben" vor? Und in der Nähe des Wortes „unten"? Wie sehen die entsprechenden Queries aus?

Für die Untersuchung von Wortkontexten bietet CATMA zwei weitere geeignete Visualisierungsmöglichkeiten: die KWIC-Anzeige (*keyword in context*) (vgl. KWIC) und den DoubleTree. Beide Visualisierungen können ausgewählte Wörter in ihren jeweiligen Kontexten anzeigen; **KWIC** macht das in Form einer Tabelle, die jeweils links und rechts des ausgewählten Wortes fünf weitere Wörter anzeigt. Diese kontextualisierende Exploration ausgewählter Wörter ermöglicht einen punktuellen und effizienten Einblick in semantische Konstellationen. Um sich ein Keyword in seinem Kontext anzeigen zu lassen, erstellen Sie sich zunächst wieder eine Wortliste und klicken dann rechts auf das „KWIC"-Symbol; Sie landen in der Anzeige für die KWIC-Visualisierung. Hier können Sie nach bekanntem Schema aus der Liste oben links Wörter auswählen und somit eine weitere Liste ausgewählter Keywords unten links zusammenstellen. Auf der rechten Seite erscheint das jeweilige Keyword als Kontextliste (vgl. Abb. 12). Weitere ausgewählte Keywords werden dieser Liste hinzugefügt; entfernen können Sie Keywords durch einen Klick auf das Radiergummi-Symbol in der Liste unten links. Sie können die einzelnen Spalten der Keywordliste mit der Maus vergrößern oder verkleinern (das gilt generell für alle Panels in CATMA) oder beispielsweise auch textchronologisch ordnen, indem Sie in der Kopfzeile auf „Start Point" klicken. Die Zahlen bei „Start Point" und „End Point" sind die sog. *character offsets*, d. h. die Nummer, die der erste Buchstabe des ausgewählten Wortes im durchnummerierten Text bekommen hat, und die des letzten Buchstabens. Bei einem Keyword mit fünf Buchstaben wird daher die Zahl bei „End Point" immer genau fünf Werte über der Zahl bei „Start Point" liegen.

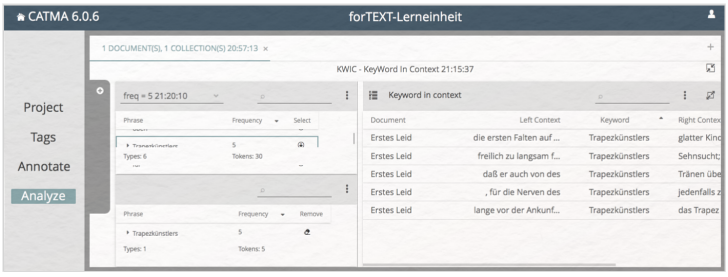


Abb. 12: Die KWIC-Liste in CATMA mit ausgewähltem Keyword

Wenn Sie beim Durchschauen der KWIC-Liste eine Stelle finden, die Ihnen interessant erscheint oder an der Sie mehr Kontext brauchen, können Sie von dieser Stelle übrigens immer direkt an die entsprechende Textstelle im

Annotate-Modul springen, indem Sie einfach doppelt in die Zeile der KWIC-Liste klicken, die Sie interessiert. Das Keyword (Token) (vgl. Type/Token) wird zweifarbig unterstrichen hervorgehoben. Ein Klick auf „Analyze“ im Navigationsbereich links bringt Sie zurück zur Liste.

Aufgabe 5: Analysieren Sie das Wort „Leid“ in seinem Kontext mithilfe der KWIC-Visualisierung. Was fällt Ihnen auf?

Eine interaktivere Form der Kontextexploration ermöglicht die **DoubleTree**-Visualisierung. Schließen Sie die KWIC-Liste durch einen Klick auf das Pfeilsymbol oben rechts und klicken anschließend auf die „DOUBLETREE“-Schaltfläche. Die Erzeugung der Visualisierung funktioniert nach bekanntem Schema: Links wählen Sie das gewünschte Keyword aus, das dann rechts als DoubleTree visualisiert wird (vgl. Abb. 13). Der Unterschied ist, dass jeweils immer nur ein Keyword betrachtet wird, das in der Liste unten links per Klick gewechselt werden kann.

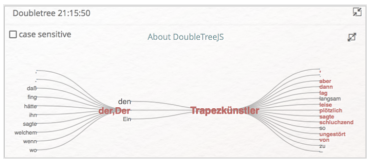


Abb. 13: Die DoubleTree-Visualisierung in CATMA

Sie sehen das Keyword im DoubleTree einmalig und rot hervorgehoben in der Mitte und die Kontexte werden als Worttypen links und rechts davon angezeigt. Wenn ein Wort häufiger im Kontext des Keywords auftaucht, ist es größer gestaltet, als wenn es nur einmalig in diesem Kontext vorkommt. Klicken Sie auf eines dieser größeren Wörter, entfaltet sich der weitere Kontext (Sie können den DoubleTree übrigens mit der Maus frei auf der Fläche bewegen). Das ausgewählte Wort wird zudem rot und auf der gegenüberliegenden Kontextseite werden diejenigen Wörter ebenfalls rot, die im Textzusammenhang mit diesem Wort vorkommen, sodass Sie die Satzstruktur nachvollziehen können. Die DoubleTree-Visualisierung lässt somit einen strukturierten Überblick über die Art und Weise zu, wie in einem spezifischen Text Vokabular zum Einsatz kommt, und bietet dadurch beispielsweise die Möglichkeit, Rückschlüsse auf Figurencharakterisierungen zu ziehen. Für diese Visualisierung lässt sich die in CATMA übliche „case sensitivity“ übrigens ausschalten, d. h. groß- und kleingeschriebene Wortvarianten können zusammengefasst werden.

Aufgabe 6: Explorieren Sie die Kontexte der Wörter „Trapezkünstler“ und „Impresario“ mithilfe der DoubleTree-Visualisierung. Können Sie etwas über die Charakterisierung der beiden Figuren sagen?

Wie lassen sich nun die in der vorherigen Lerneinheit erstellten Annotationen im *Analyze*-Modul untersuchen? (Wenn Sie die Lerneinheit Manuelle Annotation mit CATMA (Horstmann 2024) nicht durchlaufen haben sollten, können Sie hier zum Absatz nach Abbildung 15 springen.) Schließen Sie die DoubleTree-Visualisierung. Die gesamte Liste all Ihrer im Annotationsprozess vergebenen Tags können Sie mithilfe des vorgeschlagenen Querys tag=„%“ generieren. Ihnen werden zunächst die annotierten Phrasen angezeigt. Möchten Sie die Ergebnisse stattdessen nach Tag-Kategorie sortieren, wählen Sie im Dreipunktemenü „Group by Tag Path“. Im Ergebnis sehen Sie, welcher Tag wie häufig vergeben wurde (vgl. Abb. 14).

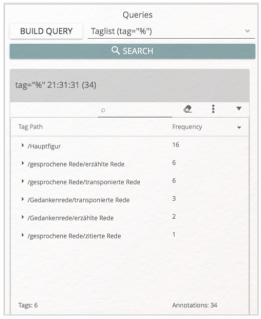


Abb. 14: Tag-Liste in CATMAs Analyze*-Modul*

Sie können zusätzlich die spezifischen Annotationen und Properties durch die Optionen „Display Annotations as flat table“ bzw. „Display Properties as columns“ anzeigen lassen. Fahren Sie anschließend mit dem Mauszeiger über die einzelnen Annotationen, wird Ihnen noch etwas mehr Kontext angezeigt (vgl. Abb. 15).

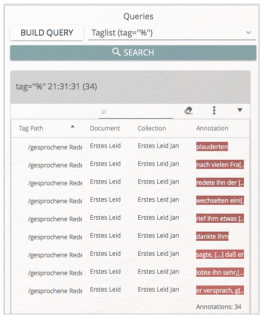


Abb. 15: Tag-Liste in CATMAs Analyze*-Modul mit Annotationen*

Sie haben die Möglichkeit, Verteilungskurven für Wörter oder Tags zu generieren - die vierte Visualisierungsmöglichkeit in CATMA. Diese sog. **Distribution Charts** werden auf die gleiche Art und Weise erstellt wie die anderen Visualisierungen, d. h. durch die Überführung von einzelnen Zeilen aus dem Query-Ergebnis in eine visualisierungsspezifische zweite Liste, die wiederum zur Bearbeitung der Visualisierung manipuliert werden kann. Wir wollen in diesem Fall die Verteilung für vergebene Tags ausprobieren. (Sollten Sie keine Annotationen gesetzt haben, erstellen Sie einfach eine Wortliste und wählen einzelne Wörter für die Distribution-Visualisierung aus.)
Erstellen Sie sich eine Tagliste mit der vorgegebenen Query und wählen im Dreipunktemenü die Option „Group by Tag Path“. Klicken Sie anschließend auf die Schaltfläche „DISTRIBUTION“ auf der rechten Seite. Im Dreipunktemenü wählen Sie nun wie schon zuvor „Select all“, um sämtliche Zeilen des Queryergebnisses in die Visualisierungsliste zu übertragen. Rechts erscheint ein Koordinatensystem, in dem die Verteilungen der von Ihnen vergebenen Tags angezeigt werden (vgl. Abb. 16). Die x-Achse bezieht sich auf die gesamte Textlänge, die hier in 10%-Segmente unterteilt wurde (anhand der Buchstabenanzahl). Auf der y-Achse erscheint die Anzahl der Vorkommnisse. Über den Zoom-Slider können Sie die gesamte Visualisierung vergrößern.

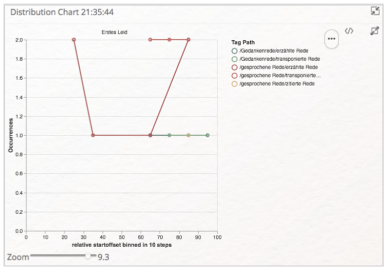


Abb. 16: Die Distribution-Visualisierung für vergebene Tags in CATMA

Aufgabe 7: Visualisieren Sie Distributionsgraphen für die von Ihnen vergebenen Tags. Was finden Sie auffällig? Wofür eignet sich diese Form der Visualisierung? In welcher Hinsicht wäre eine andere Visualisierung sinnvoller?

Sie haben auch die Möglichkeit, mit „BUILD QUERY“ **nach einzelnen Tags zu suchen**. Dort wählen Sie die Option „by Tag“. Nach einem Klick auf „CONTINUE“ erscheint das Tagset, aus dem der gewünschte Tag ausgewählt werden kann. Eine entsprechende Query sähe beispielsweise so aus: tag=„/gesprochene Rede/erzählte Rede%“. Die Query tag=„/gesprochene Rede%“ hingegen sucht nach allen Annotationen der drei dem Bereich „gesprochene Rede“ zugeordneten Tags (erzählte Rede, transponierte Rede und zitierte Rede).
Es ist außerdem möglich, **komplexere Abfragen** zu generieren, indem Sie am Ende des Build-Query-Dialogs nicht auf „FINISH“ sondern auf „CONTINUE“ klicken. Sie werden dann gefragt, ob Sie weitere Query-Ergebnisse hinzufügen wollen („add more results“), eine Teilmenge der vorherigen Ergebnisse aus den neuen Ergebnissen ausschließen wollen („exclude previous results“) oder die vorherigen Ergebnisse verfeinern möchten („refine previous results“). In den beiden letzten Fällen müssen Sie zudem bestimmen, in welcher Art vorherige und weitere Ergebnisse zueinander in Beziehung stehen sollen. Es gibt die Optionen „exact match“, „boundary match“ und „overlap match“ (vgl. Abb. 17). Die Option „exact match“ zeigt Ergebnisse, die exakt an derselben Stelle vorkommen (d. h. beispielsweise zwei Annotationen). Mit „boundary match“ können Sie Stellen suchen, die einander beinhalten (d. h. etwa alle Vorkommnisse eines Tags innerhalb einer anderen Tagkategorie). Die dritte Option („overlap match“) schließlich ermöglicht, Überlappungen anzuzeigen (wie verschiedene Annotationen, die einander überlappen können, oder auch überlappende Wörter/Phrasen und Annotationen).

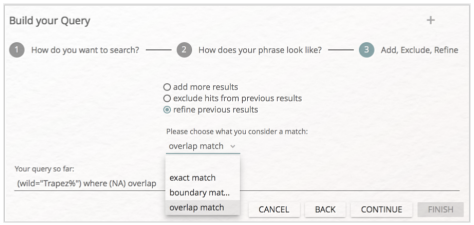


Abb. 17: Erstellen einer komplexen Query in CATMA

Aufgabe 8: Lassen Sie sich alle Vorkommnisse anzeigen, in denen Wörter, die mit „Trapez“ beginnen, mit den Tags der gesprochenen Rede überlappen. Wie sieht die entsprechende Query aus? Wie viele Ergebnisse bekommen Sie?

Die letzte Funktion des *Analyze*-Moduls, der wir uns in dieser Lerneinheit nähern wollen, ist die des **halb-automatischen Annotierens**. Ihnen wird vielleicht schon per Zufall aufgefallen sein, dass auch die Wordcloud- und Distribution-Chart-Visualisierungen, die Sie in CATMA generiert haben, interaktiv sind, d. h. Sie können auf einzelne Stellen in den Visualisierungen klicken und erhalten weitere Optionen. Wir wollen dies beispielhaft anhand einer einfachen Wordcloud-Visualisierung aller im Text enthaltenen Wörter ausprobieren.
Öffnen Sie erneut die zuvor erstellte Wordcloud-Visualisierung und fügen (sollte das nicht der Fall sein) sämtliche Wörter hinzu. Sollten Sie zwischenzeitlich den entsprechenden *Analyze*-Tab geschlossen haben, erstellen Sie sich eine neue Wortliste, klicken anschließend auf „Wordcloud“, dann auf „Select all“ und erhöhen die Anzahl der gezeigten Wörter (Types). Wenn Sie nun auf einzelne Wörter in der Wordcloud klicken, öffnet sich eine KWIC-Liste unterhalb der Visualisierung (vgl. Abb. 18). Tun sie dies mit den Wörtern „Trapezkünstler“ und „Trapezkünstlers“.

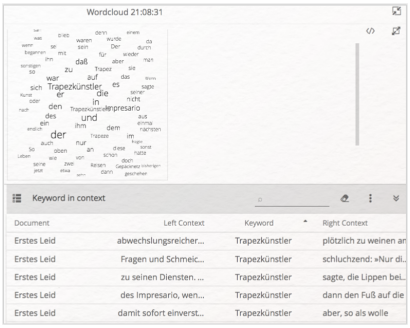


Abb. 18: Wordcloud mit ausgewählten Keywords

Mit KWIC-Listen lassen sich in CATMA generell halb-automatisch Annotationen für die ausgewählten Keywords erstellen. Dafür klicken Sie auf das Selektionssymbol links von der Überschrift „Keyword in context“ und wählen alle Zeilen durch einen Klick in das oberste Kästchen neben „Document“ aus. Alternativ können Sie einzelne Zeilen der Liste auswählen. Sind so sämtliche Zeilen ausgewählt, finden Sie in im Dreipunktemenü die Option „Annotate selected rows“, die Sie anschließend anklicken.
Es öffnet sich ein Dialogfenster (vgl. Abb. 19), das Ihnen ermöglicht, existierende Tags auszuwählen, oder auch neue Tagsets bzw. Tags zu erstellen, die für die halb-automatische Annotation genutzt werden sollen. Erstellen Sie hier nun ein Tagset „Figuren“ und in diesem Tagset den Tag „Hauptfigur“ (vgl. Abb. 20). Wählen Sie in der Tagliste nun diesen neuen Tag aus und klicken auf „CONTINUE“.

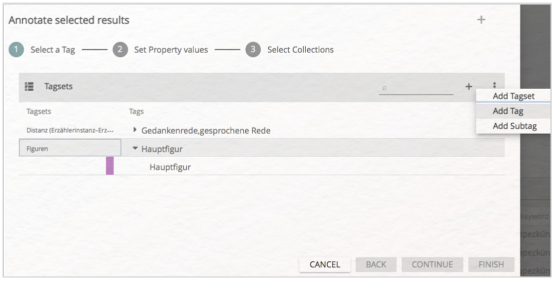
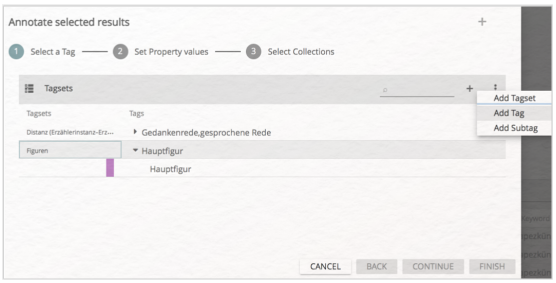


Abb. 19: Dialogfenster zum halb-automatischen Annotieren in CATMA



In einem letzten Schritt werden Sie gefragt, in welcher Annotation Collection die neu zu erzeugenden Annotationen gespeichert werden sollen. Hier wählen Sie ihre eigene Annotation Collection aus (vgl. Abb. 21) und klicken anschließend auf „FINISH“. Die neuen Annotationen wurden gesetzt.

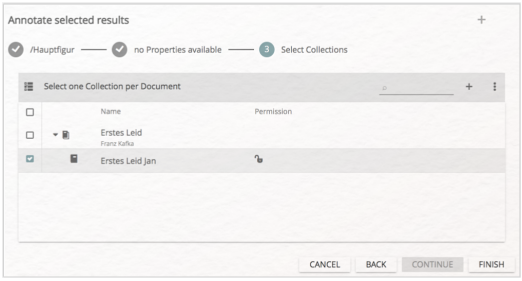


Abb. 21: Auswahl der Annotation Collection zur halb-automatischen Generierung von Annotationen

Sie haben nun alle Keywords der Liste mit dem Tag „Hauptfigur“ annotiert. Ein Doppelklick in die Liste wird Sie zurück in das *Annotate*-Modul bringen (vgl. Abb. 22), in dem Sie die neuen Annotationen sehen können (evtl. müssen Sie dafür noch einmal auf das Augensymbol neben dem entsprechenden Tagset auf der rechten Seite klicken, sollte es durchgestrichen sein).

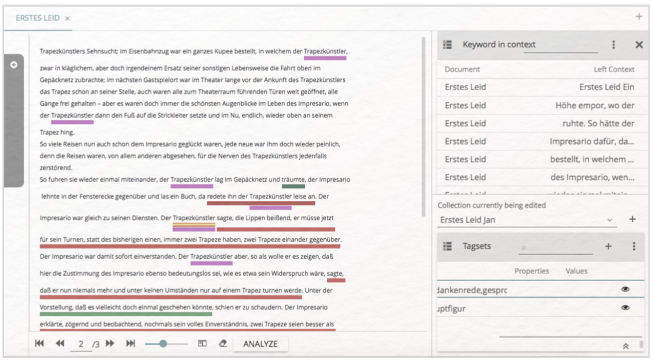


Abb. 22: Kafkas Erstes Leid *mit halb-automatisch generierten Annotation der Hauptfigur*

Aufgabe 9: Für welche Annotationsaufgaben bietet sich die Funktion des halb-automatischen Annotierens besonders an? In welchen Fällen ist Vorsicht geraten?

Sie haben nun sämtliche Grundfunktionen des *Analyze*-Moduls in CATMA kennengelernt: Sie können einfache und komplexe Queries erstellen, Listen für Visualisierungen erzeugen und manipulieren, Visualisierungen verändern und KWIC-Listen für die halb-automatische Annotation generieren, um Ihre Annotationsdaten anzureichern.

4. Lösungen zu den Beispielaufgaben

Aufgabe 1: Schauen Sie sich die erstellte Wortliste genauer an. Wie viele Wörter enthält Kafkas Erzählung? Welches sind die drei am häufigsten vorkommenden Inhaltswörter (d. h. Wörter mit „mehr“ Bedeutung als

Funktionswörter wie Artikel, Pronomen etc.)? Was fällt Ihnen außerdem an den angegebenen „Wörtern“ auf? Und wie müsste eine Abfrage lauten, die alle Wörter des Textes ausgibt, die mehr als fünfmal vorkommen?

Query-Ergebnisse werden immer auch in der grauen Spalte über der Liste in Klammern angezeigt. Die Erzählung *Erstes Leid* enthält demnach 1337 Wortvorkommnisse (Tokens) und 585 Worttypen (Types), die am Ende der Liste angezeigt werden. Aber Achtung: Hierbei werden die Wörter aus dem Paratext mitgezählt, der Informationen über die Edition, den Text, seinen Autor und den Herausgeber enthält. Die Wörter „Franz“ und „Kafka“ kommen daher innerhalb der Erzählung *Erstes Leid* nicht wie angegeben neunmal vor. Behalten Sie bei der Analyse von Worthäufigkeiten daher immer im Blick, wie das zugrunde liegende Dokument gestaltet ist.
Das häufigste Inhaltswort ist „Trapezkünstler“ (16 Mal), gefolgt von „war“ (13 Mal) und „Impresario“ (12 Mal). Bereits aus diesem schnellen Blick in die Wortliste ließe sich auch ohne Textkenntnis auf das Milieu der Erzählung, ihre Hauptfiguren sowie das Verhältnis von Erzählzeit und erzählter Zeit (späteres Erzählen) schließen.
Auch Satzzeichen werden in CATMA als „Wörter“ gezählt (wenn Sie unter dem Dreipunktemenü „Filter punctuation“ deaktivieren, können Sie sich diese auch anzeigen lassen). Der Grund dafür ist, dass ein Wort im Grunde durch die Leerzeichen vor und hinter dem Wort definiert werden könnte. Satzzeichen (die immer ohne Leerzeichen neben einem Wort stehen) würden bei dieser Regel zu den Wörtern gezählt (d. h. „interpretieren“ und „interpretieren,“ wären zwei unterschiedliche Wörter). Gleichzeitig möchte man auf die Angabe der Häufigkeit von Satzzeichen nicht verzichten, kann es doch wertvolle Impulse für die Interpretation eines Textes geben, dass dieser bspw. mehr Frage- als Ausrufezeichen oder mehr Punkte als Kommas etc. enthält.
Schließlich könnte Ihnen noch aufgefallen sein, dass CATMA groß- und kleingeschriebene Wörter nicht zusammen zählt (z. B. kommt „der“ 29 Mal vor, „Der“ fünfmal) und auch Wortformen getrennt zählt (z. B. kommt „Trapezkünstler“ 16 Mal, „Trapezkünstlers“ fünfmal vor). Möchte man nun also herausfinden, wie häufig ein Wort in seiner Grundform genannt wird (genau genommen ein Lemma), sollte man alle Formen des Wortes addieren. Dazu eignet sich die Suchleiste in der Kopfzeile des Query-Ergebnisses. Die Abfrage für alle Wörter, die mehr als fünfmal im Text vorkommen, müsste analog zur Wordlist-Query lauten: freq>5.

Aufgabe 2: Wie viele Wörter im Kafka-Text beginnen mit dem Buchstaben „a“, wie viele mit „b“?

Es werden 28 Tokens angezeigt, die mit „a“ beginnen, und 23 mit „b“. Die zweite Query lautet: wild = „b%“.

Aufgabe 3: Wie viele Wörter erscheinen in Kafkas Erzählung fünfmal oder häufiger, aber weniger als 16 Mal? Wie sieht die entsprechende Query aus?

Es erscheinen in der Textvorlage 42 Wörter (Types) zwischen fünf und 15 Mal (auch hier werden die Wörter der Paratexte mitgezählt). Die Query für diese Abfrage lautet: freq = 5-15. CATMA zählt die Wörter, die fünfmal bzw. 15 Mal vorkommen, hierbei mit.

Aufgabe 4: Welche Wörter in Kafkas Erzählung haben eine 60-prozentige Ähnlichkeit mit dem Wort „Kunst“? Wie viele sind es bei 65 % und 70 %? Wie häufig kommt das Wort „Kunst“ in der Nachbarschaft von zehn Wörtern des Wortes „oben“ vor? Und in der Nähe des Wortes „unten“? Wie sehen die entsprechenden Queries aus?

Acht Wörter (Types), die insgesamt zwölfmal vorkommen (Tokens), haben eine 60-prozentige Ähnlichkeit mit dem Wort „Kunst“, darunter „unter“, „unten“, „sonst“ und „Ankunft“. Bei 65 Prozent sind es noch zwei („Kunst“ und „Ankunft“), bei 70 % bleibt nur noch „Kunst“. Die Query lautet: simil=„Kunst“ 60%. Es ist zeitsparender, in der Query-Leiste einfach die Prozentzahl anzupassen, als für jede Abfrage erneut die BUILD-QUERY-Funktion zu bemühen.
Das Wort „Kunst“ kommt einmal in der Nähe von „oben“ vor, keinmal jedoch in der Nähe von „unten“. Das passt zwar zum Kafka-Text; allgemein sind Kollokations-Abfragen bei diesem Text jedoch kaum erkenntnisfördernd, da die Ergebnismengen immer äußerst gering ausfallen - ein Indiz nicht nur für die Kürze des Textes, sondern auch für Kafkas variantenreiches Formulierungsvermögen. Die Query hierfür lautet: „Kunst“ & „oben“ 10.

Aufgabe 5: Analysieren Sie das Wort „Leid“ in seinem Kontext mithilfe der KWIC-Visualisierung. Was fällt Ihnen auf?

Das Wort „Leid“ kommt dreimal vor, ein Blick in die KWIC-Liste verrät jedoch unmittelbar, dass es sich ausschließlich um Vorkommnisse aus dem Paratext handelt. In Kafkas Erzählung *Erstes Leid* kommt das Wort „Leid“ damit kein einziges Mal vor.

Aufgabe 6: Explorieren Sie die Kontexte der Wörter „Trapezkünstler“ und „Impresario“ mithilfe der DoubleTree-Visualisierung. Können Sie etwas über die Charakterisierung der beiden Figuren sagen?

Bei beiden Wörtern handelt es sich nicht um Namen, sondern um Funktionsbeschreibungen. Aus diesem Grund birgt der rechte Kontext sehr viel mehr bedeutungstragende Information als der linke, der ausschließlich aus bestimmten oder unbestimmten Artikeln besteht. Der Trapezkünstler erscheint bereits in dieser überblicksartigen Perspektive als sensible, schwache, kränkelnde Figur, während dem Impresario unterstützende, organisierende und erklärende Funktionen zufallen.

Aufgabe 7: Visualisieren Sie Distributionsgraphen für die von Ihnen vergebenen Tags. Was finden Sie auffällig? Wofür eignet sich diese Form der Visualisierung? In welcher Hinsicht wäre eine andere Visualisierung sinnvoller?

Auffällig ist, dass aufgrund der sehr geringen Textlänge die 10%-Segmente entsprechend kurz ausfallen, was zur Folge hat, dass in jedem Segment höchstens zwei Annotationen einer Kategorie vorkommen. Die Distribution zeigt daher kaum mehr als die beim Annotieren bereits gewonnene Erkenntnis, dass Rede- und Gedankenwiedergabe insbesondere gegen Ende des Textes zunehmen und am Anfang nicht vorhanden sind. Distributionsgraphen werden i. d. R. interessanter, wenn Vorkommensverhältnisse (etwa zwischen Tags und Wörtern) visualisiert und untersucht werden sollen. Die Verbindungen zwischen den einzelnen Punkten der Distribution Chart sollten zudem kritisch betrachtet werden, suggerieren sie doch eine kontinuierliche Entwicklung, die in Texten so nicht gegeben ist, wenn voneinander getrennte Wörter die Datengrundlage bilden. Verteilungskurven bieten sich daher insbesondere für längere Texte an. Für eine bloße Darstellung der Häufigkeiten vergebener Tags ist die Wordcloud die geeignetere Visualisierungsoption, die in diesem Fall etwa so aussehen könnte:

Abb. 23: Tag-Cloud für Rede- und Gedankenwiedergabe in Kafkas Erstes Leid

Aufgabe 8: Lassen Sie sich alle Vorkommnisse anzeigen, in denen Wörter, die mit „Trapez“ beginnen, mit den Tags der gesprochenen Rede überlappen. Wie sieht die entsprechende Query aus? Wie viele Ergebnisse bekommen Sie?

Die entsprechende Query lautet: (wild=„Trapez%“) where (tag=„/gesprochene Rede%“) overlap. Die Ergebnismenge hängt etwas davon ab, welche Entscheidungen im Annotationsprozess gefällt wurden. In unserem Fall erhalten wir sechs Übereinstimmungen.

Aufgabe 9: Für welche Annotationsaufgaben bietet sich die Funktion des halb-automatischen Annotierens besonders an? In welchen Fällen ist Vorsicht geraten?

Insbesondere für die Annotation benannter Entitäten (wie in unserem Beispiel die Hauptfigur) bietet die Funktion des halb-automatischen Annotierens eine sehr zeitsparende Unterstützung. Aber auch Zeitformen von Verben könnten beispielsweise mit den entsprechenden Tags (etwa „Vergangenheit“, „Gegenwart“, „Zukunft“ oder auch „Präteritum“, „Perfekt“, „Präsens“, „Futur I“ etc.) recht zügig halb-automatisch annotiert werden. Bereits hier sollten Sie etwas Vorsicht walten lassen, gibt es doch Wörter, die je nach Kontext sowohl Verben in einer bestimmten Zeitform als auch andere Wortarten sein können (das gilt noch stärker für nicht-deutschsprachige Texte). Auch der Tag „Hauptfigur“ könnte zu Schwierigkeiten führen, wenn Sie beispielsweise eine Erzählung mit homodiegetischem Erzähler annotieren wollen, in der das „ich“ nicht in jedem Fall für den Erzähler, sondern in direkter Rede auch für jede andere Figur stehen könnte. Der Kontext sollte in der KWIC-Liste gerade beim halb-automatischen Annotieren daher unbedingt beachtet werden.

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)

Bibliographie

Horstmann, Jan. 2024. 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.

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.

Jacke, Janina. 2024b. Methodenbeitrag: Kollaboratives literaturwissenschaftliches Annotieren. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 4. Manuelle Annotation (7. August). doi: 10.48694/fortext.3749, https://fortext.net/routinen/methoden/kollaboratives-literaturwissenschaftliches-annotieren.

———. 2024a. 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: CATMA. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 4. Manuelle Annotation (7. August). doi: 10.48694/fortext.3761, https://fortext.net/tools/tools/catma.

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).

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

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

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.

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.

JSON JSON ist die englische Abkürzung für *JavaScript Object Notation*. Dabei handelt es sich um ein kompaktes Textformat, das insbesondere zum Datenaustausch entworfen wurde. Es ist für Menschen einfach zu lesen und zu schreiben und für Maschinen einfach zu analysieren und zu generieren. JSON ist ein Format, das unabhängig von Programmiersprachen ist.

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.

POS PoS steht für *Part of Speech* , oder „Wortart“ auf Deutsch. Das PoS- Tagging beschreibt die (automatische) Erfassung und Kennzeichnung von Wortarten in einem Text und ist of ein wichtiger Preprocessing-Schritt, beispielsweise für die Analyse von Named Entities.

Preprocessing Für viele digitale Methoden müssen die zu analysierenden Texte vorab „bereinigt“ oder „vorbereitet“ werden. Für statistische Zwecke werden Texte bspw. häufig in gleich große Segmente unterteilt (*chunking*), Großbuchstaben werden in Kleinbuchstaben verwandelt oder Wörter werden lemmatisiert.

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

Query *Query* bedeutet „Abfrage“ oder „Frage“ und bezeichnet eine computergestützte Abfrage zur Analyse eines Textes. Um Datenbestände zu durchsuchen, werden Abfragesprachen eingesetzt, die *Queries* (Anfragen) an den Datenbestand senden. So bilden alle möglichen Queries zusammen die *Query Language* eines Tools.

Reguläre Ausdrücke Reguläre Ausdrücke, auch *Regular Expressions* oder *RegEx* genannt, sind standardisierte Zeichenketten zur Beschreibung von Mengen von Zeichenketten mit Hilfe bestimmter syntaktischer Regeln, die in Abfrage- und Programmiersprachen (z.B. in Wort, CATMA, Python, R usw.) für unterschiedliche Problemlösungen verwendet werden. Sie können beispielsweise als Filterkriterien in der Textsuche oder in Texteditoren (z.B. in Word oder OpenOffice) zum „Suchen und Ersetzen“ von bestimmten Begriffen genutzt werden.

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.

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.

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.