

Einführung in CATMA

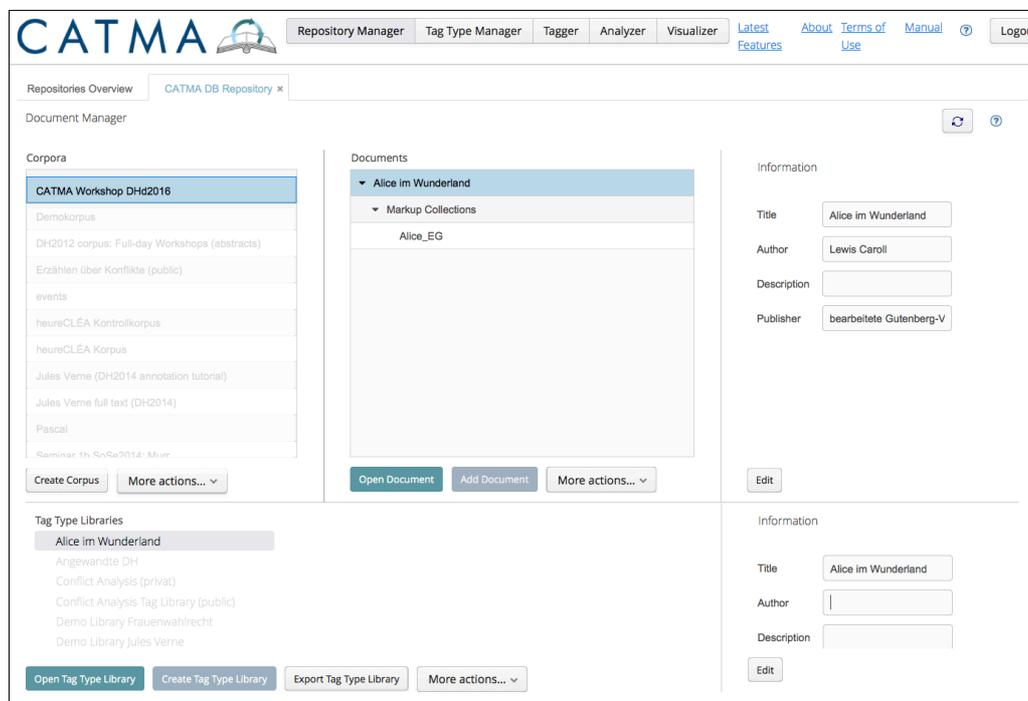
- I. Erste Schritte in CATMA
- II. Texte analysieren – Suchabfragen in CATMA
- III. Texte (kollaborativ) annotieren mit CATMA
- IV. Weitere Suchabfragen: Annotationen auswerten

I. Erste Schritte in CATMA

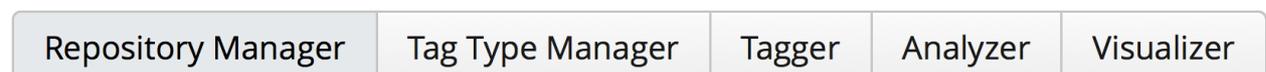
Einloggen unter www.digitalhumanities.it/catma

Im Repository Manager:

1. Anlegen eines Korpus (z.B. „CATMA Workshop“) über „Create Corpus“,
2. Anlegen von Dateien im Korpus (eigenen Dateien oder Texte aus dem DTA, Project Gutenberg etc.) über „Add Document“. Text für die Hands On-Session: Lewis Caroll „Alice im Wunderland“ (bearbeitete Version aus Project Gutenberg eBook),
3. Erzeugen einer Tag Type Library über "Create Tag Type Library". Wählen Sie einen geeigneten Namen für Ihre Library und speichern Sie sie. Anschließend können Sie Ihre Namen und eine Beschreibung ergänzen (Library auswählen und "Edit" klicken),
4. Erzeugen einer Markup Collection durch Auswählen des Textdokuments und Klicken von "More Actions -> Create Markup Collection".



In CATMA gibt es fünf Module, die Sie durch Anklicken des jeweiligen Reiters nutzen können:



II. Texte analysieren – Suchabfragen in CATMA

Wählen Sie im Repository Manager Modul das erstellte Korpus bzw. Textdokument aus und wählen Sie "More Actions -> Analyze Corpus" bzw. "More Actions -> Analyze Document".

Wortlisten

5. Klicken Sie auf den "Wordlist" Button.
6. Sortieren Sie die Wortliste nach abfallender Frequenz. Welches ist das häufigste Inhaltswort? (= ein Wort mit „mehr“ semantischer Bedeutung als Funktionswörter wie Artikel, Pronomen etc.)?

KWIC Visualisierung (KWIC = KeyWord In Context)

7. Für das gewählte Wort: Klicken sie auf das "Visible in KWIC" Kästchen und betrachten Sie die KWIC-Anzeige. Ändern Sie ggf. den Umfang des angezeigten Kontexts über den Schieber unter (Einstellung auf 1-30 Tokens möglich).
8. Doppelklicken Sie auf eines der Keywords in der KWIC-Anzeige, um zu seiner Position im Volltext zu springen.

Double Tree Visualisierung

9. Gehen Sie zurück zur Wortliste im Analyzermodule. Markieren Sie die Zeile des ersten (= häufigsten) Inhaltsworts und klicken Sie auf den DoubleTree-Button (2. Button links unten).
10. Klicken Sie auf die Wörter in der Double Tree-Visualisierung und versuchen Sie herauszufinden, wie der Double Tree funktioniert.

Distributionsgraph

11. Gehen Sie wieder zum Analyzermodule und markieren Sie erneut die Zeile des gewählten Worts und klicken Sie auf den Distributionsgraph-Button (1. links unten).
12. Gehen Sie zurück zum Analyzermodule und wählen Sie ein anderes Wort, das Ihnen interessant erscheint. Lassen Sie sich auch dieses im Distributionsgraphen anzeigen. Gibt es etwas Interessantes, das sich aus der Darstellung der beiden Wörter erkennen/vermuten lässt?
13. Gehen Sie zurück zum Analyzermodule und wählen Sie in der Wortliste eine Gruppe von Wörtern, die Ihnen interessant erscheint. Markieren Sie die entsprechenden Zeilen (durch Halten der command-Taste) und lassen Sie sich die Wortgruppe im Distributionsgraphen anzeigen.

Querybuilder

14. Gehen Sie wieder zum Analyzermodule und öffnen Sie den Querybuilder, indem Sie auf den entsprechenden Button klicken.
15. Suchen Sie mithilfe des Querybuilders nach einem der gewählten Wörter – oder mehreren

Wörtern, die mit denselben Buchstaben enden.¹

16. Suchen Sie mithilfe des Querybuilders nach allen Wörtern, die öfter als zehn Mal vorkommen.
17. Suchen Sie mithilfe des Querybuilders nach allen Wörtern mit 70% Ähnlichkeit zu „verwirrt“. Öffnen Sie einen neuen Analyzer-Tab (durch Klicken des Buttons „+ New Query“ rechts oben) und erhöhen Sie die Ähnlichkeit auf 80%. Machen Sie dasselbe nochmal mit 75%.

III. Texte (kollaborativ) annotieren mit CATMA

*Hinweis zum **Sharing***: Wenn Sie mit einem geteilten Korpus arbeiten, wird jedes Textdokument und jede Markup Collection, die Sie zum Korpus hinzufügen, *automatisch mit allen anderen geteilt*, die Zugriff auf das Korpus haben. Wenn Sie etwas aus dem Korpus löschen, ist es *nur für Sie nicht mehr sichtbar*, alle anderen sehen es weiterhin.

a. Was wollen Sie untersuchen? -> Tagerstellung (zu zweit oder dritt)

Für die Annotation brauchen Sie eine **Tag Type Library** – entweder eine bereits bestehende eine neue, die Sie erstellen. Tag Type Libraries können für andere Dokumente wiederverwendet und mit anderen Usern geteilt werden.

Diskutieren Sie mögliche Textanalysen in der Gruppe und überlegen Sie, welche Konzepte dafür interessant sein könnten. Wenn Sie z.B. an Figurenanalyse interessiert sind, könnten Sie das Vorhandensein der Figuren im Text, ihr Verhalten, Charakterzüge o.ä. analysieren; wenn sie geographische Besonderheiten interessieren, könnten Sie Typen geographischer Entitäten analysieren – Länder, Städte, Gewässer, Inseln, aber auch besondere Ort etc.; wenn Sie an Themen interessiert sind, die entsprechend Themen...

Die folgenden Aktivitäten brauchen im Falle eines geteilten Texts/Korpus nur von einem Gruppenmitglied durchgeführt werden:

18. Markieren und öffnen Sie Ihre Tag Type Library und erstellen Sie ein Tagset (z.B.: "Figuren", "geographische Entitäten", „Themen“)
19. Klicken Sie auf das Tagset und erzeugen Sie einige Tags (z.B.: "Figurenname", "Verhalten", "Land", "Gewässer", „Natur“, „Dunkelheit“ etc.). Achten Sie dabei darauf, dass sich die Farben der Tags gut voneinander unterscheiden lassen.
20. Sharen Sie die Tag Type Library mit Ihrer Gruppe, indem Sie auf “More Actions -> Share Tag Type Library” im Tag Type Library-Bereich des Repository Manager-Moduls klicken und die Emailadressen Ihrer Gruppenmitglieder einzeln eingeben (Achtung: das funktioniert nur mit registrierten CATMA-Nutzer/innen).

¹ *Hinweis*: Wenn der "Finish" Button nicht anklickbar ist, klicken Sie auf den "Show Preview" Button.

b. Texte manuell (kollaborativ) annotieren

Markup Collections speichern Ihre Annotationen des Textes als Standoff Markup, also unabhängig vom Text. Jeder Text kann mehrere Markup Collections haben und diese können mit anderen geteilt werden. Wenn Sie ein Dokument, zu dem eine Collection gehört, bereits geteilt haben, wird die Collection automatisch geteilt.

Die folgenden Aktivitäten brauchen im Falle eines geteilten Texts/Korpus nur von einem Gruppenmitglied durchgeführt werden:²

21. Öffnen Sie die Markup Collection Ihrer Gruppe, indem Sie sie im Repository Manager-Modul auswählen und auf "Open Markup Collection" klicken. Die Collection wird sich zusammen mit dem Textdokument im Taggermodul öffnen.
22. Unterteilen Sie den Text nach Anzahl der Gruppenmitglieder und weisen Sie jedem Gruppenmitglied einen Teil zu.
 - Passen Sie den Seitengrößenzoom im Tagger-Modul an 50% (wenn Sie zu zweit arbeiten) bzw. 33% (wenn Sie zu dritt arbeiten) an und gehen Sie zu Ihrem Textteil.
 - Alternativ können sie auch die zu analysierenden Kategorien (=Tags) auf die Gruppenmitglieder aufteilen und dann den gesamten Text annotieren.
23. Gehen Sie zum Tab "Active Tagset" und klicken Sie auf "Open Tagset". Wählen Sie Ihre Tag Type Library aus und markieren Sie das gewünschte Tagset. Klicken Sie auf "Load Tagset into currently active document".
24. Lesen Sie den Text und achten Sie auf die Phänomene, für die Sie Ihre Tags erstellt haben. Wenn Sie ein Phänomen identifizieren, markieren Sie die entsprechende Zeichenkette und klicken Sie im „Active Tagsets“-Tab auf den Button neben dem Tag, das sie dafür vergeben wollen (ggf. müssen Sie vorher des entsprechende Tagset durch Klicken auf das Pfeilsymbol davor aufklappen).
25. Machen Sie eine Weile mit der Annotation weiter.

c. Suchergebnisse annotieren: Halb-automatische Suche nach direkter Rede

Wenn direkte Rede in Ihren Texten in Anführungszeichen steht, kann sie folgendermaßen halbautomatisch annotiert werden:

26. Klicken Sie auf den "Analyze Document"-Button (oder wählen Sie das gesamte Korpus über das Repository Manager-Modul durch Klicken auf "More Actions -> Analyze Corpus") und geben Sie folgende Abfragesyntax ein, die nach allen Zeichenketten zwischen öffnendem und schließendem Anführungszeichen sucht: `reg="(?!<=)(\w.*?)(?=\")`. Passen Sie die Art der Anführungszeichen ggf. an den Text an, indem Sie sie aus dem Text in die Abfrage kopieren (die Anführungszeichen stehen jeweils hinter "=").
27. Lassen Sie sich die Ergebnisse in der KWIC-Anzeige anzeigen.
28. Stellen Sie sicher, dass alle Dokumente im Korpus eine Markup Collection haben (Erzeugen Sie ggf. Markup Collections für Dokumente, die noch keine haben).

² *Hinweis:* Wenn Sie Ihre Collection außerhalb eines Korpus erzeugt haben: Teilen Sie die Collection mit den anderen Gruppenmitgliedern, indem Sie auf "More Actions -> Share Markup Collection" im Tag Type Library-Bereich des Repository Manager-Moduls klicken geben Sie die Emailadressen ein.

29. Markieren Sie die Zeilen mit passenden Fundstellen der KWIC-Anzeige (oder wählen Sie alle aus, indem Sie den "Select All" Button klicken) und taggen Sie die Ergebnisse, indem Sie das entsprechende Tag aus dem Active Tagset im Taggermodul (oder aus der Tag Type Library) zur Markierung ziehen. Falls Sie ein neues Tag vergeben wollen, müssen Sie zuerst es in der Tag Type Library erzeugen. Wählen Sie im erscheinenden Fenster die Markup Collection aus, in die die Annotationen gespeichert werden sollen.

Hinweis: Dieses Vorgehen kann in vielen Fällen, die an der Textoberfläche eindeutig erkennbar sind, genutzt werden. So etwa für die Identifikation von Eigennamen in Sprachen, die nur wenig Großschreibung nutzen (Englisch, Französisch, Italienisch etc.), indem Sie sich alle Wörter mit Großbuchstaben am Anfange anzeigen lassen: `reg="[A-Z][a-z]*"`. Wollen Sie nur häufig vorkommende Namen finden, können Sie Ihre Ergebnisse weiter einschränken, etwa mit `reg="[A-Z][a-z]*" where freq>5` für mindestens fünfmal im Text vorkommende Wörter mit einem Großbuchstaben am Anfang.

d. Automatische Annotationen

30. Klicken Sie im Repository Manager-Modul durch auf "More Actions -> Generate Annotations" und wählen Sie im erscheinenden Dialog die Art der Annotationen, die Sie automatisch erstellen wollen
31. Lassen Sie sich die Annotationen anzeigen, indem Sie die erzeugte Markup Collection eines Dokuments im Korpus öffnen (der Name beginnt mit „UIMA ...“)

Hinweis: das automatische Annotieren kann je nach Textumfang und Korpusgröße eine Weile dauern. Sobald die Annotationen erstellt sind, erscheinen sie als neue Markup Collection im Repository Manager.

IV. Weitere Suchabfragen: Annotationen auswerten

Die Analyse Ihrer annotierten Text(e) hängt von Ihrem Erkenntnisinteresse und Ihren Annotationen ab. Die folgenden Schritte sollen Vorschläge für Suchen sein. Suchen Sie gerne nach anderen Aspekten, die Ihnen passender für Ihre Analyse erscheinen!

32. Suchen Sie mithilfe des Querybuilders nach einem von Ihnen vergebenen Tag oder geben Sie direkt folgende Abfrage ein: `tag="[Ihr Tagname]"`. Wenn Sie mehrere Tags unter einem Obertag subsumiert haben, können Sie alle Tags über den Querybuilder oder folgende Suchabfrage finden: `tag="[Obertagname]%"`.
33. Betrachten Sie die Ergebnisse sowohl in Bezug auf die Ergebnisphrasen als auch das Markup (über die entsprechenden Tabs im Analyzermodule).
34. Betrachten Sie die Ergebnisse im Distributionsgraphen.
35. Suchen Sie nach weiteren Tags und lassen Sie sich auch diese im Distributionsgraphen anzeigen.
36. Nutzen Sie den Querybuilder, um komplexe Abfragen zu erstellen (= zum Verfeinern, Kombinieren und Ausschließen von Suchergebnissen).

Sie können jederzeit einen neuen Tab im Analyzer-Modul öffnen, indem Sie auf den "+"-Button oben rechts klicken.