A Short Introduction to CATMA

Outline:

I. Getting Started

II. Analyzing Texts - Search Queries in CATMA

III. Annotating Texts (collaboratively) with CATMA

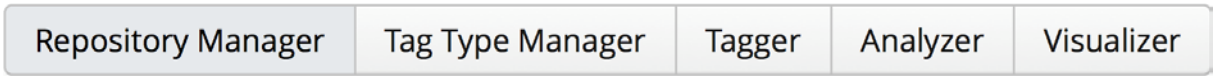
IV. Further Search Queries: Analyze Your Annotations

I. Getting Started

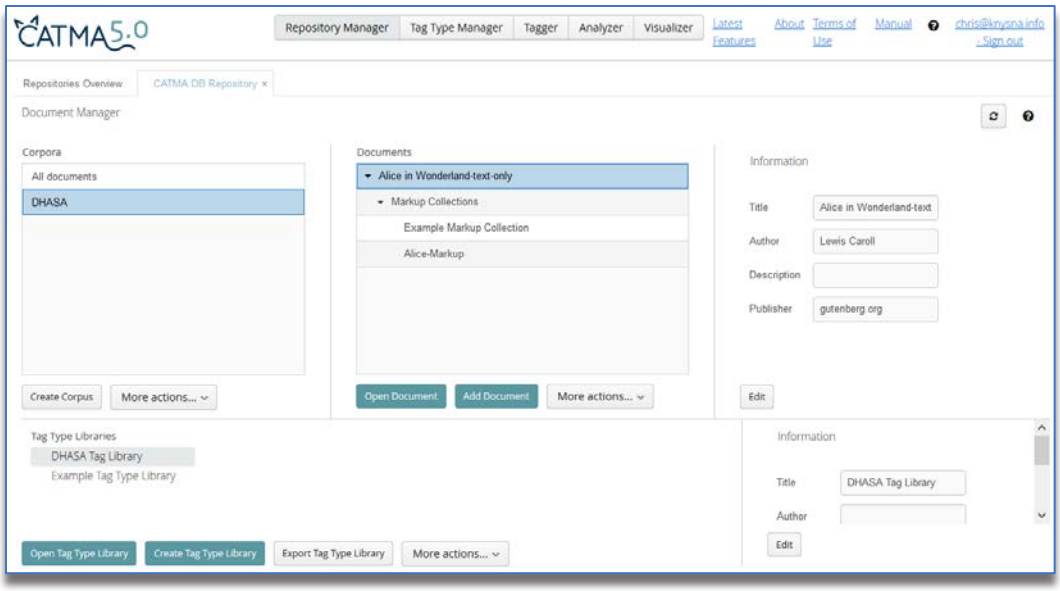
Go to www.catma.de and click this 

button:

In CATMA there are five modules, which you can use by clicking on the respective tab:



After registering and login you will automatically be taken to the “Repository Manager”:



Please note: if you are using the Tutorial Guest Login at http://catma.de/documentation/tutorials/

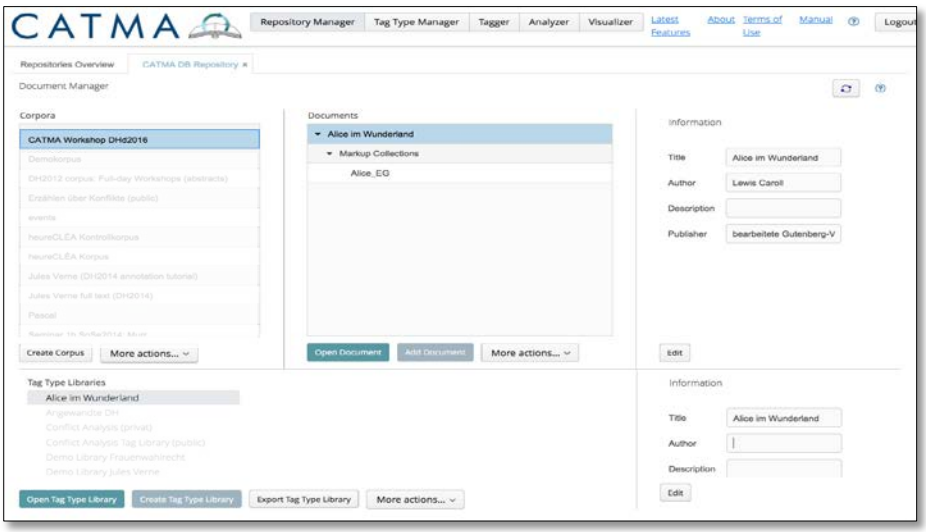
you can skip the following steps 1 to 4 – CATMA will open with Lewis Caroll's "Alice in Wonderland" and all other demo files pre-loaded.

In the Repository Manager you can:

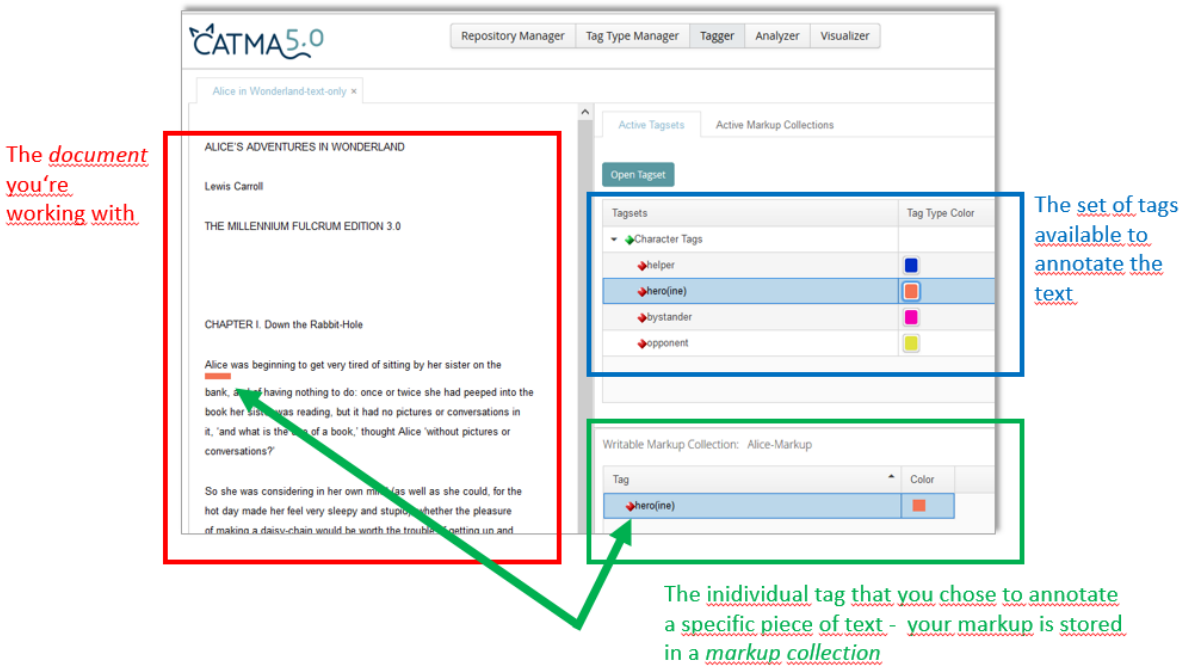
1. Create a corpus (e.g., "CATMA Workshop") using "Create Corpus".
2. Add documents to your corpus (your own files or texts from a text repository, such as Project Gutenberg etc.) via "Add Document".
3. Create a Tag Type Library using "Create Tag Type Library". Choose a suitable name for your library and save it. You can also add your name and a description to the Tag Type Library (select library and click on "Edit").

➔ **"What's a *Tag Type Library*?"** – Let's start the other way round: *tags* are the descriptive labels which you can assign to any piece of text loaded into CATMA. Tags are grouped in *tag sets* – for example, for "Alice in Wonderland" we pre-loaded two tag sets, *Character tags* and *Rhetorical tags*. Every tag set is part of a *Tag Type Library* which can contain as many tags and tag sets as you need.

4. Create a Markup Collection by selecting the text document and clicking "More Actions → Create Markup Collection".



➔ **"What's a *Markup Collection*?"** – When you annotate a text in CATMA using tags, you create what is referred to as *Markup*. You can have different types of markup for one text, or you might annotate a text in a team where everybody creates their own version of the markup. The *Markup Collection* is where all annotations pertaining to one text or one corpus are stored.

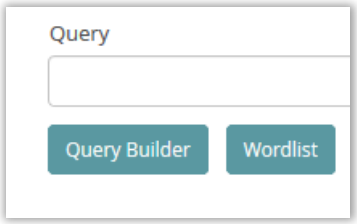


II. Analyzing Texts - Search Queries in CATMA

Highlight a text document which you uploaded in the Repository Manager and click on the "Open document“ button.

Your document will now open in the Tagger Module. Below the text window on the left, click on the “Analyze Document“ button.

You will be taken to the Analyzer Module and see two buttons in the top left section.



There are a couple of interesting operations which you can now do immediately:

Word lists

5. Click the "Wordlist" button.
6. Sort the word list in descending frequency. What is the most common content word (= a word with "more" semantic meaning than function words such as articles, pronouns etc.)? *
7. What is the second most common word? How frequent is it?*

KWIC visualization (KWIC = keyword in context)

8. For the selected word: Click on the "Visible in KWIC" box and look at the KWIC display on the right hand side of the screen. If necessary, change the scope of the displayed context using the slider at the bottom (setting to 1-30 tokens possible).
9. Double-click on one of the keywords in the KWIC – CATMA will now jump to the selected keyword's full-text location in the Tagger Module.

DoubleTree visualization

10. Return to the word list in the analyzer module. Look for the word "hatter" in the list and click on the DoubleTree button (2nd button at the bottom left).
11. Click on the words in the DoubleTree visualization and try to figure out how the DoubleTree works. Can you tell something about the character? *
12. Look for how often the term "hatter" occurs. Then analyze it in the DoubleTree and describe your results. *
13. Try the same with "Alice".

Distribution graph

14. Go back to the analyzer module, select the line that says "Alice" again and click on the distribution graph button (1st left below).
15. Go back to the analyzer module and select a different word that is interesting to you. You can also display this in the distribution graph. Is there something interesting that can be seen from the representation of the two words?
16. Go back to the analyzer module and select a group of words that will be interesting to you in the word list. Select the appropriate lines (by holding the command key) and display the word group in the distribution graph.
17. Compare the occurrence of the terms "Alice" and "curious" in the distribution graph. What can be observed? *

Query Builder

18. Return to the Analyzer module and open the Query Builder by clicking on the corresponding button.
19. Use the Query Builder to search for one of the selected words – or multiple words ending with the same letters. Use the "by word or phrase" query option.
20. Use the Query Builder to find all words that occur more than ten times. Use the query function "by frequency". *
21. Use the Query Builder to find all words that are between forty and fifty times. Are there animals among them?
22. Using the Query Builder, search for all words with 70% similarity to "confused". Use the "by grade of similarity" query option. Open a new Analyzer tab (by clicking the "+ New Query" button on the upper right) and increase the similarity to 80%. Do it again with 75%. *
23. Using the Query Builders, find how often the word "timidly" appears near the word "Alice" (with a span of ten words). Use the "collocation" function. *

III. Annotate Texts (collaboratively) with CATMA

*Note on **Sharing***: When you work with a shared corpus, each text document and markup collection you add to the corpus is automatically shared with everyone who has access to the corpus. If you delete something from the corpus, you can no longer access or see it, but it remains visible for everybody else.
Also, note that everything that you share with others in "Write" mode can be edited by them. If you do not want these other users to change your data, you can either share your items in the "read" mode or export them (click on the document / markup collection / tag library → Click the "More actions" button → "Export Document" / "Export Markup Collection" | "Export Tag Library"), send the file to your colleagues by e-mail and have the file imported by them (click the "More actions" button → "Import Document" / "Import Markup Collection" / "Import Tag Library"). The changes that your colleagues make to the thus 'copied' element will not affect your version and vice versa.

a. What do you want to find out? → Tag Creation (in groups of two or three)

For the annotation you need a Tag Type Library – either an existing or a new one that you create. Tag Type Libraries can be reused for other documents and shared with other users.

Discuss possible approaches to a textual analysis in the group and think about concepts fpr the analysis that might be of interest. For example, you might be interested in analyzing the presence of the characters in the text, their behavior, their character traits or the like; if you are rather interested in geographical features, you could analyze types of geographic entities - countries, cities, waters, islands, but also special places, etc.; If you are interested in topics, identify relevant topics, etc.

In the case of a shared text or corpus the following activities need only to be performed by one group member:

24. Select and open your Tag Type Library and create a tag set (for example, "Characters", "Geographic Entities", "Themes").
25. Click on the tag set and create some tags (for example: "Character name", "Behavior", "Land", "Waters", "Nature", "Darkness", etc.). Make sure that the colors of the tags can be clearly distinguished from each other.
26. Share the Tag Type Library with your group by clicking " More Actions → Share Tag Type Library" in the Tag Type Library area of the Repository Manager module, and entering the email addresses of your group members (Note: This only works with registered CATMA users). Decide whether your group members should only be able to use or edit the Tag Type Library and choose "Read" or "Write".

b. Annotating texts manually (collaboratively)

Markup Collections save your annotations of the text as stand-off markup, that is, independent of the text. Each text can have multiple markup collections and these can be shared with others. If you have already shared a document to which a collection belongs, the collection is also shared automatically.

In the case of a shared text / corpus, the following activities only need to be carried out by a group member:

27. Open the Markup Collection of your group by selecting it in the Repository Manager module and clicking "Open Markup Collection". The collection will open together with the text document in the Tagger module.
28. Divide the text by the number of group members and assign a part to each group member.
    - Adjust the page size zoom in the Tagger module to 50% (if you are working in pairs) or 33% (if you are working in pairs) and go to your text part.
    - Alternatively, you can share the categories (= tags) to be analyzed to the group members and then annotate the entire text.
29. Go to the "Active Tagset" tab and click "Open Tags". Select your tag type library and select the desired tag set. Click "Load Tagset into currently active document".
30. Read the text and pay attention to the phenomena for which you have created your tags. If you identify a phenomenon, select the regarding text string and click on the button next to the tag in the "Active Tagsets" tab (you may need to open the respective tag set by clicking on the arrow symbol).
31. Continue to annotate for a while.

c. Annotating search results: semi-automatic search for direct speech

If direct speech is enclosed in quotation marks in your texts, it can be annotated semi-automatically as follows:

32. Click on the "Analyze Document" button (or select the entire corpus of the Repository Manager module by clicking "More Actions Analyze Corpus") and enter the following

query syntax between all the strings between opening and closing quotation mark:

reg = "(?<=\W[‘])(.*?)(?=[’]\W)"

33. If necessary, adjust the type of quotation marks to the text by copying them from the text into the query (the quotation marks are always behind "= \").
34. Display the results in the KWIC display.
35. Make sure that all documents in the corpus have a markup collection. (If necessary, create Markup Collections for documents that do not already have one).
36. Select the rows with matching KWIC display (or select all by clicking the "Select All" button) and tag the results by selecting the appropriate tag from the Active Tag set in the tagger module (or from The Tag Type Library). If you want to assign a new tag, you must first create it in the tag type library. Select the Markup Collection in which the annotations should be saved in the window that appears.

Note: This approach can be used in conjunction with many phenomena that are clearly recognizable on the text surface, such as the identification of nouns in languages that use little capitalization (English, French, Italian, etc.) by searching for all words starting with uppercase letters: reg = "[AZ] [az] *". If you only want to find frequently occurring names, you can further restrict your results, for example with reg = "[A-Z] [a-z] *") where freq> 5 for at least five words in the text with a capital letter at the beginning.

d. Automatic annotations

NB: At current this works for German language texts only – English will follow soon.

37. In the Repository Manager module, click "More Actions → Generate Annotations" and select the type of annotations that you want to create automatically in the dialog that appears.
38. Display the annotations by opening the generated markup collection of a document in the corpus (the name begins with "UIMA ...").

Note: Automatic annotation can take a while, depending on the text size and corpus size. Once the annotations are created, they appear as a new markup collection in the Repository Manager.

IV. Further Search Queries: Analyze Your Annotations

The analysis of your annotated text(s) depends on your research interest and your annotations. The following steps are suggestions for searches. Think about other aspects that are (more) suitable for your analysis!

39. Use the Query Builder to search for a specific tag, or type the query directly: tag = "[your tag name]". If you have subsumed several tags under a tag, you can find all tags using the Query Builder or the following search query: tag = "[parent tag name]%".
40. Examine the results for both the results by phrase and the results by markup (use the corresponding tabs in the analyzer module).
41. Examine the results in the distribution graph.
42. Look for other tags and display them in the distribution graph.

43. Use the Query Builder to create complex queries (= to refine, combine, and exclude search results).

You can always open a new tab in the Analyzer module by clicking the "+" button in the upper right corner.

ANSWERS

Question 6:
"said", 456 times (or, in case you don't consider pronouns functions words, "she", 502 times).

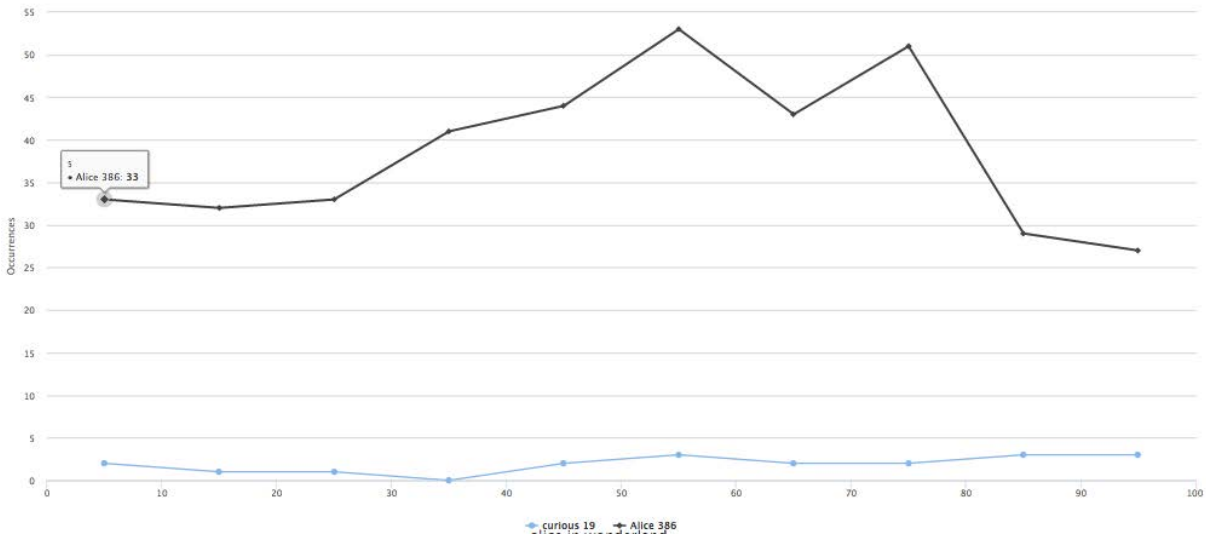
Question 7:
"Alice", 386 times.

Question 11:
The DoubleTree seems to draw a fairly clear picture, for example, "The miserable hatter," "sighed the Hatter," "muttered the Hatter, "said the Hatter with a sigh".

Question 12:
54 times.
The hatter is associated with (partly emotionally) negative verbs (examples: growled, sighed, interrupted). The same applies to adjectives (examples: sad, unhappy, sharp). This may lead to conclusions about the character of the hatter as well as its relationship to and / or its behavior towards Alice.

Question 13:
In the case of Alice, a clear picture cannot be seen at first glance - there are significantly more and clearly more variable results.

Question 17:



Here you can see where "Alice" and "curious" co-occur, and thus where she might be described as curious. However, a restriction is that other similar / related terms such as "curiosity" were not included in this analysis.

Question 20
There are 377 words that occur more than ten times.

Question 21:
There are 31 different words. The rabbit is among them.

Question 22:
Results for 70% Similarity to "confused":
"consented", "dunce", "offend", "concluded", "closed", "caused", "confusing", "confused", "sound", "second", "confusion", "continued.

With 80% similarity only "second" and "confused" is listed.

75%: "second", "confused", "dunce" and "sound".

Question 23:
Six times.