A Short Introduction to CATMA

Outline:

I.   Getting Started
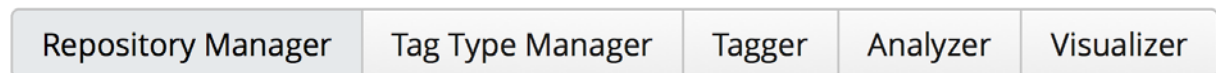
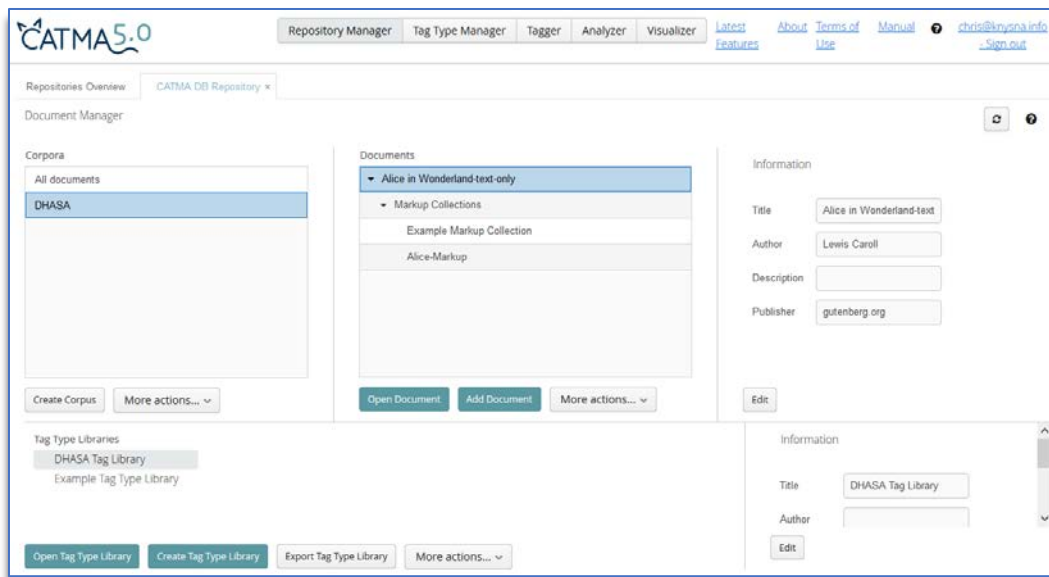Go to www.catma.de and click this   button:

Note: You can always click on the little question marks that can be found in CATMA for further help.

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: since you are using the Tutorial Guest Login at http://catma.de/documentation/tutorials/, several preparatory steps have already been carried out automatically – CATMA will open with Lewis Caroll's "Alice's Adventures in Wonderland", and a Tag Type Library as well as a Markup Collection have been created for you.

➔ "**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's Adventures 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.

➔ "**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.

the *document* you're working with

the set of tags available to annotate the text

the individual tag that you chose to annotate a specific piece of text – your markup is stored in a *markup collection*

II. Analyzing Texts – Search Queries in CATMA

Since you are working with the preloaded version, your document "Alice's Adventures in Wonderland" is already displayed in the <u>Tagger Module</u>. Below the text window on the left, click on the "Analyze Document" button.

You will be taken to the <u>Analyzer Module</u> and see two buttons in the top left section.
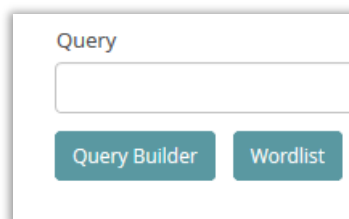
There are a couple of interesting operations which you can now do immediately:

Word lists

1. Click the "<u>Wordlist</u>" button.

2. Sort the word list in <u>descending frequency</u>. What is the most common content word (= a word with "more" semantic meaning than function words such as articles, pronouns etc.)?* (→ For all questions with asterisk, there are answers provided at the end of the tutorial.)

3. What is the second most common word? How frequent is it?*

KWIC visualization (KWIC = keyword in context)

4. Now select the word "Alice" in the wordlist. Tick the "Visible in KWIC" box for "Alice" and look at the <u>KWIC display</u> 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).

5. Double-click on one of the keywords in the KWIC – CATMA will now jump to the selected keyword's <u>full-text</u> location in the Tagger Module.

DoubleTree visualization

6. Return to the word list in the analyzer module. Un-tick "Alice". Now look for the word "Hatter" in the list, highlight it and click on the <u>DoubleTree</u> button (2nd button at the bottom left).

7. Click on the <u>words</u> in the DoubleTree visualization and try to figure out how the DoubleTree works. Can you tell something about the character?*

8. Try the same with "Alice".

Distribution graph

9. Go back to the analyzer module, select the line that says "Alice" again and click on the <u>distribution graph</u> button (1st left below).

10. Go back to the analyzer module and select the word "curious". If you now click the distribution graph button, it will be displayed in the same diagram. Is there something interesting that can be seen from the representation of the two words?

11. Close the visualizer tab, go back to the analyzer module and select <u>a group of words</u> that will be interesting to you in the word list. This is done by selecting the appropriate lines (by holding the command key). If you now click the distribution graph button, the words will be displayed in a single distribution graph.

Query Builder

12. Return to the Analyzer module and open the <u>Query Builder</u> by clicking on the corresponding button.

13. Use the Query Builder to search for the word "Queen". Use the "by word or phrase" query option and click "Show in preview" to test whether your query generates a result, then click "Finish".

14. Use the Query Builder to find all words that occur more than ten times. Use the query function "by frequency".*

15. Use the Query Builder to find all words that occur between forty and fifty times. Are there animals among them?

16. 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%.*

17. Using the Query Builder, find how often the word "timidly" appears near the word "Alice" (within a span of ten words). Use the "collocation" function.*

III. Annotating own texts with CATMA

Whenever we engage with texts in order to explore and discuss their meaning, two operations go hand in hand: analysis and interpretation. Their mutual dependency results in what is known as the *hermeneutic circle*.

CATMA supports this approach by combining text analysis and text annotation in a cyclical, iterative workflow. Once you have uploaded your digital text to the CATMA repository, you will typically proceed along the following phases during each of these cycles:
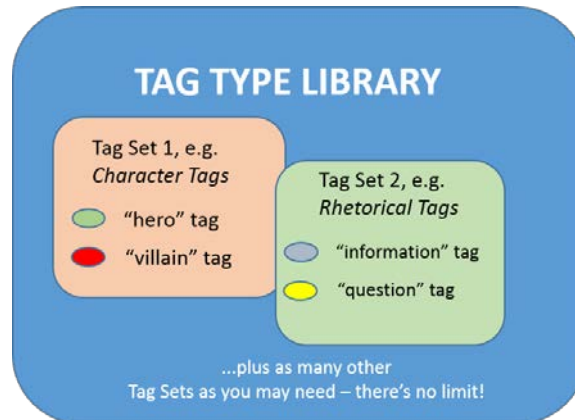
PHASE 1: Clarify your problem – what do you want to find out about the source text?

18. Take some time to identify the questions and problems that you would like to investigate upfront. 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 more 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.

19. Now look at your list of questions – they normally fall into three different categories:

   a. Questions that can be answered directly by investigating the actual words and phrases that constitute the text. For example, which character in 'Alice in Wonderland' is mentioned *verbatim* more often – Alice or the Hatter? A WORDLIST analysis with the CATMA analyzer already answers this one: Alice is mentioned more often than the Hatter!

   b. Questions that require you to pre-interpret textual elements in order to find out what they mean. For example: in addition to names, characters in 'Alice in Wonderland' are also identified by pronouns, such as *she.* But which of these point to Alice and which to the Queen? You can make those instances of *she* that refer to Alice explicit by annotating them with CATMA tags.

   c. Questions that require you to combine both strategies – direct textual analysis and pre-analytical text interpretation. So if you really want to find out <u>every</u> mention of Alice in the book, you can now use the CATMA analyzer to do a combined search for literal occurrences of the name *Alice* plus a search for all occurrences of *she* that you annotated as an indirect reference to Alice.

Type (b) and (c) questions are the ones where TAGS come into play – that is, interpretive annotation labels that you attach to a word in order to make its implicit meaning or function explicit to CATMA.

PHASE 2: Tag creation

In CATMA tags are managed by grouping them into Tag Sets, which in turn form part of a Tag Library:

TAG TYPE LIBRARY

Tag Set 1, e.g. *Character Tags*

⬭ "hero" tag

🔴 "villain" tag

Tag Set 2, e.g. *Rhetorical Tags*

⬭ "information" tag

⬭ "question" tag

...plus as many other Tag Sets as you may need – there's no limit!

This hierarchical structure guarantees that you can extend and modify your tags at any time – in other words, you don't have to worry about some missing tags when you start; you can always add them later. Also, the structure enables you to share selected tag sets with other CATMA users and/or to import theirs. (We are not dealing with the SHARING function in the current tutorial though.)

20. Upload your own text in the CATMA Repository Manager. Then open your document: you will be taken to the Tagger window.

21. In the Tagger window, click on "Open Tag Set". First create your own Tag Type Library (e.g., "My Tag Library") – this is a virtual folder where you collect all your (reusable) Tag Sets.

22. Within that Tag Type Library, now define a first Tag Set. Tag Sets allow you to group a set of related annotation labels (for example, "Characters", "Geographic Entities", "Themes").

23. Click on your tag set and now create the actual tags that you require (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.

PHASE 3: Annotating texts

24. Go to the "Active Markup Collection" tab and click "Open Markup Collection". In the pop-up box, click on "Create Markup Collection". This is where all your annotations relating to this particular source text will be collected.

25. Go to the "Active Tagset" tab and click "Open Tags". Select your tag type library and the desired tag set. Then click "Load Tagset into currently active document" – you can now use all tags within the tag set to annotate your text.

26. In the Tagger window, read your source text and pay attention to the phenomena for which you have created your tags. If you identify a phenomenon, select the corresponding 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).

27. Continue to annotate manually for a while.

28. Repetitive manual tagging of the same word across an entire text can be avoided by running a query first and then annotate the search results. First, run a query as described in Part II; steps 1–4 above. Then activate the KWIC display, highlight the relevant KWIC lines and click on "Tag selected results": a box displaying your Tag Set and its Tags will open. Select and draw the relevant tag onto the KWIC lines: all the selected search results have now been tagged with the same tag.

PHASE 4: 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!

29. Use the Query Builder to search for a specific tag, or type the query directly: **tag = "tag name"**.

30. Examine the results for both the results by phrase and the results by markup (use the corresponding tabs in the analyzer module).

31. Examine the results in the distribution graph.

32. Look for other tags and display them in the distribution graph.

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

When you run the next analysis, a new tab in the Analyzer module will open – so you don't lose your previous results. You can also open a new tab 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", 509 times).
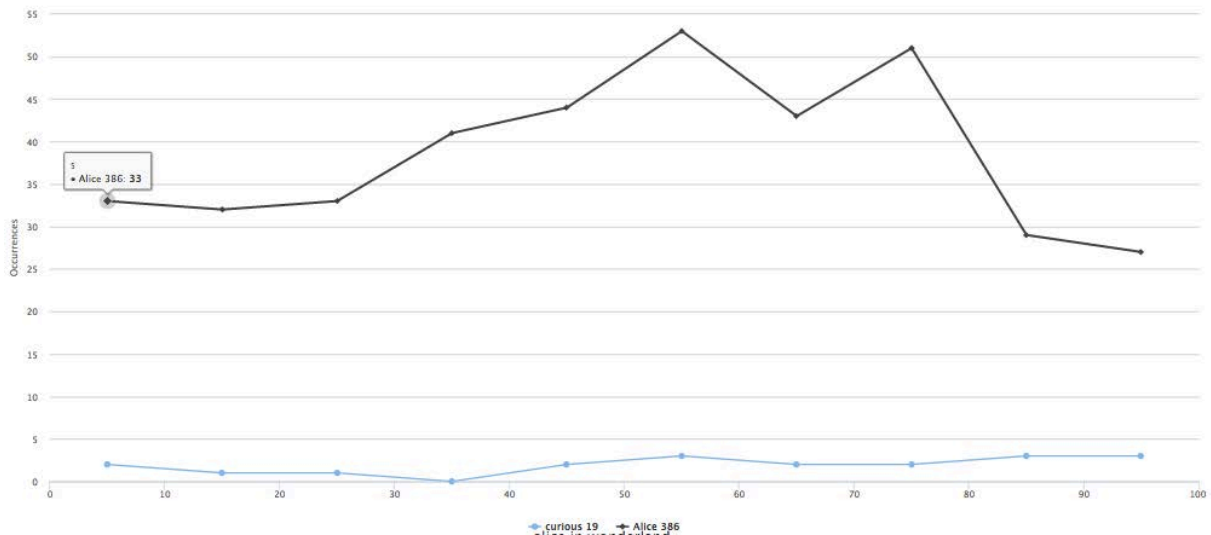
Question 7:
"Alice", 395 times.

Question 11:
The DoubleTree allows you to explore a word in its various contexts. For example, the Hatter is portrayed in the text as "The miserable Hatter", "sighed the Hatter", "muttered the Hatter", "said the Hatter with a sigh". 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 15:

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 18
There are 379 words that occur more than ten times.

Question 19:
There are 32 different words. The rabbit is among them.

Question 20:
Results for 70% Similarity to "confused": 14 words.

With 80% similarity, only "second" and "confused" are listed.

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

Question 21:
Six times.