First Steps in CATMA

Here are some tasks that were put together as a short introduction to the most important functionalities in CATMA 4 (CLÉA).

Before you start:
Log on at www. digitalhumanities.it/catma. After logging on you should see the tutorial corpus in the "Corpora" section of the Repository Manager window.

Part I
Intro: basic CATMA functionalities

Wordlists

1. Mark the corpus and choose "More Actions -> Analyze Corpus" in the corpus section of the repository manager window.
2. Click on the "Wordlist" button
3. Sort the word list by descending frequency. Which is the most frequent content word? (e.g. word with “more” semantic meaning than function words like articles, pronoun etc.)?

KWIC visualization (KWIC = KeyWord In Context)

4. For the chosen word: check the "Visible in KWIC" box and look at the KWIC display
5. Choose one of the words in the KWIC display and double click on it in order to jump to its position in the full text

Double Tree visualization

6. Close the Tagger window and go back to the left side of the Analyzer window. Mark the row of the first content word and click the DoubleTree button (second from left on the bottom)
7. Click on the words in the Double Tree visualization and try to figure out, how the Double Tree works
8. Go back to the Analyzer Window and click on the arrow on the left of the chosen word. In which chapter does the word occur most frequently? (you can sort the chapters by frequency if you click on one of them before sorting the results by frequency)

Distribution Graph

9. Go to the Repository Manager window, choose the chapter with the most occurrences of the chosen content word (in the "Documents" section) and click on "More Actions -> Analyze Document"
10. Build a word list for the chapter. Mark the row with the chosen word in the corpus and click on the distribution graph button (first from left on the bottom).
11. Go back to the Analyzer window and choose another word you consider interesting from the word list. Display that word in the distribution graph, too. Look at the graph: Is there anything interesting you could tell from the distribution of these two words?
12. Go back to the Analyzer window and choose a group of words you consider interesting from the word list. Mark the rows (hold the command key while clicking on the rows) and display the word group in the distribution graph.

Query Builder

13. Go to the Repository Manager window, choose the corpus again (in the "Corpora" section) and click on "More Actions -> Analyze Corpus"
14. Open the Query Builder by clicking on the respective button in the Analyzer window.
15. Use the query builder for searching for one of the chosen words–or more words starting or ending with the same letters

Note: If the "Finish" button is not clickable after you have built a query, click on the "Show Preview" button before going on.

16. Search for all words that appear more frequently than 7 times
17. Search for all words with 80% similarity to the word "island"

Any findings?

Discuss your results with your neighbor/s in order to prepare the next step (e.g. the annotation):

- Did one or more of your actions produce possibly interesting results?
- Did you find one or more starting points for the further text analysis?
- What could be the next steps for the analysis of the text?

Part II
Annotating texts collaboratively

Note: We work with a shared corpus and therefore every text document or markup collection that is put in our corpus is shared automatically with those who have access to the corpus. If you delete a document or a collection it will be deleted only from your view of the corpus, everybody else will still see the deleted item in their views.
(If you feel very uncomfortable with this you can avoid the automatic sharing by creating your User Markup Collection outside the corpus, e.g. in your "all documents" corpus.)

What are you interested in? Creating tags (in pairs/groups of three)

For the annotation you need to choose or build a tag library that contains the tags you want to use. Tag libraries can be reused for other documents as well as shared with other users.
Choose one of the analysis possibilities you discussed before and try to think about related concepts (e.g. if you are interested in character analysis you could analyze the presence of the characters in the text as well as their behaviour, character traits etc.; if you are interested in the geographic peculiarities you might analyze types of geographic entities–lands, cities, rivers, seas, islands, etc.,– or the geographic area the entities belong to, and so on…)
The following actions need to be performed by one group member only:

18. Create a new tag library by clicking on the button "Create Tag Library" in the Repository Manager window. Choose an expressive name for your library and save it. If you wish, you can add your name and a description afterwards (cf. the "Edit" button next to the tag libraries).
19. Open your tag library and create a tag set (for example: "characters" or "geographic entities")
20. Click on the tag set and create some tags (for example: "professor Liedenbrock", "Axel", "Martha", "Hans Bjelke" etc., or "island", "volcano", "river", "city"; "Germany", "Iceland"). Make sure that the colors of your tag are easily distinguishable.
21. Share the tag library with the person(s) you are working with be clicking on "More Actions -> Share Tag Library" in the Tag Library section of the Repository Manager window and entering their email addresses one by one

Creating a User Markup Collection

User Markup Collections store your annotations of a text as standoff markup. Every text can have more User Markup Collections and you can share your collections with other users, if you wish.
Here again just one group member needs to create the collection:

22. Choose Chapter 10 form the corpus in the Repository Manager window and create a User Markup Collection by clicking "More Actions -> Create User Markup Collection". Type in an expressive name (for example "Chapter 10 [your names]").
    - If you built your collection outside the corpus: Share your collection with your group members by clicking on “More Actions -> Share User Markup Collection” in the Tag Library section of the Repository Manager window and entering their email addresses one by one. (If you built in inside the corpus it will be visible automatically)

Annotating the text manually (and collaboratively)

23. Open the User Markup Collection of your group by selecting it in the Repository Manager window and clicking on “Open Markup Collection”. (It will open in the Tagger window, together with the text it belongs to)
24. Open your Tag Library
25. Drag the tag set you want to use to the Tagger window and drop in the "Active Tagset" tab.
26. Split the text in as many portions as your group has members and assign every member a portion of text.
27. Adjust the page size zoom in the Tagger window to 50% (if working in two) resp. 33% and go to your text portion.
28. Read through the text and look for the phenomena you described with your tags. If you encounter one (for example one of the characters you included in your tag set), mark the respective string and click on the button next to the tag you want to use for it.
29. Keep on annotating for some time.

Annotating query results: Semi-automatic detection of proper names

30. Click on the “Analyze Document” button and enter the following query syntax for searching for all words starting with a uppercase letter followed by any number of lowercase letters and occurring at least five times: `(reg="[A-Z][a-z]*") where freq>5`
31. Search the result list for proper names and check the box for the KWIC display for them. (If you ar not sure, check the KWIC representation and double click on the keyword in order to get to the text)
32. Mark the whole list in the KWIC display by clicking the "Select All" button and tag the results by dragging the appropriate tag from the Tag Manager window and dropping it in the marked KWIC section. Choose your markup collection, if necessary.

Note: You can use this functionality in many cases. For example, in our text direct speech could be identified as a string between two quotation marks that starts with a word character. It can be found with the following regular expression: `reg="(?<=\")(\w.*?)(?=\")"`

Any findings or observations?

Discuss your work with your group again

Part III
Analyzing the annotated text(s) – More Queries

The analysis of your text(s) depends very much upon your research interest and the annotations you have. The following steps are meant as suggestions. Feel free to search for other aspects!

Step 1: Analyzing your annotations in Chapter 10

33. Use the query builder for searching for one of the tags you used in your annotations or type your query directly into the query line of the analyzer tag="[your tagname]".
34. View the results both in the phrase and the markup representation (cf. the respective tabs in the analyzer window)
35. View the results in a distribution graph
36. Search for more tags and view the results in the distribution graph. Use the query builder for building complex queries (= refining, combining, excluding results), if you wish

Note: You can open a new tab in the Analyzer window by clicking on the "+" button on its upper right

Step 2: Analyzing annotations in the full text

37. Open the markup collections that have been added by the tutorial facilitators.
38. Figure out possible connections between your tagging and the tagging in the added collections and build queries in order to analyze them (e.g, how many of the superlatives are used within direct speech?[1]; how many adverbs occur close to one of the figures?[2])

[1] Queries:

- [1]superlatives: tag="superlative"
- [1]superlatives within direct speech: tag="superlative" - (tag="superlative" where tag="direct_speech" boundary)

[2] Use collocation queries for this, for example: tag="Liedenbrock" & tag="adverb"

[2]You can adjust the span of words within which CATMA searches for the collocations by adding the respective number (the default is 5), e.g. tag="Liedenbrock" & tag="adverb" 3 searches for collocations within three words.