TXM workshop
–
a gentle introduction to TXM key concepts in 1 ½ hour

IQLA-GIAT Summer School in
Quantitative Analysis of Textual Data
University of Padua, 16-20 September 2013

Serge Heiden

This workshop will introduce you to the TXM text analysis tool and particularly to some of its key features:

- the Graphical User Interface environment with hypertext and window manager (desktop and web portal versions);
- how to search for words or n-gram patterns (based on word form, part of speech, lemma...) with the Query Assistant to list or count them, and the underlying powerful CQL query language;
- how to build a table of linguistic pattern frequencies, for example most frequent "ADJ NOUN" lemma sequences, to compare fictional texts in the Brown corpus with specific word patterns analysis (build a sub-corpus, a partition, a CQL based Index and word specificity analysis);
- how to exploit with TXM tools the text structures and the word properties encoded in XML sources;
- how to transfer results to the R environment embedded in TXM and call R scripts on those objects, a way to prepare complex linguistic data tables to be processed by existing R packages (eg Stylo).

We will use the Brown corpus (500 samples of English-language text, totaling roughly one million words, compiled from works published in the United States in 1961).

Foreword

This introduction is focused on specific features for a very brief presentation. If you have time, for a complete introduction please read the TXM manual (in French)[1] or watch the TXM introductory workshop in Youtube (in French).

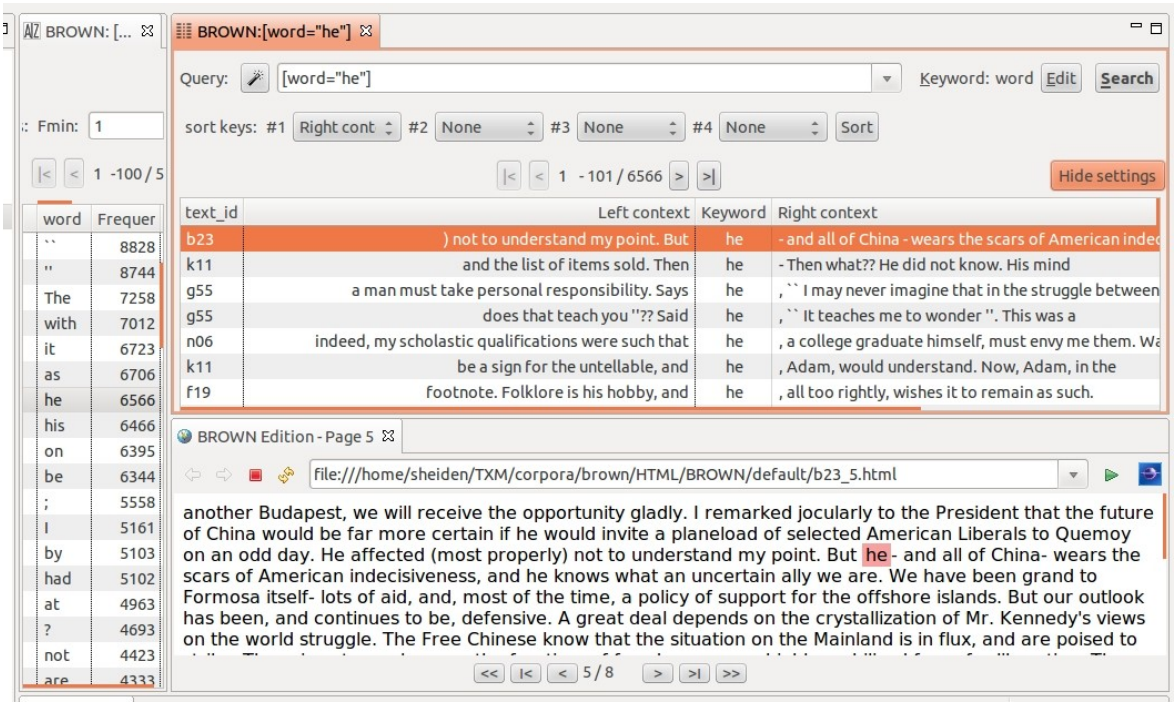
Graphical User Interface

TXM combines qualitative tools (word lists, concordances...) with quantitative tools (specific words analysis, collocate words analysis...) through a standard GUI:

- the interface consists of: the view on the left to select corpora, sub-corpora or results, main menus on top to apply tools on selected corpora, result windows on the right and a messages console on the lower right;
- first, lets look at the Brown corpus properties we are going to work on:
  - select the Brown corpus icon;
  - launch the Description command by clicking on the "'I' in circle" button in the toolbar → the numbers of words, of word properties, of structures, etc. are displayed in a new window on the right panel.

1 Or the old TXM 0.5 manual English version.

- let's calculate a word frequency list of Brown corpus → Lexicon tool;
- from the lexicon result window, let's calculate a Concordance of “he” word form with frequency 6566 down the list, by double-clicking on the “he” hypertextual line in the lexicon result;
- the interface layout is managed by an intuitive window manager:
- let's display the concordance next to the lexicon by dragging its tab on the right of the results area and releasing when the phantom window cuts half the area by a vertical line;
- let's sort by right context by clicking on the “Right context” column header;
- let's display the text page where the first pivot occurs by double-clicking on the first line of the concordance (“not to understand my point. But”, “he”, “- and all of China - wears the scars of American indecisiveness”). The text Edition page is displayed with the match highlighted;
- let's display the edition under the concordance by dragging its tab down the results area and releasing when the phantom window cuts half the concordance area by a horizontal line:



- one can browse the concordance contexts by double-clicking on each concordance line and updating the text edition window;
- one can repeat the same scenario with the “she” word form with the frequency 1949 down the lexicon and display its concordance and text pages;
- then “he” and “she” window groups can be organized together or side by side (with four windows).

The demo portal has an equivalent user interface: corpora, lexicon, concordance and edition view by hypertextual double-click (http://portal.textometrie.org/demo/?locale=en).

The Query Assistant and the CQL query language

Example 1: Searching for a word ending

In the “he” concordance, we can see the [word="he"] expression in the query field.
We can change the query using the query assistant launched by clicking the wand button.
In the assistant, we can build a query to search for words ending with “ing”:

- change the line - Word n°1 with its property [enlemma] 'equals to' [] - by - Word n°1 with its property [word] 'ends with' “ing” - and click OK

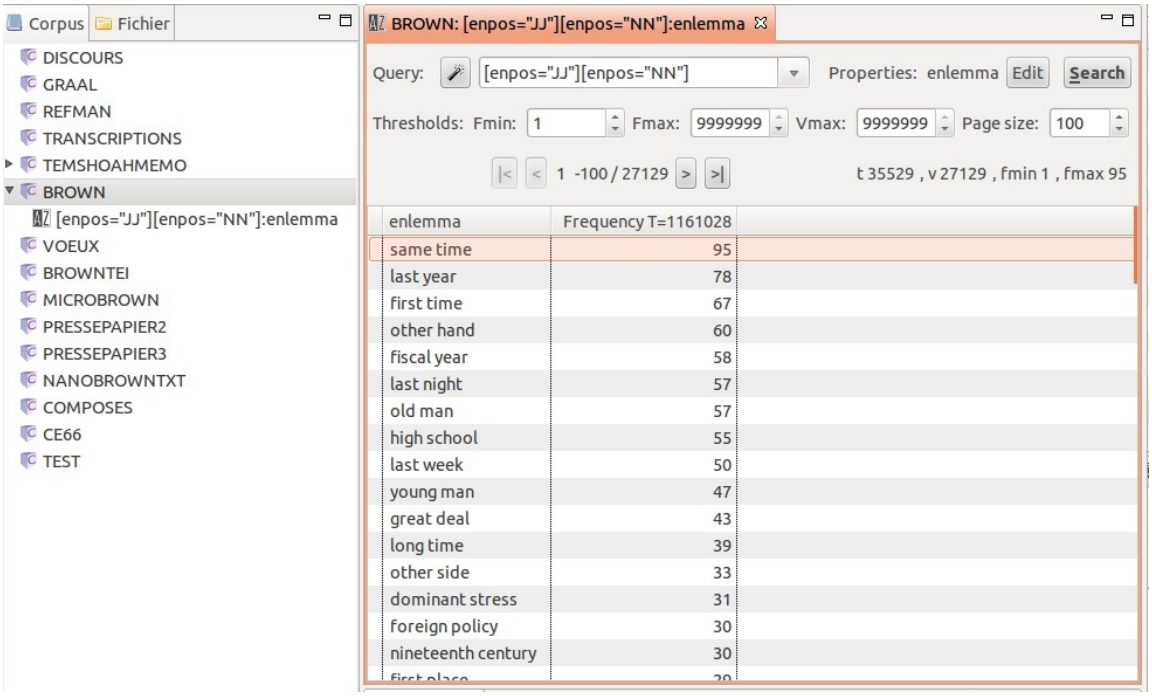
→ a new query expression - `[word=".*ing"]` - should appear in the query field[2];

- you can now build its concordance by clicking “Search”;
- now, to build the frequency list of all the words matching that query, we can use the Index tool directly on the Brown corpus instead;
- to reuse the previous query expression, use the “arrow down” button right to the query field to select the query in the history list then click “Search”.

Example 2: Searching for the lemmas of a sequence pattern

In a new Index window on the Brown corpus, we can search for the lemmas of the “an ADJECTIVE immediately followed by a NOUN” sequence pattern:

- launch the query assistant by clicking the magic wand button;
- change the line - Word n°1 with its property [enlemma] 'equals to' [] - by - Word n°1 with its property [enpos] 'equals to' “JJ”[3];
- click on the “Add a word” button;
- change the line - Word n°2 with its property [enlemma] 'equals to' [] - by - Word n°2 with its property [enpos] 'equals to' “NN” - and click OK
  → a new sequence query expression - `[enpos="JJ"][enpos="NN"]` - should appear in the query field;
- in the Index parameters, change the “Properties” field from “word” to “enlemma”;
- you can now build the frequency list by clicking “Search”:



2 When you will know the query syntax, you will be able to directly type the expression in the query field without using the query assistant.

3 The enpos property has been added by TreeTagger. The tagset is described here http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Penn-Treebank-Tagset.pdf.

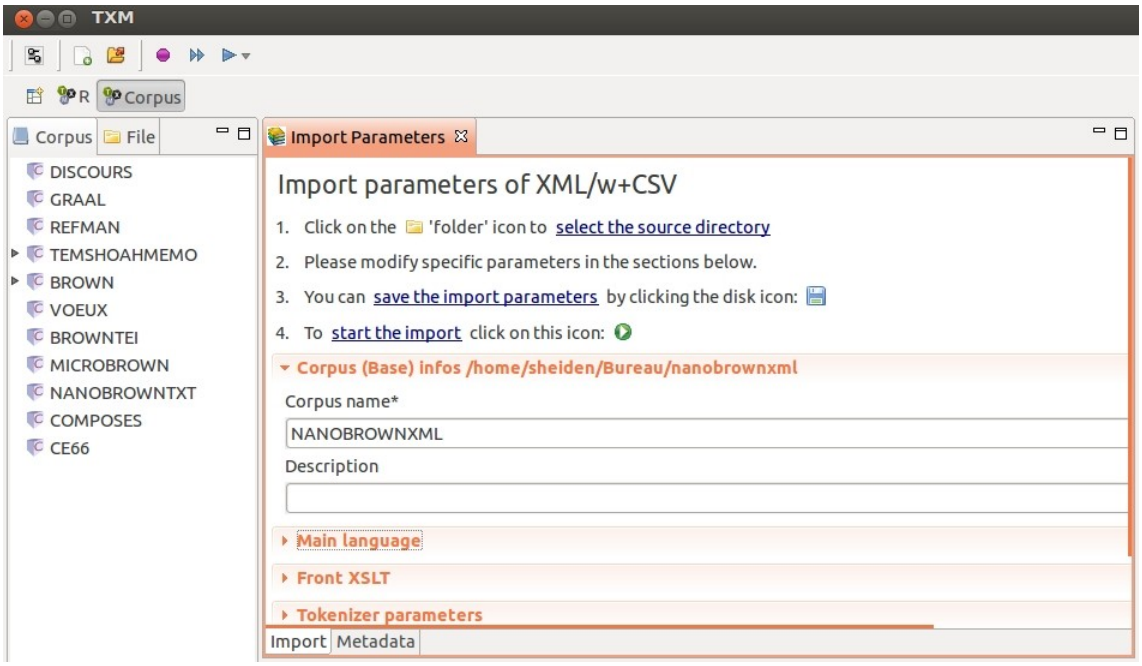
Exploiting XML encoding

To illustrate the use of XML tags, let's use an XML version of the Brown corpus with texts encoded like the “a01” press article sample of the Brown corpus ("Atlanta Primary …", The Atlanta Constitution, November 4, 1961, p.1), stored in the “a01.xml” file:

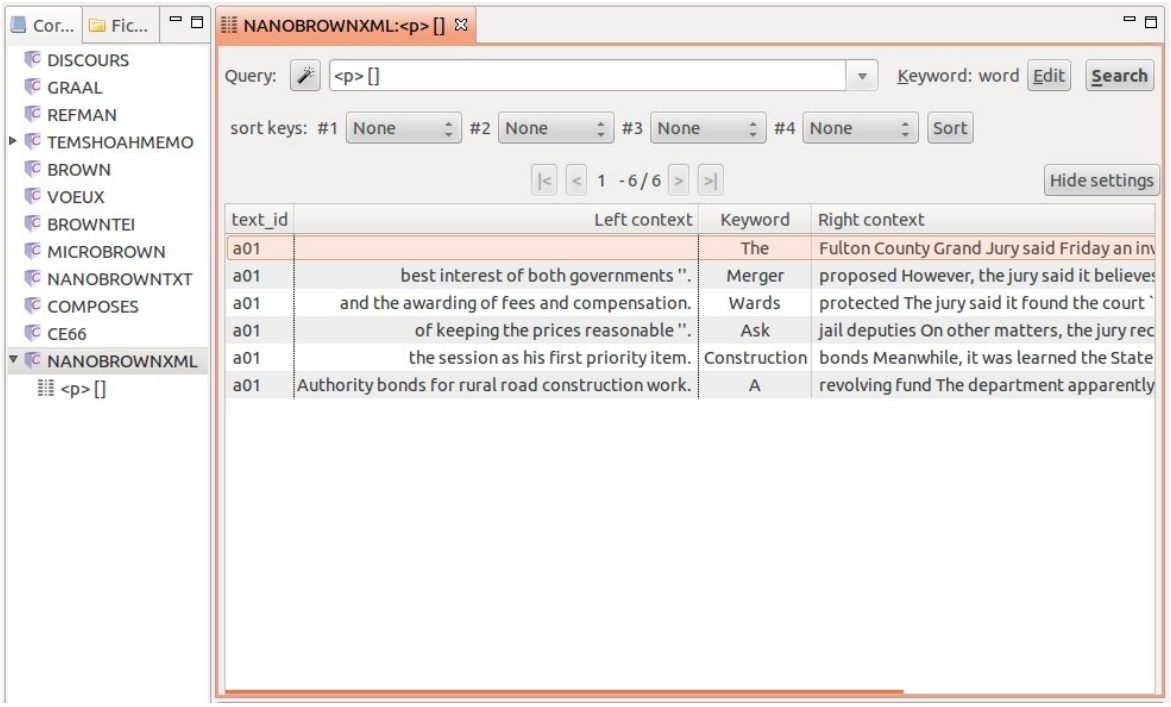
a01.xml

```
<doc>
<p n="1">The Fulton County Grand Jury said Friday an investigation of Atlanta's
recent primary election produced `` no evidence '' that any irregularities took
place.
The jury further said in term-end presentments that the <w t="institution">City
Executive Committee</w>, which had over-all charge of the election, `` deserves
the praise and thanks of the City of Atlanta '' for the manner in which the
election was conducted.
The September-October term jury had been charged by <w t="institution">Fulton
Superior Court</w> Judge Durwood Pye to investigate reports of possible ``
irregularities '' in the hard-fought primary which was won by Mayor-nominate
Ivan Allen Jr..
`` Only a relative handful of such reports was received '', the jury said, ``
considering the widespread interest in the election, the number of voters and
the size of this city ''.
The jury said it did find that many of Georgia's registration and election laws
`` are outmoded or inadequate and often ambiguous ''.
It recommended that Fulton legislators act `` to have these laws studied and
revised to the end of modernizing and improving them ''.
The grand jury commented on a number of other topics, among them the Atlanta and
Fulton County purchasing departments which it said `` are well operated and
follow generally accepted practices which inure to the best interest of both
governments ''.
</p>
<p n="2">Merger proposed
However, the jury said it believes `` these two offices should be combined to
achieve greater efficiency and reduce the cost of administration ''.
The City Purchasing Department, the jury said, `` is lacking in experienced
```

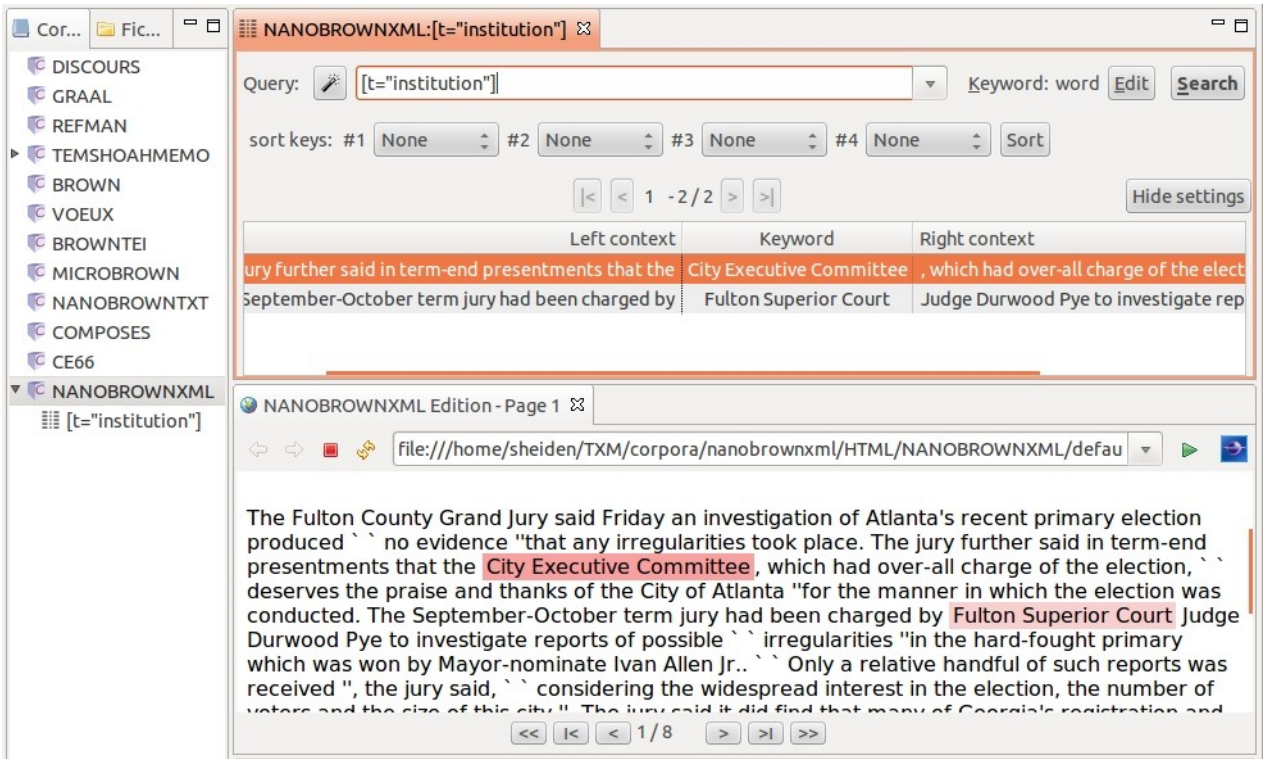
– we import the corpus into TXM by indicating the “nanobrownxml” source directory to the “File / Import / XML/w+CSV” import command:



– now, with the NANOBROWNXML corpus, we can do a concordance of all the first words of paragraphs by searching for <p> []:

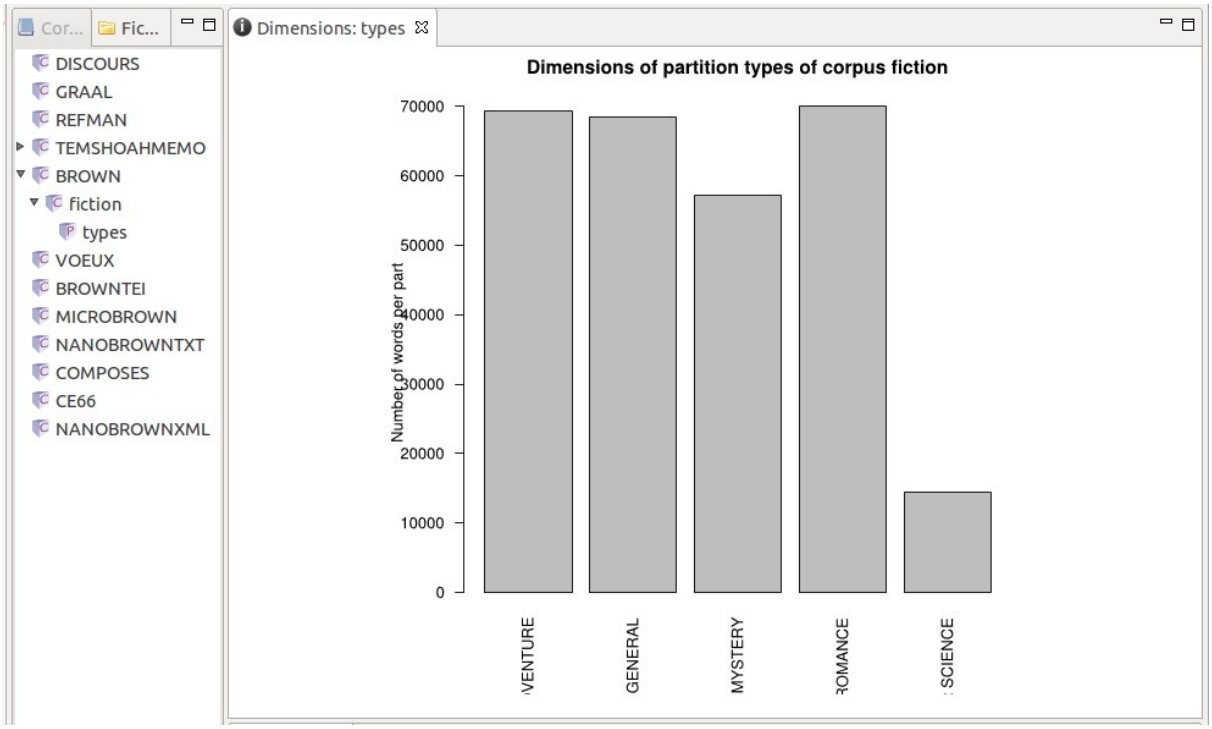


– or search for words with “t” property set to “institution”: [t=”institution”]

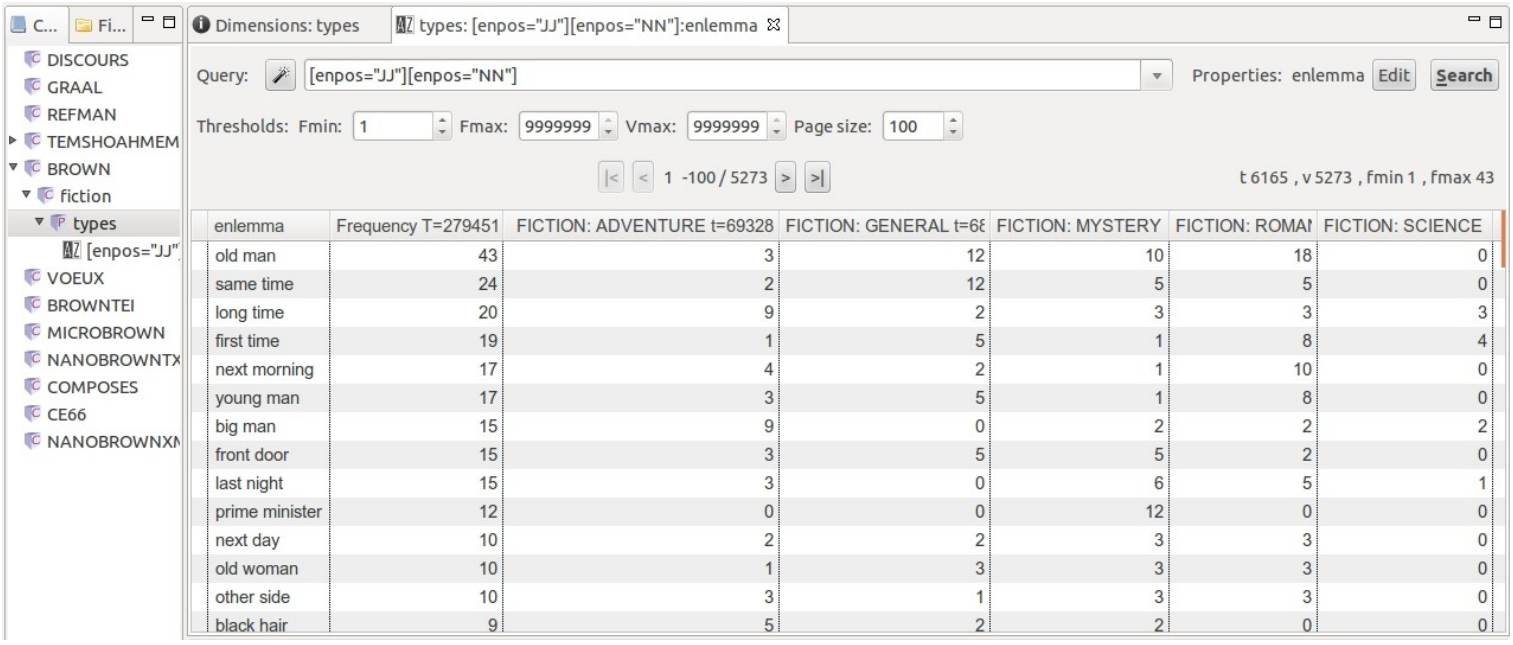


Comparing fictional text types by specific patterns analysis

- to work exclusively on fictional texts, we need first to build a sub-corpus: apply the Sub-corpus command on the Brown corpus. In the parameters dialog box:
  - give a name to the sub-corpus, for example “fiction”;
  - select the “type” property of the “text” structure
  - then select all the text types beginning with “FICTION:...” with control-click on each (from “FICTION:ADVENTURE” to “FICTION:ROMANCE”);
  - click on OK → a new fiction sub-corpus, descendant of Brown, is added to the Corpus view ;
- to compare fiction text types, we need to build a partition: apply the Partition command on the fiction sub-corpus. In the parameters dialog box:
- give a name to the partition, for example “types”;
- select the “type” property of the “text” structure
- click on OK → a new types partition, descendant of fiction sub-corpus, is added to the Corpus view ;
- we can check part size by applying the Description tool on the types partition:

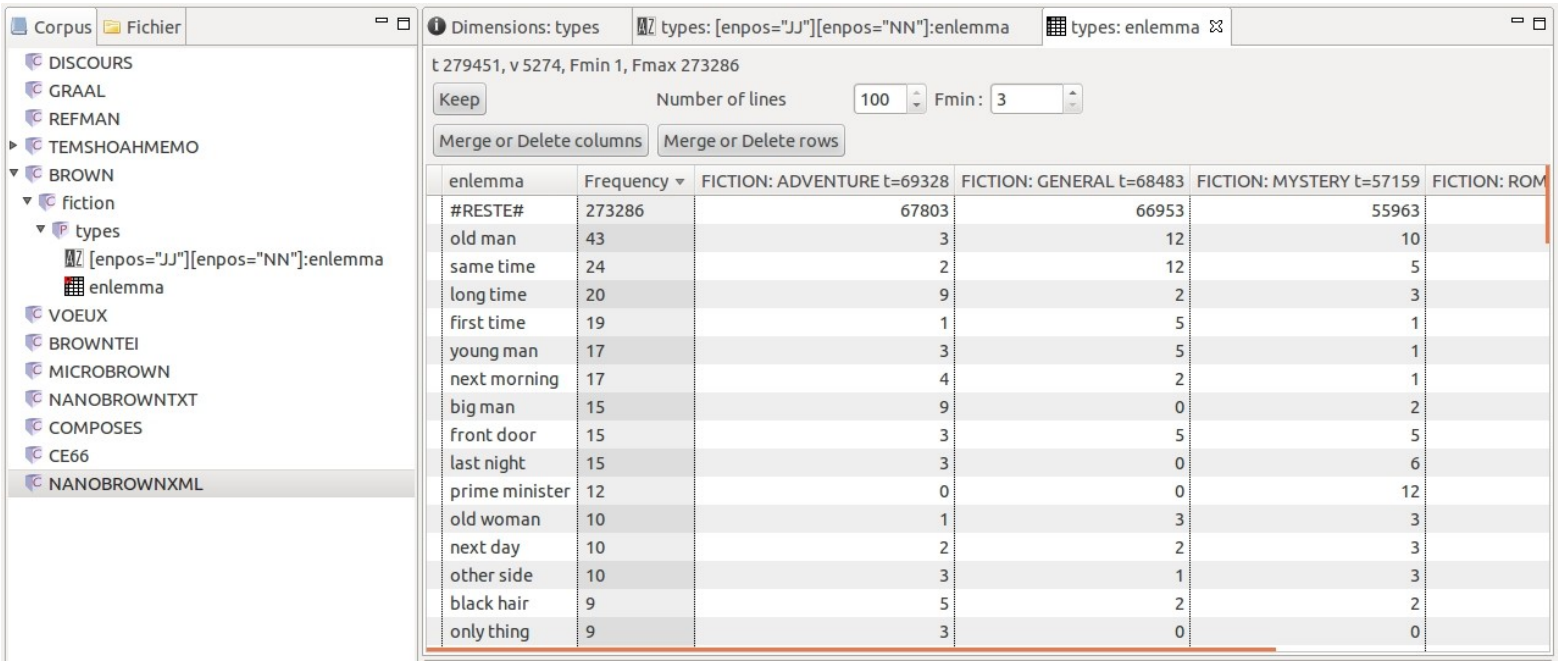


– to compare the repartition of the “an ADJECTIVE immediately followed by a NOUN” sequence pattern between text types, let's use the Index command on the types partition. For this, follow the same operations as in the **Example 2** of the “The Query Assistant and the CQL query language” section above except that you will work on the types partition instead of the Brown corpus:



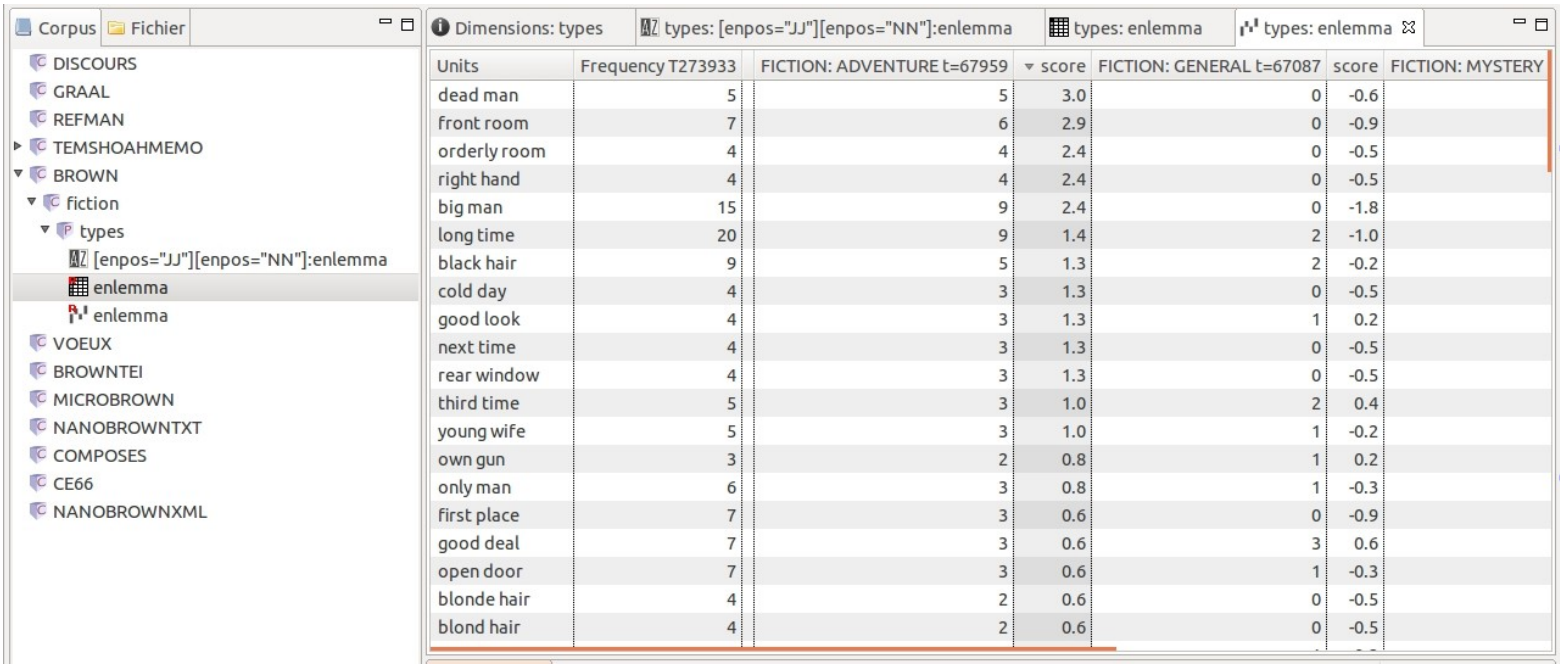
enlemma	Frequency T=279451	FICTION: ADVENTURE t=69328	FICTION: GENERAL t=68	FICTION: MYSTERY	FICTION: ROMAI	FICTION: SCIENCE
old man	43	3	12	10	18	0
same time	24	2	12	5	5	0
long time	20	9	2	3	3	3
first time	19	1	5	1	8	4
next morning	17	4	2	1	10	0
young man	17	3	5	1	8	0
big man	15	9	0	2	2	2
front door	15	3	5	5	2	0
last night	15	3	0	6	5	1
prime minister	12	0	0	12	0	0
next day	10	2	2	3	3	0
old woman	10	1	3	3	3	0
other side	10	3	1	3	3	0
black hair	9	5	2	2	0	0

– in order to use the *specificity* statistical model to analyze the specificity of sequences, we need to first transform the index in a lexical table: apply the LexicalTable command to the [enpos="JJ"][enpos="NN"]:enlemma index, select “Use all occurrences” for margins and “Keep” the 100 most frequent sequences:



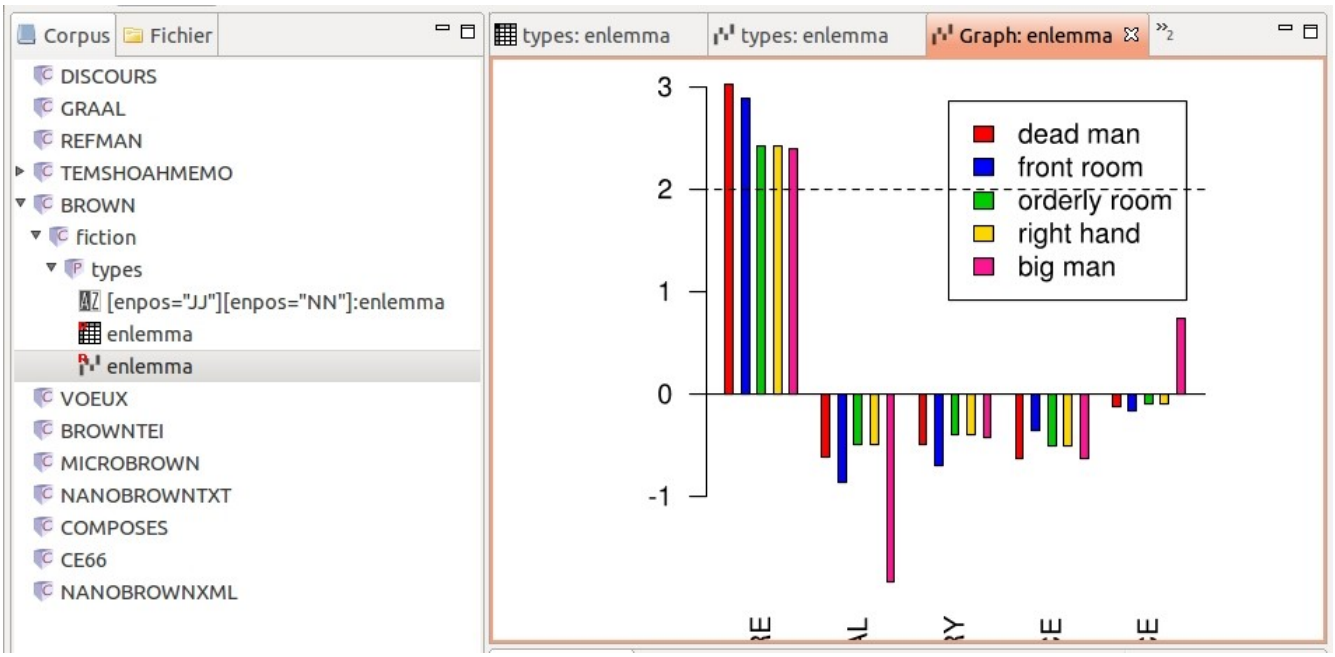
enlemma	Frequency	FICTION: ADVENTURE t=69328	FICTION: GENERAL t=68483	FICTION: MYSTERY t=57159	FICTION: ROM
#RESTE#	273286	67803	66953	55963	
old man	43	3	12	10	
same time	24	2	12	5	
long time	20	9	2	3	
first time	19	1	5	1	
young man	17	3	5	1	
next morning	17	4	2	1	
big man	15	9	0	2	
front door	15	3	5	5	
last night	15	3	0	6	
prime minister	12	0	0	12	
old woman	10	1	3	3	
next day	10	2	2	3	
other side	10	3	1	3	
black hair	9	5	2	2	
only thing	9	3	0	0	

– now we can calculate the specific sequences of the FICTION:ADVENTURE type: run the Specificities command on the enlemma lexical table and sort it descending on the “FICTION:ADVENTURE” specificity score column:



Units	Frequency T273933	FICTION: ADVENTURE t=67959	score	FICTION: GENERAL t=67087	score	FICTION: MYSTERY
dead man	5	5	3.0	0	-0.6	
front room	7	6	2.9	0	-0.9	
orderly room	4	4	2.4	0	-0.5	
right hand	4	4	2.4	0	-0.5	
big man	15	9	2.4	0	-1.8	
long time	20	9	1.4	2	-1.0	
black hair	9	5	1.3	2	-0.2	
cold day	4	3	1.3	0	-0.5	
good look	4	3	1.3	1	0.2	
next time	4	3	1.3	0	-0.5	
rear window	4	3	1.3	0	-0.5	
third time	5	3	1.0	2	0.4	
young wife	5	3	1.0	1	-0.2	
own gun	3	2	0.8	1	0.2	
only man	6	3	0.8	1	-0.3	
first place	7	3	0.6	0	-0.9	
good deal	7	3	0.6	3	0.6	
open door	7	3	0.6	1	-0.3	
blonde hair	4	2	0.6	0	-0.5	
blond hair	4	2	0.6	0	-0.5	

– we can get a general overview of the specificity scores for the first five sequences for each text type by drawing their bar graph: in the types:enlemma specificity table, select the first five lines with the mouse and right-click to call the “Histogram”:



Using TXM objects in R scripts

Let's display the bar plot of <ADJ NOUN> sequence frequencies for the “FICTION:ADVENTURE” text type:

– first, switch to the “R perspective” by clicking on the “R perspective” button of the perspective toolbar (or from the “View / Perspectives / R” menu item) and organize the display to superpose the Corpus and the “R variables” views;

- then transfer the <ADJ NOUN> sequence index by text types to R: right-click on the "[enpos="JJ"][enpos="NN"]:enlemma" index and run the Send to R command;
- open a new R session script by clicking on the "New session" button in the toolbar;
- write the following script in the session1.R script:

```
svg("/tmp/test.svg")
barplot(t(Index1$data), space=c(1,35), horiz=F, las=2, beside=T)
dev.off()
```

- and execute it by clicking on the "Submit" button;
- an SVG file is produced and you can open and display it in TXM from the File view: