The Accolades Project on Large Textual Corpora - Working Paper

The content of this Working Paper was first presented at the University of Edinburgh
19th - 22nd May 2010
thanks to a network research meeting financed by a grant from:
The British Academy, UK and the Centre National de la Recherche Scientifique, France under their scheme: Grants for British-French Joint Projects
Universities in this Project: Edinburgh, Liverpool, Sheffield, Nancy.

Author: Charlie Mansfield

To cite:
Mansfield, Charlie (2010) The Accolades Project on Large Textual Corpora - Working Paper, Edinburgh Research Archive (ERA), Edinburgh.

Contents

Defining the Research & Software Development

Through my software development and computer programming in PHP and JavaScript I continue to develop analysis tools for use on huge corpora of text (typically 250 000 to 400 000 words). These corpora include longer literary texts and large files of written web content. By writing new search tools using the PHP scripting language this part of my on-going research will aim to automate the identification of instances of space and movement in the content.

The applied aspect of the new software created will be the development of this web-tool for use in an advisory or consultancy capacity for specialists in textual analysis and for professionals concerned with the management of urban destinations in the tourism industry. It is particularly aimed at town planning officers and tourist managers who wish to develop tourism using their literary and built heritage.

My software development takes as a starting point the concerns of academic researchers working with corpora of Middle French texts. These transcribed texts, that is re-keyed texts, are often about to be tagged or encoded with XML, the extensible markup language. One of the key aims of researchers in this field is finding ways of attaching their knowledge of a word or phrase to the item itself so that later interrogations of the large text file will return a

more complete presentation of the knowledge stored during the painstaking process of encoding (sometimes called tagging or marking-up). I have been a research engineer on two projects to digitise, transcribe and encode large Middle French texts; this type of project is to protect and at the same time make available precious heritage artefacts which are too fragile for constant use. In tourism and heritage studies this digitisation of artefacts is in the new field of e-Tourism.

LAMP Technologies for Interactivity

The list of web-page and server technologies referred to as L.A.M.P. have been deployed since the beginning of 2000 to develop more responsive and interactive systems for extending the use of the web from simply displaying content keyed-in by the server owner without any control by the user or viewer of the web-page. My Accolades Project draws on these technologies to provide a web-based tool for the user interested in analysing a large corpus of text by testing their own hypotheses concerning some aspect of the text.

A Preliminary Case Study and Trial

By looking at the work of palaeographers and those interested in historical language use a trial can be set up where the specialists already have a set of requirements from any analytical tool. In particular, two groups of researchers are valuable, language researchers who are looking at diachronic shifts, that is, the changes in orthography over a certain time period, and, those interested in detecting what are called hand-changes in a large corpus to detect the orthographic signature (spelling style) of individual scribes.

I completed the design and coding of a sub-system of my Accolades tool-kit with a range of PHP programs called Loceme. They are stored and may be launched on a Unix/Linux Apache server at this location (web address)
http://eserve.org.uk/loceme/

The functional component of the PHP code uses these two instructions to find string-within-string matches in each line of text and then replace them with the same text but highlighted in a colour scheme to provide visual clues to the user about their hypotheses:

strstr( )
str_replace()

The first of the two experiments will be published as:
Anthony Lodge (2011 forthcoming) Variation and change in the Montferrand account-books (1259-1367). In: Yuji Kawaguchi, Makoto Minegishi, Wolfgang Viereck (eds), *Corpus Analysis and Diachronic Linguistics*, Amsterdam/Philadelphia : John Benjamins.

The key to this trial is that the PHP is written to handle the file whether it contains XML tags or not. The only parameters required are that it is in UTF-8 encoding and that the lines terminate with CR/LF (carriage return, line feed). This position, of working with untagged, non-XML texts stems from the consideration that these researchers in linguistics may wish to run a series of experiments on the keyed-in corpus before they undertake the painstaking task of encoding certain features of the text with XML elements and their own specialised notes.

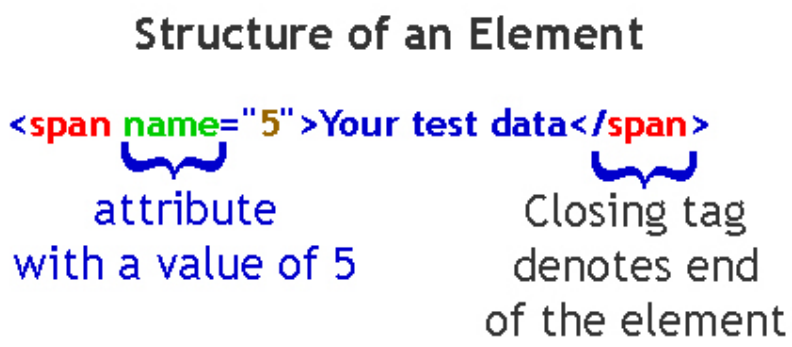


Figure 1 – Structure of an XML Element (tag)

SIFT in JavaScript (ECMAScript)

Mansfield had written an earlier tool for trying out concordance-type searches on a large text in JavaScript, called SIFT (Script Informatisé pour Faire le Tamisage). A working copy of this is held and available for free use at the UK Universities LLAS Subject Centre at this url http://www.llas.ac.uk/materialsbank/mb082/zola.htm

SIFT is documented at ERA (Edinburgh Research Archive) at url
http://hdl.handle.net/1842/1983
Mansfield, Charlie (2007) Modéliser l'espace dans les textes en moyen français : le développement d'un script informatisé pour faire le tamisage (SIFT), ERA, Edinburgh.

A web browser capable of reading the Document Object Model (DOM) is required to make use of SIFT, Mozilla Firefox is capable of this.

The purpose of SIFT is to determine what can be discovered in an un-tagged text and so offers a preliminary step for researchers before they begin adding XML tags, and can even help in deciding which tags and what type of attributes may be needed to provide the later display or analysis required by the researchers.

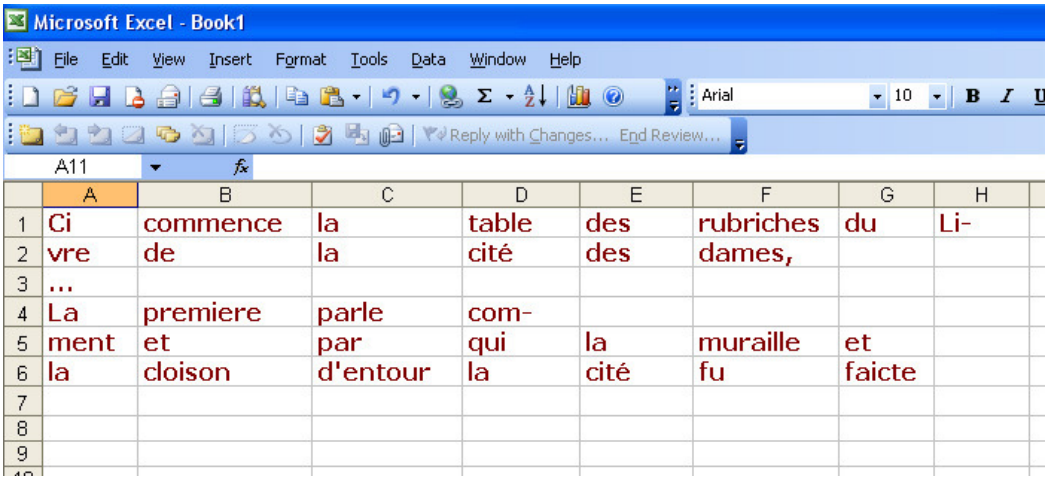
SIFT is limited by the size of the text that the web browser can handle in a single session. If the users follow these instructions then the browser can handle the complete text of a novel the size of Zola's *L'Assommoir*:

```
Type (or copy & paste) the following into your Firefox location bar
(where you type in the web-address or url) and press Enter:

about:config

Now type or copy & paste the following into the Filter textbox:

dom.max_script_run_time

Now double-click on the line displaying

dom.max_script_run_time
```

```
Change the value to the number of seconds you want Firefox to wait
before time-out.
Use 120 (2 minutes) initially. Just enter the number 120

Now please close the about:config tab and then restart your Firefox
browser. The new setting will become effective.
```

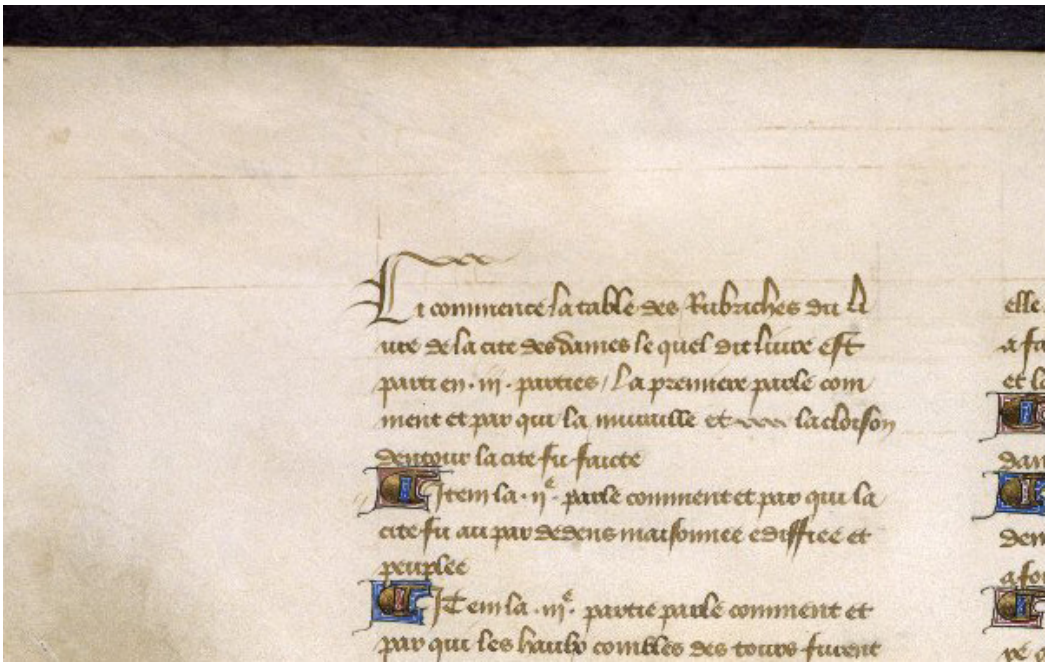
An Account of XML Design for Literature Corpora

The Montferrand account-books of Professor Lodge's work remind the literary scholar, who is planning to use XML, that in our post-Norman, tax-recording culture keeping financial accounts was a parallel activity to giving accounts of historical events; *un conte* and *un compte* are quite different to the modern mind. The story, *le conte*, belongs to the word processor and the account belongs to *le tableur* or spreadsheet. However, the passage of time represented by the unrolling of the text is an aspect shared by the two discourses, accounting and recounting. In realist fiction we expect the chronology of the text to unfold or unroll the further we read, in the account books the passing months and tax years are recorded or journalled to represent passing time. In approaching bound collections of writing, for example the Harley MS 4431 collection of thirty works of Christine de Pizan in the British Library it is tempting to believe initially that these are in chronological order, too; it is part of our acculturation. In approaching XML design, though, as a researcher attempting to discover a new aspect of the literary text, we attempt to put aside this eleventh-century acculturation. A first step may be to put the literary text into a spreadsheet (*un tableur*):

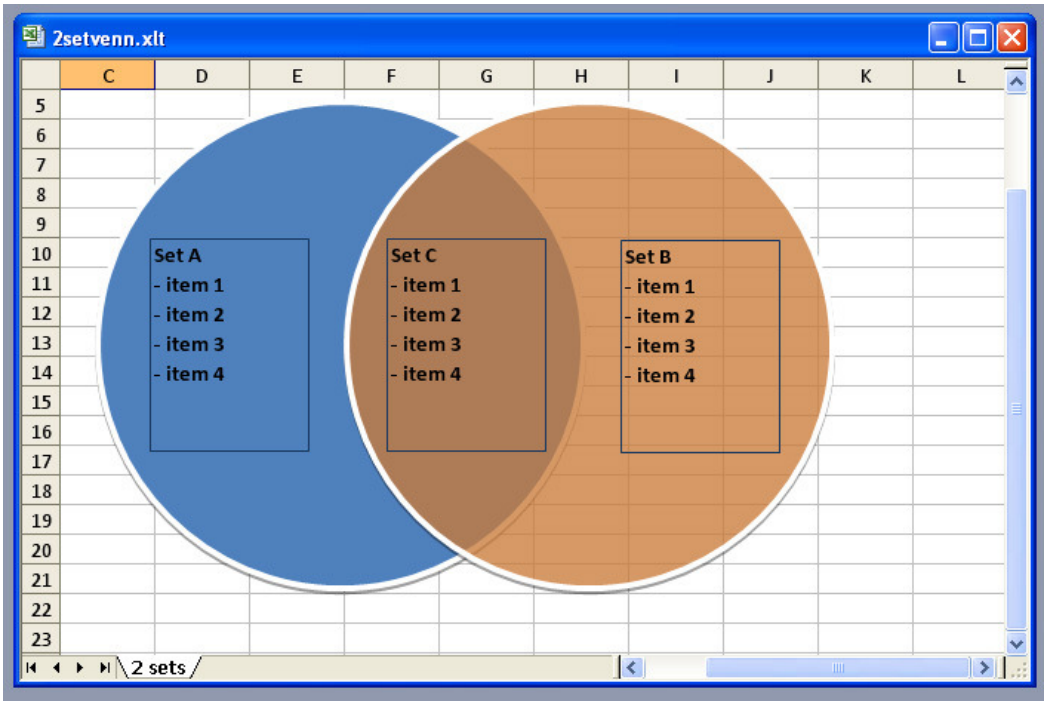


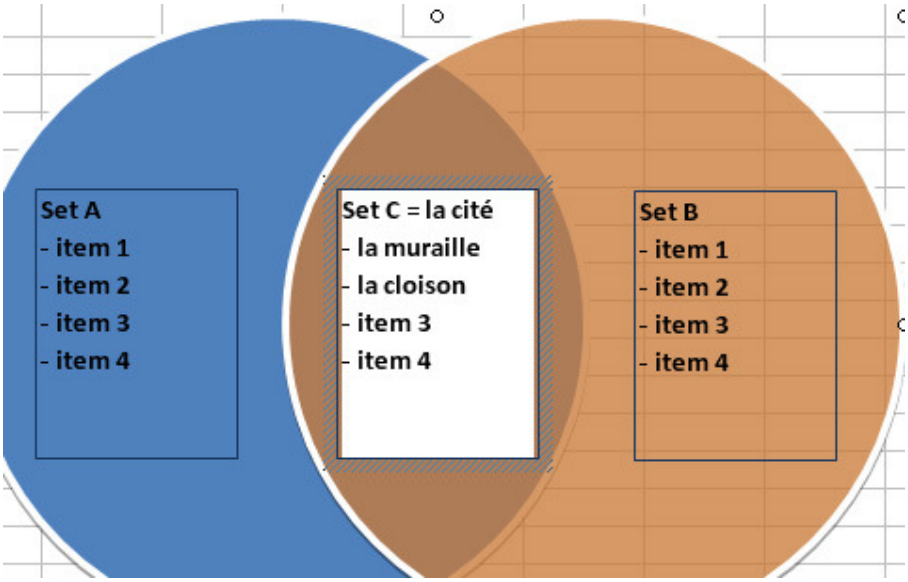
	A	B	C	D	E	F	G	H
1	Ci	commence	la	table	des	rubriches	du	Li-
2	vre	de	la	cité	des	dames,		
3	...							
4	La	premiere	parle	com-				
5	ment	et	par	qui	la	muraille	et	
6	la	cloison	d'entour	la	cité	fu	faicte	
7								
8								
9								

In a very rudimentary way this laying-out of the literary text on a table, like the patient on the table of the clinic, does classify the text into lines. However, this classification has already taken place either by the medieval author in her thinking or by the copyist endeavouring the fit the story into the ruled confines of the manuscript page:



With XML, though, we have a system that offers us a structured way of classifying based on the thinking of Saussure and Lévi-Strauss. As the linguist or literary scholar chooses the XML elements they plan to use and the attributes and codes for these attributes a class structure is at work similar to the set lists of Venn diagrams where items can belong to a class or even to several classes depending on which attributes the scholar uses to classify them.





Set C here could be said to be a set of building components of *la cité*, these include: *la muraille* and *la cloison*. When we begin to turn that analysis into XML tags we have to consider whether we are, in fact, building a structure with windows left that will allows us to see inside, only to find what we have designed or, at worse, to see nothing but empty space.

BL, London, Harley MS 4431 290r La Cité des dames. Christine de Pizan circa 1413.

We may arrive at the same discovery that we could make by reading and understanding the Middle French story and its use of language. For example, by the painstaking classification of the components of *la cité* we begin to understand that for Christine this was not a modern city but an enclosed space.

What are Corpus or Web-Content Researchers Seeking?

My work continues in the Accolades Project initially to make the hypothesis-testing speed of the Loceme software tools easier and easier to use. One method is to attempt to determine what the users are looking for in the large literary corpora or in the large collections of web-content. Accolades is a piece of software that collects the search terms and processes them in a way to detect patterns for the software developer to re-work the interface.

Working Paper
Charlie Mansfield, University of Plymouth, School of Tourism & Hospitality.
c.mansfield AT plymouth.ac.uk