Glimpses though the clouds: collocates in a new light

This paper demonstrates a web-based, interactive data visualisation, allowing users to quickly inspect and browse the collocational relationships present in a corpus. The software is inspired by tag clouds, first popularised by on-line photograph sharing website Flickr (www.flickr.com). A paper based on a prototype of this Collocate Cloud visualisation was given at Digital Resources for the Humanities and Arts 2007. The software has since matured, offering new ways of navigating and inspecting the source data. It has also been expanded to analyse additional corpora, such as the British National Corpus (http://www.natcorp.ox.ac.uk/), which will be the focus of this talk.

Tag clouds allow the user to browse, rather than search for specific pieces of information. Flickr encourages its users to add tags (keywords) to each photograph uploaded. The tags associated with each individual photograph are aggregated; the most frequent go on to make the cloud. The cloud consists of these tags presented in alphabetical order, with their frequency displayed as variation in colour, or more commonly font size. Figure 1 is an example of the most popular tags at Flickr:

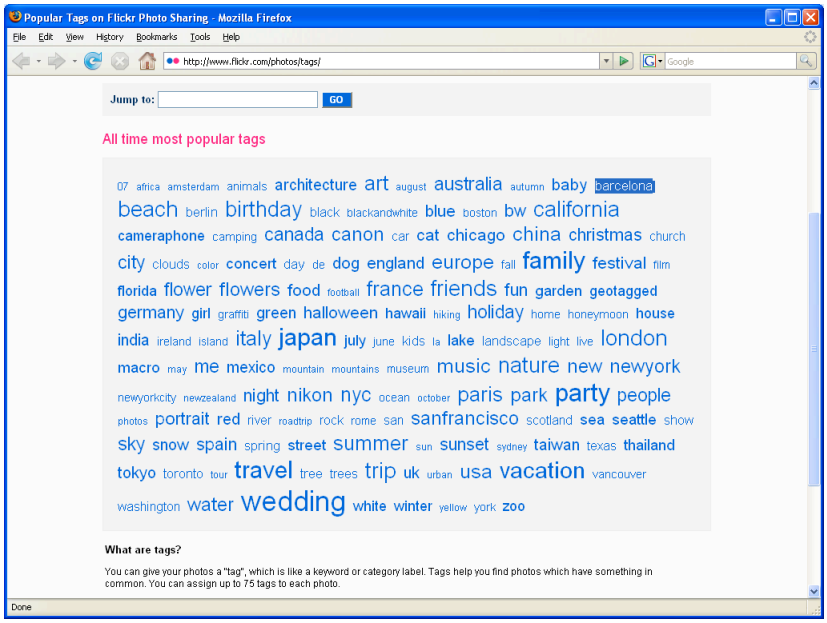

Figure 1. Flickr tag cloud showing 125 of the most popular photograph keywords http://www.flickr.com/photos/tags/ (accessed 23 November 2007)

The cloud offers two ways to access the information. If the user is looking for a specific term, the alphabetical ordering of the information allows it to be quickly located if present. More importantly, as a tool for browsing, frequent tags stand out visually, giving the user an immediate overview of the data. Clicking on a tag name will display all photographs which contain that tag.

The cloud-based visualisation has been successfully applied to language. McMaster University's TAPoR Tools (http://taporware.mcmaster.ca/) features a 'Word Cloud' module, currently in beta testing. WMatrix (http://ucrel.lancs.ac.uk/wmatrix/) can compare two corpora by showing log-likelihood results in cloud form. In addition to other linguistic metrics, internet book seller Amazon provides a word cloud, see figure 2.

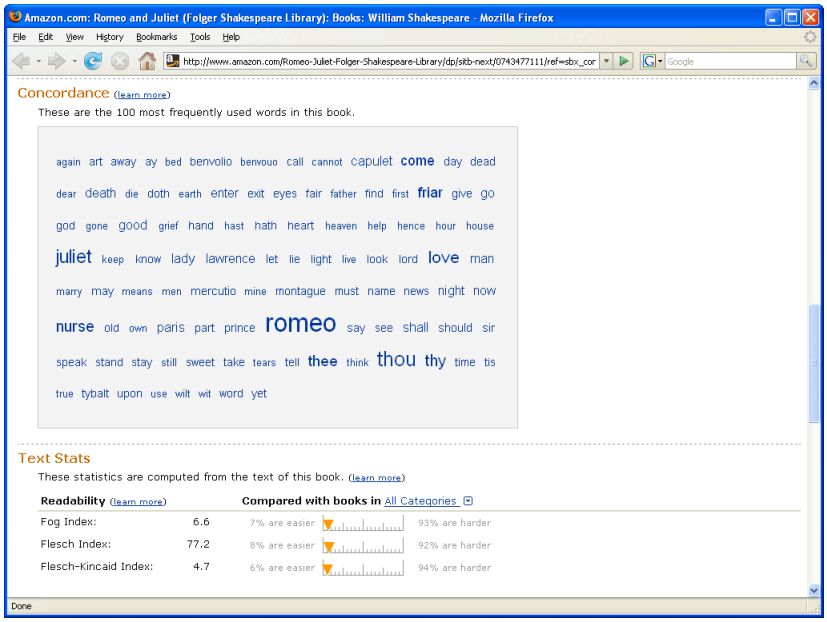

Figure 2. Amazon.com's 'Concordance' displaying the 100 most frequent words in Romeo and Juliet http://www.amazon.com/Romeo-Juliet-Folger-Shakespeare-Library/dp/sitb-next/0743477111/ref=sbx_con/104-4970220-2133519?ie=UTF8&qid=1179135939&sr=1-1#concordance (accessed 23 November 2007)

In this instance a word frequency list is the data source, showing the most frequent 100 words. As with the tag cloud, this list is alphabetically ordered, the font size being proportionate to its frequency of usage. It has all the benefits of a tag cloud; in this instance clicking on a word will produce a concordance of that term.

This method of visualisation and interaction offers another tool for corpus linguists. As developer for an online corpus project, I have found that the usability and sophistication of our tools have been important to our success. Cloud-like displays of information would complement our other advanced features, such as geographic mapping and transcription synchronisation.

The word clouds produced by TAPoR Tools, WMatrix and Amazon are, for browsing, an improvement over tabular statistical information. There is an opportunity for other corpus data to be enhanced by using a cloud. Linguists often use collocational information as a tool to examine language use. Figure 3 demonstrates a typical corpus tool output:

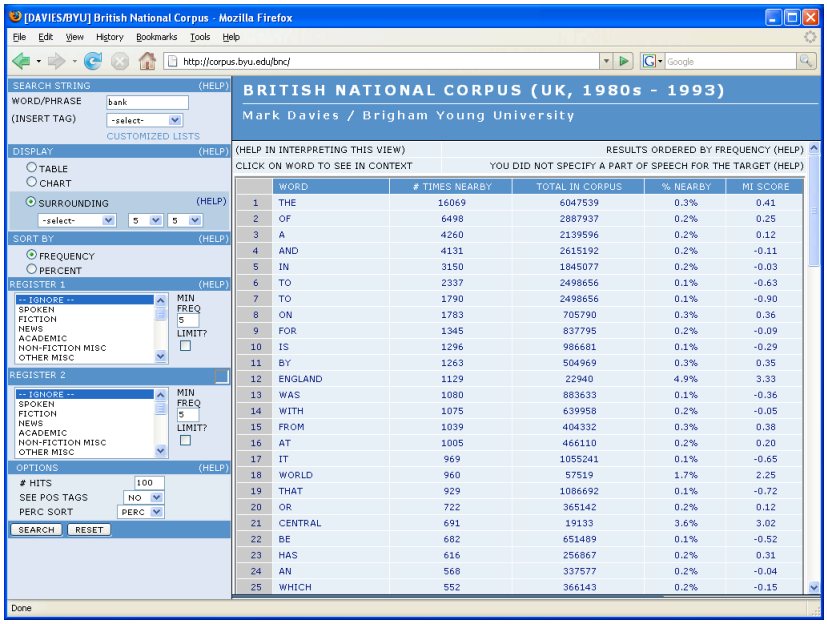

Figure 3. British National Corpus through interface developed by Mark Davies, searching for 'bank', showing collocates http://corpus.byu.edu/bnc/ (accessed 23 November 2007)

The data contained in the table lends itself to visualisation as a cloud. As with the word cloud, the list of collocates can be displayed alphabetically. Co-occurrence frequency, like word frequency, can be mapped to font size. This would produce an output visually similar to the word cloud. Instead of showing all corpus words, they would be limited to those surrounding the chosen node word.

Another valuable statistic obtainable via collocates is that of collocational strength, the likelihood of two words co-occurring, measured here by MI (Mutual Information). Accounting for this extra dimension requires an additional visual cue to be introduced, one which can convey the continuous data of an MI score. This can be solved by varying the colour, or brightness of the collocates forming the cloud. The end result is shown in figure 4:

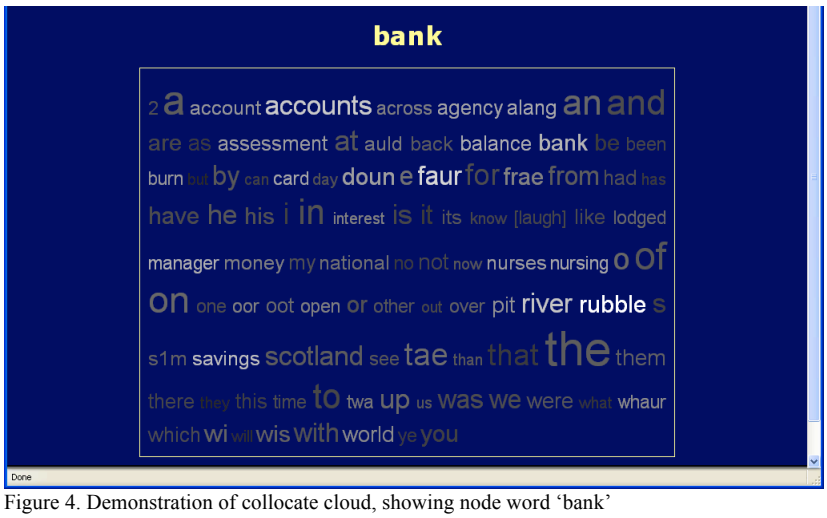

Figure 4. Demonstration of collocate cloud, showing node word 'bank'

The collocate cloud inherits all the advantages of previous cloud visualisations: a collocate, if known, can be quickly located due to the alphabetical nature of the display. Frequently occurring collocates stand out, as they are shown in a larger typeface, with collocationally strong pairings highlighted using brighter formatting. Therefore bright, large collocates are likely to be of interest, whereas dark, small collocates perhaps less so. Hovering the mouse over a collocate will display statistical information, co-occurrence frequency and MI score, as one would find from the tabular view.

The use of collocational data also presents additional possibilities for interaction. A collocate can be clicked upon to produce a new cloud, with the previous collocate as the new node word. This gives endless possibilities for corpus exploration and the investigation of different domains. Occurrences of polysemy can be identified and expanded upon by following the different collocates. Particular instances of usage are traditionally hidden from the user when viewing aggregated data, such as the collocate cloud. The solution is to allow the user to examine the underlying data by producing an optional concordance for each node/collocate pairing present. Additionally a KWIC concordance can be generated by examining the node word, visualising the collocational strength of the surrounding words. These concordance lines can even by reordered on the basis of collocational strength, in addition to the more traditional options of preceding or succeeding words.

This visualisation may be appealing to members of the public, or those seeking a more practical introduction to corpus linguistics. In teaching use they not only provide analysis, but from user feedback, also act as stimulation in creative writing. Collocate searches across different corpora or document sets may be visualised side by side, facilitating quick identification of differences.

While the collocate cloud is not a substitute for raw data, it does provide a fast and convenient way to navigate language. The ability to generate new clouds from existing collocates extends this further. Both this iterative nature and the addition of collocational strength information gives these collocate clouds greater value for linguistic research than previous cloud visualisations.