## Visualising corpus linguistics

*Paul Rayson and John Mariani*

Department of Computing

Lancaster University

*{p.rayson, j.mariani}@lancaster.ac.uk*

1. Introduction

Corpus linguists are not unusual in that they share the problem that every computer user in the world faces – an ever-increasing onslaught of data. It has been said that companies are drowning in data but starving for information. How are we supposed to take the mountain of data and extract nuggets of information? Perhaps corpus linguists have been ahead of the pack, having had large volumes of textual data available for a considerable number of years, but visualisation techniques have not been widely explored within corpus linguistics. In the application of standard corpus linguistics methodology to ever increasingly large datasets, as often collected in the web-as-corpus paradigm (Kilgarriff and Grefenstette, 2003), researchers have to deal with both practical and methodological problems. For example, faced with corpora containing hundreds of millions of words, the analysis of key words, concordance lines and n-gram lists is not practically possible without some pre-filtering, sorting or further selection by applying a cut-off value. This cut-off value is usually chosen for reasons of practicality in order to reduce the large number of lines or entries to a suitable level and represents some sort of compromise made by the researcher. As corpora become even larger, the amount of data rejected starts to impinge on the quality of the analysis that can be carried out.

Information visualisation is one technique that can be usefully applied to support human cognitive analysis of data allowing the discovery of information through sight. Commercially available systems provide the ability to display data in many ways, including as a set of points in a scatterplot or as a graph of links and nodes. There are many different established techniques for visualising large bodies of data and to allow the user to browse, select and explore the data space. Such techniques may allow us to explore the full set of results extracted by corpus tools from very large datasets.

In this paper, we consider two aspects of information visualisation in relation to corpus linguistics. First, visualisation *in* corpus linguistics, i.e. what tools and techniques are currently used. Second, visualisation *of* corpus linguistics, shown through an example of techniques that we propose; by visualisation of the papers presented at the corpus linguistics conferences.

The remainder of this paper begins with an overview of information visualisation techniques in section 2. In section 3, we briefly review current techniques in corpus linguistics that can be seen as using some type of visualisation. Section 4 contains a case study where we apply our proposed key word cloud and dynamic tag cloud techniques to visualise the development of the corpus linguistics conference series. Finally, we conclude in section 5.

2. Information Visualisation

Information Visualisation is one response to the problems of information overload. There are so many sources of data in the modern world that this means we are simply swamped by it. Businesses have the problem that they are engulfed by a mountain of data which may well contain nuggets of useful information, if only they could get their hands on it. Information Visualisation is one approach in the Data Mining toolbox; a way of processing data in order to extract useful information, normally in the form of trends and/or patterns.

The problem is that of transforming data into information. If we can present that data in a useful way, then people can detect and extract information from it by spotting patterns in the presentation that normally wouldn’t be visible. Unfortunately, this is often not that simple.

An early example was the application of StarField (Ahlberg and Shneiderman, 1999) to estate agent property data. This allowed users to dynamically interact with the data to manipulate the display in order to find houses that suited, and perhaps more importantly, nearly suited their needs.

Visualisation did not begin with the Computer age, however. In the 1850’s, Florence Nightingale invented the pie chart in order to present mortality statistics. Napoleon’s mapmaker, M. Minard, produced a graphic showing Napoleon’s march and subsequent retreat from Moscow, shown in figure 1.

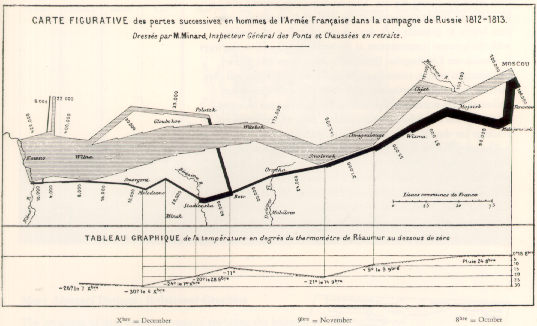


Figure 1: Napoleon’s March. The thickness of the line shows the number of soldiers, the colour the temperature, the time and the route itself.

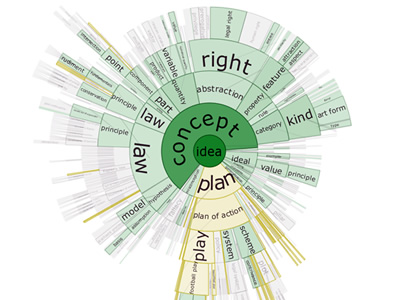


Figure 2: DocuBurst

We now turn to three modern examples of visualisation systems. DocuBurst (Collins, 2006) provides a front end to WordNet. To use the system, the user loads a document into the visualisation tool and chooses a WordNet node to root the visualisation. Here in figure 2, we see “idea” was chosen as root, and the concepts that fall under “idea” appear in the visualisation. The gold colour nodes show the results of searching for nodes that begin with “pl”.

The technique used is a radial, space-filling layout of hyponymy (IS-A relationship). Alongside we have the interactive controls for zoom, filter and details-on-demand as shown in Shniderman’s work. With regard to the latter, we can drill down into the document. When a node is selected, a fish-eye view of the paragraphs show which paragraphs contain the text of the node. Selecting a paragraph causes the paragraph to be displayed, with highlighted occurrences of the node’s value.

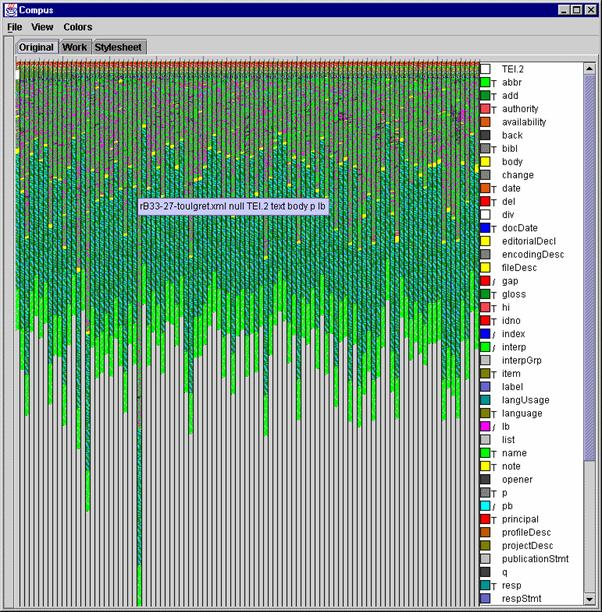


Figure 3: Compus

In the Compus visualisation (Fekete, 2006) shown in figure 3, each vertical bar represents an XML document. Each XML element is associated with a colour. The length of a colour within a bar is derived from the length of the element in the document. Finally, colours are allocated within a bar by a space-filling algorithm.

Finally in this section, we look at the Many Eyes website provided by IBM. (<http://manyeyes.alphaworks.ibm.com/manyeyes/>). This site aims to promote a community of visualisation users and creators. Within the community, you can browse existing visualisations, upload your own data for visualisation, and add visualisations of your own. Some examples of these are shown in figure 4.

|  |  |
| --- | --- |
| Picture 12.png | Picture 13.png |
| Picture 16.png | Figure 4: Some word-based visualisations from Many Eyes |

3. Techniques already used in corpus linguistics

Although they are not necessarily viewed as such, some existing techniques in corpus linguistics can be considered as visualisation. First and foremost, the concordance view with one word aligned vertically in the middle of the text and the left and right context justified in the middle, is a way of visualising the patterns of the context of a particular word. By sorting the right and left context, we can more easily *see* the repeated patterns.

Concgrams (Cheng et al, 2006) takes this visualisation one step further by automatically highlighting repeated patterns in the surrounding context, as shown in figure 5.



Figure 5: Concgrams

Another tool in the corpus toolbox is collocation, and Beavan (2008) has explored visualisation techniques already. By taking the collocates of a word, ordering them alphabetically and altering the font size and brightness, the collocate cloud shown in figure 6 allows us to see collocates in a new light. Here, font size is linked to frequency of the collocate and brightness shows the MI score. In this way, we can easily see the large and bright words that are frequent with strong collocation affinity.

Also, in the area of collocations, McEnery (2006: 21-24) employs a visualisation technique to draw collocational networks. These show key words that are linked by common collocates, as shown in figure 7. McEnery’s work is influenced by Phillips (1985) who uses similar diagrams to study the structure of text.

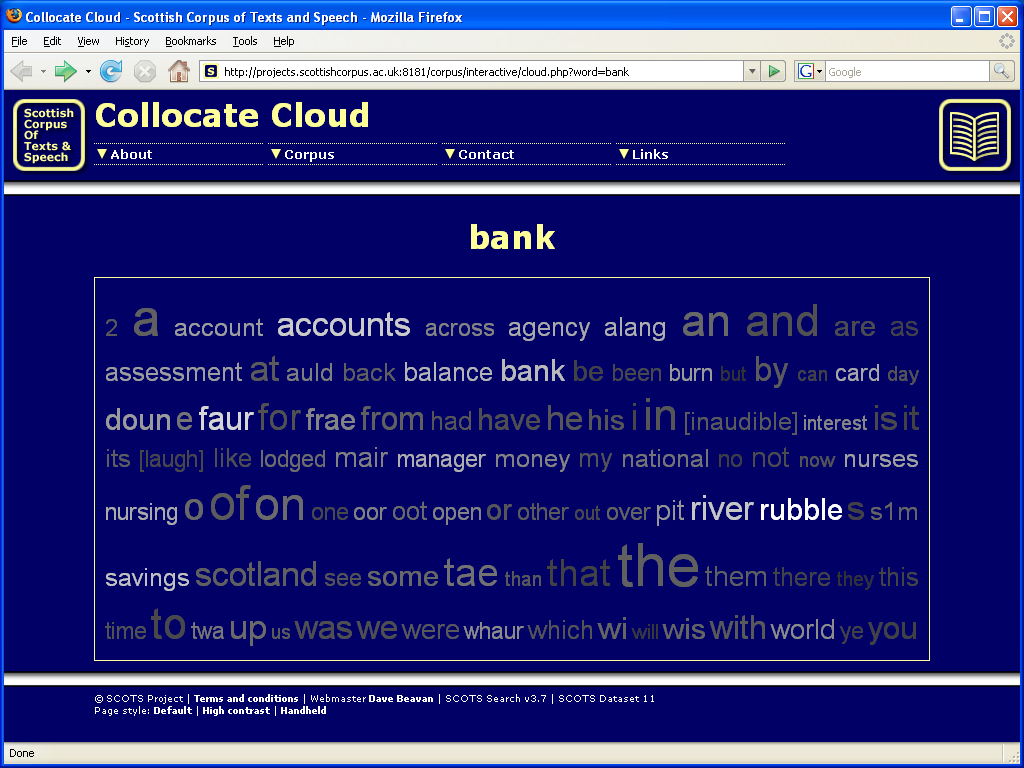


Figure 6: Collocate cloud

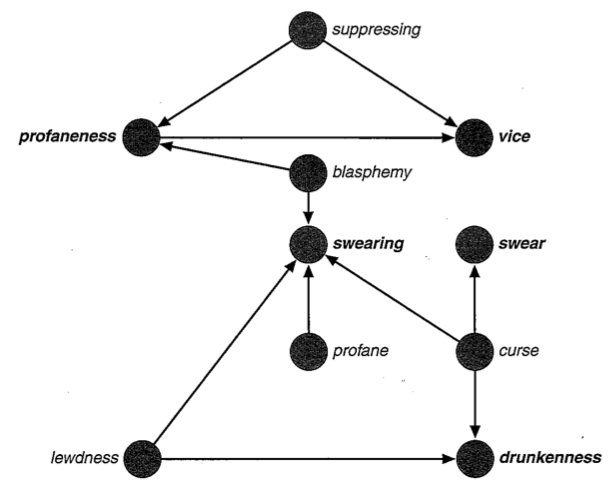


Figure 7: Collocational networks

4. Proposals for new techniques: key word clouds and dynamic tag clouds

In the previous section, we presented the techniques currently used in corpus linguistics which go some way towards information visualisation. In this section, we propose a technique for visualisation of key words results both statically and dynamically. The key words technique (Scott, 1997) is well known in corpus linguistics to users of WordSmith and other tools. By comparing one corpus or text to a much larger reference corpus, we can extract those words that occur with unusual frequency in our corpus relative to a general level of expectation. A keyness metric, usually chi-squared or log-likelihood is calculated for each word to show how ‘unexpected’ its frequency is in the corpus relative to the reference corpus. By sorting on this value we can order the words by their keyness and see the most ‘key’ words at the top of a table. The Wmatrix software (Rayson, 2008) includes a visualisation of the key words results in a ‘key word cloud’. Influenced by tag clouds in social networking and other sites such as Flickr, where the frequency of a word is mapped to its font size, the key word cloud maps the keyness value onto font size. By doing so, we can quickly ‘gist’ a document by viewing the words in the key word cloud. This also avoids the need to use a stop word list to filter out the most frequent closed class words. By way of an example, we describe here a case study using data drawn from the proceedings of the Corpus Linguistics conference series. Through this example, we show the key word cloud visualisation in practice.

Prior to the CL2009 conference in Liverpool, there have been four Corpus Linguistics conferences, occurring bi-annually since 2001. All the proceedings are online1, and are therefore amenable to machine processing. For this case study, we created a key word cloud for each conference. The procedure applied for each set of conference proceedings was:

1. Extract paper titles and authors
2. Create a word frequency list
3. Apply the keyness measure to extract key words by comparing the frequency list to a standard written reference (BNC written sampler)
4. Generate a key word cloud from the key word list

The resulting four tag clouds are shown in figures 8 – 11 below.

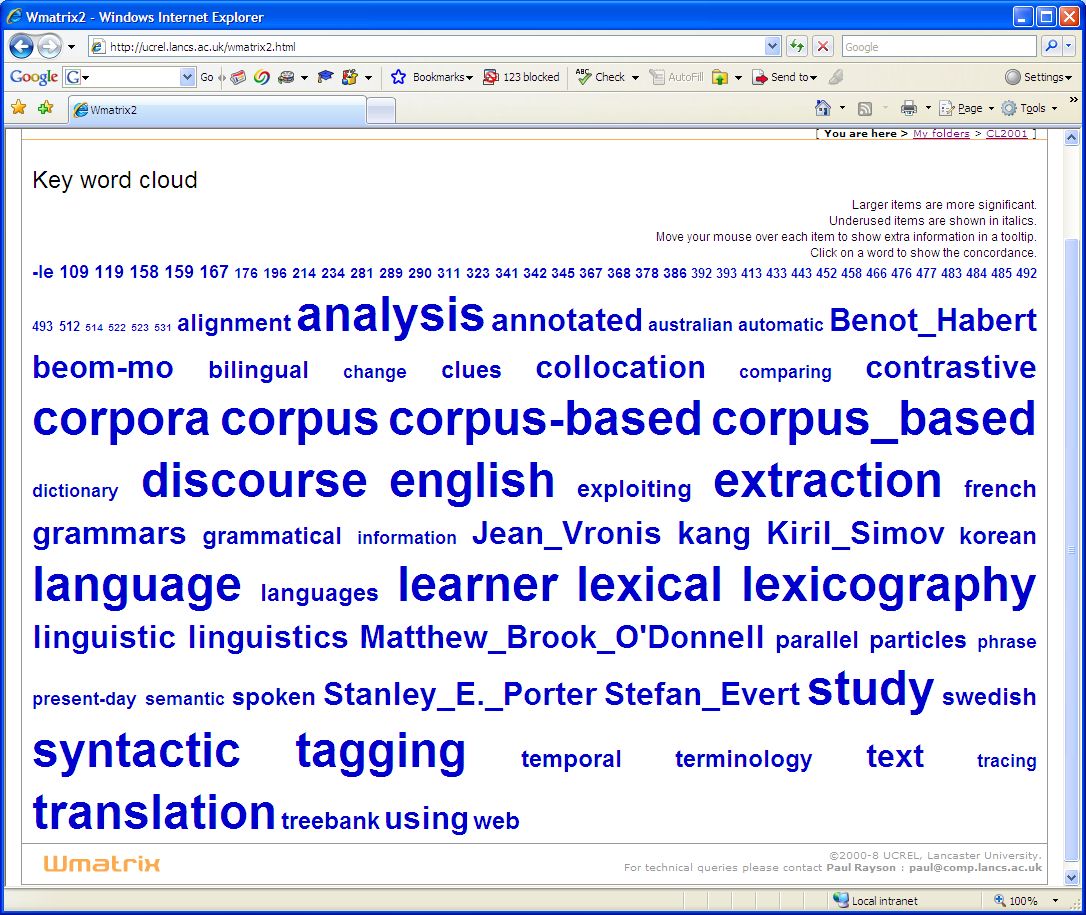


Figure 8: CL2001 key word cloud

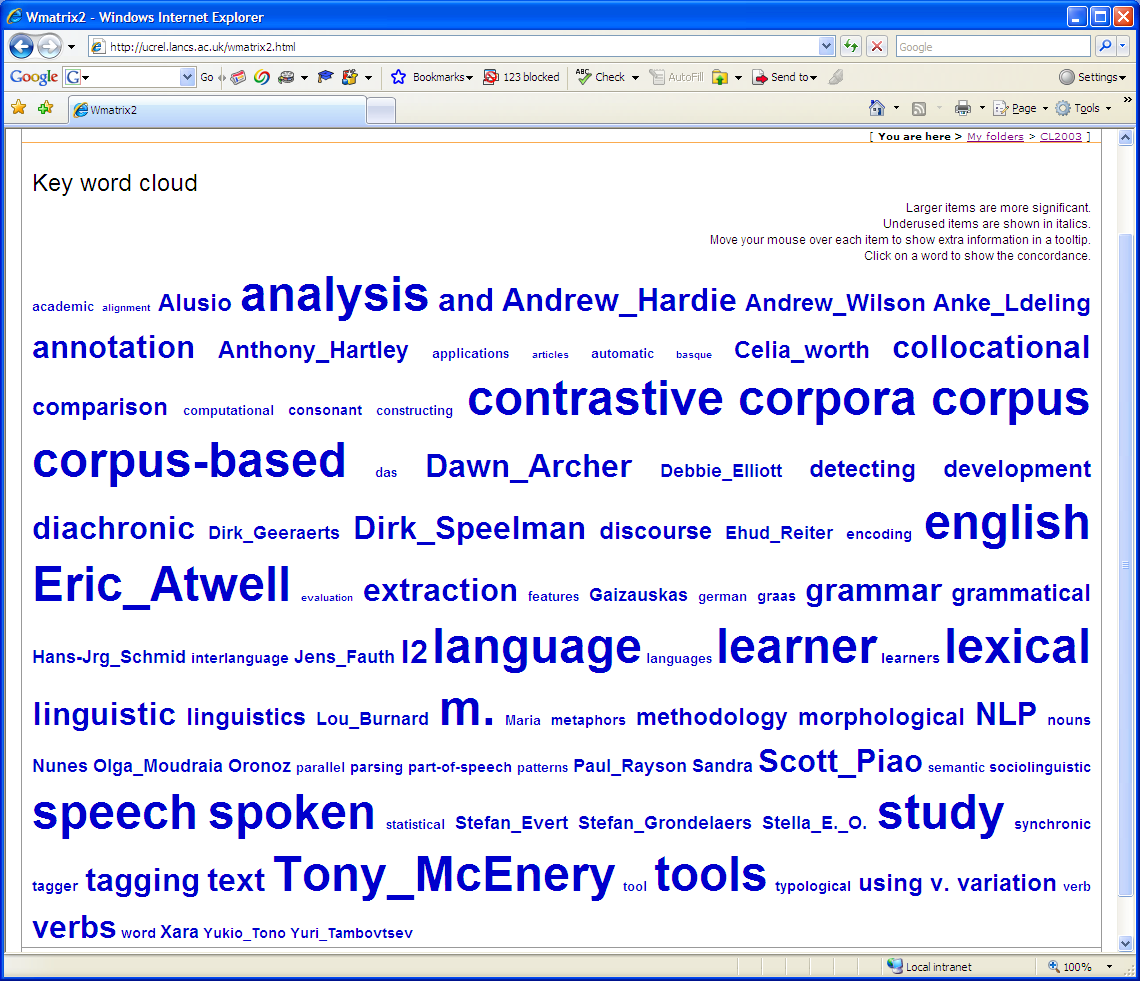


Figure 9: CL2003 key word cloud

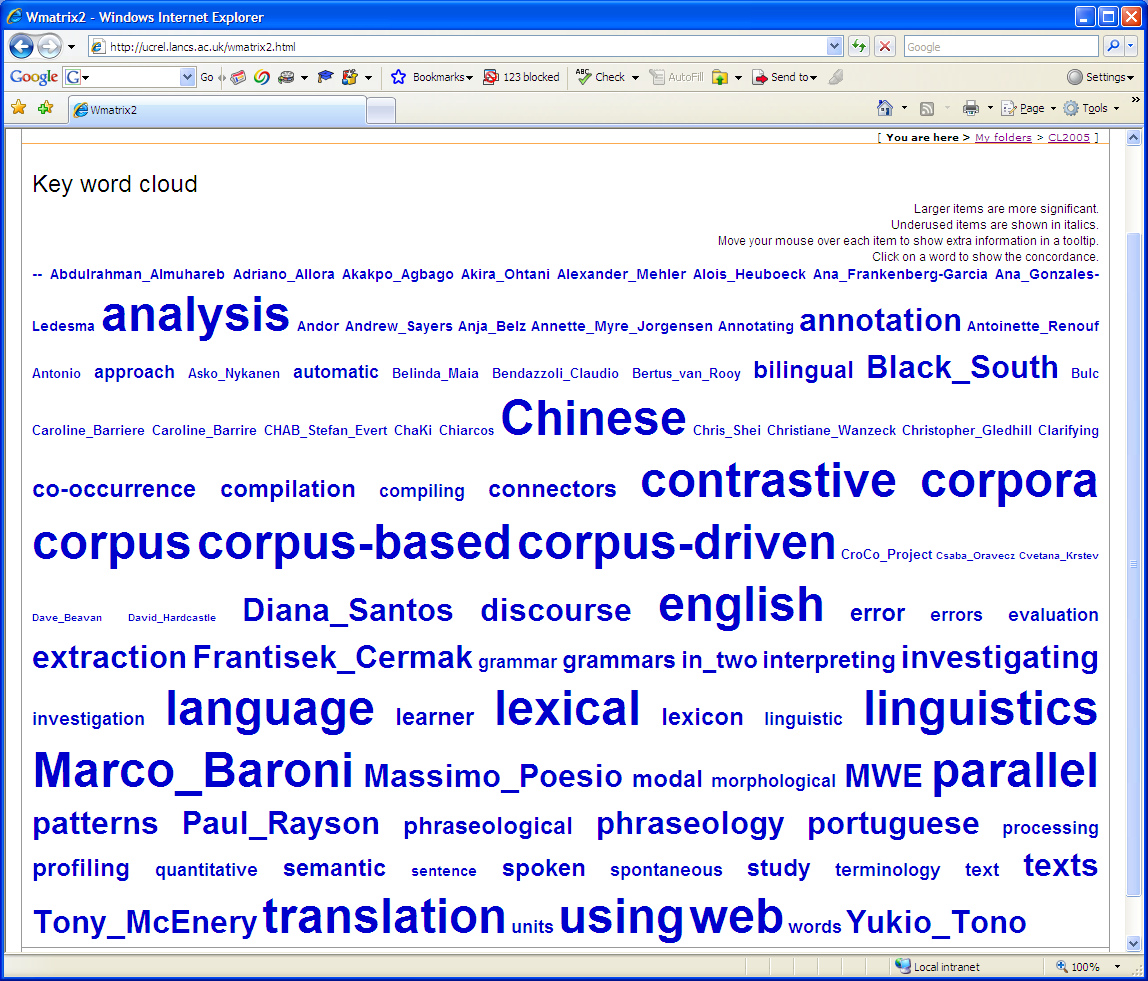


Figure 10: CL2005 key word cloud

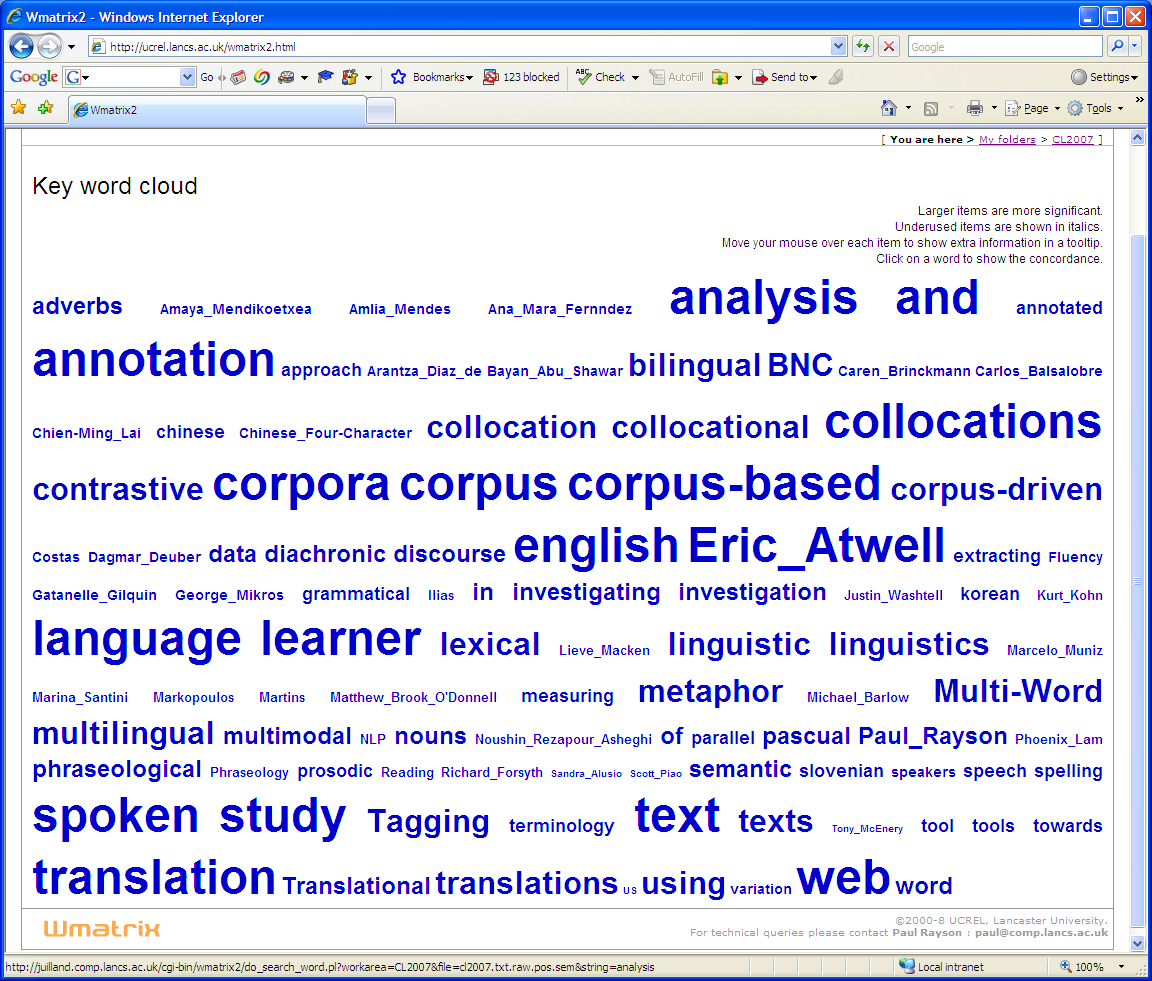


Figure 11: CL2007 key word cloud

Examination of the four key word clouds can be carried out very quickly, showing the advantage of the visualisation method. We do not need to consider the key word list itself. A summary of observed patterns is as follows:

* Methodological steps such as *Annotation* and *tagging* well represented throughout
* There is a move over time from *grammar* to *semantic* and *phraseology*
* There are a number of languages represented in each conference cloud
  + *English*, *French*, *Korean*, *Swedish* in 2001
  + *English*, *German* in 2003
  + *Chinese*, *English*, *Portuguese* in 2005
  + *Chinese*, *English*, *Slovenian* in 2007
* We can see that papers related to *Spoken* corpus analysis has been increasing since 2003
* *Translation* has been a strong theme (apart from 2003)
* *Web* appears as a major theme from 2005 onwards, marking the focus on web-as-corpus research

While it is possible to analyse the clouds by looking at each in isolation, we felt that we could extend the word cloud approach by supporting a form of animation which would assist the user to observe the changing trends in a set of data over a period of time. Furthermore, by considering the word clouds as a set rather than as individual objects, we can highlight conditions changing between clouds within the visualisation in a way that is obviously not possible otherwise. We have implemented a prototype of a Dynamic Tag Cloud (DTC) viewer as shown in figure 12.

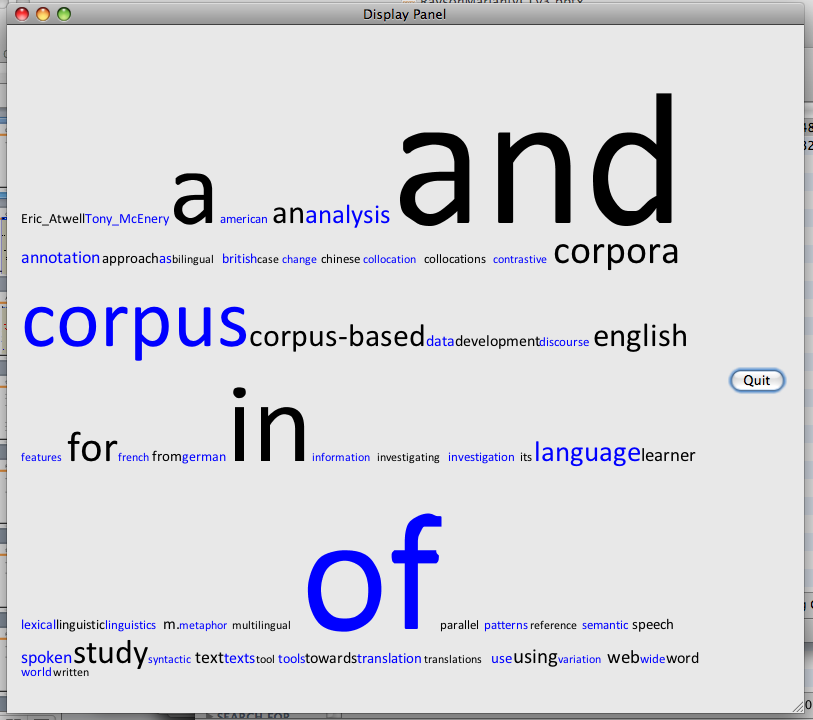


Figure 12: The Initial Display

The DTC system consists of two parts. The first part (the calculator) does the majority of calculations and saves the results in a binary file. The second part is the visualiser and it provides the end-user interface to the data. A word is admitted to the system only if it has a frequency of higher than 5 (the current system threshold). The smallest font size currently used is 12. The Calculator looks at the word cloud information for each conference and finds the maximum and minimum frequency for a word. It then uses this to work out the maximum screen space required by each individual word. We take the maximum frequency, subtract 5 and add 12. (We could experiment with more complicated formulae if required). From the font size and the number of letters, we can calculate the maximum screen area the word requires. The initial display (as shown in figure 12) shows the maximum word cloud.

When displaying a word in a cloud at less than its maximum value, we calculate its size at this font size and position it in the middle of its maximum area. This means if viewing a word at its less than maximum value, the user can see that there is room for growth and instantly knows that this is not its maximum value. As can be seen in figure 12, there is no room and hence all words are being shown at maximum frequency.

For each conference, then, the Calculator works out the font size and position of each word (relative to its maximum). Further, it finds out if the word is about to grow or shrink relative to the next conference in the sequence. This data is then saved in binary format for the Visualiser. When the Visualiser displays a cloud, words coloured red are in a growing phase, and blue means they are shrinking. A word does not necessarily have to appear in every conference. If a word is missing, it appears in a small font size (currently 5) just to indicate to the user there is a space for a word here but it is absent at the moment.

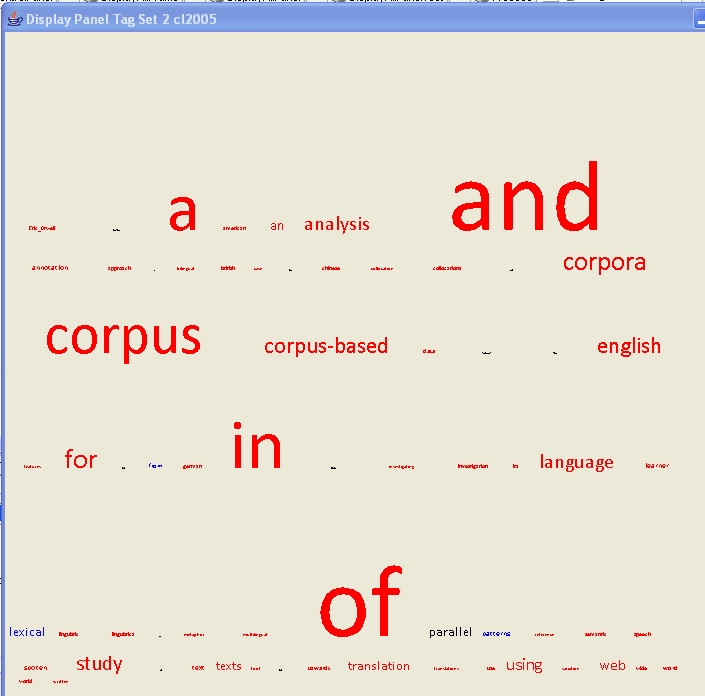


Figure 13: A Frame of Animation

In our case study example, we have four conferences as our sets of data. The initial frame would be CL2001, the second CL2003, the third CL2005 and the final CL2007. This is only four frames and would convey the information in a very jerky manner. This means the Visualiser has to do some calculations to create “in-between” frames. The issues that therefore arise are : (a) how much time should elapse between frames (i.e. how long should an individual frame stay on screen) and (b) how many “in-between” frames should there be? It was decided that these should be specified by the end-user.

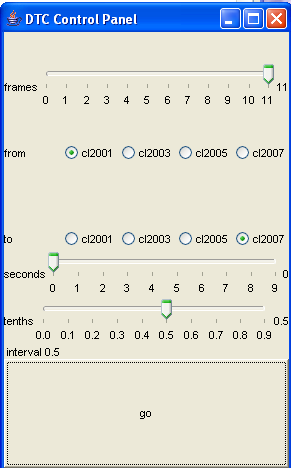


Figure 14: The DTC Control Panel

As can be seen in figure 14, the end-user can set the number of frames between clouds anywhere between 0 and 11. The number of seconds elapsing between frames can be set between 0 and 9 seconds, and 0.1 to 0.9 seconds. We also decided it might be useful to allow users to specify a subset of clouds, hence the “from” and “to” radio buttons. This is a prototype system, and the interface needs work. It would seem sensible to have controls users would expect with continuous media i.e. play, stop, rewind, forward, maximising the kinds of exploration of the tag sets available to the end user.

5. Conclusions and future work

In this paper, we have proposed the idea of using information visualisation techniques for corpus linguistics. We have highlighted tools and techniques that are already used in corpus linguistics that can be considered as visualisation: concordances, concgrams, collocate clouds, and collocational networks. We described the key word cloud approach as implemented in the Wmatrix software and have shown how it can be extended using a dynamic visualisation technique.

In future work, we plan to explore the addition of a dynamic element to the existing visualisations described in section three which are currently rather static. This would enhance their “data exploration” nature even further. To paraphrase Gene Roddenberry2, we wish to allow linguists to explore their data in ‘strange’ new ways and to seek out new patterns and new visualisations. In this enterprise, we can assess the usefulness or otherwise of the new techniques.

For the case study itself, we will extend the corpus to include abstracts and then full text of the papers instead of just the titles and authors. The same technique could also be applied to chart the changing trends in corpus and computational linguistics journals over time.

With significantly larger corpora being compiled, we predict that the need for visualisation techniques will grow stronger in order to allow interesting patterns to be seen within the language data and avoid practical problems for the linguist who currently needs to analyse very large sets of results by hand.

Notes

1. At the following URLs: http://ucrel.lancs.ac.uk/publications/CL2003/CL2001%20conference/index.htm; http://ucrel.lancs.ac.uk/publications/CL2003/index.htm; http://www.corpus.bham.ac.uk/PCLC/; http://ucrel.lancs.ac.uk/publications/CL2007/

2. http://en.wikipedia.org/wiki/Gene\_Roddenberry

References

Ahlberg, C. and Shneiderman, B. (1999) Visual information seeking: tight coupling of dynamic query filters with starfield displays, in *Readings in information visualization: using vision to think,* Morgan Kaufmann Publishers Inc.  San Francisco, CA, USA, 1-55860-533-9, 1999, pp 244-250

Beavan, D., ‘Glimpses though the clouds: collocates in a new light’. *Proceedings of Digital Humanities 2008*, University of Oulu, 25-29 June 2008.

Cheng, W., C. Greaves and M. Warren. 2006. ‘From n-gram to skipgram to concgram’, *International Journal of Corpus Linguistics* 11 (4), pp. 411–33.

Collins, C. (2006). DocuBurst: Document Content Visualization Using Language Structure. IEEE Information Visualization Symposium 2006.

Fekete, J-D. (2006). Information visualisation for corpora. In proceedings of Digital Historical Corpora, Dagstuhl-Seminar 06491, International Conference and Research Center for Computer Science, Schloss Dagstuhl, Wadern, Germany, December 3rd-8th 2006.

Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics* 29 (3), pp. 333-347.

McEnery (2006) *Swearing in English*. Routledge, London.

Phillips, M. (1985). *Aspects of text structure*. Elsevier, Amsterdam.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*. 13:4 pp. 519-549.

Scott, M. (1997) PC analysis of key words – and key key words. System 25.2. Amsterdam: Elsevier, pp. 233-245.