ComPair: Compare and Visualise the Usage of Language

Introduction

This paper will demonstrate ComPair, a new tool to investigate and compare word usage, encouraging new ways to explore language variation. While remaining focussed on the usability and the promotion of navigation, this tool represents an evolutionary step forward from the author's previous award winning visualisation applications. This paper will introduce the methods and technologies at its core, perform a demonstration of the tool and discuss opportunities for further collaboration.

Collocation

Firth in 1957 tells us 'You shall know a word by the company it keeps' leading to a contextual investigation of language which remains with us today. Identifying a word of interest and examining its collocates, often tells us more than a traditional dictionary definition ever could. Traditional corpus tools display collocates in tabular format, providing rich statistical data at the expense of giving the user an opportunity to see the overall linguistic landscape. Tools such as Beavan's Collocate Clouds present this information very differently, visualising the collocates in cloud form, as in figure 1.

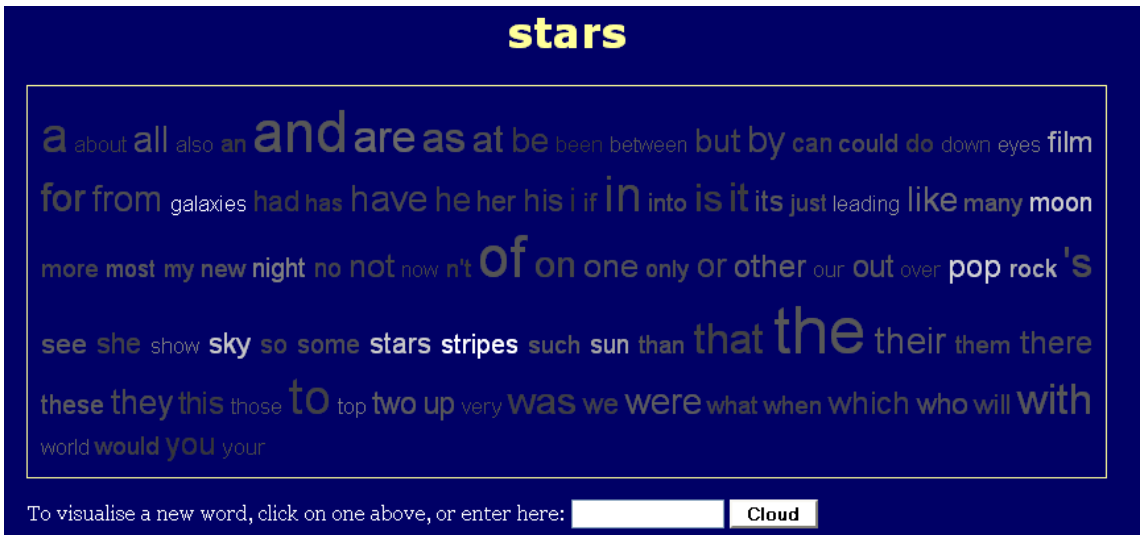

Figure 1. Collocate Cloud of node word ‘stars’

http://www.scottishcorpus.ac.uk/corpus/bnc/collocatecloud.php?word=stars (accessed 1 November 2010)

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.

Comparison

Louw introduced us to semantic prosody, which describes how synonymous words can actually take on positive or negative connotations. A natural way to investigate this would be to separately compare the collocates of each node word of interest. This can be performed by looking at multiple collocate clouds side by side, or by using statistical tools presenting tabular data. While these methods may be best suited to the comparison of many node words of interest, ComPair provides a solution to the comparison of two words, while keeping true to the aims of collocate clouds.

Semantic prosody is illustrated in figure 2, comparing the collocates of ‘utterly’ vs. ‘absolutely’. Negative terms cluster near ‘utterly’ where as positive terms cling to ‘absolutely’. At face value

these words are synonymous, but they are clearly used in different contexts and are not simply interchangeable. These are often issues which challenge learners of English as a foreign language.

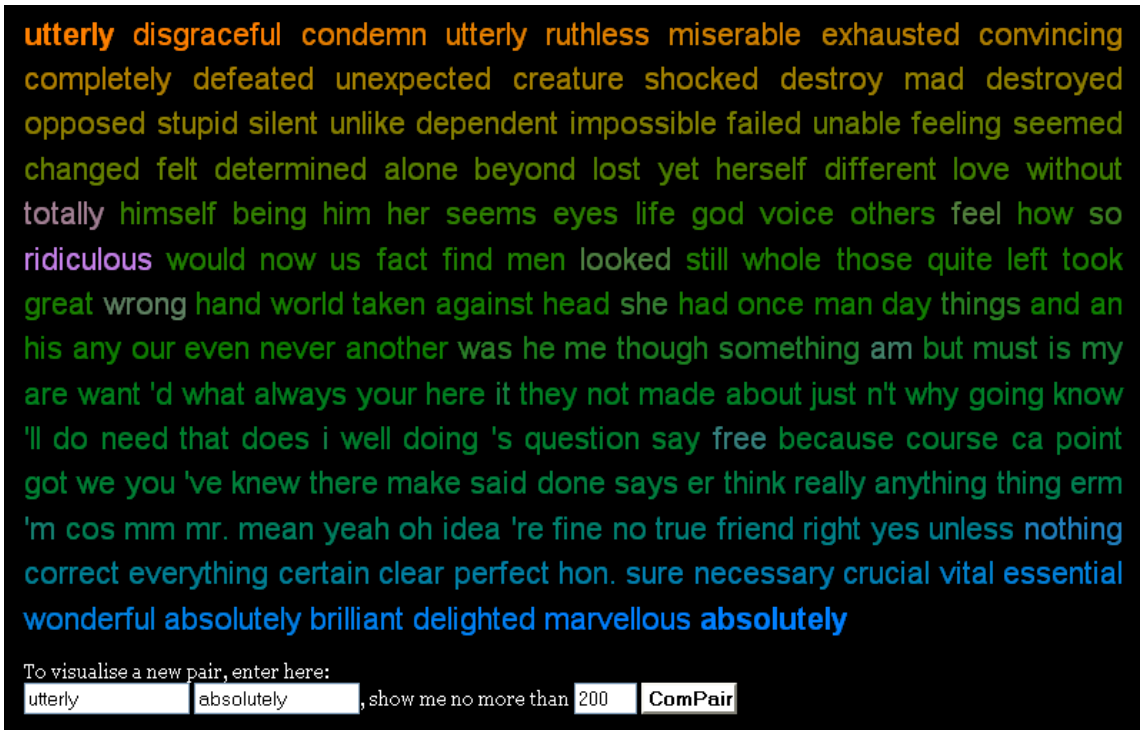

Figure 2. ComPair visualisation of ‘utterly’ vs. ‘absolutely’ in the British National Corpus

Method

ComPair calculates the collocates of both node words, ranking the results using a combination of frequency of co-occurrence, and by collocational strength (adopting the Mutual Information (MI) measure). A continuum is formed, with each extremity representing the separate node words (imagine a piece of string, with a label representing each search term at each end). The collocates are then distributed along this continuum, using the relative pull of collocational strength towards each node (the words near the end of the string are strongly associated with those end labels, those in the centre much less so).

Visualisation

ComPair displays this continuum, displaying each node word and the collocates between them. The display uses a spectrum of colour, to further enforce the ordering of the collocates. In the MI tug of war, the words in the centre share similar MI scores with each input term. Typically these are fairly low MI figures and appear green. In the ‘utterly’ vs. ‘absolutely’ example above ‘ridiculous’ appears in pink, this indicates that while the MI scores (utterly- ridiculous’ vs. absolutely-ridiculous) are roughly the same, they are much higher than the surrounding collocates. Ridiculous is therefore a word used strongly with both utterly and absolutely. Those collocates appearing close to each node, and sharing its colour are used very strongly with that node word, and only that word. Figure 2 tells us that things can be ‘absolutely marvellous’ but not ‘utterly marvellous’. In comparison, someone can be ‘utterly ruthless’, but not ‘absolutely ruthless’.

Future directions

At present ComPair allows for the comparison of two separate words in a single corpus. One possible extension would be the facility to search for the same words across two corpora. Imagine two corpora of differing political parties. With a single search term, ComPair would help expose the views and attitudes towards that concept. Another avenue would contrast word usage in British vs. American English.

Other applications would involve its use as a learning tool, allowing users to go beyond dictionaries and thesauri, to see in detail how different words actually operate. Visualisation of more than two node words should also be possible given different display techniques.

Bibliography

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

Firth, John R., 1957. *Modes of meaning*. Oxford: Oxford University Press.

Louw, B., 1993. *Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies.* In Baker, M., Francis, G. & Tognini-Bonelli, E. (eds) "Text and Technology". Philadelphia/Amsterdam: John Benjamins.