Visualisation of textual data through collocate clouds

David Beavan

Corpus of Modern Scottish Writing
Department of English Language
University of Glasgow

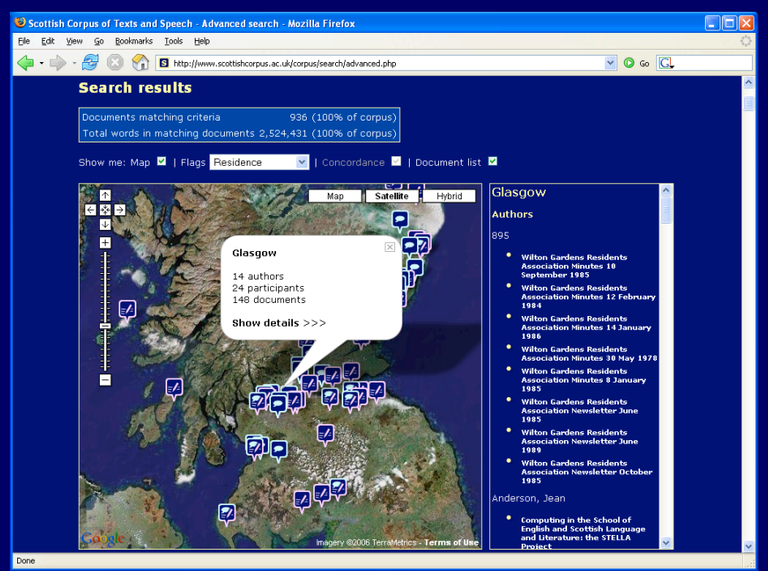
Scottish Corpus of Texts & Speech

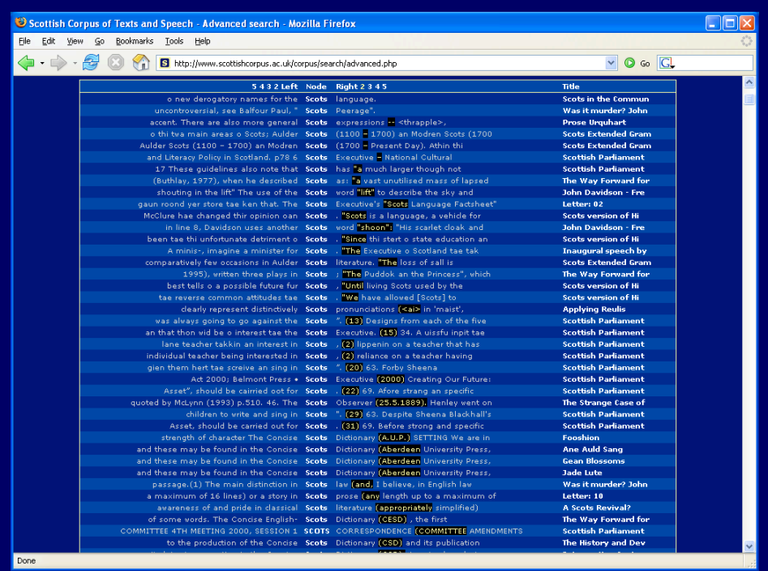
Features

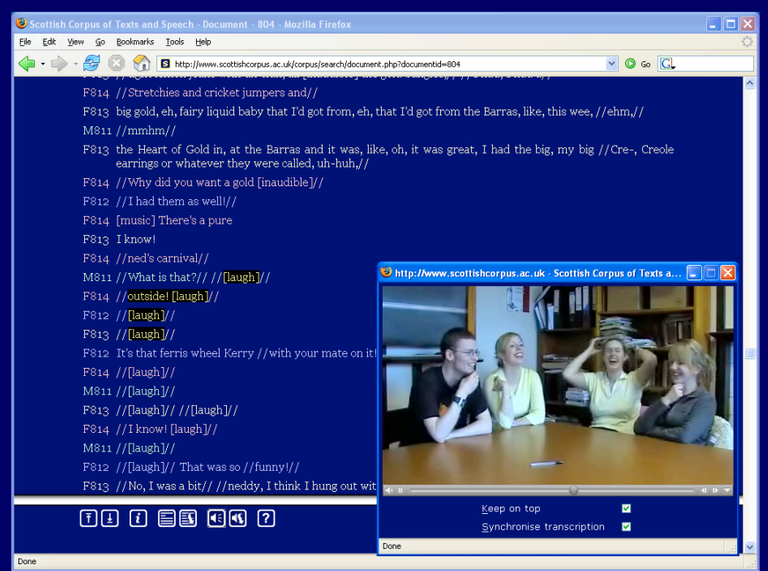
- Scots and Scottish English
- 4 million words, 20% spoken
- 1945 – present day
- Diverse range of genres e.g. conversations to legal documents
- Wide geographic spread
- Free and publicly available at www.scottishcorpus.ac.uk

Challenges

- Variant spellings e.g. how do we link
  home = hame = haim
  potatoes = tatties
- How to place a document on a continuum from Scots to Standard Scottish English?
- Can regional or dialectal words be reliably found?







Collocate Cloud - Mozilla Firefox

File Edit View History Bookmarks Tools Help

blood

_ a all also amount an and aneurysm any are arterial artery as be becomes been being but by carried cells changes circulating clot diseases effused escapes extravasated fig flow fluid for form formation fresh from hæmorrhage has if in into is it its known large leucocytes loss lost lymph may more most new not of on only or other part parts pigment poisoning present pressure quantity rapidly red result s sac serum should skin slowly so some stained stream such supply than that the these they this through tissue tissues to usually veins vessel vessels way when which with

Small = low frequency, big = high frequency. Dim = low MI, bright = high MI

Click on word above or visualise new word: yourword Go

Done

Collocate Cloud - Mozilla Firefox

File Edit View History Bookmarks Tools Help

activation

- a ab abstract activation activity after also an and {ap-1} are as be binding both but by ca2 can cell cells contrast control cytotoxicity dependent during expression factor factors for from gene have {hiv-1} {il-2} important in induce induced induces inhibit inhibited involved is it kinase lipopolysaccharide lps may mechanism mediated {nf-kappa-b} {nf-kappab} no not {nuclear-factor-kappa-b} of on only or pathway production proliferation promoter {rap1} rapid required response results role show signal signalling {stat1} {stat4} {stat5} {stat6} {t-cell} {t-cells} {tcr} that the these this through ti to transcription transcriptional tyrosine ui upon vivo was we were which while with

Small = low frequency, big = high frequency. Dim = low MI, bright = high MI

Done

Collocate Cloud - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://projects.scottishcorpus.ac.uk:81/corpus/interactive/custom/vis.php?url=http%3A%2F%2Fprojects. tattie

glycogenin-2

34850-6 46202-5122 52 a ability abolished accumulation after also and another based but can characterization contained described detectable determine ewing extracts found glycogen glycogenin-1 glycogenin-2beta human in indiana indianapolis indicates initiator interact is isoform level liver medicine most mutation not of one only phe recently residue s self-glucosylate self-glucosylating that the to treatment tyr-196 usa was with

Done

Collocate Clouds visualisation

Benefits

- Clouds familiar to many users e.g. Flickr shows shared photograph keywords as a cloud
- Good introduction to analysis for new users and beginners
- Acts as a window into what may be complex documents or data sets
- See how lexical terms interact
- Visualise data subsets and vocabularies
- Promotes data exploration and browsing by allowing collocate words to trigger new clouds using selected word as node
- Could be used as a gateway into more specific statistics in tabular form

Method

1. Choose node word e.g. ‘blood’
2. Search entire document for node word
3. Collect collocates (surrounding words) five words before node, five words after
4. Total up occurrences of each type
5. Keep 100 most frequent collocates
6. For each collocate calculate MI (mutual information).
   The measure of how likely the node and the collocate occur together.
7. Display the 100 collocates alphabetically
8. Scale font size to frequency of occurrence
9. Scale brightness to MI score