Paul Rayson, John Mariani, Bryce Anderson-Cooper, Alistair Baron, David Gullick, Andrew Moore, Stephen Wattam

Towards Interactive Multidimensional Visualisations for Corpus Linguistics

We propose the novel application of dynamic and interactive visualisation techniques to support the iterative and exploratory investigations typical of the corpus linguistics methodology. Very large scale text analysis is already carried out in corpus-based language analysis by employing methods such as frequency profiling, keywords, concordancing, collocations and n-grams. However, at present only basic visualisation methods are utilised. In this paper, we describe case studies of multiple types of key word clouds, explorer tools for collocation networks, and compare network and language distance visualisations for online social networks. These are shown to fit better with the iterative data-driven corpus methodology, and permit some level of scalability to cope with ever increasing corpus size and complexity. In addition, they will allow corpus linguistic methods to be used more widely in the digital humanities and social sciences since the learning curve with visualisations is shallower for non-experts.

1 Introduction

Corpus linguistics is a methodology for the study of language using large bodies (corpora, singular corpus) of naturally occurring written or spoken language (Leech, 1991). Corpus linguistics has collected together a number of computer-aided text analysis methods such as frequency profiling, concordancing, collocations, keywords and n-grams (also called clusters or lexical bundles) which have been utilised over the last forty years or so for language analysis in a number of areas in linguistics e.g. vocabulary, syntax, semantics, pragmatics, stylistics and discourse analysis. Corpus methods are inherently data driven, largely exploratory and allow the analyst to carry out empirical investigations, to discover patterns in the data that are otherwise difficult to see by other means e.g. by intuition about language (Sinclair, 2004).

The corpus linguistics methodology is based on comparing corpora or subsets of a corpus with each other in order to discover differences in the language represented by those corpora or sub-corpora. Many standard reference corpora have been collected to represent specific language varieties or genres. With the availability of more powerful computers and larger data storage facilities, these standard reference corpora have increased in size

over the years from the one million word LOB corpus (Johansson et al., 1978), 100 million word British National Corpus (BNC) (Leech, 1993), 385 million word COCA (Davies, 2009) and the two billion word Oxford English Corpus. However, corpus methods have largely remained the same over this time period. As a result, compromises have to be made with each type of corpus analysis e.g. higher cut-off values are used to filter key words and collocation results based on a need to reduce analysis time rather than for any specific level of significance. Concordance lines are thinned by large factors in order to fit with time scales of analysis rather than by variation and relevance factors. With the web-as-corpus paradigm (Kilgarriff and Grefenstette, 2003) gaining prominence, even larger collections of textual data sourced from websites are becoming available (Baroni et al., 2009) so the problem will continue to worsen. In addition, corpus linguistics methods are spreading to other research areas in linguistics, digital humanities and social sciences e.g. discourse analysis (Baker, 2006), sociolinguistics (Baker, 2010), conceptual history (Pumfrey et al., 2012), and psychology (Prentice et al., 2012). For these disciplines, it is imperative that the corpus tools and methods have a shallow learning curve (Rayson, 2006) and we hypothesise that interactive visualisation technologies will help with this expansion.

Basic visualisation techniques (e.g. bar charts for relative frequency plots) have been used in the past in corpus linguistics but these have focussed on one level (e.g. lexical, grammatical, semantic) or method of analysis at a time. Very few publications discuss specific requirements for extending corpus retrieval software (c.f. Smith et al. (2008)), and this paper goes some way to address this deficiency. The main contributions of this paper are the novel interactive and dynamic techniques that we have developed for extending advanced corpus linguistics methods. We also propose a framework to combine all these separate multiple dimensions together. We describe an interactive key word cloud for visualising keyness statistics, an interactive and dynamic method for visualising collocation statistics and a method for contrasting social network relations with language comparisons. These visualisation methods are an improvement on current state of the art in at least four ways. First, they are designed to support the data-driven multidimensional iterative exploration embodied in the corpus linguistics methodology. Second, they address the shortcomings of current static one dimensional corpus methods. Third, they are scalable in order to cope with increasing corpus size and complexity. Finally, they contribute to enabling the analysis methods of corpus linguistics to be accessible to a variety of audiences, for example, non-technical users in the wider social sciences and humanities.

2 Related Work

The corpus linguistics methodology for the study of language using large corpora consists of five core steps (adapted from Rayson (2008)):

1. Question: devise a research question
2. Build: corpus design and compilation
3. Annotate: manual or automatic analysis of the corpus
4. Retrieve: quantitative and qualitative analyses of the corpus
5. Interpret: manual interpretation of the results

The methodology is inherently data-driven and empirical, exploiting the collections of real language samples to drive the analysis and direct the results as opposed to the use of manually constructed language examples driven by intuition. Corpus retrieval software, our focus here, is intended to facilitate exploration of the annotated corpus data using a variety of quantitative techniques. These techniques include frequency profiling: listing all of the words (types) in the corpus and how frequently they occur, and concordancing: listing each occurrence of a word (token) in a corpus along with the surrounding context. The n-gram technique (also called clusters or lexical bundles) counts and lists repeated sequences of consecutive words in order to show fixed patterns within a corpus. A typical corpus investigation would proceed with a large number of retrieval operations conducted through the corpus retrieval software (e.g. to check the frequency of a particular word or linguistic feature, or to search for an item or pattern using the concordancing view), guided by the research question and the quantitative results obtained in earlier searches. Although this iterative process is often not reported in final publications, it is evident from the many textbook descriptions of corpus linguistics. Typically, the research question itself (step 1) is refined in the light of categorisation and analysis of concordance results and comparison operations between corpora, and then the stepwise process begins again. This refinement process specifically corresponds to the interactive exploratory approach that we propose here to be aided by improvements in visualisation methods. Although they are not necessarily viewed as such, some existing techniques in corpus linguistics can be considered as visualisations. In this and the next section we will consider three of the most prominent examples: concordances, collocations and key words.

First and foremost, the concordance view with one word of interest 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, and is the main way that corpus linguists engage with corpora. 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 1.

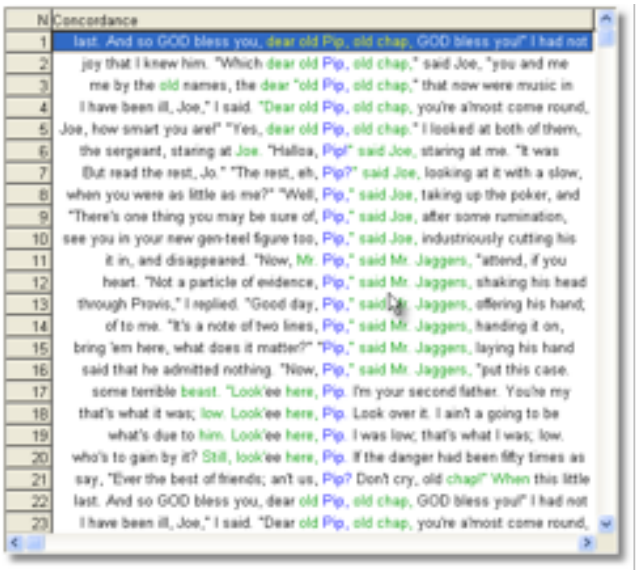

Figure 1: Concordance concgrams.

Another method in the corpus retrieval toolbox is collocation, for which Beavan (2008) has already explored visualisation techniques. Collocations are pairs or sequences of words that co-occur in a text more often than would be expected by chance, usually within a window of five words of each other. By taking the collocates of a word, ordering them alphabetically and altering the font size and brightness, the collocate cloud shown in figure 2 provides an intuitive view of a set of collocates. Here, font size is linked to frequency of the collocate and brightness shows the Mutual Information (MI) score (a statistical measure of the strength of association between the words). 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) employs a visualisation technique when manually drawing collocational networks (figure 3). These show key words that are linked by common collocates. McEnery's work is influenced by Phillips (1985) who uses similar (again, manually created) diagrams to study the structure of text.

Visualisation is finding application in many areas of the modern world; in science, arts, social media, and the news. The cognitive principles behind visualisation are well summarised by Meirelles (2011) when she writes "to record information; to convey meaning; to increase working memory; to facilitate search; to facilitate discovery; to support perceptual inference; to enhance detection and recognition; and to provide models of actual and

Figure 2: Collocate Cloud.

Figure 3: Collocational network created manually.

theoretical worlds". Linguistics is no exception to the rule. An interesting aspect of the work is a willingness to use existing tools, not only those specifically designed for corpora, but also more general visualisation toolsets.

Siirtola et al. (2010) foresaw some of this development when they discussed the use of the R statistical language and the Mondrian data visualisation tool. However, the R language is generally thought to have a steep learning curve and "the dreaded command line interface" (ibid). They argue for interactive tools, but of course, when using various tools from different sources, it can be difficult to link together the tools so that changes made in one are reflected in the other. Scrivner and Kubler (2015) describe a multi-dimensional parallel Old Occitan-English corpus. They use the ANNIS (Zeldes et al., 2009) search engine to provide graphical querying and displaying of multi-layered corpora. The user can specify their query graphically. In their example, they convert the retrieved data into the R data frame format and produce a motion chart, using GoogleViz.

The Text Variation Explorer (TVE) (Siirtola et al., 2014) harkens back

to earlier work done in the visualisation field by Ben Shneiderman, and the concepts of direct manipulation, continuous and immediate feedback and linked visualisations (see, for example, the Film Finder). This, as indicated above, can be difficult to achieve when using a tool chain. They refer back to the 2010 paper when they observe that while Mondrian can supply interactive graphs (for quickly formulating hypotheses about data and perhaps even managing to verify them in some cases) it lacks one essential: a connection to the text itself. The graphical aspect of TVE is a line graph; in the paper, they use James Joyce's "Ulysses" as an example. They split the text into windows (the size of which is specified by the user) and calculate three parameters, each one of which is represented as a line within the graph. So, going from left to right, we move from the first window of the novel to the last. The user can position themselves anywhere on the line graph, and the underlying text of the window they are selecting is also displayed (in context with the rest of the text). By accessing the text display, and selecting (a) word(s), their new position in the text is reflected in the line graph. This is known as "brushing", the ability to interact with one visualisation and have that interaction reflected in all other associated visualisations.

The WordWanderer (Dork and Knight, 2015) extends tag clouds into a navigational interface for text. Beginning with an alphabetically ordered tag cloud showing frequency. By moving the pointer over a word (in their "Hansel and Gretel" example), say "forest", common collocates are highlighted (this indicates that "children" is a common collocate). If the user now selects "forest", its collocates are organised according to their relative proximity in the text. Finally, if the user draws a line between two words, we get a comparison view, arranging collocates according to their relative strength of association to each of the two words. Hilpert (2011) proposed the use of motion charts (a series of time ordered scatter plots) to dynamically visualise language change in a diachronic corpus. This type of visualisation requires relatively large corpora.

A novel direction has emerged recently in two distinct areas: dialectology and spatial humanities. The common thread between these two approaches is map-based visualisations of language data. In order to understand regional linguistic variation in the US, Huang et al. (2015) collected a year of geo-tagged Twitter data. County-based results were plotted and hierarchically clustered dialect regions were derived from the analysis. In order to showcase the newly emerging area of spatial humanities which combines Geographic Information Systems with natural language processing and corpus linguistics, Murrieta-Flores et al. (2015) carried out an analysis of the UK Registrar General's Reports containing descriptions, census data and other information to examine how mentions of various diseases correlated with place names in the data (see figure 4 for an example of their results). Map-based

visualisations are derived and were compared over decade spans. In general, the spatial humanities method allows a researcher to ask three main types of questions of a dataset (a) where is the corpus talking about, (b) what is the corpus saying about these places, and (c) what is the corpus saying about specific themes e.g. health and disease, money and finance, in proximity to these places? In contrast, Knowles et al. (2015) have explored 'inductive visualisation' techniques that allow the exploration of time and space in holocaust testimonies which do not lend themselves to regular geographical and sequential time-based representations.

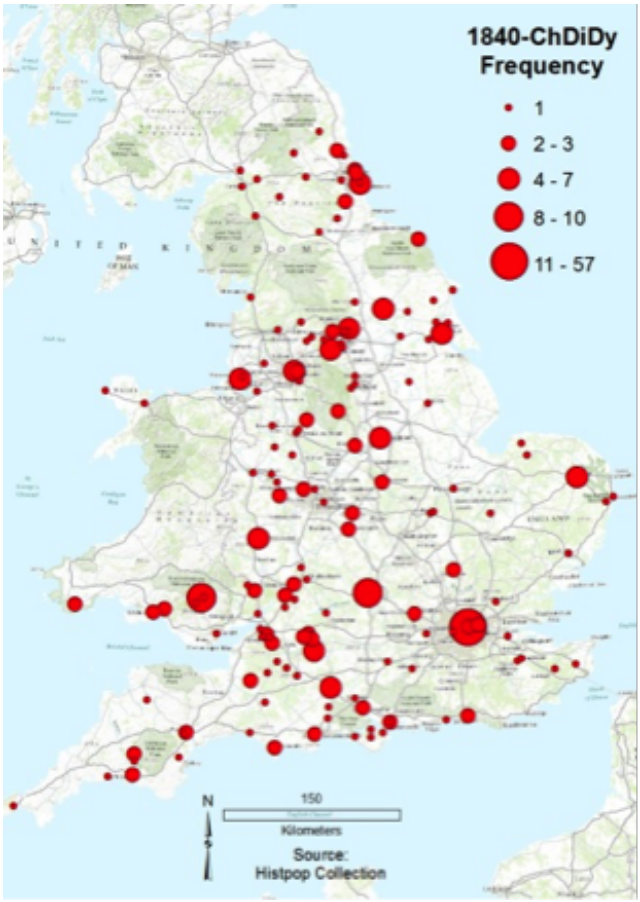

Figure 4: Frequency map of diseases in Registrar General data.

In other cognate areas, such as digital humanities and literary analysis, visualisation approaches are gaining ground. Keim and Oelke (2007) and Oelke et al. (2012) develop the idea of a literature fingerprint which is a pixel-based visualisation to view fine-grained detail of one particular value, where each pixel corresponds to one word. This value could represent the occurrence of a particular character name, function words, average sentence length or hapax legomena. Voyant Tools (Sinclair and Rockwell, 2016) provides a web-based text reading and analysis environment, complemented

by a variety of visualisations including: bubbles and cirrus (similar to word clouds), bubblelines (word repetitions), links (collocation relationships), and RezoViz (relationships between people, places and organisations). Wattenberg and Viégas (2008) present the Word Tree as an interactive version of the concordance view, first implemented in the IBM Many Eyes system. It provides a branching view of words and contexts occurring to the right of the word in the centre of the concordance largely preserving the linear view of the text. Culy and Lyding (2010) extended this (to be closer to the corpus linguistics approach) in their Double Tree implementation to include words on the left of the concordance line, frequency and part-of-speech information. Two further surveys of text visualisation techniques and related taxonomies were created by Kucher and Kerren (2015) and Jänicke et al. (2015).

The discussed examples allow us to highlight some further issues with the current technologies used to visualise large bodies of text. The first problem is the static nature of many of the technologies. This often presents users with far more information than necessary and offers no mechanism to limit the data to those aspects the user is interested in. This static nature can cause a significant amount of information overload, rather than reduce its impact and this is the second issue to be faced. This problem was partly tackled by TextArc amongst other tools which allow the interrogation of the data. However, this technology still displays the whole block of textual data at the same time, which will leave the graphic cluttered and possibly unclear. More generally, static and full text representations do not sit well with the iterative and data-driven nature of the corpus linguistics methodology. Very few of the existing techniques are tailored for the specific methods in corpus linguistics, and in addition, the existing corpus visualisations do not scale to large bodies of texts, a key requirement to tackle the growing size of corpora. All these reasons call for new visualisation techniques, or at least the adaptation of existing ones, in order to specifically address the particular needs of corpus linguistics in terms of scalability, and support for iterative exploration.

3 Case Studies

With the case studies presented in the following three subsections, we examine complementary aspects of visualising different dimensions of language corpora. Our case studies cover three of the five main methods in the corpus linguistic methodology: frequency lists, key words, and collocations. A fourth method, concordancing, is included in our multidimensional visualisation framework as proposed in section 4.

3.1 Case Study 1: key word and tag clouds

In this first case study, we propose a method that can be applied at multiple linguistic levels for the visualisation of key words results. The key words technique (Scott, 1997) is well known in corpus linguistics to users of WordSmith[1], Wmatrix[2], AntConc[3] and other tools. By comparing one corpus or text to a much larger reference corpus (or another comparable text), 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, along with an effect size is calculated for each word to show how 'unexpected' its frequency is in the corpus relative to the reference corpus. By sorting on this keyness value we can order the words and see the most 'key' words at the top of a table. In the Wmatrix software (Rayson, 2008), we have included a visualisation of the key words results in a static but interactive 'key word cloud'. In contrast to tag clouds in Flickr and other social networking websites, 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. In addition, we can apply the same comparison approach at other levels of linguistic analysis. Instead of comparing two word frequency lists, we can compare two part-of-speech frequency lists, or two semantic tag frequency lists. This extends the existing method and permits gisting by stylistic profile and key concepts. Previous work has used word clouds for visualising texts (Heimerl et al., 2014; Xu et al., 2016), but these have not exploited the keyness measures used in corpus linguistics. Vuillemot et al. (2009) does use the log-likelihood measure to compare sub-corpora but then relates word size to frequency rather than keyness. Our method also avoids the need for stop word removal of frequent closed class words which may well result in the loss of significant items of linguistic interest.

Here, we describe a case study using data drawn from the set of UK General Election 2015 Manifestos from the seven main political parties. Via this example, we show the key word and key concept cloud visualisation in practice. First, the seven manifestos for Conservatives, Labour, Liberal Democrats, Green Party, Plaid Cymru, Scottish National Party (SNP) and UKIP were downloaded from their websites in May 2015. Each file was converted from PDF by saving as text from Adobe Reader. Minor editing was required to format headers, footers and page numbers in XML tags, and converted n-dashes, pound signs, begin and end quotes to XML entities. Next, the resulting files were run through the Wmatrix tag wizard pipeline which assigns part-of-speech tags (Garside and Smith, 1997) and semantic

[1] http://www.lexically.net/wordsmith/
[2] http://ucrel.lancaster.ac.uk/wmatrix/
[3] http://www.laurenceanthony.net/software/antconc/

tags (Rayson et al., 2004a) and prepares word frequency and semantic tag frequency lists. Key word and semantic tag clouds are produced by comparing the frequency lists with the BNC Written Sampler corpus[4]. In these visualisations, the larger the font, the higher the log-likelihood score, so larger items are more significantly overused compared to the reference corpus.

The first two visualisations show the key words and key semantic categories for the Conservative party. Figure 5, at the word level, shows their focus on EU, tax, NHS and schools, amongst other items. Figure 6, at the semantic tag level, expands this and highlights their discourse on law and order, business, and employment, in particular.

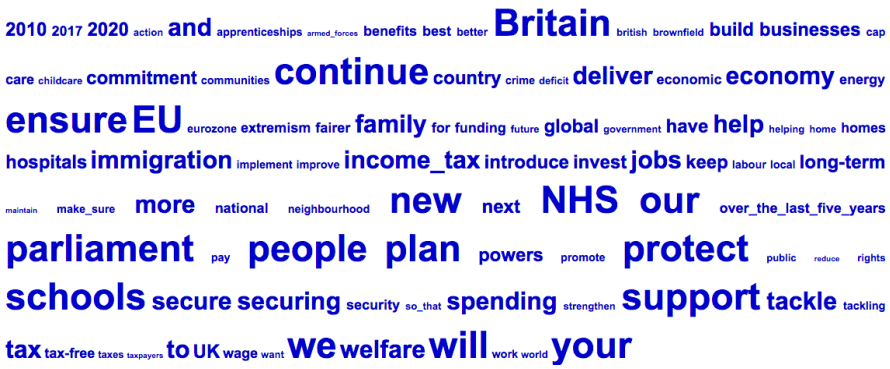

Figure 5: Word cloud for Conservative manifesto.

Figure 6: Semantic tag cloud for Conservative manifesto.

These two clouds can be contrasted with all those from the other parties.[5] For reasons of space here, we include only one other party. For the Green

[4] http://ucrel.lancaster.ac.uk/bnc2sampler/sampler.htm
[5] See http://ucrel.lancaster.ac.uk/wmatrix/ukmanifestos2015/ for the full set.

Party, the six most key words in their manifesto are green_party, we, local, tax, energy and climate, as shown in figure 7. Alongside green issues, their key semantic cloud in figure 8 focusses on money and government.

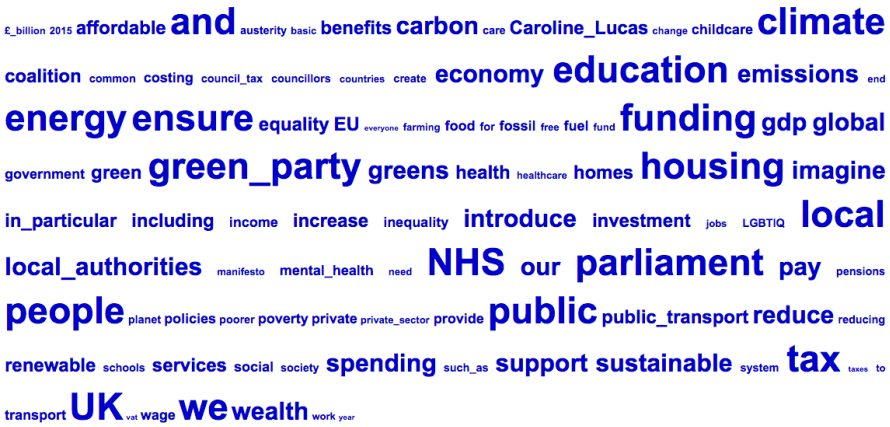

Figure 7: Word cloud for Green Party manifesto.

Figure 8: Semantic tag cloud for Green Party manifesto.

The Wmatrix software allows a user to click through the cloud in order to view concordance lines for a specific word or semantic tag, and by hovering over an item, the frequency and log likelihood statistic can be viewed. Thus the word and tag clouds do have interactive elements and represent multidimensional or multi-level visualisations.

3.2 Case Study 2: collocation networks

In the second case study, we propose to use interactive visualisation techniques to improve the interpretation and exploration of the collocation method in corpus linguistics. We have implemented these methods in both

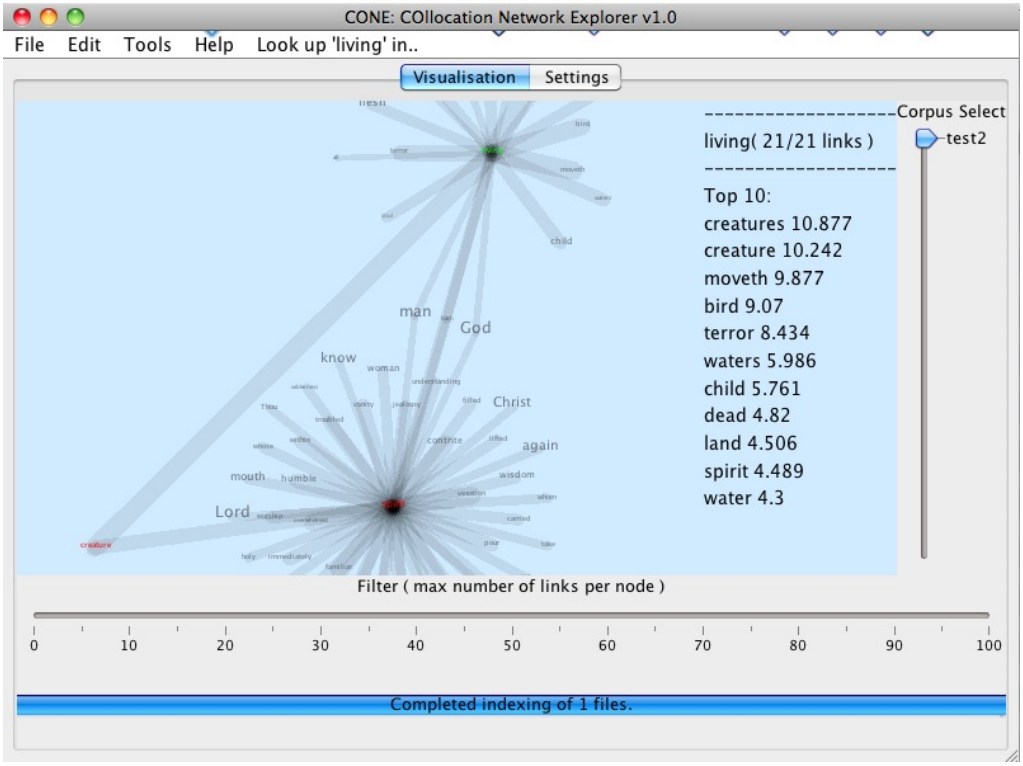

Figure 9: CONE

CONE (Gullick et al., 2010) and GraphColl (Brezina et al., 2015) (Figures 9 and 10 respectively) which provide visualisation of text collocation for all terms within a corpus simultaneously, presenting a graph that the user can manipulate and explore. The concept of collocational networks is a natural extension of collocation, and was first proposed before computing hardware was generally sufficient to provide an interactive visualisation (Phillips, 1985).

Collocation networks are generated by computing a statistical measure of association between all terms within the corpus. Such terms form the nodes of the graph, with edges being drawn between those with a significant tendency to co-occur. The exact measure and policy for graph construction varies between implementations. Early implementations used the mutual information (MI) score (Williams, 2002). CONE implements the commonly-used log-likelihood score as a measure of significance (Rayson et al., 2004b), whereas GraphColl supports a number of measures, as well as implementation of bespoke approaches.

Graph exploration presents a number of design challenges. Firstly, the choice of statistical measure (and the significance or effect size threshold chosen) dramatically affects the resulting graph. This is compounded by the tendency of the constraint-based graph layout algorithms used in both CONE

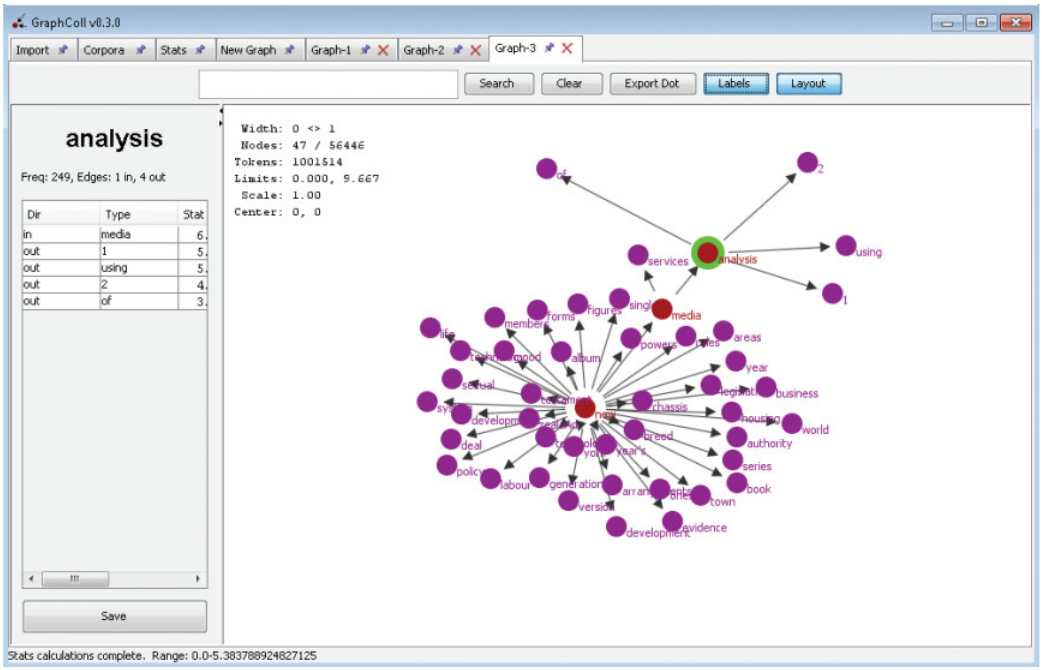

Figure 10: GraphColl

and GraphColl to produce non-deterministic layouts: even topologically similar graphs may appear different.

Graph layout is also a significant challenge to scalability. The Zipfian nature of linguistic data often yields graphs with high centralisation, leading to a dense mass of edges for diverse corpora, a problem also faced in a similar approach by Perkuhn (2007). This is mitigated in both tools by allowing the user to pan and zoom around the graph whilst rendering features at the same scale, essentially making them less dense at higher zoom levels.

Higher level visualisations such as those produced by CONE and GraphColl also present challenges to scientific replicability in that they present large amounts of data in a very dense manner, with the potential to embody many study designs. Both CONE and GraphColl permit partial exploration of graphs, accentuating this issue: a user chooses which nodes to expand (and thus compute collocates for), and this means it is possible to deliberately or unintentionally miss significant links to second-order collocates (or symmetric links back from a collocate to a node word). GraphColl's design attempts to minimise these issues by colouring links according to their "completed" state. This issue is also addressed in documentation, which presents a standardised method for reporting results from graph explorations which is intended to illustrate which design choices have been made during graph creation.

The ease-of-interpretability that visualisations offer presents scientific challenges: when the graph's generating function can be changed *during*

exploration, data dredging becomes as simple as moving a slider. To this end, GraphColl prohibits what many see as desirable features: wildcard searching, stoplists, and on-the-fly adjustment of statistical thresholds are all disallowed by design.

This issue is primed to affect any interactive visualisation. High-level visualisation tools such as CONE and GraphColl must walk a fine line between offering a useful perspective on data (which would not be possible otherwise) and providing such a strong lens as to render any observations largely dependent upon the tool itself. This tendency is evidenced in many areas of science already (such as genetics, which often relies on proprietary machinery), but is readily solvable through responsible reporting and efforts such as the open data movement (Kauppinen and de Espindola, 2011).

The high level from which data are seen also pose technical challenges for interchange formats, leading to a situation where data exported from such tools is either presented in a relatively arcane proprietary format, or stripped of much of the information from the data structures used for analysis. The solution to this lies in an approach of layered formats, which may yield further data where required: something that may take the form of an API to provide live interconnections between tools, or advancements in database representations.

Finally, it should be noted that GraphColl has a concordance feature built-in so that users can use the interface to more closely examine specific collocations in context. Either these things have to be built-in to support richer interaction, or there must be an interchange format to communicate with other corpus tools (a corpus data connector of some kind).

3.3 Case Study 3: social network relationships

This case study proposes the extension of an existing network visualisation based on 'follow relationships' in an online social network (Twitter) to instead be based on distances between language profiles. The overall aim of the study was to analyse potential political defections in the United Kingdom parliament. Using the Twitter REST API[6], the last 3,600 tweets (the maximum available) from verified UK Members of Parliament (MPs), and the list of verified MPs that each of these follow were collected. In total, 426 MPs of the 650 MPs in parliament were present in our dataset, the remaining MPs were not verified or not on Twitter.

The list of follow relationships were converted into a list of one-directional links between each MP who followed another MP, finding 29,345 links in total. If two MPs followed each other, two links were listed.

The words were collected from all tweets and a frequency list created for each MP. We removed URLs and user mentions from the list of words as

[6] `https://dev.twitter.com/rest`

URLs were very rarely repeated and were mostly auto-created short URLs for Twitter, and user mentions were removed to avoid overlap with follow relationships. All punctuation was removed and all words were converted to lowercase. A random sample of 2,000 words was taken for each MP, with MPs excluded who had used less than 2,000 words (thereby removing only 3 MPs). Each MP sample was compared against every other MP using two similarity measures: Jaccard and Log Likelihood. Jaccard looks at the similarity in the set of words used, whereas Log Likelihood looks at frequency differences. This process was repeated 10 times with a different 2,000 word random sample each time.

Both the follow relationships and the language relationships were visualised using force-directed graphs in D3.js[7]. In force-directed graphs, nodes are pushed away from each other while simultaneously pulled towards the centre of the graph. This allows any node's location to be based on their relative positions to one another, attempting to minimise crossing links and balance link distances. For follow relationships, all one-way links were of equal length but bi-directional links were set as half as long to represent a closer relationship. For language relationships, the link length was determined by the value that the similarity measure produced between the two nodes for that link. The more similar two nodes were, the lower the link lengths, bringing the nodes closer together. The closest nodes to any particular node are those that have the closest relationships. The resulting network graphs are shown for follow relationships in Figure 11, for Jaccard word similarity in Figure 12, and for log-likelihood word similarity in Figure 13. Note that the graphs are interactive, allowing particular parties to be highlighted. MP names and links between MPs can also be displayed. In all graphs, the positions of certain MPs and the orientation of the entire graph may vary as nodes are initially randomly placed, resulting in multiple possible stable arrangements. However, the overall pattern is consistent.

The follow relationship graph more visibly splits the MPs into distinct clusters related to political party. This may possibly be due to links not being present between all nodes, unlike in the word similarity graphs where a link is always present, but more or less distant depending on similarity. The word similarity graphs both do show the current three biggest UK political parties (Conservative: blue, Labour: red and Scottish Nationalist: yellow) generally clustered together, with outlier MPs (i.e. clustered closer to other parties) indicating possible interesting cases for further analysis. The interactive visualisation approach in this case study is a vital exploratory tool when developing the method (e.g. selecting appropriate distance measures) and analysing results (e.g. choosing subsets of MPs). Thus, our third case study shows that the existing visualisation technique previously used for exploring

[7] `https://d3js.org`

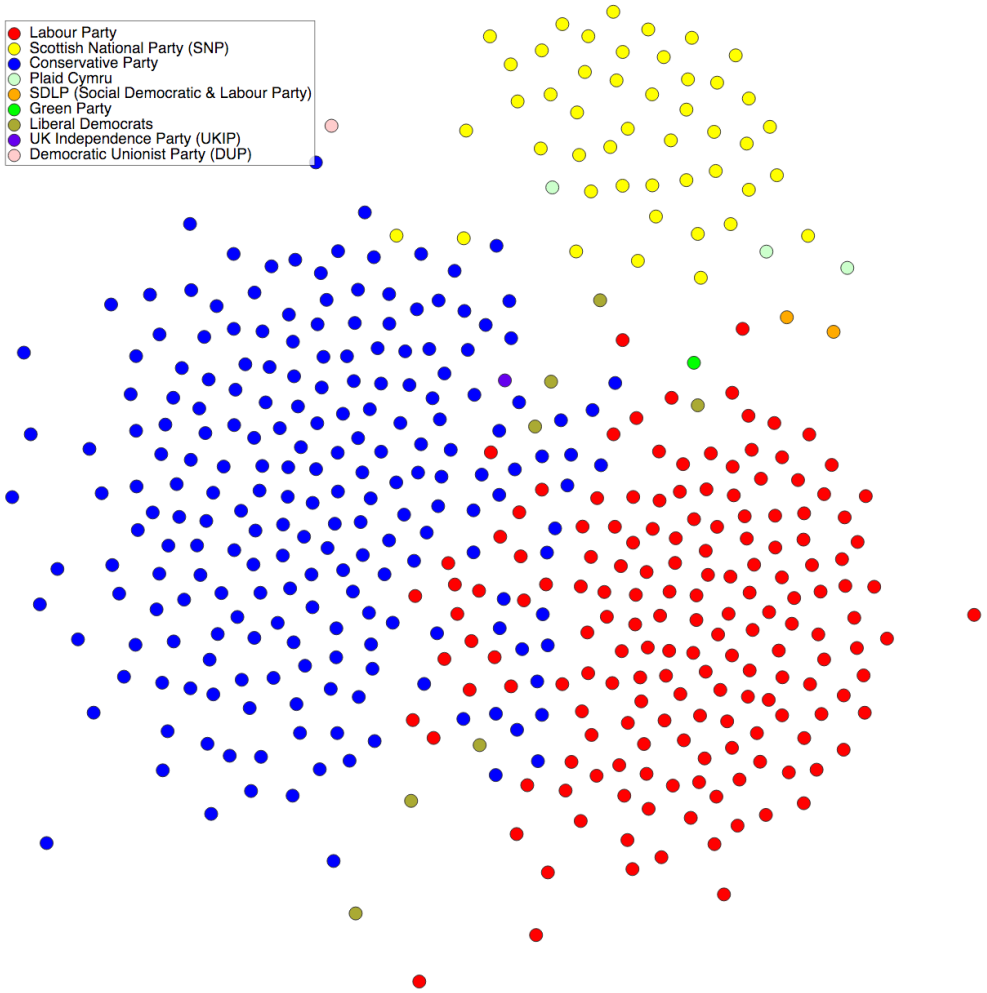

Figure 11: Follow relationships network of UK MPs on Twitter.

the network of relationships in an online social network can also be used to explore the linguistic similarity of specific subcorpora at the word level.

4 Proposal for Multidimensional Visualisation

Using the case studies demonstrated in the previous section, a proposal for putting these concepts into a multidimensional framework is described here. Our framework splits along three orthogonal dimensions: linguistic (lexical, grammar/syntax, semantics), structural (to permit sub-corpora) and temporal (for diachronic corpora). Our proposal for multidimensional visualisation explicitly supports key tenets of interactive visualisation such as navigating from a high level overview of the dataset, via filtering on specific dimensions to view slices or subcorpora (Heer and Shneiderman, 2012).

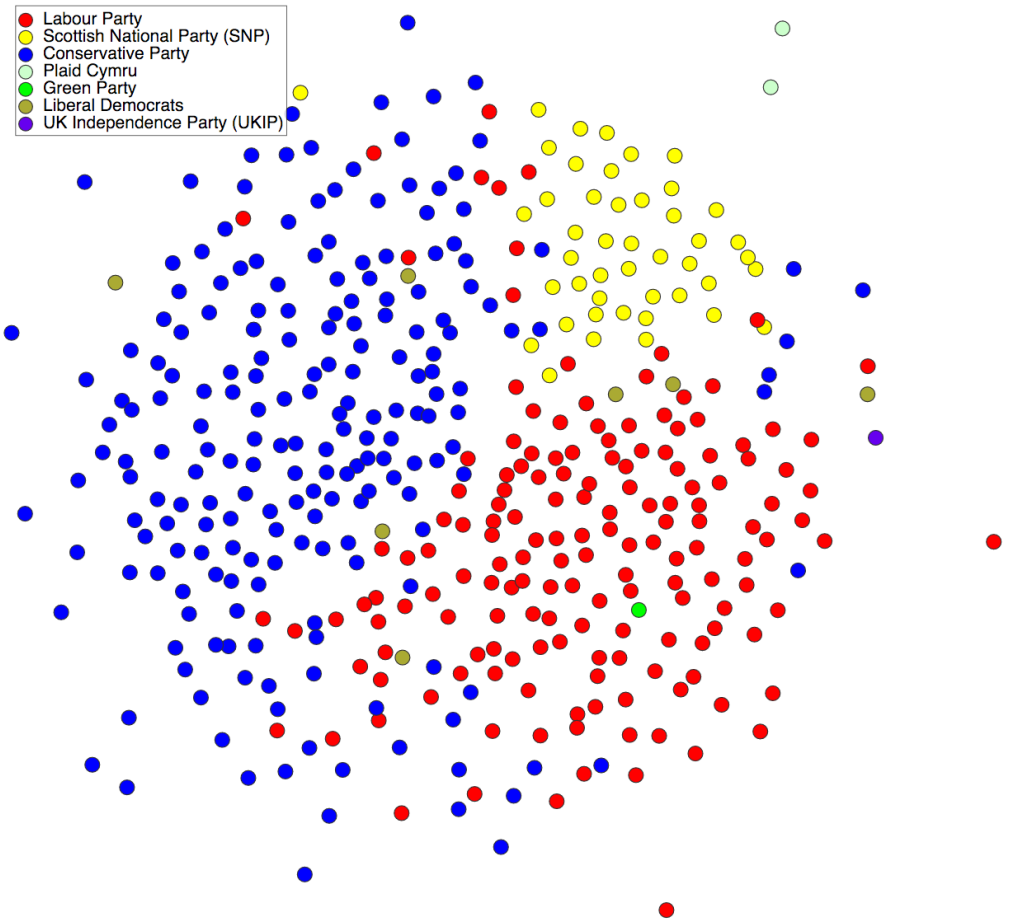

Figure 12: Jaccard word similarity network of UK MPs on Twitter.

Within the structural dimension there is a natural layering beginning with the whole corpus and subdividing into meaningful subsets dependent on the type of data (e.g. documents, chapters, tweets, person). At the top level, the whole corpus level can be analysed via the word clouds shown in section 3.1 where a user can find the most over/under-used words relative to a reference corpus. Incorporating the collocation network approach in section 3.2 the user would be able to click on a word in the key word cloud to explore the collocates for that word or tag, and from there to the concordance view. Furthermore a user should be able to select a group of words within the cloud and visualise collocates for those words to explore further similarities between the words. Second and subsequent layers would permit selection of subcorpora in order to exploit structure within the corpus, e.g. tweets as used in section 3.3. In addition, using the network visualisations shown in section 3.3 a user should be able to define subcorpora and visualise their similarities, differences alongside other relationships drawn from the dataset. A specific use case for our proposed framework can be extended from the

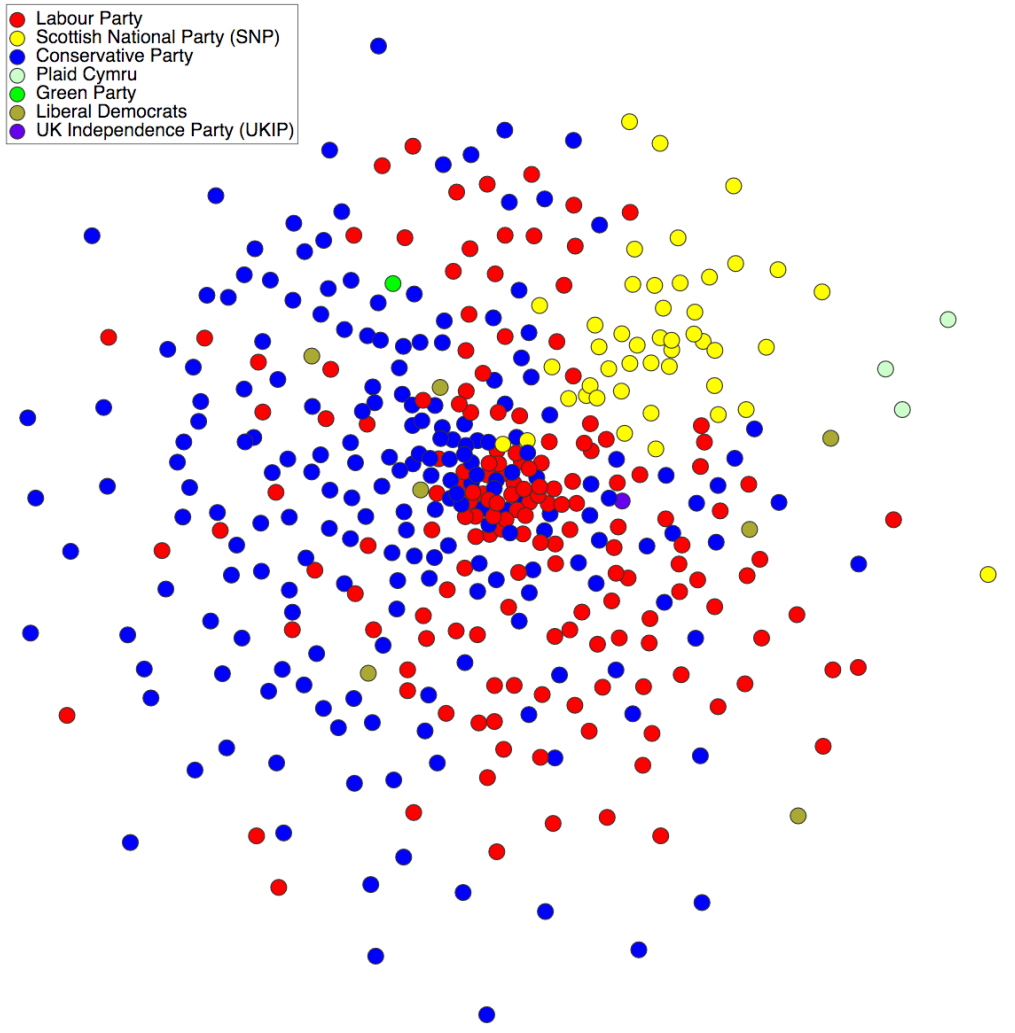

Figure 13: Log-likelihood word similarity network of UK MPs on Twitter.

case study described in section 3.3. A user would explore the MP network and explore the differences between two or more political parties or groups within those parties.

Within the linguistic dimension there are at least three prominent levels: lexical, grammatical and semantic levels, as exemplified in our case studies. The exploration could proceed as described for the structural use case above but would now be extended to cover other levels of linguistic annotation assuming that they were represented in the corpus.

The final dimension incorporated into our proposed framework is time which will assist with the exploration and visualisation of diachronic corpora. A prototypical example of this would be a Twitter corpus that has been collected over a number of months or years. The social network data would be visualised as points on a 2D time series graph. From this graph a user

can select groups of data to compare against within the key word clouds, collocation networks and social network relationships, and how each of these aspects varies over time.

This combination across three dimensions will therefore allow a user to explore the corpus on many different interconnected levels and visualisations. Employing multiple visualisations is of utmost importance to counter deficiencies in some methods (such as information loss and uncertainties in force-based methods) and to ensure the various model abstractions align with analysis tasks (Chuang et al., 2012). We envisage our framework would be developed by achieving interoperability between the existing tools rather than developing a new standalone system.

5 Conclusion and Future Work

In this paper, we have proposed the idea of using interactive information visualisation techniques for supporting the corpus linguistics methodology at multiple levels of analysis. We have highlighted tools and techniques that are already used in corpus linguistics that can be considered as visualisation: concordances, concgrams, collocate clouds, and described new methods of collocational networks and exploratory language analysis in social networks. In addition, we described the key word and semantic cloud approaches as implemented in the Wmatrix software.

With the CONE and GraphColl prototypes, we have proposed and illustrated a highly dynamic way of exploring collocation networks, as an example of our wish to add dynamic elements to both existing and novel visualisations. This would enhance their "data exploration" nature even further. To paraphrase Gene Roddenberry[8], 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. We have shown how the dynamic techniques align more closely to the iterative data-driven corpus linguistics methodology. 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 corpus linguist who currently needs to analyse very large sets of results by hand. In future work, we will explore techniques which are able to support longer explorations in order to avoid corruption or 'messiness' in the interface which still persists after a prolonged period of use. There is clearly a need for new static analysis techniques as well; to extract the data required as well as novel methods for displaying and exploring the data.

[8] See `http://en.wikipedia.org/wiki/Gene_Roddenberry`

Acknowledgements

Francois Taiani was involved in supervision of the original CONE project. This work was partly funded by the EPSRC vacation bursary grant awarded to David Gullick at Lancaster University. GraphColl software development was supported by the ESRC Centre for Corpus Approaches to Social Science, ESRC grant reference ES/K002155/1. The UCREL research centre supported the development of the integrated visualisation framework.

References

Baker, P. (2006). *Using Corpora in Discourse Analysis.* Continuum.

Baker, P. (2010). *Sociolinguistics and Corpus Linguistics.* Edinburgh University Press.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–231.

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

Brezina, V., McEnery, T., and Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2):139–173.

Cheng, W., Greaves, C., and Warren, M. (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics*, 11(4):411–433.

Chuang, J., Ramage, D., Manning, C. D., and Heer, J. (2012). Interpretation and trust: Designing model-driven visualizations for text analysis. In *ACM Human Factors in Computing Systems (CHI).*

Culy, C. and Lyding, V. (2010). Double Tree: An Advanced KWIC Visualization for Expert Users. In *14th International Conference Information Visualisation*, pages 98–103.

Davies, M. (2009). The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2):159–190.

Dork, M. and Knight, D. (2015). WordWanderer: A navigational approach to text visualisation. *Corpora*, 10(1):83–94.

Garside, R. and Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In Garside, R., Leech, G., and McEnery, T., editors, *Corpus Annotation: Linguistic Information from Computer Text Corpora.*, pages 102–121. Longman.

Gullick, D., Rayson, P., Mariani, J., Piao, S., and Taiani, F. (2010). CONE: COllocational Network Explorer [Computer Software]. `http://ucrel.lancaster.ac.uk/cone/`.

Heer, J. and Shneiderman, B. (2012). Interactive dynamics for visual analysis. *Queue*, 10(2):30:30–30:55.

Heimerl, F., Lohmann, S., Lange, S., and Ertl, T. (2014). Word cloud explorer: Text analytics based on word clouds. In *2014 47th Hawaii International Conference on System Sciences*, pages 1833–1842.

Hilpert, M. (2011). Dynamic visualizations of language change: Motion charts on the basis of bivariate and multivariate data from diachronic corpora. *International Journal of Corpus Linguistics*, 16(4):435–461.

Huang, Y., Guo, D., Kasakoff, A., and Grieve, J. (2015). Understanding U.S. regional linguistic variation with twitter data analysis. *Computers, Environment and Urban Systems*, 59:244–255.

Johansson, S., Leech, G., and Goodluck, H. (1978). *Manual of information to accompany the Lancaster-Oslo/Bergen corpus of British English, for use with digital computers.* Department of English, University of Oslo.

Jänicke, S., Franzini, G., Cheema, M. F., and Scheuermann, G. (2015). On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. In Borgo, R., Ganovelli, F., and Viola, I., editors, *Eurographics Conference on Visualization (EuroVis) - STARs*. The Eurographics Association.

Kauppinen, T. and de Espindola, G. M. (2011). Linked open science-communicating, sharing and evaluating data, methods and results for executable papers. *Procedia Computer Science*, 4:726 – 731. Proceedings of the International Conference on Computational Science, ICCS 2011.

Keim, D. A. and Oelke, D. (2007). Literature fingerprinting: A new method for visual literary analysis. In *IEEE Symposium on Visual Analytics Science and Technology, 2007. VAST 2007.*, pages 115–122.

Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3):333–347.

Knowles, A. K., Westerveld, L., and Strom, L. (2015). Inductive Visualization: A Humanistic Alternative to GIS. *GeoHumanities*, 1(2):233–265.

Kucher, K. and Kerren, A. (2015). Text visualization techniques: Taxonomy, visual survey, and community insights. In *2015 IEEE Pacific Visualization Symposium (PacificVis)*, pages 117–121.

Leech, G. (1991). *The state of the art in corpus linguistics.*, pages 8–29. Longman.

Leech, G. (1993). 100 million words of English: a description of the background, nature and prospects of the British National Corpus project. *English Today*, 33(9).

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

Meirelles, I. (2011). Visualizing data: new pedagogical challenges. In *Selected Readings of the 4th Information Design International Conference*, pages 73–83.

Murrieta-Flores, P., Baron, A., Gregory, I., Hardie, A., and Rayson, P. (2015). Automatically analysing large texts in a GIS environment: the Registrar General's reports and cholera in the nineteenth century. *Transactions in GIS*, 19(2):296–320.

Oelke, D., Kokkinakis, D., and Malm, M. (2012). Advanced visual analytics methods for literature analysis. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 35–44. Association for Computational Linguistics.

Perkuhn, R. (2007). Systematic exploration of collocation profiles. In *Proceedings of Corpus Linguistics 2007, Birmingham, UK.*

Phillips, M. (1985). *Aspects of Text Structure: An investigation of the lexical Organisation of Text*, volume 52. North Holland.

Prentice, S., Taylor, P., Rayson, P., and Giebels, E. (2012). Differentiating act from ideology: evidence from messages for and against violent extremism. *Negotiation and Conflict Management Research*, 5(3):289–306.

Pumfrey, S., Rayson, P., and Mariani, J. (2012). Experiments in 17th century english: manual versus automatic conceptual history. *Literary and Linguistic Computing*, 27(4):395–408.

Rayson, P. (2006). *AHRC e-Science Scoping Study Final report: Findings of the Expert Seminar for Linguistics.* AHRC e-Science Scoping Study (eSSS) project report.

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

Rayson, P., Archer, D., Piao, S., and McEnery, T. (2004a). The UCREL semantic analysis system. In *Proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004), 25th May 2004, Lisbon, Portugal.*, pages 7–12.

Rayson, P., Berridge, D., and Francis, B. (2004b). Extending the cochran rule for the comparison of word frequencies between corpora. In *7th International Conference on Statistical analysis of textual data (JADT 2004)*, pages 926–936.

Scott, M. (1997). Pc analysis of key words – and key key words. *System*, 25(2):233–245.

Scrivner, O. and Kubler, S. (2015). Tools for digital humanities: Enabling access to the old occitan romance of flamenca. In *Proceedings of NAACL-HLT Fourth Workshop on Computational Linguistics for Literature*, pages 1–11, Denver, Colorado.

Siirtola, H., Räihä, K.-J., Säily, T., and Nevalainen, T. (2010). Information visualisation for corpus linguistics: Towards interactive tools. In *IVITA'10*, pages 33–36, Hong Kong. ACM.

Siirtola, H., Säily, T., Nevalainen, T., and Räihä, K.-J. (2014). Text variation explorer: Towards interactive visualization tools for corpus linguistics. *International Journal of Corpus Linguistics*, 19:3:418–429.

Sinclair, J. (2004). *Trust the text: language, corpus and discourse.* Routledge.

Sinclair, S. and Rockwell, G. (2016). Voyant tools. `http://voyant-tools.org/`.

Smith, N., Hoffmann, S., and Rayson, P. (2008). Corpus tools and methods, today and tomorrow: Incorporating linguists' manual annotations. *Literary and Linguistic Computing*, 23(2):163–180.

Vuillemot, R., Clement, T., Plaisant, C., and Kumar, A. (2009). What's being said near "Martha"? Exploring name entities in literary text collections. In *IEEE Symposium on Visual Analytics Science and Technology, 2009. VAST 2009.*, pages 107–114.

Wattenberg, M. and Viégas, F. B. (2008). The word tree, an interactive visual concordance. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1221–1228.

Williams, G. (2002). In search of representativity in specialised corpora: Categorisation through collocation. *International Journal of Corpus Linguistics*, 7(1):43–64.

Xu, J., Tao, Y., and Lin, H. (2016). Semantic word cloud generation based on word embeddings. In *2016 IEEE Pacific Visualization Symposium (PacificVis)*, pages 239–243.

Zeldes, A., Ritz, J., Lüdeling, A., and Chiarcos, C. (2009). Annis: A search tool for multi-layer annotated corpora. In *Proceedings of Corpus Linguistics*, Liverpool, UK.