Semantic Pathways: A novel visualisation of varieties of English

Claire Brierley,[1] Eric Atwell,[1] Chris Rowland[2] and John Anderson[2]
[1] School of Computing, University of Leeds
[2]Duncan of Jordanstone College of Art and Design, University of Dundee

Abstract

Semantic Pathways *is a corpus exploration tool with a unique visual interface in which keyword extraction and keyword-based document clustering have been implemented in order to facilitate insight forming.* Semantic Pathways *combines corpus comparison techniques from Corpus Linguistics with aesthetically-driven design and interaction, to produce fluidly interactive information exploration. In addition, users can access* Semantic Pathways *via a command-line interface, where integration with Python and NLTK offers additional benefits. We describe system operation from the user's perspective, and then use the tool for corpus comparison of different varieties of English with the LOB and Brown corpora as test and reference sets, demonstrating its novelty in gisting an entire document collection and speedy and intuitive exploration of lexical usage across the document set.*

1 Introduction

2012 saw the first international conference on the application of visualisation tools and techniques for Linguistics: Advances in Visual Methods for Linguistics – AVML 2012. Popular topics included: the use of maps and Geographic Information Systems (GIS) for studying regional dialectology and dialect change (Montgomery and Stoeckle 2012; Priestley *et al*. 2012); phonetics-based visual representations such as vocal tract sections, speech waveforms, and acoustic spectra (Fabricius *et al*. 2012; Huckvale 2012); visual methods for discourse analysis and the study of conversational interactions (Angus *et al*. 2012); and visualising complex ontologies for knowledge representation, management and reasoning (Dukes and Atwell 2012).

At the cutting-edge of AVML were new tools and techniques for visualising the properties of linguistic corpora. These are exemplified by the University of Dundee's *Semantic Pathways* system. *Semantic Pathways* is a visualisation tool for *gisting* or summarizing large document collections, enabling speedy disclosure of high-level themes and topics, and intuitive exploration and query of the entire document set. Information visualisation as a response to information overload, and as a means of supporting human reasoning, is central to the new and emerging field of Visual Analytics. Methods and systems for data handling, data mining and sense-making, particularly in the context of national security, have been discussed at the recent *Visual Analytics Workshops* at Imperial College London in 2010 and 2011. These incorporate statistical methods for *text* data mining within large, unstructured text collections. Developing innovative tools for visualising the experimental results of Text Analytics is also a priority for the Corpus Linguistics research community (Rayson and Mariani 2009).

The novelty of *Semantic Pathways* from the perspective of Corpus Linguistics is that it enables text analytic investigation of a set of unstructured documents via an interactive visual interface. This is important for applications which perform cluster analysis, relationship modelling and sense-making on large, free text data collections: the typical case load for intelligence analysts, for example. Thus the focus of *Semantic Pathways* is on *document* space as opposed to word space. However, instead of confronting the user with abstract representations such as a histogram of concept clusters, the typology of document space is initially represented and summarised by a finite set of statistically significant keyword triggers, computed via the standard log-likelihood metric, in an aesthetically-driven design approach.

In this paper we evaluate a prototype version of *Semantic Pathways* via corpus comparison of American and British English in the Brown (Francis and Kučera 1979) and LOB (Johansson *et al.* 1978, 1986) corpora. Our experiments serve as a 'real-world' use case (Maguire and Bevan 2002) for formulating and understanding user requirements and potential user interaction with the system. It is often the case that users do not know in advance exactly what they want from a new system (Olphert and Damodaran 2002), and trials have led to insightful recommendations for both contributors: the visualisation team at Dundee and the text analytics team at Leeds. The paper is structured as follows: we describe widely-used tools for visualising linguistic data and their underpinning theory (Sections 2 and 3); we then give a detailed description of *Semantic Pathways* (Section 4) before reporting on two sets of experiments (Section 5) and drawing conclusions (Section 6).

2 Information visualisation for Corpus Linguistics: Concordance

While corpus linguists have pioneered web data analysis (Kilgarriff and Grefenstette 2003) and information extraction from large datasets, they have been more conservative in their use of visualisation (Rayson and Mariani 2009). However, use of concordance software is a well-developed Corpus Linguistics technique for visualising words in context and their lexical patterns of association or collocates. A concordance lists all occurrences of a search word or character span within a text on separate lines, embedded centrally in a window of *N* preceding and subsequent words/spans. Collocations may extend to lexical phrases as evidence of formulaic language in a given domain such as English academic writing (Oakey 2002) or idiomatic language use (McEnery and Hardie 2011). Widely-used, state-of-the-art toolkits for Corpus Linguistics incorporating concordance software are WordSmith Tools (Scott 2012) and Wmatrix (Rayson 2008). There is further discussion of these toolkits in Section 3.2 of this paper.

3 Research paradigm for corpus analysis and comparison

The Corpus Linguistics research paradigm involves compilation and automated, quantitative analysis of a *corpus* or sample of naturally-occurring language texts capturing empirical data on the concept(s) or behaviour(s) being studied. It is therefore versatile and has cross-disciplinary applications, since *any* machine-readable text-based data is amenable to this approach. It is also scalable and can accommodate large quantities of data from different fields: *web-as-corpus* is now an established technology (Baroni *et al*. 2006). Finally, computational, Text Analytics techniques for quantitative analysis and modelling can complement introspective approaches based on manual annotation of text data, thus facilitating knowledge exchange and collaboration from different methodological standpoints and traditions.

Corpus Linguistics, like Machine Learning, is concerned with pattern-seeking, where significant linguistic patterns are determined via counts and via comparison with linguistic norms. Thus, if the researcher can define the concept or behaviour being studied via a set of countable linguistic features, then even complex stylistic phenomena such as metaphor (Culpeper *et al.* 2009) and genre (Abu Shawar and Atwell 2003) can be explored through the standard Corpus Linguistics approach based on corpus comparison.

The basic comparison is to measure deviation from linguistic norms by comparing the text(s) under investigation (i.e. the test set) with the norms of the language as a whole, represented by a general reference corpus. *Keyness* is a central concept: the researcher is looking to identify whether phenomena of

interest *only* occur in the test set, or occur *significantly more* or *less* in the test set than the reference set. This Corpus Linguistics technique is known as *Keyword Extraction*.

3.1 Keyword Extraction

Keyword Extraction relates in principle to Information Retrieval in that the starting point for both methods is to conceive of the text as a bag of words, and also in that high frequency words specific to that text then perform a discriminatory function. Ascertaining keywords is a statistical process effected via formal comparison of word frequency distributions for a dataset of interest (i.e. observed frequencies in a test set) with their expected frequencies inherent in a suitable reference dataset (i.e. a representative cross-section of the language). Keywords are thus words of unusual frequency or infrequency in the test set relative to the reference set, where unusualness is defined by some pre-determined confidence level of statistical significance. Keywords in this context differ, therefore, from common parlance, where the notion of keyness denotes words viewed *subjectively* as key.

3.2 WordSmith and Wmatrix

WordSmith and Wmatrix expedite corpus comparison via a statistical approach to Keyword Extraction, where keywords in the dataset are *not* purely intuitive but instead represent lexical items of uncommon frequency or infrequency relative to a reference corpus. In addition to extracting keywords, Wmatrix automates frequency profiling of key syntactic categories and key semantic domains, and visualises significant differences via relative font sizes in word and tag clouds in a graphical user interface. In both applications, keywords are retrieved via a staged procedure whereby: (i) wordlists and word frequency distributions are computed for the test and reference sets; and (ii) apparent overuse or underuse is verified via the log likelihood ratio test. Keywords resulting from this comparison mitigate against researcher bias. They are taken to be revelatory of significant lexical differences between the two datasets in terms of *aboutness* and style, and have been used in genre analysis. For example, WordSmith has been used in the analysis of illness (Koteyko and Carter 2008) and refugee/asylum seeker narratives (Baker *et al.* 2008); while Wmatrix has been used in studies on entrepreneurship (Doherty *et al.* 2006), political science (Beigman Klebanov *et al.* 2008), and human-computer chatbot dialogues (Abu Shawar and Atwell 2005). A further point is that the meaning of language patterns uncovered through this approach still depends on human interpretation.

3.3 Metrics: Log likelihood

The likelihood ratio test compares the fit of two models, in this case the test and reference corpora, and expresses how many times more likely the data (i.e. words or other lexical items) are under one model than the other. The resulting log likelihood (LL) statistic is also used to compute a *p-value* denoting the probability of obtaining a test statistic at least as extreme as the one observed, assuming that there is no difference in the test and reference sets. In Corpus Linguistics, it is usual to set a very challenging LL statistic and p-value to determine whether there is a *significant* difference between the two datasets: WordSmith and Wmatrix implement a cut-off value of LL = 6.63 (corresponding to a p-value of 0.01) to generate keywords; and this standard is retained in *Semantic Pathways*.

4 Semantic Pathways

Semantic Pathways is a visual corpus exploration tool in which keyword extraction and keyword-based document clustering have been implemented in order to facilitate insight forming. *Semantic Pathways* combines corpus comparison techniques from Corpus Linguistics with aesthetically-driven design and interaction to produce fluidly interactive information exploration. In the following sections, a description of how the system operates from the user's perspective is given, followed by a discussion of the design strategy that was adopted in the development of the system. Finally an overview of the technical implementation of *Semantic Pathways* is discussed.

4.1 System description

On starting *Semantic Pathways* the user is first presented with an overview of the corpus which is to be explored. This overview is depicted as a ten-keyword summary, or 'gist', of the corpus in a region of the display called the *Semantic Space*. The name of the corpus is displayed in the centre of the *Semantic Space* to signify that the user is looking at a 'collection level gist' and the extracted keywords are arranged radially in a layout which is termed a 'fan'. The lower region of the display is termed the *Source Space,* and here a bar graph representation of the entire corpus is initially shown with genre boundaries denoted by colour. This initial state is shown in Figure 1:

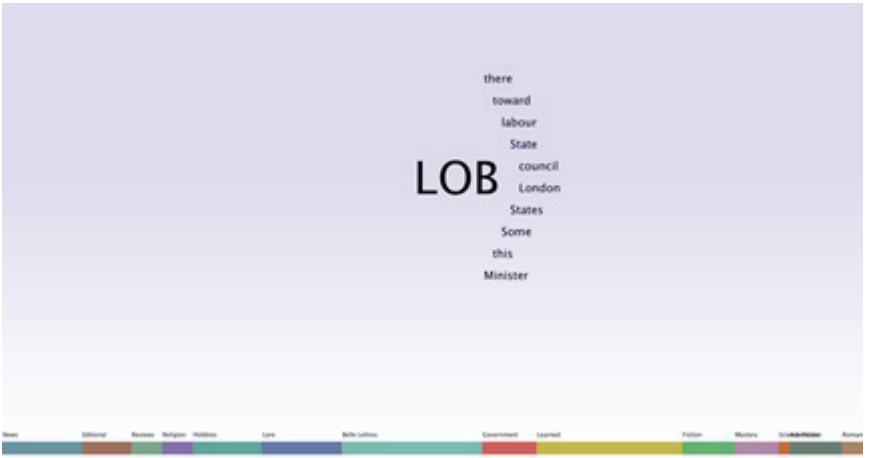

Figure 1: Initial state of Semantic Pathways – collection level gist

The user may increase the number of radial keyword fans displayed in order to progressively reveal more detail in the keyword gist. Inner fans contain keywords with a higher value for the log-likelihood statistic, and outer fans contain keywords with progressively lower values (Figure 2):

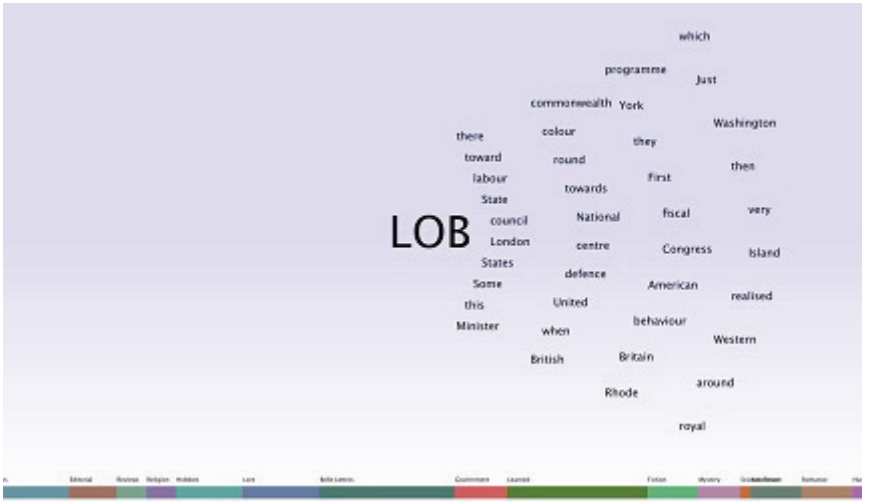

Figure 2: Radial fans for progressive disclosure of keywords

On seeing a keyword of interest, the user may select it, and the keyword then animates into the centre of the display to become the current 'keyword of interest'. The original collection level gist recedes into the background and the source space is updated to show documents, depicted as coloured rectangles, which cluster around the keyword of interest. The system now shows a 'document level gist' in which keywords are extracted from the document cluster and drawn radially around the new keyword of interest in the *Semantic Space* (Figure 3):

Figure 3: Document level gist

Each keyword in this gist is the top ranking keyword, by log-likelihood statistic, of each document in the cluster and the keywords are rendered using the same colour as their respective documents. Again the user may increase the number of keyword fans to reveal more keywords for each document in the cluster. The user may then select a new keyword, chosen from the document level gist in order to define a new keyword of interest and a new document level gist. Interaction with the corpus may proceed in this fashion in order to build up a *Semantic Pathway.* A history of the corpus exploration accumulates as 'waypoints' along the z-axis serving as a visual reminder to the user of the *semantic pathway* that has been taken (Figure 4). It is easy for the user to retrace the *semantic pathway* by stepping back through the waypoints and to initiate new or alternate *semantic pathways* at any juncture.

Figure 4: History 'waypoints' along the z-axis

In the source space the user may easily access the full text of any document in the current cluster. Text in the document view is interactive so the user may make any term in the text a keyword of interest. A concordance view shows the keyword of interest in the context in which it appears in the corpus (Figure 5). The concordance text is also interactive, and selecting any term in the concordance view causes an in-place redraw to show the new keyword of interest in context.

Figure 5: The concordance view

4.2 Design

The aesthetic of the design has been a primary concern for the designers of *Semantic Pathways* and it is a considered blend of typography, spatial layout and layering. The designers have investigated an interaction style that directly relates both to the visual aesthetic and to the task of corpus exploration. The interaction style is based on the tenet of *data as interface*. The following sections will discuss the visual aesthetic and the interaction style.

4.2.1 Visual aesthetic

In Schneiderman (1996) the assertion is made that information exploration should be a joyous experience, and this may be interpreted as alluding to the aesthetic of the software environment in which the information exploration takes place. In Viégas *et al.* (2009) the importance of visual aesthetic is explored in the context of a study of users of the popular text visualisation tool *Wordle*. Viégas *et al.* (2009) shows that attention to aesthetic properties such as typeface, font size, font colour and layout composition can have a profound effect on extending the reach of information visualisation to a mass audience. The same study reveals that *Wordle* visualisations have the capability to assist memory and learning and that this capability is related to their aesthetic qualities. So an aesthetic approach should be considered an essential element of effective information exploration.

Semantic Pathways uses spatial layout of typographic elements to subtly delineate conceptually and functionally separate areas of the visualisation. The 2-dimensional layout of the *keyword of interest* and the radial keyword *fans* and the 3-dimensional layout of the *waypoints* that record the user's exploration of the corpus combine to give subtle and implicit visual cues as to how the system should be used, and largely removes the need for traditional GUI elements which would clutter the display and compromise the aesthetic experience.

Transparency and animation effects are used to emphasise the logical separation between what is current, what is history and where *waypoints* reside relative to each other along the user's chosen *semantic pathway.* Waypoint transitions are smoothly animated in order to convey with motion the sense that the user is progressing along a pathway.

4.2.2 Interaction style

An observable characteristic of many information exploration systems is their reliance on traditional graphical user interfaces drawn from the *Windows, Icons, Mouse, Pointer* (WIMP) interaction paradigm. In Endert *et al.* (2012: 473) it is noted that these traditional interaction modalities are often used to control visualisation parameters that users "do not understand and do not relate to their analytic process". Graphical interface components such as sliders, menus and text fields are external to the visual metaphor of the visualisation. A need is therefore identified to develop an interaction style which "leverages the cognitive connection formed between the user and the spatial layout" (Endert *et al.* 2012: 475). This form of fluid interaction in which the visual properties of the display and the interaction properties of the system are essentially one and the same – *data as interface* – is the defining characteristic of *Semantic Pathways'* interaction style.

4.3 Technical implementation

Semantic Pathways is implemented as a typical Model-View-Controller (MVC) architecture. The *model* layer handles text analysis and is implemented as a single Python class, the *Pathways Python Module*. The *controller* layer is implemented in Objective-C using the Python C API and is responsible for mediating between the *model* and *view* layers. The *view* layer is implemented in OpenGL and *Quartz Composer*, a proprietary visualisation framework available exclusively for the Mac OS X platform. The *view* layer handles all display and user interaction functionality. The architecture is shown in Figure 6. The entire application may also be executed as a stand-alone Python script in a Command-Line Interface (CLI) which permits its use on any platform which can run a Python interpreter. Each of these layers is discussed in the following sections.

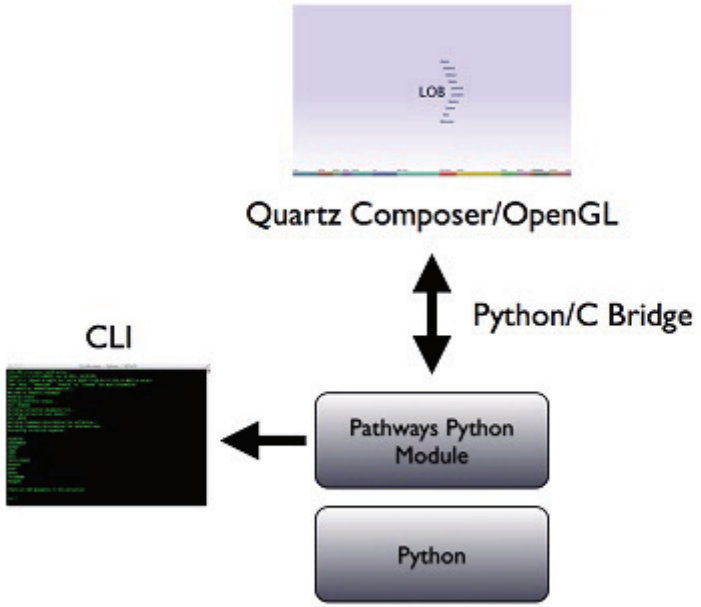

Figure 6: Semantic Pathways Architecture

4.3.1 Model layer

The *model* layer comprises a single Python class called *Collection*. In the class constructor the open source Natural Language Tool Kit (NLTK) is accessed in order to load the test and normative corpora. These corpora may be selected from corpora supplied with the NLTK (e.g. Brown) or alternatively the `PlaintextCorpusReader` class of the NLTK may be used to load external corpora. The Collection class constructor performs a number of important initialisation routines. The test corpus and normative corpus (also known as the reference corpus) are first assigned to the instance variables `test` and `normative`. Then frequency distributions for each corpus are computed using the NLTK probability module and assigned to the instance variables `test_fdist` and `normative_fdist`. A list of the filenames of the documents in the test corpus is extracted using the NLTK `fileids()` method and these are stored in the instance variable `documentNames`. The vocabulary (i.e. set of words) for the test corpus is computed and assigned to a local variable `vocab`. Then all of

the collection level keywords are computed by performing the log-likelihood test (described in Section 3) for each term in `vocab`. Terms are tested both for their log-likelihood score and for their *keyness* (as described in Section 3) according to a *keyness preference* supplied as an argument to the Collection class constructor. This argument may be '+' (positively key terms preferred), '-' (negatively key terms preferred) or '?' (mixed keyness preferred). Terms which score 6.63 or greater in the log-likelihood test *and* match the preferred keyness are retained as keywords and all other terms are discarded. A minimum term length is also specified and by default this is four, i.e. terms of three or less characters are also discarded, thus helping to minimise noise. *Semantic Pathways* does not utilise stemming, stopwords, lemmatisation, downcasing or any other text preprocessing or noise removal technique. The extracted keywords are stored in the instance variable `collectionLevelKeywords` as an array ordered by log-likelihood score descending. Pseudocode for the initialisation of the Collection class is shown in Code Listing 1:

Code Listing 1

```
class Collection:
    init (args):
        self.normative = args[normative]
        self.test = args[test]
        self.preferred_keyness = args[keyness]
        self.normative_fdist = FreqDist(self.normative)
        self.test_fdist = FreqDist(self.test)
        self.documentNames = self.test.fileids()
    self.collectionLevelKeywords = []
    vocab = set(self.test)

    for word in vocab:
        if len(word) > 3:
            #frequency in test corpus
            a = test_fdist[word]
            #frequency in normative corpus
            b = normative_fdist[word]
            #word count of test corpus
            c = len(self.test)
            #word count of normative corpus
            d = len(self.normative)

            #compute log-likelihood
            LL = self.computeLogLikelihood(a, b, c, d)

            #compute keyness
            keyness = self.computeKeyness(a, b, c, d)

            if LL >= 6.63:
                if keyness == self.preferred_keyness:

self.collectionLevelKeywords.append(word)
```

The methods `computeLogLikelihood()` and `computeKeyness()` compute the log-likelihood score and keyness respectively for a given term according to the method described in Section 3. Once initialisation of the Collection class has completed all functionality required to drive the visualisation is provided via the *model* layer public API which comprises three instance methods:

- `Collection.collectionKeywords()`
  This method simply returns the array of collection level keywords;
- `Collection.documentsForQuery(word)`
  This method returns the cluster of documents in which the term word is a keyword (using the same method described above but computed from the document text with respect to the test corpus);
- `Collection.keywordsForDocument(document)`
  This method returns all keywords for a given document (using the same method described above but computed from the document text with respect to the test corpus).

4.3.2 Controller layer

The *controller* layer is a lightweight application runtime implemented in Objective-C on the Mac OS X platform. The Python C API is used in order to wrap the model layer public API methods and make them accessible from the runtime. The *controller* layer requests data from the *model* layer and forwards it to the *view* layer in response to user actions. Additionally the *controller* layer manages all of the application-level functionality such as history maintenance and traversal, managing the instantiation of various views (collection level gist, document level gists, concordance) and issuing requests to the *model* layer in response to user input.

4.3.3 View layer

The *view* layer is the part of the system with which the user interacts. This layer is implemented in *Quartz Composer* which is a proprietary visualisation framework and part of the Mac OS X operating system. The *view* layer comprises multiple *Quartz Composer* visualisation graphs, one each for the collection level gist, document cluster and concordance views and an arbitrary number for the document level gists. These graphs are constructed using a dataflow programming paradigm an example of which is shown in Figure 7. The graphs are rendered in an OpenGL context and managed by private methods of the *controller* layer.

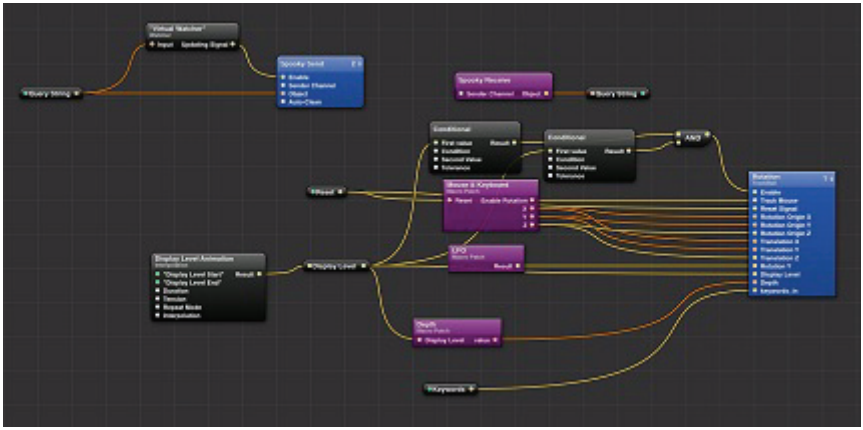

Figure 7: Example of a Quartz Composer graph (collection level gist)

5 *Trialling and experimenting with* Semantic Pathways*: Visualising varieties of English*

Semantic Pathways can be run on other platforms, but in these demos the tool is located in a Windows Desktop folder in which we then run Python and NLTK scripts from a command- line interface. Various versions of the LOB and Brown corpora, representing British and American English, have been placed alongside other linguistic datasets in NLTK, and NLTK's corpus reader source code has been modified slightly to read in this new material. In the first experiment (Sections 5.1 to 5.4 inclusive), we evaluate functionality in the *Semantic Pathways* prototype from the perspective of Corpus Linguistics, to suggest modifications and improvements. In the second experiment (Section 5.5) we trial the updated version.

5.1 Experiment 1: Tokenization issues

Having imported NLTK and the LOB and Brown corpora, we can import and call the *Semantic Pathways* `Collection()` method on both LOB and Brown as reference sets, to generate collection-level keywords for LOB with respect to Brown and vice versa, shown in Table 1 and Figure 8. We can see that differences in manual pre-processing and editing of these standardised corpora result in keywords of *apparent* significance only, which on further investigation, are attributable to differences in use of case and tokenization of enclitics, following manual editing of Word-Initial Capitals and enclitics in LOB (Atwell 1981, 1982).

Table 1: Corpus comparison at collection level (cf. Figure 1) in *Semantic Pathways*: challenges posed by standardised corpora for Keyword Extraction

Collection KWs: Brown as reference set	**Collection KWs: LOB as reference set**
There	This
toward	There
labour	They
State	When
council	Mrs.
London	What
States	didn't
Some	don't
this	program
Minister	That

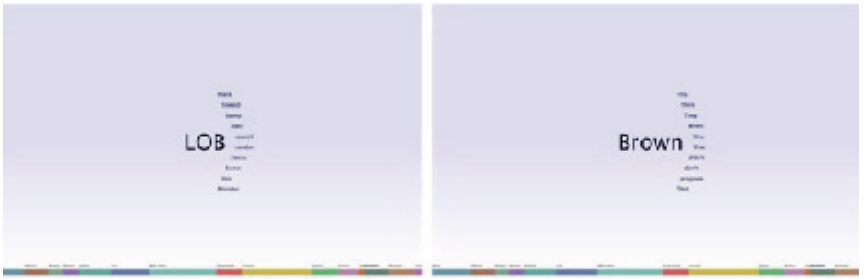

Figure 8: Collection gists for LOB with respect to Brown and vice versa

LOB, for example, reserves upper case for proper nouns. Hence the appearance of *Some* (capitalised) as a keyword with Brown as reference set is puzzling until we inspect concordance lines for this item in LOB and discover it is actually the name of a horse: *Some Alibi*. Similarly, most items generated via LOB as reference set are due to design decisions in corpus build: Brown retains encliticised forms {*didn't; don't*}, plus sentence-initial capitalisation resulting in a preponderance of function words as apparently 'key': {*This; There; When; What; That*}. So, we tried to reduce such artificial results by reading in lowercase versions of both datasets: `brown_lc and lob_lc`.

5.2 Experiment 1: Overuse and underuse at collection-level

Table 2 and Figure 9 display the ten most significant collection-level keywords recalculated from lowercase versions of our datasets. We are immediately struck by the duplication of some keywords in both lists: *toward*; *labour*; *states*; *london*.

Table 2: Collection-level keywords recalculated from lowercase versions of LOB and Brown for more informative comparison of these corpora

Collection KWs: Brown as reference set	**Collection KWs: LOB as reference set**
toward	mrs.
labour	don't
states	didn't
london	program
state	**toward**
colour	it's
towards	**labour**
round	cannot
centre	**states**
federal	**london**

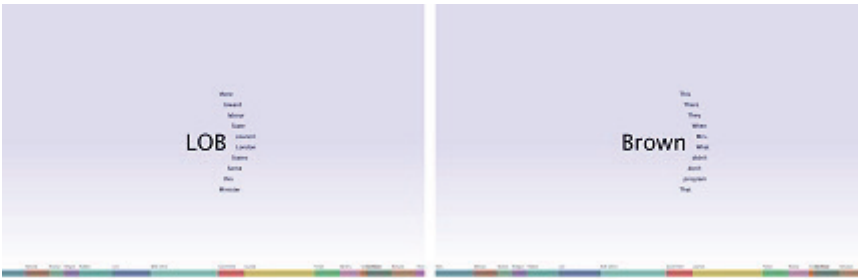

Figure 9: LOB (lowercase) with respect to Brown (lowercase) and vice versa

On further investigation, this occurs because the initial *Semantic Pathways* prototype generates items which are *positively* and *negatively* key, but does not as yet distinguish these for the user in the same way as Wmatrix, which inserts + or

- to denote unusual term frequency or infrequency vis-à-vis the reference set. *Toward* in Table 2 is an interesting case in point. When we inspect comparative statistics for this term, we discover that it is significantly *under*-used in LOB and *over*-used in Brown (Table 3):

Table 3: Significant under-use of *toward* in LOB in comparison to Brown

Brown as reference set					
Term: toward	Observation (LOB): 14	Expected (Brown): 386	WC (LOB): 1131976	WC (Brown): 1161192	LL: 423.73

On the other hand, British English favours *towards* as a preposition, as comparative statistics again clearly show (Table 4). This can be further investigated by calling NLTK's `concordance()` method on the corpus comparison object within *Semantic Pathways* to review usage in concordance lines.

Table 4: Significant *over*-use of *towards* in LOB in comparison to Brown

Brown as reference set					
Term: towards	Observation (LOB): 318	Expected (Brown): 64	WC (LOB): 1131976	WC (Brown): 1161192	LL: 190.80

5.2.1 Hands-on definition of user requirements to enhance system functionality
This exercise is a collaboration between researchers in visualisation and natural language processing to develop a Visual Analytics tool for exploring large document collections. So far, our hands-on approach has demonstrated that differences in tokenization and the treatment of enclitics even in standardised corpora are challenging for corpus comparison via keyword extraction. We have also demonstrated the importance of distinguishing between words that are positively and negatively key, and the importance of user access 'on demand' to collection and document-level word frequency statistics. Both recommendations have now been implemented in the latest version of *Semantic Pathways*: a method call (i.e. mouse click in the GUI version) on the corpus comparison object `data1.stats()` displays the following information on highlighted words of interest; see Table 5:

Table 5: Statistics for a sample of collection-level keywords from the corpus comparison object `data1` (LOB *wrt* Brown)

Term	Test Count	Ref Count	Test Total	Ref Total	LL score	Keyness
toward	14	386	1131976	1161192	423.73	(-)
labour	276	4	1131976	1161192	353.21	(+)
states	123	603	1131976	1161192	333.72	(-)
london	492	89	1131976	1161192	318.24	(+)

5.3 Experiment 1: Iterative exploration of Semantic Space *at collection-level*

Semantic Pathways fan() method analyses data comprising the reference versus test set: in this case, either data1 (with Brown as reference set) or data2 (with LOB as reference set). It displays keywords in descending order of significance in *N* batches of ten words where batch size corresponds to the value of the argument supplied to the method. For example, data1.fan(5) results in a display of the top fifty collection-level keywords for the LOB corpus with respect to Brown (Table 6 and Figure 10), thus presenting the researcher with a gist of the entire collection.

Table 6: 50 collection-level keywords for LOB versus Brown in descending order of significance

KWs 1-10	KWs 11-20	KWs 21-30	KWs 31-40	KWs 41-50
toward	commonwealt	britain	chicago	should
labour	h	rhode	providence	nigel
states	defence	washington	favour	organisation
london	british	which	lord	jazz
state	programme	around	corps	congress
colour	york	realised	whilst	behaviour
towards	he’s	united	favour	entire
round	fiscal	cent	been	mercer
centre	american	minister	negro	minister
federal	behaviour	very	alan	dollars
	council			

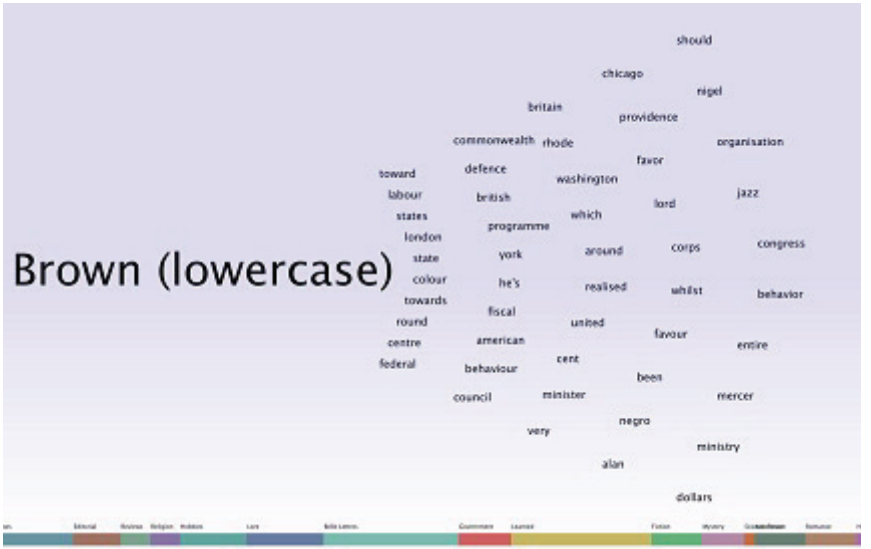

Figure 10: Collection lever keywords for LOB versus Brown

5.4 Experiment 1: Zooming in on particular keywords

Our research interest in this demo is to trace the 'semantic pathway' of particular lexical items across genres, looking at usage and word sense; and we have chosen words that emerge as positively and negatively key in both corpus comparison objects: `data1` (LOB wrt Brown), and `data2` (Brown wrt LOB). Hence, in addition to *toward* (Section 5.2), we are looking at *states* and *london* in the initial collection level gists from Table 2.

5.4.1 Zooming in on states

The intuitive and transparent interface in *Semantic Pathways* initiates corpus comparison at *document* level simply by mouse click: selecting a word of interest in the display initiates a recalculation of the visualisation such that the item is then drawn at the centre of the semantic space and a new gist is drawn radially round it (cf. Figure 3). Using the tool with Python and NLTK, this same functionality is available via the `query()` method. For example, `data2.query ('states')` returns a list of all documents in Brown in which the term *states* is significant, and where each document is represented by its most significant keyword, as well as its filename. Colour coding in the *Semantic Pathways* GUI, mapping between filenames in the source space and keywords in the semantic

space, provides instant insight into usage of lexical items across genre (Table 7 and Figure 11):

Table 7: A sample *only* of Brown corpus documents (represented by their most significant keyword mapped to their filename) in which the term *states* is also significant

Brown (A)	**Brown (B)**	**Brown (E)**	**Brown (F)**	**Brown (G)**	**Brown (H)**	**Brown (J)**
administration a04	trujillo b01	boat e06	fort f17	south g01	vehicles h04	hypothalamic j17

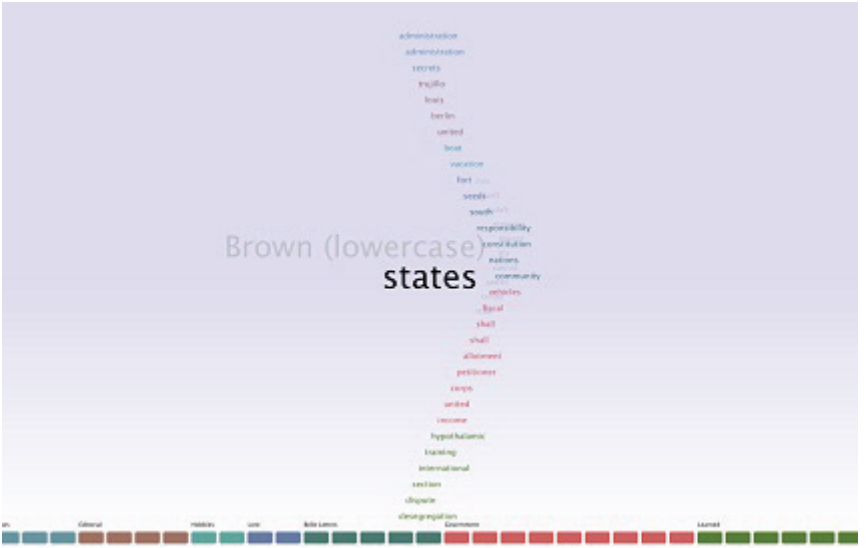

Figure 11: Document cluster for the term states

In the corpus comparison object `data2` (Brown *wrt* LOB), *states* emerges as a keyword in the following categories: press reportage (A); press editorials (B); skills, trades and hobbies (E); popular lore (F); belles lettres (G); miscellaneous (H); learned and scientific writing (J). Assuming *states* refers to the US (or states within the US) in documents from categories A and B, and accessing the entire text of a document via the following method call `data2.more(file-`

`name)` on the `data2` object, we skim-read the Brown collection, focusing mainly on genre boundaries, and find that, in all probability, the term *states* generally refers to the US (or states within the US) until we encounter learned and scientific writing (Table 8):

Table 8: A rapid snapshot of word use and changes in word sense across genre boundaries in the Brown corpus given by *Semantic Pathways*

Brown Category	**Document KW and Filename**	**Sample Text**
Brown (E)	**boat** (ce06)	...many **states** have laws regulating the use of boat trailers...
Brown (F)	**seeds** (cf34)	...corn and wheat supply most of the starch in the united **states**...
Brown (H)	**allotment** (ch14)	...the appropriations, allotment base, federal grants to **states** and state matching funds for this part of the grant program...
Brown (J)	**hypothalamic** (cj17)	...the emotional **states** produced by drugs influence the cortical potentials in a characteristic manner...

5.4.2 Zooming in on london

We have queried *london* in `data1` (LOB *wrt* Brown) via a similar method call to that used for *states* in the previous section, returning in this case twenty documents in which *London* appears with unusual frequency, and where each document is represented by its most significant keyword mapped to its filename: {*police/a02; clore/a09; ascot/a10*} and so forth. A further method call of `data1.queryFan(10)` then displays a gist of each of these twenty texts, seeming to give different 'views' or cultural perspectives on *London*, as sampled in Table 9:

Table 9: *Semantic Pathways* gists documents via their ten most significant keywords

Provenance	**Gist**
ascot/a10	*ascot; plane; duke; queen; stokowski; baudouin; baby; royal; queen's; honeymoon*
middlesex/f16	*middlesex; labour; party; london; executive; parties; county; regional; merger; national*
muggeridge/g14	*muggeridge; satirist; punch; satirize; confidence; that; engine; malcolm; piece; hitler*
jewish/g15	*jewish; palestine; zionist; judaism; yishuv; ahad; hilfsverein; nationalism; zionism; weizmann*
physics/j15	*physics; college; philosophy; science; scientific; experimental; mathematics; textbook; natural; hooke*
ware/j39	*ware; pottery; polychrome; jugs; lesnes; abbey; saintes; base; fragments; lustre*

We focus on document *cg15* where the most significant term is *jewish*. At present, *Semantic Pathways* does not display statistics for a particular word in a given document *wrt* the overall count for that term in the entire corpus: this could be useful. However, we can view the entire text of our document via the `data1.more('cg15')` call, and obtain the counts: {*london (7); jewish (20); palestine (14); zionist (14); zionism (3); weizmann (3)*}. This is an interesting, if dated, biographical text (Simon 1961) complete with biblical insertions: '*...anglo-jewry...was for him no better than a whited sepulchre...*'; '*...old friends...saved him from complete isolation. but he remained a stranger in a strange land...*' and phrases/standpoints which today may seem politically incorrect: '*...the emergence of a new hebrew type of life in palestine...*'; '*...the prestige of jewish palestine in the middle east*'.

5.5 Experiment 2: Corpus comparison with the updated version of **Semantic Pathways**

The updated version of *Semantic Pathways* allows the user to specify whether they want the system to display words that are positively key only, or negatively key only, or whether they want mixed results as in the first experiment (Sections 5.1 to 5.4 inclusive). In the command-line interface, this is achieved via the initial call to the tool's `Collection()` method. For example, to create the corpus

comparison object LOB *wrt* Brown, we use the following code: `data1 = Collection(brown_lc, lob_lc, '+')`, where the final parameter specifies keyness. Keyword outputs for both comparisons: Table 10 and Figure 12 immediately highlight differences in British and American spelling {*centre/center; programme/program*}:

Table 10: Collection-level keywords for LOB wrt Brown (and vice versa) once items which are negatively key have been filtered out

Collection KWs: Brown as reference set	**Collection KWs: LOB as reference set**
labour	mrs.
london	don't
colour	didn't
towards	program
round	toward
centre	it's
commonwealth	cannot
defence	states
british	center
programme	state

Figure 12: Collection-level keywords for LOB wrt Brown (and vice versa) with negatively-key items filtered out

5.5.1 Zooming in on round *in LOB*

The LOB top ten keywords include UK versus US spelling variants (*labour, colour, centre, defence, programme*), lexical variation (*towards* versus *toward*),

UK-specific placenames (*london, commonwealth, british*), and one unexpected term: *round*. The Brown keywords are still dominated by artefacts of differences in tokenization (*don't, didn't, it's, cannot, mrs.*) but also spelling variants (*program, center*), US-specific placenames (*states, state*) and lexical variants (*toward*).

Round has come up as a keyword in LOB *wrt* Brown (data1) in both versions of the *Semantic Pathways* tool. If we now query its usage in LOB via `data1.query('round')`, we get documents spanning the following genres: skills, trades and hobbies (E); popular lore (F); belles lettres (G); general fiction (K); mystery, detective fiction (L); adventure, western fiction (N); romance and love story (P), with each document represented by its most significant keyword. Sample output is given in Table 11:

Table 11: A sample *only* of LOB corpus documents (represented by their most significant keyword mapped to their filename) in which the term *round* is also significant

LOB (E)	**LOB (E)**	**LOB (F)**	**LOB (F)**	**LOB (G)**	**LOB (H)**	**LOB (J)**
crochet ce01	fish ce06	poultice cf33	song cf37	christmas cg19	irish cg25	farm cg62

Scanning the first two texts via the `text.more('filename')` method, we find that in the text about crocheting (ce01), *round* is most often used as a particle in phrasal verb constructions {*laced round*; *stitch round*; *thread round*}, whereas in the text about fishing (ce06), it is always used as an adjective: {*round fish*; *round fish like haddock*}. However, rather than (or in addition to) scanning entire texts, we would recommend extra built-in functionality whereby the user can call up concordance lines for a particular word in a particular text, but this is as yet not implemented in *Semantic Pathways*.

5.5.2 Zooming in on cannot *in Brown*

Cannot is an interesting Americanism in Brown *wrt* `LOB (data2)` at collection-level. Again, the term spans a number of genres: `data2.query('cannot')` returns the following document head-words {*china; Khrushchev; teachers; anti-slavery; born; farm; corporations; change; policy; artists*} and genre categories {press editorials - (B); religion (D); popular lore (F); belles lettres (G); miscellaneous (H); learned and scientific writing (J). In sampling two of these documents, we do not find this word being used as a prohibition, despite the political and social connotations suggested by the topmost significant terms or head-words in the text (Table 12):

Table 12: *Cannot*: a significant term in Brown *wrt* LOB without denoting prohibition

Brown (B)	**china** (cb23)	...cannot themselves resist... ...cannot take the initiative... ...cannot so much as try... ...cannot even introduce...
Brown (G)	**change** (cg25)	...cannot be derived... ...cannot be done... ...cannot distinguish... ...cannot save...
Brown (J)	**artists** (cj62)	...cannot entitle me... ...cannot be content... ...cannot be said... ...cannot expect...

6 Conclusions

We have presented *Semantic Pathways*, a Visual Analytics tool for supporting human reasoning by facilitating interactive query of large document sets. The tool is customised for all platforms (Mac, Linux, Windows) and users can opt to navigate via the *Semantic Pathways* Graphical User Interface (GUI), enabling speedy and intuitive exploration of a document collection, or a Command-Line Interface (CLI), where integration with Python and NLTK offers additional benefits.

The prototype versions trialled on the Brown and LOB corpora in this paper gist an entire collection of documents as a set of keywords extracted via corpus comparison, where significant items can be called in batches of ten keywords at a time, colour-coded in the GUI to denote provenance with respect to source text. From here the user can follow semantic pathways of lexical usage across the collection by clicking on an interesting word (in the GUI version) to identify individual documents in which it is key. Each document thus called is represented by its most significant term mapped to its filename; and from there the user can selects an item of interest to initiate corpus comparison of a single document with respect to the collection as a whole. Each new document retrieved is denoted by its ten most significant keywords, prompting iterative query, and offering access to source texts for more detailed inspection.

Trialling via prototype has led to insightful recommendations for system developers at Dundee and computational and corpus linguists at Leeds, where the latter have defined user requirements more closely through hands-on interaction with the system, and creating use case scenarios. An early finding is that differences in manual pre-processing and editing (*e.g.* tokenization schema) even in standardised corpora like Brown and LOB affect results, yielding keywords of apparent significance only. Further, we have made the following recommendations, which have now been implemented in the current version of *Semantic Pathways*. Users are now able to: restrict system display to positive keywords; and access raw frequency statistics on demand at collection and document levels. Through experimentation, we have demonstrated the tool's novelty, ease of use and intuitiveness in: (i) gisting an entire document collection; (ii) instant and aesthetic visualisation of term usage across genres; (iii) gisting a document via its ten most significant keywords; (iv) uncovering usage of selected items in terms of form, function and sense across a collection; (v) and recording exploratory user narratives as a history or chain or semantic pathway of key terms queried during the session.

Note

For availability of *Semantic Pathways* please contact c.rowland@dundee.ac.uk.

Acknowledgements

The authors would like to acknowledge the support of the Engineering and Physical Sciences Research Council, the Economic and Social Research Council, and the Centre for Protection of National Infrastructure, through the EPSRC/ESRC/CPNI research grant EP/H023135/1 **IDEAS Factory - Detecting Terrorist Activities: Making Sense**
http://gow.epsrc.ac.uk/NGBOViewGrant.aspx?GrantRef=EP/H023135/1;
and we would also like to acknowledge the contributions of our collaborators within the EPSRC funded Making Sense project led by Prof. Chris Hankin of the Institute for Security Studies, Imperial College, London.

References

(Web versions accessed on 13/02/2013.)

Abu Shawar, Bayan and Eric Atwell. 2003. Using dialogue corpora to train a chatbot. *Proceedings of Corpus Linguistics 2003*, Lancaster, UK. http://ucrel.lancs.ac.uk/publications/cl2003/papers/shawar.pdf.

Abu Shawar, Bayan and Eric Atwell. 2005. A chatbot system to animate a corpus. *ICAME Journal* 29:5–24. http://icame.uib.no/ij29/ij29-page5-24.pdf

Angus, Daniel, Janet Wiles and Andrew Smith. 2012. Generating visual insights into effective doctor-patient consultations. In *Proceedings of Advances in Visual Methods for Linguistics (AVML) 2012*, 22. York, UK. http://avml2012.files.wordpress.com/2012/08/avml2012_abstracts_final.pdf.

Atwell, Eric. 1981. *LOB Corpus tagging project: Manual pre-edit handbook*. Departments of Computer Studies and Linguistics, Lancaster University.

Atwell, Eric. 1982. *LOB Corpus tagging project: Manual post-edit handbook*. Departments of Computer Studies and Linguistics, Lancaster University.

Baker, Paul, Costas Gabrielatos, Majid Khosravinik, Michal Krzyzanowski, Anthony McEnery and Ruth Wodak. 2008. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse and Society* 19 (3): 273–305. http://www.lancs.ac.uk/fass/doc_library/linguistics/wodakr/bakeretal2008.pdf.

Baroni, Marco, Adam Kilgarriff, Jan Pomikalek and Pavel Rychly. 2006. WebBootCaT: A web tool for instant corpora. In *Proceedings of Euralex 2006*, 123–131. Turin, Italy. http://www.euralex.org/elx_proceedings/Euralex2006/.

Beigman Klebanov, Beata, Daniel Diermeier and Eyal Beigman. 2008. Automatic annotation of semantic fields for political science research. *Journal of Language Technology and Politics* 5 (1): 95–120. http://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID1208145_code292153.pdf?abstractid=1026961&mirid=3.

Culpeper, Jonathan, Dawn Archer and Paul Rayson. 2009. Love – 'A familiar or a devil'? An exploration of key domains in Shakespeare's comedies and tragedies. In D. Archer (ed.). *What's in a word-list? Investigating word frequency and keyword extraction*, 136–157. Farnham: Ashgate.

Doherty, Neil, Nigel Lockett, Paul Rayson and S. Riley. 2006. Electronic-CRM: A simple sales tool or facilitator of relationship marketing? In *Proceedings of 29th Institute for Small Business and Entrepreneurship Conference*, Cardiff-Caerdydd, UK.
http://eprints.lancs.ac.uk/12852/.

Dukes, Kais and Eric Atwell. 2012. Visual methods for understanding the language of the Quran. In *Proceedings of Advances in Visual Methods for Linguistics (AVML) 2012*, 33–34. York, UK.
http://avml2012.files.wordpress.com/2012/08/avml2012_abstracts_final.pdf.

Endert, Alex, Patrick Fiaux and Chris North. 2012. Semantic interaction for visual text analytics. In *Proceedings of CHI'12 ACM conference on Human Factors in Computing Systems*, 473–482.
http://infovis.cs.vt.edu/sites/default/files/PDF_2.pdf.

Fabricius, Anne, Charlotte Vaughn and Tyler Kendall. 2012. Plotting speakers' vowel systems in real-time interaction: A first approach. In *Proceedings of Advances in Visual Methods for Linguistics (AVML) 2012*, 14. York, UK.
http://avml2012.files.wordpress.com/2012/08/avml2012_abstracts_final.pdf .

Francis, W. Nelson and Henry Kučera. 1979. *Brown corpus manual – revised and amplified.* ICAME. http://icame.uib.no/brown/bcm.html.

Huckvale, Mark. 2012. Speech as visible patterns of sound. In *Proceedings of Advances in Visual Methods for Linguistics (AVML) 2012*, 17. York, UK.
http://avml2012.files.wordpress.com/2012/08/avml2012_abstracts_final.pdf.

Johansson, Stig, Geoffrey Leech and Helen Goodluck. 1978. *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers*. ICAME. http://khnt.hit.uib.no/icame/manuals/lob/index.htm.

Johansson, Stig, Eric Atwell, Roger Garside and Geoffrey Leech. 1986. *The tagged LOB corpus users' manual.* ICAME. http://khnt.hit.uib.no/icame/manuals/lobman/index.htm.

Kilgarriff, Adam and Gregory Grefenstette. 2003. Introduction to the Special Issue on Web as Corpus. *Computational Linguistics.* 29 (3): 333–347.
http://acl.ldc.upenn.edu/J/J03/J03-3001.pdf.

Koteyko, Nelya and Ronald Carter. 2008. Discourse of 'transformational leadership' in infection control. *Health: An Interdisciplinary Journal for the Social Study of Health, Illness and Medicine* 12 (4): 479–499. http://www.tara.tcd.ie/bitstream/2262/52185/1/PEER_stage2_10.1177%252F1363459308094421.pdf.

Maguire, Martin and Nigel Bevan. 2002. User requirements analysis: A review of supporting methods. In *Proceedings of IFIP 17th World Computer Congress,* 133–148. Montreal, Canada. http://nigelbevan.com/papers/WCC_UserRequirements.pdf.

McEnery, Tony and Andrew Hardie. 2011. *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Montgomery, Chris and Philipp Stoeckle. 2012. Visualising perceptual dialectology data using Geographical Information Systems. In *Proceedings of Advances in Visual Methods for Linguistics (AVML) 2012*, 19. York, UK. http://avml2012.files.wordpress.com/2012/08/avml2012_abstracts_final.pdf.

Oakey, David. 2002. Formulaic language in English academic writing: A corpus-based study of the formal and functional variation of a lexical phrase in different academic disciplines. In R. Reppen, S. Fitzmaurice and D. Biber (eds.). *Using corpora to explore linguistic variation,* 111–129. Amsterdam: John Benjamins.

Olphert, Wendy and Leela Damodaran. 2002. Getting what you want, or wanting what you get? – beyond user centred design. In D. McDonagh, P. Hekkert, J. van Erp, D. Gyi (eds.). *Design and emotion: Proceedings of 3rd International Conference on Design and Emotion*, 126–130. Loughborough, UK.

Priestley, Janne Bondi Johannessen, Kristin Hagen, Anders Nøklestad and André Lynum. 2012. Maps as a central linguistic research tool. In *Proceedings of Advances in Visual Methods for Linguistics (AVML) 2012*, 20. York, UK. http://avml2012.files.wordpress.com/2012/08/avml2012_abstracts_final.pdf.

Rayson, Paul and John Mariani. 2009. Visualising Corpus Linguistics. In *Proceedings of Corpus Linguistics 2009*, Liverpool, UK. http://ucrel.lancs.ac.uk/publications/cl2009/426_FullPaper.docx.

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

Rowland, Christopher, John Anderson, Eric Atwell and Claire Brierley 2012. Real time aesthetic visualisation of NLP-driven semantic pathways. In *Proceedings of Advances in Visual Methods for Linguistics (AVML) 2012*, 12. York, UK.
http://avml2012.files.wordpress.com/2012/08/avml2012_abstracts_final.pdf.

Scott, Mike. 2012. *WordSmith Tools* version 6. Liverpool: Lexical Analysis Software. http://www.lexically.net/wordsmith/.

Shneiderman, Ben. 1996. The eyes have it: a task by data type taxonomy for information visualisations. *VL'96 Proceedings of IEEE Symposium on Visual Languages*, 336–343.
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=545307.

Viegas, Fernanda, Martin Wattenberg and Jonathan Feinberg. 2009. Participatory visualisation with Wordle. *IEEE Transactions on Visualisation and Computer Graphics* 15 (6): 1137–1144.