Visualising Textual Concordances with Word Mosaics and Graphs: Prototypes and Evaluation Methods

Shane Sheehan

Computer Science Technical Report
School of Computer Science and Statistics
Trinity College Dublin

August 25, 2014

Abstract

The visualisation of textual concordances has seen little change since the proposal of the keyword-in-context indexing technique in the 1950s. This report details an investigation of the current state of the art in concordance visualisation. The investigation lead to the the expansion of context trees to a new graph based representation of concordance lists. This representation is explained in detail and two visualisations, of the graph, are explored. It is proposed that one of these visualisations may offer significant advantages over existing concordance visualisations for the investigation of keyword collocation frequencies. An evaluation plan to test this hypothesis is outlined as future work in this project.

Contents

Chapter 1

Introduction

In the last decades, empirical and corpus-based text analysis methods have come to the fore. This is, in part, due to the vast amounts of text available in digital format, and also changing theoretical perspectives. Computational and statistical methods for the analysis of textual resources are referred to collectively as text analytics by language researchers and, increasingly, end users of this technology. Text analytics are often combined with interactive visualisation methods which can enhance and complement them. Examples of such text visualisation methods range from the popular "word clouds" and other forms of "vernacular" visualisation [16], to full-fledged interactive systems [18].

Concordancing, or the arranging of passages of a textual corpus in alphabetical order according to user-defined keywords, is one of the oldest and still most widely used forms of text analysis. The advent of computers allowed the systematic creation of concordances through the "keyword in context" indexing technique, first proposed by Hans Peter Luhn in the 50s. The traditional visual representation of these concordances lays out text samples so that a keyword of interest is aligned centrally, an example of which can be seen in Figure 2.1. This form of visualisation, enhanced in interactive systems by features such as search, context sorting and statistical analysis, is still widely used by corpus linguists, lexicographers, translators and others [8].

Word Tree [18] is an alternative form of concordance visualisation which has recently been proposed. It displays the keyword and the words appearing to either the left or right of the keyword as a tree structure Figure 2.2. The nodes in the Word Tree are rendered as labels which are scaled according to frequency in a manner which does not break the sentence structure in the chosen half of the context. Since only one side of context can be displayed, due to the underlying structure being a suffix tree, users are prevented from reading the full concordance lines in which the keyword appears. To view both contexts on the keyword, requires switching between trees.This is because a new suffix tree must be generated each time a new context is displayed. This causes a loss in visual continuity for the user.

For certain corpus linguistics tasks, such as detection of phrases that span left and right contexts(e.g. as in the expression "run the whole *gamut* of ..."), frequency information for words occurring on each context is usually more useful to the analyst than the linear structure of a single context [13]. In the traditional keyword-in-context visualisation, word frequencies are visualised by sorting the concordance lines alphabetically at a specified word position. This doesn't allow for a good overview of the word frequencies at each position, since sorting more than one position at a time is not possible without resulting in the deconstruction of the concordance lines linear structure. Inspecting frequency information in Word Tree visualisations becomes less useful at nodes further from the keyword, since they are given less weight due to the branching required to remain consistent with the tree structure(i.e. there is at most one path between any two nodes).

In this report these issues are addressed by the proposing a unified graph based data abstraction of keyword-in-context concordance lists, where the keyword is defined in terms of eccentricity and concordance lines can be reconstructed from the graph. Three interactive visualisations are presented, a bi-directional hierarchical display, an interactive mosaic (space-filling) display and a composite mosaic and keyword-in-context display. The path to implementing these visualisations is described in terms of the data state model (or information visualisation reference model) [1] and details are provided of the familiar information visualisation techniques which have been integrated into these visual interfaces.

Chapter 2

Background

2.1 Corpus and Concordance

A corpus can be defined as "A collection of naturally occurring language text, chosen to characterize a state or variety of a language" or, more simply, "a large collection of texts" where texts are defined as "samples of language in use" [13, 12]. In the study of corpus lexicology the ability to select small samples of the corpus is desired since corpora typically contain more information than an individual can efficiently analyse. A concordance is one such sample which contains a chosen word or phrase (keyword) and the contexts in which it appears in the corpus. Simply put it is an index pointing to each occurrence of the keyword in the corpus

2.1.1 Keyword-in-context

Concordances are typically created using the keyword-in-context (KWIC) indexing technique. The traditional visualisation of the KWIC involves a layout which aligns occurrences of the keyword vertically and displays a chosen number of words from both the left and right contexts. The left and right contexts are, simply, the words which occur before and after the keyword in the corpus, respectively. Position and color are most often the visual variables [17, p. 137] used to differentiate between the keyword and both contexts. The simplified example of the KWIC for the keyword "eye" is shown in Figure 2.1. In this example we arbitrarily chose to show five words form the left context and four from the right.

```
       is invisible to the naked eye. From egg to  egg
   simply invisible to the naked eye. It crawled  without a
         sort out with the naked eye the blur of bodies
      diagnose it with the naked eye, and there are two
           at him with her naked eye, almost with curiosity.
are therefore keeping a watchful eye on both the reach
         we will keep a watchful eye open to ensure that
```

Figure 2.1: KWIC display for the word "eye"

2.1.2 Word Tree

A recent addition to the visualisation of concordances called *Word Tree* displays the keyword and, either the left or right contexts [18]. This visualisation, as the name suggests, takes the familiar form of at tree structure, in which the keyword is displayed as the root vertex and additional word vertices are connected in text order to each other. The vertices are scaled on the square root of the frequency of the words they represent i.e. the number of leaf vertices they connect to. This interactive visualisation enables exploration of the corpus since a large number of concordance lines can be displayed on the screen at once. Figure 2.2 an example of a Word Tree with keyword "eye" and the right context. This overview can then be investigated by zooming and panning to see detail .The keyword can be redefined by clicking on vertices to focus only on branches exiting that vertex.

An important property of word trees is that there is at most one path between any two vertices. The combining of multiple word occurrences at a concordance position is permitted provided this property is not violated in doing so.

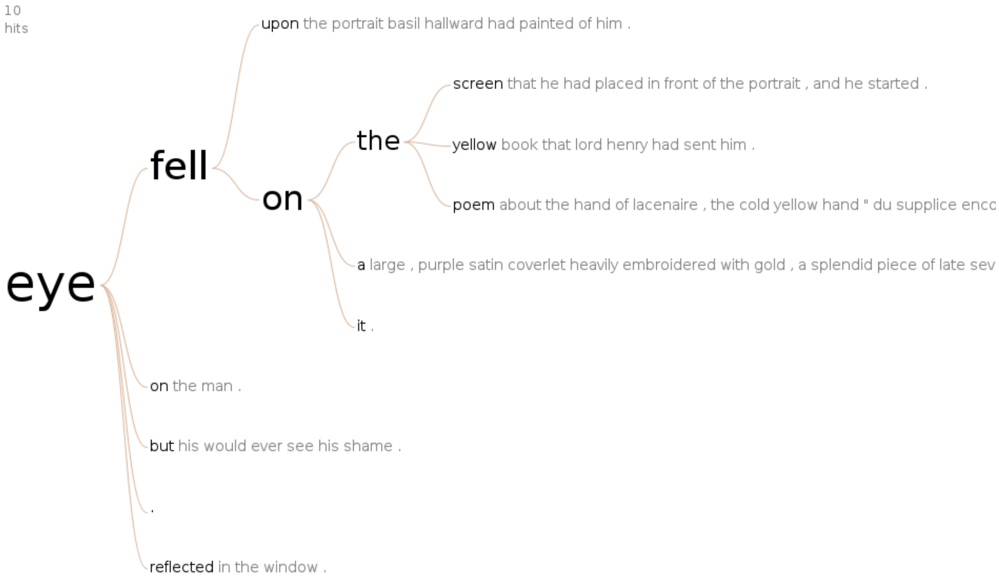

10
hits
eye
fell
upon the portrait basil hallward had painted of him .
on
the
screen that he had placed in front of the portrait , and he started .
yellow book that lord henry had sent him .
poem about the hand of lacenaire , the cold yellow hand " du supplice encc
a large , purple satin coverlet heavily embroidered with gold , a splendid piece of late sev
it .
on the man .
but his would ever see his shame .
.
reflected in the window .

Figure 2.2: A Word Tree visualisation generated by the free service ManyEyes

However, a major drawback of this method is that only one of either the left or right contexts can be displayed at a time. This is due to the underlining representation of the data being based on a suffix tree. The display of only one side of the context means it is not possible to view full sentences, unless the keyword happens to be at the beginning or end of a sentence.

2.1.3 TEC Corpus Browser

To develop additional visualisations of concordances it is useful to leverage pre-existing corpus browsing technologies. The MODNLP-tec corpus management suite contains a corpus/concordance browser which, importantly, allows for plugins. The browser can connect to local or remote corpus, and enables keyword search, sorting, sub-corpus selection and other features, such as access to metadata. Searching for a keyword returns a concordance which is displayed as a traditional KWIC visualisation. The suite also contains an indexer and server which we use to index and serve our test data. The visualisations presented in this study are all created as plugins for the TEC corpus browser. An example of the KWIC view available in the TEC browser can be seen a part of the composite visualisation in both Figures 4.4 and 4.5.

This study focuses on visual techniques which can be applied to corpora, specifically visualising concordances. Thankfully the MODNLP-tec corpus management suite provides us an extensible platform with which to perform our visualisations, and removes from us the errand of creating a web based corpus browser, corpus indexer and server. [8, 7]

2.2 Information Visualization Reference Model

The information visualization reference model (or data state model)[1] is a conceptual framework that enables the concise description of the visualisation construction process. The framework allows the classification of data states in to four stages and enables the description of transfer between these states by the use of data operators. The four stages a data state can be classified under are *data*, *analytical abstraction*, *visual abstraction* and *view*. This model is typically represented as a diagram where nodes represent data states and edges represent operators. Operators are also subject to the same stage classifications, however between stage operators also exist. As an example the model for the visualisations created for this study is shown in Figure 3.1.

At the first stage (Data), the states contain data representations or raw data which is operated on in ways such as combining, adding, deleting or filtering the data. A between stage operator structures this data in some way (usually by extraction), moving the data to a state which can be considered an “analytical abstraction” e.g. a graph, tree or metadata etc. This process is referred to as “data transfer”. Operations within the “analytical abstraction” stage often take the form of selection operations where a subset of the data is selected by some analytical process. Visualisation transfer now takes place transforming the data state to one which can be directly mapped to a visual representation. Operations at this visualisation abstraction stage deal directly with how the data will be visualised e.g. combining nodes in a cluster representation. The visual mapping transfer is done by operators which create views from visual abstractions. Each state at the view stage is a visualisation of the data. View stage operators modify an existing view to produce an altered one e.g. zooming, highlighting panning etc.

An extensive review , under this model, of information visualisation applications exists [2]. This provides a taxonomy of visualisation techniques, and serves to illustrates the descriptive power of the data state model. Additionally the data state model has been shown to be functionally equivalent to the data-flow model which is accepted as ”an industry standard way of constructing visualization for large scientific data sets”. [3]

2.2.1 Prefuse

Prefuse is an extensible interface toolkit for interactive visualisations written in the Java programming language. The fundamental data structure used in prefuse is a graph, since our new concordance representation is a graph prefuse is a natural choice of visualisation technology. In addition to this the main reason we use prefuse is that its design is based on the information visualisation reference model [5]. Several built-in data structures are available for data states at the analytical abstraction level, to which filters can be applied to create the *VisualItems* of the visual abstraction level. A large number of operators are available at this stage, which perform *actions* such as layout, coloring, sizing and interpolation. These VisualItems can then be rendered to a display and various “Controls” can be applied to perform interactions such as distortion, zooming and panning [14].

Using prefuse, we can us design and document our visualisation process using the data state model and create a software implementation which also relates directly to the operators and states outlined in the model.

Chapter 3

Implementation

3.1 Model

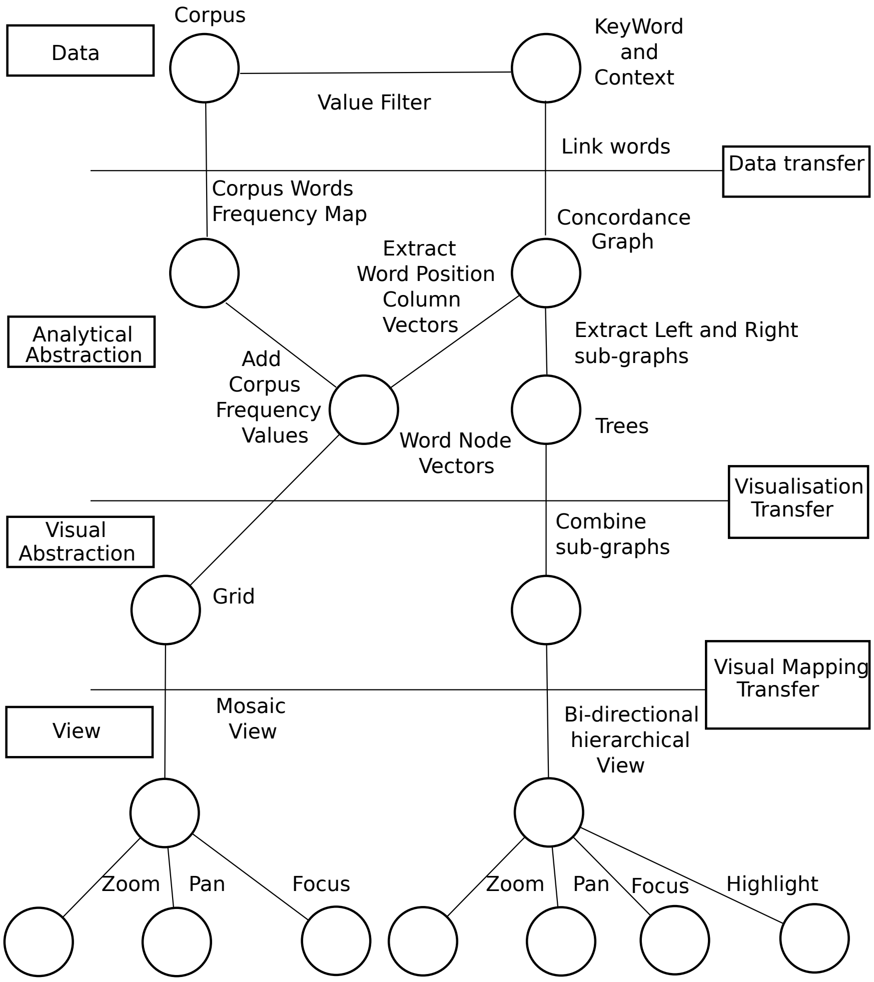


Figure 3.1: Concordance visualisation data state model

The concordance visualisation process is described by the data state model diagram Figure 3.1. The data stage contains two states the first is the indexed corpus which can be accessed through the TEC browser and the second is the KWIC index which is created when a user defined keyword is supplied to the browser. Information about the frequencies of words in this KWIC are required in the creation of one of our visualisations. Through the Corpus browser we can extract this frequency information to create the first state of the analytical abstraction stage. The second state available at this stage is the concordance graph, a new graph based representation of textual concordances. The Concordance graph was developed with the express purpose of creating a new representation of concordance lists, from which new visualisation options would be available. Before continuing the description of the visualisations, a formal description of the concordance graph is presented, since it is a original and key element of the concordance visualisation process.

3.2 Concordance Graph

A KWIC indexing system works by circularly shifting each element of a set of ordered lines of text. For concordancing, the presentation is such that the keyword is placed at the centre of the line. We shall designate the keyword by k and its left and right contexts by $L = (l_1, \ldots, l_n)$ and $R = (r_1, \ldots, r_n)$, respectively, where l_i (respectively, r_i) denotes a word i positions to the left (right) of k. The index can then be represented by a set $\mathcal{C}$ of triples of the form $C = (L, k, R)$.

This structure encodes a high degree of redundancy in that it disregards the fact that for a given position, many different word occurrences (*tokens*) across the concordance lines are simply instances of the same word (*types*). This is illustrated by the words "naked" and "the", respectively l_1 and l_2 in the concordance shown in Figure 2.1. Since according to Zipf's Law a small number of types tends to dominate the distribution of tokens at a particular position, the bulk of the data in $\mathcal{C}$ will consist of such repetitions.

One can devise a more economical representation by exploiting the linear structure of C. The approach we propose does this by representing the concordance set as a graph, where vertices correspond to word types and the linear order is encoded by the edges, as follows.

Definition A *concordance graph* is a quadruple $\mathcal{G} = (V, E, V_l, E_l)$ where V is a set of vertices, $E \subseteq V \times V$ is a set of edges (v_s, v_t) connecting vertices, $V_l : V \to \mathit{Types}$ is a labelling of vertices with words and $E_l : E \to \mathbb{R}$ is a labelling of edges with word frequency information.

The word frequency labels in E_l are assumed to indicate the number of concordance lines between the two ends of the edges. A concordance graph can be built through an algorithm that takes a KWIC index $\mathcal{C}$ (encoded, say, as a tabular structure) as input and

1. cycles though each lexicographically sorted column, starting from the centre (corresponding to k, with index $i = 0$) and expanding over L and R,

2. creates a vertex $v_{i,j}$ for each type (in row j of column i), labelling the vertex with the appropriate string,
3. recursively connects each vertex to the next column's vertices $v_{i+1,m}$, labelling edges according to the number of strings $v_{i,j}$␣$v_{i+1,m}$ in the concordance,
4. and, finally, creates edges linking each vertex v_n^l for each row in the leftmost column of C to the corresponding vertices v_n^r for the rightmost column. We will refer to such edges as *contextual edges*.

To aid understanding of the concordance graph it is possible to conceptualise it as the combination of the left and right context trees (Figure 3.2), with the contextual edges added to enable the reconstruction of concordance lines from the graph (Figure 3.3). Since each of the “word trees” (or context trees) observe the principle of having at most one path between tree vertices, it is clear we can reconstruct partial concordance lines up to and after the keyword separately in the case of concatenated context trees Figure 3.2. However Figure 3.3 shows an example of concordance graph for a subset of the fragment seen in Figure 2.1 with word count labelling, from which the entire concordance line can be reconstructed thanks to the contextual edges. Note that the edges that connect the left to the rightmost vertices (contextual edges) guarantee that the entire set of concordance lines going through any vertex $v_{i>0}$ is retrievable by traversing the concordance graph starting from v_i, which is not possible in a concatenation of Word Trees.

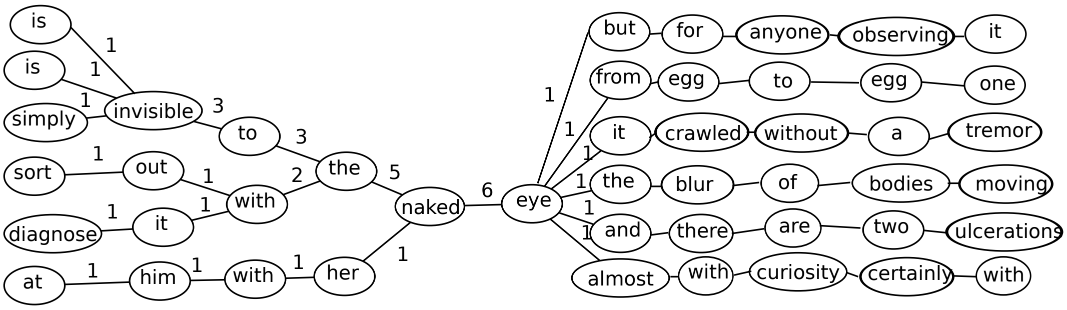


Figure 3.2: Concatenated context trees

Figure 3.4 shows an example of concordance line reconstruction for the node labelled “invisible” (highlighted in yellow). From the diagram it is clear that traversing the branches from the selected node to the keyword “eye” reconstructs all concordance lines passing through the selected node “invisible” .

To determine the keyword from the graph we define graph distance $d(v, u)$ as the minimum length of the paths connecting vertex v to u in concordance graph $\mathcal{G}$ and an operation $P(\mathcal{G})$ which removes all contextual edges (v_n^l, v_n^r) from $\mathcal{G}$, then we can retrieve the keyword vertex through its eccentric property.

Definition The *eccentricity* $\epsilon(v)$ of a vertex v in a concordance graph is defined as $\epsilon(v) = \max_{u \in V \setminus \{v\}} d(v,u)$. The minimum graph eccentricity ($\min_{v \in V} \epsilon(v)$) is the *graph radius.*

Given a concordance graph $\mathcal{G}$, the keyword k is the label $V_l(v_k)$ of the vertex v_k whose eccentricity $\epsilon(v_k)$ is equal to the graph radius of $P(\mathcal{G})$. The above described graph construction algorithm guarantees that this vertex is unique and corresponds to k in the KWIC representation.

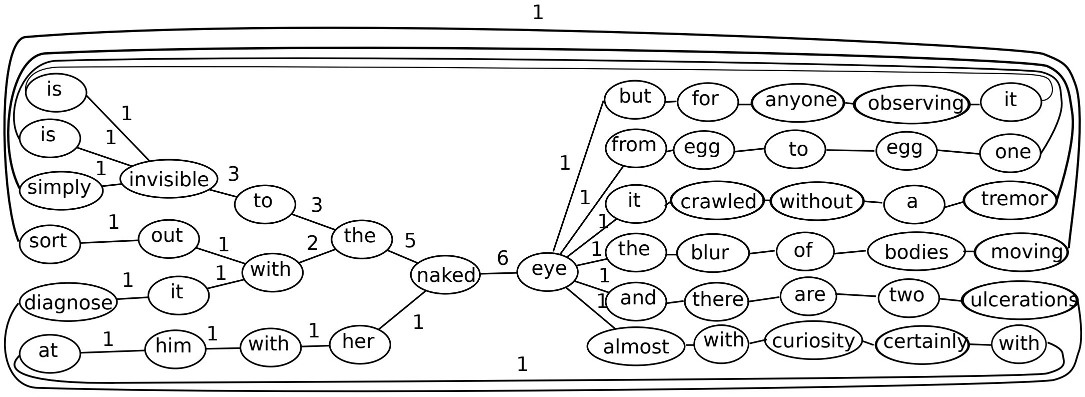


Figure 3.3: Sample concordance graph for the word "eye".

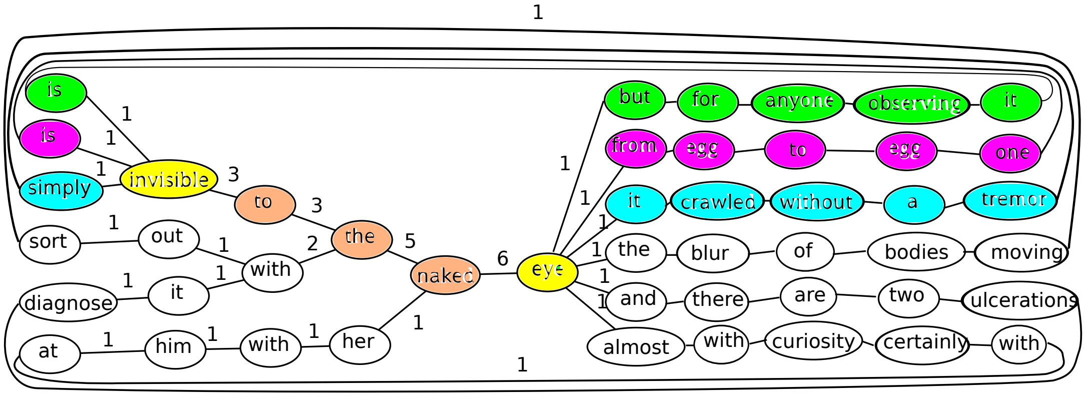


Figure 3.4: Concordance line reconstruction example

Chapter 4

Visualisations

4.1 Bi-directional Hierarchical Display

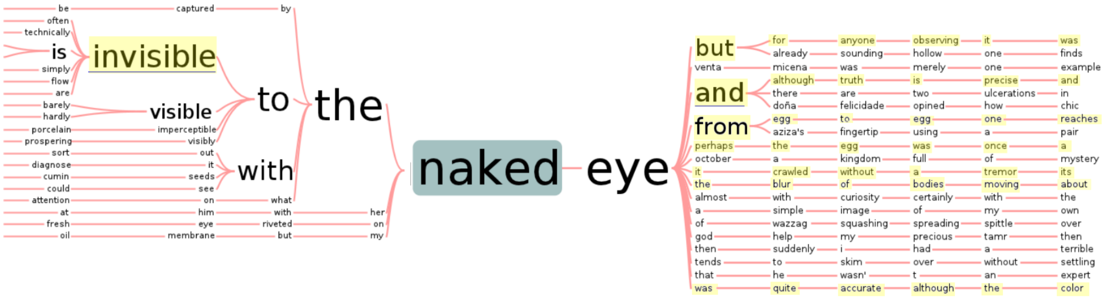

be captured by
often
technically
is
invisible
simply
flow
are
barely
hardly
visible
porcelain imperceptible
prospering visibly
sort out
diagnose it
cumin seeds
could see
with
to
the
attention on what
at him with her
fresh eye riveted on
oil membrane but my
naked
eye
but
for anyone observing it was
already sounding hollow one finds
venta micena was merely one example
and
although truth is precise and
there are two ulcerations in
doña felicidade opined how chic
from
egg to egg one reaches
aziza's fingertip using a pair
perhaps the egg was once a
october a kingdom full of mystery
it crawled without a tremor its
the blur of bodies moving about
almost with curiosity certainly with the
a simple image of my own
of wazzag squashing spreading spittle over
god help my precious tamr then
then suddenly i had a terrible
tends to skim over without settling
that he wasn' t an expert
was quite accurate although the color

Figure 4.1: Bi-directional hierarchical view of a concordance for the compound “naked+eye”

Our initial exploration in visualising the concordance graph draws inspiration from the Word Tree visualisation [18]. While word tree shows only the keyword and one side of the context, we wish to display both sides simultaneously such that no loss of words occur in comparison with the KWIC display.

The process for creating the bi-directional hierarchical display (BDH) is shown in Figure 3.1. To begin, form the concordance graph we extract trees corresponding to the left and right contexts, this is done to perform graph layout, in data state terms this is our visualisation transfer. These layouts are then recombined to form the visual abstraction from which we can build the display. At the visual abstraction stage other information is encoded that will be used in the visual mapping transformation, the label font of each vertex is scaled proportional to the maximum frequency label of the edges that are incident on it (i.e. the edge which connects to a vertex closer to the keyword). Figure 4.1 illustrates the BDH for a subset of the concordance for the keyword “eye”, with a focus on the expression “naked eye”.

In visualisation design of the eight known visual variables position is regarded as the most important [17, p. 137]. We use positional encoding to show vertices concordance positions, relative to the keyword, by placing them on the same vertical alignment, thereby retaining some resemblance to the KWIC display. This is in contrast to Word Tree where word position information is lost (e.g. compare positioning in Figure 2.2 and

Figure 4.1). In the BDH the position of a word relative to the keyword is much clearer than the Word Tree where, for most words, it cannot be determined at a glance.

A selection interaction enables highlighting of a vertex, and results in the highlighting of the vertices in the opposite context which form sentence fragments with the selected Vertex. Only vertices in the opposite context are highlighted as it is easy to see from the hierarchy if vertices are connected in the same context. The difficulty occurs in following sentences through the keyword as all *left to right* directed paths in the graph pass through this vertex. In this case all words in the opposite context would be highlighted as candidate concordance line matches. This situation is avoided by using the contextual edges of the concordance graph. Using the contextual edges we remove the need to highlight vertices in the opposite context which do not pass through a contextual edge to the selected vertex. Figure 4.1 shows the highlighted vertex with the label "invisible", and the vertices in the right context which form concordance lines with eye are also highlighted.

4.2 Mosaic Display

The concordance mosaic visualisation which we developed is visually similar to other space-filling visualisations such as the temporal mosaic [9]. Treemaps [11] are a rectangular space-filling visualisation for hierarchies where position encodes vertex relations. However the concordance mosaic differs from these visualisations by using vertical position and scale to encode word statistics and horizontal position to encode the position of each vertex (word) relative to a single reference vertex (the keyword).

Returning to our description of the visualisation from the view point of the data state model Figure 3.1, we show that at the analytical abstraction stage a new data state is created from the corpus frequency information and the concordance graph.

This data set takes the form of an ordered collection of vectors of word objects. These word objects are vertices form the concordance graph combined with the word frequency at its concordance position and also the frequency of the word in the entire corpus. The vectors are extracted by traversing the context trees of the concordance graph in a breadth first manner and counting token occurrences at each level of tree depth. The vectors are ordered so that if vector x contains the keyword then vector $x + 1$ contains all words which occur one position to the right of the keyword (in the corpus) and vector $x - 1$ all words one position to the left. Visualisation transfer maps these vectors to a grid onto which word position and frequency information can be encoded.

The two visual variables we use encode the information are position and scale, while color is used only to differentiate between the boundaries of word objects. The vertex of each word object is represented as a rectangle of equal width within the grid. The horizontal position of each rectangle encodes its concordance position relative to the keyword in a similar way to the KWIV and BDH positional encoding. Vertical position of the rectangles is determined by scale i.e. ordering the rectangles in each column by decreasing height.

Two different scaling schemes were explored.

4.2.1 Concordance Position Word Frequency

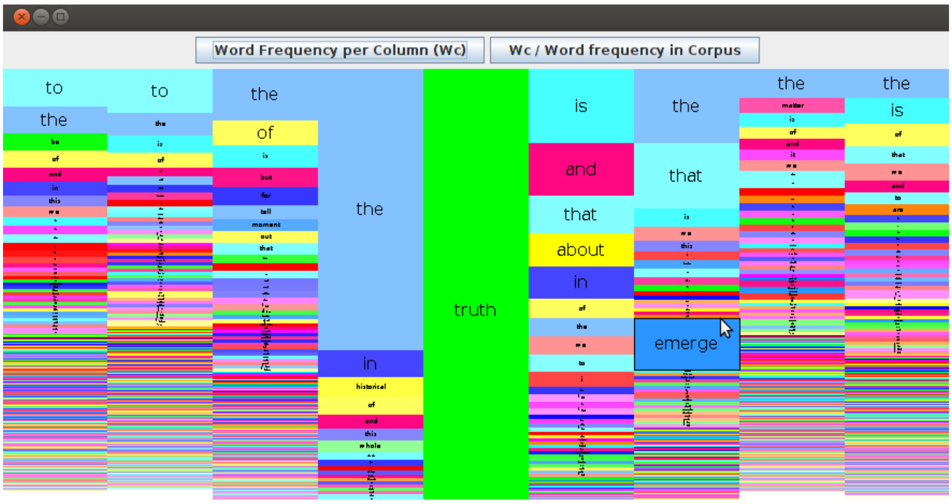


Figure 4.2: Mosaic scaled according to concordance position word frequency

One scaling scheme, Figure 4.2, scales the height of rectangles according the word frequency per concordance position (column). These frequencies are mapped then to heights to normalise them to the screen height. We call this scaling scheme *Concordance Position Word Frequency*. This representation provides at a glance review of frequent collocations of the keyword and surrounding contexts. This feature is not available in the previously discussed hierarchical displays, where as distance from the keyword increases frequency information becomes less evident. In the case of KWIC positional frequency information is also difficult to obtain, since sorting of a single position at a time provides a restricted view of frequency and for most concordances scrolling is necessary for exploration of concordance.

Figure 4.2 shows the concordance mosaic for the keyword “truth”. We can see that at most positions the word “the” is one of the two most frequent words. This is not surprising as “the” is the most common word in the corpus we used for testing. We should expect that under this scaling scheme each position will be dominated by words with high corpus frequency, and that words which are less frequent in the corpus but which co-occur very frequently with the keyword will also be prominent.

4.2.2 Concordance to Corpus Relative Frequency (Collocation Strength)

To give more emphasis to words with many keyword co-occurrences relative to their corpus occurrence frequency we selected a scaling scheme where rectangle height is directly proportional to a words concordance frequency at a position, while inversely proportional to its word frequency in the entire corpus.

Figure 4.3 shows the same concordance as Figure 4.2 with the height of the rectangles scaled on their column frequency relative to the word frequency in the corpus. This rela-

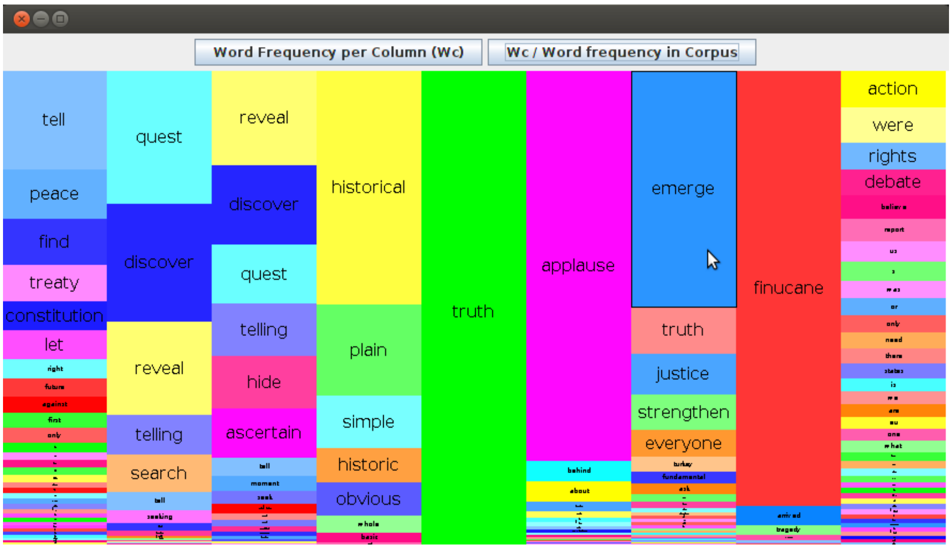


Figure 4.3: Mosaic scaled on word frequency relative to corpus word frequency

tive frequency representation has the effect of reducing the size of words which have high frequency in the corpus and increasing words of lower frequency (e.g. comparing figures 4.2 and 4.3 the word "the" has been reduced and the word "review" has increased in the corpus relative frequency view). In the relative frequency representation we have also reduced the size of rectangles which have relative frequencies below a certain threshold, this gives more screen space to the words which more often occur with the keyword.

This representation can be thought of as a visualisation of relative collocation strength with thekeyword i.e. how often a word occurs with the keyword compared with its total corpus occurrences relative to the other words at that position. An example from [13] explains that the phrase "kith and kin" are the only occurrences of the word "kith" in the chosen corpus. When looking at a concordance of the word "kin" it may not be obvious how strong the collocations pattern is if there are many more occurrences of "kin" in the corpus, but when using the relative frequency view the collocations strength should be apparent.

4.2.3 Interactions

To aid investigation of the concordances we implemented a bifocal distortion [15] which is activated on mouse over of rectangles which are below a chosen height threshold. Figure 4.2 shows an example of this distortion. The rectangle with label "emerge" has been expanded so that the label can be read. Before distortion this rectangles size was approximately equal to the height of the rectangles directly above and below it.

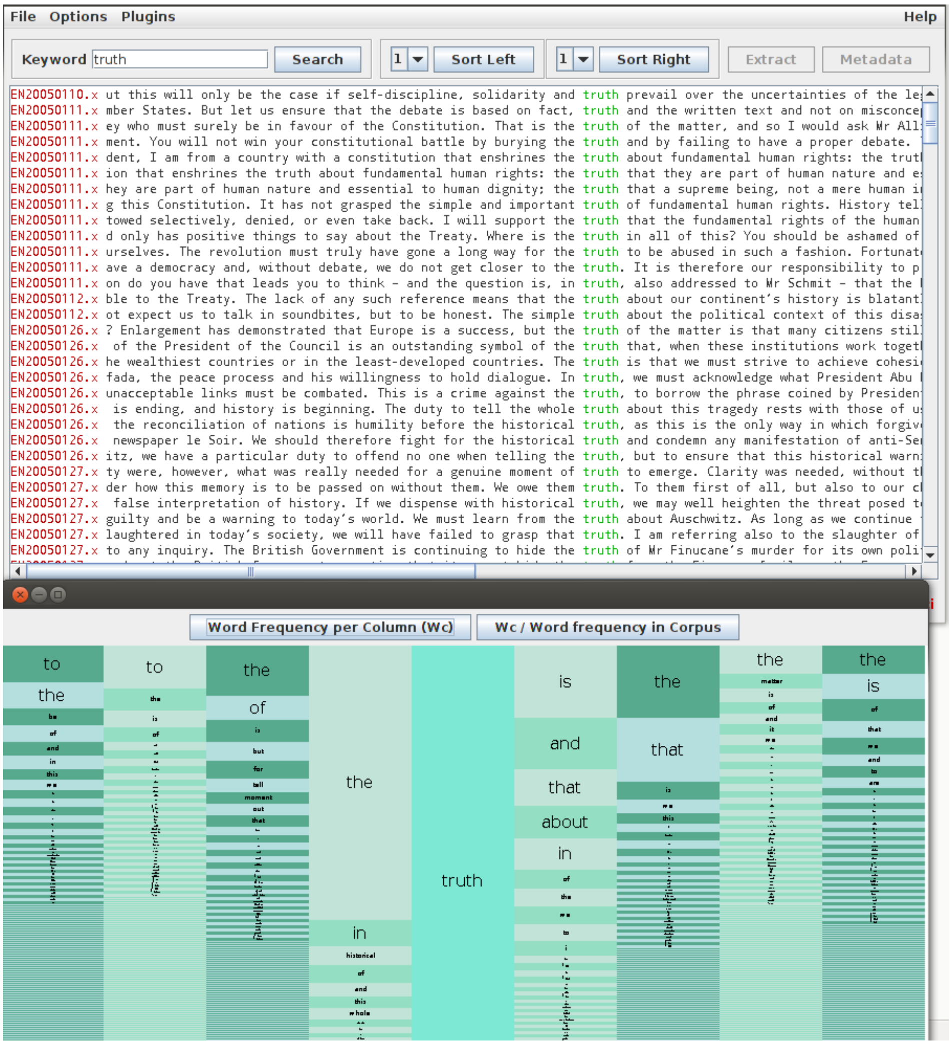
Word selection is also enabled and has the effect of highlighting the selected rectangle and all rectangles which form concordance lines with the selected word. This highlighting is done by adding a border to the selected word and changing the background color to white as seen in the mosaic in Figure 4.5 where the word “about” at position $x + 1$ has been selected.

These interactions help in the investigation of collocations patterns, however, even when highlighting the words which form concordance lines with the selected word, it is ambiguous which words link together to form the exact lines found in the KWIC view.

4.3 Juxtaposed Views

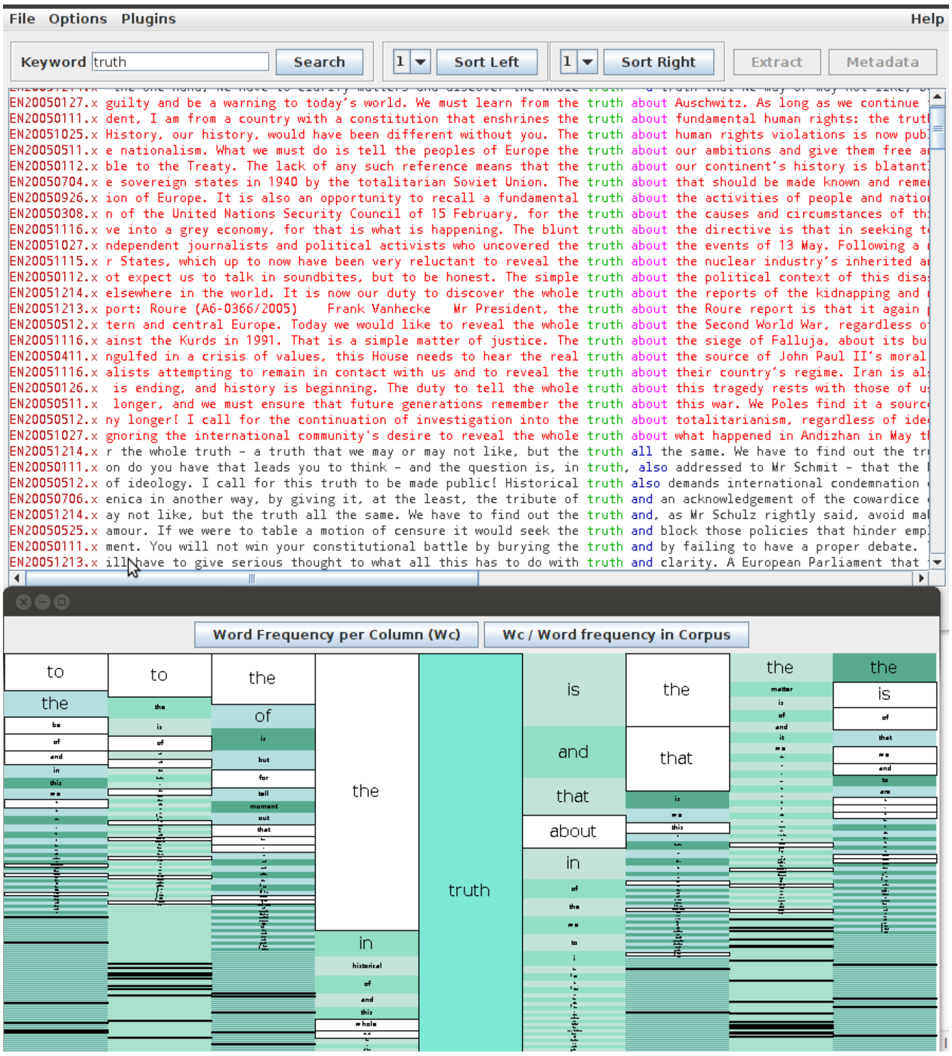
We used juxtaposed views [6], a design pattern of composite visualization views, to make use of the advantages of both the traditional and mosaic concordance views. This design gives the user both the overview and detail removing the concordance line ambiguity of the individual mosaic view.

Since the data is implicitly linked, interaction with either view can affect the other. We demonstrate this by a applying a filter interaction to the juxtaposed view shown Figure 4.4. Selecting a word on the mosaic view filters the traditional KWIC view, such that only concordance lines which contain the highlighted word in the selected column are highlighted. The KWIC view is sorted by the selected column and auto scrolls to the first occurrence of the selected word in that column. Figure 4.5 shows the result of the the filter interaction where the word “about” has been selected in column $x+1$. The KWIC view is scrolled to the first occurrence of the word “about” in the column, making both the concordance lines and frequency information available to the user simultaneously.



```
EN20050110.x ut this will only be the case if self-discipline, solidarity and truth prevail over the uncertainties of the le
EN20050111.x mber States. But let us ensure that the debate is based on fact, truth and the written text and not on misconce
EN20050111.x ey who must surely be in favour of the Constitution. That is the truth of the matter, and so I would ask Mr All
EN20050111.x ment. You will not win your constitutional battle by burying the truth and by failing to have a proper debate.
EN20050111.x dent, I am from a country with a constitution that enshrines the truth about fundamental human rights: the trut
EN20050111.x ion that enshrines the truth about fundamental human rights: the truth that they are part of human nature and e
EN20050111.x hey are part of human nature and essential to human dignity; the truth that a supreme being, not a mere human i
EN20050111.x g this Constitution. It has not grasped the simple and important truth of fundamental human rights. History tel
EN20050111.x towed selectively, denied, or even take back. I will support the truth that the fundamental rights of the human
EN20050111.x d only has positive things to say about the Treaty. Where is the truth in all of this? You should be ashamed of
EN20050111.x urselves. The revolution must truly have gone a long way for the truth to be abused in such a fashion. Fortunat
EN20050111.x ave a democracy and, without debate, we do not get closer to the truth. It is therefore our responsibility to p
EN20050111.x on do you have that leads you to think - and the question is, in truth, also addressed to Mr Schmit - that the
EN20050112.x ble to the Treaty. The lack of any such reference means that the truth about our continent's history is blatant
EN20050112.x ot expect us to talk in soundbites, but to be honest. The simple truth about the political context of this disa
EN20050126.x ? Enlargement has demonstrated that Europe is a success, but the truth of the matter is that many citizens stil
EN20050126.x  of the President of the Council is an outstanding symbol of the truth that, when these institutions work toget
EN20050126.x he wealthiest countries or in the least-developed countries. The truth is that we must strive to achieve cohesi
EN20050126.x fada, the peace process and his willingness to hold dialogue. In truth, we must acknowledge what President Abu
EN20050126.x unacceptable links must be combated. This is a crime against the truth, to borrow the phrase coined by Presiden
EN20050126.x  is ending, and history is beginning. The duty to tell the whole truth about this tragedy rests with those of u
EN20050126.x  the reconciliation of nations is humility before the historical truth, as this is the only way in which forgiv
EN20050126.x  newspaper le Soir. We should therefore fight for the historical truth and condemn any manifestation of anti-Se
EN20050126.x itz, we have a particular duty to offend no one when telling the truth, but to ensure that this historical warn
EN20050127.x ty were, however, what was really needed for a genuine moment of truth to emerge. Clarity was needed, without t
EN20050127.x der how this memory is to be passed on without them. We owe them truth. To them first of all, but also to our c
EN20050127.x  false interpretation of history. If we dispense with historical truth, we may well heighten the threat posed t
EN20050127.x guilty and be a warning to today's world. We must learn from the truth about Auschwitz. As long as we continue
EN20050127.x laughtered in today's society, we will have failed to grasp that truth. I am referring also to the slaughter of
EN20050127.x to any inquiry. The British Government is continuing to hide the truth of Mr Finucane's murder for its own poli
```

Figure 4.4: Juxtaposed views filtering example



```
EN20050127.x guilty and be a warning to today's world. We must learn from the truth about Auschwitz. As long as we continue
EN20050111.x dent, I am from a country with a constitution that enshrines the truth about fundamental human rights: the trut
EN20051025.x History, our history, would have been different without you. The truth about human rights violations is now pub
EN20050511.x e nationalism. What we must do is tell the peoples of Europe the truth about our ambitions and give them free a
EN20050112.x ble to the Treaty. The lack of any such reference means that the truth about our continent's history is blatant
EN20050704.x e sovereign states in 1940 by the totalitarian Soviet Union. The truth about that should be made known and reme
EN20050926.x ion of Europe. It is also an opportunity to recall a fundamental truth about the activities of people and natio
EN20050308.x n of the United Nations Security Council of 15 February, for the truth about the causes and circumstances of th
EN20051116.x ve into a grey economy, for that is what is happening. The blunt truth about the directive is that in seeking t
EN20051027.x ndependent journalists and political activists who uncovered the truth about the events of 13 May. Following a
EN20051115.x r States, which up to now have been very reluctant to reveal the truth about the nuclear industry's inherited a
EN20050112.x ot expect us to talk in soundbites, but to be honest. The simple truth about the political context of this disa
EN20051214.x elsewhere in the world. It is now our duty to discover the whole truth about the reports of the kidnapping and
EN20051213.x port: Roure (A6-0366/2005)    Frank Vanhecke   Mr President, the truth about the Roure report is that it again
EN20050512.x tern and central Europe. Today we would like to reveal the whole truth about the Second World War, regardless o
EN20051116.x ainst the Kurds in 1991. That is a simple matter of justice. The truth about the siege of Falluja, about its bu
EN20050411.x ngulfed in a crisis of values, this House needs to hear the real truth about the source of John Paul II's moral
EN20051116.x alists attempting to remain in contact with us and to reveal the truth about their country's regime. Iran is al
EN20050126.x  is ending, and history is beginning. The duty to tell the whole truth about this tragedy rests with those of u
EN20050511.x  longer, and we must ensure that future generations remember the truth about this war. We Poles find it a sourc
EN20050512.x ny longer! I call for the continuation of investigation into the truth about totalitarianism, regardless of ide
EN20051027.x gnoring the international community's desire to reveal the whole truth about what happened in Andizhan in May t
EN20051214.x r the whole truth - a truth that we may or may not like, but the truth all the same. We have to find out the tr
EN20050111.x on do you have that leads you to think - and the question is, in truth, also addressed to Mr Schmit - that the
EN20050512.x of ideology. I call for this truth to be made public! Historical truth also demands international condemnation
EN20050706.x enica in another way, by giving it, at the least, the tribute of truth and an acknowledgement of the cowardice
EN20051214.x ay not like, but the truth all the same. We have to find out the truth and, as Mr Schulz rightly said, avoid ma
EN20050525.x amour. If we were to table a motion of censure it would seek the truth and block those policies that hinder emp
EN20050111.x ment. You will not win your constitutional battle by burying the truth and by failing to have a proper debate.
EN20051213.x ill have to give serious thought to what all this has to do with truth and clarity. A European Parliament that
```

Figure 4.5: Juxtaposed views filtering example

Chapter 5

Evaluation

Reviewing collocations frequencies using traditional concordance tools is an action common to many corpus analysis tasks [13]. The mosaic interface has been designed with the goal of making collocations frequency or strength visible at a glance. To evaluate this design we chose to compare the performance of a traditional concordance KWIC browser, the mosaic interface and juxtaposed views which combines the two into a single interactive interface. We chose to define performance as the speed and accuracy with which a participant completes an assigned corpus analysis task. These tasks will be adapted from the tasks detailed in [13] and the exemplified keywords will be used where possible.

Before performing user evaluation it is prudent to investigate any new user interfaces in a less costly manner, in order to minimise interface issues which could confuse, slow or irritate the participants. A common way to do this is with a heuristic evaluation.

5.1 Heuristic Evaluation

The implementations of the mosaic and juxtaposed views have gone through several iterations the latter of which were driven by the results of a heuristic evaluation. We had three interface experts evaluate the design, in the context of the heuristics we have chosen, and report any usability issues with reference to these heuristics.

Once the experts had been introduced on the operation of the system they were given a list of suggested keywords to investigate briefly, then asked to read the heuristics and finally to spend 20 minutes investigating the interface fully looking for usability issues and report any findings in terms of the heuristics.

We chose a total of six heuristics, three (Visibility of System Status, Recognition rather than recall, User control and freedom) are traditional heuristics which have been used for interface evaluation since the method was first proposed by Nielsen [10]. The other three were selected from a list of heuristics proposed more recently for the evaluation of visualisations [4].

5.1.1 Heuristics

Visibility of System Status

Users should be kept informed of system state, progress or changes.

Recognition rather than recall

Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.

User control and freedom

Users should be able to easily navigate to, act on and exit system components.

Information coding

Perception of information is directly dependent on the mapping of data elements to visual objects. This should be enhanced by using realistic characteristics/techniques or the use of additional symbols.

Spatial organization

Concerns users orientation in the information space, the distribution of elements in the layout, precision and legibility, efficiency in space usage and distortion of visual elements.

Match between system state and real world

Information should be presented in a logical/natural manner. Concepts and language should be familiar to the user.

5.1.2 Results

The issues raised must be considered in regard to the domain problem the design is trying to address, since the evaluation participants are interface experts and not users from the target domain some suggestions may not improve the design for the target users. However, many issues were raised which improved the interface.

After each evaluation time was taken to implement the changes suggested which we felt were most important. After the third evaluation we decided that we would end the evaluation as the issues being raised overlapped significantly with what was suggested by the first two participants.

To give an example of a change which was made to the interface as a result of the heuristic evaluation see the difference in color scheme between the mosaic visualisation presented in Figures 4.3 and 4.4. Two participants suggested that the color scheme in Figure 4.3 violated the heuristic of *information encoding*. They reported that using separate colors for each word added to visual clutter without encoding any additional information, since labels already conveyed this information. Both suggested the main function of the colouring is to make the borders between rectangles clear, and so a minimal color scheme which achieves this would be a more appropriate use of the visual variable color.

After making the required changes to the interface, teased out by the heuristic evaluation, we can move to testing the design with users in the problem domain with a greatly reduced chance of user frustration or confusion.

5.2 User Evaluation Plan

This section outlines a a user evaluation which we are planning to perform in the coming months. The details are still at the planning stage but we expect the eventual experiment to approximately follow what is described here. The question of the evaluation is "On which of the three conditions does the population have the best performance on tasks involving identification of collocations or colligations".

The population of interest is users with experience in corpus linguistics and in using traditional concordance tools, for this reason we have chosen to select our participants from students studying in this area.

The performance will be measured in terms of the speed of completion of the questions in each task the overall task completion time and the number and type of errors made.

The three experimental conditions we are going compare are: performing a task using the traditional KWIC interface, the mosaic interface and the juxtaposed views interface. The evaluation will be a within (same) user study where each user will perform each condition. To ensure the order in which the users are presented the interfaces does not bias the experiment we will counterbalance the order in which the conditions are presented. This will neutralise the effect of users learning from the previous tasks.

Task selection is an important design decision for the experiment. The tasks created must only vary slightly. If we create three different tasks that are of equal difficulty for a single condition we will have turned our study into a between user study, since natural groups are would be formed by the combinations of conditions and tasks. Similarly, creating multiple tasks and having users perform them all for each condition requires that we address the task order bias. We instead chose to have the users perform the same set of tasks for each condition, with a different concordance list and keyword. We must still ensure that these concordance lists are of equal difficulty for the tasks. The choice of the keywords is important for controlling the size of the concordance lists and the frequency profiles per word position. The size of the concordance lists we chose may affect the experiment, as we believe frequency information will be more difficult to quickly acquire from the KWIC interface as the number of concordance lines grows and should remain constant for the mosaic and juxtaposed interfaces.

We do not wish to have the participants talent as a corpus linguist affect their ability to perform the tasks we design. We will design the tasks so that the user performs an action from which they will have a quantitative result to report and which requires minimal linguistic knowledge to perform.

The concordance analysis tasks detailed by Sinclair [13] cover a broad range of linguistic analysis tasks such as hidden meanings or "semantic prosody", examining literal and metaphorical word usage, different usages of the singular and plural word forms, and reconsidering common word usage. The majority of the tasks require counting collocation or colligation frequencies with a keyword a various positions. We will take a collection of these counting actions and combine them with hand picked keywords to form our tasks. Our starting point for selecting the keywords will be the keywords used in Sinclair's tasks.

Before performing an experiment a number of practical issues must be addressed. Firstly, the experimental set-up has to be consistent. Our current prototype must be made more rigid so that possible user interactions are reduced to the set of interactions required to perform the tasks. Window placement, minimizing and closing windows must be restricted and consistent machine and screen specifications must be used when collecting the data. We must also develop the instrumentation of the prototype, logging participant interactions (clicks) and timing participants. An interface for participants to give there answers to each action must be included as well as a interface to display the questions. Finally, a briefing and perhaps practise task must be designed to introduce the participants to the designs and concepts they will encounter in the experiment.

What follows is a outline of how the experiment might be performed.

5.2.1 Example Experiment

The experiment will begin with the briefing and practise task, which should take approximately ten minutes. The participant will then perform a task for each of the three conditions. The tasks are designed to take approximately fifteen minutes for the slowest interface, a pilot study will be performed to determine if the length of the tasks need to be reduced to fit the fifteen minute threshold. For each condition we will us a different keyword but the task actions will remain the same. The current proposed list of actions are listed here the *italicised* items are selected based on the keyword.

1. For the keyword "*keyword*", list the two most common words at position n-1 and the number of occurrences of each word.

2. For the keyword "*keyword*", list the two most common words at position n+2 and the number of occurrences of each word.

3. For the keyword "*keyword*", list the 3 most common (*verb, adjectives ...*) from the words at positions n-1 and n-2 combined

4. Of the concordance lines which contain the word "*word*" at position n-1 what is the most common word at n-2 second? Possibly remind users they can search "*word*"+"*keyword*"

5. At position "position" suggest a word which has high collocation strength? Possibly remind them that collocation strength is collocation frequency divided by corpus frequency. The button labelled "Wc/ Word frequency in Corpus" in Figure 4.2 will be relabelled as "Concordance Strength"

The motivation for actions one and two are we wish to test speed and accuracy (correct count) of finding high frequency words at a position, under each condition. Having the second task be very similar to the first should show some indication if the participants have increased performance due to learning from the first task.

The third actions tests a more difficult task as the frequencies the participate must examine are split across two positions. It is a task to find the three most common colligations of a specific grammatical class across two word positions. We expect to see a greater difference between the conditions for more difficult actions. We should make sure the participant knows what the grammatical class we wish them to identify is. Perhaps we should give typical example words.

The fourth action requires using both frequency information and linear structure of the concordance lines. The information will not be available in the mosaic view without a new search for "*word*"+"*keyword*" and this search should make the action easier in both of the other conditions. We must decide if we should prompt the user with the availability of a new search. The search feature will have been explained in the introduction.

The final action, is a question we propose can best be answered using the collocation strength scaling of the mosaic view. We ask the user to suggest the strongest collocation at a position. Sinclair's sixth task [13] makes reference to uncommon words and how some of them only occur when specific subjects are discussed, while some others usually only appear in certain idiomatic phrases (e.g. kith in"kith and kin"). So we propose when examining a keyword it may be useful to determine the strength of its collocations. A problem with this action is that in the KWIC view the participant may have difficulty performing this action without many searches and counts. Perhaps a answer option to say it is too difficult to determine could be included or the action failed after a certain time limit.

Chapter 6

Discussion

The task of creating a data abstraction of KWIC concordance lists enabling graph based visualisations has been accomplished. The concordance graph gives us a method of identifying the keyword (using graph eccentricity) and allows the reconstruction of the concordance lines passing through any chosen vertex (using the contextual edges). This is beneficial since it enables exploration of these concordances in previously impractical ways. Examples of this are the interactive mosaic and bi-directional hierarchical visualisations which we created from these graphs.

The BDH lets us view the entire context of the selected keyword as an overview, while focusing on a selected word vertex gives the detail i.e. which concordance lines include that vertex. This has the screen space saving advantages of the Word Tree visualisation while not suffering from the restriction of only displaying a single side of the context. Thus making certain corpus linguistic tasks, such as identifying phrases which include words from both contexts, easier. However, due to the manner in which nodes are merged and scaled, frequency information is difficult to interpret at a glance at. This is because multiple vertices with the same label can appear at a position, as merging them would have caused ambiguity in reconstructing concordance lines.

In contrast, the mosaic (space-filling) interface throws out the linear structure of the concordance lines, so that word frequency per position (or other metrics such as collocation strength) can be determined at a glance. Viewing the concordance in this way immediately makes clear any high frequency co-occurrences with the keyword. We believe this interface will allow faster and more accurate location of high frequency collocations when compared with pre existing concordance visualisation tools.

We also implemented a scaling scheme for the mosaic which emphasises collocation strength over raw frequency. We have defined this positional collocation strength as a words frequency at a position relative to a keyword divided by its total occurrences in the selected corpus. This measure increases the visibility of words that dont occur often when the keyword is not present. This is useful for finding patterns of concurrences that only appear in the context of certain topics or phrases.

The main disadvantage the mosaic interface has in comparison to traditional concordance tools is the missing linear structure of the concordance lines. To try and make of for this highlighting of sentence components which join a selected word was enabled by traversing the concordance graph. However, even with the components of the concordances lives visible it is not possible to determine the exact concordance lines from the visualisation. To overcome this limitation without resorting to tactics such as colouring selections to show sentence structure we created a composite visualisation of the mosaic and traditional KWIC view.

This juxtaposed view not only consisted of two views on the data it implicit links the views and the data such that interactions on either view affects the other. selection of word rectangles on the mosaic sorts the KWIC alphabetically on the selected position and scrolls to the the first occurrence of the selected word now highlighted on both views. This composite view should be greater than the sum of its parts due to these new linked interactions.

To investigate if the mosaic or juxtaposed views offer any improvements over the TEC KWIC concordance browser we have decided to conduct a user evaluation using linguistics students who are already familiar with concordance analysis using KWIC browsers. This evaluation has not been conducted yet and the evaluation plan is detailed in this report. To briefly summarise, the evaluation will take the form of a within (same) participant study where the three conditions will be evaluate. The evaluation will be based on a set of common corpus analysis actions involving the investigation of collocations. All interactions will be logged and timestamped so a rigorous statistical analysis an be performed on the results.

Beyond the evaluation, future work involves investigating other mosaic scaling schemes and using the mosaic for corpus comparison. Using the mosaic for corpus comparison could be done by visualising the difference in collocation occurrences (word usage) between two corpus or sub-corpus. This corpus comparison could have many applications for example comparison of author writing style or comparison of translated and native language texts.

Chapter 7

Conclusions

Representing a keyword-in-context concordance list as a connected graph offers several benefits over the traditional aligned list of sentence fragments approach. The graph based abstraction gives us the ability to use a wide rang of graph based techniques in analysis and visualisation of the concordance lists that do not have well defined counterparts in dealing with lists of sentence fragments. These graphical techniques include operations such as edge contraction to merge similar nodes, breadth first traversal to perform word position operations, and applying various graph layout operations for visual transfer.

Two important characteristics of this graphical abstraction are that the keyword can be uniquely identified using the property of graph eccentricity and that the linear structure of the entire concordance lines can be recovered by including contextual edges in the graph. The first property allowed us build visualisations where the keyword was clearly identified and other vertices are categorised by their distance and direction from the keyword. The second property enables the the recovery of the individual concordance lines from the graph, this puts it on par with traditional keyword-in-context displays and improves on the tree based abstraction, which is used in Word Tree, as both contexts can be linked and recovered using the contextual edges.

To exemplify the additional concordance visualisations possible due to this abstraction two new concordance visualisations were created. One, which we call a bi-directional hierarchical display, is a natural extension of the Word Tree visualisation where position is given higher importance and both context trees are joined on the keyword instead of being displayed separately. This join makes sentence structure ambiguous across the keyword in the case of simply joining context trees, however when the underlying representation is a concordance graph we can enable selection operations which highlight the concordance lines.

The second visualisation was designed to emphasise frequency information by creating a mosaic (space-filling) representation where vertical scale represented some relative score at that horizontal word position. The scores we chose to implement were word position concordance frequency and word position concordance strength. We propose that this mosaic visualisation will improve speed and accuracy identifying collocation and colligation patterns. To test this proposal a detailed plan for a user evaluation of the mosaic, traditional keyword-in context view and a composite of both views, has been included in this report. We hope to carry out this work in the coming months.

Bibliography

[1] E. H Chi and J. T Riedl. An operator interaction framework for visualization systems. In *Procs. of the IEEE Symposium on Information Visualization*, pages 63–70, 1998.

[2] Ed H. Chi. A taxonomy of visualization techniques using the data state reference model. In *Proceedings of the IEEE Symposium on Information Vizualization 2000*, INFOVIS '00, pages 69–, Washington, DC, USA, 2000. IEEE Computer Society.

[3] Ed H. Chi. Expressiveness of the data flow and data state models in visualization systems, 2001.

[4] Camilla Forsell and Jimmy Johansson. An heuristic set for evaluation in information visualization. In *Proceedings of the International Conference on Advanced Visual Interfaces*, AVI '10, pages 199–206, New York, NY, USA, 2010. ACM.

[5] Jeffrey Heer, Stuart K. Card, and James A. Landay. prefuse: a toolkit for interactive information visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '05, pages 421–430, New York, NY, USA, 2005. ACM.

[6] W. Javed and N. Elmqvist. Exploring the design space of composite visualization. In *Pacific Visualization Symposium (PacificVis), 2012 IEEE*, pages 1–8, 2012.

[7] Saturnino Luz. A software toolkit for sharing and accessing corpora over the Internet. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhauer, editors, *Proceedings of the Second International Conference on Language Resources and Evaluation: LREC-2000*, pages 1749–1754, May 2000.

[8] Saturnino Luz. Web-based corpus software. In Alet Kruger, Kim Wallmach, and Jeremy Munday, editors, *Corpus-based Translation Studies – Research and Applications*, chapter 5, pages 124–149. Continuum, 2011.

[9] Saturnino Luz and Masood Masoodian. Visualisation of parallel data streams with temporal mosaics. In *Procs. of the 11th International Conference on Information Visualisation*, pages 196–202, Zurich, 2007. IEEE.

[10] Jakob Nielsen. Usability inspection methods. In *Conference Companion on Human Factors in Computing Systems*, CHI '94, pages 413–414, New York, NY, USA, 1994. ACM.

[11] Ben Schneiderman. Tree visualization with tree-maps: 2-D space-filling approach. *ACM Transactions on Graphics*, 11(1):92–99, January 1992.

[12] J. Sinclair. *Corpus, concordance, collocation.* Describing English language. Oxford University Press, 1991.

[13] J.M. Sinclair. *Reading Concordances: An Introduction.* Longman Publishing Group, 2003.

[14] Robert Spence. *Information visualization.* Addison-Wesley Harlow, 2nd edition, 2007.

[15] Robert Spence and Mark Apperley. Bifocal display. *The Encyclopedia of Human-Computer Interaction, 2nd Ed.*, 2013.

[16] F.B. Viégas and M. Wattenberg. Tag clouds and the case for vernacular visualization. *Interactions*, 15(4):49–52, 2008.

[17] Matthew Ward, Georges Grinstein, and Daniel A. Keim. Interactive data visualization: Foundations, techniques, and application. May 2010.

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