Information Visualisation Applied to Corpus Linguistic Methodologies

Shane Sheehan

A dissertation submitted to Trinity College, University of Dublin for the degree of Doctor of Philosophy

School of Computer Science and Statistics
Trinity College, University of Dublin

August 2023

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Shane Sheehan

Summary

This thesis uses established visualisation design methods to characterize problems in corpus linguistics. The identified problem areas are concordance collocation patterns, frequency list comparison, and concordance meta-data analysis. The identification of these problems required collaboration with researchers from corpus linguistics. These collaborations explored example methodologies and research questions in the domain.

Each of the three identified problem areas was addressed by designing visualisation tools. The three visualisations described in the thesis are:

1. **Mosaic** visualisation of positional collocation patterns in concordance.

2. **ComFre** visualisation for frequency list comparison

3. **MetaFacet** visualisation for exploring meta-data facet distributions of concordance lists

A mix of encoding justifications, methodological impact/adoption, and laboratory study are used to validate the visualisations.

Mosaic effectively visualized collocation patterns showing improved speed and accuracy over established methods. Concordance Mosaic's methodological impact was also high as corpus linguistic researchers adopted it to improve the efficiency of analysis.

The ComFre visualisation was effective in comparing frequency lists even in situations where the lists are of vastly different sizes. The methodological impact of the technique had to be assessed as low since its only evidence of methodological adoption was in the form of an example method created by a domain expert to demonstrate the tool's usefulness.

MetaFacet was not available during the methodological review process. It does, however, show clear advantages in task time for methodologies revealed during the review.

Acknowledgements

This thesis could not have been completed without the help of my mentors, colleagues, family and friends.

I want to thank Dr Saturnino Luz and Dr Martin Emms for their help, supervision, and patience during this process.

The members of the Genealogies of Knowledge project are owed a debt of gratitude for the time, resources, and expertise they so generously shared.

I offer my heartfelt thanks to my office mate and friend, Dr Derek Kelleher.

I must thank the ADAPT centre and Prof. Vincent P. Wade for the opportunity to conduct this work. The ADAPT centre and all its members I consider not just colleagues but friends.

I want to thank my parents for their support during this period and throughout my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The result of visualisation is not a picture or an interface but a mental model which can be focused, expanded, or navigated through interaction. Visual encoding without justification must either be viewed as an artistic expression of the data or analysed by the user to determine how the visualisation lens distorts the underlying data. A visualisation should be grounded in the problem domain characterisation. It should link the problem characterization to data abstractions and fully explain its visual encoding choices. Mistakes at earlier stages propagate through the design process, often leading to unsuccessful visualisation designs.

With this in mind, I set out in this thesis to explore the corpus linguistics domain and find problems with effective visual solutions. Concordance analysis is a corpus linguistic methodology wherein every instance of a keyword and its surrounding context are extracted from a document or corpus and presented for inspection by an analyst. The context consists of the words preceding and following the keyword instances. The linguistic properties of the concordance list are then examined by the analyst, using word frequencies, part of speech frequencies (colligations), patterns of occurrence, or other linguistic properties [Sinclair, 2003, Scott, 2010].

Since Hans Peter Luhn invented the keyword-in-context indexing technique in the 50's [Luhn, 1960], concordance analysis has grown in popularity in fields such as corpus linguistics, classical studies, and translation studies. The keyword-in-context (KWIC) visualisation technique is the most used tool for analyzing concordance lists. This visualisation consists of fragments of text containing the keyword, which have been arranged vertically so that each occurrence of the keyword is aligned centrally in the view.

Concordance analyses focus on analyzing keywords and their context in a corpus. A natural question is how does one choose what keywords to investigate. The domain expertise and understanding of the corpora under investigation play an important role in finding useful keywords. Frequency lists are usually used to gain further insight into the structure of a corpus. It is often useful to view the frequency in reference to usage in some reference corpus, perhaps of general language usage. These comparisons are generally performed using a side-by-side comparison of the rank and frequency of two frequency lists.

In the last decades, empirical and corpus-based text analysis methods have come to the fore. This is partly due to the vast amounts of digital text and changing theoretical perspectives. Computational and statistical methods for analysing textual resources are collectively referred to as text analytics by language researchers and, increasingly, end users of this technology. One of the key features of textual resources in digital formats is the addition of meta-data to supplement the raw text. In concordance tools, this meta-data seems to be mostly ignored for anal. The filename is usually supplied with a concordance, but more detailed information is often only available as an afterthought.

The work presented in this thesis is the characterization of these three areas of corpus analysis and the design and evaluation of visualisations to support them.

1.1 Research Questions

The research questions which drive the research presented in this thesis are:

1. Which methods employed by corpus linguists are not well served by visualisation?

2. How would visualisation tools affect the workflows of corpus linguists?

3. Can collocation patterns be visually summarized effectively?
4. Can unequal-length word frequency lists be effectively compared using visualisation?
5. Can meta-data visualisation supplement the concordance methodology?

By answering these questions, it should be possible to arrive at problems in corpus linguistics which are not adequately served by visualisation. These problems can be investigated regarding visualisation's effect on existing methods and where relevant new visualisation methods can be developed. Finally, the effectiveness of the proposed visualisation solutions can be evaluated in relation to the identified problems.

1.2 Contributions

The main contributions of this thesis are three visualisation tools for the analysis of corpora:

1. **Mosaic** visualisation of positional collocation patterns in concordance.
2. **ComFre** visualisation for frequency list comparison
3. **MetaFacet** visualisation for exploring meta-data facet distributions of concordance lists

A secondary contribution is an investigation corpus analysis methodologies which were observed during the research.

1.3 Publications

Part of the work in this thesis is based on the following two papers

COMFRE: A visualisation for Comparing Word Frequencies in Linguistic Tasks [Sheehan et al., 2018]

A Graph-Based Abstraction of Textual Concordances and Two Renderings for Their Interactive Visualisation [Luz and Sheehan, 2014]

1.4 Thesis Structure

The remainder of this thesis is structured as follows:

In **Chapter 2**, I present an overview of background work and review the related literature on corpus linguistics and visualisation design. I begin with a discussion of corpus linguistics in general, then narrow the focus to concordance analysis. The literature on visualisation design focuses on visual variable encoding and visualisation design frameworks.

In **Chapter 3**, I present a Domain characterization that includes questions corpus linguists would like to ask a corpus, example methodologies, and a literature-based task analysis.

In **Chapter 4**, I present a conceptual data model of keyword-in-context concordances. In combination with the domain characterisation, this conceptual model enables structured encoding comparison of existing keyword-in-context visualisations. In addition, a structured encoding comparison of frequency list visualisation is presented.

In **Chapter 5**, I present the design of the Concordance Mosaic interface for concordance collocation pattern analysis. A graph-based data abstraction and details of the encoding design are included here.

In **Chapter 6**, I present the ComFre interface for the comparison of frequency lists. This chapter includes requirements analysis, encoding design and implementation details.

In **Chapter 7**, I present an evaluation of the Concordance Mosaic. The evaluation takes the form of a laboratory study. The study measures the time to complete tasks and the correctness of the achieved result to compare the Concordance Mosaic with the traditional Keyword-in context visualisation.

In **Chapter 8**, I present a review of the methodological impact of the Mosaic and ComFre visualisations.

In **Chapter 9**, I present the MetaFacet Visualisation for analyzing metadata counts from a concordance.

In **Chapter 10** contains a discussion of the presented material from which conclusions are drawn.

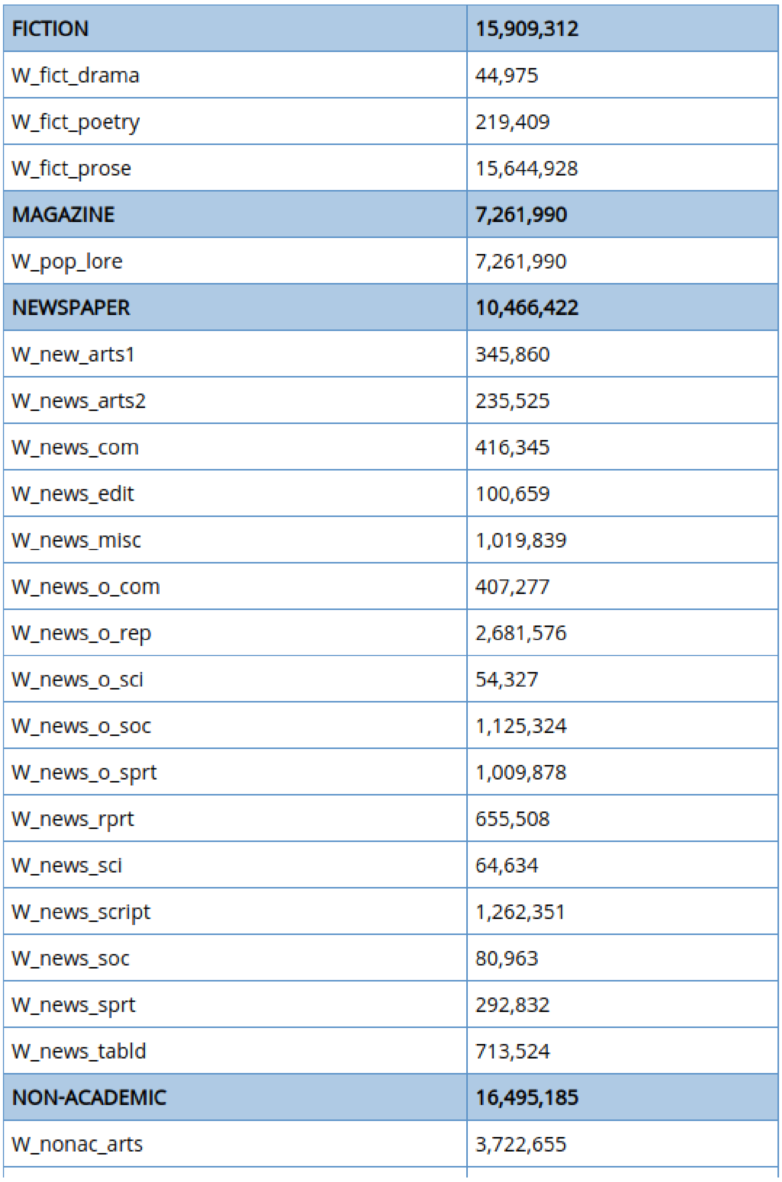
Chapter 2

Background

2.1 Corpus Linguistics Background

Corpus linguistics is fundamentally the analysis of language using collections of texts. These collections are known as corpora. The corpus used for analysis must be carefully chosen or created as “the results are only as good as the corpus” [Sinclair, 1991, p. 13]. Corpus linguistics is not strictly a branch of linguistics; it is sometimes described as a methodology for performing linguistic analysis. Another view of corpus linguistics is that it goes far beyond a simple methodology and should be treated as a discipline in applied linguistics [Tognini-Bonelli, 2001]. Describing a general corpus-based research methodology is difficult as there is no agreed-upon research methodology nor a common set of research questions [Thompson and Hunston, 2006] and, as Mahlberg [Mahlberg, 2005] points out, there is no sign of a unifying theory binding together corpus linguistics analyses [Barlow, 2011]. One of the reasons put forward by Gries [Gries, 2010] for the problematic relationship between Corpus linguistics and linguistic theory is that “corpus linguists differ as to what they think CL [corpus linguistics] is: a tool, method, discipline, theory, paradigm, framework, ...;”[Gries, 2010, p. 1]

When corpora are used for analysis in linguistic branches such as semantics or syntax, they are referred to as corpus-based, i.e. corpus-based semantics or corpus-based syntax. This distinction serves to differentiate the approach from introspective linguistic techniques.



FICTION	**15,909,312**
W_fict_drama	44,975
W_fict_poetry	219,409
W_fict_prose	15,644,928
MAGAZINE	**7,261,990**
W_pop_lore	7,261,990
NEWSPAPER	**10,466,422**
W_new_arts1	345,860
W_news_arts2	235,525
W_news_com	416,345
W_news_edit	100,659
W_news_misc	1,019,839
W_news_o_com	407,277
W_news_o_rep	2,681,576
W_news_o_sci	54,327
W_news_o_soc	1,125,324
W_news_o_sprt	1,009,878
W_news_rprt	655,508
W_news_sci	64,634
W_news_script	1,262,351
W_news_soc	80,963
W_news_sprt	292,832
W_news_tabld	713,524
NON-ACADEMIC	**16,495,185**
W_nonac_arts	3,722,655

Figure 2.1: British National Corpus (BNC) sample of register breakdown and word counts. Davies, Mark. (2004-) BYU-BNC. (Based on the British National Corpus from Oxford University Press). Available online at https://corpus.byu.edu/bnc/ .

While any collection of more than one text is technically defined as a corpus, in modern linguistics, a corpus is expected to be well defined in terms of its sampling and representativeness, size, and machine-readable format.[McEnery and Wilson, 1996]. For example, the British National Corpus (BNC), created in the 1980s and early 1990s, contains over 100 million words from texts in various genres. The Corpus is separated into registers such as *Fiction* or *Newspaper*, and word counts for each register are supplied. An even more fine-grain breakdown of the registers is provided to define the representativeness of the texts further, as seen in Figure 2.1. The corpus is exposed as a web interface that enables various forms of analysis.

Pioneering work in corpus creation by Kučera and Francis in 1967[Francis and Kucera, 1967] led to the Brown Corpus of written American English, the first computer-readable general corpus of texts prepared for linguistic research on modern English. The corpus contains over one million words of text and has been tagged to add grammatical information about each individual word (token) in the corpus. The analysis of the Brown corpus, which appeared in *Frequency Analysis of English Usage: Lexicon and Grammar* [Francis et al., 1982], is the earliest use of a modern corpus in linguistics.

Modern computers have made corpus analysis quite popular at a previously impossible scale [Sinclair, 1991]. However, corpus-based approaches to linguistics have been used as early as the nineteenth century, where studies of infant language were collected and laboriously analyzed by hand [Taine, 1877, Preyer, 1898]. Several later works, up until the mid-1950s, are clearly corpus-based. Boas's work in "Linguists of the structuralist tradition" [Boas, 1941], Fries and Traver's [Fries and Traver, 1940] work in language pedagogy, and Eaton's [Eaton, 1940] study of word meaning frequency in several languages are all examples of early corpus-based studies.

In the late 1950s, the direction of linguistics quickly moved away from corpus-based approaches and toward rationalist theories of linguists. Extremely influential works from Chomsky [Chomsky, 1957, Chomsky, 1965] were heavily critical of corpus-based approaches to linguistics, favouring an approach based on theory and introspective judgment. The criticisms can be summed up using two key arguments. The first is that corpus-based methods encourage the modelling of the wrong thing. They encourage the enumeration and description of linguistic phenomena, while the goals of linguistic analysis should be introspection and linguistic competence. The second criticism was that even if the enumeration of linguistic phenomena were the correct goal, a finite corpus can never represent an infinite language. These criticisms were viewed at the time as fatal, and the techniques went through a period of neglect during the 1960s and 1970s. However, the methodology was not entirely abandoned. Over the passage of time, some of the criticisms were taken on board while others proved to be irrelevant.[McEnery and Wilson, 1996] After this period of drought in corpus usage, the techniques began to gain popularity; in the period 1976-1991, the number of studies published doubled every five years from 10 studies prior to 1965 to 320 in the five years before 1991 [Johansson, 1991]. A citation analysis study of corpus linguistics references in the period 1997 to 2016 identified a growing number of corpus linguistic publications; in the period 1997-2001, 286 documents were identified; in 2002-2006 there were 669 publications; in 2007-2011, there were 1,792 corpus linguistic publications identified, and in 2012-2016 there were a total of 2,853 documents [Park and Nam, 2017].

The practitioners of corpus linguistics range across many diverse disciplines. For example, McEnery and Wilson [McEnery and Wilson, 2001] introduce corpus linguistics by covering topics such as lexical studies, grammar, semantics pragmatics and discourse analysis, sociolinguistics, stylistics and text linguistics, historical linguistics, dialectology, variation studies, psycholinguistics, the teaching of languages and linguistics, cultural studies, and social psychology. While the exact methodology differs in each case, the use of computer-generated quantitative information, the investigation of lexical or grammatical patterns in a corpus, and the qualitative discussion of the quantitative and textual information are consistent components of the techniques.

One example area in which corpus-based methods have grown in popularity is Translation studies. Baker's [Baker et al., 1993] early advocacy for the use of corpus-based methods in the study of translation has led to its adoption in various sub-fields of translation [Olohan, 2002, Baker, 1995, Rabadán et al., 2009, Zanettin, 2001, Zanettin, 2013]. The re-emergence of corpus-based methods had a transformative effect. The corpus-based methodology has been described as one of the most important gate openers to progress in translation studies.[Hareide and Hofland, 2012] Corpora, for use in translation studies, were developed more slowly than in other linguistic fields. Translations were excluded from language reference corpora, such as the BNC. This exclusion would suggest translations were not viewed as representative of language in use by other linguists.[Olohan, 2002] Olohan's introduction to corpus-based translation studies [Olohan, 2004] puts corpus design as a key focus. The types of corpora useful for translation studies, such as parallel and comparable corpora, must be carefully constructed, taking into issues of corpus design and representativeness. The choices of which translations and versions of a text to include in a corpus implicitly affect the assumptions of any research carried out using the corpus.

The design of corpora for translation studies is an active area that has to deal with additional design challenges not often found in traditional corpora, such as representing related texts (for example, translations of the same source text) and designing multi-lingual corpora [Zubillaga et al., 2015]. The Translational English Corpus (TEC) [Luz, 2000] is one of the oldest corpora designed specifically for translation studies. TEC contains a collection of English texts translated from a variety of languages, both European and non-European. The corpus is currently maintained by the University of Manchester. TEC made possible empirical investigations into *the universals of translation* and was quickly followed by monolingual comparable corpora in languages such as Finnish, Swedish, and Portuguese. *Simplification* "the idea that translators subconsciously simplify the language or message or both" [Baker, 1996, p. 176] was originally investigated using the TEC corpus [Laviosa, 1997, Laviosa, 1997, Laviosa, 1998] before being replicated using a more recent corpus [Xiao, 2010].

The use of corpora in translation studies has now come of age in both descriptive and applied research [Laviosa, 2010], and ambitious projects such as *Genealogies Of*

Knowledge continue to create modern corpora for the translation research community.

While the exact definition and research questions of corpus linguistics vary, the fundamental methods at the core of corpus-based investigations are consistent and underpin the quantitative and qualitative justification for observable linguistic phenomena.

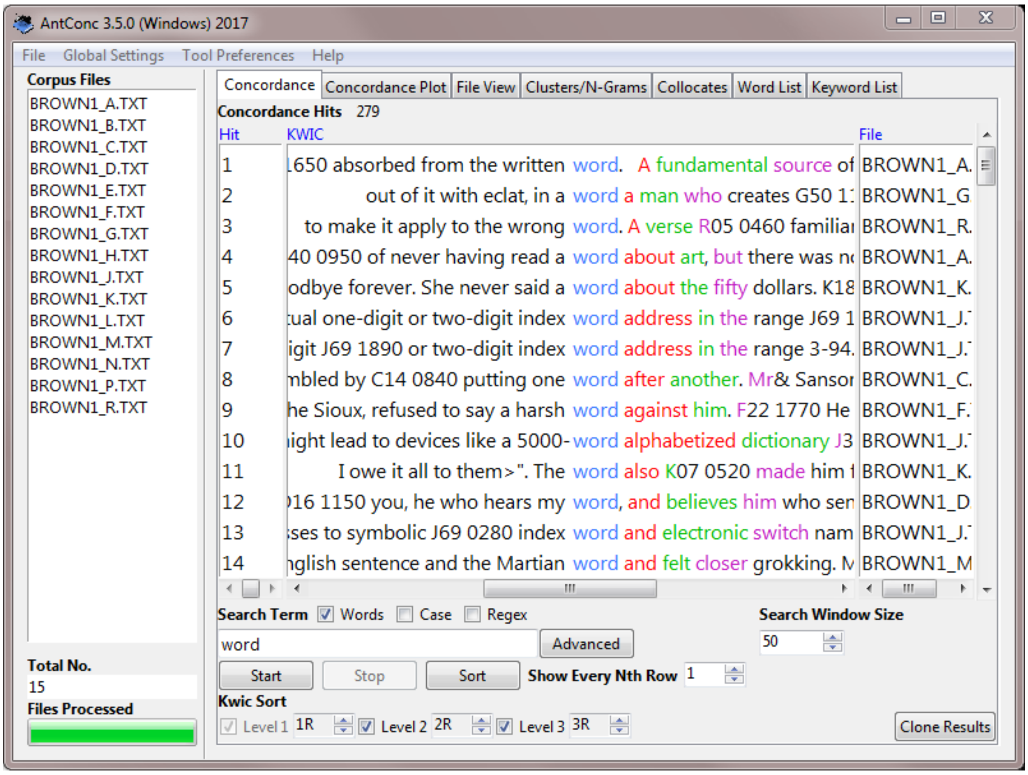
2.1.1 Corpus linguistic methodology

Biber, Conrad, and Reppen [Biber et al., 1998] make a distinction between the use of corpus linguistics for the study of language structure and language use. Topics such as grammar, lexico-grammar, and discourse characteristics are discussed under the umbrella of analyzing the characteristics of language features. Historical and stylistic investigations, language acquisition and development, and English for specific purposes are used to exemplify techniques for investigating the characteristics of varieties. This second group of examples emphasizes analysing and comparing registers and dialects. Registers are defined as varieties defined by their situational characteristic. These registers can be very specific such as the works of an individual author, or more general, such as conversational language. Registers differ from dialects; dialects are defined by their association with a speaker group, such as regional or social groups. The analysis of the corpus in the examples treats each register as its own corpus (sub-corpus). Characteristics within each register are interesting, as are the between-register comparisons.

Corpus linguistics makes use of quantitative techniques, employing frequency and statistical analysis to collect evidence of language structure or usage. Many of the methods can be thought of as empirical linguist techniques. However, a common misconception is that corpus-based approaches are entirely quantitative and do not require any qualitative input [Baker, 2006]. Biber presents a collection of quantitative corpus-based methods. In each case ".. a great deal of space is devoted to the explanation, exemplification, and interpretation of the patterns found in quantitative analyses. The goal of corpus-based investigations is not simply to report quantitative findings but to explore the importance of these findings for learning about the patterns of language use"[Biber et al., 1998, p. 5]. These qualitative interpretations are important "... a crucial part of the corpus-based approach is going beyond the quantitative patterns to propose functional [qualitative] interpretations explaining why the patterns exist"[Biber et al., 1998, p. 9]. In a linguistic study, before the application of quantitative techniques, the formulation of hypothesis and research questions is often informed by qualitative analysis and/or prior knowledge of the texts under investigation. A qualitative approach must precede a good quantitative study if anything beyond a simple description of the statistical properties of the corpus is to be achieved [Schmied, 1993].

2.1.2 The Concordance

Concordance analysis is a core activity of scholars in a number of humanities disciplines, including corpus linguistics, classical studies, and translation studies, to name a few. Through the advent of technology and the ever-increasing availability of textual data, this type of structured analysis of text has grown in importance [Sinclair, 1991, Bonelli, 2010]. In the public domain, concordance programs are the most common way corpora are manipulated, and many commercial concordance tools are available. Some of the most popular tools which have concordance browsing at their core include *WordSmith Tools* [Scott, 2008], *SketchEngine* [Kilgarriff et al., 2008] and *AntConc* [Anthony, 2005]. An example of the concordance browser AntConc in use can be seen in Figure 2.2. The keyword "word" is highlighted in blue, and the concordance is sorted at the position immediately to the right of the keyword (highlighted in red). While there is some variation in advanced features across the range of concordance browsers, each provides a windowed concordance that can be explored (via scrolling or multiple pages) and is usually sortable at word positions. This simple feature set is the key to supporting the traditional corpus linguistic methodology of concordance analysis.



Corpus Files
BROWN1_A.TXT
BROWN1_B.TXT
BROWN1_C.TXT
BROWN1_D.TXT
BROWN1_E.TXT
BROWN1_F.TXT
BROWN1_G.TXT
BROWN1_H.TXT
BROWN1_J.TXT
BROWN1_K.TXT
BROWN1_L.TXT
BROWN1_M.TXT
BROWN1_N.TXT
BROWN1_P.TXT
BROWN1_R.TXT

Total No.
15
Files Processed

Concordance Hits 279

Hit	KWIC	File
1	650 absorbed from the written word. A fundamental source of	BROWN1_A.
2	out of it with eclat, in a word a man who creates G50 1:	BROWN1_G
3	to make it apply to the wrong word. A verse R05 0460 familia	BROWN1_R.
4	40 0950 of never having read a word about art, but there was n	BROWN1_A.
5	odbye forever. She never said a word about the fifty dollars. K18	BROWN1_K.
6	ual one-digit or two-digit index word address in the range J69 1	BROWN1_J.
7	igit J69 1890 or two-digit index word address in the range 3-94.	BROWN1_J.
8	nbled by C14 0840 putting one word after another. Mr& Sanso	BROWN1_C.
9	he Sioux, refused to say a harsh word against him. F22 1770 He	BROWN1_F.
10	ight lead to devices like a 5000-word alphabetized dictionary J3	BROWN1_J.
11	I owe it all to them>". The word also K07 0520 made him t	BROWN1_K.
12	16 1150 you, he who hears my word, and believes him who sen	BROWN1_D
13	ses to symbolic J69 0280 index word and electronic switch nam	BROWN1_J.
14	nglish sentence and the Martian word and felt closer grokking. M	BROWN1_M

Search Term ☑ Words □ Case □ Regex
word
Search Window Size 50
Show Every Nth Row 1
Kwic Sort ☑ Level 1 1R ☑ Level 2 2R ☑ Level 3 3R

Figure 2.2: AntConc Concordance Browser.

In concordance analysis, every occurrence of a keyword of interest in a corpus is displayed along with its context. The context is an ordered list of words that precede and follow the keyword. The analyst then seeks to discover the linguist properties of the keyword and the contextual patterns which predict them by observing the frequencies of occurrence, in the keyword's context, of words (collocations), word combinations, parts of speech (colligations) or the various other lexical classifications [Sinclair, 2003, Scott, 2010].

The most widely used tool in this kind of analysis is a form of tabular visualisation tool known as a concordance browser. Hans Peter Luhn first proposed the creation of concordances through the keyword in context indexing technique in the 50's [Luhn, 1960]. KWIC displays are enhanced in interactive systems by features such as search and context sorting and are widely used not only by academics and scholars but also by professional translators and post-editors [Karamanis et al., 2011].

```
       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 2.3: KWIC display for the word “eye”.

Corpora typically contain more information than an individual can efficiently analyze without sampling. A KWIC-indexed corpus can efficiently retrieve every occurrence of a keyword and the contexts in which it appears in the corpus. The retrieved concordance is typically displayed with the keyword aligned centrally. A very simple representation of a KWIC display for a sample concordance of the keyword “eye” is shown in Figure 2.3. The keyword is aligned vertically, and five words of the left context are displayed, and four words of the right context are in view. The choice of five and four words is arbitrary in this case. However, Sinclair [Sinclair, 2003, p. 174] mentions a span of four or five words is “sufficient for most descriptive purposes and not so large that a great deal of extraneous material is also collected”. In this oversimplified example two lexical patterns in the concordance are apparent, the patterns “the naked eye” and “the watchful eye”.

Analyzing a corpus by using a concordance, is very different from traditional analysis of an individual text [Tognini-Bonelli, 2001, p. 2-4]. Table 2.1 lists several points of difference between analyzing language use in an individual text and a corpus. When examining a corpus, evidence found relates to a generalized form of usage based on the formal definition of the corpus, evidence gleaned from a text, while found in a (verbal) context, is also evaluated in terms of the cultural and situational context of the text. Traditional analysis involves reading horizontally (from left to right in English), boundaries such as sentences and paragraphs are important for the

A Text	A Corpus
1. read whole	1. read fragmented
2. read horizontally	2. read vertically
3. read for content read	3. read for formal patterning
4. read as a unique event	4. read for repeated events
5. read as an individual act of will	5. read as a sample of social practice
6. instance of *parole* (the actual linguistic behavior or performance of individuals, in contrast to the linguistic system of a community.)	6. gives insights into *langue* (a language viewed as an abstract system used by a speech community, in contrast to the actual linguistic behaviour of individuals.)
7. coherent communicative event	7. not a coherent communicative event

Table 2.1: Differences between a text and a corpus. Adapted from *Corpus Linguistics at Work* [Tognini-Bonelli, 2001, p. 2-4].

analysis. "Reading" a corpus in the KWIC format, is done vertically by scanning the contexts around the keyword for repeated patterns. The concordance makes the context around the keyword available, and the individual instances can be read horizontally in the broader context of the vertical patterning. The concordance can be viewed as represented on the horizontal axises syntagmatic structure, and the vertical axis represents the pragmatic availability (the meaning choices available to a writer) in the corpus under investigation. Furthermore, many corpora include information about the texts they contain (meta-data), and linking this information to the concordance narrows the gap between the quantitative and qualitative analysis available in a concordance browsing tool. The TEC corpus browser, for example, has a feature which displays all of the available information (meta-data) about a concordance line, such as the source file name, author, publication date and more.

Paul Baker explicitly sets out a step-by-step approach to a typical concordance analysis in the discipline of discourse analysis. Table 2.2 shows these steps. Between steps one and two, the expertise of the linguist is essential in deciding the starting point for analysis. Many of the steps call for the identification of patterns using the

Concordance analysis of discourse

1. Build or obtain access to a corpus
2. Decide on a search term/terms
3. Obtain a concordance of the search terms
4. Clean the concordance (by removing repetitions or irrelevant lines)
5. Sort the concordance repeatedly on different words to the left and right while looking of evidence of grammatical semantic or discourse patterns
6. Look for further evidence of such patterns in the corpus
7. Investigate the presence of particular terms more closely (explore collocates or distribution in reference corpora of general language)
8. Once no more patterns can be found, carry out a close analysis of the remaining concordance lines. (Look for similarities or patterns in terms of meaning or discourse)
9. Note rare or non-existent cases of discourses based on your own intuitions. (See if such discourses occur in other more general corpora
10. Attempt to hypothesize why the patterns appear and relate this to issues of text production and reception

Table 2.2: Condensed version of a "Step by step guide to concordance analysis" found in *Using corpora in discord analysis* [Baker, 2006, p. 92].

concordance. The close analysis of patterns in concordance lines is used to form a hypothesis in relation to the linguistic property of interest.

One of the foundational works in concordance analysis is *Reading Concordance* by John Sinclair[Sinclair, 2003]. In *Reading Concordance*, eighteen tasks are presented as examples of concordance analysis. The tasks are presented in a workbook-like manner with example concordances and questions to answer. Each task comes with a detailed answer key that explains the analysis steps required to answer the questions and the linguistic interpretations of the analysis.

Table 2.3 gives an overview of the tasks in *Reading Concordance* . These tasks are organized into four difficulty levels and cover a wide range of analyses. The reader is taught how to use corpus evidence to understand word meaning by looking for many different types of linguistic features. Many tasks require analysis of the concordance in terms of word frequency relative to a keyword. For example, in task four of level

Level 1	level 2	level 3	level 4
1. How meanings are shown 2. Underlying regularity 3. Words as liabilities 4. Literal and metaphorical 5. Meaning focus	1. Specialised meaning 2. Subtle distinctions 3. Meaning flavour 4. Extensions of grammar 5. Meaning and context	1. Words difficult to define 2. Ad-hoc meaning 3. Grammatical frames 4. Hidden meanings	1. Closely related meanings 2. One and one is not exactly two 3. Common words 4. Singular and plural

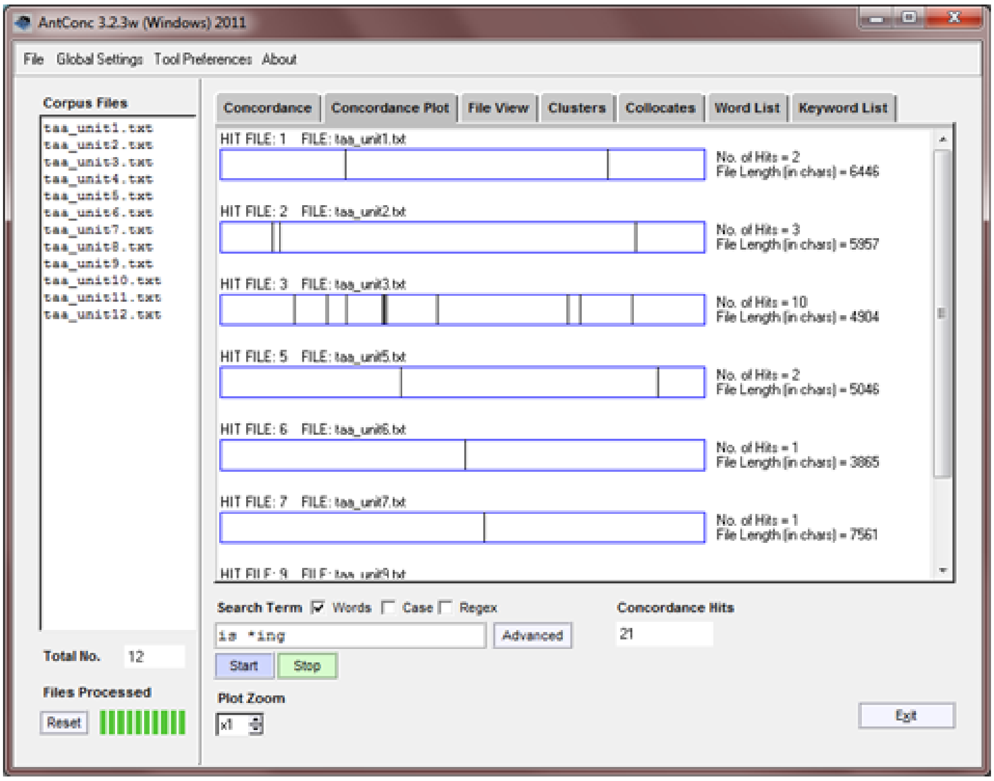
Table 2.3: The tasks presented in *Reading Concordance* [Sinclair, 2003]. Task levels are in increasing order of difficulty.

one, a concordance of “free hand” is analysed in detail to discover the difference between its usage as a literal or metaphorical phrase (the concordance of “free hand” used in the task can be seen in Figure 3.3). The reader must answer twelve questions that focus mainly on understanding the patterns of words surrounding the keyword. These patterns, known as collocation patterns, are analysed by investigating the frequency of the words which appear with a keyword. Collocation patterns are the key to the analysis and understanding presented in the majority of the tasks.

Some advanced tools for the analysis of concordance can be seen in AntConcs’ dispersion plot Figure 2.4 and WordSmiths Tools Pattern windowFigure 2.5. The dispersion plot shows a keywords position in every file of the concordance. The Pattern view displays a frequency-ordered list of words for each position in a concordance.

2.1.3 Genealogies of Knowledge Project

The Genealogies of Knowledge Project (GOK) focuses on translation phenomena and other sites of mediation involving three distinct lingua francas: medieval Arabic, early Latin and modern English. It engages with key historical moments that



Corpus Files

```
taa_unit1.txt
taa_unit2.txt
taa_unit3.txt
taa_unit4.txt
taa_unit5.txt
taa_unit6.txt
taa_unit7.txt
taa_unit8.txt
taa_unit9.txt
taa_unit10.txt
taa_unit11.txt
taa_unit12.txt
```

Total No. 12

Files Processed

```
HIT FILE: 1   FILE: taa_unit1.txt    No. of Hits = 2
                                     File Length (in chars) = 6446
HIT FILE: 2   FILE: taa_unit2.txt    No. of Hits = 3
                                     File Length (in chars) = 5957
HIT FILE: 3   FILE: taa_unit3.txt    No. of Hits = 10
                                     File Length (in chars) = 4904
HIT FILE: 5   FILE: taa_unit5.txt    No. of Hits = 2
                                     File Length (in chars) = 5046
HIT FILE: 6   FILE: taa_unit6.txt    No. of Hits = 1
                                     File Length (in chars) = 3865
HIT FILE: 7   FILE: taa_unit7.txt    No. of Hits = 1
                                     File Length (in chars) = 7561
HIT FILE: 9   FILE: taa_unit9.txt
```

Search Term ☑ Words □ Case □ Regex

is *ing

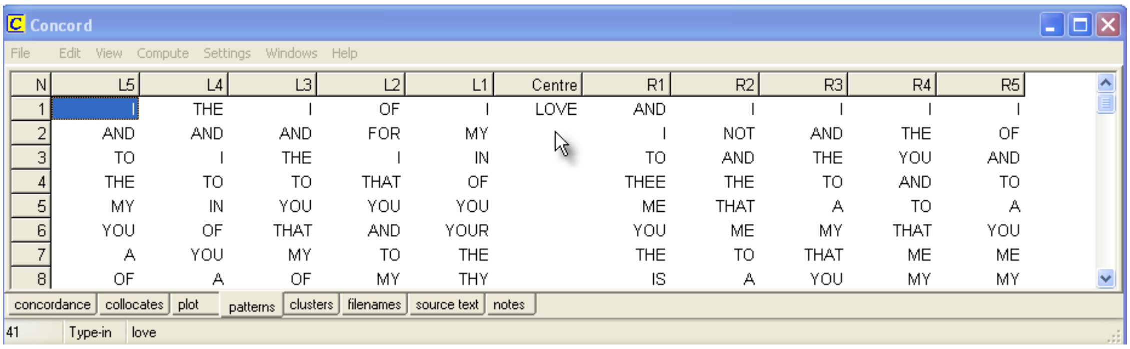
Concordance Hits 21

Plot Zoom x1

Figure 2.4: AntConc Concordance Plot. Useful for identifying keyword position and dispersion within and across the files of a concordance.

have brought about transformations in the interpretation of two constellations of concepts across the last 2500 years. The first constellation relates to the body politic and includes concepts currently expressed by the following lexical items in English: polis, polity, democracy, civil society, citizenship, nation, state, natural law, and human rights. The second constellation consists of concepts that underpin scientific, expert discourse (including medical discourse as a case in point), such as experiment, observation, evidence, proof, episteme, truth, falsehood, aetiology, causation, justification, fact, validity, and expertise.

Collaboration with researchers from this project was the source of the domain-specific knowledge used to guide visualisation design in this thesis. All visualisations created are available as plugins for the GOK Concordance Browser.



N	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	I	THE	I	OF	I	LOVE	AND	I	I	I	I
2	AND	AND	AND	FOR	MY		I	NOT	AND	THE	OF
3	TO	I	THE	I	IN		TO	AND	THE	YOU	AND
4	THE	TO	TO	THAT	OF		THEE	THE	TO	AND	TO
5	MY	IN	YOU	YOU	YOU		ME	THAT	A	TO	A
6	YOU	OF	THAT	AND	YOUR		YOU	ME	MY	THAT	YOU
7	A	YOU	MY	TO	THE		THE	TO	THAT	ME	ME
8	OF	A	OF	MY	THY		IS	A	YOU	MY	MY

41 Type-in love

Figure 2.5: Concordance patterns for the keyword “love” from WordSmith Tool [Scott, 2008].

2.2 Information Visualisation Background

2.2.1 Visual Variables

Jacques Bertin proposed an original set of “retinal variables” (visual variables) in Semiology of Graphics (1967) [Bertin, 1983]:

- Position
- Size
- Shape
- Value (lightness)
- Color hue
- Orientation
- Texture

The list has since been further expanded to include :

- Color saturation
- Arrangement

- Crispness
- Resolution
- Transparency

Each of these variables can be used to encode information on the 2-dimensional plane. Using the original variables, a visual item can be positioned, sized, shaped, and given lightness, colour, orientation and texture without visual interference. Each of these variables could be used to encode a separate piece of information, or multiple variables can be used for the same information to emphasise the visual encoding of that information.

2.2.2 Visual Variable Ranking

In 1984 William S. Cleveland and Robert McGill wrote a foundational paper on graphical perception [Cleveland and McGill, 1985] , in this paper, they discuss many perceptual limitations and advantages of the human visual system. This paper is considered a breakthrough in visualization design theory. The paper established an accuracy ranking of quantitative perceptual visual variables, as shown in the Figure 2.6. Higher visual variables are more accurate than lower variables for quantitative tasks. This accuracy is sometimes referred to as perceptual efficiency in visualisation literature. For quantitative data, the ranking of these tasks was empirically verified by experiment and reported in the paper.

In 1987 Jock D. Mackinlay wrote *Automating the Design of Graphical Presentations of Relational Information.* Mackinlay extended the work of Clevland and McGill by suggesting an ordering of perceptual efficiency of visual variables for Ordinal and Nominal data types. The diagram in Figure 2.7 has been used as a guide for effective visualisation design since its first appearance in Mackinlays work.

In this thesis, I use Mackinlay's ranking to design and compare visualisations. In particular, in section 4.2 and section 4.3, structured encoding comparison relies

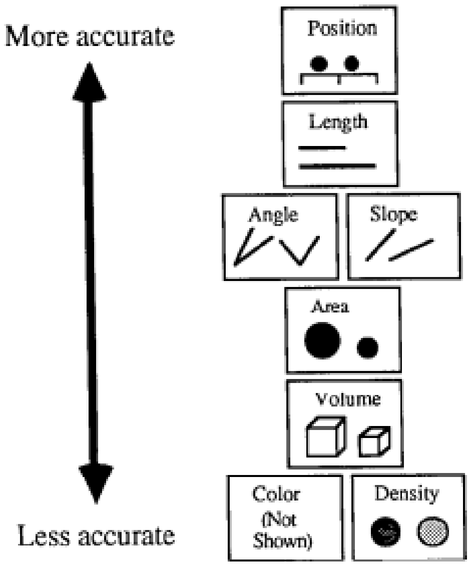


Figure 2.6: Quantative ranking of visual variables for quantitative tasks proposed in [Cleveland and McGill, 1985].

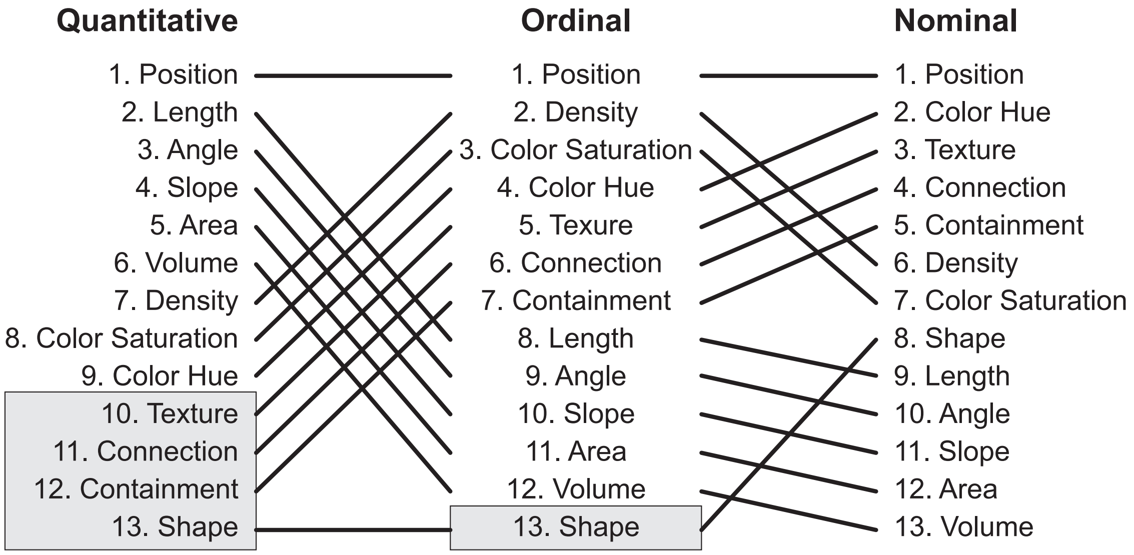


Figure 2.7: Ranking of Visual Variables for Nominal Ordinal and Quantitative Data (adapted from [Mackinlay, 1986]).

on this ranking of visual variables to compare and evaluate the perceptual efficiency of the existing visualisation designs.

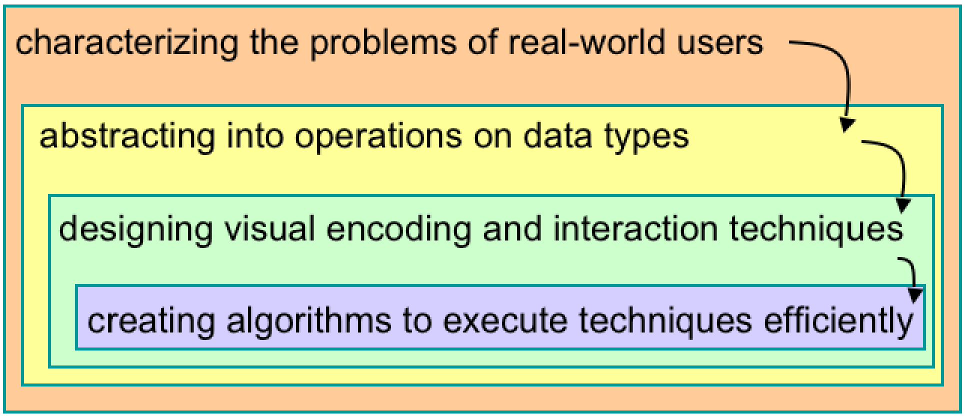

characterizing the problems of real-world users
abstracting into operations on data types
designing visual encoding and interaction techniques
creating algorithms to execute techniques efficiently

Figure 2.8: Nested model of visualisation design.

2.2.3 Nested Model of Visualisation Design

The nested model of visualisation design [Munzner, 2009] gives clear guidelines for visualisation design. The model, as seen in Figure 2.8, splits the design of visualisations into four cascading levels. It argues that a real-world problem must be identified to design an effective visualisation. If that is done correctly, the next stage is creating a data abstraction that correctly models the information required to address the problem from level one. Only after the first two steps are complete can a visualisation be created through visual encoding techniques. The visual encoding technique uses the visual variables, seen in Figure 2.7 and their rankings [Mackinlay, 1986] to map the data abstraction in the second level into a visual representation.

The nested model also offers guidance in evaluating visualisation design. Figure 2.9 shows the possible threats to visualisation validity at each model level. The techniques for validating the visualisation to help mitigate these threats are also supplied. The position of validation steps in the model shows what is required at deeper levels before the current level can be validated using that technique. For example, observing adoption rates requires completing every stage of the nested model before it can be used to validate the domain characterization. Justification of encoding choice is a particularly good validation technique, justification can give strong validation evidence for a visualisation [Ellis and Dix, 2006].

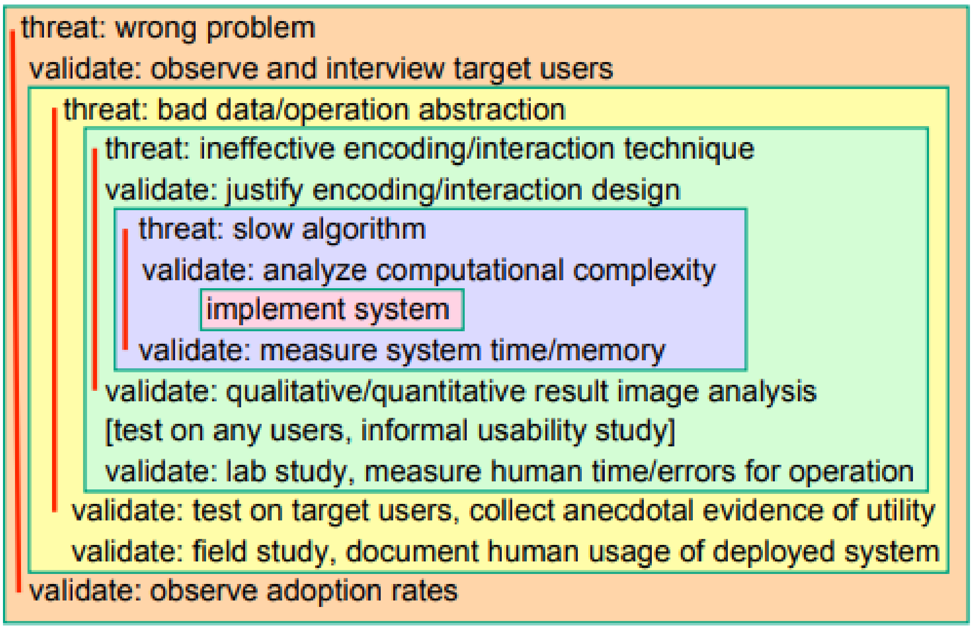

threat: wrong problem
validate: observe and interview target users
threat: bad data/operation abstraction
threat: ineffective encoding/interaction technique
validate: justify encoding/interaction design
threat: slow algorithm
validate: analyze computational complexity
implement system
validate: measure system time/memory
validate: qualitative/quantitative result image analysis
[test on any users, informal usability study]
validate: lab study, measure human time/errors for operation
validate: test on target users, collect anecdotal evidence of utility
validate: field study, document human usage of deployed system
validate: observe adoption rates

Figure 2.9: Nested threat model of visualisation design.

2.2.4 Activity-Centered Network Model

The Activity-centered network model for domain characterization in problem-driven scientific visualisation [Marai, 2018] suggests adding requirement specification analysis to the validation of domain characterization in the nested model. The technique is an extremely detailed process for characterizing a domain based on interaction with experts in a scientific field. It suggests an interview and review of questions the users would like to answer about their data. Several probes are also described to ensure visualisation is valid, such as investigation of related visualisation to ensure the visualisation is necessary.

One technique from this model, which is used in section 3.2, is the 20 questions approach for elicitation of requirements at the domain characterisation stage. This technique asks domain experts to envision their dataset as an entity to which they can ask questions. By asking them to come up with 20 questions to ask the dataset, subquestions and hierarchies often emerge, leading to productive discussions about requirements which go beyond high-level desires and towards detailed specifications.

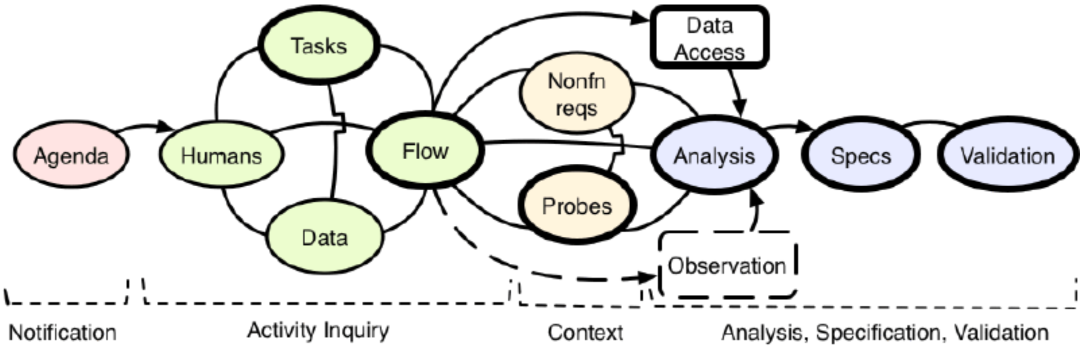


Figure 2.10: Activity-centered network model for domain characterization in problem-driven scientific visualisation. The model has four chronological steps, indicated by colour in this figure. A heavier outline of a node marks it as critical. These critical components must be completed, or the validity of the visualisation design is threatened. Dashed nodes are optional. In this network model, arrows indicate unidirectional flow, while arcs indicate bidirectional flow. The complete specification of the model([Marai, 2018]) provides detailed actions for completing each stage of the model in order to create a set of requirements for a visualisation design.

2.2.5 Information Visualization Reference Model

The information visualization reference model (or data state model)[Chi and Riedl, 1998] is a conceptual framework that enables the concise description of the visualisation construction process. The framework allows the classification of data states into 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. The reference model for the Mosaic visualisation, an interface developed as part of the work described in this thesis, is shown in Figure 5.6.

In 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 some analytical process selects a subset of the data. 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.

Using this model, an extensive review of information visualisation applications has been created [Chi, 2000]. This provides a taxonomy of visualisation techniques and serves to illustrate 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”. [Chi, 2002]

2.3 Conclusion

This chapter provides background knowledge on corpus linguistics and information visualisation. These two topics areas are essential to the interpretation of the contributions of this thesis. In the coming chapters, the material in this chapter will be used to describe the requirements, designs, and evaluations of visualisation contributions.

In the next chapter (chapter 3), efforts at domain characterisation (a technique described in subsection 2.2.3) are presented.

Chapter 3

Domain Characterisation

This section explores the domain of corpus linguistics to identify problems and methods which will benefit from visualisation. Visualisations which address the needs of corpus linguists are much more likely to be effective if those needs are well understood. The inclusion of domain experts in this visualisation design stage is very beneficial. However, talking to users is typically insufficient in forming a full and accurate domain characterization. Expert users help define the domain's high-level goals and tasks and rank the importance of tasks. The characterization can be made more detailed by using methods such as examination of domain literature, contextual studies [Sedlmair et al., 2012], and requirements gathering [Marai, 2018].

By performing a domain characterization, as outlined in the nested model [Munzner, 2009], the methodologies used to achieve the identified goals can be investigated. The aim is to extract the low-level tasks which are performed in the process of working towards the higher-level goals. This analysis can be arranged as a hierarchy of goals, tasks, and low-level actions. The hierarchy can then be used to gain insight into the challenges faced by corpus linguists and how they have been previously addressed.

The characterization presented in this section will be used in later chapters to guide encoding design. Once an adequate characterization of the domain is presented here, visualisation prototyping can be used to explore the identified problems. Before prototyping, further problem-specific domain characterization may be useful for the individual problem. Comparing existing techniques and visualisations for a particular problem specification and establishing expert user familiarity with these techniques can help avoid the main threat to validity at this stage (visualizing the wrong problem). Possibilities Exploration, as described by the activity-centred model Figure 2.10, is the identification of other desirable features that the users did not previously consider. If these desirable features can be identified early in the design stage, there will be less deviation from the functional specification in the final visualisation.

Requirement specification and validation has been shown to be a worthwhile endeavour when creating scientific visualisations [Marai, 2018], the activity-centred model, shown in Figure 2.10, established an approach to domain characterization, which makes used of detailed requirements gathering techniques. The domain characterization was not originally based on the activity-centred model. However, the majority of the techniques described were covered by the analysis. Describing the contribution within this framework helps with creating a user-centric description of corpus analysis, from which functional specifications were extracted. Through a multi-year collaboration with the researchers of the GOK project, the visualisation requirements were established and revised. Project meetings, regular video conferences, and informal discussions about the goals of the project led to the majority of the insights. Where clarification was needed, formal requests for written feedback were made.

3.1 User Expressed Requirements

Initial conversations with domain experts of the GOK project focused on the goals and requirements of the expert users for the initial prototype. Their main requirement was to create visual tools that support the analysis of the unique corpus created by the GOK project. General corpus linguistic workflow should be supported to enhance the analysis of this unique data source. In addition, the project seeks to establish new quantitative methodologies for the analysis of corpora. The tools should support analysis of the corpus's temporal aspect and multilingual nature. The visualisation tools must be created as plugins for the GOK corpus browser.

The GOK software was adapted from the TEC corpus browser with an established user base of several thousand unique users. The target users are corpus linguists in general and translation studies researchers as a specific subgroup. The typical methodology of these users is based on concordance analysis in the style of *John Sinclair* [Sinclair, 2003]

3.2 Tasks for Corpus Analysis

Two researchers from the GOK project (who will be anonymously named Daisy and Dave) were requested to list 20 questions which they would like to be able to answer about a corpus. This 20-question technique for requirements elicitation is established in the activity-centred network model [Marai, 2018], as described in subsection 2.2.4.

The researchers were asked to list the questions them in order of importance where possible and made themselves available to discuss the lists. The request for the lists was made a week before the meeting to discuss the results. Possible methods for answering the questions were discussed with the researchers, and the proposed methods were included with the questions.

3.2.1 Daisys questions to ask a corpus

Daisy created the list with a very loose ranking system. The categories and questions are in order of when they were thought of; she suggests this may be implicitly correlated with question importance. The three categories she identified were keywords, collocational patterns and temporal spread.

3.2.1.1 Keywords

Pros

1. How many times is the chosen keyword used across all of the corpus texts as a whole?

   • Method: Frequency list

2. Is the keyword used with more or less uniform frequency in each of the corpus texts individually, or are there significant imbalances in the dispersion of the keyword?

   • Method: Look at filenames associated with the concordance lines. This can be done in the GOK concordance browser.

3. Which specific corpus texts use a given keyword proportionally greater frequency? And lesser? What patterns can we see if we rank the corpus texts by the number of hits for this keyword?

   • Method: Look at the filename for the concordance lines and compile a spreadsheet

4. Which linguistic-grammatical form(s) of the term/concept under investigation (e.g. singular vs plural, forms suffixed with -ship, -like or -ly) is/are more common across the corpus as a whole?

   • Method: Use the concordance browser to search terms and record the number of lines returned

5. To what extent are the relative proportions of these different word forms the same or different within each corpus text?

   • Method: Sub-corpus selection, look at the concordance lines, and compile a spreadsheet containing each of the terms and their frequencies per text.

6. Are there other related keywords we might study in order to expand our investigation? Can the software suggest important keywords to these texts that we might not have considered?

   • Method: Currently, this is performed by the expert suggesting her own alternative keywords.

7. If so, are the frequencies of these terms similar or different to the first keyword, both across all of the corpus texts as a whole and in each corpus text individually?

   • Method: Search these terms and analyze the number of concordance lines and file names associated with them. A spreadsheet is often useful.

3.2.1.2 Collocational patterns

"N.B. These questions assume the keyword is a noun. From my experience so far, I don't believe much would change in terms of the tools required for analysis if the keyword were an adjective or verb etc... This is because we would still be interested in finding patterns in the words that appear in close proximity to the keyword."

1. What are the adjectives that most commonly modify the chosen keyword (LEFT +1) across all of the corpus texts as a whole?

   • Method: Sort concordance browser and visually scan while scrolling.

2. What adjectives most commonly modify the chosen keyword (LEFT +1) in each text?

   • Method: Sub-corpus selection and estimate frequency at a position in concordance.

3. Are there any adjectives that modify the chosen keyword significantly more frequently in one text when compared with the others?

   • Method: Investigate the results of question 3, possibly using a spreadsheet

4. Are these adjectives only used to describe this keyword, or are they connected to other keywords in this text?

   • Method: Calculate some measure of collocation strength between the identified context words and the keyword.

5. What verbs are most commonly associated with the keyword (normally, RIGHT +1, RIGHT +2) across all of the corpus texts as a whole?

   • Method: Concordance analysis investigating frequency at several positions

6. What verbs are most commonly associated with the keyword (RIGHT +1, RIGHT +2) in each of the corpus texts individually?

   • Method: Sub-corpus selection or look at concordance sorted on filenames for short concordance lists. Followed by an analysis of frequency at the desired positions

7. Are there patterns of interest in any of the other word positions relative to the keyword? For example, if the keyword is a label used to describe a particular kind of political agent, we might be interested in examining what collective nouns are used to group and characterize these political agents (e.g. a mob of citizens, a tribe of politicians: LEFT +2).

   • Method: Try to get a global overview of context word frequencies by repeatedly sorting different positions and trying to identify frequent patterns.

3.2.1.3 Temporal spread

"These questions assume the keyword is a noun. From my experience so far, I don't believe much would change in terms of the tools required for analysis if the keyword were an adjective or verb etc... This is because we would still be interested in finding patterns in the words that appear in close proximity to the keyword."

1. In what ways do these patterns correspond to the temporal spread of these texts (i.e., given that some of these texts were published in 1850, others in 2012)?

   • Method: Select temporal sub-corpora and perform keyword and collocation pattern analysis

2. Is a particular keyword more frequent in one time period or another (e.g. within a specific year, decade, or longer historical period, e.g. the Victorian era, post-1945, pre-1989, etc.)?

   • Method: Split corpus into temporal sub-corpora and examine keyword frequency. This may be very time-consuming as the number of temporal sub-corpora (slices) could be large.

3. Are there time periods when the keyword does not feature?

   • Method: Answering question 2 should suffice

4. Can certain adjectives/nouns/verbs be found to collocate more frequently with the keyword (in a particular word-position) in those corpus texts produced within one time period versus those produced earlier or later?

   • Method: Comparison of collocation patterns in these sub-corpora (temporal splits). Could be very slow.

5. Are the changes in the relative frequency of a keyword over time similar or different to the patterns observed with regard to other keywords?

   • Method: Temporal keyword analysis (question 3) for multiple keywords.

6. To what extent can these patterns be explained by other factors (especially those pertaining to the construction of the corpus itself, e.g. the uneven distribution of tokens across the corpus as a whole)?

- Method: Expert analysis of the results in the context of domain knowledge and understanding of the limitations of the data source (the corpus).

3.2.2 Daves questions to ask a corpus

Dave decided to split the questions into four categories Keyword, Text, Author and Corpus. The order of the sections is by importance. Within each section, the questions are also ranked by importance. Dave was keen to point out that the questions from the sections will often intertwine, and the importance ranking is only an approximation.

3.2.2.1 Keyword

"i.e. in case one wants to study a specific keyword in any number of texts"

1. How frequent is the keyword, and where is it ranked in a frequency list?

   - Method: Question one can be answered by consulting a frequency list and looking at the number of concordance lines returned for a search. This is possible in most concordance browsers, including the GOK corpus browser.

2. With which words is the keyword most frequently combined, in a span of 4 positions to the left and right?

   - Method: The second question requires an investigation of word frequencies at positions relative to the keyword. This can be done using the concordance browser but is time-consuming for large concordance lists.

3. What is the approximate strength of the collocational patterns observed?

   - Method: Question three could be answered by calculating a statistical association between the keyword and a context word. One measure used in the field is Mutual Information.

4. Are there intuitive variations of the keyword (both formally and semantically) that occupy similar positions and display similar collocational patterns?

   • Method: Question four is answered by the users understanding of the keywords meaning and linguistic properties.

5. Which position does the keyword take in the clause, the sentence, and the text?

   • Method: Question five would require information from the individual text in the form of an extract or full text.

3.2.2.2 Text

"i.e. in case one wants to uncover the properties of a certain text"

1. What are the most frequent content words in the text, and how does this compare to other texts of a similar character?

   • Method: The first question could be answered by comparing frequency lists for texts. The exact method to achieve representative comparison is unclear.

2. What are the most frequent function words and connective elements in the texts, and with which of the content words above do they recurrently combine?

   • Method: The first part of the second question could be answered in a similar manner to question one. The second part would require some other technique, such as concordance analysis, by searching and analyzing the contexts of the high-frequency keywords of interest.

3. What are the most common proper names in the text?

   • Method: Difficult to answer without the tagging of these names either manually or automatically

4. Having identified all the above, do they vary in their dispersion across the document?

- Method: Extracting the full text and analyzing keyword position or dispersion analysis tool such as AntConcs Concordance Plot Figure 2.4

5. Having established all the above, where are (dis-)continuities situated in the text? (For instance, does the introduction display a different textual character than the body of the text)

   - Method: Similar to question four.

3.2.2.3 Author

“i.e. in case one wants to construct a profile for an author with multiple texts in the corpus”

1. Which words are the most frequent in each individual text written by the author in question, and how does this compare to the overall frequency of words in all the author’s texts combined?

   - Method: Use sub-corpus selection to investigate frequency lists for individual authors or files.

2. Which words does the author use significantly often in comparison to other authors similar in a temporal, spatial, linguistic, or social context?

   - Method: Use domain knowledge to split the corpus into relevant sub-corpora and compare frequency lists

3. Who does the author frequently cite?

   - Method: Citations would need to be tagged or automatically identified in some way for this to be feasible for large sub-corpora

4. Which multi-word expressions occur significantly often?

   - Method: Identify multi-word expressions for sub-corpora. Concordance analysis of frequent terms Nlp techniques exist for extraction of candidate multi-word expressions.

5. Given all the above, are there temporal changes to be observed in the author's textual profile?

   • Method: Expert interpretation of the results

3.2.2.4 Corpus

"i.e. in case one wants to interrogate a corpus varied in textual material"

1. What are the corpus's most frequent words, collocations, and other multi-word expressions?

   • Method: Frequency lists, concordance analysis, automatic methods.

2. Can the frequency of the above be attributed to a limited number of texts, or is it characteristic of the corpus as a whole?

   • Method: Search the keywords and regular expressions of interest. Sort by filename and estimate the frequency of each lexical item of interest per file.

3. If the texts in the corpus display varied patterns regarding the above, how are relevant keywords, collocations, and multi-word expressions distributed across the corpus in terms of the publication date, source language, author, etc?

   • Method: This is difficult to do as meta-data for each file has to be investigated individually. The best method is to perform sub-corpus selections for the meta-data attributes of interest and perform concordance and keyword analysis.

4. What can one say about the specificity of the corpus in question compared to another specific corpus? (For instance, do certain keywords occur very often in all texts studied while being very low-frequency in another varied corpus set)

   • Method: Comparison of keyword lists for corpora and concordance analysis based on any discovered differences.

5. Are the texts in the corpus explicitly connected through quotation or other types of reference?
    - Method: Tagging required to identify references

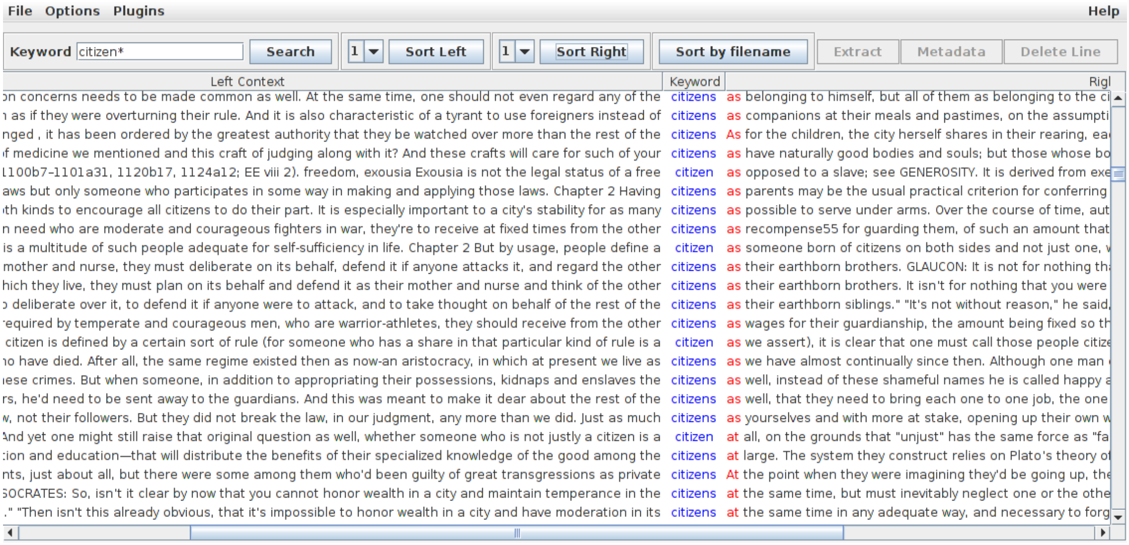
3.3 Example Methodologies

Our collaborators in the GOK project put a great deal of effort into corpus creation; prior to the collection of the full corpus, these language scholars could not perform full and rigorous scholarly analysis. However, preliminary studies and exploratory hypothesis generation/testing began as the corpus matured during collection. At this stage of the project, it became possible to investigate the practical requirements of these expert users in relation to the GOK project goals.

3.3.1 Daisys example methodology

A member of the GOK project gave a brief presentation outlining an example methodology and its challenges (the researcher was previously anonymously referred to as Daisy) to help with the initial definition of visualisation goals for the project. The methodology was explained in the form of a case study. The case study made use of the portion of the GOK English corpus, which was available at the time. The task was defined as comparing the patterning around the keyword "*citizen**". The * represents a regular expression search for continuations of the word citizen, such as citizens and citizenship. The patterns identified are compared across two large sub-corpora.

- **Sub-corpus 1** A sub-corpus of modern English translations from Classical Greek (1850 onwards);
- **Sub-corpus 2** A sub-corpus of translated and non-translated texts written by contemporary authors, published between 1992 and the present day.



Left Context	Keyword	Rigl
on concerns needs to be made common as well. At the same time, one should not even regard any of the	citizens	as belonging to himself, but all of them as belonging to the ci
n as if they were overturning their rule. And it is also characteristic of a tyrant to use foreigners instead of	citizens	as companions at their meals and pastimes, on the assumpti
nged , it has been ordered by the greatest authority that they be watched over more than the rest of the	citizens	As for the children, the city herself shares in their rearing, ea
f medicine we mentioned and this craft of judging along with it? And these crafts will care for such of your	citizens	as have naturally good bodies and souls; but those whose bo
1100b7–1101a31, 1120b17, 1124a12; EE viii 2). freedom, exousia Exousia is not the legal status of a free	citizen	as opposed to a slave; see GENEROSITY. It is derived from exe
aws but only someone who participates in some way in making and applying those laws. Chapter 2 Having	citizens	as parents may be the usual practical criterion for conferring
th kinds to encourage all citizens to do their part. It is especially important to a city's stability for as many	citizens	as possible to serve under arms. Over the course of time, aut
n need who are moderate and courageous fighters in war, they're to receive at fixed times from the other	citizens	as recompense55 for guarding them, of such an amount that
is a multitude of such people adequate for self-sufficiency in life. Chapter 2 But by usage, people define a	citizen	as someone born of citizens on both sides and not just one, w
mother and nurse, they must deliberate on its behalf, defend it if anyone attacks it, and regard the other	citizens	as their earthborn brothers. GLAUCON: It is not for nothing tha
hich they live, they must plan on its behalf and defend it as their mother and nurse and think of the other	citizens	as their earthborn brothers. It isn't for nothing that you were
o deliberate over it, to defend it if anyone were to attack, and to take thought on behalf of the rest of the	citizens	as their earthborn siblings." "It's not without reason," he said,
required by temperate and courageous men, who are warrior-athletes, they should receive from the other	citizens	as wages for their guardianship, the amount being fixed so th
citizen is defined by a certain sort of rule (for someone who has a share in that particular kind of rule is a	citizen	as we assert), it is clear that one must call those people citize
no have died. After all, the same regime existed then as now-an aristocracy, in which at present we live as	citizens	as we have almost continually since then. Although one man
ese crimes. But when someone, in addition to appropriating their possessions, kidnaps and enslaves the	citizens	as well, instead of these shameful names he is called happy a
rs, he'd need to be sent away to the guardians. And this was meant to make it dear about the rest of the	citizens	as well, that they need to bring each one to one job, the one
w, not their followers. But they did not break the law, in our judgment, any more than we did. Just as much	citizens	as yourselves and with more at stake, opening up their own w
And yet one might still raise that original question as well, whether someone who is not justly a citizen is a	citizen	at all, on the grounds that "unjust" has the same force as "fa
tion and education—that will distribute the benefits of their specialized knowledge of the good among the	citizens	at large. The system they construct relies on Plato's theory of
nts, just about all, but there were some among them who'd been guilty of great transgressions as private	citizens	At the point when they were imagining they'd be going up, the
SOCRATES: So, isn't it clear by now that you cannot honor wealth in a city and maintain temperance in the	citizens	at the same time, but must inevitably neglect one or the othe
." "Then isn't this already obvious, that it's impossible to honor wealth in a city and have moderation in its	citizens	at the same time in any adequate way, and necessary to forg

Figure 3.1: Visualisation proposed by GOK researcher.

The method itself consisted of two techniques. The first technique identifies explicit definitions of ’citizenship’ within each sub-corpus. The researcher wants to compile a list of frequently used verbs and prepositions at position ’keyword+1’ to find these definitions. To achieve this, the GOK corpus browser is used. Sub-corpus 1 was selected using the sub-corpus selection tool, the regular expression ’citizen*’ was searched, and the concordance was sorted at position ’keyword +1. The researcher then spends time scrolling through the concordance and compiling a list of relevant frequent words at the position of interest, Figure 3.1 shows the concordance window sorted and scrolled to the preposition *as*. With this list in hand, more accurate searches can be run, such as:

- citizenship+“(is|as|was|defined|conceived|are|equals |considered|appears|means)”
- citizenship+“(has|should|must|will|may)”
- citizen+“(is|as)”
- citizens+“(are|as)”

Definitions can be extracted by reading the concordance lines generated by these new searches. Some examples of the definitions found are:

- Citizenship is a status bestowed on those who are full members of a community.
- As well as enjoying rights, citizens are required to undertake responsibilities such as paying taxes and jury or military service.
- citizenship should be based purely on residency
- US citizenship has represented a safe haven from oppressive regimes around the world

The second technique is the observation of patterns in the adjectives used to modify "citizenship" and constructions such as "citizens+of+*". The researcher explained that this technique is more difficult and time-consuming than a concordance browser. To quote the researcher.

> "Specifically, it is difficult to get a quick overview of such patternings using the concordance, given that the number of lines returned for my searches are quite large: e.g. 4420 hits for "citizen*" in my subcorpus of translations from Classical Greek."

The researcher had some experience with linguistic visualisation and, in the past, had used word clouds, such as Wordle [Viegas et al., 2009] to present research results. To use word clouds for the methodology, there are some challenges to overcome. First, the researcher noticed that stop-words dominate the frequency distributions of the word positions, so some technique has to be applied to get meaningful results. The suggested technique was to use a stop-word list to filter the visualisation. The concordance would need to be processed to extract the words at particular positions for visualisation since the concordance is structured as a list of aligned text extracts. The researcher's reasoning resulted in an interface for displaying positional word clouds with the option to exclude stop words. The presentation included a mock-up of what a visualisation to solve this problem would look like Figure 3.2. The mock-up displays a word cloud for either a full concordance or a chosen word position

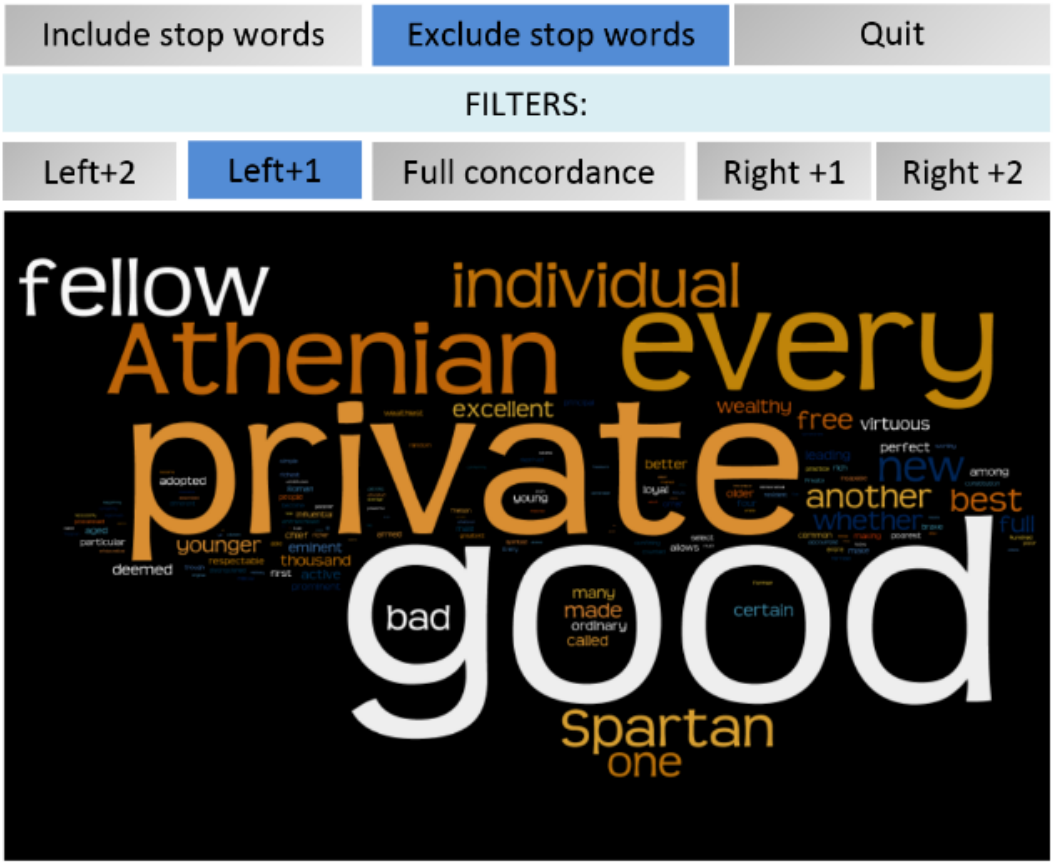


Figure 3.2: Visualisation proposed by GOK researcher.

and has the option to remove stop words. Looking at the mockup in Figure 3.2, the words modifying citizen are presented to emphasise frequency and provide an overview of a position relative to the keyword on a single screen.

The idea and its feasibility were discussed at the end of the presentation, and some questions were asked to clarify the methodology. Notes were taken during this discussion and later discussed with Daisy. Daisy then prepared written answers to the core questions arising from the discussion session. The questions as Daisy interpreted them and her responses follow:

- What is the domain in which the case study is situated?

> “Translation and Reception studies. How we have received classic Greek texts. How has translation shaped this reception? The role of translation is often overlooked.”

- Is this methodology (excluding the proposed visualisation) typical of the field?

> “Translation Studies as a discipline tends to encourage close qualitative analysis of a small selection of examples chosen from specific texts to illustrate a particular argument.
>
> Corpus analysis enables the translation scholar to identify and investigate with significantly greater ease differences between and patterns within translations, taking into account the full length of each work as a complete text.
>
> Corpus analysis has been extensively used in translation studies before (e.g. within the TEC project and many others), but the field has tended to focus mainly on more micro-level linguistic concerns rather than the socio-political implications of translators’ word choices etc.”

- How did the idea for this example arise?

> “Gok seeks understand the constellation of concepts related to the body politc across time and space. Citizenship is a lexical item in that constellation. Comparing meaning, frequency and usage of related terms is an exploratory process used to discover obvious patterns”

This final answer from Daisy provides a succinct description of the requirements of Dasiy as a target user; she wishes to compare meaning, frequency and usage of related terms (keywords) in an exploratory process where the identification of patterns is the goal.

3.3.2 Daves example methodology

The second methodological example gives an overview of the general technique employed by the researcher and is not based on a specific case study. The researcher,previously referred to as Dave, presented this methodology as a slideshow with a follow-up question and answer session.

The methodology was described as the search for the largest "unit of meaning" related to a keyword in the Genealogies of Knowledge corpus. Meaning, in this case, should be constructed from the evidence present in the corpus. The corpus is central to the analysis, and the technique is in the style of *John Sinclair* [Sinclair, 2003]. In this style, the analyst investigates collocation, colligation, semantic preference and semantic prosody. The analyst must also make an effort to ignore personal bias and experience with the texts under investigation. A summary of the steps used to perform the analysis follows:

Construction of Meaning:

- Sample
- Describe Patterns
- Sample
- Compare Patterns
- Hypothesis

In the steps, the term "Sample" refers to a sub-corpus selection and keyword search in a concordance browser; if the concordance is large, the samples may be thought of as subsets of the full concordance list.

"Describe Patterns" refers to analyzing the positional frequencies around the keyword. Dave explained that the analysis begins by looking at the patterns of words occurring beside the keyword and expands to additional positions until the discovered patterns describe the meaning of most of the concordance lines. The remaining lines would also be analyzed after the core units of meaning had been established.

"Compare Patterns" examines the differences and similarities between the "described patterns" of the samples.

The following clarifying question was posed to Dave:

- Can you give practical details of your typical methodological approach?

> "Investigate the keyword and its neighbouring collocates (Left and right +1) Investigate deviations from frequent patterns, then expand the analysis horizon and repeat until the largest unit of meaning is found. The largest unit of meaning should be read as 'overarching' in the sense that the point is not to necessarily go beyond the concordance line but to construct an abstract unit that can account for as many concordance lines as possible. If interesting patterns which lead to a hypothesis are discovered, pursue these. Typical corpus linguistic method applied to unique corpus."

Dave reported some difficulties. The unique nature of the corpus makes the generalization and the representativeness of a hypothesis more difficult to explain. Viewing meta-data for the concordance lines is useful, but as it is line specific, it is impractical for analyzing large numbers of concordance lines. The concordance browsers used for identifying patterns in large numbers of concordance lines require sampling multiple times. A broader picture of the concordance, which can examine broader and restricted contests, would be desirable. Any visualisation needs to integrate into the research flow.

Daves' statement that the unique nature of the corpus caused problems required clarification, and the expert group agreed these problems are also some of the corpus linguistics research opportunities. The following question was posed to Dave in relation to this issue:

- How and why does the corpus influence the methodology?

> " If speaking about 'typical/traditional' corpus linguistics (which is always a bit of a stretch), one finds actual/practical lexicography, and analysis of register/text type, etc., drawing on corpora that are con-

> structed to serve as a sample of the full language or sub-language under investigation. Think of the British National Corpus, for example.
>
> Our corpus hosts a variety of texts, but it would be difficult to make the case that it is representative of anything outside the corpus itself. Our Internet corpus, for instance, is not a sample that can tell something meaningful about the Internet 'as a whole'. Therefore, rather than making exhaustive analyses of a certain word across the corpus or tracing a grammatical pattern across its contents, the corpus urges one to study a specific subset of texts and to complement the findings with sources outside of its confines to make hypotheses about conceptual developments. Consequently, the method will be less repetitive than one would traditionally see and more meandering, to a certain extent. "

Dave described his current research as focused entirely on the GOK Internet corpus. The concepts he investigates are community, politics, and democracy. Dave was asked to clarify the domain in which he would situate his research:

- How would you describe your research area?

> "In principle, the field is corpus-based translation studies. As the translational component is fairly minimal in my research, and as I focus a lot on linguistic categorizations and interactions, with the aim of having something meaningful to say about societal developments. You could define the field as Discourse, which in this sense, can be read as language as social practice determined by social structures. "

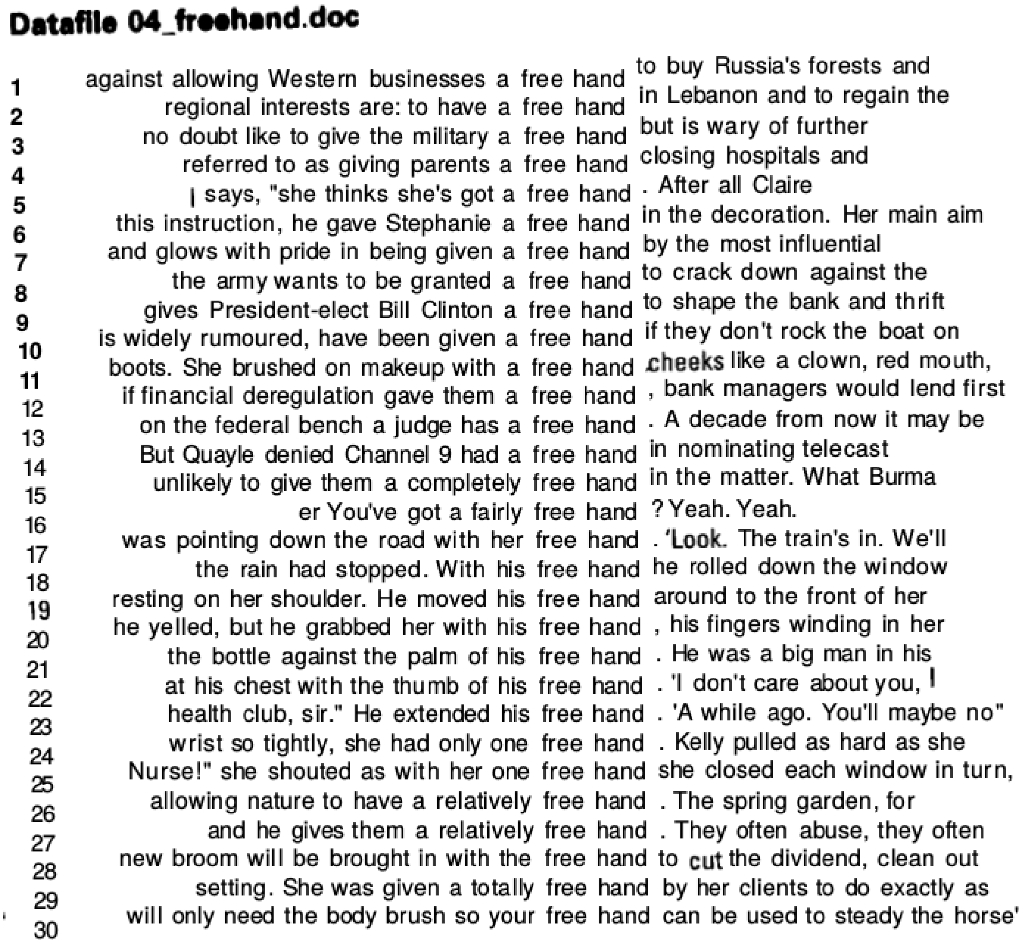
3.4 Domain Literature-Based Task Analysis

Consultation and collaboration with the language scholars of the GOK project who interrogate corpora as an essential part of their analytical work, corpora lead to the natural discussion of visual tools to support analysis. These collaborations revealed how integral the KWIC-based concordance display is the work of the text analyst. These visual representations provide an essential view of the keyword's context. However, examining the relative frequencies of the words which surround the keyword is also a commonly performed task using these tools, for which it would appear these tools are not well suited. In practice, the analyst usually complements the textual information provided by the KWIC display with lists of words sorted by frequency of occurrence in the subcorpus under examination, as well as other statistics. Different processes and sub-tasks mediate the analysis as a whole.

To study this type of concordance analysis in a practical context we turned to a reference work entitled *Reading Concordances: An Introduction* [Sinclair, 2003]. This book is intended as a tutorial on how to look for certain linguistic properties of a keyword (such as word sense, phrasal usage, part of speech and many others) using a KWIC concordance list. The reader is invited to perform eighteen tasks Table 2.3 which introduce the key practical actions and usage of linguistic knowledge required to make decisions about the properties of a word or collocation. For each of these tasks, I performed a hierarchical task analysis by combining or splitting the steps into a series of actions and sub-actions.

Each of the eighteen tasks was analysed and tagged to assist with the classifying and counting of the actions and sub-actions. Before explaining the exact meaning of the tags, an example of the tagging procedure for task 4 is given. This tagging procedure can allow a visualisation researcher with limited knowledge of the problem domain to extract meaningful actions.

Task 4 is concerned with identifying literal and metaphorical usage phrases. The preamble to the task provides some linguistic insight explaining that "some idiomatic phrases in English are recognizable because they contain a word which is not found anywhere else, like *at loggerheads*". They may also be recognizable because the literal meaning is absurd. But others are more subtle and don't have



Datafile 04_freehand.doc

1 against allowing Western businesses a free hand to buy Russia's forests and
2 regional interests are: to have a free hand in Lebanon and to regain the
3 no doubt like to give the military a free hand but is wary of further
4 referred to as giving parents a free hand closing hospitals and
5 | says, "she thinks she's got a free hand . After all Claire
6 this instruction, he gave Stephanie a free hand in the decoration. Her main aim
7 and glows with pride in being given a free hand by the most influential
8 the army wants to be granted a free hand to crack down against the
9 gives President-elect Bill Clinton a free hand to shape the bank and thrift
10 is widely rumoured, have been given a free hand if they don't rock the boat on
11 boots. She brushed on makeup with a free hand cheeks like a clown, red mouth,
12 if financial deregulation gave them a free hand , bank managers would lend first
13 on the federal bench a judge has a free hand . A decade from now it may be
14 But Quayle denied Channel 9 had a free hand in nominating telecast
15 unlikely to give them a completely free hand in the matter. What Burma
16 er You've got a fairly free hand ? Yeah. Yeah.
17 was pointing down the road with her free hand . 'Look. The train's in. We'll
18 the rain had stopped. With his free hand he rolled down the window
19 resting on her shoulder. He moved his free hand around to the front of her
20 he yelled, but he grabbed her with his free hand , his fingers winding in her
21 the bottle against the palm of his free hand . He was a big man in his
22 at his chest with the thumb of his free hand . 'I don't care about you, I
23 health club, sir." He extended his free hand . 'A while ago. You'll maybe no"
24 wrist so tightly, she had only one free hand . Kelly pulled as hard as she
25 Nurse!" she shouted as with her one free hand she closed each window in turn,
26 allowing nature to have a relatively free hand . The spring garden, for
27 and he gives them a relatively free hand . They often abuse, they often
28 new broom will be brought in with the free hand to cut the dividend, clean out
29 setting. She was given a totally free hand by her clients to do exactly as
30 will only need the body brush so your free hand can be used to steady the horse'

Figure 3.3: Concordance of free hand used in Task 4 of *Reading Concordance* [Sinclair, 2003].

the aforementioned identifying marks. As an example, the phrase *he got cold feet* seems to be a literal way of saying that his feet are cold. How do we, as readers, know when it means he is cowardly? The task studies the example of the phrase "free hand". A concordance of 30 lines is provided, and a set of twelve directions on how to analyse the concordance are given to the reader. An answer key is also provided, which expands on the analysis and the insights that can be gained.

The first direction tells the reader to at the position directly to the left of the phase, which have been sorted alphabetically "and list them in order of frequency. Can you associate any of the SINGLETONS with any of those that recur?"

I tagged this action with the *frequency* tag, *word position* tag, *group* tag and *expert decision* tag. The key gives a breakdown of the words at the position and notes that “her, your” are in the same word class as “his” and that “completely, fairly, and totally” are in the same word class as ”relatively”.

Step two asks the reader to

> “Look again at the five lines where N—1 is an adverb of degree. What is the word at N—2? Then consider the two lines where N—1 is one. What is the word at N—2? Can you associate these seven lines with the two big groups of a and his . . . ?”

The positional notation N-2 means the set of words two positions to the left of the keyword. The same tags are applied to this action as word position, exact frequency counts and linguist knowledge are used. The answer key states

> “Where N - 1 is an adverb of degree, N—2 is a; so these five lines join the group of the indefinite article. Where N—1 is the word one, in no. 25 N - 2 is her and so this line joins those with possessive adjectives. The other one, no. 24, has only at N - 2 , which is unlike all the other lines in this sample, so we will fit it in later on.”

Step three starts by explicating that in the previous step, 28 of the 30 lines were extracted and divided into two groups based on the “choice of determiner in front of the noun hand”. The reader is then told “here the difference is not just the type of determiner; consider the meaning of free hand in the two types of line and comment on the distinction in meaning.”. This task is tagged with *Similar Meaning*, *expert decision* and *read context.* For this example, the meanings of the keyword must be analysed by reading the contexts and using linguist knowledge to compare the meanings. The answer key explains that a possessive adjective is the determiner the word “free” means “available” and the word “hand” is a part of the human body. When the determiner is a phrase “a free hand” means “an unrestricted opportunity”.

Skipping forward to step seven, the reader is narrowing in on the linguistic patterns used to determine the literal or metaphorical usage of the phrase free hand. The reader is asked to group concordance lines according to whether the verb is active or passive and to examine if this accounts for the use of the word "given" exclusively before "a free hand". Tags *group, read context, and expert decision* all apply. Step 8 then combines the previous analysis to describe an algorithm for determining metaphorical or figurative usage of the phrase "free hand". Many of the lines which have been discarded as not matching any patterns are not included in the construction of the algorithm.

Condition 1 of the algorithm is that there is a form of the word "give" or a similar meaning word to the left of the phrase. If not, there is an occurrence of the verb "have" or "get", or one with a similar meaning and use?

Condition 2 is that the indefinite article precedes the core phrase, either directly or with only an adverb of degree in between.

If both conditions hold, the phrase "free hand" means "to be set a task without restrictions on resources or methods to accomplish it."

Steps nine to twelve examine all lines that had not previously been examined in the concordance. The word frequencies and patterns to the right of the keyword are analyzed and used to help account for the lines which the left context analysis could not explain.

This example should help clarify how the tags were assigned to the individual steps of the tasks. There was significant variation across the tasks, but the core actions could be described with a relatively small set of tags.

The actions and sub-actions are used to generalize the descriptive analysis steps in Sinclair's tasks into operations common to many tasks. Taking an overview of the tags that describe the tasks and the common action chains they form, we created the hierarchy shown in Figure 3.4.

The primary actions (second level) are split into quantitative and qualitative

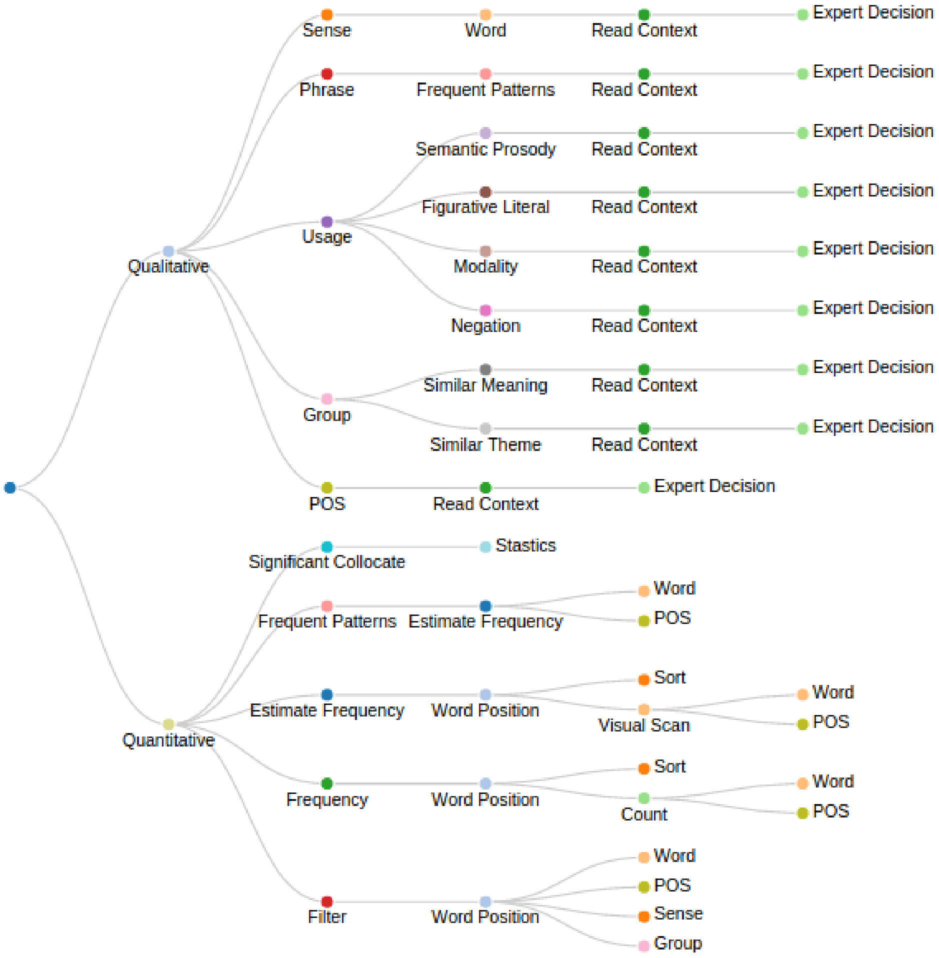


Figure 3.4: Hierarchical visualisation of Concordance Based Corpus Analysis Actions.

groups at the first level of the hierarchy. Qualitative actions are classified on the criteria that a decision that would be possible for experts to disagree on needs to be made to complete the action. These experts could be human users or algorithmic classification processes. Quantitative actions may form a part of a qualitative action; for example, frequent patterns must be identified before they can be classified as phrasal or non-phrasal usage [Sinclair, 1991].

Quantitative actions are those in which the steps involved in the action can be clearly stated, and given the classifications have already been made, the results will be the same when performed by a reliable analyst. For example, word frequencies at a specific word position can be accurately and repeatably determined. Quantitative actions often make use of the results of qualitative actions, such as estimating the frequency of words to the left of a meaning group where the group has to be first identified by expert decision.

The second level of the hierarchy contains the primary actions. These are the actions which most often describe the spirit of the instructions given in the eighteen tasks. Deeper into the hierarchy the sub-actions required to perform these primary actions are presented.

At the third level of the hierarchy, the *area of analysis* is displayed; this is the level at which the primary action is performed. Looking first at the quantitative actions, I found that in three primary actions (filter, frequency and estimate frequency), a word position relative to the keyword is the area at which the actions are applied. A fourth quantitative action, frequent patterns, has an area of analysis, estimate frequency, which is one of the other primary actions. This means the action is performed on a collection of results from the *estimate frequency* actions, i.e. the analysis is performed on frequency estimations across word positions. It is worth noting that in four of the five quantitative tasks, identified word position or multiple word positions is the area at which the action is performed. The final action identified, *significant collocates*, uses the results of statistical analysis of the keyword and its context from the corpus under investigation. This analysis is usually undertaken as a separate analysis, with its results reported as a list of frequent collocations with a keyword.

Turning to the qualitative actions and, again, looking at the area of analysis at level three, the analysis always occurs at the sentence level, which is implied by the read context action. This contrasts with quantitative actions, where positions are the most common analysis area. For qualitative actions, it appears the horizontal structure of the KWIC list is emphasised while the qualitative makes better use of the vertical alignment. Each of the actions requires an expert (or algorithm) who evaluates the context of individual occurrences of the keyword and makes a classification decision based on the semantic and syntactic content of the concordance

Table 3.1: Action counts from task analysis. Total numbers of actions found in the 18 tasks and number of the 18 tasks which feature the action.

Tag	No. of tasks in which action appears	Total action appearances
expert decision	18	60
estimate frequency	16	34
read context	16	31
frequent patterns	15	21
frequency	14	18
word position	13	24
POS: Part of speech	11	23
filter	11	18
sense	10	19
group	7	9
significant collocate	5	7
usage	5	6
phrase	5	6

line. This *Expert Decision* can often be the result of a combination of reading the individual contexts (the linear structure of the text) and performing some of the quantitative actions (positional statistics of the text). In essence, the *Expert Decision* action encapsulates the process of using linguistic knowledge extracted by other primary actions to answer questions about the keyword.

While most tags represent actions, a few additional tags were chosen to help clarify and add information about the tasks and sub-tasks. The tags *word*, *semantic prosody*, *Similar Meaning* and others are not themselves actions but are useful in clarifying the objective or operation of the sub-actions. The part of speech (*POS*) tag is a primary action tag and a clarifying tag. The *POS* primary action is to determine the part of speech of a word occurrence. The POS clarifying tag represents the use of part of speech information in another action. The purely clarifying tags are omitted from the analysis of tag frequency.

I recorded the distribution of the tags according to the number of tasks in which they appeared and the total number of actions which received the tag as shown in Table 3.1. At a high level, this table tells us that both qualitative actions enabled by reading concordance lines and quantitative actions which require positional statistics are necessary for the style of concordance analysis outlined by Sinclair [Sinclair, 2003].

3.5 Discussion

From the process of domain characterization, it has become clear that the use of concordance browsers is central to the analysis of the corpus. It is used in both methodologies and is useful for the vast majority of the corpus questions provided by domain experts. The concordance browser makes reading all text fragments in a corpus related to a keyword easy and, by alignment, makes reasoning about relative word position possible.

Positional word frequencies seem similar to the ability to read the concordance lines, as frequency calculation, estimation, and patterns rank as the fifth, fourth and second most used tags in the analysis of *Sinclair's* tasks. Of the eighteen tasks, sixteen require estimation of frequency using the concordance. The word position tag is the fifth most used tag and is found in thirteen of the tasks. The first and third most used tags are expert decision and read context. *Expert decision* is a qualitative action which captures any consideration of evidence in the context of domain knowledge and is often supported by positional frequency analysis and reading the concordance lines in a concordance browser (read context).

Positional word frequency in the concordance is often discussed in both the methodological examples and questions. There is evidence that corpus tools do not well support estimating or calculating these positional frequencies. For example, in the ranking of questions to ask a corpus, one researcher (Dave) ranked positional word frequencies as the second most important question to answer. The first was the calculation of raw frequencies in the corpus, which the expert confirmed is well supported by frequency lists. In his methodology description, Dave again mentioned the difficulty in using the concordance for identifying patterns when a large number of lines are returned.

Similarly, Daisys' methodology example very explicitly calls for visualisation of positional word frequencies; after describing gaining an overview of positional frequencies as difficult when the browser returns many concordance lines. This is highlighted even further as seven of the twenty questions proposed by Daisy are under the heading "collocational patterns" and require positional frequency analysis as the method of answering. In addition, the questions under the heading of Temporal spread often also require collocational pattern analysis to answer the questions.

Another aspect of collocation patterns, which Dave ranked as the third most important question, is collocation strength. Collocation strength was discussed and understood to mean a quantitative measure of how strongly related context words are to the keyword. An example of the type of questions you could ask about collocation strength would be "Do the context words appear more frequently with the keyword than with any other word?". Daisy includes in her questions"Collocation patterns" the question " Are these adjectives only used to describe this keyword or are they connected to other keywords in this text?". A measure of collocation strength for each word and position would be useful to answer this question.

Keyword frequency is calculated in both Daisy and Daves's methodology examples by selecting a sub-corpus and searching for a keyword to reveal the number of concordance lines. However, in Dave's methodology, question frequency lists are often needed to identify the rank of the frequency of the keywords. Frequency lists have a further use in comparing corpora and sub-corpora. In Dave's questions, frequency list comparison could be used in four of the twenty questions. Daisys questions called for the use of frequency lists in the first question. Additional questions, such as the fifth question and five of the six questions in the "Temporal spread" section, can be answered by comparing frequency lists. Frequency and rank in the lists are important for most analysis, however, the corpora should be of comparable size to draw conclusions from the comparison of the lists.

The sub-corpus selections seem to be regarded as an adequate method of partitioning the corpus for analysis. The problems are associated with the analysis of each being time-consuming as quantitative values are difficult to estimate and compare for large concordances.

3.6 Conclusion

Based on the domain characterisation discussion, we have established that using concordance browsers is central to the corpus analysis and that these do not fully meet the requirements for positional analysis of collocation frequencies and statistics.

Additionally, there is some evidence that using frequency lists in the analysis of corpora is an inadequate comparison method.

In the next chapter (chapter 4), attempts at mitigating visualisation design validity risk at the data abstraction and encoding levels of the nested model of visualisation design (as described in subsection 2.2.3) are presented.

Chapter 4

Data Abstraction and Relevant Encodings

The motivation to create a conceptual model of the KWIC concordance list comes from a need to formalize the data structures, attributes and relationships inherent to the concordance list so that visualisations can be evaluated and designed in terms of their effectiveness at representing the model. The model design seeks to structure the data entities to support best the actions described in the task analysis. This is done in section 4.1.

The establishment of the conceptual data model provides a reference point for the comparison visualisation, which realizes the model. Any existing visualisations which fully or partially realize the model can be compared in terms of their visual variable encoding to determine the benefits and drawbacks of the choices made. In section 4.2, a comparison of existing visual encodings is presented.

Finally, while frequency list comparison does not have a place in the conceptual model a structured encoding comparison of relevant visualisations is an appropriate method for mitigating threats to validity at the data abstraction and encoding levels of the nested model of visualisation design (subsection 2.2.3). This structured comparison is presented in section 4.3.

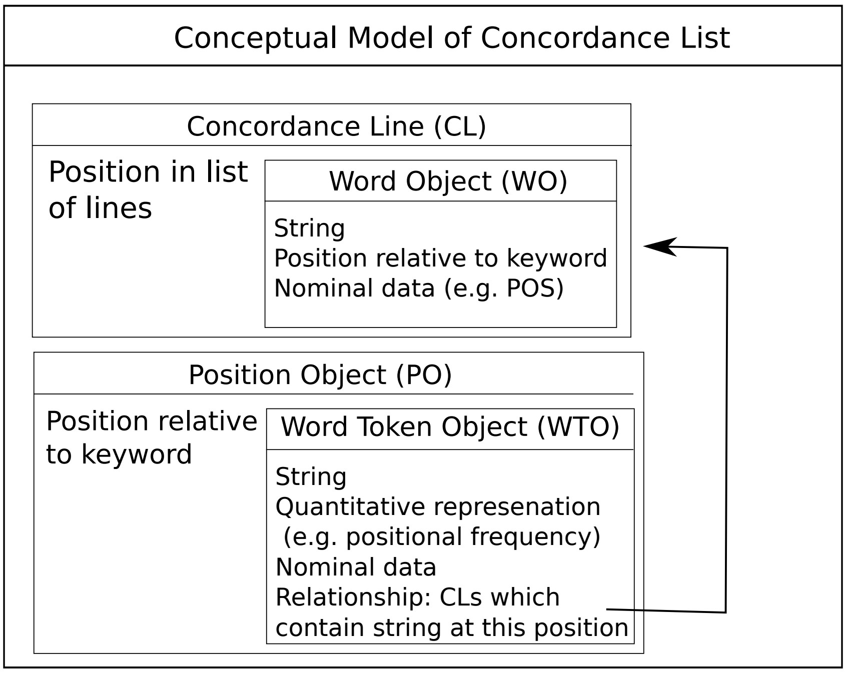

Conceptual Model of Concordance List
Concordance Line (CL)
Position in list of lines
Word Object (WO)
String
Position relative to keyword
Nominal data (e.g. POS)
Position Object (PO)
Position relative to keyword
Word Token Object (WTO)
String
Quantitative represenation (e.g. positional frequency)
Nominal data
Relationship: CLs which contain string at this position

Figure 4.1: Conceptual Data Model of Concordance Lists. Position objects contain word token objects that are a single word representing many occurrences. These may link to several concordance lines where that word occurs at the specified position. The arrow in this diagram represents the one-to-many relationship between the word token objects and the concordance lines.

4.1 Conceptual Data Model

The traditional rendering of a KWIC concordance list evokes a conceptual model consisting of a list of aligned sentence fragments (concordance lines). In this model, each concordance line (CL) has an attribute representing its position in the list (concordance lists are usually presented in alphabetical order) and contains an ordered set of word objects (WO). The word objects represent an individual occurrence of a word in a concordance line and contain its string representation (nominal data), position relative to the keyword and any other nominal variables (meta-data) available, e.g. part of speech tags. Put simply, one possible data model of a concordance list is an ordered set of text fragments where each text fragment contains words. These words have attributes such as the position in the fragment relative to some keyword.

All qualitative actions identified in the task analysis, presented in chapter 3, require reading of the concordance lines ("read context"). This action is the third most common action, only surpassed by the use of expert knowledge to perform a classification and frequency estimation. In order to read the context, the text fragments must be available. Since the concordance lines' linear structure (sentence structure) is emphasized in the concordance line model, it is included in the KWIC conceptual model to facilitate this *read context* action. The data model now offers strong support for the qualitative actions identified in the task analysis.

Since the word objects in the concordance line model are representative of a single occurrence of a word, they do not have as an attribute a quantitative variable such as word frequency. The frequency values are available by counting similar word objects within all concordance lines, but the frequencies are not attributes of any entities in the model. For example, in Figure 3.3, word frequencies in the list can be calculated by counting. I would like the KWIC model to contain these quantitative variables as attributes of some entity since estimating frequency, frequent patterns and frequency values all feature in the top five concordance analysis actions and are the three most common quantitative actions. I also found that word position was often required in conjunction with these frequency-orientated tasks. For instance, estimating word frequency at a position relative to the keyword is more common than estimating frequency in the context window. Again looking at Figure 3.3 it would be beneficial to have the frequencies of the words directly to the left of the keyword immediately available for the types of analysis performed by domain experts, as discussed in chapter 3.

With this in mind, I now conceptualize the concordance lists as an ordered set of position objects (PO). Within each position object, there is an attribute for the position relative to the keyword and a set of word token objects (WTO). If you imagine the concordance as a horizontal list of positions relative to the keyword, where each position contains a word list.

The positional ordering of these WTO within the position objects is an additional positional attribute. These word token objects differ from the word objects in the concordance line model in several ways. The most important way they differ is that these objects represent all occurrences of a string (or string and nominal variable) at the position in which they reside. That is to say, there will be, at most, one object with a particular string and nominal variable (meta-data) combination. For example, if part of speech tags are available, there will be one object representing the noun "date" and one representing the verb "date". Each word token object inherits its position as an attribute. Figure 2.5 shows the pattern viewer from WordSmith Tools. This representation is similar to position objects, where a single occurrence of each word used is presented per position.

Quantitative attributes which represent positional count, frequency or other statistics of the word token object in the KWIC are included in the model. Finally, an attribute (relationship) which maps each word token object to the concordance lines in which it occurs is also available. This attribute and the position attribute of the concordance line word objects provide a link between the models and unify the KWIC conceptual model as seen in Figure 4.1. This linking of the conceptual models is especially useful for the *frequent patterns* action, where word combination frequencies between positions could be used. An analyst may want to view the concordance lines where a particular word occurred directly to the left of the keyword, questions similar to "which lines in Figure 3.3 have the word "a" directly to the left of the keyword?" were identified as important in chapter 3.

The attributes of the KWIC conceptual model (Figure 4.1) can be referenced using a concatenated orthography to identify the Object.attribute or Object.Subobject.attribute. For example, the strings that identify the position object's word token objects can be referred to using the concatenation PO.WTO.String. In Table 4.1, Table 4.2, and Table 4.3, the concatenation is extended to identify the data type of each attribute between Nominal, Ordinal and Quantitative types. This is useful when assessing the encoding choices. Mackinlay's ranking of visual variables Figure 2.7 provides a framework for reasoning about the perceptual efficiency of an encoding of a conceptual model in terms of these three data types.

To practically realize this model and generate appropriate data structures to represent the elements of this conceptual model, I developed the concordance graph described in subsection 5.1.1

4.2 Structured Encoding Comparison: KWIC Visualisations

Text visualisation encompasses many different visual methods. I am interested in comparing visualisations which have been designed for keyword search results represented as a concordance list. However, I also wish to investigate techniques which, while not designed with the concordance in mind, could potentially have applications in concordance analysis.

Evaluation of the selected visualisations at level three (Figure 2.9) of the nested model [Munzner, 2009] should enable validation of the visual encoding for the identified tasks. At level three of the model, I considered two main validation approaches. One is to test the systems in a laboratory experiment. None of the KWIC visualisations proposed as extensions of the traditional interface had experimental data or results. In addition, most of the systems had no usable implementation at the time of writing and have not been integrated into existing corpus tools. For these reasons, evaluation and comparison using traditional experimental methods were not attempted.

The second approach is a qualitative discussion of images or videos of the system. I have chosen to present a qualitative discussion of the systems in a semi-structured manner. I hope structuring the discussion and presenting tabular descriptions of visual encoding and data abstraction will provide clarity in the comparison and become a standard encoding evaluation technique. To paraphrase Munzer's statement, while visualisation experts may draw the same conclusions from the inspection of a system, the validation is strongest when there is an explicit discussion pointing out the desirable properties in the results.

Mackinlay proposed a ranking of visual variables[Bertin, 1983] per data type [Mackinlay, 1986] (see Figure 2.7). This ranking of variables is in agreement with a ranking proposed for quantitative data by Cleveland and McGill [Cleveland and McGill, 1985]. Mackinlays ranking of visual variables for the 3 data categories should help guide variable choice. A justification should be given if a variable is chosen from a data attribute instead of a higher-ranked one. Encodings are often presented, leaving the reader to guess the author's reasoning [Green, 1998].

I suggest that to evaluate or design a visual encoding, one should perform a semi-structured encoding discussion using Mackinlay's ranking of visual variables. Once this has been done, easier comparison of visualisations within and across problem domains would be possible. Taxonomizing current visualisations in this way could lead to much clearer comparisons and aid in adoption across problem/domain barriers. To structure the discussion, the importance of the attributes of the data model should be emergent from a detailed domain characterization.

The task analysis, described in chapter 3, has identified a split in KWIC tool requirements and this split is reflected in the conceptual model Figure 4.1. Qualitative actions often operate on concordance lines (CL), where the readability and linear structure of the lines are emphasized. While quantitative actions require positional statistics (PO), they do not often require the readability of the individual concordance lines.

Looking at the conceptual model and task analysis together, I concluded that the most important attribute to present for qualitative primary action is the ordered list of words making up the concordance line (CL.WO.position.ordinal) so that the fragments are readable. All other attributes are much less important. Once the sentence fragments are presented to the analyst in a readable manner, all qualitative analysis is facilitated. Encoding CL.position.ordinal, the order in which the lines are presented in a logical manner, can likely aid the qualitative concordance analysis. An example of this would be sorting the lines alphabetically at a position. Other attribute encodings, such as part of speak aid analysis but are not strictly required.

Quantitative actions often require the estimation of frequency or word statistics. The calculation of exact word frequencies is also a common action. Clearly, to facilitate the quantitative actions attributes PO.position attribute, PO.WTO.quantitative attribute and, to a lesser extent PO.WTO.position attribute should be prioritized. The PO.position attribute describes the grouped word positions relative to the keyword. The PO.WTO.quantitative attribute represents word token statistics and the PO.WTO.position attributes describe how the tokens are arranged within the positional groupings.

The analysis presented explores the mappings of the ranked visual variables to the conceptual model, providing a framework within which it is possible to compare the various concordance visualisations. The ranking of the visual variables for different data types helps us reason about the advantages and limitations of the visualisations for the actions I have identified. However, when applied to each data attribute, this ranking of the variables does not fully quantify the validity of the visualisations. This process does not capture other factors. The discussions of each visualisation try to capture as many additional limitations and advantages as possible. I do not wish to comment on visual design as it pertains to the beauty of a visualisation, and will instead stick to an evaluation in terms of long-established visualisation patterns, such as overview+detail on demand [Shneiderman, 1996], and those design choices which restrict or improve the information encoding of the conceptual data model.

The method used to evaluate each candidate visualisation is first to map each attribute of the KWIC conceptual model to the relevant visual variables in that visualisation. Then make note of the number of visual variables mapped to the attribute and the ranking of the visual variables for the data type it represents in the visualisation. The attributes of the KWIC conceptual model are further expanded by categorizing them as nominal, ordinal or quantitative data types. For example, position objects (PO) can be both positionally ordinal and nominal by being nominal groupings of words which also have order relative to the other position objects. The mapping of these attributes to visual variables for the examined visualisations is presented in Table 4.1, Table 4.2 and Table 4.3. Comparison of the mappings presented in these tables gives an overview of the usefulness of each visualisation for the identified quantitative and qualitative actions.

4.2.1 Keyword-In-Context (KWIC)

To begin with, let's examine more formally and evaluate the most widely used concordance visualisation, the traditional KWIC visualisation. The position of the concordance lines is represented by the vertical positioning of the word objects in each line. The horizontal position of the word objects is used to represent the position relative to the keyword. In some traditional KWIC renderings, the beginning of each string (representing a word object) is aligned with all other strings at the same position relative to the keyword, creating a grid-like view. Other versions of these visualisations trade this vertical alignment for increased horizontal readability by rendering the left and right contexts similarly to how they would appear in the source texts. When this unaligned context version is presented, the option to highlight, using colour, word objects with the same positional value is often available. This creates an *integral* (combination of visual variables which are perceived together) combination of the visual variables horizontal position and colour hue for the attribute word object position.

Table 4.1 contains a summary and ranking of the visual variables used by the traditional KWIC visualisation. The concordance line position is represented by the vertical position of its enclosed strings. These enclosed strings are rendered

Table 4.1: (A) Analysis of text visualisations in terms of the KWIC conceptual model mapped to the visual variables available in the visualisation. Q = quantitative, O = ordinal and N = nominal. The numbers combined with the data type letter represent the position of the visual variable in Mackinlay's visual variable ranking. For example, (N 2) represents the second-best visual variable for nominal data.

Attributes	KWIC visualisation	Mosaic	interHist
CL.position.ordinal	vertical pos (O 1)		
CL.WO.String.nominal	color hue (N 2)		
CL.WO.position.ordinal	horizontal pos (O 1)		
CL.WO.position.nominal	color hue (N 2)		
CL.WO.meta-data.nominal	color hue (N 2)		
PO.position.ordinal	horizontal pos (O 1)	horizontal pos (O 1)	horizontal pos (O 1)
PO.position.nominal	color hue (N 2)	horizontal pos (O 1)	
PO.WTO.position.ordinal		vertical pos (O 1)	vertical pos (O 1)
PO.WTO.meta-data.nominal		color hue (N 2)	color hue (N 2)
PO.WTO.quantitative		length (Q 2)	length (Q 2)
PO.WTO.relationship.nominal			

horizontally, left to right, in the order they appear in the text fragments. Put plainly. The sentence fragments are presented in a familiar linear style. Both concordance lines and word objects are mapped to the best visual variables for the ordinal position data types. Because of this, I expect it will be easy to identify individual concordance lines and find where they rank in the chosen ordering scheme (usually alphabetically by a selected or default word position). Similarly, it will not be difficult to identify the word objects in the order in which they appear in the text fragments. The concept of a position object can only be loosely applied in this visualisation. The word positions across concordance lines can be easily identified if an associative visual variable, such as colour, is encoded on words of the same position. The variable colour hue is mapped to three attributes. This could cause a perceptual issue as the meaning of the variable is inconsistent. Practically only one of these attributes would be mapped to the variable at a time, and different views could be used for each attribute.

Since the KWIC display is designed with the readability of concordance lines in mind, the inability to gain an overview of a large concordance is a necessary trade-off. In this rendering, the detail is presented at all times. An overview of

Table 4.2: (B) Analysis of text visualisations in terms of the KWIC conceptual model mapped to the visual variables available in the visualisation. Q = quantitative, O = ordinal and N = nominal. The numbers combined with the data type letter represent the position of the visual variable in Mackinlay's visual variable ranking. For example, (N 2) represents the second-best visual variable for nominal data.

Attributes	Word Tree	Corpus Clouds	Parallel Coordinates
CL.position.ordinal		vertical pos (O 1)	
CL.WO.String.nominal		color hue (N 2)	
CL.WO.position.ordinal		horizontal pos (O 1)	
CL.WO.position.nominal		color hue (N 2)	
CL.WO.meta-data.nominal		color hue (N 2)	
PO.position.ordinal	horizontal pos (O 1)	horizontal pos (O 1)	horizontal pos (O 1)
PO.position.nominal	connection (N 4)	color hue (N 2)	horizontal pos (O 1)
PO.WTO.position.ordinal	vertical pos (O 1)		vertical pos (O 1)
PO.WTO.meta-data.nominal			color hue (N 2)
PO.WTO.quantitative	area (Q 5)	length (Q 2)	length (Q 2)
PO.WTO.relationship.nominal	connection (N 4)		connection (N 4)

the entire concordance list is only available for concordances which fit within the screen at a readable font size. Windowing the concordance and scrolling is the usual solution. This works well for viewing individual concordance lines, but a higher-level view would be better to get an overview of the positional frequencies and patterns. Larger screen sizes and higher resolutions can improve the situation, but the scale required becomes impractical as more data becomes available.

Clearly, this visualisation contains no explicit representation of the word token objects, so I expect visual assessment of exact or estimated word frequency to be difficult. Nevertheless, this visualisation is the most common tool used for concordance analysis, where, as I have shown, positional frequencies are regularly used. The observational study and task analysis found that counting the strings is the usual way to calculate these positional word frequencies. While this visualisation is very effective for reading concordance lines, it would seem to be of limited use for quantitative concordance actions.

To summarize, looking at the conceptual model and variable mapping Table 4.1, I noticed that the traditional KWIC interface facilitates the CL objects extremely well due to its encoding mapping the CL object of the conceptual model to high-ranking

Table 4.3: (C) Analysis of text visualisations in terms of the KWIC conceptual model mapped to the visual variables available in the visualisation. Q = quantitative, O = ordinal and N = nominal. The numbers combined with the data type letter represent the position of the visual variable in Mackinlay's visual variable ranking. For example, (N 2) represents the second-best visual variable for nominal data.

Attributes	Double Tree	Bi-Directional
CL.position.ordinal		
CL.WO.String.nominal		
CL.WO.position.ordinal		
CL.WO.position.nominal		
CL.WO.meta-data.nominal		
PO.position.ordinal	position(O 1) & connection(O 6)	position(O 1) & connection(O 6)
PO.position.nominal	position (N 1)& connection (N 4)	position (N 1)& connection (N 4)
PO.WTO.position.ordinal	vertical pos (O 1)	vertical pos (O 1)
PO.WTO.meta-data.nominal		
PO.WTO.quantitative	color sat (Q 8)	area (Q 5)
PO.WTO.relationship.nominal	color hue(N 2) & connection (N 4)	color hue(N 2) & connection (N 4)

visual variables. There are many available implementations of this visualisation, and it has been widely adopted as the main corpus interface of linguistic analysis in systems such as WordSmith, MonoConc, the Stuttgart workbench, Manatee and the TEC corpus browser [Scott et al., 2001, Luz, 2011, Kilgarriff et al., 2008]. However, no encoding of quantitative information is attempted in the KWIC interfaces. This leaves the user needing external tools or counting word occurrences by hand [Scott, 2010].

4.2.2 Tree Representations

The visualisation of concordances called *Word Tree* displays the keyword and one side of the context tree, either the left or right context [Wattenberg and Viégas, 2008]. As the name suggests, this visualisation takes the familiar form of a tree structure, in which the keyword is displayed as the root vertex and additional word vertices are connected in text order to each other. An example of a right context Word Tree for the keyword "eye" is shown Figure 4.2.

Table 4.2 contains a summary and ranking of the visual variables used in the Word Tree visualisation. The main benefits of this visualisation are that the linear structure and readability of the concordance lines are maintained through the combination of the visual variable's connection and horizontal position, in addition to the inclusion of some indicators of word statistics. Connection defines the word position by the number of edges from the root, and horizontal position per branch provides partial positional groups (ordered positions in a sub-tree). These positional groups allow the frequencies at a position along a branch to be easily estimated since the words are rendered proportional size of their sub-tree. However, while frequency in a branch/sub-tree is easy to estimate, the frequency at a word position is less clear. Looking at word positions as they move away from the root (keyword), positional frequencies become increasingly difficult to estimate. This is because any token can occur for each node in the proceeding tree level. This leads to the possibility of multiple occurrences of a word object at a position. So at a position/depth of one from the root/keyword, each rendered word represents a positional word token object (WTO), but deeper into the tree, each rendered word is a partial WTO which only represents each occurrence of a token at that position in the sub-tree. A combinatorial explosion causes the estimation of frequency to be more difficult when viewing positions deeper into the tree. At the first position from the keyword, Word Tree does allow for a choice of sorting of the vertical position of the words. One of the sorting schemes is by frequency which is an advantage for the quantitative actions identified.

An additional problem with the estimation of frequency (or other word statistics) using this visualisation is that variation in word length causes the variable (Area) representing the quantitative information to be inconsistent. The square root scale used by the visualisation should make word Area roughly proportional to frequency if not for these word length variations. While it may initially seem natural to include quantitative information about a word by scaling the font representing that word, it is worth noting that the visual variable Area ranks fifth for the display of quantitative information under Mackinlay's ranking scheme and, additionally, variations of word length complicate the interpretation of the quantitative values.

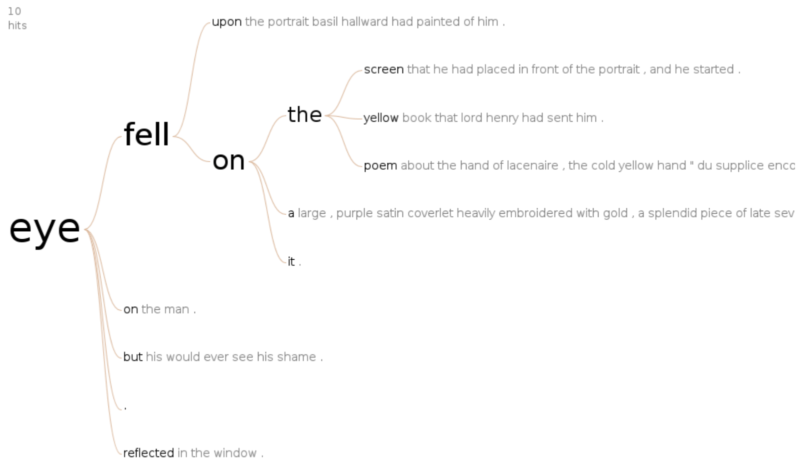

10
hits
eye
fell
upon the portrait basil hallward had painted of him .
on
the
screen that he had placed in front of the portrait , and he started .
yellow book that lord henry had sent him .
poem about the hand of lacenaire , the cold yellow hand " du supplice encc
a large , purple satin coverlet heavily 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 4.2: WordTree for keyword eye and its right context. Generated by the free online service ManyEyes.

Word Tree allows for rendering an entire side of a concordance list (either left or right context). This provides an overview which can be explored through interactive techniques to investigate the detail. Using font size to display quantitative information influences the scaling that is used. If the frequency differences are very large much of the context will be unreadable in the fully expanded overview.

In conclusion, Word Tree provides a useful interface for investigating word frequencies at a distance of one position from the keyword, with a drop-off in usefulness the further from the keyword you go. An overview of the context is available along with the detail, which allows for investigation of how the concordance lines diverge after the keyword, but reading the individual concordance lines in this view offers few advantages over a sorted KWIC list. The biggest limitation of the visualisation is that only a single context, either left or right, of the concordance can be viewed at a time.

Double Tree [Culy and Lyding, 2010] extends and refines the Word Tree visualisation to create a tool designed for the “linguist’s task of exploratory search using linguistic information”. This exploratory search task has different objectives than those encountered in Sinclair’s concordance analysis tasks. So, it is unsurprising that some design decisions hinder the actions that should be enabled by concordance visualisation tools.

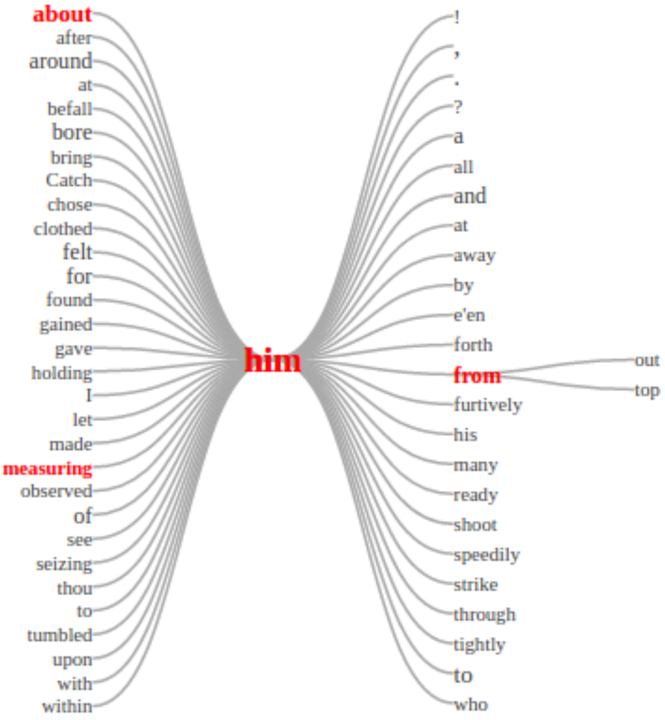


Figure 4.3: DoubleTreeJS demo for keyword "him" from the text *Robin Hood*.

Table 4.2 shows a breakdown of the visual variable encoding used in the Double Tree visualisation, mapped to the conceptual model. As I have previously stated, quantitative actions are important for concordance analysis, and the ability to estimate and calculate positional word frequency distributions is vital to these actions. Double Tree encodes the branching factor of a word/node by the variable colour saturation. In the ranking of quantitative variables, colour saturation ranks eighth. This branching factor is not a quantity of interest in this analysis (though it can be useful for looking at frequent combinations within sub-trees). Positional frequency is given no visual representation. In fact, the visualisation only completely displays the positions one to the left (k-1) and one to the right (k+1) of the keyword (k). All other positions are displayed only when individual branches are expanded. The full left or right context trees are never displayed fully expanded.

Position in the Double Tree is encoded independently by two variables connection and horizontal position. This is different than the encoding of Word Tree, where horizontal position only indicates position relative to its parent and child nodes, and connection encodes absolute positional values. The combination of these variables clearly defines the positional groups of the displayed word token objects no matter how far into the sub-trees you have expanded. Compared with the Word Tree representation of the nominal position groups, using the top-ranked visual variable position to represent the groupings increases the associativity of the positional groups.

The major advantage of Double Tree (Figure 4.3) over Word tree is the display of both contexts simultaneously. Connecting both context trees at the keyword root node creates the double tree structure. This connected multi-tree structure does not, however, link the contexts. From one context to the other, the continuity of the text fragments is broken. Double Tree links these contexts using the variable colour hue to show which words in the first position of one context form concordance lines with words in the first position of the other context. This is only a partial solution to this problem; as the trees branch from these connected words, it's not possible to tell which branches connect to which other branches in the opposite context. This is because the branches expanding from a node contain all sentence fragment continuations from that word. In Figure 4.3, the keyword "him" is linked to the word "from" at position (k+1). At position (k-1), the words "about" and "measuring" are highlighted, indicating a connection through the keyword. However, the continuation of the fragments to the words "out" and "top" is ambiguous either could connect through the keyword. The paper has not revealed the process or data structure used to link the contexts.

The graph-based abstraction of concordance lists joins the two context trees to produce a concordance graph. This graph includes additional *contextual edges*, which connect the two contexts and allows the recovery of all of concordance lines that connect to a particular node in the graph. As previously stated, the keyword node is characterised in terms of graph eccentricity properties and positions (WPOs) and consists of all nodes at a certain distance and direction from the keyword node.

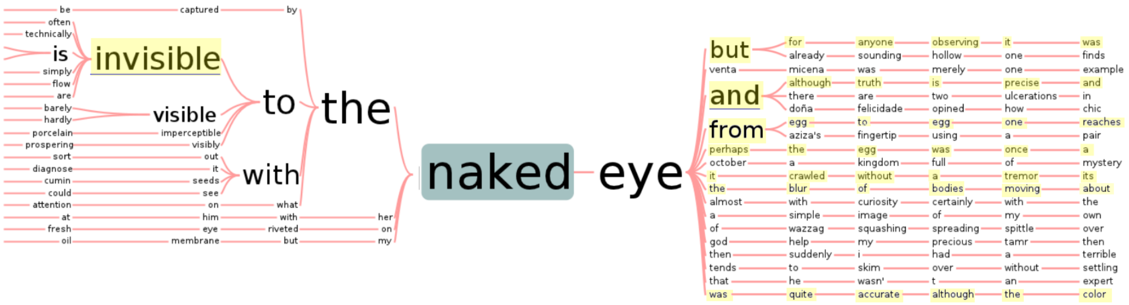

be captured by
often
technically
is
invisible
simply
flow
are
to
the
barely
hardly
visible
porcelain imperceptible
prospering visibly
sort out
diagnose it
cumin seeds
could see
with
attention on what
at him with her
fresh eye riveted on
oil membrane but my
naked
eye
but for anyone observing it was
already 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 4.4: Bi-directional hierarchical view of a concordance for the compound "naked+eye".

In addition to the mosaic visualisation, this new structure was used to create a Double Tree visualisation titled *bi-directional hierarchical display*. Figure 4.4 shows the visualisation for the keyword *naked-eye*. This visualisation differs from Double Tree in having a more frequency-focused encoding of its visual variables (see Table 4.2). The size of each word is proportional to its positional frequency. These font scale values are not per-branch frequencies, as in Word Tree and Double Tree, but are positional frequencies relative to the keyword. The combination of this scaling scheme with the display of the fully expanded contexts makes this a strong candidate for use in the concordance quantitative actions encountered in the task analysis. Using the visual variable area, applied to the text label, to represent the quantitative information has the same drawbacks as discussed for Double Tree.

In all tree-based and double-tree-based visualisations, only a single position (or one position per context) can be sorted at a time. Otherwise, the edges connecting the words might cross. Crossing the edges makes it more difficult to investigate the linear structure of the concordance lines, so each of these visualisations does not allow it. The type of order does not matter if you sort at a global position, not simply per branch. The same issues persist for any ordering scheme. The traditional KWIC visualisation can also only be sorted by a single position since the conservation of the concordance lines breaks down when more than one position is sorted.

4.2.3 Other Relevant Visualisations

Similar to the Concordance Mosaic, interHist [Lyding et al., 2014] is a complementary visualisation that is used to display quantitative information about its accompanying KWIC view. In this case, the visualisation is of stacked bars analogues to mosaic tiles where height is used to display the quantitative information. This interface differs from the mosaic in that it is designed for part of speech information and does not represent individual tokens or WTOs. Instead, each rectangle is a positional part of a speech object. It is not difficult to imagine these rectangles as representing WTOs without changing the visual representation. The visual variable breakdown for interHist, in Table 4.1, is constructed under these false assumptions that the rectangles represent WTOs. In interHist vertical positioning of the rectangles is not encoded with meaning, making it more difficult to perceive quantitative differences between the rectangles than in the Mosaic where an integral combination of variables position and length is used. The scaling also differs from Concordance Mosaic, where a space-filling approach was used. In interHist the total length of the positional bar explains the difference in total quantity between positions. For word frequency, all bars will be the same height due to requiring concordance lines of consistent length.

In both interHist and Concordance Mosaic, it was shown that colour could be used to effectively represent nominal meta-data, such as part of speech tags or labels. The Double Tree visualisation makes no use of the visual variable colour hue. This visualisation does not display visual encoding of any meta-data. It does seem fair to assume that colour hue could be used for meta-data in these visualisations. The authors not explicitly outlining this obvious potential encoding choice isn't a good argument against this visualisation's ability to visually encode metadata.

Corpus Clouds [Culy and Lyding, 2011] is a frequency-focused corpus exploration tool which consists of composite views of a corpus query. The main display is a word

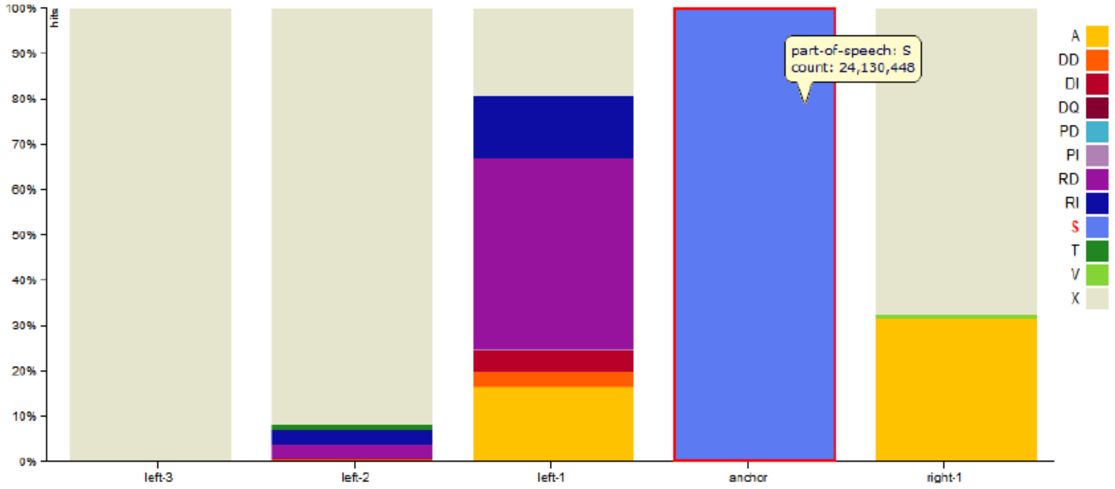
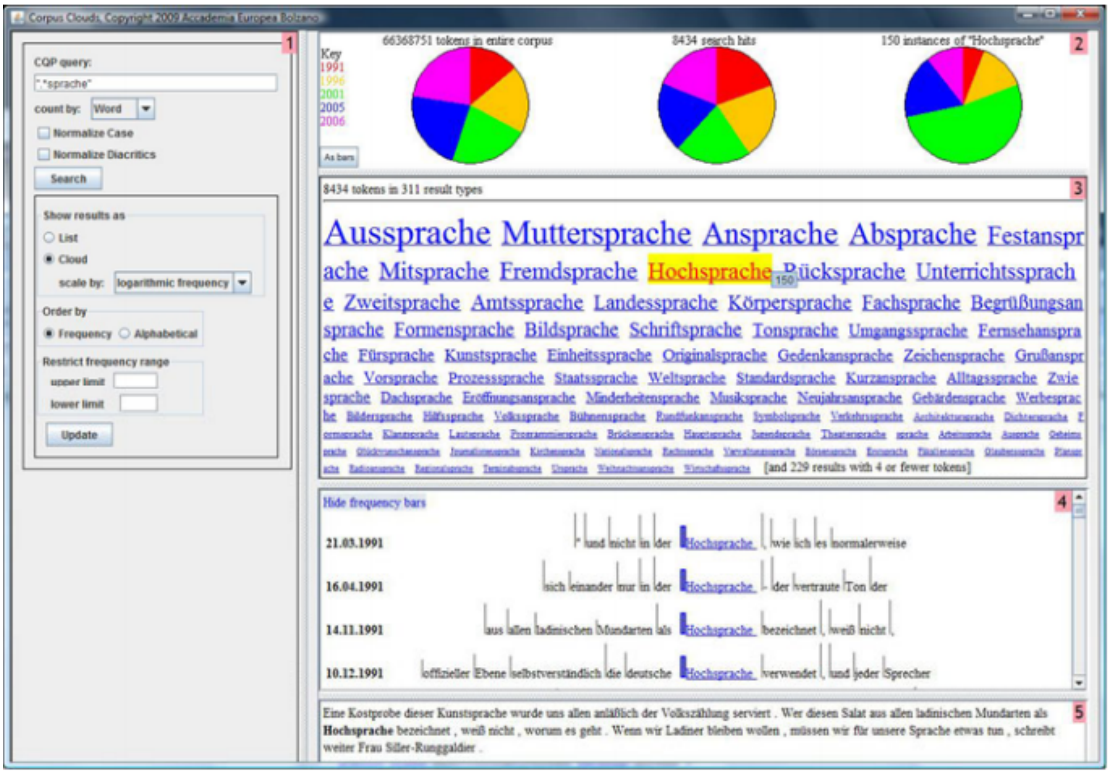


Figure 4.5: 'interhist' interface showing part of speech positional frequencies relative to *nouns* in Italian.

cloud, based on the tag cloud visualisation [Viégas and Wattenberg, 2008], where the absolute frequencies of all the words returned by a corpus query are displayed, see Figure 4.6. These word clouds map this quantity to an area using font size, the limitations of which were previously discussed. This visual encoding does not translate to the conceptual model since positional concordance frequencies are the quantity of interest, not global frequency lists.

Another view in the interface presents a modified KWIC display. The modification is the addition of a small vertical bar, similar to a sparkline, beside each word token in the KWIC view. This makes use of the variable length (Q 2) for frequency information, but the effectiveness of the variable is reduced for several reasons. The main limitation is that the size of the bars is restricted, causing only large differences in frequency to be perceived easily. Additionally, comparisons between lines take place in both planes, vertically across concordance lines and horizontally within lines, again making it difficult to perceive small variations in frequency. Also, the number of KWIC lines which can be displayed per screen is practically limited if readability is to be maintained. The visual variable breakdown for this design is presented in Table 4.3.

The final concordance-based visualisation I surveyed is called Structured Parallel



Corpus Clouds, Copyright 2009 Accademia Europea Bolzano

CQP query: ".*sprache"

count by: Word

☐ Normalize Case

☐ Normalize Diacritics

Show results as: ○ List ● Cloud

scale by: logarithmic frequency

Order by: ● Frequency ○ Alphabetical

Restrict frequency range: upper limit, lower limit

66368751 tokens in entire corpus

8434 search hits

150 instances of "Hochsprache"

8434 tokens in 311 result types

[and 229 results with 4 or fewer tokens]

21.03.1991 - und nicht in der Hochsprache , wie ich es normalerweise

16.04.1991 sich einander nur in der Hochsprache - der vertraute Ton der

14.11.1991 aus allen ladinischen Mundarten als Hochsprache bezeichnet , weiß nicht ,

10.12.1991 offizieller Ebene selbstverständlich die deutsche Hochsprache verwendet , und jeder Sprecher

Eine Kostprobe dieser Kunstsprache wurde uns allen anläßlich der Volkszählung serviert . Wer diesen Salat aus allen ladinischen Mundarten als **Hochsprache** bezeichnet , weiß nicht , worum es geht . Wenn wir Ladiner bleiben wollen , müssen wir für unsere Sprache etwas tun , schreibt weiter Frau Siller-Runggaldier .

1
2
3
4
5
Key
1991
1996
2001
2005
2006
As bars
Search
Update
Hide frequency bars
Aussprache Muttersprache Ansprache Absprache Festansprache Mitsprache Fremdsprache Hochsprache Rücksprache Unterrichtssprache Zweitsprache Amtssprache Landessprache Körpersprache Fachsprache Begrüßungsansprache Formensprache Bildsprache Schriftsprache Tonsprache Umgangssprache Fernsehansprache Fürsprache Kunstsprache Einheitssprache Originalsprache Gedenkansprache Zeichensprache Grußansprache Vorsprache Prozesssprache Staatssprache Weltsprache Standardsprache Kurzansprache Alltagssprache Zwiesprache Dachsprache Eröffnungsansprache Minderheitensprache Musiksprache Neujahrsansprache Gebärdensprache Werbesprache
150

Figure 4.6: CorpColud visulisation [Culy and Lyding, 2011].

Coordinates [Culy et al., 2011]. The breakdown of the encoded variables for this visualisation is also given in Table 4.3. Several uses of the parallel coordinates visualisation for different types of structured language data are presented, one of which is a KWIC plus frequency visualisation seen in Figure 4.7. This visualisation places WTOs, rendered as text labels, on the parallel axis, which represents ordered word positions. The concordance line structure is maintained using a connection between the words positioned on the parallel y-axes. Statistical information, such as frequency, is then placed on the axes, and the connection height between the position axis and the quantitative y-axes is used to express the statistical quantities, such as frequency. An individual quantitative axis is required for each quantity and word position pair. As with all parallel coordinate visualisations, the choice of y-axes orderings is important. In this case, the choice was to order the word positions in concordance list order and create the statistical axes to the right of the collection of position axes. This positioning makes it difficult to follow connections from word positions distant from the relevant quantitative axes, and quantitative comparisons between word positions can be perceptually difficult. This difficulty is due to needing information from four axes for comparison across two word positions of a single statistic. In comparison, interHist and Concordance Mosaic, in essence, combine the quantitative and word position axes to enhance the link between a WTO and its associated statistics.

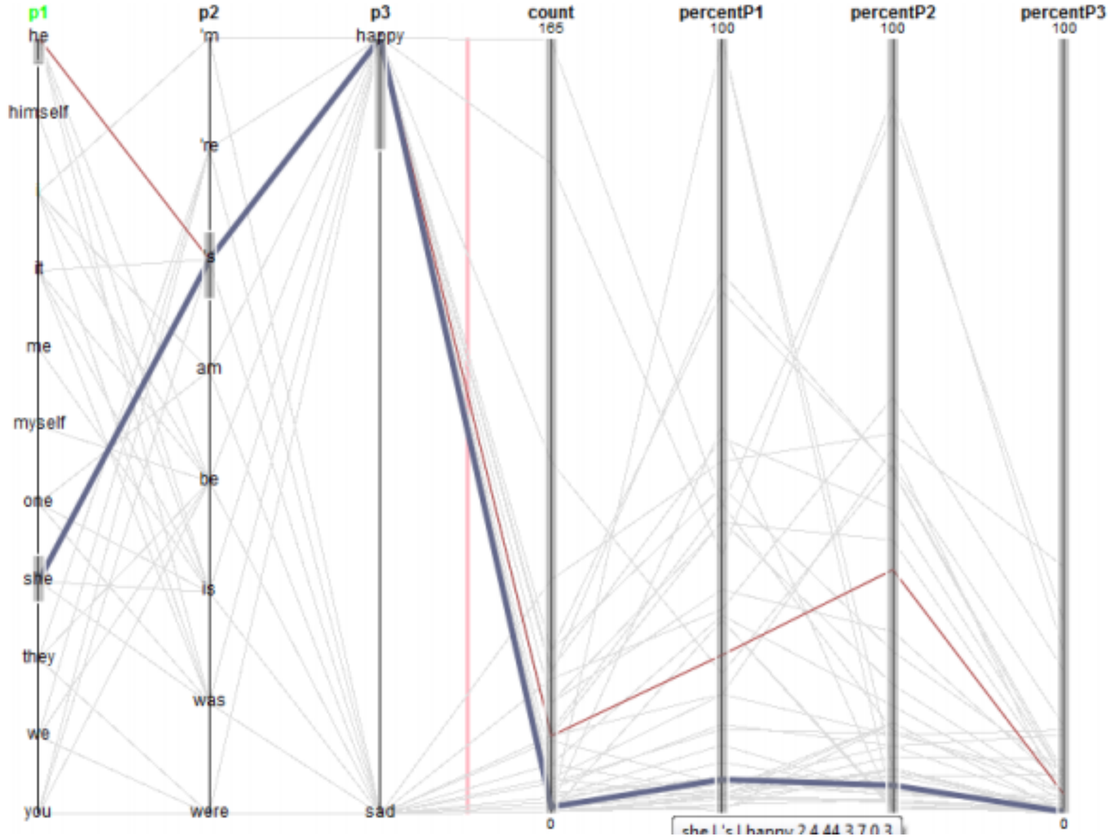


Figure 4.7: Structured Parallel Coordinates for ngrams plus frequencies of *preposition* followed by lemma “to be” followed by “happy/sad”.

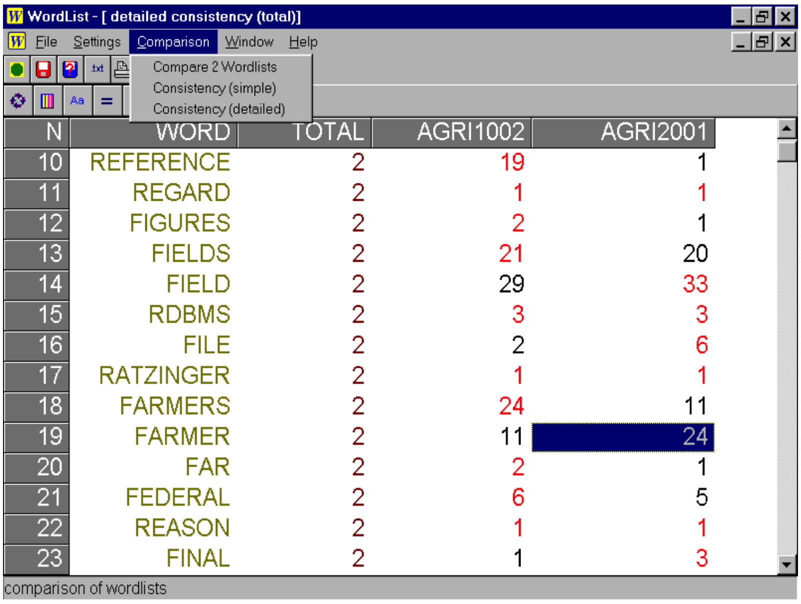
Structured Parallel Coordinates have an advantage over other visualisations: the ability to show multiple statistics on a single rendering. This option is available since, in theory, one may add as many parallel axes as one wishes. However, this advantage also comes with a trade-off, as each additional axis increases visual complexity. Even in the case of a single-word statistic and multiple-word positions, identification of the most common word is visually taxing. Reordering the axes can help but does not entirely solve the issue. In addition, in Structured Parallel Coordinates, the linear order of the sentences is partially maintained through connection. However, the sentences are lost since the connected nodes are WTOs, and only the preceding and next connections are meaningful. That is to say, for any position more than one position away from any word, the user no longer knows if the words connect to form a concordance line.

4.3 Structured Encoding Comparison: Frequency List Comparisons

The need for a tool to better facilitate frequency list comparison emerged through discussion with domain experts and exploration of their corpus analysis workflows. In chapter 3, the need to be able to compare word frequencies was identified as a core task in corpus analysis. During domain characterization, the usage of frequency lists for these comparisons was identified as a requirement; ideally, a comparison of frequency lists of unequal sizes would be possible. A structured encoding comparison of relevant literature is performed here to identify if existing systems meet these requirements.

Currently, only a limited number of visualisations are widely used for frequency-based linguistic tasks. The most commonly used of them generally provide only tables or "Word Lists". 4.8 shows an example of comparing word frequencies using *Wordsmith tools* [Scott et al., 2001], a popular corpus linguistics analysis software suite. Tabular structures are fine if the user is only interested in frequencies of specific words, but they are not suitable for providing overviews or for allowing more complex comparisons across different parts of a table.

Another popular visualisation tool for comparing word frequencies is "Word clouds", where the general idea is to draw each word in a font size proportional to its frequency of occurrence in a given text corpus. This type of visualisation

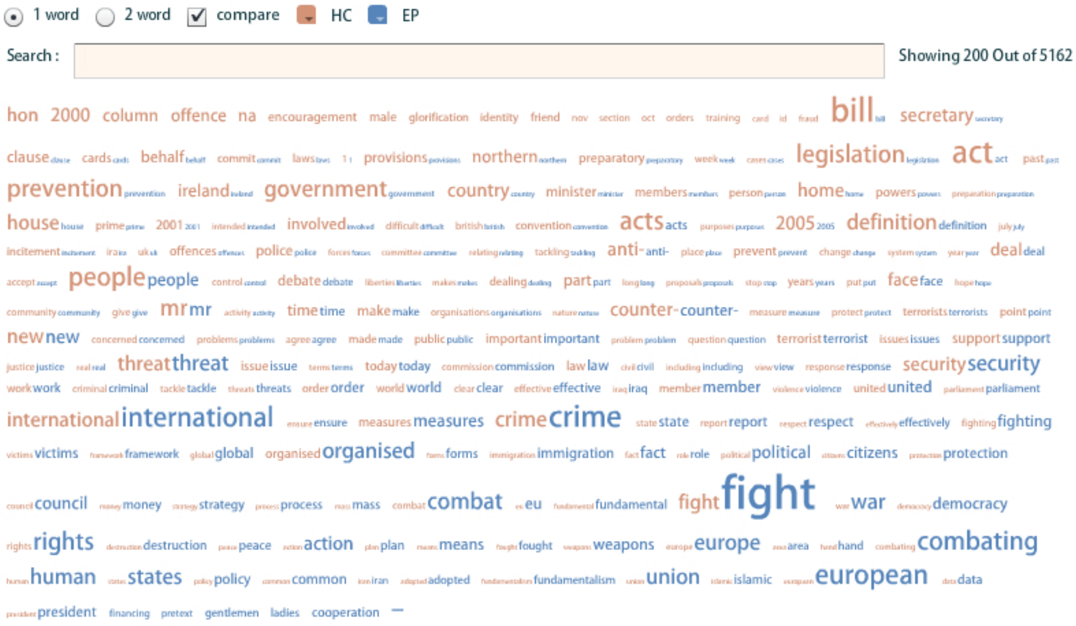


N	WORD	TOTAL	AGRI1002	AGRI2001
10	REFERENCE	2	19	1
11	REGARD	2	1	1
12	FIGURES	2	2	1
13	FIELDS	2	21	20
14	FIELD	2	29	33
15	RDBMS	2	3	3
16	FILE	2	2	6
17	RATZINGER	2	1	1
18	FARMERS	2	24	11
19	FARMER	2	11	24
20	FAR	2	2	1
21	FEDERAL	2	6	5
22	REASON	2	1	1
23	FINAL	2	1	3

Figure 4.8: Word frequency comparison in *Wordsmith*.

is sometimes useful in allowing the user to get "the gist" of a text and has been used for comparing texts and sub-corpora. For instance, words or bi-grams of concordances for the same word from different corpora could be presented side-by-side to allow some level of frequency comparisons. Figure 4.9, for example, shows the most frequent words among the collocates of the word "terror" in the European Parliament (blue colour) and the UK House of Commons (brown colour) speeches [Calzada-Pérez and Luz, 2006], rendered as word clouds using IBM's *ManyEyes* tool [Viégas et al., 2007].

There are a number of well-known limitations with Word cloud visualisations. Seminal work on visual variables [Bertin, 1983] and perceptual tasks [Cleveland and McGill, 1985, Mackinlay, 1986] has shown that the most accurate quantitative perceptual tasks are perception of *position*, followed by *length*, *angle/slope*, *area*, *volume*, and *color/density* as the least accurate. While word clouds attempt to equate font size to word length (height), which is identified as an effective visual task in the classification of [Mackinlay, 1986], the fact remain that longer words will be highlighted against shorter words, even if their frequencies are equal. This is because the salience of a word in word clouds is determined not by its height but by



● 1 word ○ 2 word ☑ compare 🟫 HC 🟦 EP

Search :

Showing 200 Out of 5162

hon 2000 column offence na encouragement male glorification identity friend nov section oct orders training card id fraud bill bill secretary secretary
clause clause cards cards behalf behalf commit commit laws laws 1 1 provisions provisions northern northern preparatory preparatory week week cases cases legislation legislation act act past past
prevention prevention ireland ireland government government country country minister minister members members person person home home powers powers preparation preparation
house house prime prime 2001 2001 intended intended involved involved difficult difficult british british convention convention acts acts purposes purposes 2005 2005 definition definition july july
incitement incitement ira ira uk uk offences offences police police forces forces committee committee relating relating tackling tackling anti- anti- place place prevent prevent change change system system year year deal deal
accept accept people people control control debate debate liberties liberties makes makes dealing dealing part part long long proposals proposals stop stop years years put put face face hope hope
community community give give mr mr activity activity time time make make organisations organisations nature nature counter- counter- measure measure protect protect terrorists terrorists point point
new new concerned concerned problems problems agree agree made made public public important important problem problem question question terrorist terrorist issues issues support support
justice justice real real threat threat issue issue terms terms today today commission commission law law civil civil including including view view response response security security
work work criminal criminal tackle tackle threats threats order order world world clear clear effective effective iraq iraq member member violence violence united united parliament parliament
international international ensure ensure measures measures crime crime state state report report respect respect effectively effectively fighting fighting
victims victims framework framework global global organised organised forms forms immigration immigration fact fact role role political political citizens citizens protection protection
council council money money strategy strategy process process mass mass combat combat eu eu fundamental fundamental fight fight war war democracy democracy
rights rights destruction destruction peace peace action action plan plan means means fought fought weapons weapons europe europe area area hand hand combating combating
human human states states policy policy common common iran iran adopted adopted fundamentalism fundamentalism union union islamic islamic european european data data
president president financing pretext gentlemen ladies cooperation

Figure 4.9: Word frequency comparison of the European Parliament (in blue) and the UK House of Commons (in brown) speeches related to the word "terror".

the total area of the word occupies. The area of a longer word will be larger than the total area of a shorter word of the same frequency. Indeed, studies have suggested that in certain situations, tag (word) clouds will be even less effective than simple lists [Rivadeneira et al., 2007] as a frequency visualisation device.

Visual tools with capabilities for comparison of textual items using sets have been previously proposed. *Parallel tag clouds* [Collins et al., 2009] enables quantitative comparisons across documents. In Figure 4.10, a parallel tag cloud which compares ten lists simultaneously is displayed. This technique is designed for faceted investigation of corpora, and when used for analyses of frequency, only very frequent or infrequent items will be displayed. Also, quantitative information is encoded using font size, which does not scale well to a global frequency view over a large number of lexical items.

The *Parallel tag clouds* visualisation is essentially a set of connected word clouds, with each cloud being represented as a column, one per frequency list. The size of each word corresponds to how unusually highly occurring the word is given the other lists as a reference. Words that occur in multiple columns are connected by edges, and the presence of edges are indicated by "edge stubs" which hint at the direction and distance to the next occurrence. When a word is hovered, the full edges connecting it to other columns are shown, and a rich tooltip provides additional information about the occurrence of the term in all corpora.

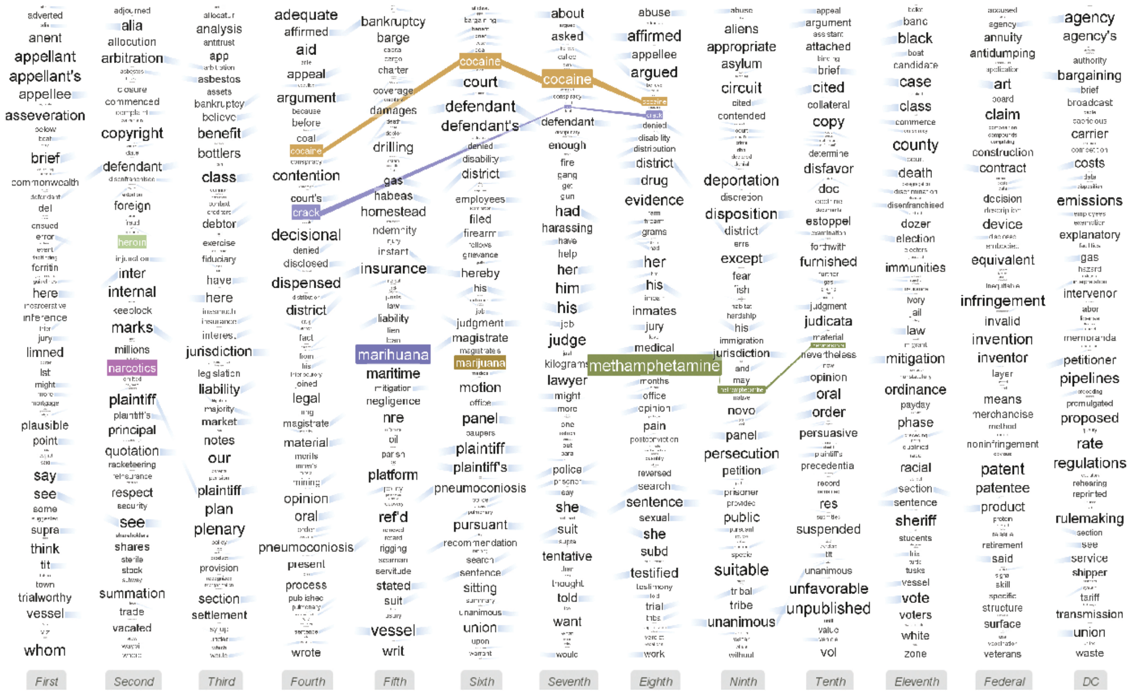


Figure 4.10: A parallel tag cloud revealing the differences in drug prevalence amongst the circuits.

The *Jigsaw* system [Stasko et al., 2007] for visualizing entity connections across documents contains a set comparison tool which uses sloped lines to compare lists of textual items. Figure 4.11 displays two lists of unequal size inked through slope lines. While the slope lines do not encode quantitative information, quantitative differences could be estimated from the visualisation if the lists were in rank order. However, the validity of these comparisons would rely on distributional similarity and comparable list lengths. In word frequency lists, this assumption of comparable length is almost never the case.

This type of set comparison is also similar to another visualisation called *BiSet* [Sun et al., 2016], which suffers from the same drawbacks.

Neither *Parallel tag clouds* nor *Jigsaw* use the positions of items in the lists to

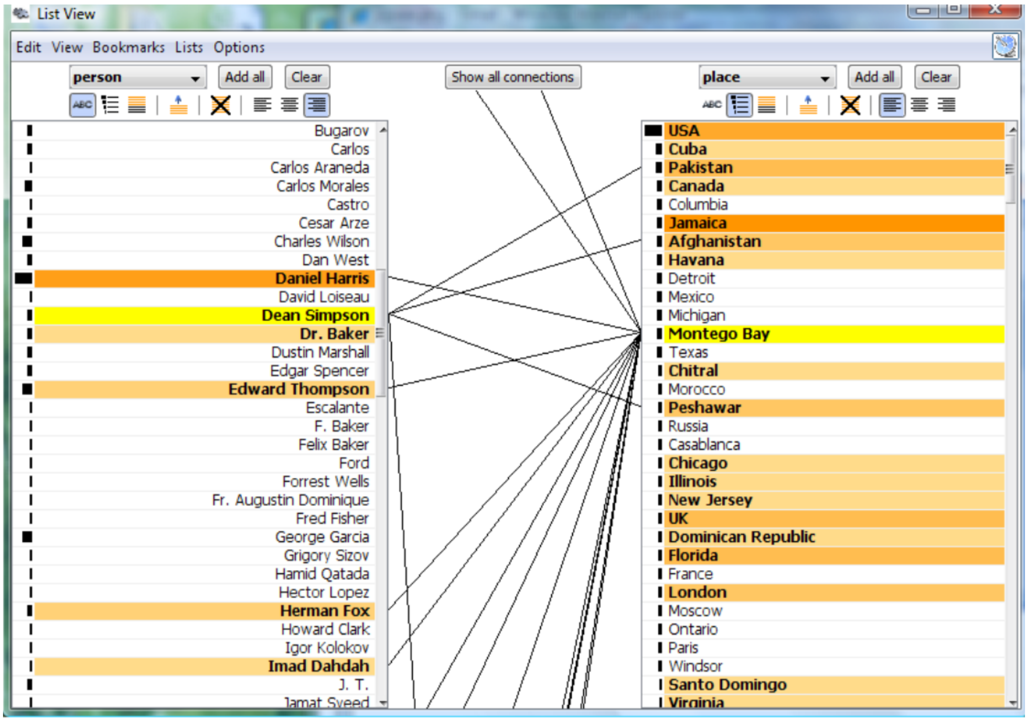


Figure 4.11: Jigsaw List View linking people with places.

encode quantitative information. In *LineUp* [Gratzl et al., 2013], a visualisation for multi-attribute rankings analysis, set items are listed in rank order for multiple data attributes, and the quantitative values are encoded as stacked bars. Figure 4.12 shows *LineUp* where the number of items in each list is the same. The interface is not designed to deal with lists of differing sizes. Even if one of the lists could be made longer than the other, a comparison of word frequency lists in this visualisation would have some drawbacks. For instance, using lists of unequal sizes makes estimating the relative distributional position of the words difficult. In a simple example, an item appearing at position 5 in a list of 10 items should be comparable to an item at position 50 in a list of 100 items. In addition, displaying rank in an equally spaced list distorts the relative frequency differences between lists and over-emphasizes low-frequency words in the shorter list.

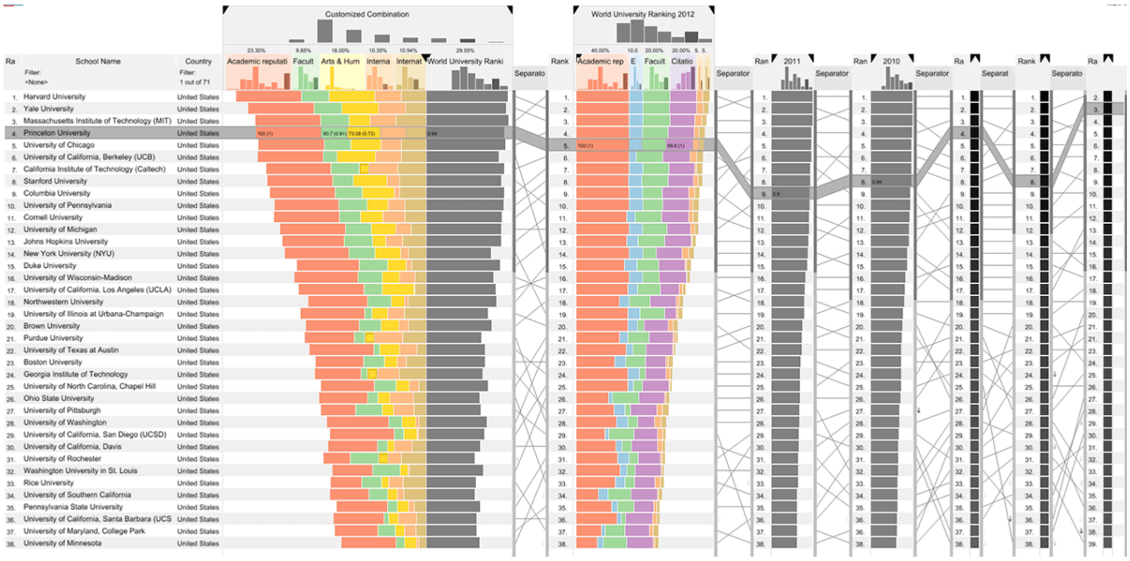

Customized Combination
23.30%
9.85%
16.00%
10.35%
10.94%
29.55%
Ra
School Name
Country
Filter:
<None>
Filter:
1 out of 71
Academic reputati
Facult
Arts & Hum
Interna
Internat
World University Ranki
Separato
Rank
World University Ranking 2012
Academic rep
E
Facult
Citatio
Separator
Ran
2011
2010
Separat
Rank
Ra
1. Harvard University
2. Yale University
3. Massachusetts Institute of Technology (MIT)
4. Princeton University
5. University of Chicago
6. University of California, Berkeley (UCB)
7. California Institute of Technology (Caltech)
8. Stanford University
9. Columbia University
10. University of Pennsylvania
11. Cornell University
12. University of Michigan
13. Johns Hopkins University
14. New York University (NYU)
15. Duke University
16. University of Wisconsin-Madison
17. University of California, Los Angeles (UCLA)
18. Northwestern University
19. University of Illinois at Urbana-Champaign
20. Brown University
21. Purdue University
22. University of Texas at Austin
23. Boston University
24. Georgia Institute of Technology
25. University of North Carolina, Chapel Hill
26. Ohio State University
27. University of Pittsburgh
28. University of Washington
29. University of California, San Diego (UCSD)
30. University of California, Davis
31. University of Rochester
32. Washington University in St. Louis
33. Rice University
34. University of Southern California
35. Pennsylvania State University
36. University of California, Santa Barbara (UCS
37. University of Maryland, College Park
38. University of Minnesota
United States
100 (1)
90.7 (0.91)
73.08 (0.73)
0.94

Figure 4.12: LineUp showing a ranking of the top Universities according to the QS World University Ranking 2012 dataset with custom attributes and weights, compared to the official ranking.

4.4 Conclusion

In this chapter, attempts at mitigating visualisation design validity risk at the data abstraction and encoding levels of the nested model of visualisation design (as described in subsection 2.2.3) were presented. This was achieved by developing a conceptual data model of the concordance list and using it to compare encoding choices in existing KWIC visualisations. The visualisations either don't encode surveyed encode positional frequency or encode it using non-optimal visual variable choices.

Further, visualisations which are useful in the comparison of frequency lists were compared and found to not encode frequencies from unequally sized lists in a manner that makes them directly visually comparable.

In the next chapter(chapter 5), the Concordance mosaic visualisation is presented. I developed this visualisation to address the lack of perceptually efficient KWIC visualisation to represent positional frequencies and statistics.

In chapter(chapter 6), the ComFre visualisation is presented. This visualisation was developed to enable visual comparison of word frequency lists of unequal size.

Chapter 5

Concordance Mosaic Visualisation

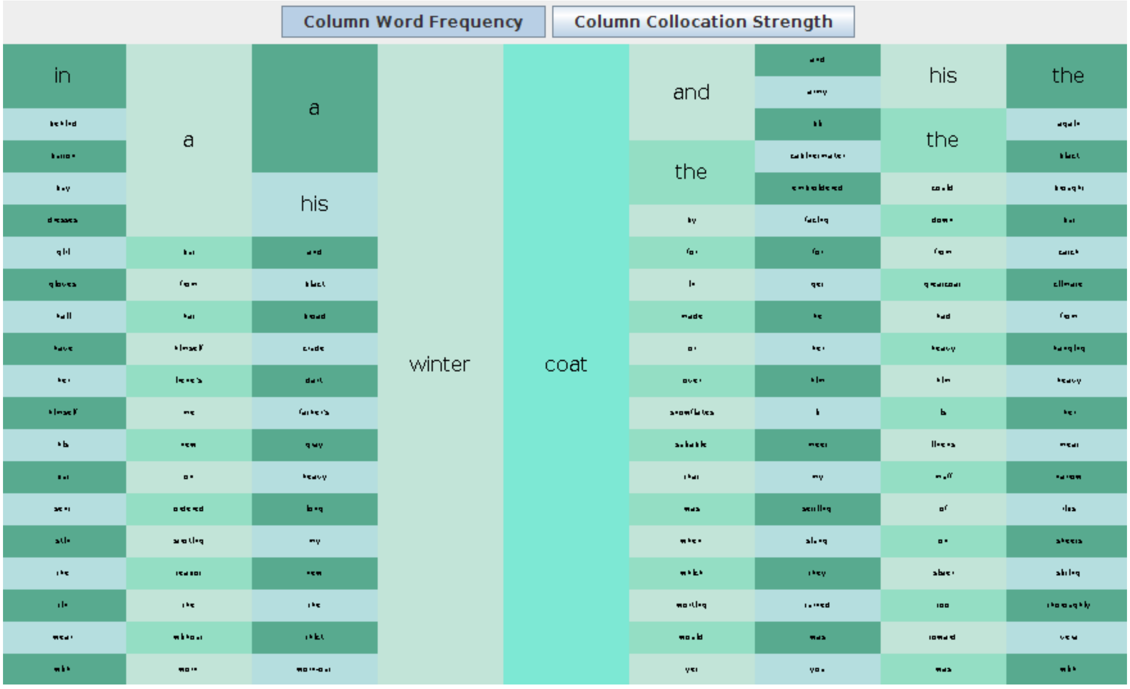


Figure 5.1: Filtered Mosaic for keyword: *Coat* filtered by *winter* at position *keyword* − 1 from Figure 5.8.

The Concordance Mosaic visualisation addresses a number of problems identified during the domain characterization chapter 3. In particular, the difficulty in dealing with positional word frequencies is eased by the use of Mosaic. This chapter begins with an overview of the Concordance Mosaics visual encoding before describing the graph-based abstraction on which the visualisation implementation relies. The visualisation encoding is then described in detail, along with the interaction designs. Finally, issues of time complexity are discussed.

5.1 Mosaic Overview

The *Concordance Mosaic* (Figure 5.1) was created to enable better the quantitative actions identified in the domain characterization chapter 3 via the conceptual data model of the concordance. I tried to overcome the limitations of the reviewed visualisations section 4.2 by mapping visual variables to the conceptual model as efficiently as possible. Using an established visual metaphor, the design also seeks to remain visually similar to the KWIC interface. For example, representing word positions (WPO) on the horizontal axis rather than on the vertical should make the visualisation more appealing to expert users.

The Mosaic renders the concordance as a collection of columns representing word positions (WPOs) relative to the keyword column. Each positional column contains rectangles (mosaic tiles) representing the conceptual model's word token objects. Each rectangle represents all occurrences of a single token at a position relative to the keyword. The position objects and their order is clearly defined using horizontal position, as is the order relative to the keyword. The width of the rectangles is constant, but variation in height is used to show quantitative differences between the WTOs. This is quantitative information's second-highest-ranking variable (length Q 2).

The variable vertical position (O 1) is used for ordering WTOs at a position. This was also done in the previously discussed tree-based visualisations. However, in this case, every position is ordered based on the quantitative values of the WTO, which was not possible in the tree-based visualisations due to the constraint of maintaining the structure (readability) of the concordance lines. So this visualisation trades the linear structure of the text fragments to enhance the perceptibility of the quantitative information. The boxes' ordering and length are perceived together, forming an integral combination of the two variables.

Mosaic also enables visualizing quantitative results of statistical analysis (such as collocation strength) in a positionally aware way. This gives an overview of the concordance list, which seems well-suited to the quantitative actions identified in the task analysis.

Using high-ranking variables required trading off the structure of the concordance lines. Since the underlying abstraction maintains this structure, the visualisation was presented in combination with a KWIC interface in a design pattern of composite visualisation views known as Juxtaposed views [Javed and Elmqvist, 2012]. This composite interface, linked through the concordance graph, enables interactions where the concordance lines which connect to a mosaic tile can be highlighted in the KWIC view. The breakdown of the visual variables for the Concordance Mosaic is found in 4.1. This table shows that while the Concordance mosaic doesn't facilitate the encoding of the concordance lines, it completely renders the position objects using high-ranking visual variables.

5.1.1 Concordance Graph

Several recent renderings of the concordance list, namely Word Trees and Double Trees, are based on an analytical abstraction that splits the concordance list into two context trees. These trees are rooted at the keyword and extend into both contexts. Studying these tree-based concordance visualisations made it apparent that no clearly defined method for linking the left and right context trees has been presented. This is a major drawback, as any visualisation built solely from this abstraction cannot fully reconstruct the concordance lines. To remedy this, I created the concordance graph.

A typical KWIC display is shown in the top window in Figure 5.12. For concordancing, the presentation is such that the keyword is placed at the centre of the line. I 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 for a given position, many different word occurrences (*tokens*) across the concordance lines are simply instances of the same word (*types*). This is illustrated, for instance, by the words “silk”, “and” and “hat” in the left context of the word “coat” on the lines highlighted in red, in Figure 5.12. Since, according to Zipf’s Law [Manning and Schütze, 1999], a small number of word types tend to dominate the distribution of tokens at a particular position, the bulk of the data in $\mathcal{C}$ should consist of such repetitions.

A more economical representation can be devised by taking advantage of the linear structure of C. The approach proposed below 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 specified in Definition 1.

Definition 1. 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, $V_l : V \rightarrow Types$ is a labelling of vertices with words and $E_l : E \rightarrow \mathbb{R}$ is a labelling of edges with word frequency information.

The word frequency labels in E_l 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. sorting each of the word position columns lexicographically.
2. iterate through each column starting from the centre and expanding over L and R, (corresponding to k, with index $i = 0$) and expanding over L and R,
3. creates a vertex $v_{i,j}$ for each type (in row j of column i), labelling the vertex with the appropriate string,
4. 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,

5. 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. I will refer to such edges as *contextual edges*.

An example of a concordance graph created according to this procedure for the word “coat” is shown schematically in Figure 5.2.

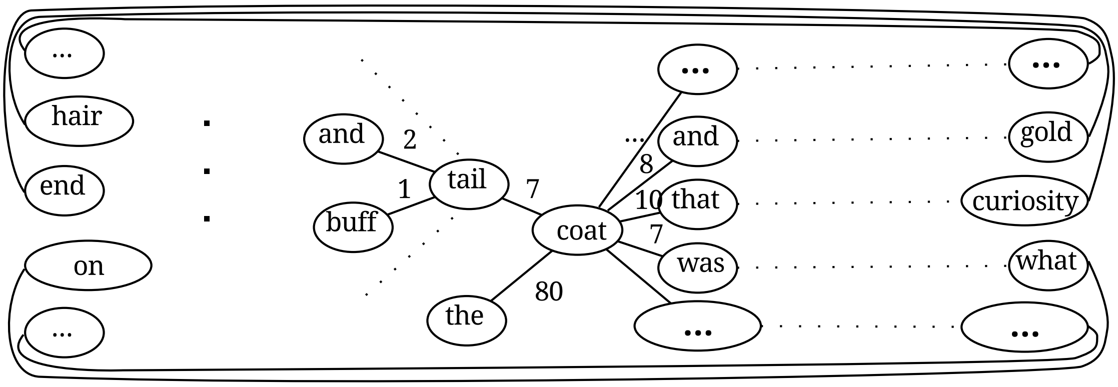


Figure 5.2: Concordance graph for the keyword “coat”.

The formal properties of Concordance graphs could be studied extensively. For the purposes of this thesis, it suffices to state that a) the keyword in a concordance graph can be uniquely identified by its node eccentricity property, b) the graph can serve as an analytical abstraction suitable for a data-state information visualisation model, and c) contextual edges link the left and right contexts to identify full concordance lines. In the following section, I present a particular realisation of a graph-enabled concordance interface.

To aid understanding of the concordance graph, it is possible to conceptualise it as the combination of the left and right context trees (Figure 5.3), with the contextual edges added to enable the reconstruction of concordance lines from the graph (Figure 5.4). Since each of the “word trees” (or context trees) observe the principle of having at most one path between tree vertices, it is clear that the partial concordance lines can be reconstructed up to and after the keyword separately in the case of concatenated context trees Figure 5.3. However, Figure 5.4 shows an example of a concordance graph for a subset of the fragment seen in Figure 2.3 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. Word trees without contextual edges do not make clear which edges in the opposite Word Tree are connected to the branches in this Word Tree. It is only by linking the leaf nodes of each tree that the text fragments are reconstructed.

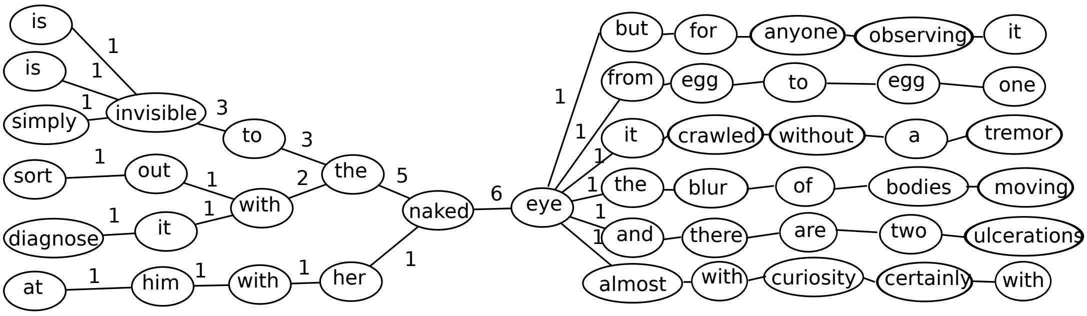


Figure 5.3: Concatenated context trees.

Figure 5.5 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, graph distance is defined $d(v, u)$ 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^l_n, v^r_n) from $\mathcal{G}$, then the keyword vertex can be retrieve through its eccentric property.

Definition 2. 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.

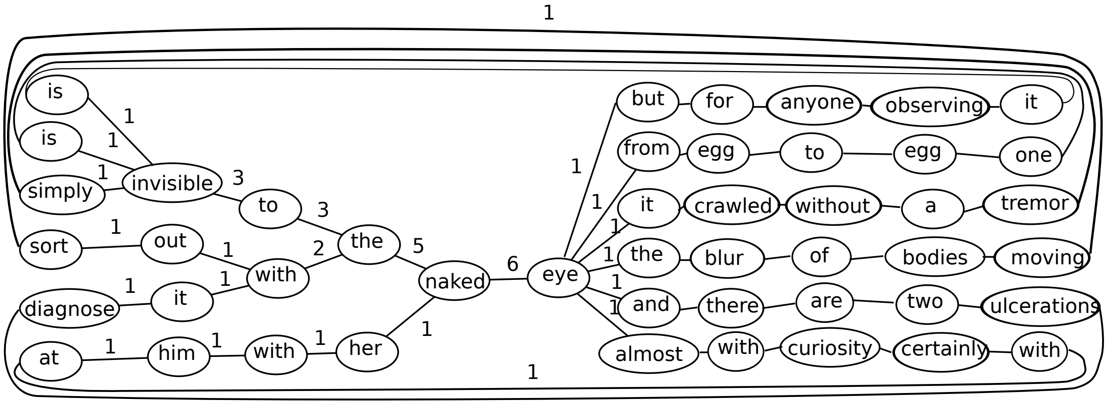


Figure 5.4: Sample concordance graph for the word "eye".

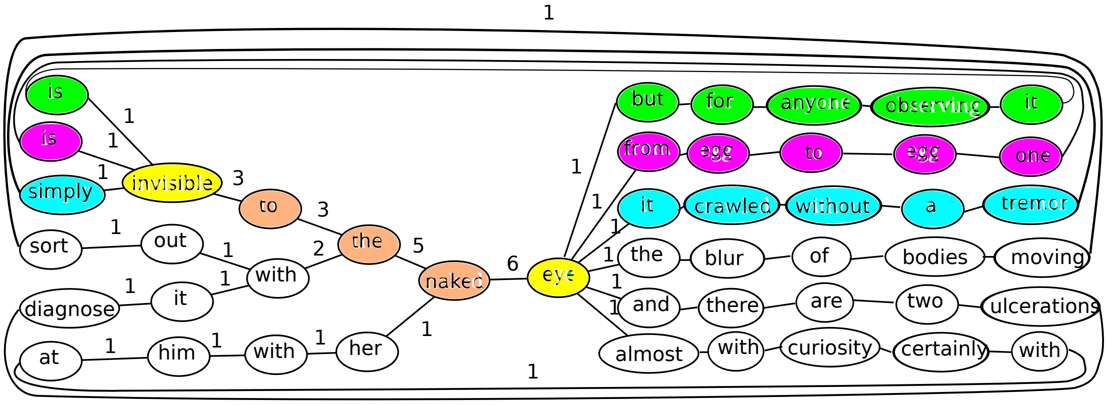


Figure 5.5: Concordance line reconstruction example.

5.1.2 Mosaic Encoding Design

The analysis of corpus linguistic actions revealed that frequency estimation at a word position is a commonly performed action. Observing these positional frequencies at multiple positions simultaneously (frequent combinations) across both contexts is difficult and often required in traditional concordance analysis. Providing a visual technique to facilitate these actions was the focus of the design of the Concordance Mosaic.

Using Chi's and Riedl's reference (data state) model [Chi and Riedl, 1998] , I show the data states and operators through which we create the visualisation (Figure 5.6). I chose to create the visualisation using the Prefuse library since its software architecture is also based on this reference model [Heer et al., 2005].

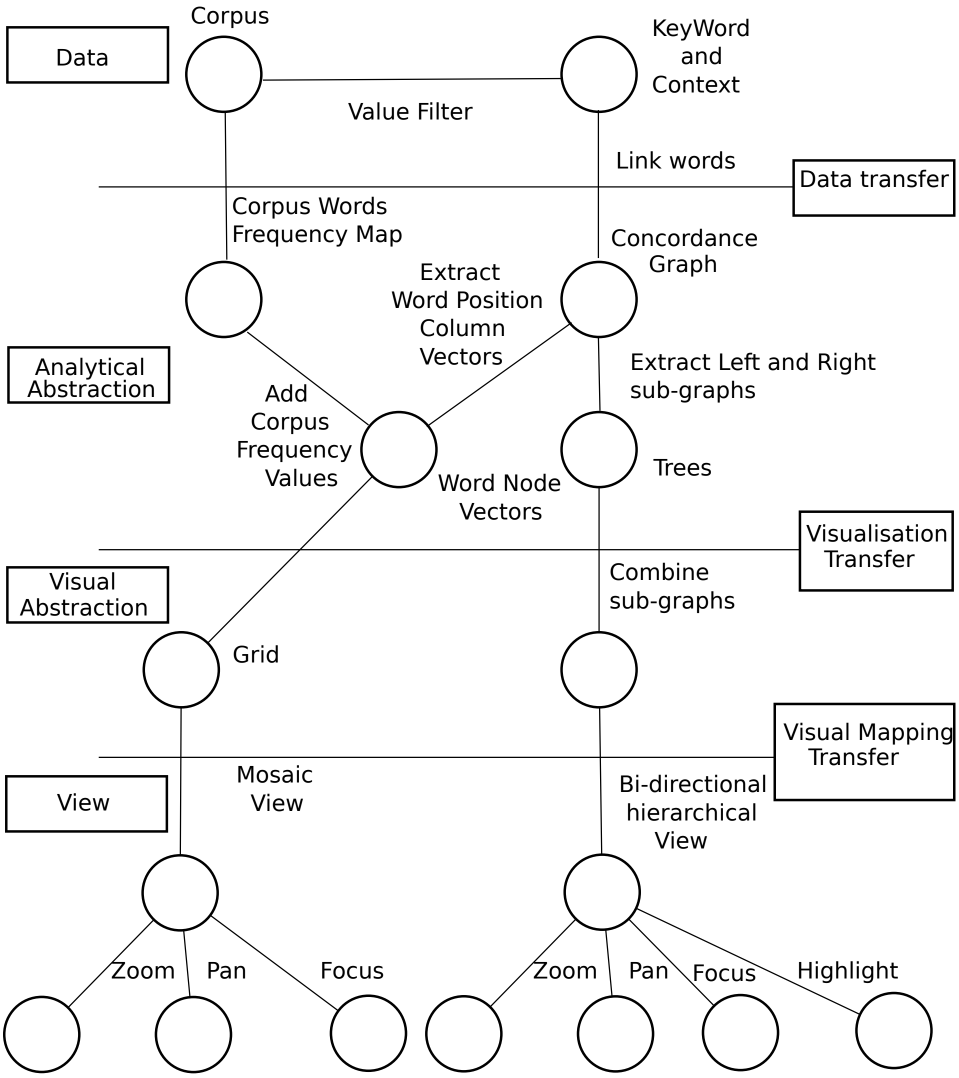


Figure 5.6: Concordance visualisation reference model diagram.

The concordance graph can be transformed into column vectors where each vector represents a position relative to the keyword. These vectors can be easily created by traversing the graph (in any manner) from the keyword node using the edge distance from the keyword node as the position identifier. This graph traversal is designated as a visualisation transfer operator, creating a new data state at the visual abstraction level of the reference model. This data state consists of a collection of vectors containing word objects. These 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. These words (word objects) consist of the word token and a quantitative value representing the frequency with which this word has occurred at this position in the concordance list.

Visual mapping these column vectors to the Concordance Mosaic entails laying out the vectors as columns in a grid according to their word position. Within these columns, each word object is represented by a fixed-width rectangle. The height of each rectangle (word object) is scaled by its frequency at the position and normalised by the required total column height. Within each rectangle, the text attribute of the word object is rendered and scaled by the height of the rectangle. An alternating four-colour scheme differentiates the columns and words, and a fifth colour is used for the keyword column (Figure 5.8). The example shown is for the keyword *coat*.

In this visualisation, the visual variable *position* is used for the word positions, represented by the ordering of the columns, and frequency by ordering the words per column from most frequent (top) to least frequent (bottom). Encoding word position into the visualisation is important for analysis actions and as a reminder of the hidden sentence structures. A second visual variable *length* is used again to represent the frequency of a word at a position. The combination of frequency-ordered columns of tiles and the length of each tile used to represent frequency strongly encodes frequency in the visualisation. Using two visual variables to encode frequency directly results from the importance of frequency-related actions identified in the task analysis.

Filtering quantitative actions were identified in the task analysis. Under the

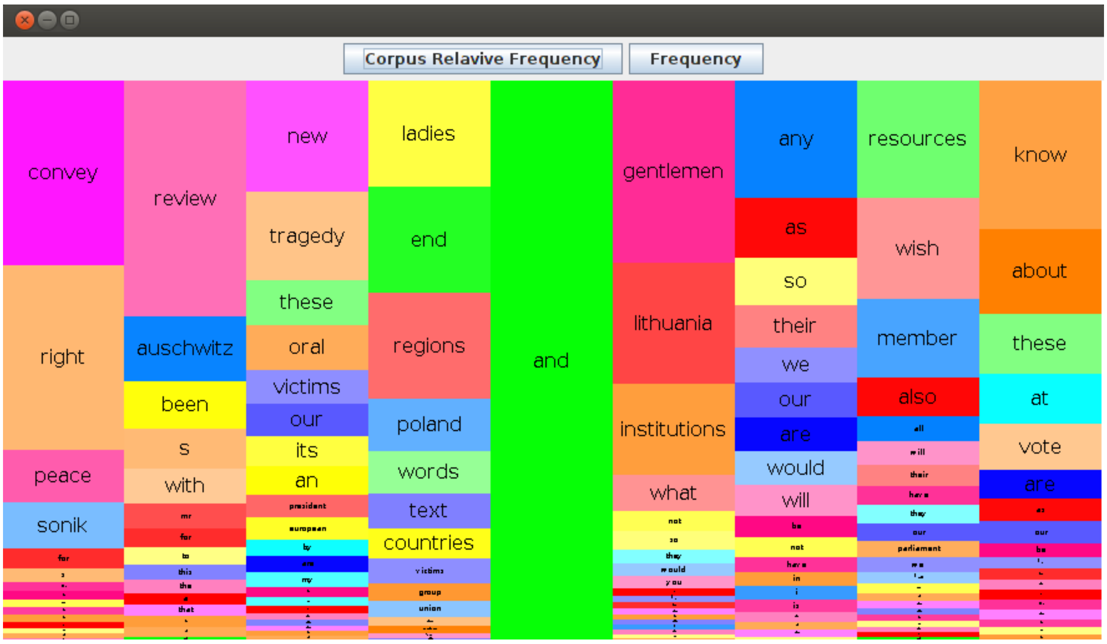


Figure 5.7: Collocation Strength Mosaic coloured by token for keyword *and*.

umbrella of filtering, many different actions were classified. For the Mosaic visualisation, I implemented a simple filter interaction which enables selecting a word at any position to create a filtered Mosaic where words which do not form sentences with the selected positional word and keyword are removed. The Mosaic for the keyword "coat" filtered by occurrences of the word "winter" at position $keyword - 1$ can be seen in Figure 5.1. A right-click interaction was used to trigger the filter.

An additional Concordance frequency view was created, which filtered out stop words. Expert users requested this filtering, and it was an implicit feature of many of the task analysis tasks. (Figure 5.9)

One of the main challenges in the visualisation of textual information is that the text itself is of interest and is often high-dimensional and difficult to map to visual variables effectively. As an example mapping each token in a corpus to a colour would still require the text to be available through some interaction or legend. For example, the analysis of a concordance using the interface presented in Figure 5.7 would be impossible if the text labels were removed. To address this issue, a bifocal distortion interaction [Spence and Apperley, 2013] is applied on mouse over to allow the inspection of the labels of rectangles which are too small to be rendered, an example of this bifocal interaction is shown in Figure 5.8. This scaling down of words with low quantitative values is desirable as it brings the words with the largest quantitative values per position into focus.

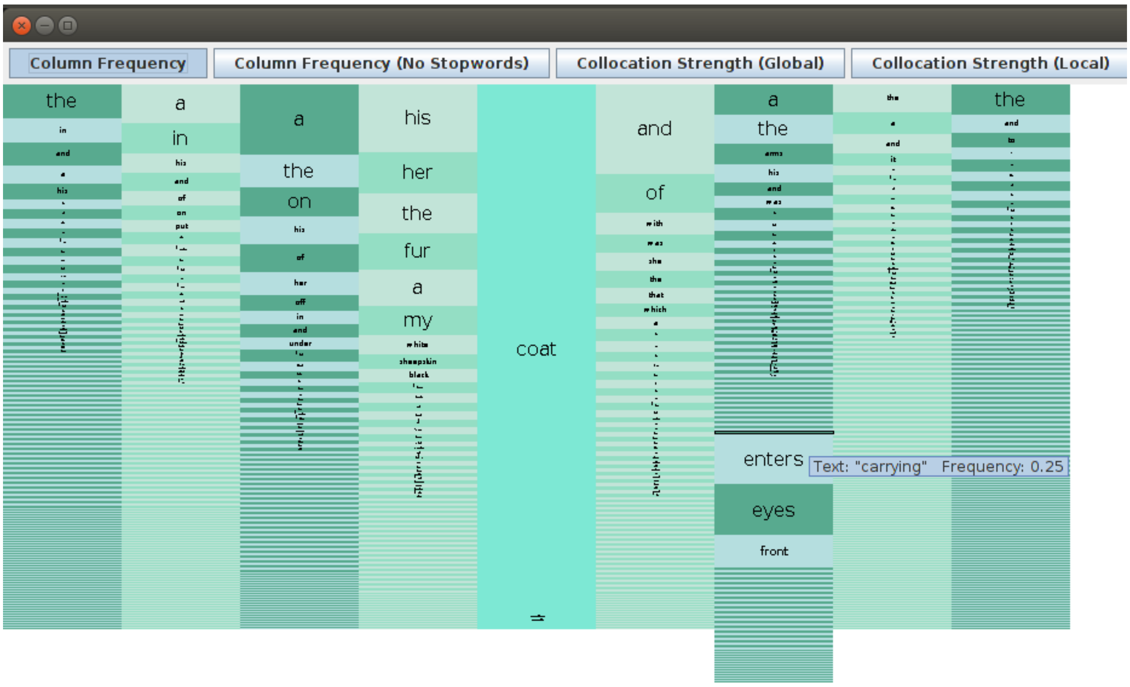


Figure 5.8: Frequency Mosaic for keyword *Coat*. Bifocal interaction and hover tooltip are shown.

The quantitative action *Significant Collocates* describes actions where a statistical measure of the significance or strength of the collocations with the keyword were investigated. These statistical measures are often calculated using external tools [Scott, 2010]. These tools return ordered lists of the significant collocates. Word position relative to the keyword is sometimes used in the calculation of significant collocates, but the output-ordered lists give no information about the positional occurrences. Using the Mosaic visualisation and concordance graph, it is possible to calculate and display similar statistics and overlay the word position information. A simple measure of positional *collocation strength* can be calculated by dividing a word's concordance positional frequency by the frequency of the word in the entire corpus (global frequency).

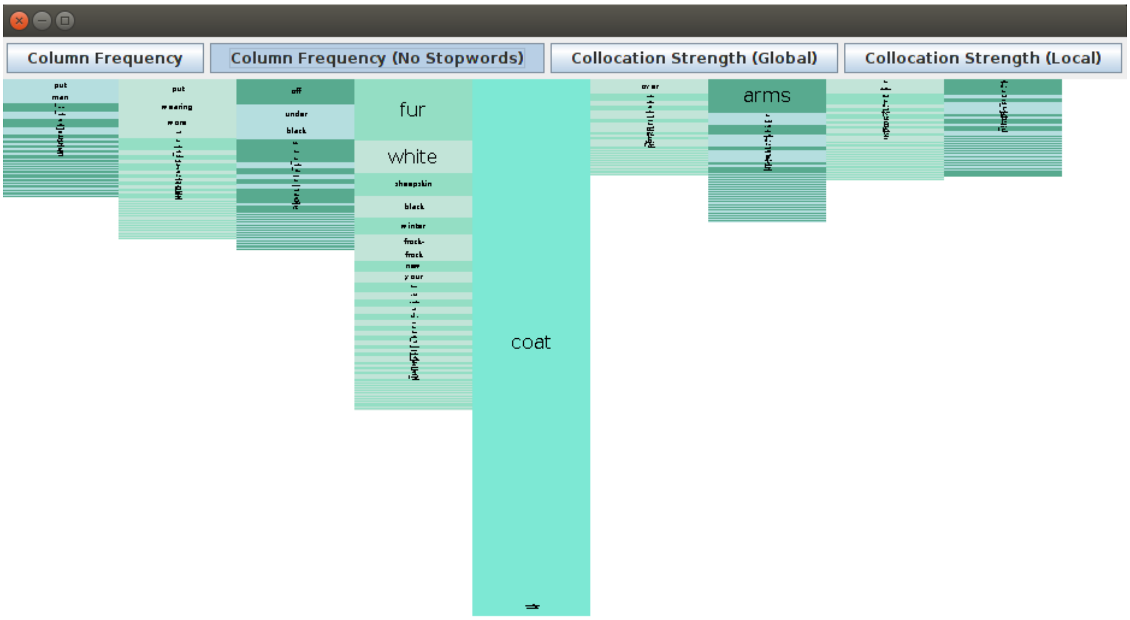
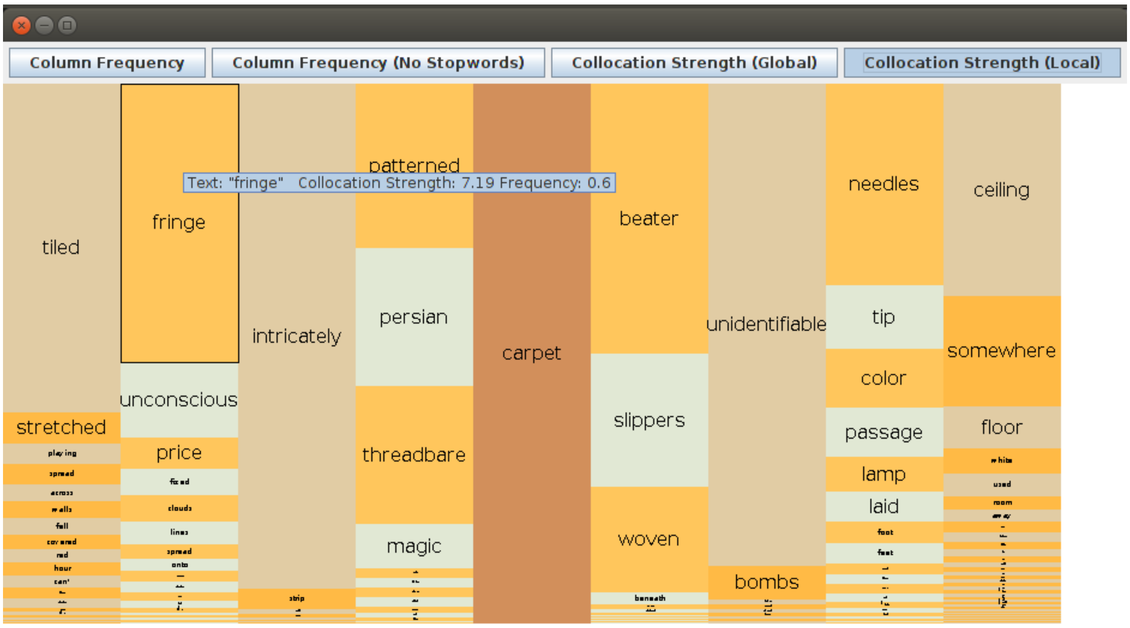


Figure 5.9: Frequency Mosaic coloured by token for keyword *and*.

The collocation strength version of the Mosaic uses a different colour palette (Figure 5.10) than the frequency rendering. The collocation strength version of the mosaic reduces the size of words with a high frequency in the corpus and increases words of lower frequency. For example, comparing Figure 5.8 and Figure 5.12 the words "his", "her" and "the", at position $keyword - 1$, have had a drastic scale reduction during the hyphenated prefixes "frock", "tail" and "rain" have been scaled up. This technique could be used to display many other scores or statistics which relate a keyword to its surrounding context, such as chi-squared or mutual information score. Due to the space-filling design of the Mosaic collocation, score heights should only be compared to other scores at the same position since the total height of each column is held constant and individual word heights are scaled appropriately. This is not an issue for the frequency view, as the number of words at each position is the same as the total number of concordance lines.

An alternative scaling scheme for the collocation strength view is shown in Figure 5.11. Under this scheme, independent of position, each word object can be directly compared to any other word object. This has the effect of adding meaning

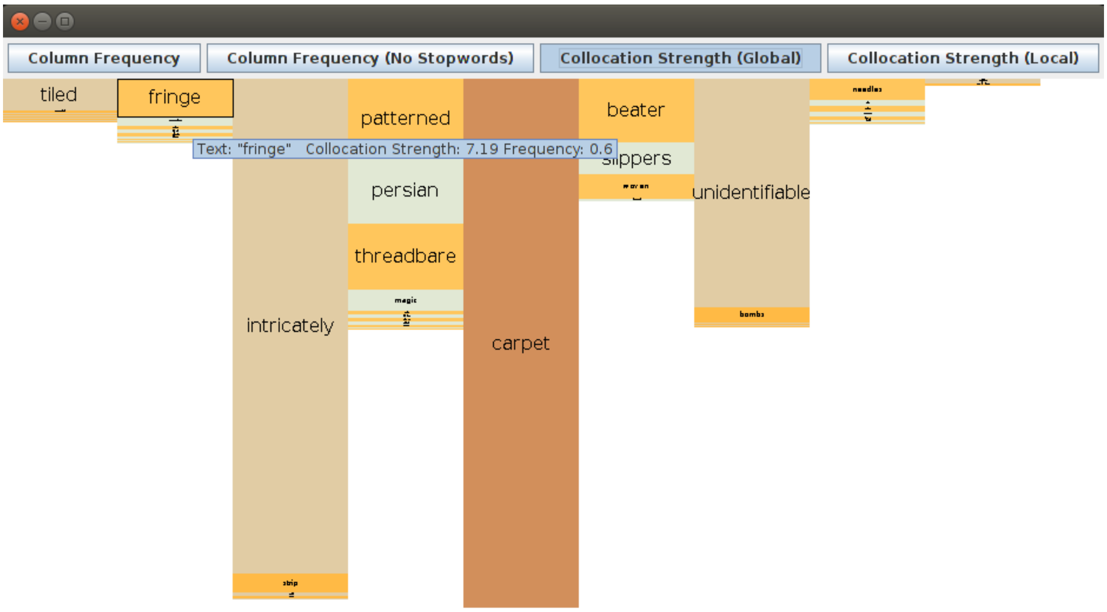


Text: "fringe" Collocation Strength: 7.19 Frequency: 0.6

Figure 5.10: Concordance Mosaic of normalised collocation strength for keyword: *Carpet*.

to the total height of each position column. This additional meaning is that the position columns now signify the total collocation strength at a position, thus giving an overview of which positions contain the strongest collocation strength set of collocates.

The addition of a hover tool-tip reveals the numerical detail of the overview provided by the WTO box heights. The Mosaic interface offers easily perceptible differences in quantitative information, but without a numerical axis, the quantitative information is comparable but not viewable. A tooltip is a message which appears when a cursor is positioned over an icon or, in this case, a mosaic tile. In Concordance Mosaic, a tooltip displays the required detail, and additional information can be made available such as word frequency in the collocation strength view. Comparing the tooltips in Figure 5.10 and Figure 5.11, the difference between the two Collocation strength scaling schemes can be seen, in the Figure 5.11, the rectangle height represents the true value of the collocation strength in relation to all other WTO rectangles.



Text: "fringe"   Collocation Strength: 7.19 Frequency: 0.6

Figure 5.11: Alternative Mosaic Collocation Strength for keyword: *Carpet*.

The *Read Context* action was always required to perform concordance analysis qualitative tasks. While the mosaic technique visually preserves word position, the links between these positions (sentence structure) are lost. However, this sentence structure is still available in the Mosaic data structure thanks to the concordance graph. To show this sentence structure, I chose to have all words which connect, through a contextual edge, to a user-selected word highlighted in white (bottom panel Figure 5.12). This interaction helps examine the context, but words with low frequency or collocation strength often will be too small to observe, as shown in the entire right context in Figure 5.12. To overcome this, we implemented a bifocal distortion [Spence and Apperley, 2013], activated on mouse-over of rectangles below a chosen height threshold. When multiple words at a position form sentences with the selected word, the choice of which highlighted words connect to each other can again become ambiguous. This ambiguity can be addressed by combining the Mosaic and KWIC techniques using the familiar patterns of overview+detail [Shneiderman, 1996] and synchronised views.

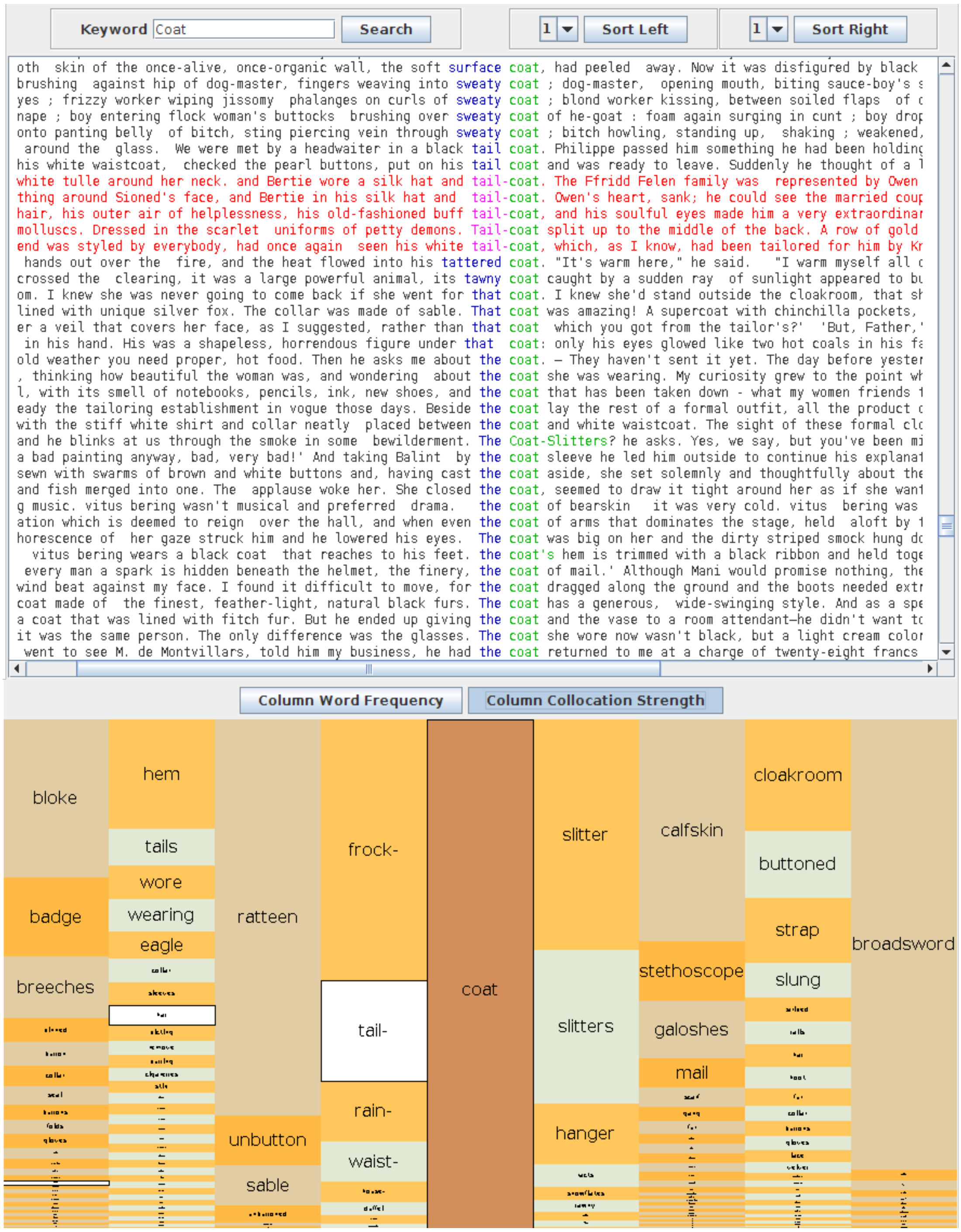
Juxtaposed views [Javed and Elmqvist, 2012], a design pattern of composite visualisation views, was used to design the combined visualisation. This design provides the user with the overview (Mosaic) and detail (KWIC). Since the data is implicitly linked, interaction with either view can affect the other. This is demonstrated by applying a focus/selection interaction on the Juxtaposed view (Figure 5.12). Selecting a word on the mosaic view sorts and scrolls the traditional keyword-in-context view such that the sentences which contain the selected word in the selected column are visible. The selected word is highlighted in pink, and the remaining context words are coloured red. The entire capabilities of Mosaic and KWIC views are available in the Juxtaposed view and are enhanced by linking the two views.

5.1.2.1 Collocation Statistics

As GOK project members used the Mosaic tool, it became apparent that the Collocation Strength view would be more useful if it were based on established collocation statistics such as Mutual Information, z-score, or log-likelihood.

Corpus linguists from the GOK project reported difficulty using the collocation strength view, explaining that the scaling of the Mosiac is not immediately transparent. An explanation of the collocation strength measure was required before an analyst could trust the collocation strength visualisation. In addition, while some GOK researchers reported using this view regularly, they would not be comfortable explaining the technique in academic publications. The analysis would be much easier to explain if established collocation strength measures were available.

Mutual Information (MI), cubed mutual information technique (MI3), z-score and log-likelihood are among the most used techniques for calculating collocation strength in corpus linguistics[McEnery et al., 2006, Manning and Schütze, 1999]. Each collocation strength measure was added to the Mosaic, and the *simple measure* of collocation strength was removed. A Mosaic displaying collocation strength using Z-score for the keyword 'Statesman" can be viewed in Figure 5.13. A Mosaic displaying collocation strength using log-likelihood for the keyword 'Statesman" can



```
oth  skin of the once-alive, once-organic wall, the soft surface coat, had peeled  away. Now it was disfigured by black
brushing  against hip of dog-master, fingers weaving into sweaty coat ; dog-master,  opening mouth, biting sauce-boy's s
yes ; frizzy worker wiping jissomy  phalanges on curls of sweaty coat ; blond worker kissing, between soiled flaps  of c
nape ; boy entering flock woman's buttocks  brushing over sweaty coat of he-goat : foam again surging in cunt ; boy drop
onto panting belly  of bitch, sting piercing vein through sweaty coat ; bitch howling, standing up,  shaking ; weakened,
 around the  glass.  We were met by a headwaiter in a black tail coat. Philippe passed him something he had been holding
his white waistcoat,  checked the pearl buttons, put on his tail coat and was ready to leave. Suddenly he thought of a l
white tulle around her neck. and Bertie wore a silk hat and tail-coat. The Ffridd Felen family was  represented by Owen
thing around Sioned's face, and Bertie in his silk hat and  tail-coat. Owen's heart, sank; he could see the married coup
hair, his outer air of helplessness, his old-fashioned buff tail-coat, and his soulful eyes made him a very extraordinar
molluscs. Dressed in the scarlet  uniforms of petty demons. Tail-coat split up to the middle of the back. A row of gold
end was styled by everybody, had once again  seen his white tail-coat, which, as I know, had been tailored for him by Kr
 hands out over the  fire, and the heat flowed into his tattered coat. "It's warm here," he said.   "I warm myself all c
crossed the  clearing, it was a large powerful animal, its tawny coat caught by a sudden ray  of sunlight appeared to bu
om. I knew she was never going to come back if she went for that coat. I knew she'd stand outside the cloakroom, that sh
lined with unique silver fox. The collar was made of sable. That coat was amazing! A supercoat with chinchilla pockets,
er a veil that covers her face, as I suggested, rather than that coat  which you got from the tailor's?'  'But, Father,'
 in his hand. His was a shapeless, horrendous figure under that  coat: only his eyes glowed like two hot coals in his fa
old weather you need proper, hot food. Then he asks me about the coat. — They haven't sent it yet. The day before yester
, thinking how beautiful the woman was, and wondering  about the coat she was wearing. My curiosity grew to the point wh
l, with its smell of notebooks, pencils, ink, new shoes, and the coat that has been taken down - what my women friends 1
eady the tailoring establishment in vogue those days. Beside the coat lay the rest of a formal outfit, all the product c
with the stiff white shirt and collar neatly  placed between the coat and white waistcoat. The sight of these formal clc
and he blinks at us through the smoke in some  bewilderment. The Coat-Slitters? he asks. Yes, we say, but you've been mi
a bad painting anyway, bad, very bad!' And taking Balint  by the coat sleeve he led him outside to continue his explanat
sewn with swarms of brown and white buttons and, having cast the coat aside, she set solemnly and thoughtfully about the
and fish merged into one. The  applause woke her. She closed the coat, seemed to draw it tight around her as if she want
g music. vitus bering wasn't musical and preferred  drama.   the coat of bearskin   it was very cold. vitus  bering was
ation which is deemed to reign  over the hall, and when even the coat of arms that dominates the stage, held  aloft by 1
horescence of  her gaze struck him and he lowered his eyes.  The coat was big on her and the dirty striped smock hung dc
  vitus bering wears a black coat  that reaches to his feet. the coat's hem is trimmed with a black ribbon and held toge
 every man a spark is hidden beneath the helmet, the finery, the coat of mail.' Although Mani would promise nothing, the
wind beat against my face. I found it difficult to move, for the coat dragged along the ground and the boots needed extr
coat made of  the finest, feather-light, natural black furs. The coat has a generous,  wide-swinging style. And as a spe
a coat that was lined with fitch fur. But he ended up giving the coat and the vase to a room attendant—he didn't want tc
it was the same person. The only difference was the glasses. The coat she wore now wasn't black, but a light cream color
 went to see M. de Montvillars, told him my business, he had the coat returned to me at a charge of twenty-eight francs
```

Figure 5.12: Juxtaposed Concordance interface showing Collocation Strength View and KWIC View for keyword: *Coat*.

be viewed in Figure 5.14. When these Mosaics for Z-Score and log-likelihood are compared, it is apparent that they give different perspectives on the positional keywords of interest.

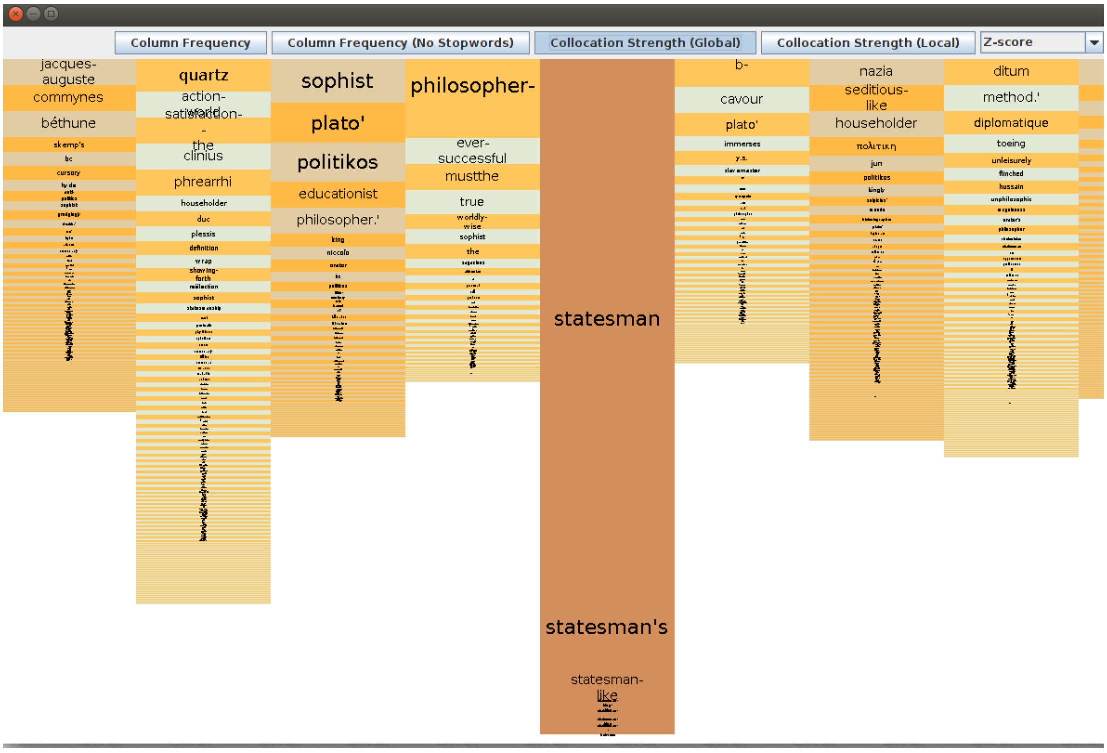


Figure 5.13: Mosaic scaled using Z-score for the keyword "Statesman".

5.1.3 Mosaic Complexity Analysis

Each column of the Concordance Mosaic has at most n tiles, where n is the total number of concordance lines. The number of word positions displayed is defined as w. Sorting and counting the wards at the word positions takes $2wn$ operations. Connecting each word token object at a position to its neighbouring position word token objects simply involves adding concordance line identifiers to each word token object for each connected line. This operation takes wn operations. Connecting the leftmost and rightmost positions to is done implicitly by recording the concordance line identifiers in the previous step. The layout algorithm simply sets each tile's height to the token's count at the position divided by the n. Each tile is placed in the appropriate column and appended in order to the end of the stack of tiles representing the word position. This will take a maximum of wn operations when each word token object at a position is unique. Hence the function representing the runtime is given by:

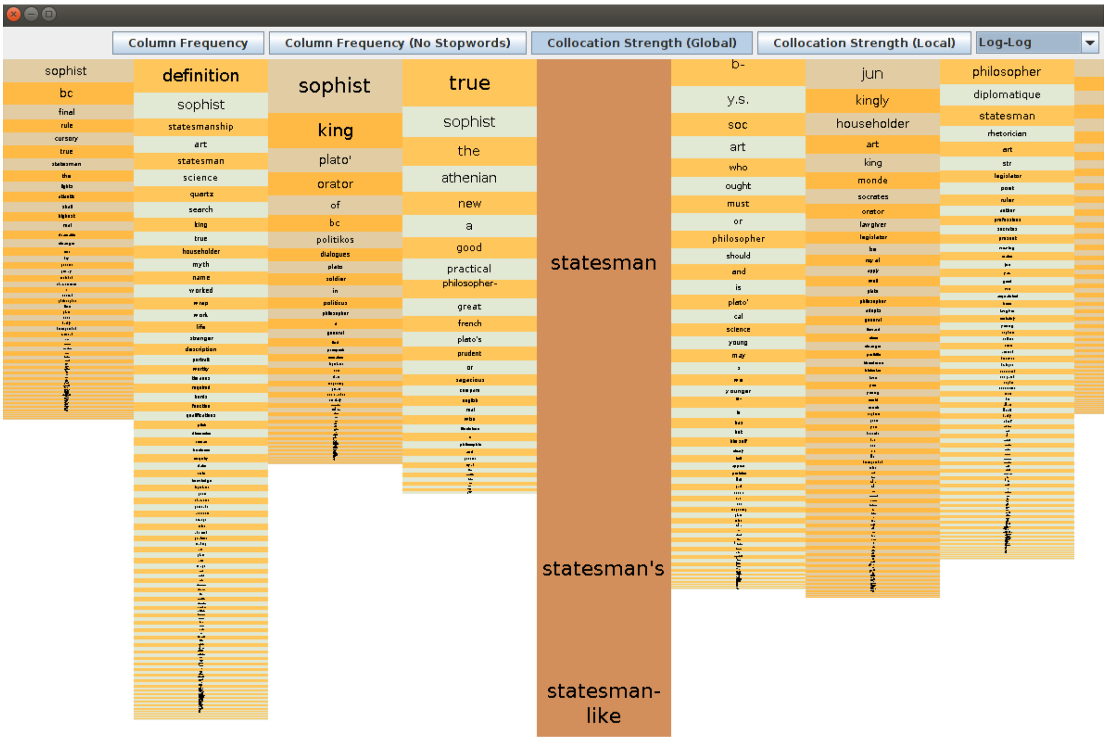


Figure 5.14: Mosaic scaled using log-likelihood for the keyword "Statesman".

$$f(n) = 4wn$$

since $w < n$ the worst-case time complexity and best-case time complexity to generate a Concordance Mosaic are given by

$$O(n) = n$$

$$\Omega(n) = n$$

5.2 Conclusion

This chapter described the first main contribution of the thesis, the Concordance Mosaic visualisation. The encoding choices and design rationale are rooted in the established requirements identified in the domain characterisation chapter (chapter 3), the conceptual data model of concordance lists (section 4.1) and the lessons learned from the structured encoding comparison of KWIC visualisations (section 4.2).

In the next chapter(chapter 6), the second main contribution of the thesis, the ComFre visualisation, is described.

Chapter 6

ComFre Visualisation

This chapter describes an interface design for comparing frequency lists (ComFre).

Frequency lists for corpora usually contain tens of thousands of lexical items. Comparing the entirety of two large lists is challenging when they are presented as ordered text with a corresponding quantity. Tasks which require pattern identification at an overview and detail on demand are good candidates for visualisation.

Instead of comparing two large lists, comparing one small and one large list presents a new range of challenges. How do frequency and rank map from one list to the other?

The ComFre visualisation meets these challenges of list comparison and enables frequency list comparisons valid for distributional similar lists. The ComFre visualisation has been designed to support comparisons of item frequencies between two sets. The primary tasks are those related word frequencies in text corpora. However, the underlying ComFre visualisation can be used for other types of tasks requiring frequency comparisons.

6.1 Requirements

The initial requirements for an analysis tool emerged through discussions with domain experts, and exploration of their corpus analysis workflows. In chapter 3, the need to be able to compare word frequencies was identified as a core task in corpus analysis. During domain characterization, the usage of frequency lists for these comparisons was identified as a requirement. These frequency lists should be from corpora of roughly equal size for the comparisons to be valid.

On discussing this point with the GOK project researchers They indicated that sub-corpus comparison, using word frequencies, formed an important part of exploitative and hypothesis-testing tasks. The limitations of traditional frequency lists and the desire to compare corpora of different sizes gave rise to the initial visualisation design sketches, which were then further refined in consultation with them. The initial requirements established for the visualisation were:

- Comparison of two frequency lists
- Comparing lists of unequal lengths should result in meaningful comparisons.
- Items in both lists should be explicitly linked.
- It should be easy to identify items with high, low and medium differences in relative rank.

An early version of the prototype was shown to a group of experts. Several users asked if they could investigate results for a collection of tokens that they were interested in. Based on this, the search capability was extended to enable the inclusion of multiple keywords (words and sub-strings) in a single search.

6.2 Visual Encoding Design

Since frequencies are quantitative data, the most accurate *visual encoding* [Bertin, 1983] according to [Cleveland and McGill, 1985] should be position, followed by length and slope. Furthermore, as pointed out above, the area does not provide a very accurate encoding for word frequencies in the case of word clouds.

Therefore, decided to design a visualisation that keeps the position (order) element of word lists, which most linguistic users are familiar with and utilize slope elements of slope charts and the length element of histograms [Tufte, 2001]. A slope chart also allows information to be shown in both integrated and separated manner. As Tufte puts it, *"... integrated through its connected content, [and] separated in that the eye follows several different and uncluttered paths in looking over the data..."* [Tufte, 2001].

Figure 6.1 shows an early sketch of the ComFre visualisation. The words from the two corpora are shown vertically on the left and right, in descending frequency orders in each corpus. The histograms of the word frequencies are shown in blue and green for each corpus. The slope lines of the selected words show their relative position, in terms of their frequency, in the two corpora.

It should be pointed out here that one of the problems in comparing relative frequencies of words are that, in natural languages, words are distributed roughly according to Zipf's Law, which states that the n$^{\text{th}}$ most frequent word occurs approximately $\frac{1}{n}$ times as frequently as the most frequent word. For instance, 6.2 shows the overall frequency distribution for the TEC corpus [Luz, 2011], a collection of translated texts widely used in corpus-based translation studies. Comparing word frequencies of different-sized corpora may lead to unfair comparisons. Fortunately, however, the size of the vocabulary (i.e. the size of the set of items to be compared,

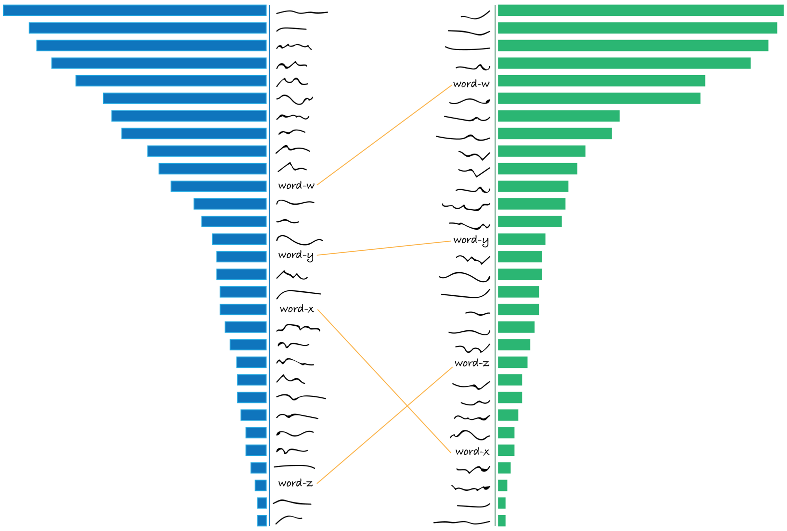


Figure 6.1: Sketch of the ComFre visualisation.

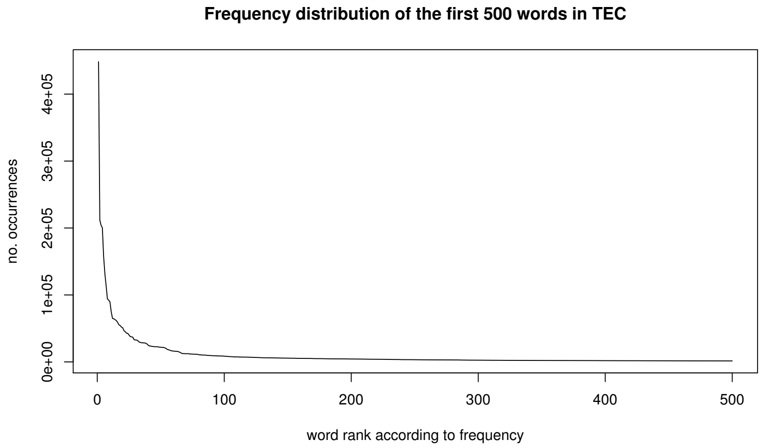


Figure 6.2: Zipfian distribution of the TEC corpus [Luz, 2011].

$|v|$), grows sub-linearly in the size of the text: $|v| = O(n^{\beta})$, where $0 < \beta < 1$ and n = text size (Heaps Law [Luz, 2011]). Therefore, as the sub-corpora to be compared grow, the problem of unfair comparisons due to the Zipfian nature of texts is mitigated. By displaying the position of the words on the distribution (represented by a histogram or a smoothed contour, for instance) as well as their relative rank orders, ComFre allows the user to visualize the most important elements of word frequency comparison at once.

6.3 Encoding Implementation

An interactive version of ComFre was created to exemplify our design. 6.3 shows the interface of this prototype, with two sample large subcorpora from the GOK corpus(pre-modern English subcorpus on the left and modern Englishsubcorpus on the right) loaded to allow comparisons of their word frequencies. ComFre displays the word frequency distributions for both corpora under examination. The distribution plots use logarithmic axis scaling. The x-axis represents the word count in the corpus, while the y-axis displays the rank of each word in the corpus. Under this scaling scheme, Zipfian distributions should appear to be approximately linear.

By fitting the distributions to the same length axis, word position within the

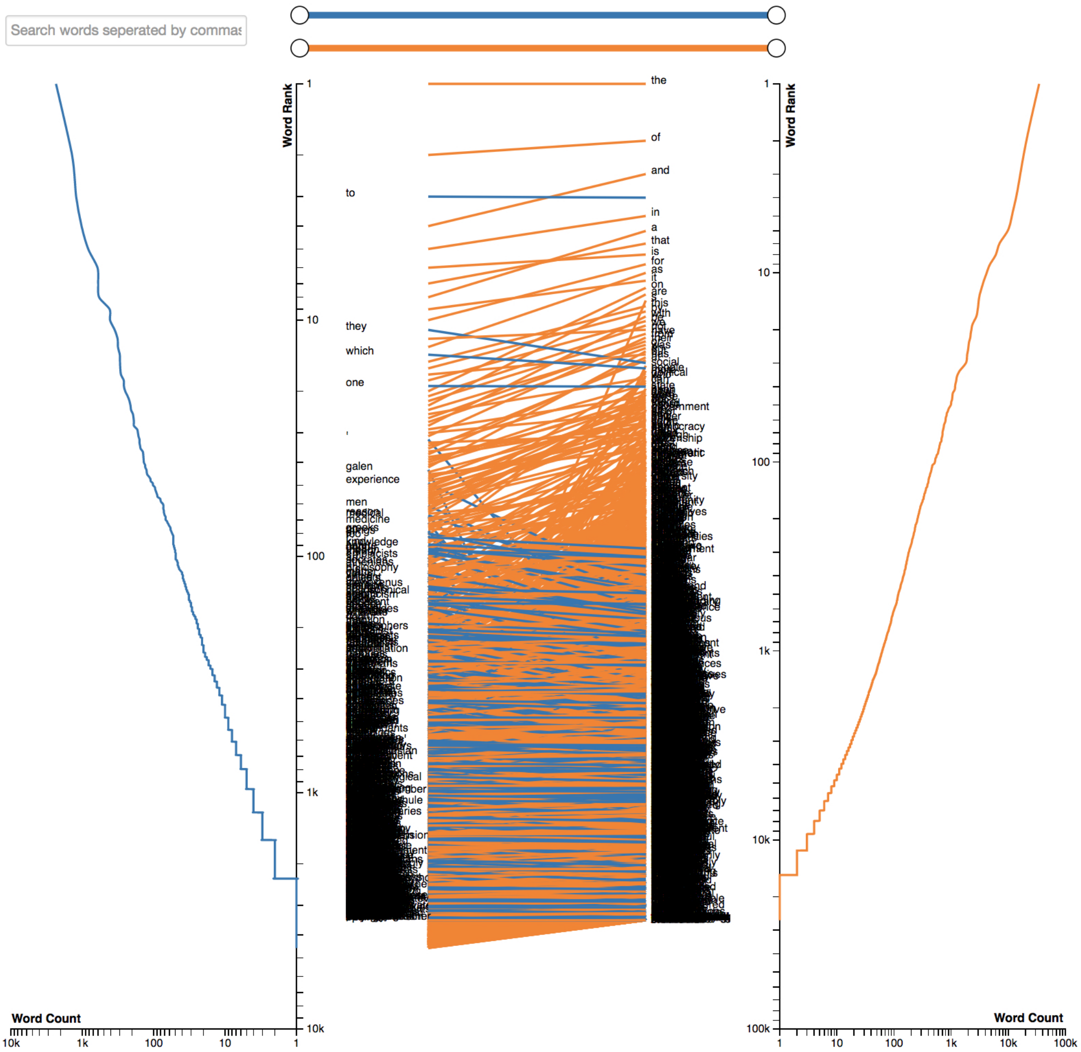


Figure 6.3: Interface of the ComFre prototype.

distributions can be compared sensibly for corpora of vastly different sizes. The corpora of interest should have similarly shaped distributions, so scaling the distributions and visualizing the difference between the word frequency profiles provides a suitable means of corpus comparison. Displaying the distributions of each corpus also gives a visual cue, even if the frequency distributions do not visually match.

The difference between a word's distributional position (scaled probability) in each corpus is displayed using a sloped line. The slope and colour of each line encode which corpus contains a particular word more frequently. Due to the log-linear nature of the distributions, under this scaling scheme, any proportional changes in position along the y-axis are equivalent. In other words, lines of equal slope represent

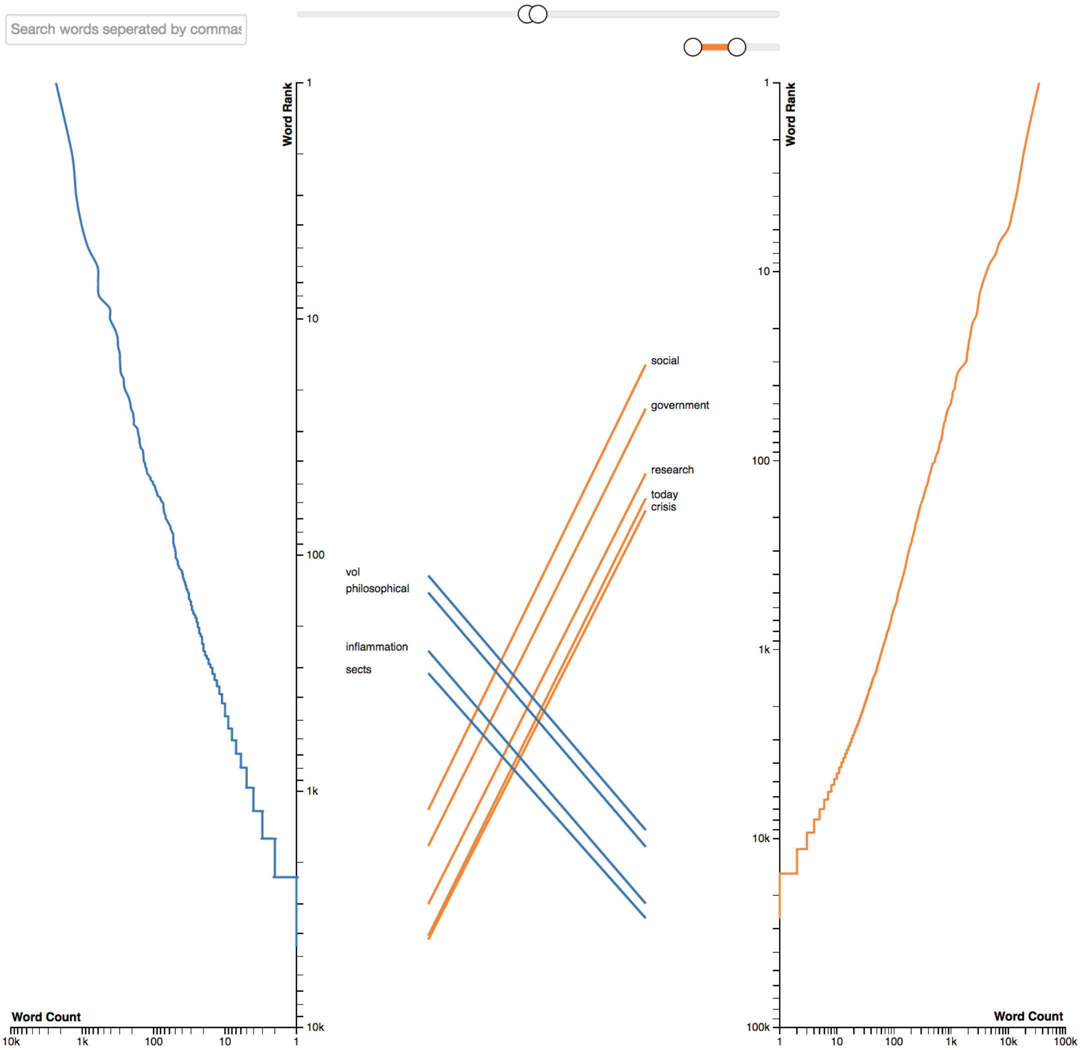
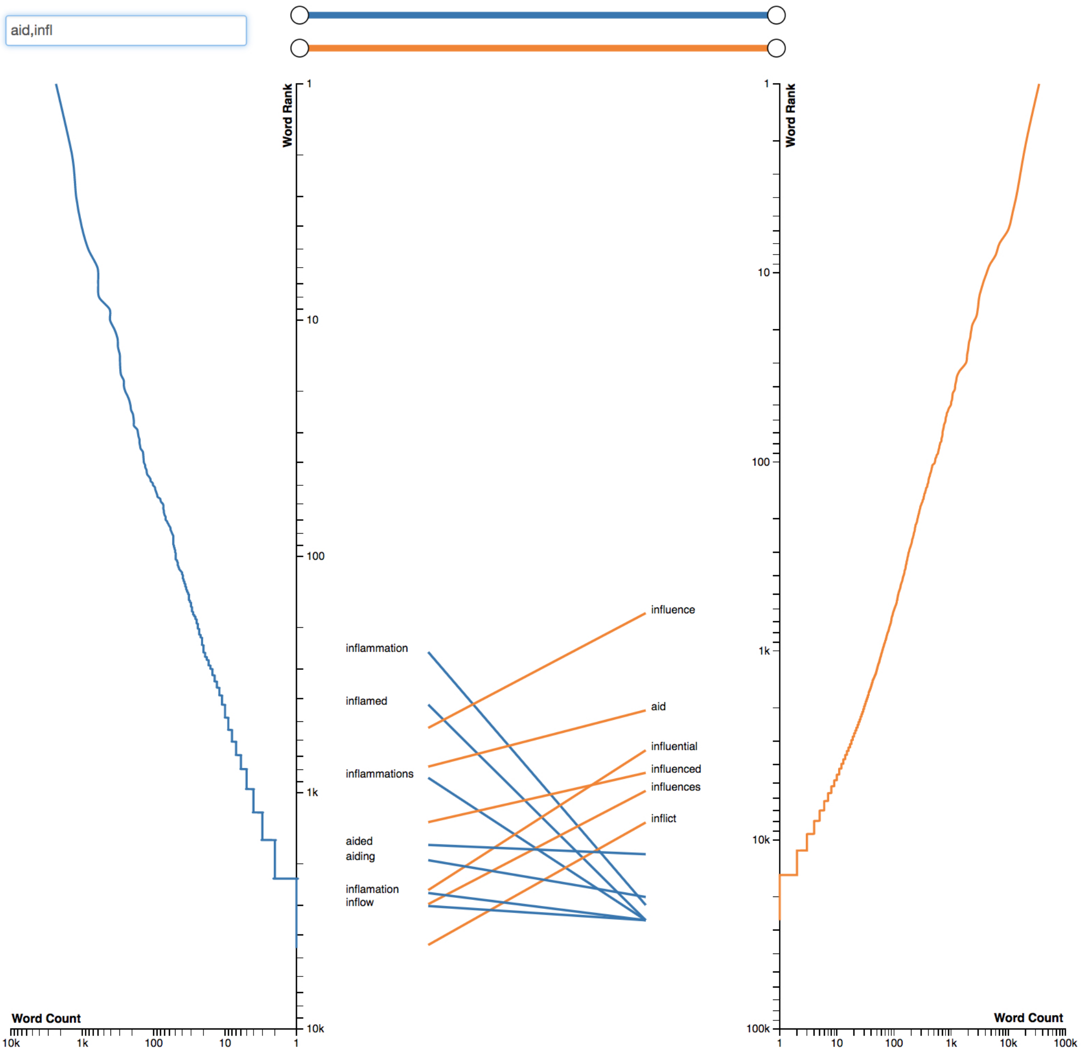


Figure 6.4: Filtering word frequency comparisons using range selections.

an equivalent change in the distributional position of a word between corpora.

The words are rendered as text labels at the higher end of each slope line. The choice of scaling means that these text labels will render very close together and on top of each other when the entire corpus of words is displayed at once (as shown in Figure 6.3). While at this level of analysis, a general sense of the variation in word frequency can be detected, the details are hidden, and user interaction is required for any fine-grained analysis. This follows the well-known visual information-seeking mantra “Overview first, zoom and filter, then details-on-demand” [Shneiderman, 1996, p. 2].

ComFre is designed to facilitate hypothesis testing and exploration between the



aid,infl

Figure 6.5: Filtering word frequency comparisons using word strings.

word frequencies of two corpora, with built-in interaction tools to support these goals. Two range sliders are provided for filtering out lines and labels (the blue and orange sliders at the top, between the y-axes, in Figure 6.3). These sliders let the user select the range of slopes which are visible. For instance, selecting a range at the far right of a slider displays the lines with the steepest slopes. 6.4 shows an example of this range manipulation and its potential for allowing the comparison of words with more specific frequency changes. This is useful for looking at the words that change the most or least between the two corpora. Of course, it is also sometimes necessary to find frequency changes for specific word(s). ComFre provides a keyword(s) Search mechanism. Comma-separated words (or partial word strings) can be typed in the search box (shown in the top left-hand side of Figure 6.3). As a search term is typed, slope lines are filtered out to display only lines which match a continuation of the current search string. 6.5 shows the result of filtering using two keyword strings (”aid” and ”infl”). One example of a use case for this type of filtering would be to see whether or not a topic, modelled as a collection of words, is more prevalent in one of the two corpora. One should also note that although the logarithmic transformation and scaling allow for more a sensible interpretation of the variations between corpora, it also de-emphasises the absolute word frequencies. Therefore, to make the absolute frequency values available to the user, a hover interaction on the slope-lines has been provided in the prototype.

As a demonstrative linguistic use case, one could imagine that a simplistic reduction of a work-flow using the tool would begin by using the range sliders to get an overview of the variation between two corpora and getting a sense of some topic that is more prevalent in one corpus. Compiling a word list that represents a topic, and then filtering using these words can then test the hypothesis.

The ComFre prototype has been implemented as a single-page web application using the JavaScript visualisation framework *D3.js* [Bostock et al., 2011]. It is included in the GOK corpus browser and works well with frequency lists extracted from the GOK corpus. It also allows users to load their own pairs of datasets to be visualized. These can be any comma-separated list of items (e.g. words) and their frequencies. The system then processes the datasets to calculate the slope lines before displaying the resulting visualisation. For large datasets, the run time increases as the number of lexical items grows. The operation of linking each item in the frequency lists has a worst-case time complexity of $O(n^2)$ where n is the length of the longer of the two frequency lists being compared. For each item in the first list, we must find the corresponding item in the second list which is unsorted. In the current version, a delay of approximately thirty seconds is required for one hundred thousand unique tokens before fully rendering the visualisation. However, pre-loaded corpora can make use of pre-calculated data to display the resulting visualisation almost instantly.

The challenges associated with visualizing a large number of lexical items are not limited to the data processing required. Rendering and interacting with a large number of visual items also affects performance and usability. The optimum static visual encoding may not lend itself to a usable implementation. Design decisions based on reducing the number of visual items had to be made. For instance, the initial design displayed the distributions as bar charts of tokens. Replacing these bar charts with two distribution lines removed at least two-thirds of the visual items required by the initial design.

6.4 Conclusion

This chapter described the first second main contribution of the thesis, the ComFre visualisation. The encoding choices and design rationale are rooted in the established requirements identified in the domain characterisation chapter (chapter 3), and the lessons learned from the structured encoding comparison of KWIC visualisations (section 4.3).

In the next chapter(chapter 7), a laboratory evaluation of the Concordance Mosaic visualisation is presented. This evaluation can help mitigate validity threats in the visual encoding choices and domain characterisation efforts concerning concordance visualisation.

Chapter 7

Evaluation

This Chapter details a laboratory study of the Concordance Mosaic interface to evaluate its performance on quantitative concordance tasks compared to the KWIC interface, which is the standard tool of concordance analysis.

7.1 Concordance Mosaic Laboratory Study

This study was designed to compare the performance of the three interfaces (Concordance Mosaic, KWIC Interface and the Juxtaposed Interface). The interfaces were developed to a point where I believe usability would not significantly affect the evaluation of the Mosaic technique. A small heuristic evaluation and pilot study were used to refine the interfaces to achieve this usability standard.

The null hypothesis in this study is that there is no significant difference in performance between the interfaces. Performance was measured by the speed and accuracy with which participants completed corpus analysis tasks. These tasks were created with the quantitative actions found during the task analysis in mind. Each participant attempted to answer five questions using each interface. The interfaces were presented in an order that was randomised and balanced across participants for every possible combination of interface orderings.

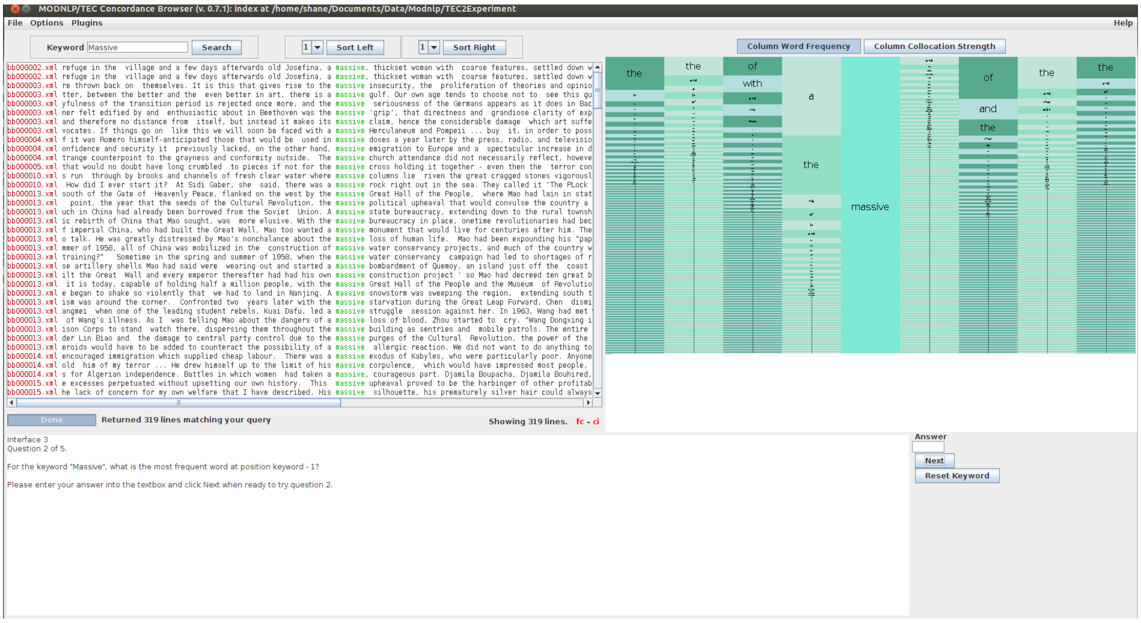
For this study, I recruited thirty-six participants ($N = 36$) from the student population via an online university noticeboard and mailing list. Since the study evaluates performance on quantitative tasks, I decided previous experience with concordance tools or corpus analysis would not be a prerequisite for participation. A pilot study was run with two additional participants. This was done to determine which areas of the interfaces, and any terminology, participants may have difficulty with. Informed by this pilot, a tutorial was designed. The tutorial took approximately ten minutes to complete. It was given to each participant immediately before they participated. In this tutorial, each of the required interface features was introduced and explained, any linguistic concepts required were also clarified, and a researcher was available to answer questions.

7.1.1 Experimental Setup

The Software I created to conduct the experiment consists of four major elements: the KWIC interface, the Concordance Mosaic interface, the question box and the answer box. The question box appears at the bottom left of the software (Figure 7.1) and is simply a text area into which the questions and instructions are rendered. The answer box (bottom right Figure 7.1) contains a button for proceeding to the next question, a button for resetting the software to the question's original state and a text box for the participant to enter the answer. The KWIC and Mosaic interfaces will be present when the software displays the Juxtaposed interface. However, when displaying either interface alone, the space where the other usually resides will be blank.

The participants were asked the same five questions on each of the three interfaces. The keywords about which they were being asked were different for each interface they encountered. I used three sets of keywords. The combination of keyword set permutations and interface orderings was balanced across the participants. This means I have an equal number of participants who used each keyword set, and I have an equal number of participants who used the interfaces in each possible order.

The selection of the keywords for each of the five questions was done in such a way as to standardise the difficulty of the question using the KWIC interface. For



```
bb000002.xml refuge in the  village and a few days afterwards old Josefina, a massive, thickset woman with  coarse features, settled down w
bb000002.xml refuge in the  village and a few days afterwards old Josefina, a massive, thickset woman with  coarse features, settled down w
bb000003.xml re thrown back on  themselves. It is this that gives rise to the massive insecurity, the  proliferation of theories and opinio
bb000003.xml tter, between the better and the  even better in art, there is a massive gulf. Our own age tends to choose not to  see this gu
bb000003.xml yfulness of the transition period is rejected once more, and the massive  seriousness of the Germans appears as it does in Bac
bb000003.xml ner felt edified by and  enthusiastic about in Beethoven was the massive 'grip', that directness and  grandiose clarity of exp
bb000003.xml and therefore no distance from  itself, but instead it makes its massive claim, hence the considerable damage  which art suffe
bb000003.xml vocates. If things go on  like this we will soon be faced with a massive Herculaneum and Pompeii ... buy  it, in order to poss
bb000004.xml f it was Romero himself-anticipated those that would be  used in massive doses a year later by the press, radio, and televisio
bb000004.xml onfidence and security it  previously lacked, on the other hand, massive emigration to Europe and a  spectacular increase in d
bb000004.xml trange counterpoint to the grayness and conformity outside.  The massive church attendance did not necessarily reflect, howeve
bb000005.xml that would no doubt have long crumbled  to pieces if not for the massive cross holding it together - even then the  terror con
bb000010.xml s run  through by brooks and channels of fresh clear water where massive columns lie  riven the great cragged stones vigorousl
bb000010.xml  How did I ever start it?  At Sidi Gaber, she  said, there was a massive rock right out in the sea. They called it 'The PLock
bb000013.xml south of the Gate of  Heavenly Peace, flanked on the west by the massive Great Hall of the People,  where Mao had lain in stat
bb000013.xml   point, the year that the seeds of the Cultural Revolution, the massive political upheaval that would convulse the country a
bb000013.xml uch in China had already been borrowed from the Soviet  Union. A massive state bureaucracy, extending down to the rural townsh
bb000013.xml ic rebirth of China that Mao sought, was  more elusive. With the massive bureaucracy in place, onetime revolutionaries had bec
bb000013.xml f imperial China, who had built the Great Wall, Mao too wanted a massive monument that would live for centuries after him. The
bb000013.xml o talk. He was greatly distressed by Mao's nonchalance about the massive loss of human life.  Mao had been expounding his "pap
bb000013.xml mmer of 1958, all of China was mobilized in the  construction of massive water conservancy projects, and much of the country w
bb000013.xml training?"   Sometime in the spring and summer of 1958, when the massive water conservancy  campaign had led to shortages of r
bb000013.xml se artillery shells Mao had said were  wearing out and started a massive bombardment of Quemoy, an island just off the  coast
bb000013.xml ilt the Great  Wall and every emperor thereafter had had his own massive construction project ' so Mao had decreed ten great b
bb000013.xml  it is today, capable of holding half a million people, with the massive Great Hall of the People and the Museum  of Revolutio
bb000013.xml e began to shake so violently that  we had to land in Nanjing. A massive snowstorm was sweeping the region,  extending south t
bb000013.xml ism was around the corner.  Confronted two  years later with the massive starvation during the Great Leap Forward, Chen  dismi
bb000013.xml angmei  when one of the leading student rebels, Kuai Dafu, led a massive struggle  session against her. In 1963, Wang had met
bb000013.xml  of Wang's illness. As I  was telling Mao about the dangers of a massive loss of blood, Zhou started to  cry. "Wang Dongxing i
bb000013.xml ison Corps to stand  watch there, dispersing them throughout the massive building as sentries and  mobile patrols. The entire
bb000013.xml der Lin Biao and  the damage to central party control due to the massive purges of the Cultural  Revolution, the power of the
bb000013.xml eroids would have to be added to counteract the possibility of a massive  allergic reaction. We did not want to do anything to
bb000014.xml encouraged immigration which supplied cheap labour.  There was a massive exodus of Kabyles, who were particularly poor. Anyone
bb000014.xml old  him of my terror ... He drew himself up to the limit of his massive corpulence,  which would have impressed most people,
bb000014.xml s for Algerian independence. Battles in which women  had taken a massive, courageous part. Djamila Boupacha, Djamila Bouhired,
bb000015.xml e excesses perpetuated without upsetting our own history.  This  massive upheaval proved to be the harbinger of other profitab
bb000015.xml he lack of concern for my own welfare that I have described. His massive  silhouette, his prematurely silver hair could always
```

Returned 319 lines matching your query

Showing 319 lines. fc - ci

Interface 3
Question 2 of 5.

For the keyword "Massive", what is the most frequent word at position keyword - 1?

Please enter your answer into the textbox and click Next when ready to try question 2.

Figure 7.1: Experimental Setup.

example, question two asks "*For the keyword KEYWORD, what is the most frequent word at position keyword - 1¿*'. The three keywords chosen for this question were *Wealthy*, *Daylight* and *Massive*. These keywords all returned a concordance with approximately three-hundred concordance lines. The most frequent word at position $keyword-1$ occurs with a frequency of between twenty-six and twenty-seven percent, and the second most frequent word at position $keyword - 1$ occurs with a frequency of twenty to twenty-two percent.

Similarly, question one is phrased exactly like question two, but the chosen keywords change the distribution of the words at the position of interest. In this case, there are again three-hundred lines in the concordance but the frequencies of the most common and second most common words at position $keyword - 1$ is approximately forty percent and between five and ten percent, respectively.

Both questions one and two are tasks which focus on frequency estimation. Question three again has the same focus, but the participant needs to identify the part of speech of the words at the position of interest. This question asks"*For the keyword KEYWORD, what is the most frequent descriptive adjective at position keyword - 1?*" and a clarifying statement and example are given"*A descriptive adjective is* *a word which describes a noun (KEYWORD is the noun in this case). e.g. In the text fragment "an old book" old is an adjective describing the noun book*".For this question, the chosen keywords return a concordances with approximately one-thousand lines where the correct adjective is the fifth most common word at position $keyword - 1$ with a frequency of about seven percent.

The fourth question asks the user to identify a frequent combination of words. Specifically, they are asked to identify the most frequent word at position $keyword-2$ when another specified word occurs at position $keyword-1$. An example of question four is "*For the keyword "standing", focusing only on concordances that contain the word "still" at position keyword - 1, what word is most frequent at position keyword - 2?*". This question becomes much easier to answer if the participant uses a filtering interaction, so a hint was provided telling them to do so for all interfaces. They had been previously instructed during the tutorial on how to perform this interaction. The frequency of the answer after the filter interaction was approximately fifty percent of 200 occurrences. The keywords used for question four were *went*, *did*, and *was*.

Finally, question five asks the participant to identify the word with the highest collocation strength at position $keyword - 1$. The correct answers have a positional collocation strength score of fifty percent: meaning the collocation strength of the word at that position is as strong as the combined collocation strength of all other words at that position. I expect this task to be the most difficult of the five when using the KWIC interface. The keywords used for question five were *jubilee*, *burlap* and *wheezing*.

Looking again at these five questions, it should be clear that questions one and two both evaluate frequency estimation actions, and I expect question two to be more difficult as the two most common words have similar frequencies. Question three is again a frequency estimation action that combines a qualitative task of identifying parts of speech, and this should also be more difficult than questions one and two because the answer is the fifth most frequent word. Question four is a mix of the filter action that is not required but makes the task much easier and is recommended to the participant. Question five is a collocation-strength action that can be performed accurately using the KWIC interface by using frequency actions for each word or expert knowledge to evaluate only the most likely candidates.

7.1.2 Results

7.1.2.1 Cross-tabulation

Table 7.1: Cross-tabulation of mean time to complete each question per interface.

Question	J	K	M
Q1	24.13	56.69	19.48
Q2	13.08	40.85	12.09
Q3	27.69	79.71	23.11
Q4	43.66	73.39	43.13
Q5	33.32	121.75	27.97

Table 7.2: Cross-tabulation of incorrect answers for each question per interface out of a possible 36 attempts.

Question	J	K	M
Q1	0	3	1
Q2	1	7	1
Q3	9	8	11
Q4	2	4	6
Q5	0	23	0

First, let us look at the cross-tabulated results of the measured values in relation to the questions and interfaces. The values being measured and used to evaluate the performance of the interfaces are the time to complete each question (t) and the correctness of the answer to each question (isCorrect). In Table 7.1, the mean time to complete each question using each interface is presented. At first glance, the mean time to complete questions using the KWIC interface (K) is longer for every question. In Table 7.2, the total number of incorrect answers for each question and interface is displayed. This table shows the least errors when using the Juxtaposed interface, slightly more errors when using the mosaic alone, and many more errors when using the KWIC interface. However, most of the errors For the KWIC interface occur in question 5, but even if question 5 were ignored, the errors in the KWIC interface are slightly more than each of the other interfaces.

Shapiro-Wilk normality tests were used to assess the normality of the times to complete each question (t) per interface. The null hypothesis of this test is that the population is normally distributed. The tests applied to response times per question and interface, in most cases, reject the null hypothesis. In Table 7.3, the only response time distributions for the interface question pairs which do not reject the null hypothesis are M:q2, K:q3, and K:q4. Taking the dataset as a whole and ignoring the split between questions and interfaces also yields a non-normal distribution.

Table 7.3: Cross-tabulation of Shapiro-Wilk normality test, the table presents p-values for the time to complete the questions per interface. The null hypothesis of this test is that the population is normally distributed. For p-values less than 0.05, the null hypothesis is rejected, meaning there is evidence that the data tested are not normally distributed. For the KWIC interface, questions 3 and 4 do not reject the null hypothesis. Question 2 using the Concordance Mosaic also does not reject the null hypothesis.

	J	K	M
1	$1.07113080201535e-03$	0.00541598858701628	$3.15653939325606e-04$
2	$2.55921250089124e-03$	0.04085259905705785	$1.37799885752188e-01$
3	$2.84871168184496e-04$	0.12309100768904353	$4.71867629281534e-09$
4	$3.14084746288457e-05$	0.15542364394019373	$2.61133378707408e-03$
5	$1.76268957731993e-08$	0.03624446087186547	$1.22349854527216e-05$

7.1.2.2 Analysis of variance

Next, let us look at the results of an ANOVA for the dependent variable t, time to complete each question measured in seconds, with respect to the categorical variables; the question being answered (q), the interface being used (i), the participants' assigned interface ordering (iOrder), the participants' assigned keyword set ordering (qOrder) and a binary variable representing a correct or incorrect answer (isCorrect). The results of the ANOVA where a significant difference ($p < .05$) was found are given in Table 7.4. It should be noted that the dependent variable (t) is not normally distributed, which might affect the results of this parametric test.

Table 7.4: ANOVA results for the dependant variable time t, where Pr(>F)<0.05 indicates significance.

	Sum Sq	Df	F value	Pr(>F)
q	78623.695388835273	4	26.2054108270503754	$8.86774908716352e-18$
i	198020.668348797597	2	132.0012482411049461	$2.15406447904375e-38$
iOrder	6984.351166801935	5	1.8623168578507050	$1.02031874105271e-01$
qOrder	280.373990821594	2	0.1868982519420692	$8.29659706452425e-01$
isCorrect	2932.384952249995	1	3.9094754830200298	$4.92632225778424e-02$
q:i	48293.087008973787	8	8.0480838436675466	$1.62770530513198e-09$
q:iOrder	10190.008199302771	20	0.6792693980437359	$8.44661539919004e-01$
i:iOrder	20394.603009683546	10	2.7190222890451836	$3.61286426896421e-03$
q:qOrder	4406.670122716925	8	0.7343753074310707	$6.61061215115642e-01$
i:qOrder	10794.107652509469	4	3.5976943610511878	$7.28688778884667e-03$
iOrder:qOrder	13167.740011112532	10	1.7555320184250351	$7.01839369225560e-02$
q:isCorrect	9712.826936671278	4	3.2373016672484667	$1.31799900714377e-02$
i:isCorrect	116.193800818961	2	0.0774551811882837	$9.25493727495571e-01$
iOrder:isCorrect	7794.221022536280	5	2.0782616534343146	$6.91801056433746e-02$
qOrder:isCorrect	2135.341021174594	2	1.4234255573715908	$2.43097054011249e-01$
q:i:iOrder	29651.732168728311	40	0.9882972548826947	$4.96821441157116e-01$
q:i:qOrder	7189.572524769057	16	0.5990741746196304	$8.82906744009495e-01$
q:iOrder:qOrder	13354.811613156053	40	0.4451181327840690	$9.98487719715081e-01$
i:iOrder:qOrder	12472.803368832771	20	0.8314412972547580	$6.73842915463997e-01$
q:i:isCorrect	274.794173166680	2	0.1831787265937906	$8.32746191351884e-01$
q:iOrder:isCorrect	6205.976097758336	7	1.1819784820435475	$3.14122729288689e-01$
i:iOrder:isCorrect	3657.920401000010	2	2.4383821291225152	$8.96583401328384e-02$
q:qOrder:isCorrect	5526.546307400014	4	1.8420072437946120	$1.21805507823843e-01$
i:qOrder:isCorrect	2777.449140500015	2	1.8514570046112679	$1.59446261468688e-01$
iOrder:qOrder:isCorrect	5809.084861125011	4	1.9361778222471682	$1.05443965652680e-01$
q:i:iOrder:qOrder	30340.261018300080	67	0.6037289779375635	$9.91788586077539e-01$
q:i:iOrder:isCorrect	0	0	0	0
q:i:qOrder:isCorrect	0	0	0	0
q:iOrder:qOrder:isCorrect	0	0	0	0
i:iOrder:qOrder:isCorrect	0	0	0	0
q:i:iOrder:qOrder:isCorrect	0	0	0	0
Residuals	165015.663174499990	220	0	0

Since the main effects q, i and isCorrect all feature in significant interactions, I have focused the posthoc analysis on interactions with these variables. I conducted Tukey's posthoc tests (HSD) to analyse the different groupings of each interaction effect, again using p < .05 to test for significance. It should again be noted that since Tukey's posthoc test (HSD) is a parametric test, the dependent variable (t) not being normally distributed might affect the results.

7.1.2.3 Posthoc analysis

The result of the HSD test for the i and qOrder interaction (i:qOrder) showed a significant difference between the two groupings. In this case, the dataset was split into nine groups by the combinations of the three interfaces and the three circularly shifted keyword set orderings. The HSD groupings simply combined these groups into data points where the KWIC interface was being used and a grouping of all data points where either the Mosaic or Juxtaposed interfaces were being used. This indicates that the interaction can be interpreted as i, and that qOrder can be safely ignored as it doesn't feature in any other significant interactions or as a main effect. This result shows, as expected, that the choice of keywords has not had a major effect on the time to complete each question.

The mean response times of the i:qOrder groups in which the KWIC interface was used were all greater than sixty-seven seconds, while the remaining groups containing the Mosaic and Juxtaposed data all had mean response times under thirty seconds. This is evidence of interface choice having an effect on response time.

The HSD test for the interaction between i and iOrder (i:iOrder) examines the data set split into eighteen groups on the combination of the three interfaces and the six possible interface orderings. The results of the HSD test found six significantly different groupings. I examined the result of the test as a scatter plot of these HSD groupings (Figure 7.2). This scatter plot shows the mean response times of the eighteen i:iOrder groups, a slight jitter from the grouping lines was applied. Looking at this scatter plot, eleven of the twelve groups which used the Mosaic (M) or Juxtaposed (J) interfaces are grouped together, and all twelve have a mean response time of less than forty seconds. The remaining groupings all have a mean response time of over sixty seconds and are the cases where the KWIC (K) interface was used. Looking at the groups where the KWIC was in use, a learning effect can be observed where in situations in which the participant had used the Juxtaposed interface before KWIC, faster response times were recorded. Interestingly, it ap-

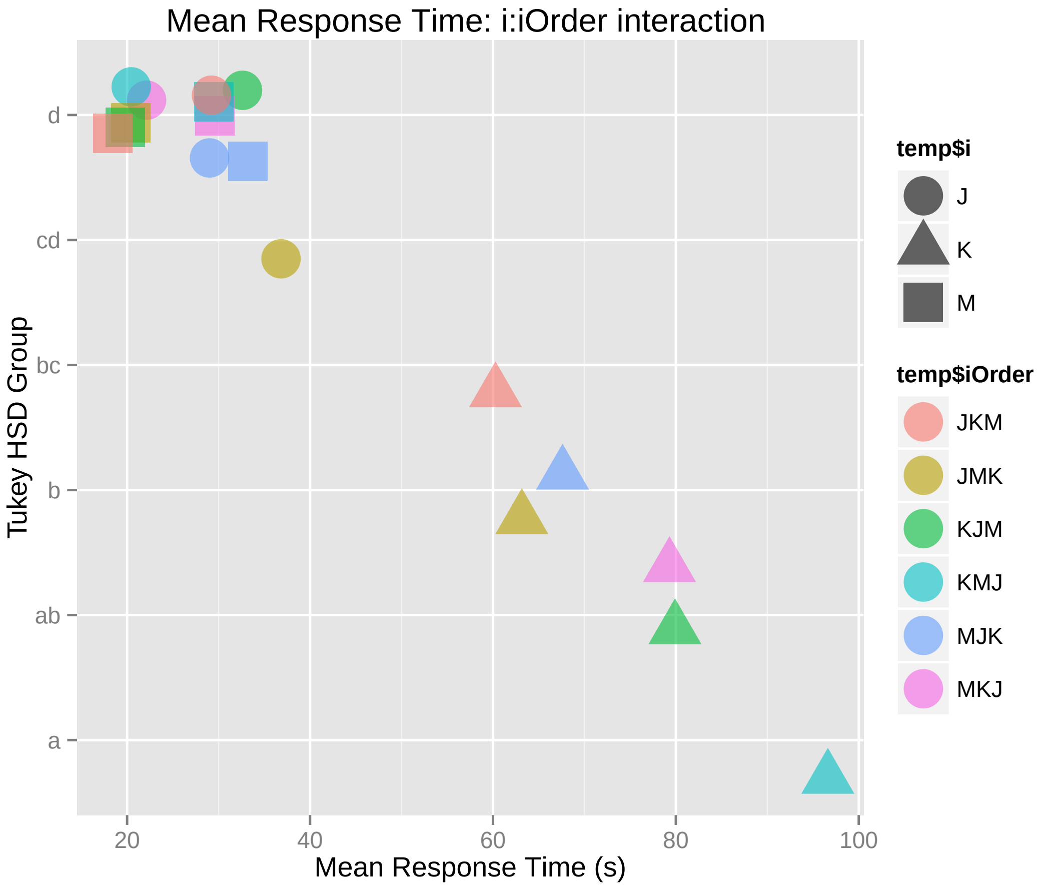


Figure 7.2: Mean Response time of i:iOrder data groups. Clustered by Tukey HSD score. The three-letter abbreviations in the legend under *tempiOrder* represent the order in which the interfaces were shown to the participants.

pears that no significantly large learning effect takes place between the Mosaic and Juxtaposed interfaces. The only data point indicating a difference between the two interfaces when changing the ordering is the J:JMK data point.

The discovery of an interaction between q and i (q:i) is of great interest since the null hypothesis states: there are no significant differences between the interfaces on a per-question basis. Analysing the groups created by splitting the data by interface and question, the Tukey HSD test found a number of significant groupings (Figure 7.3). For each question, there is a significant difference between the KWIC interface and the Mosaic and Juxtaposed interfaces. This evidence is enough to reject the null hypothesis for each question. With the null hypothesis rejected, I

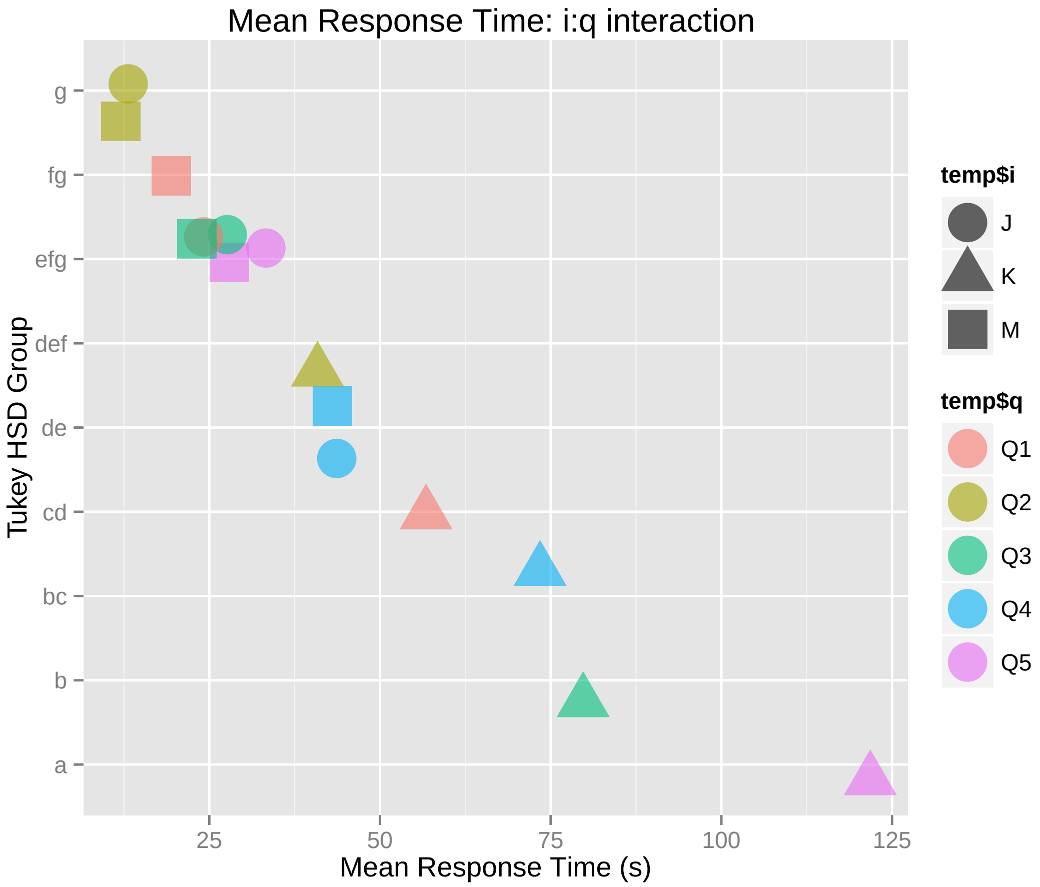


Figure 7.3: Mean Response time of i:q data groups. Clustered by Tukey HSD score.

still look further at the results to investigate these differences.

The Tukey HSD groupings of the i:q interaction show (Figure 7.3) that for all questions (with the exception of question one Figure 7.4), there was no significant difference between the response times per question for the Mosaic and Juxtaposed interfaces. For these interfaces, question two (Figure 7.5) was the quickest to complete, while questions one, three (Figure 7.6) and five (7.8) took slightly longer and question four (Figure 7.7) took even longer still. In comparison, on the KWIC interface, question four was the third fastest to complete, while five took the most time by a large margin. Again question two is the quickest to complete. These plots show the 36 data points for each question and interface combination.

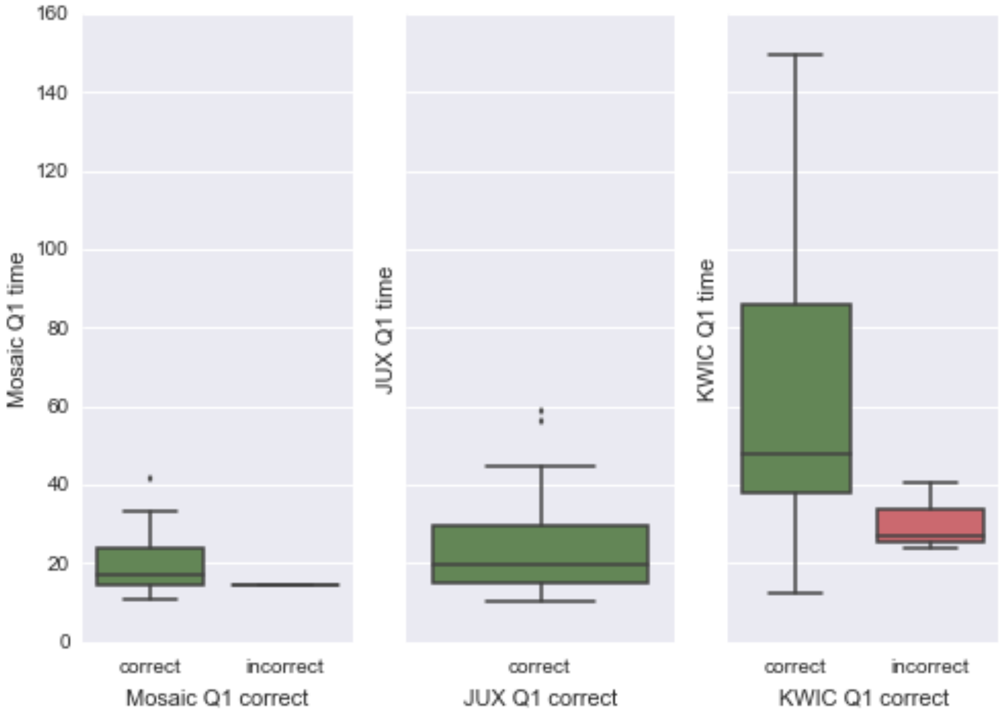


Figure 7.4: Box plots of question one response times across the three interfaces.

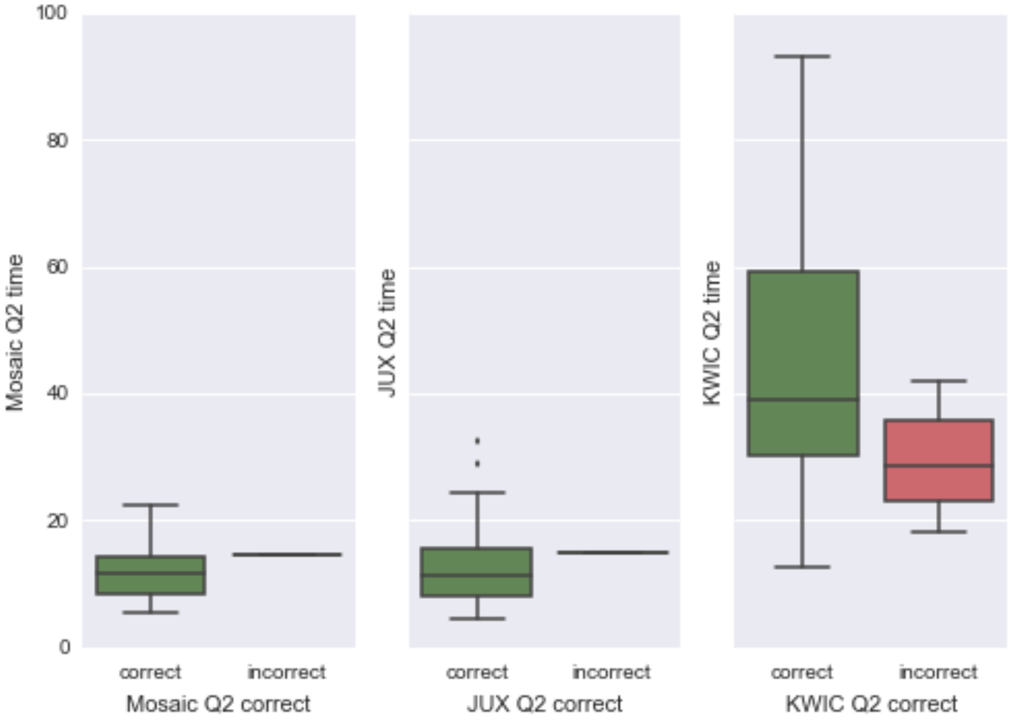


Figure 7.5: Box plots of question two response times across the three interfaces.

7.1.2.4 Analysis of correctness

The split between correct and incorrect answers can also be seen from the boxplots (Figure 7.4, Figure 7.5, Figure 7.6, Figure 7.7, and Figure 7.8). Note the differing

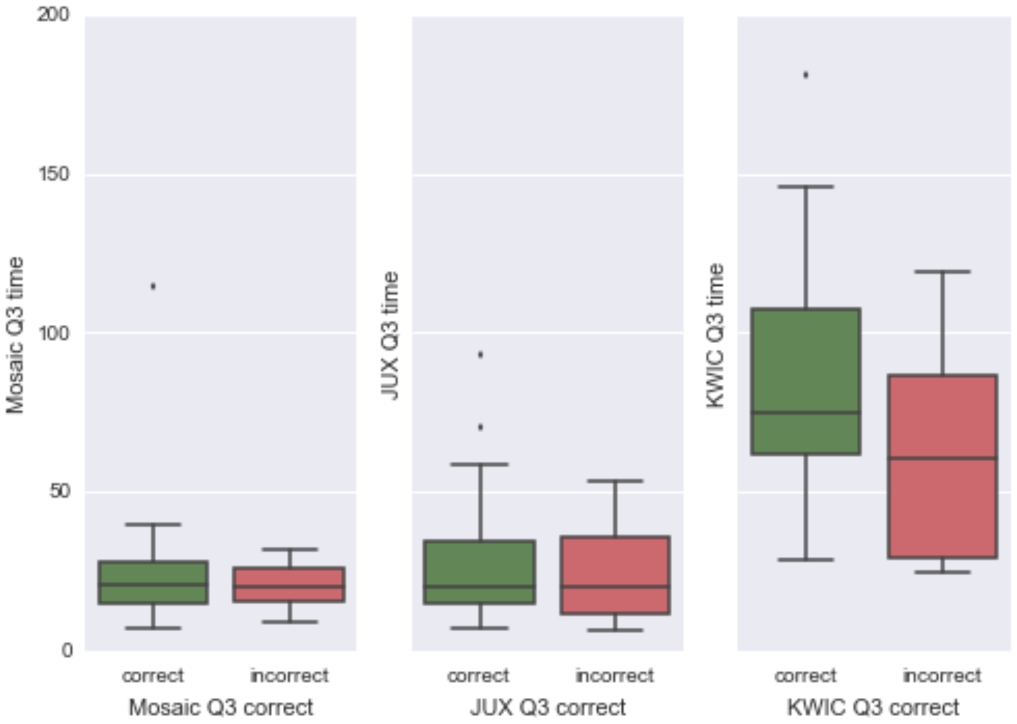


Figure 7.6: Box plots of question three response times across the three interfaces.

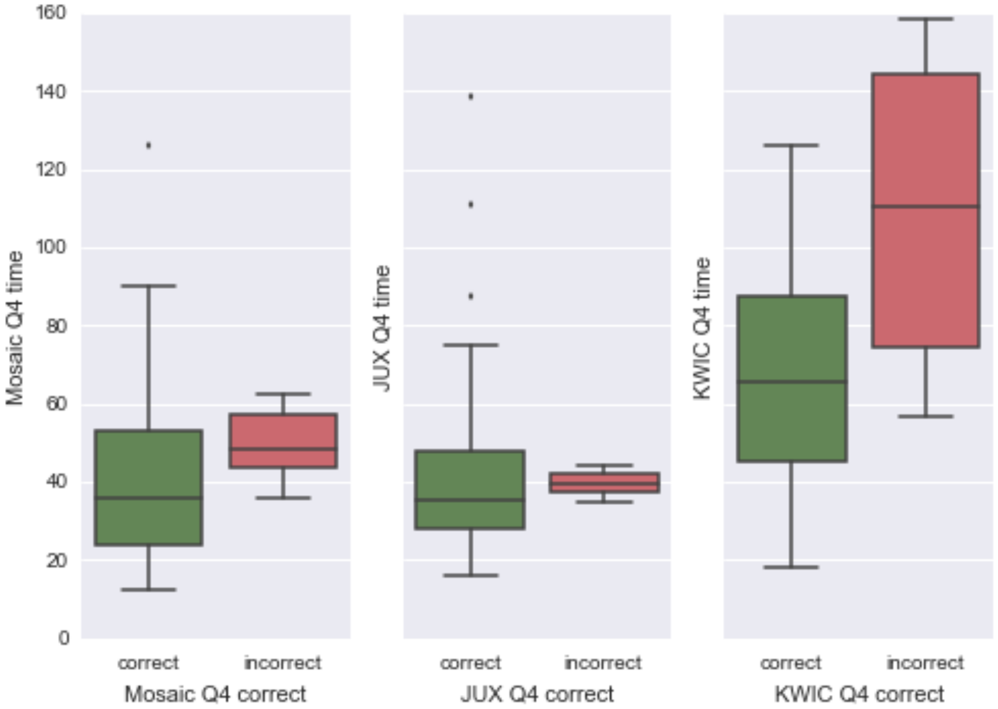


Figure 7.7: Box plots of question four response times across the three interfaces.

ranges on the y-axis between the plots. These plots show the difference between the KWIC response times and the other two interfaces. Questions two and five

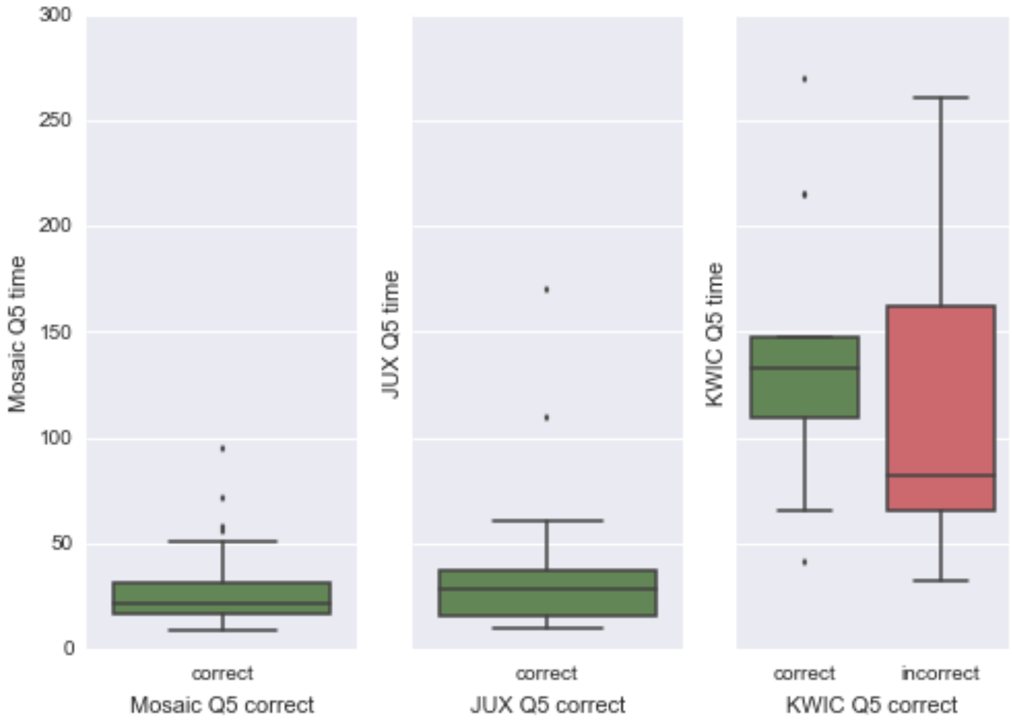


Figure 7.8: Box plots of question five response times across the three interfaces.

Table 7.5: Incorrect answers per question and interface.

	Q1	Q2	Q3	Q4	Q5	Sum
K	3	7	8	4	23	45
M	1	1	11	6	0	19
J	0	1	9	2	0	12
Sum	4	9	28	12	23	76

Table 7.6: Correct answers per question and interface.

	Q1	Q2	Q3	Q4	Q5	Sum
K	33	29	28	32	13	135
M	35	35	25	30	36	161
J	36	35	27	34	36	168
Sum	104	99	80	96	85	464

have many incorrect answers when using the KWIC interface. Table 7.5 shows the number of incorrect answers per interface and question. Question three had the most incorrect answers but was approximately evenly distributed among the interfaces. In the case of question five, there were zero errors using the Mosaic or Juxtaposed interface, while twenty-three of the thirty-six attempts using the KWIC interface were incorrect.

To test for significance in the correctness measurement, Spearman's rank correlation tests were used. The result of a test comparing the results per interface does not show a significant correlation between errors and interface.

$$S = 8, p - value = 0.3333$$

7.1.3 Discussion

The null hypothesis that there is no significant performance difference between the interfaces per question has been rejected. The Mosaic and Juxtaposed interfaces have been shown to be equivalent for the designed tasks, while the KWIC interface performs significantly worse on each of the five tasks. I speculate that this may indicate that participants using the Juxtaposed interface, which combines both the Mosaic and KWIC, have a preference for the Mosaic interface as indicated by similar performance statistics.

The five questions cover a broad section of the quantitative actions I identified in the task analysis, and each of these actions features in many corpus analysis tasks most often performed by text analysts. The fact that the Mosaic and Juxtaposed interfaces offer performance increases over the standard method in the field should be seen as a contribution to corpus analysis.

Looking at the performance between questions, I expected question two to be more difficult than one, but the opposite appeared to be true when examining completion time (see Table 7.1, Figure 7.4 and Figure 7.5), this is most likely due to participants learning from question one since both questions are the same, only the keywords and frequencies involved are different.

Using the Mosaic and Juxtaposed interfaces, participants had the worst time to complete performance on question four. However, this performance was still much better than the KWIC interface, where this question ranked third in performance. Question four involved the frequencies at two-word positions, and a filter interaction much simplified the task since the four other questions involve observing a frequency or collocation strength at a single position, and since no interaction is required, the performance decline using the Mosaic makes sense. Also, since the filter interaction simplifies the task using all interfaces, the performance against questions three and five, using the KWIC interface, where no such simplification is available, is to be expected.

Question 3 had the worst performance in terms of correctness for both the Mosaic and Juxtaposed interfaces, performing only slightly better than the KWIC interface. I suspect the additional requirement of identifying an adjective rather than a word is responsible for the drop in correctness in this question.

On the Mosaic and Juxtaposed interfaces, question five had equivalent performance; the performance of the KWIC interface in terms of both time to complete and error rate on question five is worse by a large margin, from Table 7.5 and Table 7.6 twenty-three of the thirty-six attempts at question five using the KWIC were incorrect. However, this task is less representative of a common corpus analysis task using this interface. These collocations or other statistics are usually calculated by an external tool and returned as a list without reference to word position. This question shows that using the Mosaic interface and the concordance graph, positional statistics can be calculated and included in the visual representation of the concordance in an easy-to-understand manner.

7.2 Conclusion

The null hypothesis of no significant performance difference between the interfaces per question has been rejected. The Mosaic and Juxtaposed interfaces are equivalent for the designed quantitative tasks, while the KWIC interface performs significantly worse on each of the five tasks.

This result helps to mitigate the threats to the validity of the Concordance

Mosaic design. It suggests that the combination of data abstraction and encoding choice enables faster investigation of positional collocation frequencies and statistics.

To determine if the Mosaic encoding facilitates answering useful questions for the target users, i.e. investigating positional collocation frequencies and statistics, we must use different evaluation techniques.

In the next chapter (chapter 8, the methodological impact of the Concordance Mosaic(chapter 5) and ComFre (chapter 6) visualisation are evaluated using contextual studies ([Sedlmair et al., 2012]) with target users.

Chapter 8

Review of Methodological Impact

I chose to review the impact the Mosaic visualisation (described in chapter 5) and the ComFre visualisation (described in chapter 6) were having on the methodologies of the GOK researchers. At the time of this review, the Mosiac visualisation had been available to the researchers for over a year. ComFre had been released two months before this review.

To perform the review, I conducted contextual studies [Sedlmair et al., 2012] with the researchers referred to as Daisy and Dave in chapter 3. These researchers had described example methodologies prior to the release of these visualisation tools (chapter 3). I requested that the researchers allow us to observe them performing corpus analysis. I requested that this analysis should be work that has academic publication as its goal. If this work includes the visualisation tools that enhance the researchers' speed or analytical capabilities, that is evidence of methodological impact.

An additional methodology is included in this chapter, but its relevance for determining methodological impact is less clear. This is because it was presented at a project meeting to discuss the usage of the concordance tools and was prepared as an example of how the ComFre tool can be useful for analysis. While it is still an example of methodological change, its presentation as a teaching example makes any conclusions about impact less convincing.

Finally, it is important to note the Mosaic, at the time of this review, did not include the redesigned statistical measures of collocation. It instead used a simple measure of collocation strength which was later replaced by statistical measures of Mutual information, z-score and log-likelihood.

8.1 Methodological Review/Shadowing

8.1.1 Daisys methodology: Case study of "the people"

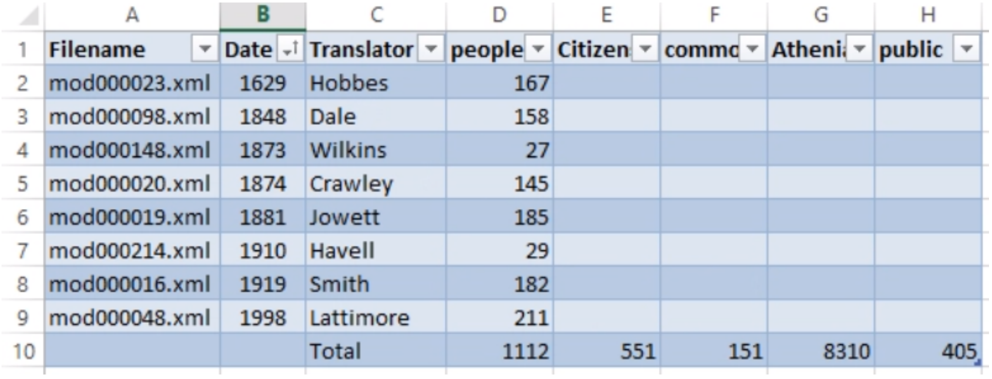
The study which Daisy shared was new work that could, with continued analysis, lead to a publication. Before the task began, a brief introduction was given. The study was on the concept of "the people" in "Thucydides". The GOK corpus contains translations of "Thucydides" from classical Greek. These translations dated from 1629 to 1998. This sub-corpus is quite small, only containing eight files. Other studies of this sort could be envisioned with a much larger corpus.

One question of interest was "Does the concept of "the people" change over time?". Another question of interest is "Who does "the people" refer to?". In this analysis, there is ambiguity about the meaning of the concept under investigation. How is the concept of "the people" presented in different time periods, texts and by individual translators?

The analysis was video recorded, and a description of the methods continuation beyond what was recorded was provided. Following the observation clarifying questions were asked of Daisy. The following is a summary of the observed analysis and the proceeding discussion.

8.1.1.1 Observation

Prior to the observation session, a spreadsheet was created with the headings filenames, date, translator, people, citizens, commons, Athenians, and public. The meta-data information related to filename, date, and Translator were added to the table. The remaining headings are keywords that will be investigated as a part of this study. The spreadsheet used in the study can be seen in Figure 8.1. Partitioning the frequencies by date, file, or translator is equivalent for this sub-corpus as each file has a unique author and date.



	A	B	C	D	E	F	G	H
1	Filename	Date	Translator	people	Citizen	commc	Atheni	public
2	mod000023.xml	1629	Hobbes	167				
3	mod000098.xml	1848	Dale	158				
4	mod000148.xml	1873	Wilkins	27				
5	mod000020.xml	1874	Crawley	145				
6	mod000019.xml	1881	Jowett	185				
7	mod000214.xml	1910	Havell	29				
8	mod000016.xml	1919	Smith	182				
9	mod000048.xml	1998	Lattimore	211				
10			Total	1112	551	151	8310	405

Figure 8.1: Spreadsheet used in Daisys study of "the people" in translations of "Thucydides" from the GOK corpus.

The first steps of the study focused on the keyword frequencies in the entire sub-corpus.

- The sub-corpus of "Thucydides" was selected.
- The keyword "people" was searched, and the total frequency in the corpus was recorded
- Regulator expressions for the other "citizens?", "commons?", "Athenians" and "public" searched, and total frequency in the sub-corpus is recorded

After the keyword frequencies had been recorded, Daisy commented that the keyword "Athenians" is much more frequent than other keywords. This is unexpected and will need to be investigated.

The next step was to gather the Keyword frequencies for individual files.

- Make a Sub-corpus selection for each individual file. Recording in the spreadsheet the number of lines returned by the keyword "people".

The analysis now turns from keyword frequency to the identification of collocation patterns. The mosaic was used extensively to identify collocation patterns and frequency of occurrence. The steps observed were:

- Make a sub-corpus selection for the first file.

- Perform a search for the first keyword “people” in the concordance browser.

- Open the Mosaic visualisation and remove stop-words.

- Examine word frequencies

- Open a document for taking notes and record the most frequent collocations directly to the left of the keyword. The words “common and “Athenian” were recorded

- Return to the sorted concordance list and check if any continuations(such as “Athenians”) are present.

- Record the counts for the frequent collocated words.(common 8, Athenian 6)

- Open the frequency mosaic with stop-words included.

- Record in notes “lots of hits for the+people (i.e. unmodified)”

- Similar analysis for second file.

- Frequent collocates directly to left of “people”(common 34, Athenian 5)

- Record “A few more different adjectives modifying this noun: entire, experienced, free, dynamic, adventurous.”

- Similarly, for the third file, the noted collocates noted were (Athenian 13, whole 13, common 5)

The recording ended, and Daisy explained how the analysis would progress. The collocation pattern method is repetitive and would continue in the same manner for each file and keyword. The next stage of the analysis would be to analyze the frequency patterns using the table. Possibly making bar charts in a spreadsheet application. Temporal patterns are expected. Identified patterns will be investigated using qualitative analysis, which involves reading the concordance lines related to the identified patterns. The goal is to understand the meaning of the concept of "the people" at different times.

This analysis is performed in the context of the knowledge Daisy has about the corpus and texts. She states that it is interesting that there are "No translations 1919-1998, during the period of huge cultural change in Britain. Possible reasons for this include Suffrage, war, or technological revolution. Daisy explained that information about the authors and texts will influence the analysis. Some examples of information that is relevant are "the political leanings of the translators, which is established relevant knowledge" and "certain texts are partial translations, abridged versions, etc". Any differences identified, temporal or otherwise, must take into account the translator's style, politics, and more.

Some questions were asked to Daisy to elicit more information about the methodology.

- How did you come up with this methodology?

> "Playing around with the corpus tools, generating concordances for interesting keywords, trying to find patterns in the data."

- How did you choose the keywords?

> "Obvious keywords associated with the concept of "the people". The idea for the study emerged through reading the literature on citizenship
> "

- Is the methodology typical of the field?

> "GoK is the first project to use corpora to attempt to understand the role of translation in the evolution and contestation of political and scientific concepts. One of the core aims of the project is to develop new methodologies which might enable researchers to study such phenomena.

Translation Studies as a discipline tends to encourage close qualitative analysis of a small selection of examples from specific texts to illustrate a particular argument.

Corpus analysis enables the translation scholar to identify and investigate with significantly greater ease differences between and patterns within translations, taking into account the full length of each work as a complete text.

Corpus analysis has been extensively used in translation studies before (e.g., within the TEC project and many others), but the field has tended to focus mainly on more micro-level linguistic concerns rather than the socio-political implications of translators' word choices, etc. ”

- Would this methodology be useful for other researchers in the field?

“Other scholars using the GOK software to investigate the role of translation in the evolution of political and scientific discourse use similar methods. Other projects developing other corpora may also adopt some aspects of the methodology. ”

- What are the barriers to the adoption of your methodology?

“Not sure. Perhaps better documentation of the corpus software, detailing what it can and can't do, with lots of example analysis. The publication of case studies by members of the team will also help demonstrate the potential of the tools. ”

- Mosaic was used in this analysis. Is this typical when you investigate collocation patterns?

“Yes. Mosaic will be very useful for this case study and any investigation of collocations because it tells you in a very quick and transparent

> way which are the most common collocates in each word position for a given keyword."

- You did not make use of collocation strength in your analysis. Do you intend to?

> "No. The collocation strength Mosaic is not immediately transparent and so (to be brutally honest) would tend to slow down analysis rather than speed it up."

- Have you used this methodology for other studies?

> "The collocation pattern aspect of this study is unique in my work. I have, in previous studies, studied keyword frequency in larger subcorpora where there are multiple files for each author and date. I can show you an example of the concept of "Statesman"."

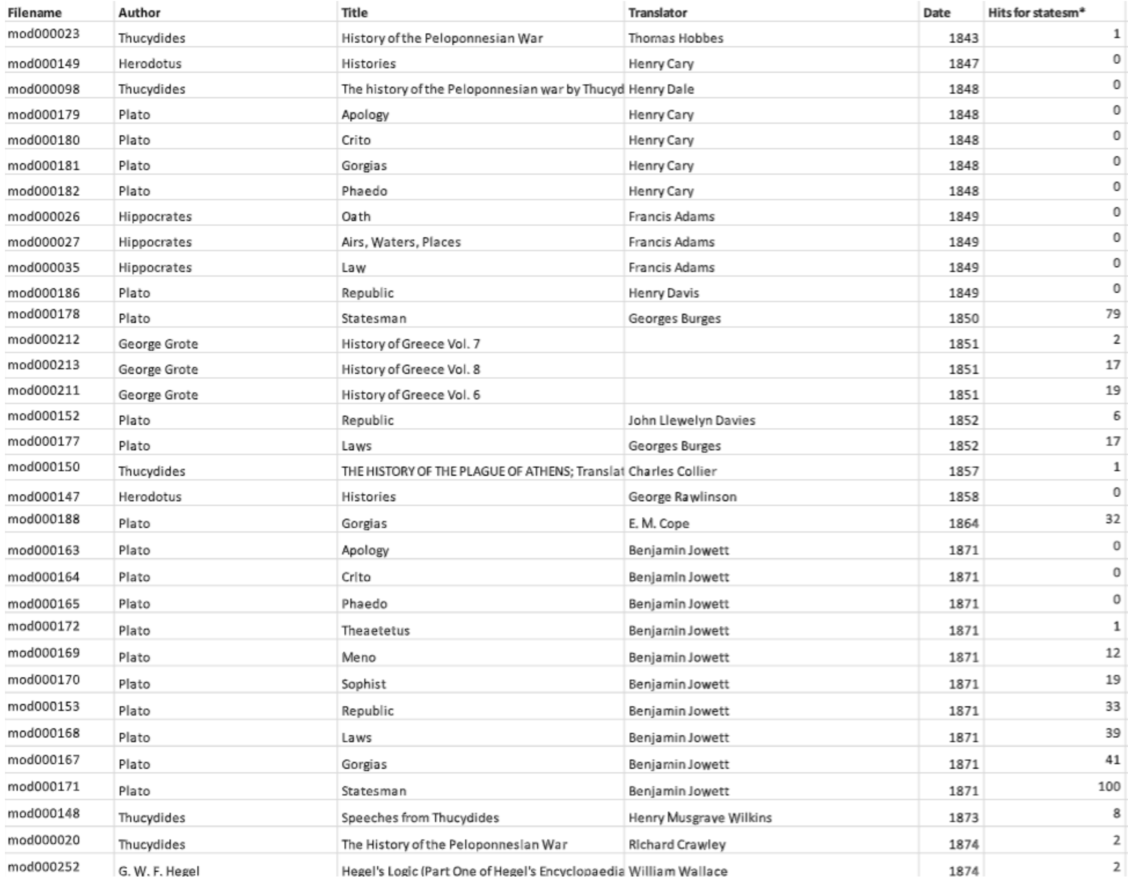
8.1.2 Daisys methodology: Case study of "Statesmanship"

An unpublished paper on a case study of the concept of "Statesmanship" was supplied by Daisy, and the major conclusions and analysis were described. The paper was later published [Jones, 2019]. Here is the abstract from the paper:

> "With its connotations of superior moral integrity, exceptional leadership qualities and expertise in the science of government, the modern ideal of statesmanship is most commonly traced back to the ancient Greek concept of politikos and the work of Plato and Aristotle in particular. Through an analysis of a large corpus of modern English translations of political works built as part of the AHRC Genealogies of Knowledge project (http://genealogiesofknowledge.net/), this case study aims to explore patterns that are specific to this translated discourse, with a view to understanding the crucial role played by translators in shaping its development and reception in society. It ultimately seeks to argue that

> the model of statesmanship presented in translations from ancient Greek is just as much a product of the receiving culture (and the social anxieties of Victorian Britain especially) as it is inherited from the classical world."

In the GOK corpus, the term "Statesman" was found to exist "almost exclusively (90%) in translations from Classical Greek". This pattern was not observed for similar keywords such as "governor", "leader", "ruler", and "citizen" which are more evenly distributed across all language pairs. The analysis which arrived at this conclusion was a simple keyword frequency comparison across the translation facets of the corpus. This involved selecting each sub-corpus individually and recording the number of concordance lines for the keywords in each sub-corpus.



Filename	Author	Title	Translator	Date	Hits for statesm*
mod000023	Thucydides	History of the Peloponnesian War	Thomas Hobbes	1843	1
mod000149	Herodotus	Histories	Henry Cary	1847	0
mod000098	Thucydides	The history of the Peloponnesian war by Thucyd	Henry Dale	1848	0
mod000179	Plato	Apology	Henry Cary	1848	0
mod000180	Plato	Crito	Henry Cary	1848	0
mod000181	Plato	Gorgias	Henry Cary	1848	0
mod000182	Plato	Phaedo	Henry Cary	1848	0
mod000026	Hippocrates	Oath	Francis Adams	1849	0
mod000027	Hippocrates	Airs, Waters, Places	Francis Adams	1849	0
mod000035	Hippocrates	Law	Francis Adams	1849	0
mod000186	Plato	Republic	Henry Davis	1849	0
mod000178	Plato	Statesman	Georges Burges	1850	79
mod000212	George Grote	History of Greece Vol. 7		1851	2
mod000213	George Grote	History of Greece Vol. 8		1851	17
mod000211	George Grote	History of Greece Vol. 6		1851	19
mod000152	Plato	Republic	John Llewelyn Davies	1852	6
mod000177	Plato	Laws	Georges Burges	1852	17
mod000150	Thucydides	THE HISTORY OF THE PLAGUE OF ATHENS; Translat	Charles Collier	1857	1
mod000147	Herodotus	Histories	George Rawlinson	1858	0
mod000188	Plato	Gorgias	E. M. Cope	1864	32
mod000163	Plato	Apology	Benjamin Jowett	1871	0
mod000164	Plato	Crito	Benjamin Jowett	1871	0
mod000165	Plato	Phaedo	Benjamin Jowett	1871	0
mod000172	Plato	Theaetetus	Benjamin Jowett	1871	1
mod000169	Plato	Meno	Benjamin Jowett	1871	12
mod000170	Plato	Sophist	Benjamin Jowett	1871	19
mod000153	Plato	Republic	Benjamin Jowett	1871	33
mod000168	Plato	Laws	Benjamin Jowett	1871	39
mod000167	Plato	Gorgias	Benjamin Jowett	1871	41
mod000171	Plato	Statesman	Benjamin Jowett	1871	100
mod000148	Thucydides	Speeches from Thucydides	Henry Musgrave Wilkins	1873	8
mod000020	Thucydides	The History of the Peloponnesian War	Richard Crawley	1874	2
mod000252	G. W. F. Hegel	Hegel's Logic (Part One of Hegel's Encyclopaedia	William Wallace	1874	2

Figure 8.2: A sample from the Spreadsheet used in Daisys study of "Statesman" in translations of Classical Greek from the GOK corpus. The full spreadsheet contains 261 lines of analysis.

The frequency of the keyword "Statesman" in the sub-corpus of Classical Greek translations was analyzed. A spreadsheet with an entry for each of the 261 files in the subcorpus was created, and meta-data (the author, the title, the translator, and the date) were entered for each file. This was done manually and was time-consuming. Daisy explained that in this form "the information could easily be (re)sorted according to each of these meta-data facets and patterns more easily identified." The number of concordance lines for each file was found by selecting a sub-corpus of a single file and searching for "Statesman". Performing this action for each of the 261 files was also time-consuming. A sample of the completed spreadsheet can be seen in Figure 8.2

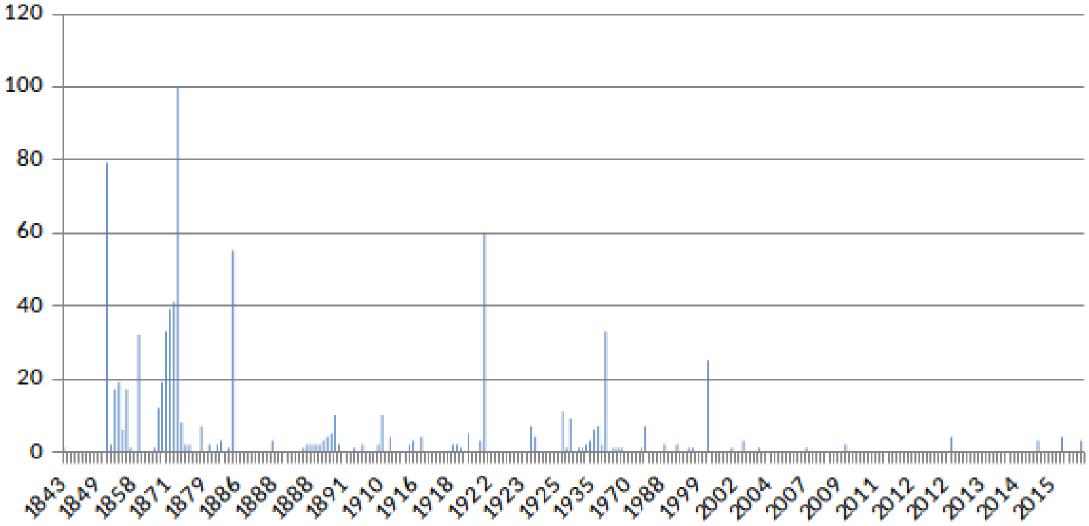


Figure 8.3: Bar chart examining temporal spread in translations of ancient Greek.

By examining the spreadsheet and generating bar charts, such as Figure 8.3, the faceted distributions of "Statesman" can be understood. Statesman seemed to be "bursty" per author and to exhibit a temporal pattern.

> The frequency of statesman in these corpora suggests most recent translations (1950-2012) of ancient Greek texts use Statesman much less frequently. This is surprising because the corpus contains several recent retranslations (published within the last seventy years) of classical

texts such as Aristotle's Politics or Plato's Dialogues, which in earlier English-language interpretations included the keyword 'statesman' very prominently.

Some clarifying questions were asked and answered:

- You mentioned the process of completing the spreadsheet was time-consuming, how long did it take?

> "Probably around 5-6 hours because of the amount of manual processing required. It would take a lot longer if I were to investigate more than one keyword."

- Where did the idea for this study and methodology come from?

> "This was exploratory. I was not trying to establish anything in particular, only to understand whether the term "statesman" was used, how frequently (in comparison with other semantically related terms), and if any obvious patterns could be found from these initial quantitative analyses.
>
> The terms "statesman" and "citizenship" which I have investigated previously, are closely related concepts, especially in classical Greek thought."

- Were the visualisation tools used in this case study?

> "My focus on the use of a single keyword (statesman) and alternative word choices did not require and collocation pattern analysis. This is more typical of translation studies research. The corpus tools lend themselves particularly well to the analysis of collocations (this is one of their clear advantages) and is why I want to push my research in this direction with my next case study."

- Are there any areas of your methodology where current or new visualisation tools could be beneficial?

> "Constructing the spreadsheets is time-consuming. A tool which can help identify patterns in the dispersion of a concept according to different meta-data facets would be extremely helpful, at least for the kinds of research I intend to carry out as part of this project."

8.1.3 Dave methodology: Case study of "Democracy"

The methodology described by Dave was based on his current research as part of the GOK project. The demonstration which was observed was a partial reenactment of the analysis which has already been performed. The concept of democracy was investigated in modern books 1970s onwards. Dave commented that this "is in line with the most fundamental goals of the GOK project.". The steps taken which were observed and recorded were:

- Begin by searching the keyword "democracy" without any bias for what will be returned.

- Open the Mosaic frequency view and see if anything stands out (It doesn't)

- Look at stop-word view. Social democracy has a very strong collocation. Click social and look at the concordance lines now highlighted in the concordance browser.

- Reading the lines reveals "Social democracy" appears in file mod8 and refers to one book title and its contents. This is only informative about this specific file, and the file is removed from the sub-corpus under investigation to gain a more balanced overview.

- The search is re-run, recording that approx 500 lines were removed from the concordance. "common and "Athenian" were recorded

- The Mosaic is investigated again. Both frequency and stop-word frequency views don't seem to show any unexpectedly frequent results.

- Navigate to the collocation strength view (investigate the words one position to the left).

- Do any of these "extreme combinations also have interesting frequency profiles (not single occurrences in the concordance). Investigate by looking for works apparent in frequency and collocation strength views.

- Did not find any particularly interesting frequent and strong collocations at position left+1.

- Search regular expression "-acy". Interested in keyword frequency and collocations.

- Note democracy is 76 percent of "-acy"occurrences.

- Looking at other frequent keywords (aristocracy, bureaucracy): they are mostly negatively framed in the concordance lines.

- Switch to concordance strength view highest ranked keyword is "mediaocracy".

- Search "mediaocracy" 10 lines returned

- Use stop-word mosaic and browser to establish the (semantic prosody) negative or positive usage of the term.

- Hypothesis: Democracy is the dominant -acy and is viewed in a positive way. All other "-acys" are presented as negative. They are presented as threats to democracy.

This seems to make heavy use of the Mosaic tool for analysis. The case study presented seemed to be a partial treatment of the problem and may have skipped some steps which were needed to reach the hypothesis. Dave was asked the following clarifying questions:

- You moved swiftly from removing the file mod8 to investigating collocation strength. After removing the file mod8 you did not reinvestigate the collocation frequency of democracy and instead moved on to collocation strength. Why?

  > 'Just for demonstration purposes. In essence, not only were pieces skipped over, the illustration was also fairly preliminary in the following sense: Removing mod8 because it creates some distortion is of course bad practice would this be the actual research. The point in doing so is to quickly weed out material unfit for my purposes, until I reach a suitable point of investigation (in this case: democracy turning from one of the competing systems of rule into the only one available, however constantly beleaguered by traits from within). Once this point of investigation is established, the analysis can start out again and I make sure to construct a suitable sub-corpus on clearly defined terms that doesn't require me to be rash at the outset of an analysis. The mosaic view can then be approached again as an entry into the data, and all the collocation patterns examined more closely. "

- How do you initially decide what sub-corpus to investigate? In this analysis, books from 1970 to the present date.

  > "Currently the first thing I do (especially when the concordance return is small) is look at overlaps in meta-data property between concordance lines, to get a sense of the whereabouts of the data."

- Would a visualisation which shows frequency of a keyword across meta-data facets be useful?

> "Yes. One could, for example, look at differences in dispersion in the use of the word 'terror' pre – and post- 9/11, look at whether a certain author evades a word (say, anarchy) that is used by all other authors writing on the same subject (say, democracy), one could look at whether a magazine has a regional, national or international outlook by comparing the proper names used with those in other magazine, etc."

- In your analysis I struggle to see why you begun analyzing the "-acy" concordance. It doesn't appear to follow from the previous steps of analysis. Is this an established next step in corpus linguistics? Is it based on experience and domain knowledge or some part of the analysis not presented?

> 'This has to do with the reduction of bias through the reliance on form. I could, for example, go look at democracy vs. totalitarianism (in my attempt to study contemporary forms of government), but I have no proof that these concepts in fact are alternatives to each other. This would be solely based on intuitions, and as a lot can be argued about language data, I would basically come to prefabricated conclusions if I wanted to (democracy is opposed to totalitarianism in the following senses). Starting out from taking the suffix –cracy and seeing what other terms it attaches to offers a more neutral entry into the data inspired by the actual linguistic form rather than preconceived oppositions."

- You didn't appear to investigate the collocations of democracy and other (-acys) to determine the usage or context in which they occur, except for meritocracy. I am assuming that this was done and just not shown?

> "Indeed, in the final analysis every term discussed merits close attention to the immediate co-text."

- You use mosaic extensively in the method? Is that typical of your work

> "I use the Mosaic every time I access the corpus. Especially at the beginning of an analysis, to get an idea where to start and to make sure I won't, in a later stage, overlook any significant patterns. "

- You appear to use the collocation strength view for analysis, what is your opinion on it?

> "Useful for analysis as it gives extreme combinations. (where the combination rarity is interesting). As it stands the analysis done using the Collocation strength view is difficult to explain. Justifications for the patterns found using this view are usually easier to re-frame as part of the qualitative analysis which involves reading the concordance lines. "

- If other statistical measures were available in the mosaic would that be useful?

> "Yes, we would benefit from a measure of confidence rather than strength, or from a commonly known measure that can simply be mentioned as such in publications. "

8.2 ComFre Usage Example

An example case study called "Sketching Women" was presented by a GOK researcher who will be given the name Paul. The case study was described as "A corpus-based approach to representations of women in online political corpora in Arabic and English". Four corpora of differing sizes and type were used in the analysis. The GOK internet corpus at the time contained approximately 900,000 words, "the data is from left-leaning (political) websites (roughly from roughly 2000 to the present)". The Arabic Political Internet Corpus (APIC)contains slightly over one million words from "Arabic 'political' websites (dating roughly between 2012 to January 2018)". Two reference corpora, the enTenTen and arTenTen corpora

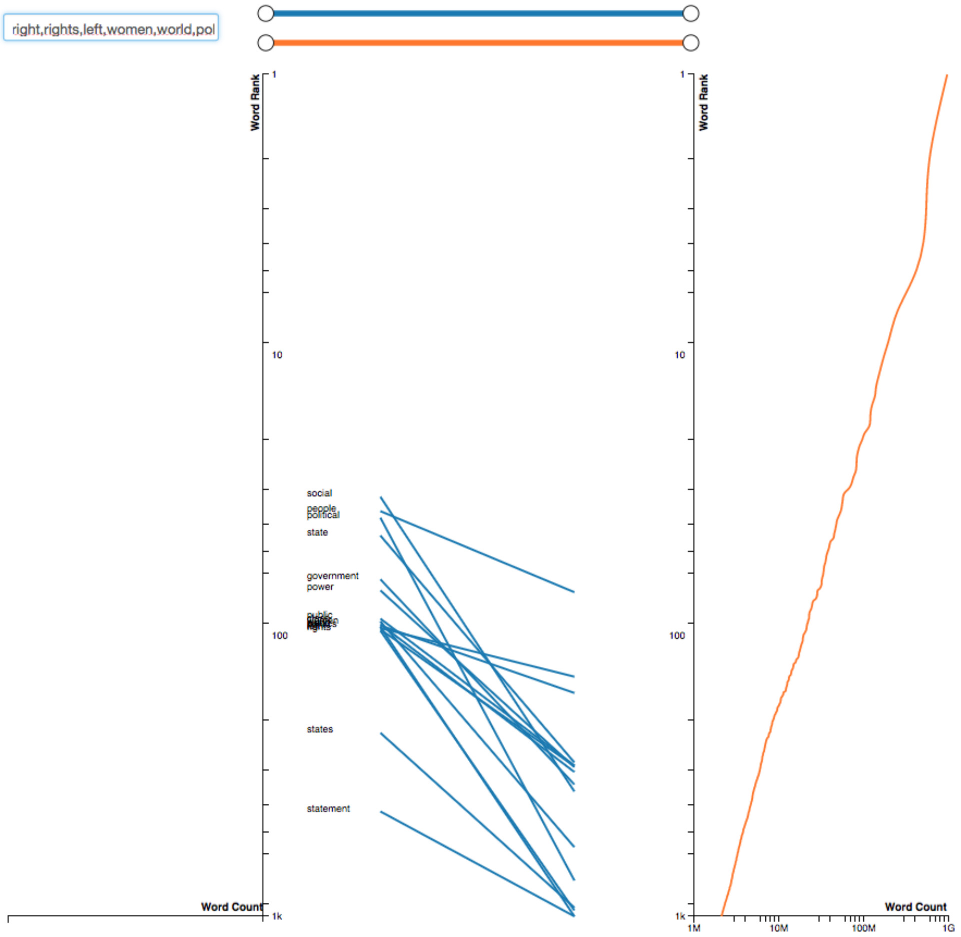


Figure 8.4: ComFre visualisation comparing high frequency political words in the GOK internet corpus(left) with the enTenTen corpus (right).

[Jakubíček et al., 2013], each contains roughly 10^{10} words. These are designed to be representative of general English and Arabic usage on the Internet.

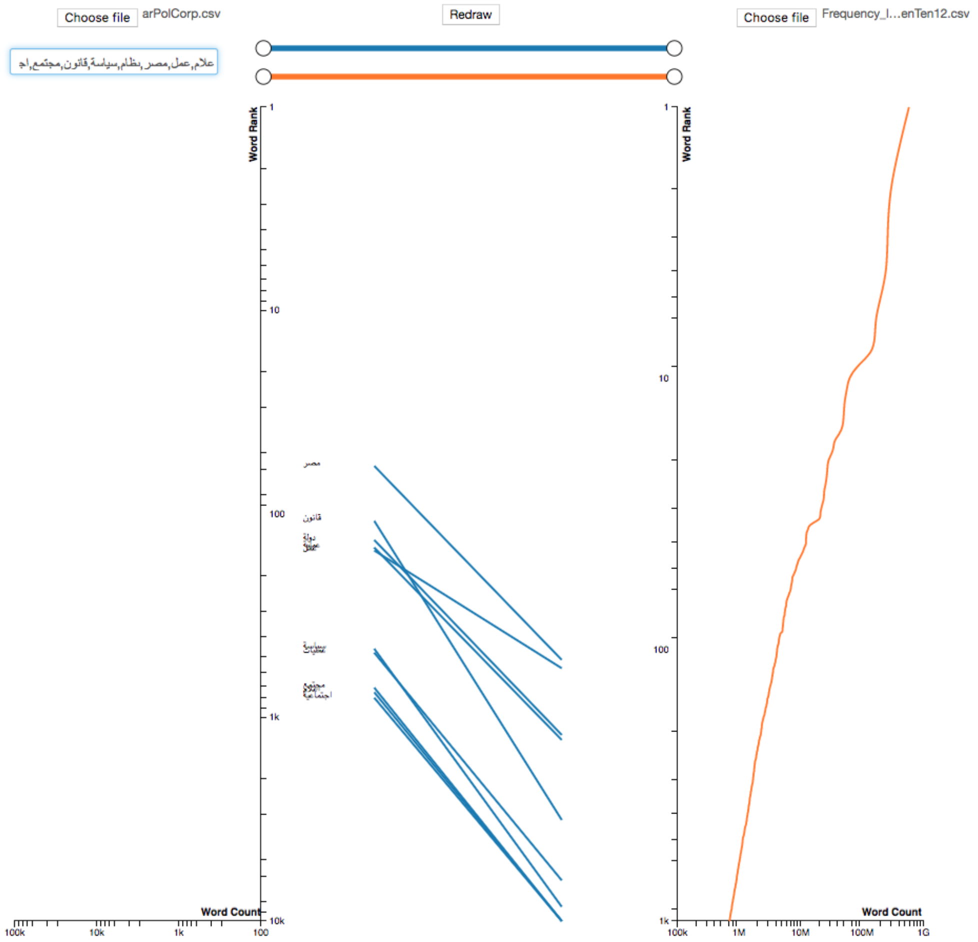
The methodology began by identifying high frequency political words in the main corpora GOK Internet and APIC. This was done by reading the frequency lists. Once these lists had been compiled frequency could be compared between the main corpora and the reference corpora. The words identified in the GOK internet corpus were:

> “ Social people political state government power public class women world politics right left rights human movement democracy states workers system movements citizenship just society working life economic process democratic”

In the APIC paper, the Arabic keywords after translation were given as:

> "state united media work Egypt regime politics Iran law society Israel Arab government authority rights states politician police Syria social Jerusalem women violence"

Since the corpus comparison is between a main corpus and a reference corpus of vastly different sizes, a direct comparison of rank in the list would be meaningless. This is where Paul turned to ComFre to aid in comparison.



علام,عمل,مصر,نظام,سياسة,قانون,مجتمع,اج

Figure 8.5: ComFre visualisation comparing high-frequency political words in the Arabic Political Internet Corpus(left) with the arTenTen corpus(right).

Since the ComFre tool enables comparison of the distributional position, meaningful comparison between usage in general language and the language specific to the main corpus could be made. Looking at Figure 8.4 and Figure 8.5 the slope lines for the keywords all have negative slopes from the main corpus to the reference corpus. This shows that the keywords represent the main corpora more than the reference ones. Paul concluded that the chosen words are all more common in the Main corpus for both Arabic and English. He rejected his null hypothesis that “The frequency of these terms is random”. Paul claimed there is strong evidence that the “The discourse in both corpora is a political discourse.”

Paul went on to describe how the analysis continued by the application of statistical tests to the main corpora. Mutual Information, t-test and log Dice association measures were used to identify salient collocations as a whole in the corpora. This analysis presents pairs of terms which have a high statistical association. Concordance analysis of the terms identified would then be used to make arguments about the language used in the corpus.

One point about Figure 8.4 and Figure 8.5 is that the frequency lists for the main corpora were not formatted exactly as required in this example. Unfortunately, this caused the distributional lines not to render. This had no effect on the slope lines and frequency comparisons, which were needed for the analysis.

8.3 Conclusion

Both researchers whose methods were examined used the Concordance Mosaic extensively in their case studies, suggesting its utility to the target user group.

However, the observed methodologies did not make use of ComFre. This may be due to the lack of familiarity of the researchers with the tool owing to its having been made available in the GOK software system two months before this review.

Daisy suggested an additional analysis where neither Concordance Mosaic nor ComFre was used. In this analysis, Concordance Mosaic and ComFre could not assist with the time-consuming task of evaluating metadata frequencies concerning the files making up a concordance list. In the next chapter (chapter 9, the domain characterisation is reexamined using this new observational evidence and the design of MetaFacet visualisation for concordance metadata frequency analysis is presented. The MetaFacet Visualisation is the third main contribution of this thesis.

Chapter 9

MetaFacet Visualisation

9.1 Introduction

Corpus often contain detailed meta-data associated with each file in the corpus. Facets of the meta-data can be displayed as part of a concordance list, filename is presented with every concordance in the GOK concordance browser, and an entire meta-data file can be extracted for each concordance line. Other concordance browsers such as Antconc [Anthony, 2005], SketchEngine [Kilgarriff et al., 2008] and WordSmith Tools [Scott et al., 2001] also display filename along with the concordance list. It is easy to imagine this filename could be replaced with any other meta-data attribute. However, this approach still doesn't make the calculation of concordance line dispersion across the facets much easier. Quantitative information is only available by counting the number of lines attributed to the meta-data tag.

In the domain characterization chapter 3, sub-corpus selections seemed to be regarded as an adequate method of partitioning the corpus on meta-data facet analysis, such as a group of authors. However, in light of the evidence provided by the methodological review chapter 8, this technique requires a large investment of time from the analyst before meaningful results can be observed. This is caused by the need to manually partition the corpus many times using sub-corpus selections and recording the keyword frequency. Only once all the frequencies have been recorded can the overview of the meta-data partition be examined. One technique I observed for this meta-data examination was the creation of bar charts in a spreadsheet application.

If large-scale quantification of concordance line distribution is required, visualisation is a good solution as the pattern of overview and detail easily maps to this task. The overview could take the form of some visualisation of keyword count in each attribute of a meta-data facet, while the detail would be the individual concordance line associated with a facet attribute. A technique based on this principle should allow quick examination of a concordance list in terms of its distribution across all meta-data facets.

I now present MetaFact, a meta-data frequency explorer which interacts with the GOK concordance browser and Mosaic plugin.

9.2 Requirements

The methodologies and comments which were presented in chapter 8 had several references to the difficulty of the task of meta-data concordance analysis. In addition, both researchers expressed interest in visualisation-based solutions to this problem. On researcher currently manually constructs bar charts to help with the analysis. Creating these bar charts in the current method is a slow and laborious task.

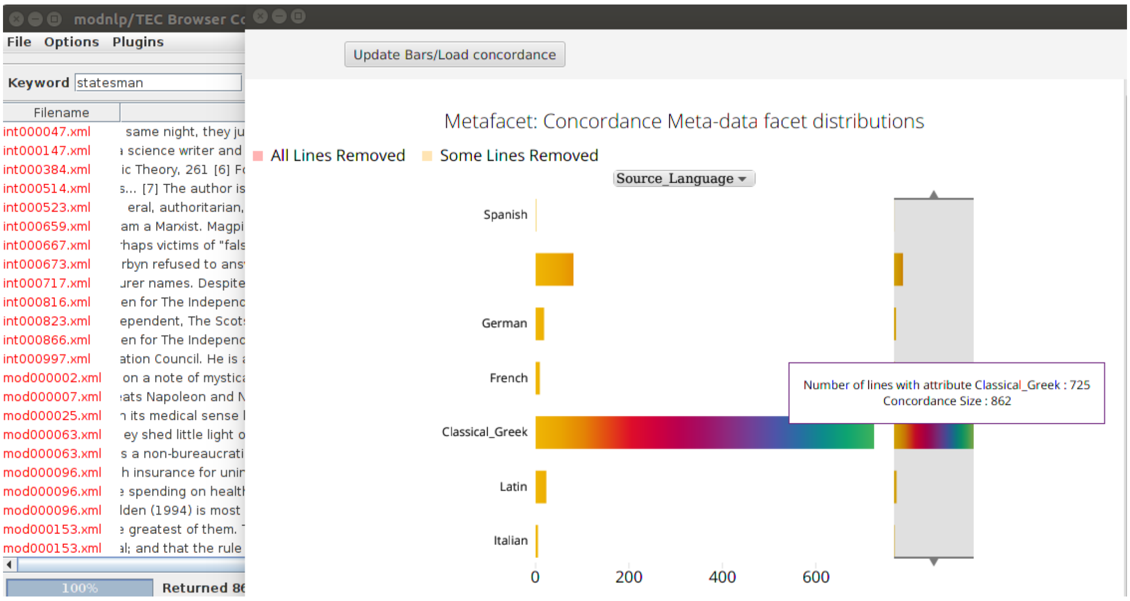
At a design meeting to discuss the construction of a visualisation to assist with meta-data concordance analysis, specific requirements were set out. They are as follows:

1. Created as a plug-in for the GOK concordance browser
2. Automatically extract meta-data for a concordance list
3. Visualize this meta-data, displaying the number of concordance lines per meta-data attribute
4. Enable filtering of the concordance list and Mosaic

In the discussion, it was also specified that a successful solution would be one which greatly reduced the time and labour required for methodologies such as the statesmanship case studies described in chapter 8.

9.3 Visual Encoding Design

To create MetaFacet, a system for meta-data extraction was added to the GOK corpus browser. Once this was in place, meta-data could be efficiently extracted for any concordance list for processing by a plug-in. MetaFacet operates on concordance lists and displays the metadata attribute counts associated with the lines of a concordance.



Keyword statesman

Filename	
int000047.xml	same night, they ju
int000147.xml	science writer and
int000384.xml	ic Theory, 261 [6] Fo
int000514.xml	s... [7] The author is
int000523.xml	eral, authoritarian,
int000659.xml	am a Marxist. Magpi
int000667.xml	haps victims of "fals
int000673.xml	rbyn refused to ans
int000717.xml	urer names. Despite
int000816.xml	en for The Independ
int000823.xml	ependent, The Scot
int000866.xml	en for The Independ
int000997.xml	ation Council. He is
mod000002.xml	on a note of mystica
mod000007.xml	ats Napoleon and N
mod000025.xml	its medical sense
mod000063.xml	ey shed little light o
mod000063.xml	s a non-bureaucrati
mod000096.xml	h insurance for unin
mod000096.xml	spending on health
mod000096.xml	lden (1994) is most
mod000153.xml	greatest of them.
mod000153.xml	al; and that the rule

Number of lines with attribute Classical_Greek : 725
Concordance Size : 862

Figure 9.1: Metafacet display of source language for the keyword “statesman” in the full GOK corpus.

The core encoding choice for the MetaFacet is a simple one. For any meta-data facet, a quantitative value is associated with each attribute of the facet. As an example, for the facet “source language” and the keyword “statesman” there is an exact number of concordance lines associated with each of the attributes “English”, “Classical Greek”, and “Latin”. From [Mackinlay, 1986] (see Figure 2.7), the best variable to map to a quantitative variable is position closely followed by length.

I decided to encode the quantitative variable using length. The GOK researcher Daisy had previously used bar charts for this type of analysis, and visual similarity to established methods should influence encoding decisions for domain-specific visualisations.

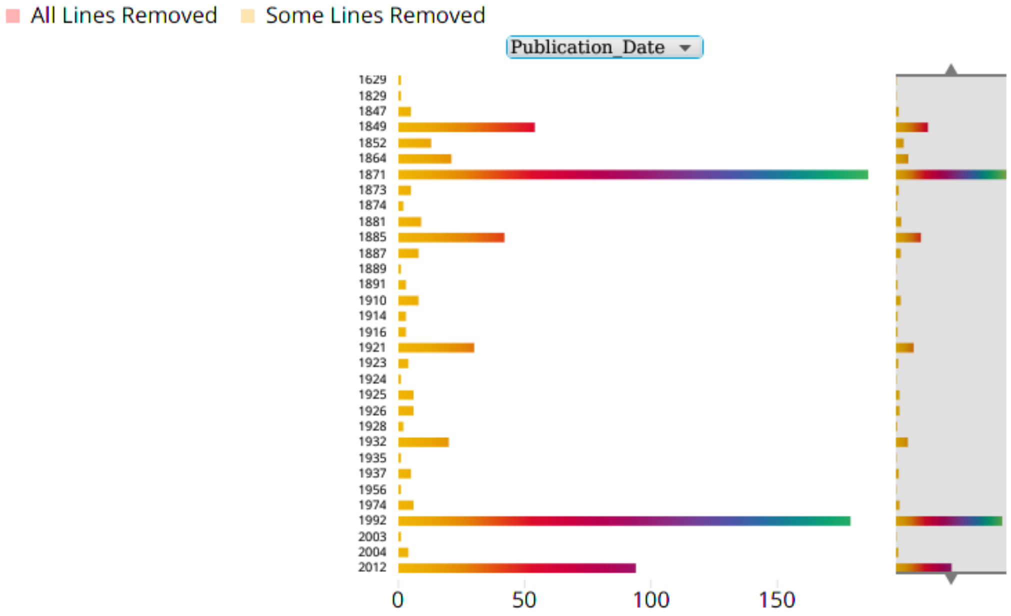


Figure 9.2: MetaFacet view of "Publication Date" facet in the GOK Classical Greek subcorpus for the keyword "statesman".

The MetaFacet interface can be seen in Figure 9.1. The encoding shown was created as a plugin for the GOK Concordance Browser using the JavaScript visualisation framework *D3.js* [Bostock et al., 2011]. The interface is based on horizontal bar charts, where the categorical variables are placed on the vertical position axis, and horizontal length is used to encode the quantitative values associated with the categories. Each positioned bar is labelled with its attribute. The attribute categories in MetaFacet are displayed in lexicographical and numerical order. This is useful for quickly identifying individual attributes in the overviews and for observing patterns in numerical categories such as dates. In Figure 9.1 the horizontal axis helps with identifying the quantities associated with each bar, in addition a hover tooltip is shown, which presents the attribute label, associated line count, and the total number of concordance lines in current concordance.

Creating multiple facet overviews on a single rendering was considered but rejected to help reduce visual clutter. Instead, a drop-down menu containing each meta-data facet was provided. This enables quick switching between meta-data facets(such as *source language* Figure 9.1 and *publication date* Figure 9.2). If a particular facet contains a large number of attributes, it can become difficult to read the labels. It would be useful to be able to move from the overview of the facet to a smaller sample of its attributes. The range window selector enables this functionality while keeping the original overview visible. The window size is adjustable and can be dragged to different sections of the overview facet overview. In Figure 9.3, a full overview of the facet translator is shown. In Figure 9.4, a smaller range of the facet has been selected, and a summary of the overview can still be seen on the range selector tool.

MetaFacet was designed in the context of the GOK corpus browser and the existing plug-ins. One of the design requirements identified by the expert users was the ability to filter the concordance list and mosaic results using the meta-data. Since the meta-data is implicitly linked to the concordance lines, this type of filtering was possible. The meta-data view presents the attributes and facets found for a concordance of a specific keyword. I created a click interaction which would filter the concordance list by removing the lines associated with the clicked attribute. The clicked attribute turns red and can be clicked a second time to return the concordance to its original state. Multiple attributes across all facets can be filtered simultaneously. If the user wants to reevaluate the filtered list's meta-data attributes, clicking the update bars button will recalculate the attributes for the filtered list. Searching the original keyword again will restore the concordance to the pre-filtered state. In Figure 9.3 a concordance list for the keyword "statesman" in the GOK Classical Greek sub-corpus has been filtered from hundreds of concordance lines down to thirteen. This was done by selecting the translators highlighted in red in the MetaFacet browser.

Similarly, the click interaction can filter the Mosaic view of the concordance. In Figure 9.5, the Mosaic is shown partially obstructed by the MetaFacet interface. This is the Mosaic of the Keyword "statesman" for the GOK Classical Greek sub-corpus with stop-words removed. Looking at the MetaFacet, the Author "Plato" accounts for the majority of the usage in the sub-corpus. Clicking on the bar representing "Plato" in MetaFacet removes lines associated with the author in both the concordance browser and Mosaic. In Figure 9.6, the Mosaic has been filtered, and analysis of frequent collocation patterns in classical Greek works not written by Plato can be Analyzed.

9.4 Requirements Fit

To demonstrate the use of the MetaFacet visualisation in supporting corpus linguistic methodologies, I reexamined a portion of the "Statesman" case study from chapter 8. In this example, MetaFacet is used to perform the previously completed analysis using the concordance browser and a spreadsheet.

The case study begins by searching for the keyword statesman in the GOK English corpus. The Original study claimed that Classical Greek translations account for over ninety percent of the concordance lines. Looking at the MetaFacet display of the facet "Source Language" in Figure 9.1 confirms that this is true . The hover tool-tip for the "Classical Greek" attribute shows that 725 of the 862 concordance lines are translations from Classical Greek. A similar method can be applied to other keywords of interest to identify patterns in the source language.

Next, the corpus is refined by selecting just the sub-corpus of "Classical Greek". Doing this with the sub-corpus selector(as shown in Figure 9.7) instead of filtering using MetaFacet, ensures all proceeding analysis will be in the correct sub-corpus even after resetting the concordance lines. In this sub-corpus, a concordance for "statesman" is, again, generated, and MetaFacet is reloaded to reflect the new data.

Analyzing the frequency attributes is the core activity of the original study. In particular, the author and publication date were found to be of interest. Investigate publication date with Metafacet. The "Publication date" facet is simply selected from the dropdown menu, and the distribution can be investigated. Figure 9.2 shows this distribution of publication dates. The "Author" facet can then be selected, as seen in Figure 9.5, to view the distribution of authors across the subcorpus for the keyword "statesman".

While the case study of the keyword "statesman" did not make use of collocation patterns, another case study on the concept of "the people" did. It is easy to show how the analysis of "statesman" could be extended to analyzing faceted collocation patterns. As an example, the Mosaic shown in Figure 9.5 is dominated by collocation patterns found in the works of the author "Plato", this can be clearly seen in the MetFacet window in Figure 9.5. Viewing the Mosaic without any concordance lines from the dominating author may be interesting. Filtering the Mosaic as shown in Figure 9.6 shows a significant change in the collocation patterns visible in the Mosaic. This filter interaction also operates on the concordance making the close inspection of concordance lines with desired meta-data attributes possible.

Including the meta-data plugin in the GOK Concordance Browser removed most of the time-consuming work required for "faceted" exploration of a concordance list. Processing of the data is performed by the tool, giving more time to the analyst to investigate the patterns and features of the corpus. The original method of analyzing meta-data facets in the case study of "statesman" was estimated to take five to six hours. Using MetaFacet, the same analysis takes minutes, not hours.

9.5 Conclusion

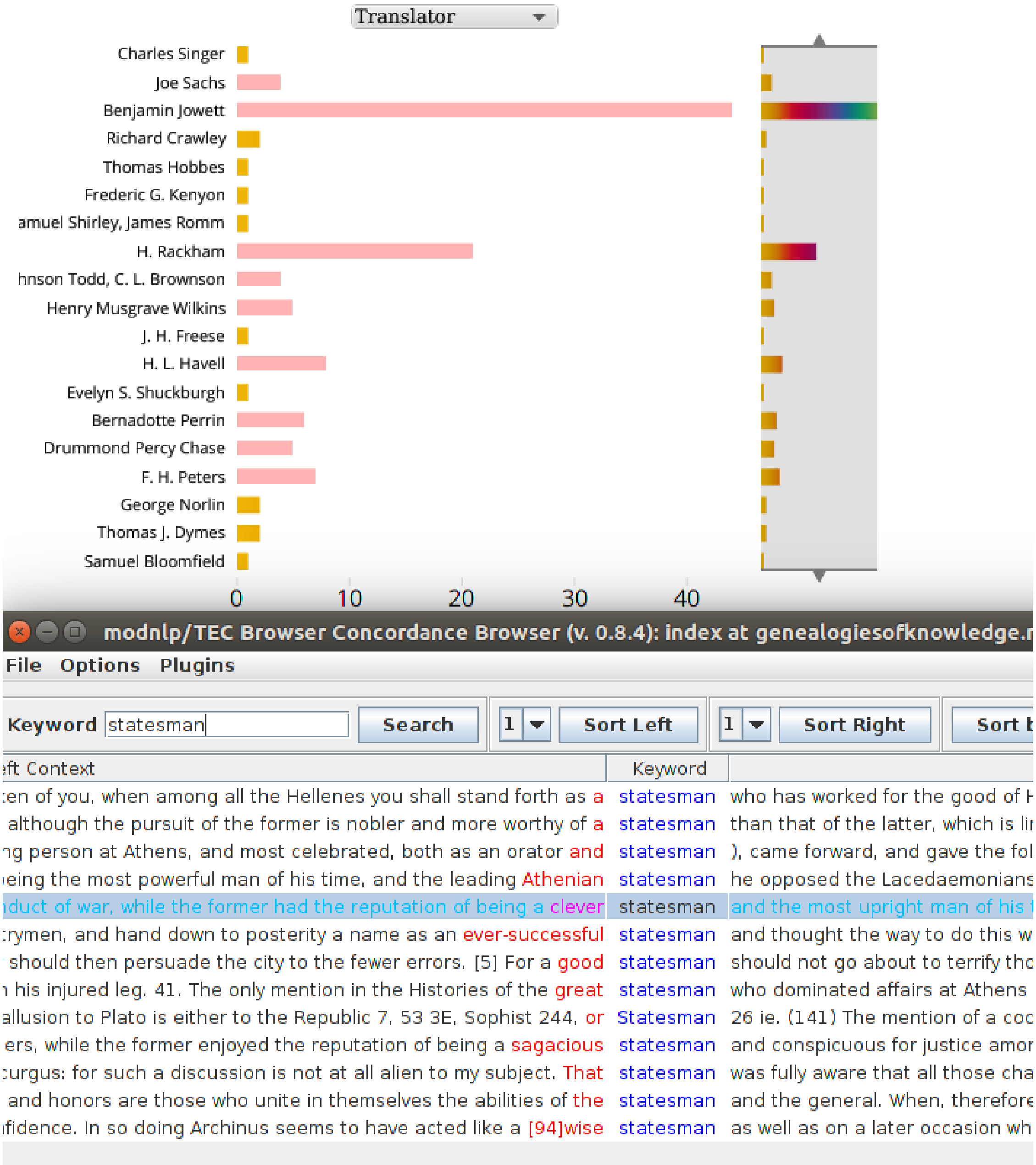
In the next chapter (chapter 9, the domain characterisation is reexamined using this new observational evidence and the design of MetaFacet visualisation for concordance metadata frequency analysis is presented. The MetaFacet Visualisation is the third main contribution of this thesis.

The MetaFacet visualisation is designed to enable the exploration of the distribution of concordance lines across meta-data facets.

The problem of quantitative evaluation of the concordance list in terms of its meta-data facets was not identified in the initial exploration of the domain characterization (chapter 3). The problem was only identified during the methodological review (chapter 8), where a method that heavily used meta-data for analysis was examined. The data abstraction for this visualisation was validated via evidence of utility in the practical steps expert users took in the methods described.

The reenactment of a method observed in the methodological review showed that MetaFacet reduces the time to complete some analysis steps. The domain experts stated that using the MetaFacet visualisation, meta-data analysis tasks take minutes instead of hours.

MetaFacet is the third and final main contribution of this thesis. In the next chapter (chapter 10), the contributions of this thesis are discussed, and conclusions are drawn.



Keyword statesman

ɛft Context	Keyword	
:en of you, when among all the Hellenes you shall stand forth as a	statesman	who has worked for the good of H
although the pursuit of the former is nobler and more worthy of a	statesman	than that of the latter, which is lir
ıg person at Athens, and most celebrated, both as an orator and	statesman	), came forward, and gave the fol
ɪeing the most powerful man of his time, and the leading Athenian	statesman	he opposed the Lacedaemonians
ıduct of war, while the former had the reputation of being a clever	statesman	and the most upright man of his t
:rymen, and hand down to posterity a name as an ever-successful	statesman	and thought the way to do this w
should then persuade the city to the fewer errors. [5] For a good	statesman	should not go about to terrify thc
ı his injured leg. 41. The only mention in the Histories of the great	statesman	who dominated affairs at Athens
allusion to Plato is either to the Republic 7, 53 3E, Sophist 244, or	Statesman	26 ie. (141) The mention of a coc
ers, while the former enjoyed the reputation of being a sagacious	statesman	and conspicuous for justice amor
:urgus: for such a discussion is not at all alien to my subject. That	statesman	was fully aware that all those cha
and honors are those who unite in themselves the abilities of the	statesman	and the general. When, therefore
ıfidence. In so doing Archinus seems to have acted like a [94]wise	statesman	as well as on a later occasion wh

Figure 9.3: MetaFacet window (Top for facet "Translator" in GOK Classical Greek sub-corpus for the keyword "statesman". Several translators (highlighted red in MetaFacet) are filtered out of concordance browser (bottom).

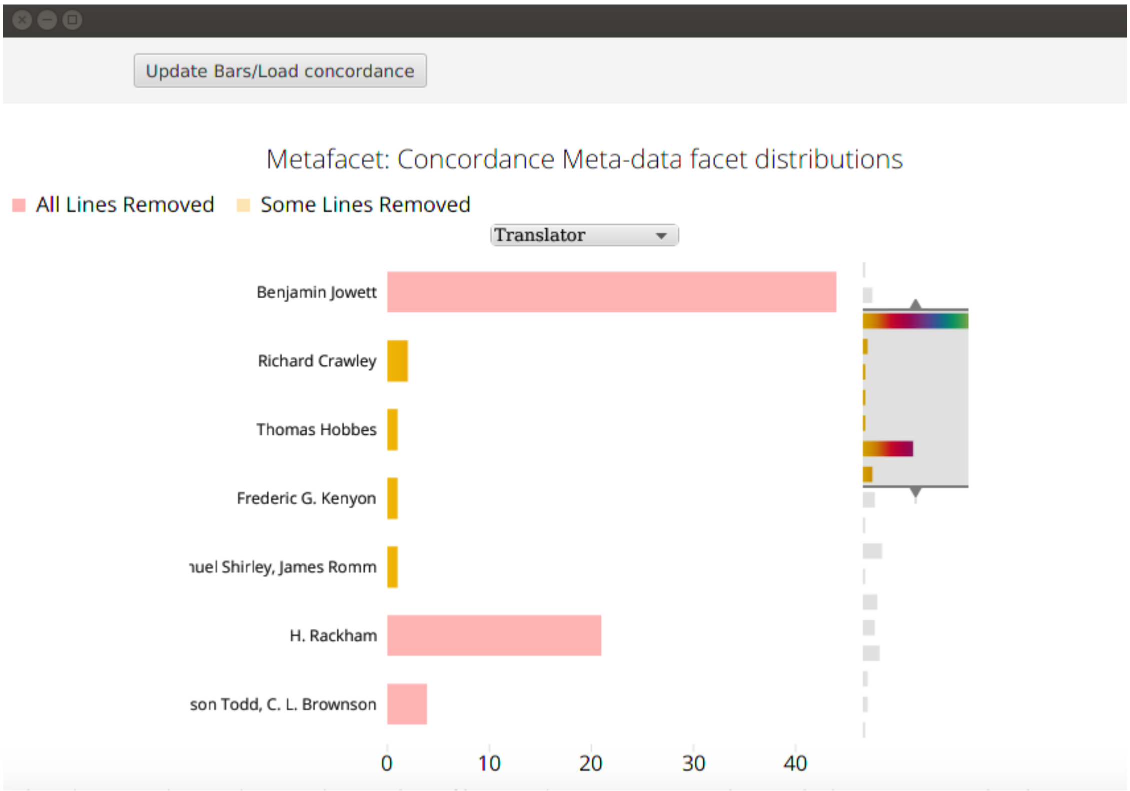


Figure 9.4: MetaFacet window for facet “Translator” in GOK Classical Greek sub-corpus for the keyword “statesman”. In this window, the range selector has been used to narrow the contest to a sample of the total “Translator” facet.

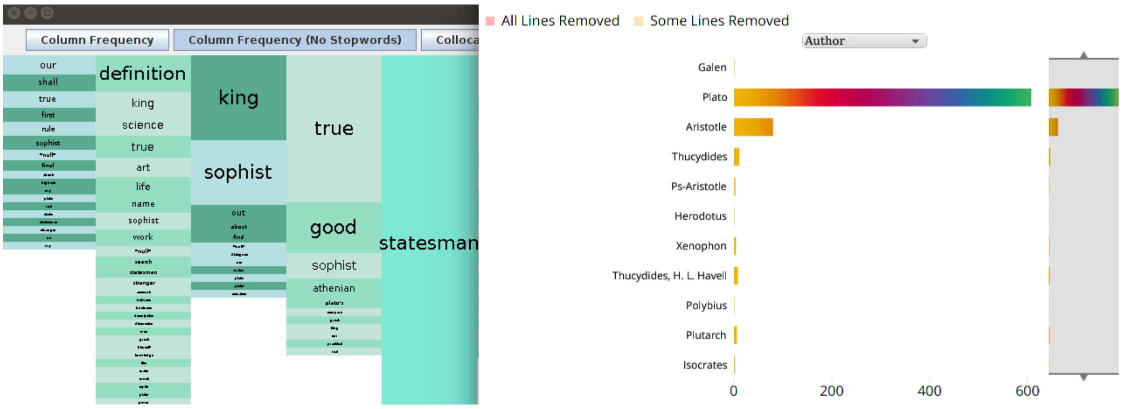


Figure 9.5: Mosaic and MetaFacet windows for the keyword “statesman” in GOK Classical Greek sub-corpus. MetaFacet for the facet “Author” is visible.

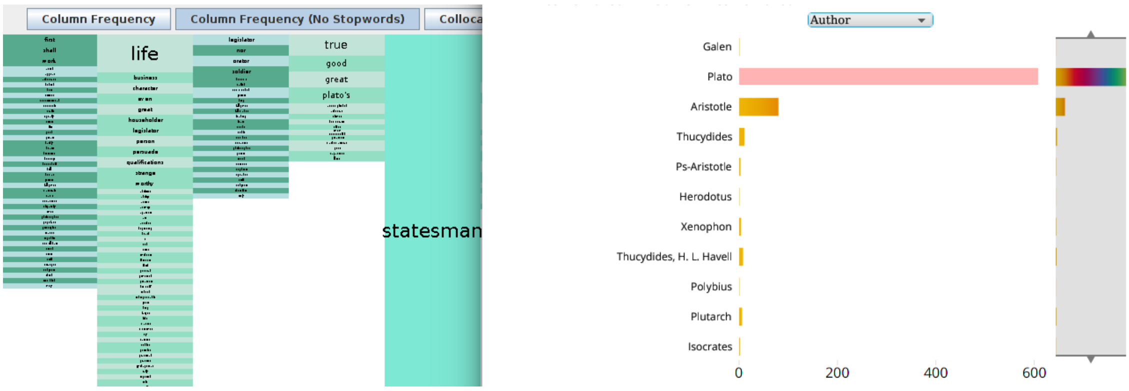
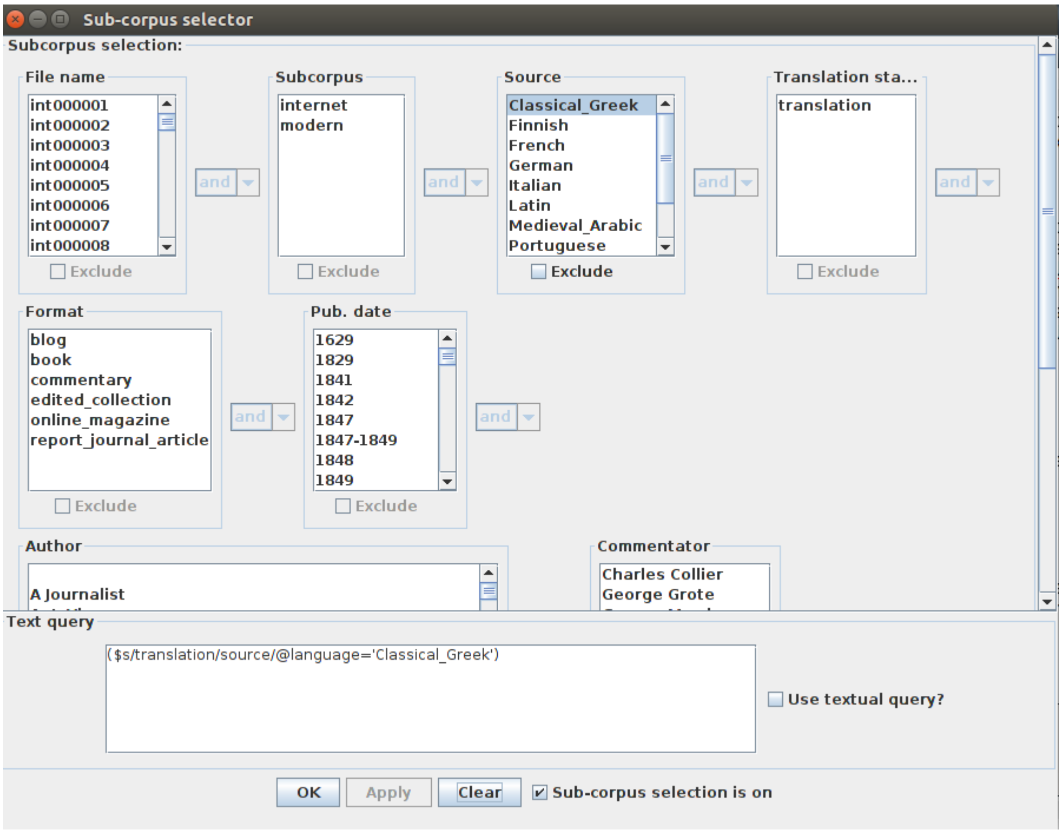


Figure 9.6: Mosaic and MetaFacet windows for the keyword “statesman” in GOK Classical Greek sub-corpus. Concordance lines associated with the author Plato have been filtered from the Mosaic by clicking on ‘Plato” in the MetaFacet display.


Sub-corpus selector

Subcorpus selection:

File name: int000001, int000002, int000003, int000004, int000005, int000006, int000007, int000008 — □ Exclude

Subcorpus: internet, modern — □ Exclude

Source: Classical_Greek (selected), Finnish, French, German, Italian, Latin, Medieval_Arabic, Portuguese — □ Exclude

Translation sta...: translation — □ Exclude

Format: blog, book, commentary, edited_collection, online_magazine, report_journal_article — □ Exclude

Pub. date: 1629, 1829, 1841, 1842, 1847, 1847-1849, 1848, 1849 — □ Exclude

Author: A Journalist

Commentator: Charles Collier, George Grote

Text query: ($s/translation/source/@language='Classical_Greek')

□ Use textual query?

☑ Sub-corpus selection is on

Figure 9.7: Sub-corpus selection window showing the selection of the Classical Greek sub-corpus.

Chapter 10

Discussion and Conclusions

10.1 Introduction

In this final chapter, I discuss the main findings of the thesis and how the approach I have taken in this thesis has led to these findings in the domain of information visualisation applied to corpus linguistic methodologies.

The literature on visualisation design argues that for a domain-specific visualisation to be successful, it must overcome four threats to validity. Each threat is exhibited at one of the four levels of the *nested model for visualisation design and validation*. As part of this research, I developed three visualisations addressing different aspects of corpus linguistic methodologies. The validity of each of these visualisations is discussed in terms of the research presented in this thesis.

10.2 Research Questions

The research questions which directed the work presented in the thesis were:

1. Which methods employed by corpus linguists are not well served by visualisation?

2. How would visualisation tools affect the workflows of corpus linguists?

3. Can collocation patterns be visually summarized effectively?

4. Can unequal-length word frequency lists be effectively compared using visualisation?

5. Can meta-data visualisation supplement the concordance methodology?

The first research question directs the search for methods in corpus linguistics which currently are not adequately addressed by visualisation tools. Through initial discussion with domain experts, concordance analysis methodology was identified as an area of corpus linguistics where visualisation tools could assist.

To better clarify which methods within concordance-based corpus analysis should be targeted for visualisation support domain characterisation efforts were undertaken. These efforts also help alleviate the validity threat from the nested model of visualisation design explored in subsection 2.2.3, of addressing the wrong problem with the designs. Analysis of Sinclair's tutorials for conducting concordance analysis helped identify actions which are fundamental components of the tasks which are often performed in the methodology. Through structured interviewing techniques based on a twenty-question model (subsection 2.2.4), domain experts helped guide the identification of which actions are most important and challenging in their analysis of corpora.

Two key findings of this process were that positional collocation frequency and corpus level frequency list comparison are fundamental aspects of the concordance-based methodology in which members of the target user group identified could benefit from visualisation support. Analysis of existing visualisations for both collocation frequency and frequency list comparison found that there are very few visualisation interfaces which address these topics. Even spreading the search wider to encompass visualisations which could be repurposed for these tasks, there was very little relevant recent literature. Those found were compared using a structured encoding comparison. The encodings which address collocations were found to either not encode positional frequency or encode it using non-optimal visual variable choices. For frequency list comparison visualisations, the encoding did not represent frequencies from unequally sized lists in a manner that makes them directly visually comparable.

To answer the second research question about how tools might influence the workflows of corpus linguists, the tools themselves were first designed and developed. This design and development process used the requirements gathered from the domain characterisation efforts. The Concordance Mosaic translated them into a visual encoding using efficient visual variables for the data attributes identified as important in the conceptual model I developed based on the actions identified in Sinclair's tutorial. In this way, the requirements were addressed in a structured manner based on visualisation design principles rather than some loose interpretation of a requirements document.

ComFre emerged as a requirement through traditional domain characterization efforts focused on target users who are domain experts and is more loosely related than Concordance Mosaic to the domain literature task analysis based on Sinclair's work. Its encoding was designed primarily to enable direct visual comparison of frequency lists of different sizes, a core requirement which emerged in the interviews and discussions.

The Metafacet visualisation requirements were identified through the methodology review, where examples of the usage of the tool by the target user group was observed, recorded and analysed. The initial requirement of metadata analysis as a part of the concordance analysis method was overlooked as it was not given great weight in discussion with the domain experts. However, observation of the tools in use made it clear that metadata analysis should also be supported by visualisation in some way. This shows how important it is to validate the assumptions that have led to the designs produced and to use design models to help validate and guide the process at each stage.

By supporting the quantitive actions identified in all of the domain characterization efforts, the interfaces enhance the concordance analysis method, which is a type of exploratory corpus analysis. The mosaic tool enhanced the existing concordance analysis method by providing a method based on the systematic study of Sinclair's concordance analysis tutorials and domain characterization of the work of experts in the field. This method was shown to effectively encode quantitative information used in concordance analysis but not readily available in the typical tools of the trade (KWIC interfaces).

Concordance Mosaics, in both frequency and statistical views, directly enable visualization of abstract patterns of occurrence and collocation in a corpus. In combination with ComFre, where keywords which have different distributional positions from a reference corpus can be easily identified, the abstract nature of not just the words themselves but their place in the overall corpus and their collocates is explorable.

The focus of the domain experts in the target user group is often on discovering surprising and unsurprising collocations around a *concept*, represented by a set of keywords. While an analyst often chooses to review each keyword individually and combine the results of each analysis, it is possible to enter each of the keywords together, separated by commas, and to view a Concordance Mosaic of the collocations of the entire concept(represented by a set of keywords).

These are examples of the impact the interfaces have had on corpus linguistic workflows. As these interfaces are used "in the wild" other examples might emerge. Continued monitoring of tool usage will help mitigate further threats to the validity of the design. Evidence of tool adoption in the domain is the key post-design measure for validating the overall domain characterisation and problem identification.

Questions three, four and five are concerned with the specific problems addressed by each interface and are explored in the next three sections where we argue each interface is at present the most effective for the identified task.

10.3 Mosaic Visualisation

The first threat to the validity the Mosaic faces is if it addresses the wrong problem. Wrong in this sense would mean that even a good solution would not benefit the target users, who, in this case, are corpus linguists. The Mosaic visualisation was created to summarize collocation patterns effectively. This problem was discovered through a task analysis of literature which described the concordance methodology for corpus analysis. In the concordance methodology, patterns of collocation were found to be a central concern for many of the analysis tasks.

There are two methods often used to validate the domain characterization of a visualisation and test that the threat has been overcome. The first is to observe and interview target users. The second is to observe adoption rates for the visualisation. Two of the researchers from the GOK project provided case studies of their research methodologies and were interviewed extensively to help identify domain-specific tasks which could benefit from visualisation. Both researchers used collocation analysis in the methods used and gave details of important questions they would like to answer about a corpus by using collocation information. The researchers also indicated that using the concordance to examine positional collocation is difficult for large concordance. One researcher even created a mock-up (Figure 3.2) of a potential visualisation based on word clouds to solve the problem of collocation patterns in the context of large concordances .

The second method of validating the domain characterization is observing the visualisation's adoption rates. As the Mosaic visualisation is included as a plug-in for the GOK Concordance Browser, the adoption of this browser will heavily influence the wider adoption of the Mosaic visualisation by corpus linguists. There is, however, clear evidence of adoption from within the GOK project. In the methodological review, both researchers described methodologies that extensively use the Mosaic tool for identifying collocation patterns. In an interview about one of these methodologies, the researcher claimed to use the Mosaic every time he accessed the GOK corpus.

The second threat to validity is that the data abstraction for the visualisation is bad. This would mean that even though the correct problem is being addressed, the information being visualized is not suitable for solving the problem. This is to validate a data abstraction suggested techniques are testing on target users to collect anecdotal evidence of utility and field studies which document human usage of the deployed system. The methodological review can be viewed as a combination of both of these techniques. The Mosiac is in active use by researchers for identifying linguistic patterns related to a keywords. One problem discovered was that for the analysis of statistical collocation strength, the original simple metric was found to be inadequate for formal linguistic research. While it was useful for collocation strength-based pattern exploration, the underlining measure was not transparent to users. Discussing this issue with the domain experts in the GOK project led to the replacement of the collocation strength measure by measures of collocation strength more familiar to domain experts (such as Z-score and mutual information).

The threat of an ineffective encoding at the third level of the nested model can be mitigated by justifying the encoding and interaction design, performing a lab study which measures time and errors for operations or informal usability studies on any users. This thesis presented a lab study comparing the mosaic encoding to the traditional KWIC concordance. The key finding of the experiment is that the Mosaic performs better, in terms of time to complete tasks and errors, for each of the five collocation pattern-based tasks of the experiment. This is not a surprising result, the mosaic was designed to enable quantitive analysis of the concordance which is something users currently struggle with due to inappropriate encoding of the quantitive information which is implicit in the positions of the concordance.

The Mosaic encoding design has been justified in relation to the choice of visual variables for the actions identified in a task analysis of concordance methods. The encoding was further justified by a structured comparison detailing related visualisations' encoding choices.

The final threat to validity is that the implementation of the visualisation will be slow. The Mosaic is currently deployed and in use by corpus linguistic researchers. The current implementation functions well and has a worst-case time complexity of $O(n)$.

Having avoided the threats to well-designed visualisation, Mosaic effectively summarizes positional collocation patterns. The visualisation has had an impact on the work of some corpus linguists, as evidenced by its use in research case studies.

10.4 ComFre Visualisation

The ComFre visualisation was designed to enable a valid comparison of sets of items of unequal size. Its envisioned application in corpus linguistics is the comparison of frequency lists. In the domain characterization, comparing keyword frequencies across sub-corpora was identified as a core task in corpus analysis. How representative comparisons of corpora of different sizes can be achieved with frequency lists was unclear. Initial design decisions and requirements were generated with the assistance of expert users who identified challenges in comparing frequency lists from two or more corpora. Identifying this problem by interaction with domain experts helps to validate the problem characterization.

In the methodological review, the observed methodologies did not make use of the compare lists interface. This may be due to the familiarity of the GOK researchers with the tool and the short time availability for methodological adaptation. At the time of the review, the ComFre interface had only been integrated into the GOK concordance browser for two months, while the Mosaic had been available for over a year. One researcher did provide an example methodology to exemplify how ComFre is useful for analysis, as this methodology was presented as an example of how to use ComFre its validity as an example of methodological adoption is less valid.

To validate the data abstraction, the ComFre usage example is useful. It provides evidence of utility as it is a description from a domain expert of how the visualisation can be used for his methods.

The ComFre Encoding was validated using encoding justification with the visualisation requirements, visual variable choice, and relevant existing visualisations.

One problem the visualisation faces is the algorithmic complexity of processing new frequency lists, which has a worst-case time complexity of $O(n^2)$. Future work should focus on attempting to improve this algorithm so that a more efficient rendering of the interface can be achieved,

10.5 MetaFacet Visualisation

The MetaFacet visualisation is designed to enable the exploration of the distribution of concordance lines across meta-data facets. The visualisation is included as a plug-in for the GOK Corpus Browser and enables the interactive filtering of the Mosaic and concordance list.

The problem of quantitative evaluation of the concordance list in terms of its meta-data facets was not identified in the initial exploration of the domain characterization. It was incorrectly assumed that a sub-corpus based on meta-data attributes, such as a date range or Author, would be selected and analyzed in detail, through concordance or keyword analysis, without requiring further meta-data information. The problem was only identified during the methodological review, where methods that heavily used meta-data for analysis were examined. In these methods, partitioning the concordance for analysis and simple visualisation was a slow process that was not well supported by corpus tools. Since the inclusion of the MetaFacet visualisation in the GOK, no new methodological review has occurred, so adoption rates have not been assessed. The problem characterization and explicit statements of interest from GOK researchers during the methodological review make a strong case for a valid domain problem identification.

The data abstraction is well-validated since there is evidence of utility in the practical steps taken by expert users. The reenactment of a method observed in the methodological review showed that MetaFacet reduces the time to complete some analysis steps. The reduction in the analysis time is significant. The domain experts stated that using the MetaFacet visualisation, meta-data analysis tasks take minutes instead of hours.

The MetaFacet Encoding was validated using encoding justification concerning the visualisation requirements and visual variable choices.

10.6 Future Research Directions

There are many potential research directions in the area of corpus analysis which could be productively addressed using information visualisation techniques. Addressing temporal collocation patterns was an idea which was identified during the domain characterisation and could be further explored.

The exploration of the collocations of sets of keywords is enabled in the Mosaic. Extending this functionality to facilitate the unique challenges of multi-keyword concordances better is an interesting problem. In addition, constructing these keyword sets dynamically or automatically could be a productive research direction.

Natural language processing techniques could also be used to enhance the analysis of concordance and to offer new opportunities for visualisation tools. These tools could aid in the generation and interpretation of machine-generated insight. Given recent trends in the field, directly asking quantitative questions of the corpora using natural language is an interesting interface where visual exploration of the output could be useful to language scholars.

Finally, adding non-textual elements (such as images, audio, and video) to corpora is becoming more common. Integrating these elements into concordance analysis will likely benefit from visualisation support in the future.

Bibliography

[Anthony, 2005] Anthony, L. (2005). Antconc: design and development of a freeware corpus analysis toolkit for the technical writing classroom. In *IPCC 2005. Proceedings. International Professional Communication Conference, 2005.*, pages 729–737. IEEE.

[Baker, 1995] Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target. International Journal of Translation Studies*, 7(2):223–243.

[Baker, 1996] Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. *Benjamins Translation Library*, 18:175–186.

[Baker et al., 1993] Baker, M., Francis, G., and Tognini-Bonelli, E. (1993). *Corpus Linguistics and Translation Studies: Implications and Applications.* John Benjamins Publishing Company, Netherlands.

[Baker, 2006] Baker, P. (2006). *Using Corpora in Discourse Analysis.* Bloomsbury Discourse. Bloomsbury Academic.

[Barlow, 2011] Barlow, M. (2011). Corpus linguistics and theoretical linguistics. *International Journal of Corpus Linguistics*, 16(1):3–44.

[Bertin, 1983] Bertin, J. (1983). *Semiology of Graphics.* University of Wisconsin Press.

[Biber et al., 1998] Biber, D., Douglas, B., Biber, P., Conrad, S., and Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge Approaches to Linguistics. Cambridge University Press.

[Boas, 1941] Boas, F. (1941). Race, language and culture. *The Journal of Nervous and Mental Disease*, 94(4):513–514.

[Bonelli, 2010] Bonelli, E. T. (2010). Theoretical overview of the evolution of corpus linguistics. *The Routledge Handbook of Corpus Linguistics*, page 14.

[Bostock et al., 2011] Bostock, M., Ogievetsky, V., and Heer, J. (2011). D3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309.

[Calzada-Pérez and Luz, 2006] Calzada-Pérez, M. and Luz, S. (2006). ECPC: Technology as a tool to study the (linguistic) functioning of national and trans-national European parliaments. *Journal of Technology, Knowledge and Society*, 5(2):53–62.

[Chi, 2002] Chi, E. H. (2002). Expressiveness of the data flow and data state models in visualization systems. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, pages 375–378.

[Chi, 2000] Chi, E. H.-h. (2000). A taxonomy of visualization techniques using the data state reference model. In *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings*, pages 69–75. IEEE.

[Chi and Riedl, 1998] Chi, E. H.-h. and Riedl, J. T. (1998). An operator interaction framework for visualization systems. In *Proceedings IEEE Symposium on Information Visualization*, pages 63–70. IEEE.

[Chomsky, 1957] Chomsky, N. (1957). *Syntactic structures.* Mouton de Gruyter.

[Chomsky, 1965] Chomsky, N. (1965). *Aspects of the Theory of Syntax.* MIT press.

[Cleveland and McGill, 1985] Cleveland, W. and McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science*, 229(4716):828–833.

[Collins et al., 2009] Collins, C., Viegas, F. B., and Wattenberg, M. (2009). Parallel tag clouds to explore and analyze faceted text corpora. In *2009 IEEE Symposium on Visual Analytics Science and Technology*, pages 91–98.

[Culy and Lyding, 2010] Culy, C. and Lyding, V. (2010). Double tree: An advanced kwic visualization for expert users. In *Information Visualisation (IV), 2010 14th International Conference*, pages 98–103.

[Culy and Lyding, 2011] Culy, C. and Lyding, V. (2011). Corpus clouds - facilitating text analysis by means of visualizations. In *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 351–360. Springer Berlin Heidelberg.

[Culy et al., 2011] Culy, C., Lyding, V., and Dittmann, H. (2011). Structured parallel coordinates: a visualization for analyzing structured language data. In *Proceedings of the 3rd International Conference on Corpus Linguistics, CILC-11*, pages 485–493.

[Eaton, 1940] Eaton, H. (1940). *Semantic Frequency List for English, French, German, and Spanish: A Correlation of the First Six Thousand Words in Four Single-language Frequency Lists.* American Council on Education. University of Chicago Press.

[Ellis and Dix, 2006] Ellis, G. and Dix, A. (2006). An explorative analysis of user evaluation studies in information visualisation. In *Proceedings of the 2006 AVI Workshop on Beyond Time and Errors: Novel Evaluation Methods for Information Visualization*, pages 1–7.

[Francis et al., 1982] Francis, W., Kučera, H., and Mackie, A. (1982). *Frequency analysis of English usage: Lexicon and grammar.* Houghton Mifflin.

[Francis and Kucera, 1967] Francis, W. N. and Kucera, H. (1967). Computational analysis of present-day american english. *Journal of Experimental Psychology: General*, 143:1065–1081.

[Fries and Traver, 1940] Fries, C. C. and Traver, A. A. (1940). English word lists: A study of their adaptability and instruction. *Committee on Modern Languages of the American Council on Education.*

[Gratzl et al., 2013] Gratzl, S., Lex, A., Gehlenborg, N., Pfister, H., and Streit, M. (2013). LineUp: Visual analysis of multi-attribute rankings. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '13)*, 19(12):2277–2286.

[Green, 1998] Green, M. (1998). Toward a perceptual science of multidimensional data visualization: Bertin and beyond. *ERGO/GERO Human Factors Science*, 8:1–30.

[Gries, 2010] Gries, S. T. (2010). Corpus linguistics and theoretical linguistics: A love–hate relationship? not necessarily. *International Journal of Corpus Linguistics*, 15:327–343.

[Hareide and Hofland, 2012] Hareide, L. and Hofland, K. (2012). *Compiling a Norwegian-Spanish parallel corpus.* Amsterdam, John Benjamins.

[Heer et al., 2005] Heer, J., Card, S. K., and Landay, J. A. (2005). Prefuse: a toolkit for interactive information visualization. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*, pages 421–430.

[Jakubíček et al., 2013] Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlỳ, P., and Suchomel, V. (2013). The tenten corpus family. In *7th International Corpus Linguistics Conference CL*, pages 125–127.

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

[Johansson, 1991] Johansson, S. (1991). Times change, and so do corpora. *English Corpus Linguistics*, pages 305–314.

[Jones, 2019] Jones, H. (2019). Searching for statesmanship: a corpus-based analysis of a translated political discourse. *Polis: The Journal for Ancient Greek and Roman Political Thought*, 36(2):216 – 241.

[Karamanis et al., 2011] Karamanis, N., Luz, S., and Doherty, G. (2011). Translation practice in the workplace: contextual analysis and implications for machine translation. *Machine Translation*, 25(1):35–52.

[Kilgarriff et al., 2008] Kilgarriff, A., Rychly, P., Smrz, P., and Tugwell, D. (2008). The sketch engine. *Practical Lexicography: a Reader*, pages 297–306.

[Laviosa, 1997] Laviosa, S. (1997). How comparable can'comparable corpora'be? *Target. International Journal of Translation Studies*, 9(2):289–319.

[Laviosa, 1998] Laviosa, S. (1998). Core patterns of lexical use in a comparable corpus of english narrative prose. *Meta: Journal des Traducteurs/Meta: Translators' Journal*, 43(4):557–570.

[Laviosa, 2010] Laviosa, S. (2010). Corpus-based translation studies 15 years on: Theory, findings, applications. *YNAPS - A Journal of Professional Communication*, 24(2010):3–12.

[Luhn, 1960] Luhn, H. P. (1960). Key word-in-context index for technical literature (kwic index). *American Documentation*, 11(4):288–295.

[Luz, 2000] Luz, S. (2000). A software toolkit for sharing and accessing corpora over the Internet. In *Proceedings of the Second International Conference on Language Resources and Evaluation: LREC-2000*, pages 1749–1754.

[Luz, 2011] Luz, S. (2011). Web-based corpus software. In *Corpus-based Translation Studies – Research and Applications*, chapter 5, pages 124–149. Continuum.

[Luz and Sheehan, 2014] Luz, S. and Sheehan, S. (2014). A graph based abstraction of textual concordances and two renderings for their interactive visualisation. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*, AVI '14, pages 293–296. ACM.

[Lyding et al., 2014] Lyding, V., Nicolas, L., and Stemle, E. (2014). interhist - an interactive visual interface for corpus exploration. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

[Mackinlay, 1986] Mackinlay, J. (1986). Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, 5(2):110–141.

[Mahlberg, 2005] Mahlberg, M. (2005). *English General Nouns: A corpus theoretical approach*. Studies in Corpus Linguistics. John Benjamins Publishing Company.

[Manning and Schütze, 1999] Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

[Marai, 2018] Marai, G. E. (2018). Activity-centered domain characterization for problem-driven scientific visualization. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):913–922.

[McEnery and Wilson, 1996] McEnery, T. and Wilson, A. (1996). *Corpus linguistics*. Edinburgh textbooks in empirical linguistics. Edinburgh University Press.

[McEnery and Wilson, 2001] McEnery, T. and Wilson, A. (2001). *Corpus Linguistics: An Introduction.* Edinburgh University Press Series. Edinburgh University Press.

[McEnery et al., 2006] McEnery, T., Xiao, R., and Tono, Y. (2006). *Corpus-based language studies: An advanced resource book.* Taylor & Francis.

[Munzner, 2009] Munzner, T. (2009). A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928.

[Olohan, 2002] Olohan, M. (2002). Corpus linguistics and translation studies: Interaction and reaction. *Linguistica Antverpiensia, New Series–Themes in Translation Studies*, 1.

[Olohan, 2004] Olohan, M. (2004). *Introducing Corpora in Translation Studies.* Taylor & Francis.

[Park and Nam, 2017] Park, H. and Nam, D. (2017). Corpus linguistics research trends from 1997 to 2016: A co-citation analysis. *Linguistic Research*, 34:427–457.

[Preyer, 1898] Preyer, W. (1898). *The mind of the child.* History of psychology. Routledge/Thoemmes Press.

[Rabadán et al., 2009] Rabadán, R., Labrador, B., and Ramón, N. (2009). Corpus-based contrastive analysis and translation universals: A tool for translation quality assessment english and spanish. *Babel*, 55(4):303–328.

[Rivadeneira et al., 2007] Rivadeneira, A. W., Gruen, D. M., Muller, M. J., and Millen, D. R. (2007). Getting our head in the clouds: toward evaluation studies of tagclouds. In *CHI '07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 995–998, New York, NY, USA. ACM.

[Schmied, 1993] Schmied, J. (1993). Qualitative and quantitative research approaches to english relative constructions. *International Computer Archive of Modern English. Conference*, pages 85–96.

[Scott, 2008] Scott, M. (2008). Wordsmith tools version 5. *Liverpool: Lexical Analysis Software*, 122.

[Scott, 2010] Scott, M. (2010). What can corpus software do. *The Routledge Handbook of Corpus Linguistics*, pages 136–151.

[Scott et al., 2001] Scott, M. et al. (2001). Comparing corpora and identifying key words, collocations, and frequency distributions through the wordsmith tools suite of computer programs. *Small Corpus Studies and ELT*, pages 47–67.

[Sedlmair et al., 2012] Sedlmair, M., Meyer, M., and Munzner, T. (2012). Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2431–2440.

[Sheehan et al., 2018] Sheehan, S., Masoodian, M., and Luz, S. (2018). Comfre: A visualization for comparing word frequencies in linguistic tasks. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces*, AVI '18, pages 42:1–42:5. ACM.

[Shneiderman, 1996] Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343. IEEE.

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

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

[Spence and Apperley, 2013] Spence, R. and Apperley, M. (2013). Bifocal display. *The Encyclopedia of Human-Computer Interaction, 2nd Ed.*

[Stasko et al., 2007] Stasko, J., Gorg, C., Liu, Z., and Singhal, K. (2007). Jigsaw: Supporting investigative analysis through interactive visualization. In *2007 IEEE Symposium on Visual Analytics Science and Technology*, pages 131–138.

[Sun et al., 2016] Sun, M., Mi, P., North, C., and Ramakrishnan, N. (2016). Biset: Semantic edge bundling with biclusters for sensemaking. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):310–319.

[Taine, 1877] Taine, T. (1877). The acquisition of language by children. *Mind*, 2:252.

[Thompson and Hunston, 2006] Thompson, G. and Hunston, S. (2006). *System and Corpus: Exploring Connections*. Functional linguistics. Equinox.

[Tognini-Bonelli, 2001] Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Linguistics Today. J. Benjamins.

[Tufte, 2001] Tufte, E. R. (2001). *The Visual Display of Quantitative Information*. Graphics Press, 2nd edition.

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

[Viegas et al., 2009] Viegas, F. B., Wattenberg, M., and Feinberg, J. (2009). Participatory visualization with wordle. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1137–1144.

[Viégas et al., 2007] Viégas, F. B., Wattenberg, M., Van Ham, F., Kriss, J., and McKeon, M. (2007). Manyeyes: a site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1121–1128.

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

[Xiao, 2010] Xiao, R. (2010). *Using corpora in contrastive and translation studies.* Cambridge Scholars Publishing.

[Zanettin, 2001] Zanettin, F. (2001). Swimming in words: Corpora, translation, and language learning. *Learning with Corpora*, page 177.

[Zanettin, 2013] Zanettin, F. (2013). Corpus methods for descriptive translation studies. *Procedia - Social and Behavioral Sciences*, 95:20 – 32.

[Zubillaga et al., 2015] Zubillaga, N., Sanz, Z., and Uribarri, I. (2015). Building a trilingual parallel corpus to analyse literary translations from german into basque. *New directions in corpus-based translation studies*, 1:71.