Matrix: A statistical method and software tool for linguistic analysis through corpus comparison

A thesis submitted to Lancaster University
for the degree of
Ph.D. in Computer Science
September 2002

Paul Edward Rayson B.Sc.
Computing Department
Lancaster University

Abstract

Matrix: A statistical method and software tool for linguistic analysis through corpus comparison

A thesis submitted to Lancaster University for the degree of Ph.D. in Computer Science

Paul Edward Rayson, B.Sc.

September 2002

This thesis reports the development of a new kind of method and tool (Matrix) for advancing the statistical analysis of electronic corpora of linguistic data. First, we describe the standard corpus linguistic methodology, which is hypothesis-driven. The standard research process model is 'question – build – annotate – retrieve – interpret', in other words, identifying the research question (and the linguistic features) early in the study. In recent years corpora have been increasingly annotated with linguistic information. From our survey, we find that no tools are available which are data-driven on annotated corpora, in other words, a tool which assists in finding candidate research questions. However, Matrix is such a tool. It allows the macroscopic analysis (the study of the characteristics of whole texts or varieties of language) to inform the microscopic level (focussing on the use of a particular linguistic feature) as to which linguistic features should be investigated further. By integrating part-of-speech tagging and lexical semantic tagging in a profiling tool, the Matrix technique extends the keywords procedure to produce key grammatical categories and key concepts. It has been shown to be applicable in the comparison of UK 2001 general election manifestos of the Labour and Liberal Democratic parties, vocabulary studies in sociolinguistics, studies of language learners, information extraction and content analysis. Currently, it has been tested on restricted levels of annotation and only on English language data.

Declaration

I declare that the work presented in this thesis is, to the best of my knowledge and belief, original and my own work. The material has not been submitted, either in whole or part, for a degree at this or any other university.

Paul Rayson.

Acknowledgements

First of all I'd like to thank Pete Sawyer, who, as we were sitting in a bar in Barcelona, convinced me that writing this up was possible, and he paid for the cerveza too. None of this work would have been possible without Roger Garside who was not only my supervisor but also ignited my interest in natural language processing when I began my third year project as an undergraduate in 1989.

During my work on this thesis, and before I started my PhD research, I have been a member of the UCREL research group at Lancaster University, and I would like to thank the members of the group, specifically, Geoffrey Leech, Jenny Thomas and Andrew Wilson, all of whom I worked with from 1990, along with Roger Garside. The early seed of this work was sown then. Nick Smith, David Lee, Simon Botley and Tony McEnery have given me support along the way. From the Centre for Applied Statistics at Lancaster University, Damon Berridge and Brian Francis have been of invaluable assistance when I posed many statistical questions to them over the past few years. My thanks also to Sylviane Granger for many interesting discussions during our work together in Lancaster and at the Université Catholique de Louvain.

I should also thank my parents who bought me a BBC Micro Computer as a combined birthday and Christmas present. I wrote my first piece of software on it. The 32K RAM seemed to be sufficient in 1982.

As I sat at the keyboard at home and in the office typing this thesis, the music of Jean Michel Jarre has assisted my concentration during the final writing stages of the work.

Finally, leaving the best 'til last, thank you to my wife, Alison, who thought I would never finish writing my thesis and enlisted members of our family to pester me about progress in the final writing up stage. Jamie and Katie might get fewer trips to Granny's house now that this is finished.

Contents

List of tables

List of figures

List of abbreviations and acronyms

ACASD	Automatic Content Analysis of Spoken Discourse
ATC	Air Traffic Control
BNC	British National Corpus
CES	Corpus Encoding Standard
CHILDES	Child Language Data Exchange System
CLAWS	Constituent Likelihood Automatic Word-tagging System
d.f.	degrees of freedom
DTD	Document Type Definition
EAGLES	Expert Advisory Group on Language Engineering Standards
FTF	Fuzzy Tree Fragment
HMM	Hidden Markov Model
HTML	Hypertext Mark-up Language
ICE	International Corpus of English
ICECUP	ICE Corpus Utility Program
ICLE	International Corpus of Learner English
IE	Information extraction
KWIC	Key Word In Context
LibDem	Liberal Democratic Political Party (UK)
LL	Log likelihood, likelihood ratio or G^2
LOB	Lancaster/Oslo-Bergen Corpus
MICASE	Michigan Corpus of Academic Spoken English
MWU	Multi-word units
NLP	Natural Language Processing
NNS	Non-native Speaker
NR	Normalised ratio
NS	Native Speaker
OCP	Oxford Concordance Program
POS	Part of speech

REVERE	Reverse Engineering of Requirements project
SARA	SGML-Aware Retrieval Application
SEC	Lancaster-IBM Spoken English Corpus
SGML	Standard Generalised Mark-up Language
TACT	Text Analysis Computing Tools
TEI	Text Encoding Initiative
TOSCA	Tools for Syntactic Corpus Analysis
UCREL	University Centre for Computer Corpus Research on Language
USAS	UCREL Semantic Analysis System
WSJ	Wall Street Journal
WWW	World Wide Web
XML	Extensible Mark-up Language
X^2	Pearson's chi-squared or χ^2 test

1. Introduction

"In the beginning was the word. But by the time the second word was added to it, there was trouble."
Simon, J. (1981: 111) *Paradigms lost: reflections on literacy and its decline.* Chatto & Windus. London.

1.1 This thesis

Traditionally, quantitative research in corpus linguistics has been *hypothesis-driven*. In other words, a specific linguistic research question, which is identified at an early stage in a research project, leads to the collection or selection of a corpus and some phenomenon is investigated using that corpus. This research process is usually focussed on investigating a small number of linguistic phenomena that have been selected prior to the investigation. The problem with this kind of approach is that during the investigation, we can search only for evidence, or lack of evidence, for what we expect to find. The alternative to hypothesis-driven research is *data-driven* research[1], in which we are informed by the corpus data itself and allow it to lead us in all sorts of directions, some of which we may never have though of. It is a process where the phenomena are identified in the course of the research project rather than at the outset. It allows us to have a wider focus on a whole corpus or text rather than on specific features contained within it. This thesis describes a technique that allows the corpus data to direct the research in sometimes new and unexpected ways, and also introduces a piece of software which implements this method. The technique allows us to explore the corpus data more completely in a shorter amount of time, and directs us to where further in-depth study should perhaps take place. It enables us to find unexpected phenomena that we would not otherwise have considered for study.

[1] We have chosen to call this *data-driven* to distinguish the methodology from corpus-driven linguistics, see section 2.3.

This thesis describes a method of statistical profiling within corpus linguistics. It focuses on a method (called Matrix, together with a piece of software implementing this method[2]) that was developed by the author and has already been used in both academic and commercial contexts. The Matrix software is principally a tool for use with annotated corpora and has the novel ability to perform statistical comparisons of corpora at multiple levels of annotation, including the lexical level. The frequency profile is the first port of call when investigating corpora and leads on to other research activities such as concordancing, and collocation analysis. These tasks can be applied to aid investigation and understanding of bodies of text in areas such as language teaching, linguistic research, content analysis, software engineering, machine translation and lexicography.

The next three sections of this first chapter introduce the notion of statistical profiling and annotation within the field of corpus linguistics. We then outline the objectives of the study. The final section describes the structure and content of the remaining chapters of this thesis.

1.2 Corpus linguistics

A *corpus* is defined in the Concise Oxford English Dictionary as a 'body, collection of writings'. Aston and Burnard (1998: 4) note that the second edition of the Oxford English Dictionary lists five distinct senses for the word. Only two of these particularly refer to language. However, preliminary standards guidelines have distinguished between the terms *corpus* and *collection* or *archive*, of which only *corpus* is related to some linguistic purpose (Sinclair, 1996). The most commonly agreed upon plural of *corpus* is *corpora*.[3] There is no accepted minimum or maximum size for a corpus, or specification of what it should contain. A corpus could contain

[2] The choice of the name *Matrix* relates to the appearance of the output of the tool: a matrix in mathematics is a rectangular array of elements set out in rows and columns. No link is intended to the film starring Keanu Reeves.

[3] The frequency and acceptability of other plural forms of the word *corpus* (e.g. *corpuses*) have been much debated on the CORPORA electronic mailing list. Aston and Burnard (1998: 63-73) devote ten pages to the question.

the entire works of Shakespeare, sets of instructions from washing powder boxes, or the text of the match-day programmes from Nottingham Forest Football Club in the season they won the League Championship. Corpora need not contain only written language; s*poken corpora* can be built by transcribing the recorded speech from, for example, news broadcasts or conversations of people giving directions in the street.

Corpora are usually collected with a particular linguistic research project in mind, such as providing frequency information for dictionary entries or advanced language learning of German (Jones 1997) or classroom teaching of French (Xunfeng and Kawecki, 2001). Sometimes corpora are collected without a specific purpose and are made available as a general language resource to linguists, social scientists, language teachers, market researchers and others. In recent years with the advent of CD-ROMs and the World Wide Web (WWW), a corpus can be a *multimedia* (or *multimodal*) *corpus*, which includes still pictures, video and sounds.

The term *corpus linguistics* has been described (McEnery and Wilson, 1996) in simple terms as the study of language based on examples of 'real life' language use. It has a relatively long history. Corpus linguistics is not a branch of linguistics such as syntax, semantics and pragmatics that concentrate on describing or explaining some aspect of language use. It is a methodology that can be applied to a wide range of linguistic study.

1.3 Statistical profiling

Frequency-sorted word lists have long been part of the standard methodology for exploiting corpora. Sinclair (1991: 30) writes, "Anyone studying a text is likely to need to know how often each different word form occurs in it". Tribble and Jones (1997: 36) outline a pedagogical methodology for using texts in the language classroom, proposing that the most effective starting point for understanding a text is a frequency-sorted word list. The frequency list records the number of times each word occurs in the text; it can provide interesting information about the words that appear (or do not appear) in a text. The list can be arranged in order of first occurrence, alphabetically or in frequency order. First-occurrence order serves as a

quick guide to the distribution of words in a text, an alphabetic listing is built mainly for reference, but a frequency-ordered listing highlights the most commonly occurring words in the text. For example, Juilland produced a series of frequency dictionaries for Spanish, Rumanian and French (Juilland et al 1964, 1965 and 1970). Even the more traditional dictionaries can make use of frequency information. The texts on which the American Heritage Word Frequency Book (Carroll et al, 1971) was built formed the citation base for the American Heritage School Dictionary.

Francis and Kučera (1982) take the simple word frequency list one stage further by reporting *grammatical word* frequencies. This gives frequencies of words with their associated part-of-speech (POS) tags in the (tagged version of the) Brown corpus (Francis and Kučera, 1964). The frequency profile for a given text can be compared to that of other similar texts or to that of large bodies of text. Since the high frequency items tend to have a stable distribution generally, significant changes to the ordering of the words in the frequency list can flag items of interest to the researcher (Sinclair 1991: 31). Such techniques can be carried out manually for a small corpus but otherwise we need the aid of a computer program. Although the computer saves us time with its processing of the texts into frequency lists, it presents us with so much information that we need a filtering mechanism to pick out significant items before the analysis can proceed. Hofland and Johansson (1982) use Yule's K statistic and the chi-squared goodness-of-fit test in their comparison to pick out statistically significant different word frequencies across British and American English. Various formulae can be applied to adjust the raw frequencies for the distribution of words within a text, or to describe the dispersion of frequencies in subsections of a corpus.

Frequency profiling is one of the two main methods in corpus linguistics, the other being the use of concordance lines. A set of concordance lines presents instances of a word or phrase usually in the centre, with words that come before and after it to the left and right. Hunston (2002: 38) devotes a whole chapter of her book to the interpretation of concordance lines. We will review software capable of this function in section 3.4.

1.4 Corpus annotation

Recently, with so much work being done on the analysis of corpora, it is seen as essential to annotate a corpus with the results of the research. Obviously, this can act as a bootstrap for an increasingly detailed and accurate analysis at the same linguistic level or for the next level of research. We can build a hierarchy of analyses from POS tagging, parsing, semantic tagging to discourse analysis. We will examine this hierarchy further in section 2.4. This enables us to revisit the results obtained from word frequency analysis and obtain frequency profiles for POS tags, semantic categories and so on. By applying the same significance testing methods, we can extend and refine our analysis based on more precise linguistic categories. This is the novel approach described in this thesis. We can perform a statistical comparison of annotated corpora and obtain results at each level of annotation contained within the corpus.

1.5 Objectives of the study

Leech and Fallon (1992) describe a two stage process in their examination of cultural differences using corpora of British and American English. Stage one is to use a comparative alphabetical list of word frequencies in the two corpora to select groups of words for further study. This stage examined the Hofland and Johansson (1982) lists of word frequencies in British and American English to select the items marked with significant differences. Stage two made use of a concordance tool to examine the contexts of the selected words from the Brown and LOB corpora. Leech and Fallon cite two main reasons for consulting the concordance lines:

1. To check whether the frequency of the graphic form actually reflected the sense of the word they were interested in.
2. To check that the high frequency of an item was not due to any obvious skewing of its distribution in the corpus.

They describe stage two as requiring “an enormous amount of human labour, and in practice the task had to simplified”. The same issues are faced by other corpus

researchers in their studies. The most used current techniques to reduce the number of concordance lines for inspection are that of random sampling, and collocation statistics (arising out of the needs of lexicographers, see Kilgarriff and Tugwell 2002).

The main research question investigated in this thesis is whether we can provide some level of automation for these two stages in terms of a method and tool support in order to assist corpus researchers. This objective breaks down into sub-objectives as follows:

1. to provide support for suggesting linguistic features to be further investigated:
    a. to investigate a data-driven method for corpus comparison which uses macroscopic analysis (the study of the characteristics of whole texts or varieties of language) to inform the microscopic level (focussing on the use of a particular linguistic feature)
    b. to investigate the use of comparison of corpora annotated with word-class and semantic tags in order to home in on particular word forms
    c. to provide range and dispersion information alongside the frequency of an item to highlight any skewing of the frequency distribution
2. to provide a method that can be used for comparison of differently sized corpora
3. to evaluate the suggested method statistically and by the results of the application of the software tool that implements the method

1.6 Thesis structure

The remaining chapters of this thesis are as follows. Chapter two gives a more detailed introduction to the methodology of corpus linguistics in which we describe background literature related to the area of research in this study: the statistical profiling of annotated corpora. The chapter includes background information on current statistical tests used in corpus linguistics, describing previous word frequency analyses, dispersion statistics and goodness-of-fit tests. We look at work already performed on word frequency distributions and the statistical theory used to describe them. We then examine work related to one of the objectives of the study: comparison

of frequency distributions in corpora. This leads on to a discussion of how representative corpora can be used as a form of control dataset with which to compare a new corpus.

Chapter three discusses the most widely available tools used for standard tasks of text analysis. It will be made clear that these tools have limitations that are remedied by the development of Matrix method and tool.

Chapter four describes the Matrix method and the software tool implementing this method. Chapter four includes a worked example showing the application of the tool to study the language used in the United Kingdom General Election manifestos of the Labour and Liberal Democratic (LibDem) parties from the June 7th 2001 election.

Chapter five is an evaluation of the Matrix method and software from two perspectives. Firstly we evaluate the statistical validity of the choice of statistic in the method, and then we go on to look at three case studies of the software and method in use.

Chapter six provides a summary of the thesis and its conclusion. We discuss limitations of the work and suggest future work on the method and software tool.

2. Corpus Linguistics

Dr. Johnson: "The dictionary contains every word in our beloved language."
Blackadder: "Every single word?"
Dr. Johnson: "Every single word, sir."
Blackadder: "May I offer the Doctor my most enthusiastic contrafibularities."

(Blackadder the Third, Ink and Incapability, British Broadcasting Corporation, 1987).

2.1 Introduction

In this chapter, we begin with some definitions of terms in order to place the work presented in this thesis in context. We will describe the typical process model of corpus linguistic research and the characteristics of the paradigm. Next, corpus annotation and encoding methods and standards employed will be surveyed. Finally, we will move on to review the previous research in the area of statistical profiling in corpus linguistics, paying particular attention to details of statistical tests for the comparison of frequency data.

2.2 Corpus Linguistics in Context

As we have described in section 1.2, a corpus can have a wide range of content and applications. Sinclair (1995) prefers to define a corpus in a less flexible way in order to make it more useful to the study of language. He defines certain linguistic criteria and characteristics that a corpus is assumed to have: quantity, quality, simplicity, and documentation. He lists various types of corpora (my examples):

1. *Reference corpus*: designed to provide comprehensive information about a language, e.g. the British National Corpus (BNC) (Aston and Burnard, 1998).

2. *Monitor corpus*: of constant size, but constantly refreshed with new material, while old material is removed to archival storage (although this is no longer strictly necessary due to the increase in computing power) e.g. the Bank of English / Birmingham Corpus (Renouf, 1987: 21).
3. *Parallel corpus*: collection of texts, each of which is translated into one or more other languages, e.g. the CRATER corpus (McEnery et al, 1997).
4. *Comparable corpus*: similar texts in more than one language, although Sinclair notes that there is no agreement on the nature of the similarity, e.g. the International Corpus of English (ICE) (Greenbaum, 1996).

Sinclair prefers to avoid using the term *multilingual corpus* in favour of the names parallel and comparable, although the term is suitable for a parallel corpus before it has been aligned. Hunston (2002: 14) adds the following types of corpora (Hunston's examples):

1. *Specialised corpus*: collection of texts of a particular type designed to be representative only of a given type of text, e.g. the Michigan Corpus of Academic Spoken English (MICASE)
2. *Learner corpus*: collection of texts produced by learners of a language, e.g. the International Corpus of Learner English (ICLE)
3. *Historical (diachronic) corpus*: texts from different periods of time, e.g. the Helsinki corpus

Biber, Conrad and Reppen (1998: 4) list the essential characteristics of *corpus-based linguistics*:

- it is empirical, analysing the actual patterns of use in natural texts;
- it utilises a large and principled collection of natural texts, known as a "corpus", as the basis for analysis;
- it makes extensive use of computers for analysis, using both automatic and interactive techniques;
- it depends on both quantitative and qualitative analytical techniques.

A related and sometimes overlapping term is *computational linguistics*, which is similarly a methodology within language study. Grishman (1986) describes it as "the study of computer systems for understanding and generating natural language". Computational linguistics concentrates on "using computers as a tool to model (and, sometimes, verify or falsify) fragments of linguistic theories deemed of particular interest" (Boguraev et al, 1995). Computational linguistics has focussed on three application areas: machine translation, information retrieval and human-computer interfaces. Building systems to perform these tasks often involves integrating software modules that analyse language at many different levels: morphology, syntax, semantics, and prosody, for example. Hence, a new term, *language engineering*[4], has appeared recently to describe this process.

The larger field of *natural language processing* (NLP), which involves the development of computer systems to imitate intelligent human linguistic behaviour, can be split into two distinct areas (see Leech 1986: 208). The first area is to provide the computer program with enough linguistic and real-world knowledge so that it can communicate within a limited domain of discourse such as the tabletop world of Winograd (1972). The second area is to process a wider range of discourse but at a restricted level of analysis (e.g. syntax or semantics). Corpus linguistics generally overlaps with natural language processing in this second category.

Two distinct phases in modern corpus linguistics have been identified by Leech (1991). The first phase, structural linguistics, ended in the late 1950s when Chomsky presented his views on the inadequacy of corpus data. The second phase, initially unfashionable, began soon after in two main locations: Randolph Quirk planned the Survey of English Usage at University College London in the UK and Nelson Francis and Henry Kučera began corpus work at Brown University in the USA. Today, we assume that corpora are machine-readable and that corpus linguists generally use

[4] The term *language engineering* seems to have another meaning which predates this sense. A branch of sociolinguistics known as 'language planning' refers to the practical and theoretical problems involved in the imposition of a standard language on a community of speakers (Aitchison, 1991: 219). Foley (1997: 398) refers to this process as 'linguistic engineering', also Sin and Roebuck (1996) use the term 'language engineering' itself in the title of their paper. For further discussion of the (software engineering) sense of language engineering used in this thesis, see Cunningham (1999).

computers in their studies. Leech (1992) describes the field as *computer corpus linguistics* "a new philosophical approach to the subject". In the same book, Francis surveys corpora before the arrival of computers (Francis, 1992). Stubbs (1993: 9, 1996: 31) attempts to contrast the approach taken by Sinclair with others in British corpus linguistics along the lines of invention of data.

Wikberg (1997) contrasts corpus studies with *discourse analysis* (*text linguistics*), where discourse analysis is to be seen as traditional text study, not based on machine readable corpora. The relevant properties of each field are identified by Wikberg as:

	Corpus studies	*Discourse analysis*
1.	language as product (static)	language as process
2.	microanalysis	micro-macroanalysis
3.	form $\rightarrow$ meaning	form $\leftrightarrow$ meaning
4.	qualitative research based on quantity	focus on quality
5.	distribution in genres and corpora	distribution in single texts

Corpora are used to derive empirical knowledge about language, which can supplement, and frequently supplant, information from reference sources and introspection (Leech, 1991; 1992). Because they are well suited to quantitative analysis, corpora can provide information about the relative frequencies of many aspects of language. These frequencies can then be employed in probabilistic analysis techniques.

Probabilistic systems, instead of using hard-and-fast rules, use frequency data along with sophisticated statistical models to make a 'best guess' about the correct analysis of a piece of language (Sampson, 1987a). Although probabilistic systems make mistakes, they often perform at a very high degree of accuracy (in the high nineties percentage accuracy for part-of-speech analysis). Compared with rule-based systems, they are exceptionally robust, and can analyse 'real' language containing performance errors (as opposed to idealised invented examples) where rule-based systems would often fail. Because of this robustness and overall accuracy, mainstream computational

linguists are now taking an increased interest in probabilistic methods and corpora (Tsujii, 2000).

According to Sinclair (1991: 36) "the most exciting aspect of long text data processing, however, is not the mirroring of intuitive categories of description. It is the possibilities of new approaches, new kinds of evidence and new kinds of description". In this thesis we present Matrix as one approach which can lead research in new directions.

Of particular relevance to this thesis is the process of *corpus annotation*. Leech (1997) defines it as the practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written language data. A typical example of this is grammatical tagging (another name for part-of-speech tagging) which associates a string (called a tag) with each lexical item in a text to indicate its grammatical class in context. Corpus annotation is a process that can be done completely by hand or in an automatic manner by computer program, or sometimes by a combination of these two methods. Usually, the computer program takes its training data from a corpus that has first been analysed by hand. Section 2.4 examines the various types of corpus annotation in more detail.

The work described in this thesis is centred on the area of corpus linguistics, providing a method and software tool to aid the investigation and study of real language. The next section looks in more depth at the standard methodologies applied in the field of corpus linguistics.

2.3 Corpus linguistic methodology

In this section, we will describe the typical linguistic research methodology using the corpus paradigm. In section 1.2 when we defined the term corpus linguistics, we described it as a methodology that can be applied to a wide range of linguistic study. This is demonstrated by Biber, Conrad and Reppen (1998) who describe corpus-based approaches in different areas of linguistics, including lexicography, grammar, discourse, register variation, language acquisition and historical linguistics. In all of

these various areas of linguistic study, there are five main steps when we examine the corpus-based approach:

1. **Question**: A research question or model is devised
2. **Build:** Corpus design and compilation
3. **Annotate**: Computational analysis of the corpus
4. **Retrieve**: Quantitative and qualitative analyses of the corpus
5. **Interpret**: Manual interpretation of the results or confirmation of the accuracy of the model

A simplified process model of the corpus-based methodology typically used by researchers is as follows: it would begin with the identification of a research question, continue with building and annotating a corpus with which to investigate the topic, and finish with the retrieval, extraction and interpretation of information from the corpus which may help the researcher to answer the research question or confirm the parameters of the model. In some cases, the process may be an iterative one, where, following the interpretation of the results some refinement is needed on the research question or annotation of the corpus.

There are two main kinds of research question (step 1 above) that can be investigated. Firstly, we can focus on the use of a particular linguistic feature, possibly a word or grammatical construction. We will call this type I. Secondly, we can examine the characteristics of whole texts or varieties of language, and we will call this type II. These two types are sometimes referred to as microscopic (type I) and macroscopic (type II), for example see Biber (1988: 61). Traditionally, studies tend to focus on type I and examine linguistic (lexical or grammatical associations of the feature), and non-linguistic aspects (distribution of the feature across different types of text or speech). Type II inverts this relationship in investigating, for example, register variation across text, by examining how certain features or groups of features characterise a text.

Increasingly, researchers no longer have to build and annotate (steps 2 and 3 above) their own corpus material, although this is usually the case with problem-oriented tagging (see Table 2.2). Instead they can use precompiled and annotated corpora that

are available 'off-the-shelf' (Meyer 1991). We will give examples of corpora and corpus annotation in section 2.4. Software tools to carry out compilation, annotation and retrieval from corpora (steps 2, 3 and 4 in our process model) will be reviewed in chapter 3. There are two major methods used in corpus linguistics to retrieve and interpret (steps 4 and 5) data from corpora. These are frequency profiling and concordancing. We will examine statistical profiling in section 2.6 and concordancing tools in section 3.4.

Our process model, as described above, is in line with Leech's (1992) view of the corpus linguistic paradigm. Leech argues that the corpus-based methodology conforms to standards commonly ascribed to 'the scientific method': falsifiability, completeness, simplicity, strength, and objectivity. In this thesis, it is hoped that these characteristics will be amply demonstrated.

There are many examples of both types of research question in the many conference publications, journals and edited collections that have appeared. Common to both is the prior selection of which linguistic features to study. The method proposed in this thesis allows a different approach: decisions on which linguistic features are important or should be studied are made on the basis of information extracted from the data itself; in other words, it is *data-driven*. We will call this type III. It combines the approaches of types I and II by first focussing on whole texts and then suggesting specific linguistic features to study in further detail. In other words, the ordering of the five main steps above will change to the following (with iteration back from step 4 to step 3, which enables refinement of the research question following a retrieval step):

1. **Build:** Corpus design and compilation
2. **Annotate**: Computational analysis of the corpus
3. **Retrieve**: Quantitative and qualitative analyses of the corpus
4. **Question**: A research question or model is devised (iteration back to step 3)
5. **Interpret**: Manual interpretation of the results or confirmation of the accuracy of the model

Our type III process model shown here is similar to that of *corpus-driven linguistics* as presented by Tognini-Bonelli (2001: 85), in which the corpus is the main informant (Francis, 1993). However, we decided to use the term data-driven to distinguish our approach from that of Sinclair, presented by Tognini-Bonelli (ibid: xi). The corpus-driven approach questions the "underlying assumptions behind many well established theoretical positions" (ibid: 48) stating that they need to be re-established or replaced based on evidence from corpora. For example, it proposes a "new unit of meaning" (ibid: 85) and states that there is "no such thing as a synonym". In this study, we will rely on currently existing POS tagsets for example. In the corpus-driven approach, Stubbs (1993: 17) notes that even the traditional POS system "is under attack".

In Table 2.1 we have summarised the number of papers in each of these three categories of research in two recently published edited volumes forming the proceedings of the ICAME conferences, ICAME-1997 in Renouf (1998) and ICAME-1998 in Kirk (2000). For each paper, we have identified the main research theme. Some papers contain other sub-themes, but this is not considered in the table. We also show papers describing corpus building or collection, and lastly, in the miscellaneous section, we count papers whose main theme is the description of tools or methods for use within the corpus linguistic paradigm. As can be seen from the figures in the table, type I research is the most frequent. Type II is less frequent, but data-driven research, type III, occurs only twice. The two papers in question are Ringbom (1998) and Hoffmann and Lehmann (2000).

Ringbom (1998) investigated advanced-learner language in the International Corpus of Learner English (ICLE) by comparing the essays produced by learners to those of native speakers. There were certain problems with this approach as identified by Ringbom. First, there was the assumption that the writing of American and British students form a reasonable norm of argumentative essay writing. Then there was the problem of the ICLE subcorpora being relatively small (roughly 100,000 words each). Ringbom thus restricted the study to high frequency items and reasoned that "if there are fewer than 20 actual occurrences of a word or phrase in such small corpora, not much can be generalised about the writer's use of this aspect of language". We have identified this as type III since Ringbom selected two verbs (*get* and *think*) for further study based on their overuse in frequency terms in the non-native speaker corpora

when compared to the native speaker data. We will argue in this thesis that studies need not be restricted to high frequency items and we will make use of more reliable statistical techniques to identify possible candidates. If we observe less than 20 occurrences of a word or phrase when we would expect to see many more than 20, then this is a phenomenon worth investigating further.

Table 2.1 Summary of papers from ICAME 1997 and 1998

Type of study	**Papers in ICAME1997**	**Papers in ICAME1998**
Question type I (focus on one, or a small number of, linguistic features) [5]	12	10
Question type II (focus on whole texts) [6]	5	4
Question type III (data-driven choice of features) [7]	1	1
Build corpora[8]	3	3
Miscellaneous[9]	1	6
TOTAL	**22**	**24**

We have classified Hoffmann and Lehmann (2000) as type III since they used collocational[10] evidence from the British National Corpus to select pairs of related words that were then used in a study to discover native and non-native speakers'

[5] Papers in the type I category from ICAME97 are Johansson and Geisler, Kjellmer, Levin, Ljung, Mair, Markus, Nevala, Nurmi, Renouf and Baayen, Sand, Wichmann, Wynne et al; and from ICAME98 are Blackwell, McEnery et al, Gerner, Kjellmer, Lindquist, de Mönnink, Paradis, Seppänen and Trotta, Stenström, Vihla

[6] Papers in the type II category from ICAME97 are de Hann, Oostdijk, Collier, Pacey, Peters; and from ICAME98 are Kennedy and Yamazaki, Ooi, Taavitsainen, Granger and Wynne

[7] Papers in the type III category from ICAME97 are Ringbom; and from ICAME98 are Hoffmann and Lehmann

[8] Papers in the build category from ICAME97 are Hasund, Keränen, Bergh et al; and from ICAME98 are Minugh, Brekke, Rahman and Sampson

[9] Papers in the miscellaneous category from ICAME97 are Aarts et al; and from ICAME98 are Lehmann et al, Mason, Oostdijk, Tapanainen and Järvinen, Voutilainen, Wallis et al

[10] Collocation is the occurrence of two or more words within a short space of each other in a text. Firth (1957) famously described collocation as 'the company [a word] keeps'.

familiarity with the word pairs. However, they did not pursue the usual type I path of performing a more in-depth linguistic analysis on the collocates that they discovered. Instead, the paper focuses on analysing the results of the familiarity questionnaire. Due to the large size of the corpus, they selected collocation pairs with less than 100 occurrences to avoid problems of excessive computation. They used the log-likelihood statistic to select 150 collocations.

Missing from this survey of techniques fitting into our type III category is the keywords method by Scott implemented in his WordSmith software (1996-99). We will examine this method in more detail in section 2.7.2. Leech and Fallon (1992) also describe a two-stage research process which we would categorise as type III. Their work is described in section 4.3

2.4 Corpus annotation and encoding

We will see in the next few sections how UCREL and other research groups have acted as corpus builders and annotators. They collect and analyse data for corpora that aim to be representative of general language, and provide the results for others to use. These corpora can then be used in any of the three types of corpus study defined in the previous section. Investigators can make use of these off-the-shelf collections in their own studies. Alternatively, researchers can collect their own corpus or select subsets from the general corpora, and use these in their analysis.

As well as for linguistic study, there are many reasons for annotating a corpus. Leech and Smith (1999) examine in detail the possible uses for wordclass tagging, and we summarise the discussion here:

1. *Adding further annotations*: wordclass tagging is a useful first step and simplifies the tasks of syntactic annotation (parsing), semantic annotation, discourse annotation (see below for examples)
2. *Information extraction*: extracting frequency information, lemmatisation, and collocations from corpora
3. *Information retrieval*: document filtering dependent on content

4. *Word processing*: spelling and grammar checkers
5. *Speech processing*: synthesis and recognition
6. *Handwriting recognition*: language modelling
7. *Machine-aided translation*: annotation of multi-lingual corpora
8. *Dictionaries and grammars*: discriminating homographs for lexicographers writing corpus-based dictionaries
9. *Language learning*: students examining real data for grammatical structures used by native speakers
10. *Development of NLP software*: training corpus for a part-of-speech tagger or evaluation corpus for a parser

In this section, we will review annotation research work performed by the corpus builders. First of all, we should distinguish between two types of corpus mark-up:

- *Annotation*: the practice of adding interpretative, linguistic information to a corpus (Leech, 1997: 2)
- *Encoding*: the insertion of symbols in to the electronic version of a corpus to represent annotation, orthographic and structural features of the text (e.g. characters, symbols, paragraphs, headings) (see section 1 of Sperberg-McQueen and Burnard, 2002[11])

The two forms are sometimes referred to as 'encoding' (Hockey 2000: 24, Edwards 1995: 34) or 'annotation', but these are the definitions we shall use in this thesis following the distinction made in the TEI and CES standards (see later in this section). The second issue really deals with formats of the strings we insert into a corpus and meta-data describing the contents of the corpus. The two types of mark-up are intertwined, and in early corpus collection the corpus annotation was recorded in a different format in different corpora. Annotation and encoding schemes developed in an ad hoc manner in each research centre, and this continues to be the case alongside standardisation initiatives as we shall see later in this section. We find that for corpora such as Brown (Francis and Kučera, 1964) and LOB (Johansson et al, 1978 and Johansson et al, 1986), a separate manual was produced describing the corpus file

[11] Available online at http://www.tei-c.org/P4X/AB.html

format and contents. Recently produced corpora such as the British National Corpus (Burnard, 1995) contain large amounts of meta-data within the electronic files themselves. It is only in the last few years that standardisation efforts have taken place and have separated corpus annotation from encoding.

Firstly, let us focus on the interpretative notion of corpus annotation. There are multiple levels of such annotation that can be applied to a corpus. We will see in the next section the POS, syntactic and semantic levels. These fit together in a hierarchy of annotation as shown in Table 2.2 (adapted from Leech, 1997). Whether, the annotation is applied manually, automatically or semi-automatically we can still attach each level to a corpus.

Table 2.2 Levels of corpus annotation (adapted from Leech, 1997: 12)

Linguistic level	**Examples of features annotated**
Orthographic	Interpretation of italics, initial capital letters or the full stop / period. Delimitation of words by spaces.
Phonological	Syllable boundaries.
Phonetic or Phonemic	Phonetic / phonemic segments: consonants and vowels.
Morphological	Prefixes, suffixes and stems.
Lemma	Roughly equivalent to dictionary headwords. For example, the lemma BE (verb) could be assigned as a tag to the following forms: *'m, 're, 's, am, are, be, been, being, is, was, were*. Lemmatisation has been carried out in Fligelstone (1995) and Sampson (1995).
Prosodic	Transcription of stress, intonation, pauses. For example in the London-Lund corpus (Peppé, 1995): `well ^very nice of you to ((come and)) _spare the !t\/ime and# ^come and !t\alk# -^ tell me a'bout the - !pr\oblems# and ^incidentally# . ^I [@: ] ^do ^do t\ell me# ^anything you 'want about the :college in ''!g\eneral`

Grammatical	Otherwise know as POS tagging or morphosyntactic annotation: assigning word-class labels for not only major parts of speech (noun, verb, preposition, etc.) but also values defining sub-classes, such as singular and plural nouns, positive, comparative and superlative adjectives, and so on. For example, POS tagging using the Penn tagset (Marcus et al, 1993): `Origin/NN of/IN state/NN automobile/NN practices/NNS ./. The/DT practice/NN of/IN state-owned/JJ vehicles/NNS for/IN use/NN of/IN employees/NNS on/IN business/NN dates/VVZ back/RP over/IN forty/CD years/NNS ./.`
Syntactic	Partial (or skeleton) parsing from for example the Lancaster/IBM Spoken English Corpus (Knowles, 1993): `[S[N Nemo_NP1 ,_, [N the_AT killer_NN1 whale_NN1 N] ,_, [Fr[N who_PNQS N][V 'd VHD grown VVN [J too RG big JJ [P for IF [N his APP$ pool NN1 [P on II [N Clacton NP1 Pier NNL1 N]P]N]P]J]V]Fr]N] , , [V has VHZ arrived VVN safely RR [P at II [N his APP$ new JJ home NN1 [P in II [N Windsor NP1 [safari_NN1 park_NNL1]N]P]N]P]V] ._. S]` Full parsing from the ICE-GB corpus (Nelson, Wallis and Aarts (2002): `<ICE-GB:W1A-001 #24:1> <#24:1> <sent> PU,CL(main,montr,pres) [SU,NP() [DT,DTP() [DTCE,PRON(dem,plu) {These}]] [NPHD,N(com,plu) {civilian][aristocrats}]] [VB,VP(montr,pres,semi,perf) [OP,AUX(semi,pres) {seem][to}] [AVB,AUX(perf,infin) {have}] [MVB,V(montr,edp) {used}]] [OD,NP() [DT,DTP() [DTCE,ART(def) {the}]] [NPPR,AJP(attru) [AJHD,ADJ(ge) {old}]] <foreign> [NPHD,N(com,plu) {civitates}]] </foreign> [A,PP() [P,PREP(phras) {as}] [PC,NP() [NPHD,N(com,plu) {power][bases}]]] [PUNC,PUNC(per) {.}]`
Semantic	Semantic field annotation (Rayson and Wilson, 1996) has applications for lexical semantic or word-sense tagging:

	There_Z5 ’s_Z5 been_A3+ more_N5++ violence_E3- in_Z5 the_Z5 Basque_Z2 country_M7 in_Z5 northern_M6 Spain_Z2 :_PUNC one_N1 policeman_G2.1/S2m has_Z5 been_Z5 killed_L1- ,_PUNC and_Z5 two_N1 have_Z5 been_Z5 injured_B2- in_Z5 a_Z5 grenade_G3 and_Z5 machine-gun_G3 attack_G3 on_Z5 their_Z8 patrol-car_M3/G2.1 ._PUNC
Discoursal	For example, anaphoric annotation showing cohesive relations in text (Fligelstone, 1992): (0) The state Supreme Court has refused to release {1[2 Rahway State Prison 2] inmate 1}} (1 James Scott 1) on bail .(1 The fighter 1) is serving 30-40 years
Pragmatic	The mark-up of speech act types (Stiles, 1992).
Stylistic	One aspect of style annotated in corpora is the linguistic representation of people’s thoughts and speech, for example, distinguishing narrative, direct and indirect speech (Short et al, 1996).
Application (or problem) oriented	Orthogonal to the above levels. For example: Error tagging showing learner errors (Meunier 1998 and Granger 1999). Dialogue act mark up for the Verbmobil project (Alexandersson et al, 1998). Annotation of swear words and terms of abuse (McEnery, Baker, and Hardie, 2000).

As can be seen from the corpus examples in Table 2.2, various annotation symbol sets and formats are used when inserting annotation symbols in a corpus. The formats use a number of different special characters such as round, curly and angled brackets and underscore to indicate the presence of an annotation symbol. If a corpus is annotated with more than one level, the resulting file becomes increasingly difficult for a human to read. One of the main reasons for annotating a corpus is to allow the interpretative information to be re-used or extracted by other users, either via some software or directly. Hence, we have seen the emergence of standards for corpus annotation and encoding formats. We will now briefly review some of these formats and standards initiatives.

The first large machine-readable corpus of British English was the LOB corpus (Johansson et al, 1978). An example of the encoding of this corpus is shown in Figure 2.1. Certain features from the original texts were omitted, such as diagrams, maps, reference lists. These were encoded in a comment tag e.g. `**[diagram**]`. Although the original physical form of the document is lost, typographical features are preserved, such as bold (`*6`) and italic (`*1`), and a vertical bar indicates paragraph beginnings. A reference system was introduced; each line in the text begins with a text identifier and a line number within the text.

```
A01  1 **[001 TEXT A01**]
A01  2 *<*'*7STOP ELECTING LIFE PEERS**'*>
A01  3 *<*4By TREVOR WILLIAMS*>
A01  4    |^A *0MOVE to stop \0Mr. Gaitskell from nominating any more Labour
A01  5 life Peers is to be made at a meeting of Labour {0M P}s tomorrow.
A01  6    |^\0Mr. Michael Foot has put down a resolution on the subject and
A01  7 he is to be backed by \0Mr. Will Griffiths, {0M P} for Manchester
A01  8 Exchange.
A01  9    |^Though they may gather some Left-wing support, a large majority
A01 10 of Labour {0M P}s are likely to turn down the Foot-Griffiths
A01 11 resolution.
```

Figure 2.1 Section from untagged LOB corpus

There are two versions of the corpus that contain POS annotation; the vertical format, with one word per line, and the horizontal format, where each word is followed by its associated tag, see Figure 2.2 and Figure 2.3. The vertical format consists of a number of columns showing, from left to right, a unique reference number, the POS tag, and the word itself, followed by orthographic formatting codes such as ‘H’ for heading. The horizontal format is similar to the untagged version, but includes POS tags linked to each word by an underscore character.

Johansson (1994) contrasts the LOB encoding and a more recent standard proposed by the Text Encoding Initiative (TEI), whose guiding principles at the time (Sperberg-McQueen and Burnard, 1990) were to:

- Provide a standard format for data interchange in humanities research;

- Suggest principles for the encoding of texts in the same format;
- Propose sets of coding conventions suited for various applications;
- Maintain compatibility with existing standards as far as possible.

```
A012001        -----   ------------------------------
A012002        *'      *'              H
A012010        VB      stop            H
A012020        VBG     electing        H
A012030        NN      life            H
A012040        NNS     peers           H
A012041        **'     **'             H
A012042        .       .               H       @
A013001        -----   ------------------------------
A013010        IN      by              H
A013020        NP      Trevor          H
A013030        NP      Williams        H
A013031        .       .               H       @
A014001        -----   ------------------------------
A014010        AT      a                       P
A014020        NN      move
A014030        TO      to
A014040        VB      stop
A014050        NPT     \0Mr                    \0
A014060        NP      Gaitskell
```

Figure 2.2 Vertical format from the tagged LOB corpus

```
A012    ^ *'_*' stop_VB electing_VBG life_NN peers_NNS **'_**' ._.
A013    ^ by_IN Trevor_NP Williams_NP ._.
A014    ^ a_AT move_NN to_TO stop_VB \0Mr_NPT Gaitskell_NP from_IN
A014    nominating_VBG any_DTI more_AP labour_NN
A015    life_NN peers_NNS is_BEZ to_TO be_BE made_VBN
```

Figure 2.3 Horizontal format from the tagged LOB corpus

The Text Encoding Initiative originated in 1987 from three Sponsoring Organisations, the Association for Computers and the Humanities (ACH), the Association for Computational Linguistics (ACL), and the Association for Literary and Linguistic Computing (ALLC). With support from the Commission of the European Communities and the Andrew W. Mellon Foundation amongst others, the TEI began the task of developing a draft set of Guidelines for Electronic Text Encoding and

Interchange. The first public draft was published in November 1990 (TEI P1). After revision, the first official (non-draft) version of the guidelines became available in May 1994 (TEI P3). At the time, the TEI claimed to be "the only systematised attempt to develop a fully general text encoding model and set of encoding conventions based upon it, suitable for processing and analysis of any type of text, in any language, and intended to serve the increasing range of existing (and potential) applications and use". In March 1999, four hosts formed a consortium to continue the work of the TEI: Brown University Scholarly Technology Group, Oxford University Humanities Computing Unit, University of Bergen Humanities Information Technologies Research Programme, University of Virginia Electronic Text Center and Institute for Advanced Technology in the Humanities.

The complete TEI P3 guidelines are over 1,000 pages long and this is not the place the summarise them. However, we will extract the main points that are relevant here. The rules and recommendations made in the TEI guidelines make use of ISO 8879, an international standard for the description of marked-up electronic texts, which defines the Standard Generalised Mark-up Language (SGML) (see van Herwijnen, 1994). The building blocks of SGML are the tags, attributes and entities. The special characters used to mark SGML tags are left and right angled brackets (<, >), and SGML entities are shown between ampersand and semi-colon (`&`, `;`). The types of tags and entities in a document are governed by a formal grammar, called a document type definition (DTD). This is perhaps easiest to illustrate by converting the example from Figure 2.1 into SGML format as in Figure 2.4. In this example, we see headings marked by the SGML tags `<head>` and `<subhead>`, paragraphs marked by `<p>` and begin and end quotes marked by entities `&bquo;` and `&equo;` respectively.

The TEI proposals also make recommendations for documentation of text. This information is usually included in the corpus header (within the `<head>` tag) and can record such meta-data as bibliographical details of a written text, identity and background of speakers for spoken text, details of sampling, and editorial principles.

```
<text id=LOBA01>
<head>&bquo;STOP ELECTING LIFE PEERS&equo;</head>
<subhead>By TREVOR WILLIAMS</subhead>
   <p>A MOVE to stop Mr. Gaitskell from nominating any more
Labour life Peers is to be made at a meeting of Labour MPs
tomorrow.
   <p>Mr. Michael Foot has put down a resolution on the subject
and he is to be backed by Mr. Will Griffiths, MP for Manchester
Exchange.
   <p>Though they may gather some Left-wing support, a large
majority of Labour MPs are likely to turn down the Foot-
Griffiths resolution.
```

Figure 2.4 SGML version of section from untagged LOB corpus

The TEI proposals for encoding of corpus annotation have changed since Johansson's paper was published and were applied in the mark-up of POS tagging in the British National Corpus (see Figure 2.5 for an example of this encoding). Here, we can see that the POS tag precedes the word it applies to, for example `<w NN1>condition` indicates that the word *condition* is tagged as *NN1* (singular common noun). Punctuation items are marked with the `<c ...>` SGML tag.

```
<text complete=Y org=SEQ  decls='CN004 HN001 QN000 SN000'>
<div1 complete=Y n=1 org=SEQ type=item>
<head type=MAIN>
<s n=001>
 <w NN1>FACTSHEET <w DTQ>WHAT <w VBZ>IS <w NN1>AIDS<c PUN>?
</head>
<p>
<s n=002>
<hi r=bo> <w NN1>AIDS <c PUL>(<w NP0>Acquired <w AJ0>Immune
 <w NN1-NP0>Deficiency <w NP0>Syndrome<c PUR>) </hi> <w VBZ>is
 <w AT0>a  <w NN1>condition <w VVN>caused <w PRP>by <w AT0>a
 <w NN1>virus <w VVD-VVN>called <w NP0>HIV <c PUL>(
 <w AJ0-NN1>Human <w NN1-NP0>Immuno <w NP0>Deficiency
 <w NP0>Virus<c PUR>)<c PUN>.
```

Figure 2.5 SGML POS encoding in the British National Corpus

TEI-conformant mark-up can also be applied at other levels of corpus annotation, for example, to the encoding of skeleton treebanks (Leech and Eyes, 1997: 51).

Following the World Wide Web Consortium's (W3C) release of the recommendation for an Extensible Mark-up Language[12] (XML) in 1998, SGML encoding initiatives for corpora have evolved into encoding in XML. The TEI Guidelines are now fully XML-compliant and published as TEI P4 guidelines (Sperberg-McQueen and Burnard, 2002). XML mark-up looks like SGML (and HTML) and XML schemas (rather than DTDs) are used to define the structure of XML documents.

The TIGER project in Stuttgart has also proposed an XML encoding scheme for syntactically annotated corpora, and this is shown in Figure 2.6 (Mengel and Lezius, 2000).

```
<s id="s1" href="#id(n1_500)"/>
<n id="n1_500" cat="S">
<edge id="edge1_1" href="#id(n1_501)"/>
<edge id="edge1_2" href="#id(n1_502)"/>
</n>
<n id="n1_501" cat="NP">
<edge id="edge1_3" href="#id(w1_0)"/>
<edge id="edge1_4" href="#id(w1_1)"/>
</n>
<n id="n1_502" cat="VP">
<edge id="edge1_5" href="#id(w1_2)"/>
<edge id="edge1_6" href="#id(n1_503)"/>
</n>
<n id="n1_503" cat="NP">
<edge id="edge1_7" href="#id(w1_3)"/>
<edge id="edge1_8" href="#id(w1_4)"/>
</n>
<w id="w1_0" word="The"/>
<w id="w1_1" word="boy"/>
<w id="w1_2" word="likes"/>
<w id="w1_3" word="the"/>
<w id="w1_4" word="girl"/>
```

Figure 2.6 TIGER XML encoding sample for a syntactically annotated corpus

[12] For further details, see the website at http://www.w3.org/XML/

The Expert Advisory Group on Language Engineering Standards (EAGLES)[13] project was funded by the European Commission. The aim of EAGLES was to "accelerate the provision of standards for (1) very large-scale language resources (such as text corpora, computational lexicons and speech corpora); (2) means of manipulating such knowledge, via computational linguistic formalisms, mark-up languages and various software tools; and (3) means of assessing and evaluating resources, tools and products". Amongst the EAGLES guidelines is one detailing recommendations for corpus encoding, the CES (Corpus Encoding Standard) guidelines (Ide 1996, 1998). This work is now being continued in XCES[14], the corpus encoding standard for XML, and it is planned for XCES to be used as the standard for encoding the forthcoming American National Corpus[15] which is partially modelled on the British National Corpus. The work of EAGLES is continuing in the International Standards for Language Engineering (ISLE) initiative[16].

So far, we have seen how annotations are usually interspersed with the base text, as part of the same composite document. Two other arrangements are possible. One is to use the form of a relational database, and the other is to hold the base text and the annotations in separate files, called *stand-off annotation*, with links relating each part of the one to each part of the other. These two possibilities are expanded upon in section 3.2.

The work on annotation graphs in the Linguistic Data Consortium at the University of Pennsylvania focuses on the logical structure of linguistic annotations and aims to cover a common conceptual core (Bird and Liberman, 1998). They contend that the file formats and the tags and attributes for describing content are secondary (Bird and Liberman, 2001). Their work is now producing open-source tools for annotation of time-series data (Bird et al, 2002).

---

[13] See the website at http://www.ilc.pi.cnr.it/EAGLES/home.html

[14] Which stands for XML CES; at the time of writing, version 0.2 of XCES appears on http://www.cs.vassar.edu/XCES/ and http://www.xml-ces.org/

[15] See the website at http://americannationalcorpus.org/

[16] Websites at http://www.mpi.nl/world/ISLE/index.html and http://lingue.ilc.pi.cnr.it/EAGLES96/isle/

As well as standards for the encoding of corpus annotation, the European Union has funded the writing of standards for tagsets. The EAGLES morphosyntactic annotation guidelines were disseminated in 1994 in the interests of interchange and reuse of annotated corpora (Leech and Wilson, 1999). The guidelines describe an extensible intermediate tagset mainly influenced by Indo-European languages.

At the time of writing, other initiatives are ongoing, such as OLAC[17] (Open Language Archives Community) which "is an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources by: (i) developing consensus on best current practice for the digital archiving of language resources, and (ii) developing a network of interoperating repositories and services for housing and accessing such resources". The Linguistic Data Consortium now documents all of its corpora using the OLAC metadata set.

2.5 The main research centres

Corpus linguistics draws on research from academics in several disciplines: statistics, linguistics, and computing. Through the work of the main research centres, we will see how modern corpus linguistics has developed over the last thirty years. Let us first focus on the UCREL (University Centre for Computer Corpus Research on Language) research centre at Lancaster University, described by Mair (1997) as "one of the hotbeds of corpus linguistics".

The software and method described in this thesis were developed as a result of the author's experiences within several UCREL research projects, and by working with other staff and researchers on related projects. It is useful at this point to detail some of the history of the UCREL research centre to put into historical context the work described in this thesis.

The research centre draws upon the expertise of the Department of Linguistics and Modern English Language and the Department of Computing. UCREL has pioneered

[17] See the website at http://www.language-archives.org/

an approach to natural language processing that is based on corpus linguistics. UCREL's work is very much focused on practical outcomes. It has engaged in corpus-based research contributing to such practical applications as speech synthesis, speech recognition, machine-aided translation, dictionary publishing, social surveys, interview analysis, and computer-aided language teaching.

UCREL began its existence in 1970 when Geoffrey Leech founded a group under the name of CAMET (Computer Archive of Modern English Texts) within the then Department of English. The CAMET group's aim was to compile a one-million word corpus of written British English in machine-readable form as a parallel to the Brown corpus of American English developed at Brown University, Rhode Island, by Nelson Francis and Henry Kučera (the first computer corpus of English). The project, which was completed in 1978, was assisted by the involvement of the Norwegian universities of Oslo and Bergen, and the completed corpus was hence called the Lancaster/Oslo-Bergen (or LOB) corpus (Johansson, Leech and Goodluck, 1978).

The Brown University team had developed computer software (TAGGIT) for assigning a part of speech (or word class) to each word in a corpus or text (Greene and Rubin, 1971). This software had a success rate of around 77% without manual intervention. In order to carry out a grammatical analysis of the corpus, the Lancaster team developed POS tagging software. This work, which involved collaboration with Roger Garside of the Computer Studies Department, culminated in the first version of the software package called *CLAWS* (Constituent Likelihood Automatic Word-tagging System) which, in much revised and improved form, is still a major component of UCREL's work (Garside, 1987; Leech et al, 1994b; Garside and Smith, 1997). The CLAWS system employed a rich blend of decision making techniques, based in particular on statistical probabilities of tag co-occurrences, using data derived from the manually-corrected tagged LOB corpus (Marshall, 1983). It achieved a success rate without manual intervention in the high 90s percentage accuracy. We will describe the CLAWS software in more detail in section 3.2.1.

The increased emphasis on corpus analysis rather than compilation, and especially the use of computational methods, led to a formalisation of the links between what were

by then called the Department of Linguistics and Modern English Language and the Department of Computing and CAMET was thus transformed into UCREL.

In 1983-84 UCREL began further work on the grammatical analysis of the LOB corpus (see Atwell, Leech and Garside, 1984). It included (i) the development of a syntactic parser using probabilistic models of analysis to determine the most likely analysis of a sentence (Beale, 1985a; 1985b; Garside and Leech, 1985; Leech, 1987) (ii) the production of a manually parsed sample of the LOB corpus as a data source for the probabilistic parser (Sampson, 1987b), now known as the Lancaster-Leeds Treebank (iii) the production of software which can derive a distributional lexicon from CLAWS tagged text, i.e., a lexicon showing the base form of each word (i.e. a *lemma*, or *head word* in a dictionary), all its inflectional variants, and the frequencies of its lexical collocations in the corpus (Beale, 1987; 1989).

A context-sensitive spelling checker project (funded by the computer manufacturers ICL) employed a probabilistic approach to the identification of spelling errors which, because they happened to form correctly-spelled English words different from the 'target' word, would not be detected by the ordinary type of spelling checker; for example the confusion of 'there' and 'their' (Atwell and Elliott, 1987).

Also in 1984, UCREL began a project with IBM UK Laboratories that involved with speech processing. The main goal of this project was to compile a corpus of modern spoken English, with a prosodic transcription showing features of stress, intonation and pauses (Taylor and Knowles, 1988). The resulting corpus of c. 53,000 words, the Lancaster-IBM Spoken English Corpus (SEC), also exists on audio-tape for instrumental analysis. Funding was later extended for a study of prosodic stylistics to examine how prosodic patterns differ between different kinds of discourse (Wichmann, 1991), and a grant was obtained to produce a speech database (MARSEC) based on the SEC. Among other things, the MARSEC project produced a digitised sound version of the corpus, and a phonetic transcription of the texts.

During IBM's ongoing research on the development of continuous speech recognition systems, UCREL's role was to perform the automatic tagging and subsequent manual parsing of large amounts of text using a fast annotation interface (Leech and Garside,

1991). The resulting 'treebank' was applied in the development of probabilistic models of language. The scope of annotation was extended to provide texts annotated with anaphoric relations between pronouns and noun phrases (Fligelstone, 1992).

In the summer of 1988, UCREL branched out into semantic analysis (tagging words based on their semantic field). In the ACASD (automatic content analysis of spoken discourse) project and others, success rates for semantic tagging of over 90% were achieved (Wilson, 1991, Wilson and Rayson, 1993, Rayson and Wilson, 1996). The resulting system called USAS is described in more detail in section 3.2.2. Work on semantic analysis was also carried out on the Lancaster Database project (Fligelstone 1995).

From 1991, UCREL participated in the British National Corpus project. UCREL was a member of a national consortium of academic and industrial partners, the other members being Oxford University Press, Oxford University Computing Services, The British Library, Longman Group Ltd. and W. & R. Chambers Ltd. The goal of the BNC project was one of compiling a representative 100-million word corpus containing a wide variety of present-day written and spoken British English (see Burnard, 1995 and Leech, Garside and Bryant, 1994a).

In the mid-1990s, UCREL moved into a new multilingual phase of development, to produce bilingual and trilingual corpora (McEnery and Daille, 1993). In order to reflect this changing nature of UCREL's research, and to emphasise its position as a research centre within the University, UCREL changed its name in June 1995. Hence the Unit for Computer Research on the English Language became the University Centre for Computer Corpus Research on Language. However, it retains the acronym UCREL. The non-English corpus work continued into UK non-indigenous minority languages in the EMILLE project, designed to build a 63 million word electronic corpus of South Asian languages, especially those spoken in the UK. In 2002, the LER-BIML project began to survey corpus work on the indigenous minority languages of the British Isles: Cornish, (Scottish) Gaelic, Irish, Manx, Scots, Ulster Scots (Ullans) and Welsh with a view to building small sample corpora of two of these languages.

There is a whole spectrum of different methodologies in the NLP domain. These range from techniques based on introspection alone, through empirical and corpus-based work, to the totally automatic statistical work carried out using neural nets. A neural network can be seen as a black box from which we can extract nothing to extend theories of language. Lancaster's approach as described above is largely a probabilistic one. Leech (1987:3) recognises the strengths and weaknesses of this approach: "it is able to deal with any kind of English language text which is presented to it: it is eminently robust" but "probability admits the possibility of error". Some mistakenly assume that this is the only methodology used within UCREL. Oostdijk (1991: 3) and Karlsson (1994) contrast it with the non-probabilistic rule-based approach used within the TOSCA group at Nijmegen University in The Netherlands. However, as described in Fligelstone et al (1996, 1997), UCREL increasingly employs a non-probabilistic technique called *template analysis* as a complement and sometimes an alternative to the statistical Markov-modelling methods. Template-based methods are used to reduce errors and/or ambiguity within CLAWS POS tagging, and more generally, without a statistical counterpart, in semantic annotation: for example in linking nouns with related adjectives or verbs (see Wilson 1993, Rayson and Wilson 1996). In any event, we can view an analysed corpus as a database that is consulted to obtain frequency and distribution information about linguistic structures. In fact, the software described in this thesis performs exactly this task.

The Nijmegen approach, as characterised by Aarts (1991), takes a corpus to be a collection of samples of running text. The samples can be spoken or written and they are subjected to syntactic analysis using a specific grammar formalism.

Other research on English corpus linguistics in the modern period has been centred in the UK and Scandinavian countries. We have already mentioned Oslo and Bergen in relation to the LOB corpus. One point of focus for activity is ICAME which is an international organisation of linguists and information scientists working with English machine-readable texts. The aim of the organisation is to collect and distribute information on English language material available for computer processing and on linguistic research completed or in progress on the material, to compile an archive of English text corpora in machine-readable form, and to make material available to

research institutions. The archive is based at the Norwegian Computing Centre for the Humanities (NCCH) in Bergen, Norway. Members of ICAME meet annually. In section 2.2, we summarised the results of 2 recent ICAME conferences.

Elsewhere in the UK, corpus research has been carried out on international varieties of English, within the International Corpus of English project co-ordinated at the Survey of English Usage at University College London. The Survey of English Usage was founded in 1959 by Professor Randolph Quirk (now Lord Quirk) to collect a million-word corpus, which samples written and spoken British English produced between c. 1955 and 1985. The corpus was originally compiled on paper, but is has since been digitised. Lexicographers and corpus linguists worked together in Glasgow and at the University of Birmingham on the Collins COBUILD project, resulting in the COBUILD series of dictionaries and grammars. Since 1980, the corpus (known as the Bank of English from 1991) has expanded to over 400 million words. The corpus mainly contains British and American newspapers, magazines and books.

Computational linguistics and language engineering research in the UK occurs on various research areas. Amongst others, this takes place on the lexicon and text generation at the University of Brighton, on natural language understanding at the University of Durham, on mark-up languages for encoding corpora at the University of Edinburgh (see related work on XML in section 2.4), on spoken dialogue processing and applications to English language teaching at the University of Leeds, on computer-aided language learning and information extraction at UMIST, on software architectures for NLP and information extraction at the University of Sheffield, and on anaphora resolution at the University of Wolverhampton.

2.6 Frequency Profiling

The Matrix method proposed in this thesis employs automatic profiling of corpora, and, more specifically, the comparison of those profiles. Therefore, this section reviews in more detail the work done in the area of frequency profiling of corpora. We look at models for the frequency distributions of words and in a subsequent section at the statistical comparison of frequency profiles.

2.6.1 Word frequency profiling

We began our examination of word frequency profiles in section 1.3 with a basic description of what they contain, and by mentioning their widespread use in corpus linguistic and classroom studies. For foreign or second language teaching, information about the frequencies of words is important for vocabulary grading and selection. Frequency studies also have applications to language teaching in such areas as syllabus design, materials writing, grading and language testing. For a recent view of the start of the art, Schmitt and McCarthy (1997) collected together many of these areas related to vocabulary. Historically, education was the driving force for frequency lists: see Thorndike (1921), (1932), Thorndike and Lorge (1944), Lorge (1949). Fries and Traver (1950) carried out an extensive survey of the English word lists available up to that time, discussed their various educational applications and compared seven of the major lists. In those early days, the source texts for the frequency lists were the ones used in the education of American children. Later counts included magazines and general reading material. A more modern and systematic project to obtain frequency counts from children's reading materials resulted in the *American Heritage Word Frequency Book* (Carroll et al, 1971). An improved kind of count (taking account of meaning but with a smaller wordlist) led to the publication of the *General Service List of English Words* by West (1953). Below the word level, Ljung (1974) published a frequency list of morphemes based on 8,000 of the most frequent words in the Thorndike-Lorge lists.

Other frequency lists have been compiled for particular varieties of English. For example, James *et al*. (1994) is a frequency book of the vocabulary of computer science; Dahl (1979) is a frequency book for the English of psychiatric interviews. The latter is one of the few existing frequency lists for spoken English, amongst others are an early list based on a limited corpus of 135,000 words (Jones and Sinclair, 1974), and that based on the spoken part of the BNC (Leech, Rayson and Wilson, 2001). The Michigan team are beginning work on a word frequency list for American academic spoken English, based on the MICASE corpus. If we consider

languages other than English, Juilland has produced a series of frequency dictionaries for Spanish, Rumanian and French (Juilland et al 1964, 1965 and 1970).

A third area of application for frequency-based word lists is that of natural language processing. NLP computer systems that process language need to know the probability of a word occurring in a text. This can be applied in, for example, machine translation or speech recognition software, where it is important to determine the most likely word to occur from a set of possible words. Finally, we can identify a fourth application for these lists, that of psychological research, where the frequency of vocabulary is valuable evidence in understanding the human processing of language.

Despite their usefulness as a starting point, there are problems with word frequency lists. The simple lists count inflectional variants of the same headword separately, so we may find the verb forms *kicked* and *kicks* high in our word count but the base form *kick* would be lower down the list. In order to study the usage of the lemma KICK as a whole we need to reduce all variants to the base headword[18] and count them together. This has to be done both for the verb lemma (*kick*, *kicked*, *kicking*, *kicks*) and for the corresponding noun lemma (*kick*, *kicks*). Leech, Rayson, and Wilson (2001) have produced lemmatised word frequency lists to overcome this problem. Simple word frequency lists often do not show frequencies for multi-word units (MWU). This usually relies on some automatic analysis to identify grammatical MWUs (e.g. the conjunction *so that*, the preposition *in spite of*, and *at least* as an adverb), or semantic MWUs (e.g. *kick the bucket*). Simple word lists also do not distinguish different words spelt the same (homographs), although this problem can be partly avoided if the lists are produced from a POS tagged corpus so that, for example, *score* as a noun is counted separately from *score* as a verb. Further ambiguities remain such as the noun *spring*, which can refer to a metal coil, a water source, or a season. This would need a fully automated word-sense analysis of the text, and such techniques are not mature enough to be used in large-scale projects as yet. Further practical problems of writing software that produces word frequency lists will be discussed in section 4.2.

[18] The headword is sometimes called the *lexeme*, and Sinclair (1999) and others call it the *definiendum* when it appears in a dictionary: “that which is to be defined”.

We will use real examples to illustrate the problems with using word frequency lists in section 4.4.

Even in a large comprehensively sampled corpus such as the BNC, the word frequency counts themselves can be misleading. This is not because we may have miscounted the words, but because of how the frequencies relate to use in the English language as a whole. If a word has a high frequency count, we may reasonably infer, due to the nature of the BNC, that the word has a similarly high currency of usage in the language. However, it is possible that the word has a high frequency not because it is widely used in the language as a whole but because it has high frequency in a much smaller number of texts, or parts of texts, within the corpus. To reveal such cases, we can calculate range or dispersion statistics. These show how widely spread the use of a word is: whether it is frequent because it occurs in a lot of text samples in the corpus or whether it is frequent because of a very high usage in only a subset of domains or genres. Frequent words with high dispersion values may be considered to have high currency in the language as a whole; high frequencies associated with low dispersion values should, in contrast, be treated with caution. In statistics, we use mean and standard deviation as summary measures. In corpus linguistics, these are analogous to frequency and dispersion. According to Fries and Traver (1950: 21), Thorndike was the first to introduce range values into frequency lists. Lyne (1985) surveys dispersion statistics in more detail and we will describe in section 4.2 how Matrix calculates the range and Juilland's dispersion statistics. An alternative approach to quoting separate dispersion and frequency statistics is to combine them into one value called *adjusted frequency* (or sometimes *coefficient of usage*). This is the method used by Francis and Kučera (1982: 464). They quote the dispersion measures by Juilland and Rosengren and describe how they can be combined with actual frequencies in order to place 'lemmas' in the Brown corpus in order ranked by their adjusted frequencies. The American Heritage Word Frequency Book (Carroll et al, 1971: xl) used a measure of relative entropy from information theory as a dispersion statistic, but similarly calculated an adjusted frequency measure from the dispersion.

Zipf (1935, 1949) established a logarithmic connection between the rank frequency of a word and the number of words at that rank. He also proposed a 'principle of least effort' for human language use. Among other things, this means that the words that

people use most often will also prove to be the shortest and simplest. In a frequency list, we can see this principle at work by looking at the lengths of words in terms of how many (spoken) syllables they contain. The BNC frequencies follow the pattern predicted by Zipf's principle (Leech, Rayson and Wilson, 2001: 121).

A lot of progress has been made since Zipf's early studies on word frequency distributions. Baayen (1993) compares three models (the lognormal law, the generalised inverse Gauss-Poisson law, and the extended generalised Zipf's law) with regard to estimating the theoretical vocabulary size. Baayen (1993: 361) writes that "the main challenge for future research in this area is to construct linguistically less naïve models that do not build on the unrealistic assumption that in language words appear at random". The three models presented are all large number of rare event (LNRE) models. Even large corpora with tens of millions of words are located in the LNRE zone (Baayen, 2001: 51). Following up his own challenge, Baayen (2001: 161) adjusts the LNRE models to take into account non-randomness in language.

Words are not selected at random in language. This has implications for carrying out statistical procedures on word frequencies, as we shall see in section 2.7. Choosing one word (or POS) constrains the choice of the following word (or POS), so that, for example, having chosen a determiner (e.g. *the*) the choices for what can grammatically follow are immediately limited (e.g. an adjective, adverb or noun). This constraint is what statistical part-of-speech taggers such as CLAWS rely on to assist prediction of the correct word-class tag (see section 3.2.1). Other factors influence word selection, such as author preference (related to language proficiency), collocations, topic, and text type. Church and Gale (1995) refer to the *bunchiness* or *burstiness* of words and show, as an example, the occurrences of the "very contagious" word "Kennedy" in the Brown corpus (because he was the president of the United States when the Brown corpus was compiled in 1961).

2.6.2 Annotation profiling

As discussed in the previous section, the usefulness of word frequency profiles increases once a corpus is annotated, because the annotation process will usually

identify sequences of words that should not generally be counted separately, but should be considered as a single unit. We can use annotated corpora to produce frequency lists of the annotation itself.

In corpus linguistics, frequency profiling of annotation has so far targeted the level of grammatical annotation. Francis and Kučera (1982: 534 – 546) provide frequency profiles of word classes and POS tags in the Brown corpus, and we will return to their comparison in section 2.7.1. Johansson and Hofland (1989: 27 – 39) show tag frequency profiles for the LOB corpus. Such profiles are mainly of use in NLP research: Johansson and Hofland (1989: volume 2) show frequencies of tag-pair combinations, and these were used to train the probabilistic model of CLAWS (see section 3.2.1 for the use of tag transition data).

In order to compare the frequency distribution for words against those for POS and semantic tags, we will examine the BNC sampler corpus. The sampler corpus contains two-million words selected from the main corpus, and includes equal proportions of written and spoken data. The BNC sampler is already CLAWS POS tagged, and we have automatically tagged the data using the USAS semantic tagger (see section 3.2.2).

According to our calculations the BNC sampler contains 2,052,440 word tokens, and 53,333 word types (distinct words). The semantically tagged corpus has a slightly reduced total frequency of 1,956,171 since we count semantic multi-word units as one item. The length of the annotation profiles are much smaller: 134 POS tags, and 499 semantic tags. Of the 53,333 word types there are 22,181 (41.6%) which occur only once in the BNC sampler. There are no POS tags that occur once, and only 17 (3.4%) of the semantic tags occur once.

Table 2.3 shows the frequency distributions for words, POS tags and semantic tags in the BNC sampler corpus. The same information is graphically represented in Figure 2.7. The numbers in the table show the percentage of types in the three frequency lists which fall within each of the frequency bands. The final four bands are single valued (4, 3, 2 and 1) since we wished to show the breakdown of the majority (68.1%) of the types in the word frequency profile which occur with a frequency of 4 or less.

Table 2.3 Frequency distribution for words and tags in the BNC sampler corpus

	Percentage of types		
Frequency band	**Words**	**POS tags**	**Semantic tags**
Greater than 500000	0.0	0.0	0.2
100000 – 499999	0.0	3.0	0.2
50000 – 99999	0.0	3.7	0.4
10000 – 49999	0.0	27.6	5.0
5000 – 9999	0.1	11.9	5.8
1000 – 4999	0.2	26.1	25.7
500 – 999	0.4	11.2	9.0
100 – 499	2.9	9.7	18.0
50 – 99	2.8	2.2	6.0
10 – 49	13.7	3.7	14.4
5 – 9	11.8	0.7	6.0
4	4.9	0.0	1.0
3	7.5	0.0	1.6
2	14.1	0.0	3.2
1	41.6	0.0	3.4

As can be seen from the table and corresponding figure, not many words in the corpus reoccur often, 79.9% of word types occur less than 10 times. We have already mentioned in section 2.6.1 that word frequency distributions are characterised by very large numbers of rare words. In contrast, the frequency distribution of POS tags is centred around the 1,000 to 50,000 frequency bands. Only 0.7% of the types occur less than 10 times. The distribution of semantic tags in the corpus is spread more widely across the frequency bands, with only 15.2% of the types occurring less than 10 times.

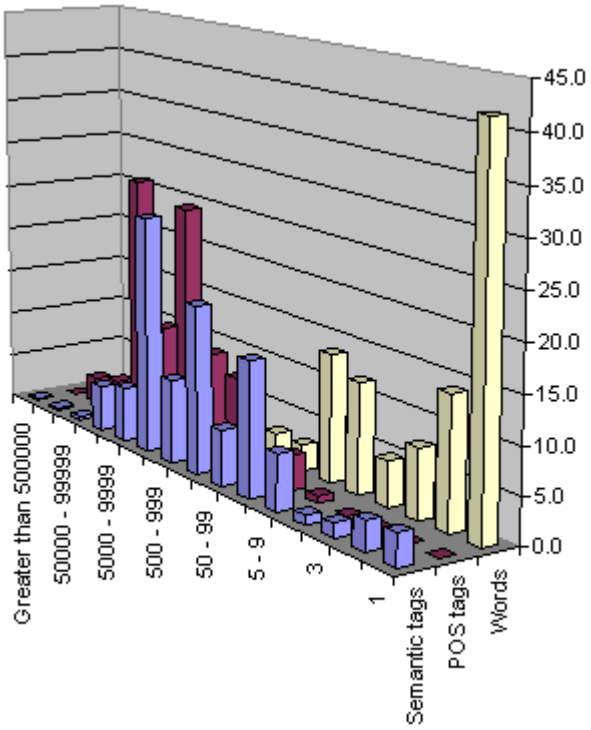


Figure 2.7 Frequency distribution for words and tags in the BNC sampler corpus

Frequency distributions for POS and semantic tags are sharply different than those for words. This information is of use if we wish to apply statistical techniques to examine tag frequencies as well as word frequencies. We will return to this data when we evaluate the Matrix method in section 5.2.

2.7 Comparison of frequency profiles

In recent years corpus-based techniques have increasingly been used to examine issues in language variation, that is, to compare language usage across corpora, users, genres, etc. We will use real examples to illustrate the practical problems with comparing word frequency lists in section 4.4. Comparison of one-million word corpora is becoming common even for beginners in corpus linguistics, with the increasing availability of corpora and the reasoning that one million words gives sufficient evidence for mid- to high-frequency words. However, with the production of large corpora such as the British National Corpus containing one hundred million words (Aston & Burnard, 1998), frequency comparisons are available across several

millions of words of text (Leech, Rayson and Wilson, 2001). Sufficient data for the investigation of relatively infrequent phenomena is still problematic (de Mönnink, 1997). However, research is still continuing on small corpora (Ghadessy, Henry and Roseberry, 2001).

There are two main types of corpus comparison:

- Type A: comparison of a sample corpus with a large(r) standard corpus (e.g. Scott, 2000b)
- Type B: comparison of two (roughly-) equal sized corpora (e.g. Granger, 1998)

In the first type (A), we refer to the large(r) corpus as a 'normative' corpus since it provides a text norm (or general language standard) against which we can compare. These two main types can be extended to the comparison of more than two corpora. For example, we may compare one normative corpus to several smaller corpora at the same time, or compare three or more equal sized corpora with each other. In general, however, this makes the results more difficult to interpret.

There are also a number of inter-related issues that need to be considered when comparing two (or more) corpora[19]:

- representativeness
- homogeneity within the corpora
- comparability of the corpora
- choice and reliability of statistical tests (for different sized corpora and other factors)

Biber (1993: 243) states that 'a corpus must be representative in order to be appropriately used as the basis for generalisations concerning a language as a whole'. Representativeness, in this sense, is seen as a particularly important attribute for a

[19] Alongside practical issues such as the cost and time taken in obtaining, or collecting such corpora.

normative corpus when comparing a sample corpus to a 'general language' corpus (such as the BNC) that contains sections from many different text types and domains. To be representative of the language as a whole, a corpus should contain samples of all major text types (Leech, 1993) and, if possible, be in some way proportional to their usage in 'every day language' (Clear, 1992). This first type of comparison (A) is intended to discover features in the research corpus which have significantly different usage (i.e. frequency) to that found in 'general' language. Representativeness can also apply to specialised corpora, whenever the researcher wants to make a claim about language use in a particular genre or domain, rather than the language as a whole.

The second type of comparison (B) is one that views corpora as equals. It aims to discover features in the corpora that distinguish one from another. Homogeneity (Stubbs, 1996: 152) within each of the corpora is important here since we may find that the results reflect sections within one of the corpora that are unlike other sections in either of the corpora under consideration (Kilgarriff, 1997). Comparability is of interest too, since the corpora should have been sampled for in the same way; in other words, the same stratified sampling method and with, if possible, randomised methods of sample selection. This is the case with the Brown and LOB corpora, since LOB was designed to be comparable to the Brown corpus, and neither corpus was designed to be homogeneous[20]. In the International Corpus of English (Greenbaum, 1996) comparability is taken a stage further, since it contains a large number[21] of comparable corpora from national and regional varieties of English around the world.

The final issue, which has been addressed elsewhere, and which we will examine further in section 5.2, is the one regarding the reliability of the statistical tests in relation to the size of the corpora under consideration. In the next section, we will consider some of the statistical tests previously used.

[20] Similarly with the 1991 replicates Frown and FLOB compiled at Freiburg

[21] Fifteen according to the ICE website at http://www.ucl.ac.uk/english-usage/ice/index.htm and twenty one according to Nelson et al (2002: 2)

2.7.1 Statistical tests

One of the largest early studies was the comparison of one million words of American English (the Brown corpus) with one million words of British English (the LOB corpus) by Hofland and Johansson (1982). A difference coefficient defined by Yule (1944) assessed the difference in the relative frequency of a word in the two corpora:

$$\frac{\text{Freq}_{LOB} - \text{Freq}_{Brown}}{\text{Freq}_{LOB} + \text{Freq}_{Brown}}$$

The value of the coefficient varies between +1 and –1. A positive value indicates overuse in the LOB corpus, a negative value shows overuse in the Brown corpus. A statistical goodness-of-fit test, the chi-squared test (χ^2), was also used to compare word frequencies across the two corpora. The chi-squared test was calculated as follows:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \text{ where } E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

where O_i is the observed frequency, E_i is the expected frequency, and N_i is the total frequency in corpus i (i in this case takes the values 1 and 2 for the LOB and Brown corpora respectively). Hofland and Johansson marked any resulting chi-squared values that indicated that a statistically significant difference at the 5%, 1%, or 0.1% level had been detected between the frequency of a word in American English and in British English. In some cases the expected frequency in at least one of the two corpora was too low for their calculation. The null hypothesis of the test is that there is no difference between the observed frequencies of a word in the two corpora. Note that even if the null hypothesis is not rejected, we cannot conclude that it is true. The cut-off value corresponding to the chosen degree of confidence may not be exceeded, but this only indicates there is not enough evidence to reject the null hypothesis (Krenn and Samuelsson, 1997: 36). Nelson et al (2002: 257) discuss further details of experimental design relevant to a parsed corpus. Critical values for the chi-squared statistic are listed in statistical tables such as those in Barnett and Cronin (1986) and Oakes (1998: 266). For example, the critical value for the 5% level, usually shown as

0.05 in the tables, is 3.84 at 1 degree of freedom. Leech and Fallon (1992) used the lists produced by Hofland and Johansson to examine evidence of cultural differences between America and Britain in 1961. We will examine their method in more detail in section 4.3.

It was Pearson (1904) who originally suggested the chi-squared test using the statistic (χ^2) described above for testing the independence of two variables. It is applicable to a general two-dimensional contingency table with *r* rows and *c* columns. The number of degrees of freedom (d.f.), which is used when looking up critical values, is the number of independent terms given that the marginal totals in the table are fixed. In corpus linguistics, we usually use a 2 × 2 table to compare frequencies of words or other linguistic features between two corpora, so d.f. as calculated by (r-1)×(c-1) is equal to 1. In this specific case, the 2 × 2 contingency table is as shown in Table 2.4. The chi-squared test can also be used to test how well a model fits the observed data, for example taking the expected values from the normal distribution (Woods et al, 1986: 135).

Table 2.4 Contingency table for the chi-squared test

	CORPUS ONE	**CORPUS TWO**	**TOTAL**
Frequency of feature	a	b	a+b
Frequency of feature not occurring	c	d	c+d
TOTAL	a+c	b+d	N=a+b+c+d

Hence, we can calculate the chi-squared statistic (X^2) as follows:

$$X^2 = \frac{N(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Francis and Kučera (1982: 534) also make use of the chi-squared statistic to compare the frequency distribution of grammatical tags across the fifteen genres in the Brown

corpus. They report that apart from low frequency tags, all the chi-squared statistics are statistically significant, even at the 1% level (at p = 0.01 for 14 degrees of freedom the critical value of chi-squared is 29.1). However, when comparing the frequency distribution of word classes across the two major subdivisions of the corpus, informative prose and imaginative prose, Francis and Kučera (1982: 544) use a normalised ratio value (NR). The ratio is normalised to take account of the fact that the informative section of the corpus is nearly three times larger than the imaginative section of the corpus. An NR value of more than 1 indicates a greater occurrence in informative prose, while a value of less than 1 points to a higher relative frequency in imaginative texts. The greater the NR deviates from 1, the greater the grouping of a particular word class in one of the sections of the corpus. Comparing NR values is problematic since they are not on a linear scale, and the calculation over-values smaller relative differences for lower frequency items compared to higher frequency items (see section 5.2). Francis and Kučera (1982: 547) also report on their use of the Mosteller-Rourke (MR) adjustment for chi-squared for large numbers. The MR value is calculated as follows:

$$MR = \frac{1000\chi^2}{n}$$

where n is the frequency of an item in the whole corpus (Mosteller and Rourke, 1973: 191). The resulting values cannot be assessed for significance in the chi-squared tables, but they are used to rank items according to their MR value. In effect, MR reduces the chi-squared values for items occurring more than 1000 times, and increases the values for items with a frequency less than 1000. This seems a rather arbitrary figure for our purposes, chosen to show 'nice numbers', and if anything the figure should be dependent on the corpus size(s).

In the field of information retrieval, word significance statistics have been used in the automatic preparation of abstracts, indexes and stop-lists. For the purposes of automatic classification of documents, Paice (1977: 75 – 79) describes the selection of key-terms by frequency and other non-statistical methods. Non-frequency based methods usually rely on thinning out words contained in a stop list, which specify function or closed class words, and allow us to concentrate on the content words in a document. Kabán and Girolami (2000) describe the use of Independent Component

Analysis for clustering and keyword identification in document collections. Berg (1997) reports on four measures for the comparison of the relative frequency of a word in a document (*f*) with the relative frequency of the same word in general usage (*r*):

$$S_1 = f - r$$

$$S_2 = \frac{f}{r}$$

$$S_3 = \frac{f}{f + r}$$

$$S_4 = \log\left(\frac{f}{r}\right)$$

As Berg reports, problems occur since his document contains words that are not found in his corpus of general English, so the value of *r* for these words is zero. This results in division by zero failures in S_2 and S_4, S_1 is always equal to *f*, and S_3 always equal to 1. His solution is to omit the words unique to his document from calculation. This is reasonable for his application which is the preparation of stop-lists, but would not be suitable for linguistic applications.

Word frequency lists have been used by corpus linguists to differentiate types of language usage. They have also been compared to find common words for learners' dictionaries and domain-independent lexicons for NLP applications. Copeck et al (1999) write that "while each frequency list is unique, ones based on written, general-purpose corpora have a statistically significant degree of resemblance when their shared vocabulary words are distinguished from proper names and other words". Their paper examines correlation between high ranking words in four corpora: BNC, Brown, LOB and WSJ (Wall Street Journal).

Kilgarriff (1996a, 1996b) points out that in the Brown versus LOB comparison many common words are marked as having significant chi-squared statistics, and that because words are not selected at random in language (as we have seen in section 2.6.1) we will always see a large number of differences in two such text collections. He selects the Mann-Whitney test that uses ranks of frequency data rather than the frequency values themselves to compute the statistic. Kilgarriff selects the Mann-

Whitney test because it "does not give undue weight to single documents with a high [frequency] count" for a particular word. However, he observes that even with the new test 60% of words are marked as significant. Ignoring the actual frequency of occurrence as in the Mann-Whitney test means discarding most of the evidence we have about the distribution of words, so the test will have lower discriminatory power. Due to problems of too many zeros in the Mann-Whitney test, Kilgarriff (2001) reports that his technique omits words with less than 30 occurrences in the joint LOB and Brown corpus. This is a major drawback with the Mann-Whitney test, here it omits 92% of the types in the joint corpus. The test assumes ordinal rating scales (Butler 1985: 98, Oakes 1998: 17) and this is why Kilgarriff has to use ranks rather than actual frequencies. One other problem is many words share ranks at the low end of frequency lists, especially for large corpora, as we have seen in section 2.6.2. Copeck et al (1999) report that 18,630 words occur six times – 10 percent of their list for the BNC. Within each rank words are ordered alphabetically. Additionally, comparing rank lists between different-sized corpora is also problematic. Copeck et al (1999) note the sizes of their frequency lists for LOB (7,950) and WSJ (4,550). This means that ranks for middle and lower frequency words in the BNC fall outside this range. These points suggest that the Mann-Whitney ranks test is suitable only for investigating mid to high frequency words when comparing corpora of the same size.

Numerous other authors have used the chi-squared test to determine significant frequency differences of individual words or other linguistics features, rather than whole frequency profiles, between two corpora (for example Woods et al 1986: 140, Virtanen 1997, Oakes 1998: 26, Roland et al 2000, Wikberg 1999). Kessler (2001: 55) performs the chi-squared test on a larger table to compare word-initial consonants in Swadesh[22] lists of pairs of languages to assist in determining ancestral connections. Many authors also apply Yates' continuity correction (1934) developed to improve the approximation of the continuous probability distribution (chi-squared) to the discrete probability distribution of the observed frequency (multinomial). The Yates' corrected chi-squared statistic (Y^2) is calculated as follows (from Table 2.4):

[22] A list of concepts supposedly basic in the vocabulary of all languages, proposed by Morris Swadesh in the 1950s for use in studies of glottochronology.

$$Y^2 = \frac{N(|ad - bc| - 0.5N)^2}{(a+b)(c+d)(a+c)(b+d)}$$

In some texts, its use has been recommended (Everitt 1992: 14, Butler 1985: 122, and Woods et al 1986: 146), but current statistical textbooks report that the correction is less important than it was once thought. Agresti (1990: 68) notes that "the corrected statistic gives P-values (from the chi-squared distribution) that better approximate hypergeometric probabilities obtained with Fisher's exact test. This adjustment is *not* intended to make the sampling distribution closer to the reference chi-squared distribution". Fisher's exact test may be used for tables with small expected frequencies as an alternative to the chi-squared test. It uses the observed frequencies themselves to find the probability (P) of obtaining any particular arrangement of frequencies a, b, c, and d (from Table 2.4):

$$P = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!N!}$$

where *a!* is 'a factorial' (the product of *a* and all the whole numbers less than it, down to one, 0! = 1). The P value is then compared directly to the probability level, e.g. 0.05 for 5%, or 0.01 for 1%, to indicate departure from the null hypothesis in a specific direction. It is a *one-tailed test* whereas the chi-squared is two-tailed. The P value may be doubled in order to compare it with the probability obtained through the chi-squared test. Fisher's exact test is computationally expensive since it involves calculating factorials, and it involves calculating P for every possible arrangement of frequencies keeping the marginal totals fixed. We can terminate the calculation without completing it if the cumulative total of the P values is larger than our chosen significance level. Mehta and Patel (1983) propose a network algorithm to reduce the time taken for the calculation and more recently this has been included in software such as R (R is a language and environment for statistical computing and graphics[23]), and their own software product called StatXact[24].

[23] For more details on R, see http://www.r-project.org/

[24] Produced by Cytel Software Corporation, see http://www.cytel.com/

Dunning (1993) reports that we should not rely on the assumption of a normal distribution when performing statistical text analysis and suggests that parametric analysis based on the binomial or multinomial distributions is a better alternative for smaller texts. Hence Dunning proposes the log-likelihood ratio as an alternative to Pearson's chi-squared test, and he demonstrates this for the extraction of significant bigrams from text. Conversely, Mosteller and Rourke (1973: 162) state that the chi-squared statistic assumes a multinomial distribution, as do Cressie and Read (1994). Woods et al (1986: 188) describe the chi-squared test for association as non-parametric and state that it makes no special distributional assumptions of normality. There seems to be some confusion in the literature. Everitt (1992: 5-8) explains the situation more clearly. It is the observed frequencies that are assumed to follow a multinomial distribution, whereas the chi-squared distribution, which is used to calculate and tabulate critical values, arises from the normal distribution. Some papers in the literature report that the chi-squared statistic becomes unreliable when the expected frequency is *too small*, and *possibly* overestimates significance with high frequency words and when comparing a relatively small corpus to a much larger one. The former of these vague terms has been taken as meaning that all *expected* values must be greater than 5 (for example, Butler 1985: 117, Woods et al 1986: 144), and sometimes the same limit is applied to the *observed* frequencies (De Cock, 1998 and Nelson et al, 2002: 277). It was Cochran (1954) who suggested a rule that 4 in 5 (80%) of the expected values in an $r \times c$ table should be 5 or more. In the 2×2 table case, this means all cells should have expected values of 5 or more. Everitt (1992: 39) cites other more recent work than Cochran which suggests that this rule is too conservative. Butler (1985: 117) suggests a solution to this is to combine frequencies until the combined classes have an expected frequency of 5 or more, likewise Nelson et al (2002: 277) for the observed frequencies, but Everitt (1992: 41) argues against this practice.

Everitt (1992: 72) also mentions that the chi-squared statistic is "easily shown to be an approximation to" the log-likelihood for large samples. The two statistics take similar values for many tables. Williams (1976) notes that the log-likelihood is preferable to Pearson's chi-squared in general. Everitt (1992: 18) also notes that the chi-squared test, Yates' corrected chi-squared and Fisher's exact test are equivalent in large

samples. The obvious question, then, is: what constitutes a large sample? Kretzschmar et al (1997) start to answer the question by estimating sample sizes for various confidence levels. Scott (2001b) uses the log-likelihood statistic in his keywords procedure, as we shall see in section 2.7.2.

For the 2 × 2 case (in Table 2.4), the log-likelihood ratio is calculated as follows:

$$G^2 = 2\,(a\ln a + b\ln b + c\ln c + d\ln d + N\ln N - (a+b)\ln(a+b) - (a+c)\ln(a+c) - (b+d)\ln(b+d) - (c+d)\ln(c+d))$$

Cressie and Read (1984) show that Pearson's X^2 (chi-squared) and the likelihood ratio G^2 (Dunning's log-likelihood) are in fact two statistics in a continuum defined by the power-divergence family of statistics. They go on to describe this family in later work (1988, 1989). Here they also make reference to the long and continuing discussion (since 1900) of the normal and chi-squared approximations for X^2 and G^2, and 2 × 2 contingency tables, during which many alternative tests have been devised (Yates, 1984).

Another test that is available is McNemar's chi-squared test (Everitt, 1992: 20) for matched samples as cited in van Halteren et al (1998) and Dietterich (1998), but this is not suitable for our present purpose since it would ignore frequencies of the same item in the two corpora.

In a different application area, that of extracting groups of associated words, Weeber et al (2000) use a combination of the log-likelihood ratio and Fisher's exact test for the full word frequency range. Hisamitsu and Niwa (2001) and Kageura (1999) compare bigram statistics and agree with Dunning that log-likelihood is preferable to chi-squared. Pederson et al (1996) mention that the two-sample t-test can be applied to the bigram difference problem, and that this is equivalent to Pearson's X^2 for 2 × 2 tables. Pederson and his colleagues opt for an exact conditional test using a Monte Carlo sampling scheme from the exact distribution, rather than using Fisher's exact test to enumerate all elements of the distribution. They decide against log-likelihood and chi-squared since 2 × 2 tables representing bigram data are always heavily

skewed (also see Pederson, 1996). Noreen (1989: 91) prefers a (computationally-intensive) approximate randomisation method over Monte Carlo sampling to test when variables are unrelated. The randomisation tests are used by Yeh (2000) for testing differences in values of metrics such as recall and precision.

The 2 × 2 tables for comparing word frequencies across corpora are, in general, not as heavily skewed as for bigram data, so the log-likelihood ratio is suitable for our use. For the reasons described in the preceding paragraphs, we chose to use the log-likelihood ratio in our work and the procedure is described in more detail in section 4.3. This choice will be evaluated in section 5.2 where we examine situations where the contingency table may become skewed.

2.7.2 Systematic approaches

Up to this point, we have described approaches that compare frequencies across corpora in an ad hoc way (on a word-by-word basis, for example). The keyword approach taken by Scott takes a more systematic approach. We will be reviewing Scott's WordSmith Tools (1996-99) in more detail in the next chapter, but for now we will focus on the Keyword module as described in Scott (1997, 1998, 2001a). Tribble (2000: 79-80) describes the way that WordSmith finds keywords as follows:

1. frequency sorted wordlists are generated for a 'reference' corpus (a collection that is larger than the individual text or collection of texts which will be studied), and for the research text or texts
2. each word in the research text is compared with its equivalent in the reference text and the program makes a judgement as to whether or not there is a statistically significant difference between the frequencies of the word in the different corpora. The statistical test evaluates the difference between counts per type and total words in each text and can be based either on a chi-squared test for outstandingness or on a log-likelihood procedure
3. the wordlist for the research corpus is reordered in terms of the 'keyness' of each word

Scott (1997) sets a minimum threshold of two occurrences for each word in the research text, although this does result in manually identified keywords being omitted from the keywords database (Scott, 2001b: 118). Other words with frequencies that violate the Cochran rule (see section 2.7.1) are still included in the keyword listing since in practice they are still interesting. The resulting keyword list contains two types of keyword: *positive* (those which are unusually frequent in the target corpus in comparison with the reference corpus), and *negative* (those which are unusually infrequent in the target corpus). These correspond to the terms *overuse* and *underuse*, as discussed in the study mentioned in section 2.3, conducted by Ringbom (1998). Tribble compares the list of positive and negative keywords against the frequency list for his corpus and demonstrates the improved usefulness of the keyword technique over simple frequencies for extracting interesting lexical items for stylistic studies. Scott also uses the notion of key-keywords. These are words that are key in all, or a large percentage, of the texts that are contained in the corpus under investigation. Tribble uses this feature to select lexical items to give pedagogical insights in the study of a particular genre.

Collins and Scott (1997: 190) extend the keyword technique to examine the lexical landscapes of business meetings, as follows:

1. prepare a word list from the text of the business meeting
2. remove from the list words which are carriers of low-level meaning, i.e. very frequent words that contribute to uses, rather than the meaning, e.g. *issue*, *get*, *make*
3. remove from the list items required for the syntactic structure of sentences e.g. *then*, *and*, *but*, *in*, *the*, *this*, *be*, *where*
4. remove from the list words which are carriers of interpersonal meaning, e.g. *must*, *may*, *very*, *hullo*
5. prepare a keyword list by comparing the resulting list to a similarly prepared list from a reference corpus
6. lemmatise the keyword list, keeping the most frequent member of the lemma in each case

The list-stripping process was intended to identify the ideational (propositional) words. Collins and Scott then looked at collocational and intercollocational links of the keywords to produce topical nets which were graphically visualised.

Scott (1997) relates his work to that of Raymond Williams in the 1970s in terms of its purpose, but not in terms of its procedure. Williams (1983: 14) selected keywords subjectively due to their use in general discussion in 'interesting or difficult ways'. Scott's motivation for his work is a text-focused one, not one aiming to 'characterise a language or a genre, but a language event', and to reveal patterns which construct texts. He argues for the study of texts in their original context with as much detail as possible recorded about the writers or speakers that produced the data. However, he is realistic about recording information about the original circumstances of the language event, such as the mood of the speaker or writer, which may be difficult to recover even for those involved in producing the language. There is no claim that the keywords would match those selected by human readers of a text (Scott, 2000a), who may specify a word not even in the text. Scott (2000b) defines an *association* relationship between words, as the co-keyness of both words within the same text, as an alternative to the standard calculation of collocation, which is based on how frequently words occur near to each other. He uses association across a large corpus to investigate the 'aboutness' or content of texts.

Tribble (2001) notes that keywords regularly occupy potential theme positions in sentences and paragraphs. We can distinguish the Matrix technique from Scott's since we use the whole corpus rather than texts within the corpus to build key-keywords. Scott usually focuses on key open-class words, although both Matrix and Scott's technique may extract closed-class words as well (Scott, 2001b: 126). Berber Sardinha (1999) points out one practical problem with the keywords technique: that it normally produces more key words than it is possible for the researcher to analyse. He proposes two techniques to reduce the set of words: by selecting a simple majority (i.e. half the number plus one), and by selecting a significant subset (by using the chi-squared test again).

Pęzik (forthcoming) examines the notion of keyness and how it links to ontology and taxonomy research[25]. He describes how it can be incorporated in a web search engine model for the automatic generation of keywords and indexes. Some researchers use the concept of keywords in a different way: they are not identified statistically. For example, Stubbs (1996: 172) describes *cultural keywords*, that is, "words which capture important social and political facts about a community" (Hunston, 2002: 117). The important feature for Stubbs is that these words occur in characteristic collocations, which show the associations and connotations they have. Stubbs (1996: 166) traces his efforts back to that of Firth (1935) on "focal and pivotal words" and to Williams' book on keywords. He writes that identification of keywords will always involve intuition, but then relies on a systematic method for searching for fixed phrases in corpora. In his study of language of Euro scepticism in Britain, Teubert (2001) manually selects keywords from a pilot corpus and supplements them with significantly frequent collocates of the keywords, in a larger corpus. Similarly, Ooi (2000) selects 10 lexical items for their supposed cultural distinctiveness and examines their collocates. The work of Wierzbicka (1997: 16) is focussed on keywords but has no "objective discovery procedure" for them. Frequency information does play a part in determining whether a candidate word is a common word.

Pre-dating the work of Scott, is that of Lyne (1985: 164) who calculates a 'registral value' for each word (instead of using a goodness-of-fit statistic) and sorts on the value to compare frequency data in two corpora. Lyne's goal was to find characteristic vocabulary of French business correspondence. Lyne also proposes a modified registral value which is adjusted for range to filter out technical items.

Keywords have been used in information retrieval and automatic abstracting. The pioneering techniques were based on frequency of words and their relative position in a sentence (Luhn, 1958), and filtering to select a subset of frequent words (Edmundson, 1969). This has recently been extended to key-phrases using a naïve Bayes learning scheme (Frank et al, 1999). Sparck Jones (1971: 74) cautions against

[25] Research on hierarchical, semantic networks, exhaustively incorporating all concepts from a given knowledge domain (Guarino, 1995)

the use of very frequently occurring keywords in information retrieval due to the retrieval of many unwanted documents.

Other relevant studies in this area are the works of Biber and Finegan, see for example, Biber (1988), Biber and Finegan (1989), and Biber (1995). These have at their core a comparison of frequency distributions across genres, but use a multi-feature, multi-dimensional methodology, grouping sets of linguistic features associated with a number of factors (called text dimensions). Biber (1988: 63) describes his approach as depending on both the type I (microscopic) and type II (macroscopic) research methodologies. He uses the type II approach to analyse the co-occurrence patterns among the linguistic features, identifying the textual dimensions, and the type I analyses to interpret these dimensions. The main methodological steps of Biber's technique are as follows:

1. review previous research to identify important linguistic features
2. collect texts
3. count occurrence of features in the texts
4. perform factor analysis: clustering of features into groups of features that co-occur with a high frequency in particular texts
5. interpret factors as textual dimensions
6. for each factor, compute a factor score for each text
7. compute an average factor score for texts in each genre
8. interpret the textual dimensions in the light of relations among genres given by the factor scores

The number of features and number of dimensions vary depending on which of Biber's publications you refer to, but Biber (1988: 73) defines a set of 67 linguistic features (and metrics for counting them) such as all adverbs, private verbs, and past tense. These 67 features were marked and counted in modified versions of the LOB corpus (written data) and the London-Lund Corpus (spoken data). Following the application of the factor analysis technique, Biber identifies seven dimensions and lists the linguistic features and their associated weights in each dimension. He then goes on to interpret six of the dimensions by examining those texts showing high co-

occurrence of those positive and negative features identified. For example, Dimension 1 is labelled 'Informational versus Involved Production' as (Biber 1988: 115):

> "The poles of this dimension represent discourse with interactional, affective, involved purposes, associated with strict real-time production and comprehension constraints, versus discourse with highly informational purposes, which is carefully crafted and highly edited."

The technique proposed by Biber has been widely cited in research articles, but also recently criticised by Lee (2000) as being linguistically and statistically unsound due to problems to do with the nature of language, the distributional properties of linguistic features and the non-representativeness of corpora. Lee attempted to replicate Biber's dimensions using the same statistical methodology on a four-million-word subset of the BNC, but found that variations in the configuration of the data (relative genre proportions), choice of variables, etc. could distinctly affect the results. This means that Biber's dimensions cannot be considered final. From the point of view of language teachers, Tribble (2000: 78) also points out the practical difficulties in actually using Biber's dimensions or applying them to new texts, due to the necessity of having the research corpus POS-tagged before any analysis can proceed. Tribble then continues his analysis using Scott's Keyword methodology described above. We can further criticise the multi-dimensional approach on grounds of inflexibility since the features/variables are chosen ahead of the research question. New features can be chosen, but then we have to repeat the complex analysis procedure, and in so doing may obtain sometimes radically different results, as Lee (2000) demonstrates. And as Altenberg (1989: 171) observes, another central problem with the technique is the interpretation of the dimensions produced by the factor analysis. He suspects that the difficulty of interpreting some of the factors "can be traced back to the linguistic features on which the study is based […] If the features are ill-defined, functionally heterogeneous, stylistically skewed, etc., this is likely to have an immediate effect on the results […] but there is little discussion of how [Biber's] choice of features may have determined the factors they produce". In contrast, the Matrix method described in this thesis does not rely on pre-selection of linguistics features for characterising texts.

Ball (1994) criticises Biber's study in terms of recall (as opposed to precision) of features, for example in relation to finding zero complementisers in subordinate clauses. Baayen (1997) expresses caution that the composition of the corpora might turn out to be the main determinant of the results obtained in the factor analysis, a finding which Lee (2000) in fact demonstrates in his research by varying the composition of his research corpus and thereby obtaining different results. The concerns of both Baayen and Lee are that Biber views the emerging dimensions as sample-independent and that the representative natures of the four language corpora being used are unclear. Baayen would prefer to see Biber's techniques used to explore correlational structure in language in a *given* corpus.

The approach presented in this thesis differs from Biber's because it is data-driven: the linguistic features worthy of microscopic analysis are suggested by the macroscopic study, rather than by intuition or previous research studies. Our approach is mainly aimed at the comparison of a small number of text corpora, usually two; one of which may be a normative corpus. Biber's approach considers frequency variation for pre-selected variables across a large number of texts and attempts to situate texts or text genres along several clines of variation.

Turning briefly to similar methods applied in areas other than corpus linguistics, we note an approach, similar to the keywords technique, described by Lebart et al (1998: 130 – 136) for extracting 'characteristic elements' (frequent words and phrases in a sub-corpus). They use the hypergeometric formula to assign probabilities to words and then sort on the probability values to expose characteristic (or significant) frequencies. Their application is that of identifying 'modal responses' to open-ended questions in surveys, i.e. those responses that characterise a group of answers since they contain the most characteristic words of the group.

In the field of information retrieval, Zhou (1999) attempts to extract topical words and phrases in a document summarisation and classification application. Zhou's Term Suggestion Toolkit uses different measures for selecting single words and multi-word terms. Multi-word terms are identified using the traditional mutual information measurement suggested by Ken Church. Single words are suggested based on the

interval between their occurrences, since Zhou's intuition was that topical words should appear more frequently and at approximately even intervals.

One final piece of work fits into this systematic frequency comparison category. In the area of authorship attribution, Burrows' (1992) method is to analyse the frequency patterns of "whatever" words occur most often in a given set of texts. He uses the Pearson product-moment method of correlation and principal components analysis to graphically visualise various authors alongside the text whose authorship is disputed.

2.8 Summary

In this chapter, we have surveyed the field of corpus linguistics and the traditional process model of 'question – build – annotate – retrieve – interpret' within which research questions are posed and investigated. We have seen that most studies decide in advance which linguistic features are to be examined, even when examining whole texts or varieties of language.

We have looked in detail at the practice of corpus annotation and seen the multiple levels at which it can be carried out. Our review then turned to frequency profiling since it is in this area that the thesis fits.

We have surveyed the various statistical techniques used to compare frequencies and frequency profiles across corpora. We have seen that keywords can be extracted statistically and manually. The advantages of the log-likelihood ratio over the other measures can be summarised as follows:

1. LL values are directly comparable
2. LL is not as expensive to compute as Fisher's Exact test, and gives similar results for large sample sizes
3. LL has been shown to be better 'in general' than the chi-squared test
4. the chi-squared statistic is an approximation to the LL for large samples
5. Fisher's Exact test requires estimation for large sample sizes

6. McNemar's chi-squared test is unsuitable since it would ignore frequencies of the same item in the two corpora
7. the selection of the Mann-Whitney test as an alternative is due to the burstiness of word occurrences
8. the Mann-Whitney test is suitable only for mid to high frequency words and for comparing corpora of the same size
9. NR values are not comparable
10. the Mosteller-Rourke adjustment for chi-squared is tied to a specific size of corpus
11. Berg's four measures are unsuitable when zero frequency entries occur

The Matrix method and tool allow corpus investigation by statistical comparison of frequency profiles at the lexical level and extend this to the word-class and semantic field levels. The Matrix method extends the whole text-focussed approach by informing the researcher as to specific linguistic features that should be studied further. This method is described in section 4.3 and evaluated in chapter 5. Although in our survey, LL has been shown to be better 'in general' than the chi-squared test, there remains a question over its specific use in the comparison of frequency profiles. In section 5.2 we examine more closely the differences between the two tests. In the next chapter, we survey the current software tools that implement the traditional corpus linguistic methodology and those that implement the data-driven approach described previously.

3. Software for Corpus Linguistics

"Only His Only Grammarian Can Only Say Only What Only He Only Means."[26]
Peter G. Neumann, ACM SIGSOFT Software Engineering Notes 9, 1, Jan 1984, pp. 6.

3.1 Introduction

By definition, corpus linguistics is a methodology that can be applied to the study of many different branches of linguistics such as morphology, syntax, and semantics. Software used during corpus linguistic study must therefore be fairly flexible unless it is to be used in one small area of research. Current software tools for corpus linguistics tend to be extremely limited in comparison with what we would like them to achieve in terms of the intelligent, comprehensive modelling of natural language. The tools needed for the creation and exploitation of corpora, in particular annotated corpora, can be classified into three major categories: corpus development (the input of information into a corpus), corpus editing (changing information in a corpus), and information extraction (the output of information from a corpus). These categories account for the major functions of corpus tools in the following way (adapted from McEnery and Rayson, 1997: 195):

1. Corpus development
    a. Text encoding
    b. Annotation
    c. Encoding of annotation
2. Corpus editing
    a. Correction
    b. Disambiguation

[26] See http://www.csl.sri.com/users/neumann/only.html

   c. Conversion of format and annotation
3. Information extraction (IE)
   a. Frequency analysis
   b. Concordancing
   c. Feedback into other tools (lexicons, grammars, etc.)
   d. Information retrieval

In the next two sections, we will summarise features of development and editing tools for corpora. We will include a more detailed examination of two example annotation tools. Finally, we will concentrate on the extraction tools and methods that are currently available for linguists to use in their research. These will be classified as either hypothesis-driven or data-driven, reflecting the distinction made in section 1.1.

3.2 Corpus development and analysis tools

Generally, in the initial stages of corpus collection and development, off-the-shelf hardware and software packages are used. To produce written corpora, these include document scanners, scanning software for optical character recognition, word processing software, and web-crawling software (if these are available for the language of interest). For spoken corpora, it would involve recording equipment such as Walkmans or Minidisc recorders, transcribing machines and word-processing software.

The encoding formats of the functions in the first and third corpus development categories (1a: text encoding and 1c: encoding of annotation) have been discussed previously in section 2.4. Their creation requires the existence of tools to aid input and validation of mark-up (typically SGML and more recently XML, see section 2.4) in accordance with some standardised system such as the CDIF[27] specification document for the BNC (Burnage and Dunlop 1993) or the EAGLES/CES guidelines (Ide 1996, 1998). Few tools are available for this task, and they tend to be developed

[27] Corpus Document Interchange Format; an application of SGML largely conforming to the TEI guidelines (see section 2.4)

on an ad-hoc per-project basis. One of the aims of the MULTEXT[28] project was to implement tools which embodied these standards. General-purpose SGML parsers, such as the public-domain *SGMLs* and *nsgmls*[29], are also available.

Annotation encoding software cannot be discussed in isolation from the 'storage architecture' question of how to represent, in an encoded corpus, the relation between the base text and the annotations. We have described in section 2.4 how the annotations are usually interspersed with the base text, as part of the same composite document. Two other arrangements are possible. One is to use the form of a relational database, where different fields of information represent the base text and different levels of annotation. This is particularly suitable for multilevel annotation, including, for example, POS tagging, syntactic annotation, and prosodic annotation. For precisely that purpose, it has been used by Knowles and Roach (Knowles 1995) in producing the MARSEC CD-ROM version of the Spoken English Corpus. Davies also used a database architecture to store the NEH Corpus del Español (Davies 2002). No special software is needed for this application, a general-purpose off-the-shelf database system being adequate. A second alternative is to hold the base text and the annotations in separate files, with links relating each part of one to the relevant part(s) of the other. This, called *stand-off annotation*, is the option favoured by Ide (1996) in the EAGLES guidelines for text representation, and by Thompson and McKelvie (1996) at Edinburgh, who have implemented this method in a toolkit (LT NSL). This method allows greater freedom than interspersing text and annotations: for example, it is possible to deal with the tags for merged words such as *du* (= *de* + *le* 'of the') in French without drawing artificial boundaries within a single orthographic word. The Thompson and McKelvie method is to make use of what is in effect a hyperlink architecture for cross-referring between the base text and different levels of annotation. In this way, overlapping hierarchies of annotation (which can be awkward to achieve in SGML) can be reasonably handled. There is need for more 'SGML application development toolkits' such as LT NSL, and particularly for the adaptation of such a toolkit in the direction of inputting and editing annotations.

---

[28] See the web site for MULTEXT at http://www.lpl.univ-aix.fr/projects/multext/

[29] See the website for nsgmls at http://www.jclark.com/sp/nsgmls.htm maintained by the software's author James Clark.

We now come to category 1b (corpus annotation software). We can distinguish between predominantly automatic and predominantly manual annotation procedures. Using examples from tools developed at UCREL, the former is represented by the tagger CLAWS (Garside and Smith, 1997) and the latter by the editor Xanadu (Garside and Rayson, 1997). The fact that tools for manual annotation input are called 'editors' shows that the boundary between annotated corpus development and annotated corpus editing in categories 1 and 2 is not a watertight one. We will also note the interaction between categories 1 and 3: automatic annotators (such as taggers and parsers) are, in effect, linguistic analysis tools, which therefore require for their operation complex linguistic information resources such as lexicons and grammars. These resources are themselves primary beneficiaries of the category 3 information extraction stage, and hence there may be an iterative cycle from category 3 to category 1, as follows: extraction of linguistic information in category 3 tools potentially enhances the input of information in category 1 tools. It is evident, already, that splitting functions into the three categories as above is somewhat simplistic.

In the next two sections, we focus on two corpus annotation tools, CLAWS and USAS in order to exemplify corpus tools in category 1b. Later, these tools will be used in the Matrix worked example in section 4.4 and in the case studies in sections 5.3.2 and 5.3.3.

3.2.1 CLAWS part-of-speech tagger

We have already described the origins of the CLAWS tagger in section 2.5; here we will describe the system in more detail. Both morphological and grammatical analysis are carried out within the CLAWS program, which has been developed continually since the early 1980s. CLAWS is a hybrid tagger using a statistical Hidden Markov Model (HMM) technique (Jelinek, 1990) and a rule-based component. In a fully automatic procedure, CLAWS assigns POS tags with 97-98% accuracy. Other POS taggers using various tagging methods quote similar success rates, such as the rule-based taggers Brill (Brill, 1992) and ENGCG (Karlsson et al, 1995), memory-based learning taggers (Daelemans et al, 1998) and the statistical Xerox tagger (Cutting et

al, 1992). Voutilainen (1999) surveys the history of the different approaches to wordclass tagging. The advantage of CLAWS is that it is a robust tool, having been trained and tested over a large amount of data, most recently the one hundred million words of the British National Corpus (Leech et al, 1994b). Figure 3.1 shows an example sentence tagged with the CLAWS4 C7 tagset; for the full tagset see Appendix III of Garside, Leech and McEnery, 1997[30]. The first letter of each tag shows the major word class: *A* for article, *D* for determiner, *I* for preposition, *J* for adjective, *M* for number, *N* for noun, *P* for pronoun, *R* for adverb, and *V* for verb. *TO* is a special tag for the infinitive marker, XX for 'not', and punctuation is tagged as itself.

The_AT **lovers**_NN2 **,**_, **whose**_DDQGE **chief**_JJ **scene**_NN1
was_VBDZ **cut**_VVN **at**_II **the**_AT **last**_MD **moment**_NN1 **,**_,
had_VHD **comparatively**_RR **little**_DA1 **to**_TO **sing**_VVI **.**_.

Figure 3.1 An example of CLAWS4 POS tagging

The functionality of CLAWS was extended for use in the British National Corpus project. Currently named CLAWS4, the system is described as having five major stages:

a) segmentation of text into word and sentence units
b) initial (non-contextual) part-of-speech assignment (using a lexicon, word-ending list, and various sets of rules for tagging unknown items)
c) rule-driven contextual part-of-speech assignment
d) probabilistic tag disambiguation, using a Markov process on bi-gram tag transition data, followed by a second pass of stage c
e) output in intermediate format (vertical, one-word-per-line, for manual post-editing) or final format (horizontal and encoded in SGML)

The pre-tagging stage (a) is not trivial, since, in any large and varied corpus, the tagger is required to deal with unusual text structures, unusual typographic features

[30] The tagset is also included as an appendix to this thesis.

(e.g. non-roman alphabetic characters, mathematical symbols), and features of conversation transcriptions: e.g. false starts, incomplete words and utterances, unusual expletives, unplanned repetitions, and (sometimes multiple) overlapping speech. Further tokenisation problems are described in section 4.2, and Garside (1995) discusses in more detail the modifications which were made to CLAWS in order to deal with spoken data.

CLAWS is best known for its probabilistic approach to tagging, and this occurs in stages b and d. The Hidden Markov Model is illustrated simply in Leech and Fligelstone (1992: 131). However, as Garside and Smith (1997) note, it should be considered as a hybrid tagger, since the rule-based component in stage c is of equal importance to the accuracy of the tagging. CLAWS now includes a two-pass application of these *idiomlist* entries. It is possible, on the first pass, to specify an ambiguous output of an idiom assignment, so that this can then be input to the probabilistic disambiguation process (d). On the second pass, however, after probabilistic disambiguation, the idiom entry is deterministic in both its input and output conditions, replacing one or more tags by others. In effect, this last kind of idiom application can be used to correct a tagging error arising from earlier procedures. The modular development of the idiom tagging is described in more detail in Fligelstone et al (1996).

3.2.2 The USAS semantic tagger

The UCREL semantic analysis system (USAS) accepts as input text which has been tagged for parts of speech using the CLAWS4 POS tagger. The tagged text is fed into the main semantic analysis program (SEMTAG), which assigns semantic tags representing the general sense field of words from a lexicon of single words and a list of multi-word combinations, called templates (e.g. ‘as a rule’). These are updated as new texts are analysed (Rayson and Wilson, 1996). Currently, the lexicon contains nearly 37,000 words and the template list contains over 16,000 multi-word units. Items not contained in the lexicon or template list are assigned a special tag, *Z99*. Figure 3.2 is an example of semantic word tagging, taken from a library system requirements definition document.

It_Z8 **is**_Z5 **anticipated**_X2.6+ **that**_Z5 **the**_Z5 **system**_X4.2
will_T1.1.3 **be**_Z5 **administered**_A9- **by**_Z5 **the**_Z5 **Library**_Q4.1/H1
,_PUNC **but**_Z5 **this**_Z8 **will**_T1.1.3 **not**_Z6 **always**_N6+++
be_the_case_A5.2+[i9.3 ._PUNC

Figure 3.2 An example of lexical semantic tagging

The semantic tags are composed of:

1. an upper case letter indicating general discourse field.
2. a digit indicating a first subdivision of the field.
3. (optionally) a decimal point followed by a further digit to indicate a finer subdivision.
4. (optionally) one or more 'pluses' or 'minuses' to indicate a positive or negative position on a semantic scale.
5. (optionally) a slash followed by a second tag to indicate clear double membership of categories.
6. (optionally) a left square bracket followed by 'i' to indicate a semantic template (multi-word unit).

For example, *A5.2+* indicates a word in the category 'general and abstract words' (A), the subcategory 'evaluation' (A5), the sub-subcategory 'true and false' (A5.2), and 'true' as opposed to 'false' (A5.2+). Likewise, *Q4.1/H1* belongs to the category 'communication' (Q), subcategory 'the media' (Q4), and refers to 'books' (Q4.1), as well as 'kinds of houses and buildings' (H1)[31].

The semantic annotation is designed to apply to open-class or 'content' words. Words belonging to closed classes (such as prepositions, conjunctions, and pronouns), as well as proper nouns, are marked by a tag with an initial *Z*.

[31] A full tagset for the USAS tagger can be found online at http://www.comp.lancs.ac.uk/ucrel/usas/ and as an appendix to this thesis. Wilson (1997) describes the background to conceptual tagging.

As in the case of grammatical tagging, the task subdivides broadly into two phases: Phase I (Tag assignment): Attaching a set of potential semantic tags to each lexical unit and Phase II (Tag disambiguation): Selecting the contextually appropriate semantic tag from the set provided by Phase I. SEMTAG makes use of seven major techniques or sources of information in phase II:

1. *POS tag.* Some senses can be eliminated by prior POS tagging. For example, consider the word *spring*. There is a lexicon entry for *spring* which specifies firstly the possibility of a noun tag or a verb tag, and secondly the possibility that the noun may have the 'coil' sense or the 'season' sense. In this sample lexicon entry, the POS tagger, by choosing the noun tag, obviously eliminates one of the senses ('to jump'). Hence the semantic tagger's task is simplified to choosing between the 'season' and the 'coil':

word form	**POS tag**	**semantic tag**
spring	noun	[season sense] [coil sense]
spring	verb	[jump sense]

2. *General likelihood ranking for single-word and template tags.* In the lexicon and template list senses are ranked in terms of frequency, even though at present such ranking is derived from limited or unverified sources such as frequency-based dictionaries, past tagging experience and intuition. For example, *green* referring to 'colour' is generally more frequent than *green* meaning 'inexperienced'.
3. *Overlapping template resolution.* Normally, semantic multi-word units take priority over single word tagging, but in some cases a set of templates will produce overlapping candidate taggings for the same set of words. A set of heuristics is applied to enable the most likely template to be treated as the preferred one for tag assignment. The heuristics take account of length and span of the idioms and how much of a template is matched in each case.
4. *Domain of discourse.* Knowledge of the current domain or topic of discourse is used to alter rank ordering of semantic tags in the lexicon and template list for a particular domain. Consider the adjective *battered* to which three candidate tags can be assigned: 'Violence' (e.g. *battered wife*), 'Judgement of Appearance' (e.g. *battered car*), and 'Food' (e.g. *battered cod*). If the topic of conversation was known to be food, then we automatically raise the likelihood of the 'Food' semantic tag, at the expense of the other two tags.

5. *Text-based disambiguation.* It has been claimed (by Gale et al, 1992), on the basis of corpus analysis, that to a very large extent a word keeps the same meaning throughout a text. For example, if a text on one occasion uses *bank* in the sense of 'side of a river', all other occurrences of *bank* are likely to have that same sense. In SEMTAG, this method works together with step 4.
6. *Contextual rules.* The template mechanism is also used in identifying regular contexts in which a word is constrained to occur in a particular sense. Consider the meaning of the noun *account*: if it occurs in a sequence such as *NP's account of NP* it almost certainly means 'narrative explanation', whereas if it occurs in a financial context, in such collocations as *savings account* or *the balance of ... account* it almost certainly has the meaning of a 'bank account'.
7. *Local probabilistic disambiguation*. It is generally supposed that the correct semantic tag for a given word is substantially determined by the local surrounding context. To return to the example of *account*: if this noun occurs in the company of words such as *financial*, *bank*, *overdrawn*, *money*, there is little doubt that the financial meaning is the correct one. However, we could identify the surrounding context not only in terms of (a) the words themselves, but in terms of (b) their grammatical tags, (c) their semantic tags, or (d) some combination of (a) - (c). This method is still under development in SEMTAG and future work includes experimentation, using a training corpus and a test corpus, to determine what weight to give each of these contextual factors in selecting the correct semantic tag for a given word or word class. Other factors which need to be determined are discussed in Garside and Rayson (1997).

After automatic tag assignment has been carried out, manual post-editing can take place, if desired, to ensure that each word and idiom carries the correct semantic classification. An additional program using template analysis techniques (see section 3.3) can then mark important lexical relations (e.g. negation, modifier + adjective, and adjective + noun combinations).

3.3 Corpus editing tools

Annotated corpus editing can refer to any procedure of changing the linguistic annotations in a corpus. Categories 2a – 2c suggests three reasons why such annotations could need to be changed. The first is to correct errors, for instance errors resulting from the use of automatic annotation tools such as a probabilistic tagger. The second is to eliminate ambiguities, such as the ambiguities left in the text by automatic annotation tools which allow ambiguous output (e.g. ENGCG, see Karlsson et al 1995, or the variant of CLAWS4 which outputs portmanteau tags, see Smith 1997). The third is to convert one set of annotations to another set for which there is a need: for example, it might be decided to adapt the grammatical tags of a corpus from one tagset to another which is more amenable to other users' requirements. The term editor may also apply to a tool for manually adding annotations to a corpus, such as the Xanadu tool (Garside and Rayson, 1997). There is a distinction, here, as with annotation input tools, between primarily automatic and primarily manual processing. Fligelstone et al (1997) describes in some detail an 'automatic editing' tool (the Template Tagger) which has diverse functions. As a general corpus editing tool, the overall purpose of the Template Tagger is simply to apply rules which change one set of annotations into another: such rules could either add, convert, or subtract annotations from the corpus. Nevertheless, our main interest here is in editors in the familiar sense of tools which allow the user to change the form of a text stored on computer. Up to a point, it is possible to rely on general-purpose text editing software such as a screen editor (for example Emacs or even the archaic Vi). But if one is trying to correct a large annotated text or an annotated corpus of any size, the need for a dedicated editor, which will aim to eliminate unnecessary human labour and error, soon becomes imperative. Moreover, much of the attention of those developing the editor will be directed to making a good graphical interface, offering the human annotator trouble-free and efficient interaction with the annotated text.

3.3.1 Manual annotation editing

In this section we discuss the facilities which would be required for a reasonable tag editor, whether the tags be of the syntactic or semantic varieties. UCREL has

constructed a variety of editors which implement this list of requirements to a greater or lesser extent, perhaps the most complete being a program called Xanthippe which in one of its incarnations has been used to edit the syntactic tags of parts of the BNC and in another has been used for the syntactico-semantic tags of the ATR project (Bateman et al, 1997). Van Halteren and Oostdijk (1993) describe similar requirements for the software involved in manual selection of word-class tags and parse trees in the Nijmegen TOSCA system.

We can generally assume that the input text has been through some form of automatic assignment of tags, and there will usually be a tag indicated as the one preferred by this automatic process, together with a list for each word of the tags rejected in the context in favour of the preferred tag. It may be that the preferred tag (and the rejected tags) are fully specified, or it may be that the automatic process is capable of assigning only an incomplete tag, or a tag to only a certain level of detail - for example, the general syntactic function might be fully specified while the detailed syntactic function and the semantic function can be added only with human intervention.

The text will typically include mark-up to indicate at least the main subdivisions of the text. This may be in some form of SGML. Since a wide variety of tagsets may be in use, a tag editor needs to be written as far as possible in a tagset-independent way. It is generally possible to have a tag editor read in the tagset to be used, but it may be necessary, because of idiosyncrasies in the tagset, to have a small amount of special-purpose code to deal with them.

The user interface, of course, needs careful consideration. UCREL has tended to work with a small team of highly-trained corpus analysts who prefer an interface which minimizes the number of keystrokes and screen redrawing for the more common functions, even if these lead to different procedures for what are conceptually similar tasks. With a larger number of less highly-trained analysts working in a less intensive way, it is possible that the user interface design criteria would have been rather different, with more commonality of procedure and more prompting from the editor.

Xanthippe's screen format (see Figure 3.3) is a series of parallel vertical columns containing (for a stretch of text) the words in one column, the preferred tags in another column (or perhaps sub-divided into a column containing that part of the tags fully specified by the automatic tagging system, and a second column containing that part only partly specified by the automatic tagging system), and further columns containing the rejected tags (or perhaps only the fully specified parts of these, to save screen space). There will often be further information associated with the words; for example, a reference code for the word; information about the automatic tagging process, including an indication of where the analyst's attention should be drawn to, places at which the process is likely to be at fault; information showing how multiword units are linked together; and so on. Although one might expect that a horizontal display would best mimic the normal process of reading, Xanthippe displays the words of the text down the screen. This is perhaps an area where the design of the user interface is influenced by the background of UCREL's analysts, since this screen format in fact mimics the hardcopy listings that most of them are familiar with from earlier projects.

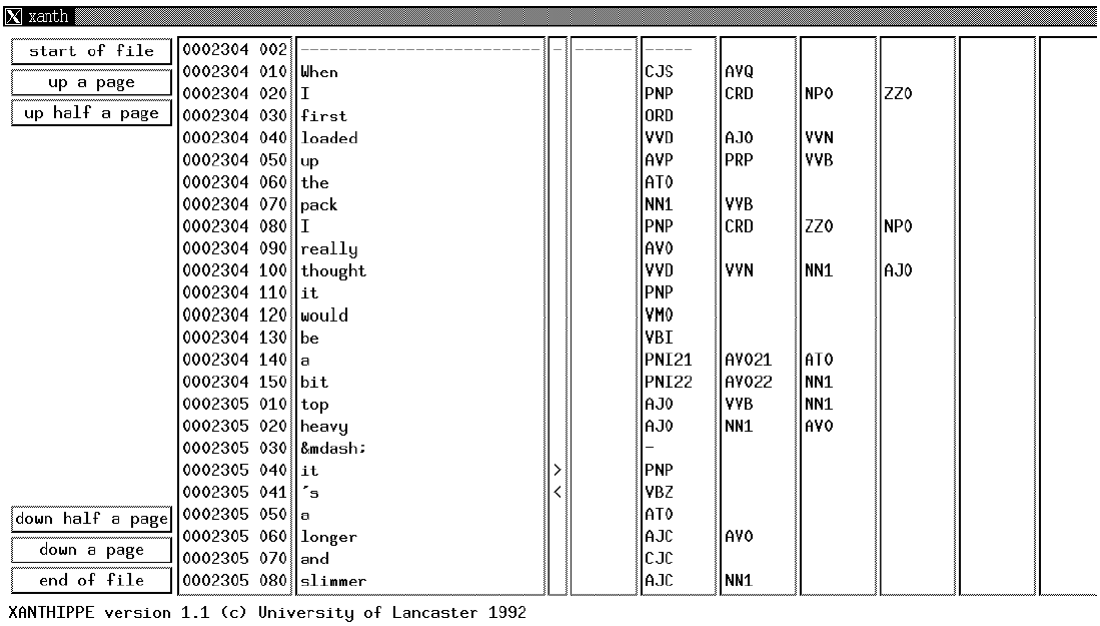


Figure 3.3 Screenshot of the main Xanthippe window

A problem with this representation is that only a small number of words of a text can be displayed at a time (perhaps twenty or twenty-five, given the size of display screen

and the choice of a font size large enough to avoid eye strain). Xanthippe therefore allows a subsidiary window to be placed alongside which shows a larger stretch of text, including the words displayed in the main tagging window, but without the annotations. As the analyst scrolls or otherwise moves through the text in the main window, this subsidiary window synchronises with the main window.

Typical user functions for a tag editor would include the following:

a) The promotion of one of the rejected tags so that it becomes the preferred tag, by far the most common type of tag correction for a reasonably competent automatic tagging process. For situations where the correct tag is not among the rejected tags a panel of tags can be displayed for user selection of the appropriate one.
b) Correction of the original words of the text. In some cases there is a requirement for the insertion into the text of a note specifying the original word, the corrected word, and an optional comment by the analyst. More generally there may be a need for the analyst to be able to insert some form of comment about a local aspect of the tag correction process. The editor generally inserts this comment into the text surrounded by a suitable SGML mark-up sequence, and indicating the identity of the analyst making the comment.
c) The insertion or deletion of markings for a multiword unit.
d) If Xanthippe is being used for correcting syntactic tags assigned by an automatic process, it is likely that occasionally an automatically assigned sentence break will need to be suppressed or a new one inserted.
e) It is useful for the analyst to be able to search the text, from the present text position or from the beginning of the file, either once or repeatedly, for words, parts of words, word sequences, word reference codes, or even tags or tag sequences, whether fully specified or not. A very useful extension to this is *global editing*. If the analyst detects a persistent pattern of error in a file of text, it is useful for them to be able to specify a pattern of words or tags to search for and a preferred tag or tag sequence to be applied throughout the file, with or without user confirmation of each matching instance. Xanthippe implements a fairly restricted form of pattern matching for global editing; a fully developed

version of this would of course be equivalent to an interactive form of the Template Tagger.

f) Finally we can log the process of manual annotation. We can simply write to a log file a list of all the tag corrections or other revisions made by the analyst, together with suitable global information as to the name of the analyst, the file being processed and the date. This is useful for extracting patterns of persistent error by the preceding automatic tagging process.

3.3.2 Automatic annotation editing

We have already made reference to the use of the Template Tagger as an automatic annotation editor, and here this function may be briefly illustrated, using a grammatically tagged corpus as an example. A tagset needs sometimes to be adapted. It may be that a tagset devised (partly) for ease of automatic tagging will later prove ill-adapted to a user's needs. For example, many of the UCREL tagsets, including C5 and C7 (see Appendix III of Garside, Leech and McEnery, 1997), do not represent auxiliary verbs as a separate category in English: something that many users may find desirable. For this purpose it is necessary to enrich the tagset by introducing new tags (for example, instead of VBZ for *is*, it will be necessary to devise a new tag, say VBAZ in addition to VBZ, for *is* as an auxiliary and *is* as a main verb respectively). To make this and other similar changes, a set of Template Tagger rules are needed enacting such changes as

> If a tag VB* is followed by a tag V*, with or without the intervention of other tags XX or AV0 (i.e. the word *not* or any adverb), then change VB* into VAB*

In fact, a comparatively small number of such rules will make the necessary change, with few exceptions (see further in Fligelstone et al, 1996). This is a relatively straightforward global edit, whereas to make other changes, say, adapting the tagging of *-ing* words in order to conform to one set of guidelines rather than another, the process is likely to be more complex. There are other reasons, apart from the needs of a specific user, why it might be desirable to adapt annotation systems. One is to

convert annotations into a form which is conformant with an externally devised standard, such as the EAGLES standard for morphosyntactic or syntactic annotation (Kahrel et al 1997).

3.4 Retrieval and extraction of linguistic information

In this section, we will survey off-the-shelf search and retrieval software for extracting linguistic information from annotated corpora. In the computer science domain, information retrieval and extraction are seen as complementary (Gaizauskas and Robertson, 1997). Other avenues are open for the researcher who can write their own software using Java or who use tools such as Perl to process their own text. These kinds of approaches are described in Burnard[32] (1992), Barnbrook (1996), and Mason (2000).

Search and retrieval software is more familiar to the general corpus user than any other: anyone who wishes to make use of a corpus is inevitably going to look for means to extract linguistic information from it. The 'naive user' is likely first to encounter a corpus through a *concordancing* facility; that is, a program for listing (a subset of) the instances of a given linguistic phenomenon (typically a word) in the corpus, together with the immediately preceding and following context. The technique predates computers by some years. In Cowden-Clarke (1881), and other volumes like it, a *concordance* has the meaning of listing a short contextual phrase for every occurrence of every word in a text corpus (alongside the location of each phrase), with the exception of a small set of words considered insignificant and occurring frequently such as *be*, *do*, *have* and some interjections. The manual effort involved in these early publications was immense, in this case taking sixteen years. However, such a task is made vastly more efficient by the use of computer software. Here, the technique is also called KWIC (key word in context), a term which, according to Paice (1977: 54) was coined by H. P. Luhn in 1960. Words need not be key in the

[32] Interestingly, Burnard (1992: 20) also predicts that "we may expect to see hypertextual interfaces replacing more conventional ways of interacting with electronic texts over the next few years". The first web browser (NCSA Mosaic) appeared in 1993.

statistical sense of section 2.7.2 but are usually the result of a search operation. Associated with the concordancer will often be other facilities: providing frequency lists of word types, listing collocations based on mutual information or other measures, and furnishing information about subdivisions of the corpus, together with the incidences of linguistic phenomena in these. Examples of frequency lists and concordances are shown in section 4.4. Many packages of this kind are available, some more advanced than others, and each tending to have its own special features. For a discussion of some of these packages, see Hofland (1991), Kirk (1994), or Hockey (2001) and for more general surveys, see Lancashire (1991) and Hughes and Lee (1994).

However, our main interest in search and retrieval packages in this chapter must focus on their treatment of annotations. Using a package with annotation awareness, it will be possible to search on annotations up to a point, but the output (in the form, say, of a concordance) is likely to be littered with annotation, with the result that no normal human user would find it easy to interpret. Hence, one useful facility is for the software to recognise annotations, and optionally to mask them from the screen interface. Examples of annotation-aware corpus exploitation tools are SARA, xkwic and ICECUP. WordSmith also offers a tag-aware facility.

In Table 3.1, we compare the requirements and capabilities of nine of the most widely cited retrieval software packages in corpus linguistics and related research. The packages we consider here are:

1. WordCruncher (Jones, 1987): WordCruncher was designed to be a sophisticated index, search and retrieval program for MS-DOS-based computers before the era of large memory and the Windows interface. It was developed at Brigham Young University for large textual corpora like ‘The Collected Works of Shakespeare’ and the Bible and the ‘Book of Mormon’. Two tools form WordCruncher: WCIndex indexes texts to be studied; WCView analyses and retrieves texts prepared with WCIndex. WordCruncher for Windows was released in 1992. Following business changes at

WordCruncher Publishing Technologies, WordCruncher for Windows has recently reappeared as DocumentExplorer[33].

2. xkwic[34] (Christ, 1994): Developed by Oliver Christ at IMS Stuttgart. This tool was part of the IMS Corpus Toolbox, version 2.2. It features a graphical user interface in X-windows Motif to access a corpus query processor (CQP). Versions of CQP linked to a web-based interface are installed at some sites as mentioned in section 3.6.
3. ICECUP (Quinn 1993 and Nelson et al 2002): The 'ICE Corpus Utility Program' is a corpus exploration program designed for parsed corpora such as ICE-GB. The ICE corpora are encoded using SGML-style tags. The main features of ICECUP are: CorpusMap, which provides an overview of the corpus, and Fuzzy Tree Fragments (FTF), which provides a way to perform grammatical queries on the corpus. Version 3.1 includes advanced features including a lexicon, grammaticon, generation of contingency tables and extensions to FTFs.
4. SARA (Aston and Burnard, 1998): The 'SGML-Aware Retrieval Application' is a client-server tool, although both client (user interface) and server (database) can now be run on one PC running Windows. SARA was developed specifically to access the BNC, but more recently, additional software was made available to index any TEI-encoded corpus for use with SARA.
5. WordSmith Tools (Scott 1996-99): The package is described as an advanced set of tools providing "an integrated suite of programs for looking at how words behave in texts". Version 3 of the package provides 6 tools: Wordlist, Concord, Keywords, Splitter, Text Converter, and Viewer. WordSmith and xkwic were compared in a review by Lee and Rayson (2000).
6. CorpusBench[35] (1993): Used by lexicographers at Longman Dictionaries. It was produced for the IBM OS/2 operating system.
7. TACT[36] (Lancashire et al, 1996): 'Text Analysis Computing Tools' is a text-analysis and retrieval system for MS-DOS that permits searches on text

[33] Available from Hamilton-Locke Inc., USA: http://www.hamilton-locke.com/DocExplorer/Index.html

[34] See website at http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

[35] See CorpusBench User's Manual. Version 1.0: December 1993. TEXTware A/S, Copenhagen

databases in European languages. Development began under the IBM-University of Toronto Cooperative in the Humanities during 1986-89. We refer to version 2.1 that consists of 15 separate DOS programs that will work within Windows 3.11, 95/98 but not with Windows NT. TACT has now been connected to the web, as mentioned in section 3.6, in an experimental service called TACTweb.

8. BNCweb (Lehmann et al, 2000): BNCweb is a web-based client program for searching and retrieving lexical, grammatical and textual data from the BNC. It relies on the BNC server program, SARA. Thanks to its integration of MySQL, a very fast SQL-database server, BNCweb is able to extend the functionality of SARA. It thus offers a whole range of additional features for corpus analysis.
9. CLAN (MacWhinney, 1995): The CHILDES (Child Language Data Exchange System) project was conceived of in 1981 to establish a system for sharing child language transcript data. It consists of a transcription and coding format (CHAT), a package of analysis programs (CLAN), and a multi-language child language database.

Other notable software not included in our table is:

1. MicroConcord (Scott and Johns, 1993): This is a concordancer, operating on PCs running DOS. The number of concordance lines is limited to around 1,500, and the user cannot save a concordance except as a text file. It is very useful for a quick analysis, but has largely been superseded by WordSmith Tools. Initially sold commercially by OUP, it is now available free from Mike Scott's website.
2. OCP (Hockey and Martin, 1987): The Oxford Concordance Program was a general purpose batch tool for generating concordances, word lists, and indexes from texts in any language or alphabet. OCP operated on an ASCII file of the text and up to 8 characters may be defined to represent one letter. OCP could handle a number of mark-up systems though frequently the text

[36] See website at http://www.chass.utoronto.ca/cch/tact.html

was tagged using the COCOA system (Word Count and Concordance Generator for Atlas). COCOA has been largely superseded by SGML.

3. MonoConc: A Windows-based PC concordance tool with a range of powerful features (Advanced Search: Full Regular Expression search, Part-of-Speech Tag Search, Collocations) published by Athelstan[37]. MonoConc Pro and WordSmith Tools were comparatively reviewed by Reppen (2001).
4. Concordance[38] (1999): Produced by R. J. C. Watt of the University of Dundee. The Windows software was designed for creating full concordances (in the sense of every occurrence of every word) to be hosted on the web for teaching English literature, but is capable of many of the normal corpus retrieval features.
5. WordPilot[39] (Milton 1999): Designed by John Milton of the Language Centre at the Hong Kong University of Science and Technology. It is intended to be used by language learners of English from within a word-processor, and the program makes available word lists and concordances. The SpeechPilot module allows learners to hear any examples they chose to be read to them by the computer.
6. Cosmas[40]: The 'Corpus Storage, Maintenance and Access System' is software with a web interface used at the Institut für Deutsche Sprache, Mannheim to access extremely large corpora, greater than 1,000 million words in size.

There are also a large number of other software tools designed for qualitative data analysis (such as NUD*IST, and The Ethnograph), and content analysis (e.g. General Inquirer and TextQuest). These tools will not be considered here[41], although in fact Matrix has already been applied in the content analysis of cancer-care doctor-patient interactions (Thomas and Wilson, 1996).

---

[37] See the website at http://www.athel.com/

[38] See the website at http://www.rjcw.freeserve.co.uk/

[39] See the website at http://www.compulang.com/

[40] See the website at http://corpora.ids-mannheim.de/cosmas/

[41] However, see http://www.textanalysis.info/ and Popping (1997) for more information

Table 3.1 Comparison of software capabilities for retrieval tools

	Wordcruncher	xkwic	ICECUP	SARA	WordSmith	Corpus Bench	TACT	BNCweb	CLAN
Commercial or share/free-ware	com	free	com	com[42]	com	com	free	com[43]	free
Operating System[44]	DOS W	UNIX	W	W[45]	W	OS/2	DOS	UNIX[46]	UNIX Mac DOS
Text encoding	Plain	SGML[47] style	ICE	SGML TEI	SGML style	SGML	SGML style COCOA	SGML TEI	CHAT
Pre-indexed	y	y	y ICE only	y BNC only	n	y	y	y BNC only	n
Annotation support	n	y	y POS Syntax	y POS only	y[48]	y POS only	y POS concept	y POS only	n
Frequency lists	y	n[49]	y[50]	n	y	y	y	y	y
Comparison of frequency lists	n	n	n	n	y	n	y[51]	n	n
Concordances	y	y	y	y	y	y	y	y	y

[42] Initially available only with the pre-indexed BNC

[43] Shipping and handling fee only

[44] In the operating system row of this table 'W' is short for Windows operating system (Windows3.X, 95, 98, NT, 2000, or XP). It indicates that the system has a graphical user interface on a PC, unlike DOS which uses a character-based interface. UNIX represents compatibility with a system like a Sun Microsystems Workstation (running SunOS or Solaris) for example.

[45] SARA client is Windows-based, the server can be UNIX/Linux or Windows-based.

[46] Server requires UNIX, but client tool is web-based so requires no particular operating system

[47] Partial support for structural annotation

[48] Annotation can be hidden from WordList

[49] Xkwic can show frequency distributions of the node word or marked collocates from a concordance

[50] Partial frequency lists via the lexicon feature in version 3.1

[51] Comparison of two lists in terms of which lines (words) are shared and which are not

Concordance sorting	n	y	y	y	y	y	n	y	n
Collocations	y	y	n	y	y	y[52]	y	y	y
Sub-corpora	n	y	y	y	y	y	n	y	y
Lemmatisation	n	n	n	y	y[53]	y[54]	y	y	n[55]

The features of the software tools recorded in the table are as follows:

a) *Commercial or share/free-ware*: Some of these software systems are sold commercially and we would expect greater flexibility and support from them. Some are aimed at corpus research rather than pedagogical applications.

b) *Operating system*: Operating system requirements are an important consideration since a large research group may have access to and systems support for UNIX machines, whereas students of linguistics departments may prefer to (or have to) use PCs running Microsoft Windows. The choice of operating system may be influenced by the need for a graphical interface, allowing mouse-driven queries.

c) *Text encoding*: The text encoding values show the different formats for recording header information (such as authorship attributes or speaker description and text type) and within-text mark-up (such as utterance delimiters, headlines, or paralinguistic features), as previously described in section 2.4.

d) *Pre-indexed*: Pre-indexing of corpora is necessary for some software. This means that a user or administrator must run an index compilation process prior to the software being capable of retrieving information from the corpus under study. The usual method is to use an inverted index of the corpus (Oakes, 1998:150). The index is usually an alphabetically sorted list of the word types in the text with some indication of where each word token occurs in the text. The benefit of pre-indexing is that retrieval times are much improved due to the nature of the inverted-index files when compared to a sequential search, especially for large

[52] Using statistical measures MI and t-score

[53] Manual grouping of words in WordList function

[54] Needs inflection dictionary

[55] Can be partly achieved using wildcard search, e.g. “kick*” matches kick, kicked, kicks and kicking

corpora such as the BNC. However, the cost of this pre-processing in terms of time taken, storage of large index files, and obtaining the necessary skills to complete the task may be off-putting for beginners and intermediate users. Some software may impose a maximum size for corpora, or practical limits beyond which it is not usable in terms of speed or memory requirements. Some software tools provide access to only pre-indexed corpora but no separate indexing tool and are therefore tied to use with specific corpora, e.g. BNCweb.

e) *Annotation support*: For annotated corpora, we would like the software to be intelligent enough to recognise the annotation and present it to us alongside the words displayed, or hide it if desired. However, many tools are unable to analyse annotation and some treat POS tags, for example, as part of a word or separate words in their own right. This often leaves the analyst in the position of having to maintain two versions of a corpus, one as raw text with the annotation stripped out and one with the tags. The ability to display multiple levels of annotations is implemented in xkwic. However, normal concordance lines display only the word level. An extended concordance view, which shows only one context line at a time, can show any of the levels of annotation recorded in the corpus alongside each word of the concordance line. Leech and Smith (1999: 30) call this feature *annotation awareness* in software.

f) *Frequency lists*: Frequency lists are standard components of 7 systems in our list. A combination of simple UNIX tools (awk, sort, uniq or perl) can be used to provide a rudimentary word frequency list for a text file (Barnbrook, 1996: 188). However, to be of any use to the user, the software should specify how word boundaries are defined, what characters signal punctuation or should be allowed within words, how the annotation in the corpus is represented and whether to count capitalised words together with their lower-case equivalents. If possible the user should be allowed to alter these settings. SARA provides word frequencies only one word at a time prior to the production of a concordance. WordSmith allows annotation attached to each word to be filtered out from the word frequency list by definition of a regular expression matching the annotation code format. None of the tools listed here allow direct frequency lists of the annotation to be produced, other than by treating it as part of a word list[56]. Frequency lists

[56] Xkwic can produce this, but it requires the prior generation of a concordance for every word in a text

showing combinations of the words and associated annotation would be extremely useful. For example, for a semantically annotated corpus a frequency list showing word and word sense tags would differentiate between homographs and allow production of a word sense lexicon.

g) *Comparison of frequency lists*: Once frequency lists are produced we may want to compare them with lists derived from other corpora or sub-corpora in an automated manner. This gives us insights into the relative over- or underuse of words (or annotation) between two or more corpora. (For further discussion of this see section 2.7). WordSmith is particularly good at this task with its KeyWords feature. It also takes the process a stage further to produce KeyKeyWords which have significantly different frequencies across a large number of files or sections in the research corpus.

h) *Concordances*: Counting occurrences of words or tags is usually only the first stage of analysis. Once we have determined which words (or groups of words) merit further investigation, we will probably want to see them in context. This allows us to examine their syntactic or semantic behaviour. Concordancing software should allow the user to set the number of words shown to the left and right of the word being studied. For example, the KWIC (Key Word In Context) and KWAL (Key Word And Line) programs of CHILDES allow one line or more than one line of context respectively. We should be able to select a key word to study on the basis of regular expressions so that inflected forms can be displayed together. For example, *kick** should display concordance lines for *kick*, *kicked*, *kicking* and *kicks*, if we adopt the asterisk as a wildcard that matches any sequence of non-space characters. However, *kick** would find *kickboxer*, *kickback*, and *kicker* if they occurred in the corpus. This technique also does not work for irregular forms of words, e.g. *be*, *am*, *are*, *been*, *being*, *was*, *is*, *were* (however, see the lemmatisation feature below). With an annotated corpus, we ought to be able to select key words on the basis of a tag or category assigned to it. This facility would allow us to refine a search of the word *table* to contain only nominal uses by placing a restriction on the POS level. We should also be able to display (or not to display) annotation within the concordance lines. The corpus encoding information should be traceable so that file headers can be displayed alongside each concordance line and within-text markers can be made visible. This would

allow us to view information about a particular speaker in a spoken corpus or bibliographical details of written texts.

i) *Concordance sorting*: Concordance lines are usually presented in order of occurrence through the original text or corpus. In order to better detect contextual patterns we might want to sort the lines to aid our investigation. If a wildcard search was performed we would probably want to sort on the keyword first so that inflectional variants are listed consecutively. Following that, we may wish to sort on the previous and/or following context in order to see syntactic patterns emerging. We might also want our contextual sort to ignore 'noise' words, for example to focus on content words and ignore closed-class words.

j) *Collocations*: Collocation analysis is a method by which we can to some extent automate the search for contextual patterns of words made possible using concordances. The software might provide statistical measures to determine the strength of association between two or more words within a short space of each other in a text.

k) *Sub-corpora*: This capability represents the ability of the software to select portions of the corpus under study for further analysis as directed by the user. Of course, this can be done manually by editing out portions of a text, or omitting certain files from the analysis in a multi-file corpus. The ability referred to here is the one in which the software uses encoding within the corpus (or perhaps contained in an associated configuration file) to determine what portions of the text should be included for a particular stage of the investigation. SARA uses the scope node of its corpus builder function to perform this task (Aston and Burnard, 1998). We can, for instance, limit our search to a particular sub-corpus containing only spoken dialogue from the demographic part of the corpus where the speaker is male and under 35 years old.

l) *Lemmatisation*: Finally, we mention the ability of the software to carry out lemmatisation itself (perhaps on the fly). This is a vitally important characteristic for many types of linguistic investigation where the user wants to look at all inflected forms of a *lexeme* (dictionary head word) by frequency or in context. As indicated in Table 3.1 the tools carry out lemmatisation to varying degrees. In English at least, we need to mark the word class correctly to distinguish the stem of some words (Beale, 1987). An attempt can be made to recover the lemma of variant word forms by using fairly simple spelling rules to take care of *-s*, *-ing* and

-ed suffixes. But special rules are needed for verbs with irregular past forms and nouns with irregular plurals. Of course, if the corpus has already been lemmatised by hand or by automatic means and the software can display the lexeme field alongside (or instead of) the word field (original form in the text) we can make use of this in our study.

In section 1.1 we made a distinction between hypothesis-driven and data-driven research. One way of combining hypothesis-driven research with data-driven research, is to use the concordancing feature of the software tools in an exploratory manner to browse a corpus before we decide on a research question. However, this is an ad-hoc approach and it may not be an efficient or successful way of finding unexpected features in a corpus. Many researchers find keywords a useful starting point in their analyses (Hunston, 2002: 68), and the only tool in our survey capable of carrying this out using statistical methods is WordSmith. Hence we classify WordSmith as a tool of possible use in our data-driven method, the remaining eight tools we classify as hypothesis-driven.

3.5 Multi-purpose tools and architectures

In sections 3.2 – 3.4 we have focused on the different functions which corpus tools fulfil. We now look in more detail at software architectures.

In the early days of corpus software development, the typical case was a program designed and written 'in house' at the users' institution, intended to perform a single task. Naturally enough, some of this software became widely used and distributed and provided a model for further software developments. A 'corpus workbench' consisting of a group of programs was the next development. We have already seen in section 3.4, the CLAN software written (by Leonid Spektor of Carnegie Mellon University) originally for use with the CHILDES database (MacWhinney and Snow 1990, MacWhinney 1991). A more advanced cluster of the same general kind is the Lexa software suite developed by Hickey (1993a, 1993b),which includes corpus pre-processing, annotation, and text retrieval. These 'toolkits' take quite a significant step

from single-function to multi-function software development, the latter also illustrated by Brodda's (1991) PC Beta software.

After the move from single-task to multi-task software development, the next logical step is to aim for modular integrated architecture. The development of tools to build and exploit corpora which may run to hundreds of millions of words is an expensive task in terms of time and money. It is hardly surprising, therefore, that concepts such as reusability have been adapted to the field of corpus-based language engineering from the field of software engineering. A useful metaphor here is 'software Lego'. Programming practices should allow small programs to be slotted together to form larger and altogether more useful programs according to need. Developing software for new functions then need not require going back to the drawing board: a couple of pieces of 'Lego' to fit to the existing architecture may be all that is required. Two initiatives which have this modular type of design are (a) the MULTEXT project (as previously mentioned in section 3.2) and (b) the GATE architecture (Cunningham et al. 1996) developed at Sheffield in the UK. In the MULTEXT work, as in related work at Edinburgh (Thompson and McKelvie 1996), the unifying principle is that it should be possible for a text stream in a standard (SGML-based) format to be pipelined between any one module and another without hindrance. Cunningham et al (2000) describe the various software requirements that guided the implementation of GATE.

The openNLP[57] initiative has some overlap with GATE and is intended to act as a coordinating structure for several open source projects in Natural Language Processing.

3.6 Summary

In this chapter, the tools needed for the creation and exploitation of corpora, in particular annotated corpora, have been categorised into three major groups: corpus development (the input of information into a corpus), corpus editing (changing

[57] See the website at https://sourceforge.net/projects/opennlp/

information in a corpus), and information extraction (the output of information from a corpus). We have looked at features and given examples of software in each of the three groups, focussing particularly on software falling into the third category.

Choosing one package over another involves decisions about machine operating system type, as not many packages are supported across the main platforms (UNIX, Linux, PC DOS, PC Windows, Apple Macintosh). Considerable advantage can be gained by using web interfaces and off-the-shelf software such as commercial database packages. Making use of a web interface for corpus software will save the end-user some of the cost of the learning curve in adopting new software, since they will usually be familiar with web browsers which provide access from most platforms. There will often be no extra software to install for the end-user since web browsers are pre-installed along with the operating system. Pioneering concordance services have been provided using the web interface to Stuttgart's xkwic[58], or the simple search of BNC Online[59], TACTweb[60] (Bradley and Rockwell 1995, Rockwell et al 1997) and BNCweb[61] (Lehmann et al, 2000). These, however, usually require separate server machines, and in the case of xkwic, for example, this server is limited to a Unix/Linux operating system. An obvious disadvantage of this approach is the requirement that the user's computer is connected to a suitable network with access to the corpus server. We mentioned one instance of commercial database packages in section 3.2, in discussing corpus storage: the database of the Spoken English Corpus. The database architecture has the advantage of using the fast indexing and data management functions already available in a commercial database package. Not all the software in the corpus toolbox has to have been developed for, and dedicated to, corpus-based research.

The functionality and usability of search and retrieval packages have been enhanced over recent years to the extent that a number of quite sophisticated functionalities are

---

[58] Available for browsing (using username and password) the ICAME corpus collection online at http://www.hd.uib.no/icame.html and the Slovene concordance service at http://nl2.ijs.si/corpus/ provided by Tomaž Erjavec.

[59] BNC Online simple search is located at http://sara.natcorp.ox.ac.uk/lookup.html

[60] See website at http://tactweb.humanities.mcmaster.ca/

[61] Information on BNCweb is located at http://homepage.mac.com/bncweb/home.html

now commonplace and expected. In this chapter, we have summarised the inclusion or exclusion of twelve important features in nine of the most widely cited retrieval software packages in corpus linguistics and related research. Many of the tools are very capable of producing word frequency lists and KWIC concordances. However, only one (WordSmith) is capable of statistical comparison of word frequency lists. None of the tools combine the annotation-awareness capability with the comparison of frequency lists. It is this combination of two features that we see as vital in defining a practical data-driven approach as discussed in section 1.1. In section 4.4 of the next chapter, we will show a worked example to illustrate why this combination of features is particularly useful in corpus studies.

4. The Matrix method and tool

A linguistics professor was lecturing to his class one day. "In English," he said, "a double negative forms a positive. In some languages though, such as Russian, a double negative is still a negative." "However," he pointed out, "there is no language wherein a double positive can form a negative." A voice from the back of the room piped up, "Yeah, right"

4.1 Introduction

In the previous two chapters, we have reviewed the field of corpus linguistics, the process of frequency profiling, comparison of frequency profiles and the software available up to now in this field. This chapter describes in detail the Matrix method and the tool that has been implemented to carry out frequency profiling of corpora, and comparison of those profiles across corpora. A worked example is included to illustrate the method with two corpora.

4.2 Frequency profiling

Producing a word frequency profile from a raw text corpus is a non-trivial task. There are many decisions to take during the conversion of a piece of text into a list of words and their frequency of occurrence. This is evidenced by the fact that one can examine frequency lists for the same corpus produced by different software tools and have different lists of words as well as different frequencies for those words.

Given a raw corpus of English in a plain text file, the first decision to make is how to delimit the words in the text, a process often called tokenisation (Grefenstette and Tapanainen, 1994, Grefenstette, 1999). This is obviously simpler in languages such as English where words are generally delimited by the space character. In Chinese text, for example, word boundaries are not marked. Word divisions can be marked by

hand, or can be generated automatically with a small error rate (McEnery, Piao, and Xin, 2000). However, even in English text, we need a large number of heuristics. We need a list of characters which can act as punctuation. These are usually the following: .,?;:’“”-()[]! So, now we can use this set in addition to the space character to delimit words. Further complications arise with the hyphen character, the full stop and the apostrophe. In some texts, particularly those displayed in multiple column formats as in newspaper pages, the hyphen character can show the continuation of a word at the end of a line. In processing such text, we need to recover the complete word before we can record it as an entry in the word frequency list. The second problem character is the full stop. This can be used to show the end of a sentence, and we can usually test whether there is a following space character and then a capital letter at the start of the next word to confirm this. The full stop also appears in acronyms which are generally sequences of capital letters, and title nouns such as Mr., Mrs., Ms., Dr. and so on. Sequences of full stops can show ellipsis in text. More recently collected corpora from the domain of electronic communication contain email addresses and website addresses. These strings have their own formats that we need to be aware of, and should be considered as one word for the purposes of a frequency list. Lastly, the apostrophe is used in contractions such as *it’s*, *he’s*, *she’s*, *that’s*, *there’s* and words such as *o’clock*. In the Brown corpus (Kučera and Francis, 1967), for example, these were considered as single words, but we may decide to treat the *’s* as a word in its own right and separate it off.

Once we have a set of heuristics we can apply to word delimitation, we have two further decisions to make which will govern the resulting frequency profile. The first is whether to include the punctuation we have detected as part of the frequency list. Normally, it would be discarded, but we have to make this decision known to the user of the frequency list. The second decision is whether to preserve the surface form of the word if it contains capital letters (i.e. should the list be *case sensitive* or *case insensitive*) or full stop characters. If we decide to leave the form unchanged, we will produce a frequency list containing two or more entries for some words. For example, *The* and *the* will be recorded separately, the first one roughly showing how many times the word occurs at the start of a sentence (although not all occurrences of *The* are at the start of a sentence since it occurs in names of newspapers for example). Given that in general we probably are not interested in recording this distinction, the

usual approach taken and implemented in software is to force all characters to lower case (or upper case in WordSmith). This does however cause problems for homographs (different words spelt the same). Consider the string *polish* in a frequency list. This represents the word count for *polish* (cleaning substance) and *Polish* (Eastern European language). Without human intervention by inspecting each occurrence of the word in the text, we cannot be certain whether *Polish* with an initial capital at the beginning of a sentence is a reference to the language. In Matrix we also remove full stop characters so that for example *B.B.C.* and *BBC* would be counted together.

If we are using an annotated corpus, such as a corpus POS tagged by CLAWS, then we can assume that part of the task described above has already been done. Before giving each word a POS label, the POS tagging software will have to decide what constitutes a word. The CLAWS software incorporates various heuristics and implements these rules in its pre-processing module (Booth, 1987). This leaves us with the much simpler task of taking the words and converting their surface forms if necessary (although some POS taggers perform this task as well).

Building the frequency list itself would seem on the surface to be a fairly simple process. For each word in the text, check whether it is contained in our list so far; if it is there we increment the frequency by one, and if the word is not there, add it to the list and give it a frequency of one. The process described so far is presented diagrammatically in Figure 4.1.

This process works well for small corpora, say up to 100,000 words. However, beyond this point we need more efficient ways of storing the frequency list. The job of checking whether a new word already appears in our current list in the most time-consuming one. If new items are added to our list at the end, we have sequentially to search the entire list each time to check for a match. In large corpora such as the British National Corpus, our list also becomes large. Leech, Rayson and Wilson (2001: 8) report that 757,087 different word forms occur in the whole corpus, 52% of which occur only once. The standard information retrieval solution to this problem is to store the list in alphabetical order since if this is the case we only need to search as far as the position in the list where the new word should appear. Hence, we can avoid

searching all the way to the end of the list. To make this more efficient, we can use a binary search mechanism (Salton and McGill, 1983: 330). The binary search eliminates one half of the list at each search step. The new item is first matched against the entry in the middle of the list. If the item matches we can increment its frequency. Otherwise, we next proceed to the middle of the top half of the list if the new item is alphabetically ordered before the compared item, or to the middle of the bottom half of the list if the new item is alphabetically ordered after the compared item. This process of dividing the list into two halves continues until a match is found or no items remain. If the latter case occurs, we can insert the new item at this last point in the list.

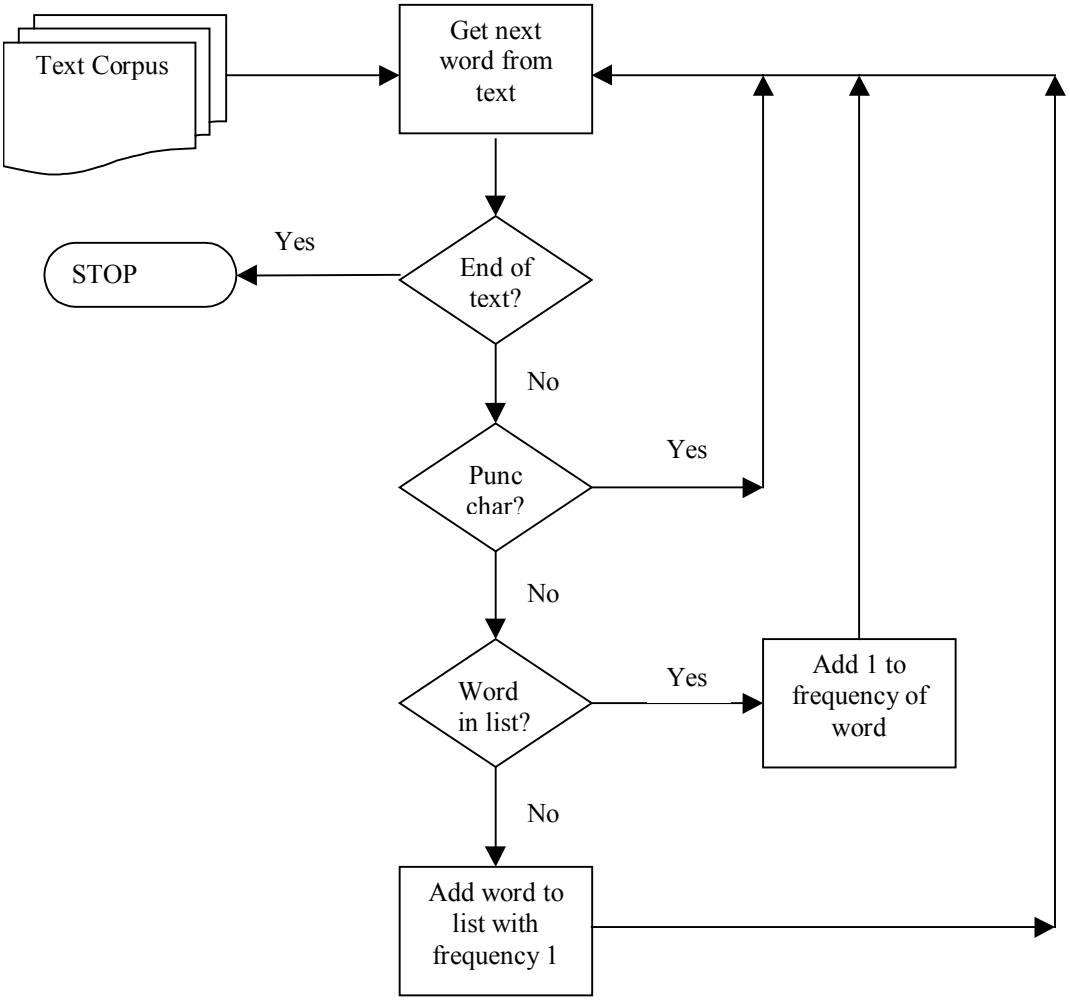


Figure 4.1 Flowchart showing basic frequency list process

In Matrix, we take this binary search process and embed it inside a simple indexed list. We split the list into twenty-seven sections. There is one section per letter of the alphabet and one extra for numbers. We maintain twenty-seven pointers to the start of each of these sections. The binary search for each new word is then limited to the appropriate section of the list, identified by examining the first letter (or numeral) in the word (or number). This search is O(log n) and the efficiency is vital when building frequency lists for large corpora.

Building frequency profiles can be more informative when we are dealing with annotated corpora. For example, if we have a corpus that is annotated at the POS and word-sense level, we can build frequency lists for POS and word-sense tags using the above technique simply by exchanging the 'word' in the process for a POS or word-sense 'tag'. The tags usually consist of sequences of letters and numbers and so are equally suitable for what is essentially an indexed list of strings. The main difference is that we know in advance that frequency lists of tags are of fixed length, usually equal to the size of the relevant tagset. The complication arises if we wish to record word frequencies by POS or word-sense tag. Assuming that each word can have more than one POS and sense tag, we might want to see the relative frequencies of 'table' as a noun and 'table' as a verb, and then compare 'table' as an item of furniture against 'table' as a mathematical object and 'table' as a speech act.

This aim is achieved in Matrix by indexing the 'word-POS-sense' triple internally at the lowest level of detail. In other words, Matrix's index records the frequency of 'table-noun-furniture' separately from 'table-noun-mathematics'. From this general frequency index, we can produce various different frequency lists. If we conflate entries by ignoring the sense tag field we obtain a frequency list containing 'table-noun' and 'table-verb'. If we conflate entries by ignoring the POS and sense tags then we obtain the overall frequency of 'table'. We can also ignore the word and sense tags to produce a POS tag frequency list, and so on. This conflation action is shown (with hypothetical frequencies for illustration) in Figure 4.2. Matrix's internal indexes are built dynamically during the first reading of a file. It does not require the files to be pre-indexed.

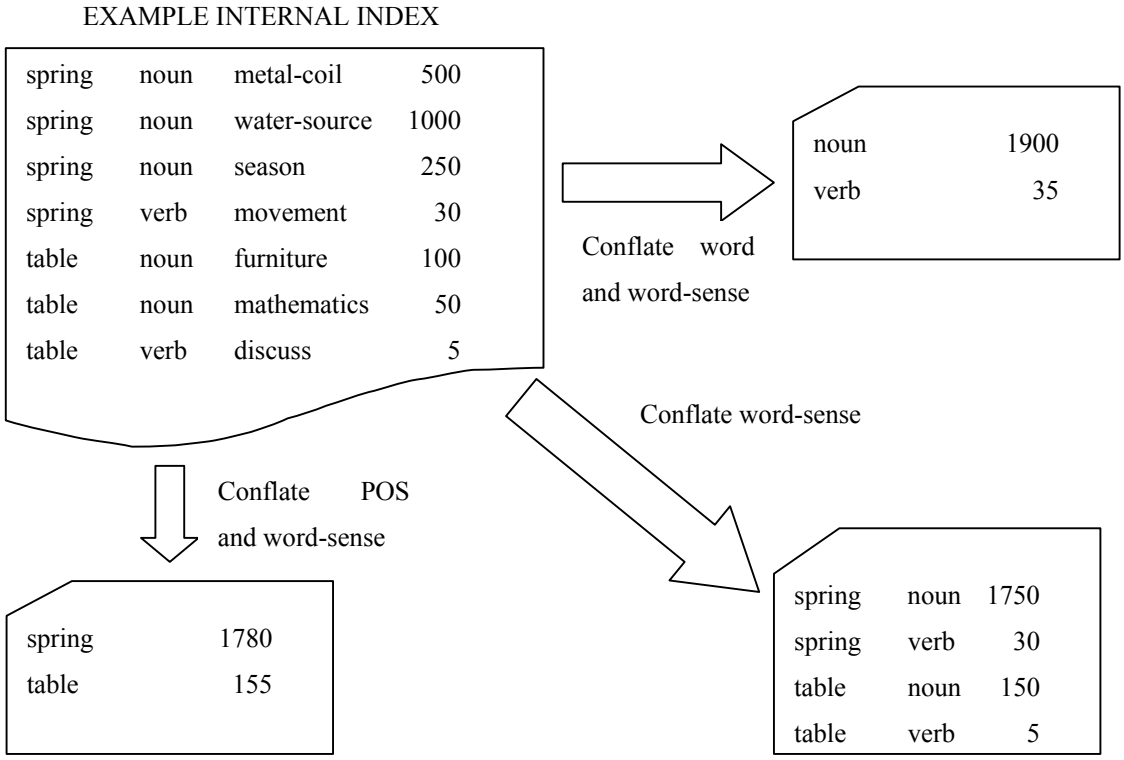


Figure 4.2 Matrix internal index

In section 2.6.1 we introduced the notion of dispersion statistics. These provide the user of a frequency list with more information about the spread of occurrences of an item in the corpus. For a large corpus or a corpus containing more than one type of text it is useful to consider the spread of occurrences of a word or other linguistic phenomena since it may be highly frequent in only a small subset of the text. In such cases, the simple frequency figure is not sufficient since it would give a false impression of the word's occurrence throughout the corpus. By splitting the corpus into a chosen number of equal-sized sectors, Matrix calculates two statistics:

1. *Range*: a simple count of how many text sectors include the word or item in question. The value varies between 1 and the number of sectors that the corpus has been divided into.
2. *Dispersion*: a statistical coefficient (Juilland's D) of how evenly distributed a word is across successive sectors of the corpus. Juilland's D is calculated as follows:

$$D = 1 - \frac{V}{\sqrt{n-1}}$$

where n is the number of sectors in the corpus. The variation coefficient V is given by:

$$V = \frac{s}{x}$$

where x is the mean sub-frequency of the word in the corpus (i.e. its frequency in each sector averaged) and s is the standard deviation of these sub-frequencies. We have selected Juilland's D as it has been shown to be the most reliable of the various dispersion coefficients that are available (Lyne 1985, 1986). The dispersion value varies between 0 and 1. The closer the value is to 1, the more equal the spread of occurrences across the sectors of the corpus. A value of 1 indicates that the frequency in each sector is the same.

Let us take an example from Leech, Rayson and Wilson (2001) to show how these statistics may be of use. The lemmas *HIV*, *keeper* and *lively* have quite similar frequencies in the whole BNC, approximately 16 occurrences each per million words. We might therefore be tempted to infer that they all have a similar currency of usage in the English language. However, when we look at how many corpus sectors they occur in (i.e. the range), we find that *lively* occurs in 97 as compared with *HIV*'s 62. *Keeper* occurs in 97. To calculate these figures, the BNC was divided into 100 equal-sized sectors, each of about 1 million words. Up to a point, this already confirms what we know: *HIV* appears to be a rather specialized term, used in a restricted number of sectors, whereas *lively* is a much more widespread word. But even these figures do not tell the whole story. If we look at the dispersion (Juilland's D) for each word, *lively* has a value of 0.92, *keeper* a value of 0.87 and *HIV* a value of 0.56. What does this difference between *lively* (0.92) and *keeper* (0.87) tell us? Remember that there was no difference in terms of how many corpus sectors they occurred in. The dispersion values, in contrast, suggest that, across corpus sectors, *lively* is more evenly distributed whereas *keeper* occurs more in bursts or clumps. Thus, by taking these dispersion values, we are able to avoid the false conclusion that these three words have roughly equivalent currency: in fact, only one is of widespread and general occurrence, with the others much more restricted to particular domains of discourse.

Matrix calculates the range and dispersion by storing a frequency vector along with each 'word-POS-sense' triple. The vector contains a list of each sector that the item occurs in, and the frequency of the item in that sector. During the internal index conflation process described above, these frequency vectors are merged to record occurrences of the resulting item. In this way, range and dispersion values are always available in lists produced by conflation operations.

4.3 The Matrix method: statistical comparison of frequency profiles

We claim that the Matrix method can be used in both types of corpus comparison (A and B, as described in section 2.7). Before applying the Matrix method it is important to consider the issues relevant to comparison of corpora as discussed in section 2.7; representativeness, homogeneity and comparability. The final issue, that of choice and reliability of statistical tests, has been addressed directly in this thesis. Our choice of the log-likelihood test (and the results produced by the method) will be evaluated in the next chapter. The method itself may assist in assessing representativeness, homogeneity and comparability of corpora, but in any experiments a user of the method should keep these issues in mind when interpreting the results. For example, if we chose to compare a written corpus with a spoken corpus, it is very likely that lexical and grammatical differences between the spoken and written language will be exposed as well as differences in domain or content that we may wish to focus on.

The Matrix method is applied using the following steps[62]. Given two corpora we wish to compare, we produce a set of frequency lists for each corpus. In previous studies, this would only be a word frequency list, but we produce a part-of-speech and semantic tag frequency list as well. Let us assume for now that we are performing a comparison at the word level (the application of this technique to POS or semantic tag frequency lists is achieved by constructing the contingency table below with tag rather

[62] A version of this method appears as Rayson and Garside (2000).

than word frequencies[63]). Due to independence assumptions, it is important that the two corpora do not overlap, or that one is a sub-corpus of the other.

For each word in the two frequency lists we calculate the log-likelihood (henceforth LL) statistic. The calculation is performed by constructing a contingency table as in Table 4.1.

Table 4.1 Contingency table for log-likelihood calculation

	CORPUS ONE	**CORPUS TWO**	**TOTAL**
Frequency of word	a	b	a+b
Frequency of word not occurring	c-a	d-b	c+d-a-b
TOTAL	c	d	c+d

Note that the value 'c' corresponds to the number of words in corpus one, and 'd' corresponds to the number of words in corpus two (N values in the formula below). The values 'a' and 'b' are called the observed values (O). We need to calculate the expected values (E) according to the following formula:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

In our case N1 = c, and N2 = d. So, for this word, E1 = c × (a+b) / (c+d) and E2 = d × (a+b) / (c+d). The calculation for the expected values takes account of the size of the two corpora, so we do not need to normalise the figures before applying the formula. We can then calculate the log-likelihood value according to this formula:

[63] The frequency distributions of part-of-speech and semantic tags are sharply different to word distributions (as we have seen in section 2.6.2). In these comparisons, we are unlikely to observe rare events such as tags occurring once.

$$-2\ln\lambda = 2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right)$$

This equates to calculating LL as follows: LL = 2×((a×ln (a/E1)) + (b×ln (b/E2))). An earlier version of the Matrix method used the chi-squared test, as described in section 2.7, instead of the log-likelihood ratio. A formative evaluation of this method appears in section 5.3.1 when it was applied to study vocabulary in the conversational spoken sub-corpus of the BNC.

The word frequency list is then sorted by the resulting LL values. Comparing LL values is possible at least for contingency tables with the same sample size, since its magnitude depends on N, as does the chi-squared statistic (Everitt, 1992: 54). Sorting the list gives the effect of placing the largest LL value at the top of the list representing the word that has the most significant relative frequency difference between the two corpora. In this way, we can see the words most indicative (or distinctive) of one corpus, as compared to the other corpus, at the top of the list. These are keywords as in Scott's statistical meaning. The words that appear with roughly similar relative frequencies in the two corpora appear lower down the list. Note that we do not recommend the reliance on the usual 5% and 1% level hypothesis tests by comparing the LL values to the critical values (3.84 and 6.63) in a chi-squared distribution table. As Kilgarriff & Rose (1998) note, even Pearson's X^2 is suitable without the 'hypothesis-testing link'. Given the non-random nature of words in a text, and the fact that we are carrying out multiple comparisons (see below), we are always likely to find frequencies of words which differ significantly across any two texts, and the higher the frequencies, the more information the statistical test has to work with. As we will see in our evaluation in section 5.2 we can rely on testing LL values at the 0.01% level with a critical value of 15.13, if a statistically significant result is required for a particular item.

Kessler (2001: 44) similarly carries out hundreds of experiments on word lists. He states "the very definition of significance testing at say, the 0.01 level, is that one is finding that the frequencies are so far from the expected value that one would expect

them to turn up by chance only one time in a hundred. So if one is doing 100 trials, approximately one of them is expected to show significance, even if in fact the results are truly random". His solution is require a higher significance value for each test by dividing by the number of tests (0.01/100 = 0.0001). Lebart et al (1998: 135) also highlight the problem of multiple comparisons. Critical values corresponding to these small levels are not usually listed in chi-squared tables, although they can be calculated.

The next stage in the Matrix method is to carry out the same comparison at the POS and semantic level. These comparisons extend Scott's technique to produce *key items* rather than key words, and allow us to find *key grammatical categories* and *key concepts*. It is at this point that the researcher must intervene and qualitatively examine concordance examples of the significant words, POS and semantic tags highlighted by this technique. We are not proposing a completely automated approach. Granger (1993) warns that we should not limit corpus investigation to what the computer can do for us automatically, and she quotes other authors who have come to the same conclusion. Woods et al (1986: 130) note that it is unsatisfactory simply to state that something was significant at the 1% level or was not significant at the 5% level, and Kretzschmar et al (1997) agree that the "analyst is always responsible for explanations" following a statistically significant result. One aim of the POS and semantic comparisons is to reduce the number of key categories that the researcher should examine (addressing a criticism of the key words approach by Berber Sardinha, 1999, see section 2.7.2).

This qualitative examination is described by Leech and Fallon (1992) in more detail as the second stage of their two stage process. The Matrix method can be substituted as an improvement on their stage one process which involved examination of the Hofland and Johansson (1982) lists of word frequencies in British and American English and selection of the items marked with significant differences. They describe the second stage using a KWIC concordance tool to examine concordance lines from the Brown and LOB corpora. They cite two main reasons for consulting the concordance lines:

1. *To check whether the frequency of the graphic form actually reflected the sense of the word they were interested in*. This issue is addressed by the Matrix comparison at the semantic tag level.
2. *To check that the high frequency of an item was not due to any obvious skewing of its distribution in the corpus*. This issue is addressed by the Matrix tool providing range and dispersion information alongside the frequency of an item.

The inclusion of range and dispersion in Matrix also reduces the justification for reusing Kilgarriff's (1996b) selection of the Mann-Whitney test which was due to the possibility of a high frequency 'burst' of a word in a single document. Scott (2000a) discusses the notion of burstiness in relation to keywords. He begins with manual segmentation of texts, but this is slow and subjective, so decides to proceed using automatic segmentation into equal sized segments. This is akin to the dispersion values reported in Matrix.

As we have described, the Matrix method relies on the total number of words in the corpus to calculate the expected frequencies of words, POS and semantic tags. Multi-word units, identified by the POS and semantic taggers, are counted as one unit for comparison and in the totals hence not affecting the counts. In their experimental design Nelson et al (2002: 270) caution against using the quantity of material in a corpus to estimate expected frequency. They give the example of 'may' which they write should be estimated against the total number of modals rather than the total number of words. They write that if we observe an unexpectedly high occurrence of 'may' in one corpus relative to another, then it might be due to (i) an increase in modals overall, or (ii) a decrease in the use of other modals. Whilst it is true "in grammar, [that] the possibility of a particular linguistic expression is dependent on whether the expression was feasible in the first place", there are practical problems in listing the possible alternatives at the semantic level (as well as the grammatical level). The Matrix method will identify any increase in modals and any decrease in the use of other modals by a combination of comparisons at the word level and the POS level.

To place the Matrix method in the context of our type III corpus linguistic methodology from section 2.3, it corresponds to steps 3 and 4 (question and retrieve), assuming that steps 1 and 2 (build and annotate) have already taken place. Step 5 (interpret) is perhaps the most important stage, and it corresponds to Leech and Fallon's stage two.

4.4 Worked example of annotation profile comparison

In this section, we show a practical example, applying the Matrix method and tool to study the language used in a pair of corpora. We will illustrate the different results obtained at the word, POS and semantic levels. This section will not contain an in depth study of any particular observed phenomena; the case studies in section 5.3 will perform that function. We have chosen to study the language used in the United Kingdom General Election manifestos of the Labour and Liberal Democratic (LibDem) parties from the June 7th 2001 election.

In a type I or II study, we might decide before looking at the corpus data what phenomena we wish to investigate. We would then collect the necessary data and examine the differences in the two manifestos to confirm or reject our hypothesis. In a type III study as pursued here, we will examine the corpus data and let the analysis direct us to suggest further items to study.

At the time of writing, the Labour and LibDem 2001 General Election manifestos were available to download from their respective websites[64]. The LibDem manifesto was available as a 248Kb Microsoft Word document (text only, no pictures) containing 57 pages and 20,402 words. This was converted to plain text HTML format using Microsoft Word 2000. The Labour manifesto was available as 4 Adobe PDF files totalling 2.6Mb, the first and last of which represented the front and back covers of the document containing pictures and one short paragraph of text. The 2 remaining documents contained 44 pages of pictures and text. These were converted

[64] The Labour Party website URL is http://www.labour.org.uk/ and the Liberal Democrat Party website is http://www.libdems.org.uk/

to plain text HTML using Adobe Acrobat 5.0 with the MakeAccessible plug-in and SaveAsXML plug-in available from Adobe (at access.adobe.com). The resulting conflated file contained 28,107 words.

4.4.1 Comparison at the word level

We began our analysis by producing a word frequency list for our two corpora. The word frequency list for the Labour manifesto had almost 4,200 entries, and the LibDem had over 3,600. In Table 4.2 we show the top 20 items in each list. The contents of the table illustrate four significant problems with using and comparing basic word frequency lists.

Table 4.2 Top 20 most frequent words in Labour and LibDem manifestos

LibDem manifesto		Labour manifesto	
Word	**Frequency**	**Word**	**Frequency**
the	1174	the	1482
and	794	to	1112
to	736	and	1100
of	632	of	719
will	461	we	669
we	428	in	546
a	345	will	515
in	320	a	506
for	308	for	491
by	196	is	330
on	166	our	272
are	128	with	242
that	123	are	226
is	119	have	209
be	109	by	194
more	107	on	185
with	107	be	173
have	97	new	165
this	94	more	162
their	93	people	160

First, the frequencies cannot be compared directly unless they are normalised. The Labour manifesto contains nearly 8,000 more words than the LibDem one, so we would expect on average the frequencies to be higher for each word. Normalising the frequencies with respect to the corpus size means converting the frequency to a percentage value, or sometimes a value per thousand (or per million) words. If we consider the word *will* in the table, it has a frequency of 515 in the Labour manifesto and 461 in the LibDem one, incorrectly suggesting higher usage by Labour. We should not compare these observed figures since the normalised values 1.83% for Labour and 2.26% for LibDem show that *will* occurs with greater relative frequency in the LibDem data. This difference is significant (LL value of 10.63 at 1 d.f. $p < 0.005$). It is worth noting at this point that the LL calculation does include normalisation as part of the expected value formula.

Second, the high frequency words at the top of any word frequency list are generally of no further interest to anyone trying to differentiate the content of two corpora. The top 20 items usually consist of closed class words, such as articles (*the*), prepositions (*to*, *of*, *in*, *for* etc), conjunctions (*and*), and auxiliary verbs (*are*, *is*, *be*, *have*). At the bottom of the top 20 items in the Labour list, we have 'interesting' words from open classes worthy of further consideration such as the adjective *new*, the noun *people* and the adverb *more*. Despite this, high frequency words are of interest to some (Sinclair, 1999).

Third, comparing the ranking of words is also misleading. The LibDem list contains *more* 3 places higher up the list than its rank in the Labour list. If we compare the relative frequencies of the word *more* in the two texts: Labour usage of 162 (0.58%) is higher than LibDem usage at 107 (0.52%). In fact, the difference is not significant (Log-likelihood value of 0.58 at 1d.f.), but we might be tempted to jump to the wrong conclusion given their relative positions in these lists.

Fourth, multi-word-units are not counted together. In the research community these are referred to under various names, sometimes called lexical bundles or prefabricated expressions. Depending on the purpose of our study, multi-word-units may be quite significant. For example, *to* in the LibDem data occurs 736 times. It can also occur in multi word prepositions, for example *subject to*, *according to* and *due to*.

Applying the Matrix method at the word level, we can compare the relative use of words between Labour and LibDem manifestos. For 1 d.f. (degree of freedom), at 99% confidence (or $p < 0.01$) the cut-off of 6.63 would indicate that there are 161 words significantly overused or underused between the Labour and LibDem data. This reduces to 65 words significantly overused or underused at the 99.99% ($p < 0.0001$) level with critical value 15.13, as suggested by our evaluation in section 5.2.2). The top 20 words (with the largest LL values) in this set are shown in Table 4.3.

Table 4.3 Top 20 most significant differences at word level between Labour and LibDem manifestos

		LibDem manifesto		Labour manifesto			
	Word	**Frequency**	**Rel. freq.**	**Frequency**	**Rel. freq.**	**O/U-use**	**LL**
1	liberal	47	0.23	0	0.00	+	81.41
2	would	70	0.34	10	0.04	+	71.89
3	democrats	40	0.20	0	0.00	+	69.29
4	our	76	0.37	272	0.97	-	63.22
5	labour	33	0.16	152	0.54	-	49.56
6	is	119	0.58	330	1.17	-	47.04
7	which	92	0.45	37	0.13	+	45.13
8	now	8	0.04	76	0.27	-	43.97
9	1997	4	0.02	54	0.19	-	36.76
10	green	26	0.13	2	0.01	+	32.81
11	environmental	47	0.23	14	0.05	+	30.98
12	establish	34	0.17	7	0.02	+	29.06
13	since	2	0.01	38	0.14	-	29.06
14	ten-year	0	0.00	25	0.09	-	27.29
15	also	88	0.43	50	0.18	+	26.30
16	Governments	15	0.07	0	0.00	+	25.98
17	britains	15	0.07	0	0.00	+	25.98
18	long_term	15	0.07	0	0.00	+	25.98
19	new	57	0.28	165	0.59	-	25.91
20	's	29	0.14	106	0.38	-	25.46

The table shows for each manifesto the frequency and relative frequency for each word in the top 20. The penultimate column indicates overuse (+) and underuse (-) of the word in the LibDem corpus with respect to the Labour corpus.

The first, third and fifth entries are unsurprising given that they show the names of the political parties. Looking at the concordance for *liberal*, there are 44 occurrences of *Liberal Democrat(s)* in the LibDem manifesto and none in the Labour one. There are some (33) references to the Labour party in the LibDem manifesto, although it has a lower frequency relative to the Labour document. It is therefore worth noting that the Labour manifesto chooses not to mention the Liberal Democrats at all.

The second most significant difference, with LL value of 71.89, alerts us to the fact that the word *would* is used almost 9 times relatively more frequently (0.04% compared to 0.34%) in the LibDem data. At this point we used the Matrix tool to look at a concordance of the key word *would* and this is shown in Figure 4.3.

mitted to training programmes which	would	bring enormous benefits to the econo
n: yes"> Companies eligible	would	include those working with Investors
ot on the unemployment register who	would	like work . <o:p> </o:p> </p
acerun: yes"> The programme	would	include environmental assessment of
nmental assessment of buildings and	would	promote the use of better insulation
/b> retained police officers . This	would	give the police more flexibility <b
nd violent offenders , so that they	would	only be released following an assess
"mso-spacerun: yes"> Reform	would	respond not only to the problems cau
rge Young Offender Institutions and	would	ensure that those young people who m
lack; layout-grid-mode:line'> which	would	limit the <span style='font-
e="mso-spacerun: yes"> This	would	extend across the UK the scheme curr
with the safety benefits that this	would	bring . <span style="mso-spacerun: y
cess in tackling poverty in Britain	would	be measured by a Quality of Life Ind
e="mso-spacerun: yes"> This	would	include a statement of the standards
pacerun: yes"> Over time we	would	ensure that a growing proportion of
"mso-spacerun: yes"> People	would	no longer have to show a history of
to claim their basic pension . This	would	eventually help around 3.4 million p
un: yes"> At present , this	would	take 1.4 million people on low incom
an> Anyone earning less than 25,000	would	pay less tax even allowing for our 1
with a Local Initiatives Fund which	would	give grants to support libraries , m
="mso-spacerun: yes"> OFCOM	would	ensure that these standards are main
asters , regulated by OFCOM , which	would	then guarantee them the right to be
setting a deadline after which they	would	automatically lapse . <span style="m
s"> This tax-free allowance	would	be set at 1500 and apply to all smal

Figure 4.3 Concordance of key word *would* from LibDem manifesto

Would, unlike *will*, denotes hypothetical or unreal events. Given the overuse of *would*, we might hypothesise that that the LibDem manifesto talks more than the Labour one about possible future plans rather than definite plans, in other words the LibDems,

unlike Labour, do not expect to win! Moreover, we can look lower down the word level comparison to remind us that *will* is also overused by the LibDems (2.26%) relative to Labour (1.83%) with a LL value of 10.63, which is significant at 99% ($p < 0.01$). We will look at the relative use of modal verbs in the next section.

The fourth most significant difference is the word *our*, which is used significantly more in the Labour manifesto (0.97%) than in the LibDem statement (0.37%). The next step is to look at concordance lines for *our* in the two documents and initially classify the occurrences into those which refer to

- the British/English nation or people, e.g. "our children", "our sense of fair play"
- the Labour party/government, e.g. "our pledge not to extend VAT", "our reforms since 1997"
- ambiguous cases between the inclusive and exclusive classes, e.g. "incentives to meet our ambitions"

This relative use of these three categories might allow us to investigate whether Labour are intentionally using ambiguous language to make the reader feel that the party shares the same goals as they do. This mirrors the investigation of the use of inclusive and exclusive *we* in the language of New Labour, see Fairclough (2000: 35).

At ninth position in the table is the number *1997* which is more frequent in the Labour manifesto (0.19% compared to 0.02%). This is unsurprising since 1997 was the year of the previous Labour victory in the General Election and the contexts for this keyword show the manifesto mentioning Labour's record in office since 1997. Labour's achievements (since 1997) are also flagged by the keyword *now*, which is the eighth most significant difference, and over six times more frequent in the Labour text (0.27% compared to 0.04%). Figure 4.4 displays a section of the concordance lines for this keyword showing this trend.

At nineteenth position in the table is the keyword *new* which as one would expect is overused in the Labour manifesto. The slogan 'New Labour, New Britain' was first used at the 1994 Labour Party conference and Fairclough (2000: 18) discusses relevant themes such as renewal and modernisation.

NT color="#000000"> Europe . Britain	now	has the best combination </FONT> </Su
rld 's first University for Industry	now	offers over 400 skills courses . For
. Safer train protection systems are	now	being installed and will be extended
e been scrapped ; all new roads must	now	be strictly appraised for maximum ben
er ten years . £ 8.4 billion is	now	being invested in local authority sch
are increasing , and over 100 towns	now	have bus services linked to train sta
ght historic wrongs . Every employee	now	has the right to four weeks ' paid ho
RDAs) have been set up and why they	now	have extra money and new freedoms . <
00000"> to £ 1.7 billion a year	now	pledged to RDAs to </FONT> </P > <P
r cent of the national workforce are	now	employed in agriculture . But the ind
velopment priorities . CAP reform is	now	more possible ; Labour 's engagement
our platform – which is why we	now	have a unified grading scheme for hot
rt services ; and the Post Office is	now	obliged to prevent closure of rural p
s of coastal and inland flooding are	now	widely appreciated , and we are commi
ONT> </B> <B> <FONT color="#6C3C8A">	Now	our ambition is for Britain to </FONT
c services are always second class .	Now	is the time to move forward . Economi
setting a clear national framework .	Now	we need to move on , empowering front
r refurbishment ; 20,000 schools are	now	connected to the internet ; there are

Figure 4.4 Concordance of key word *now* from Labour manifesto

4.4.2 Comparison at the POS level

In the previous section, we described four significant problems with using basic word frequency lists. As Barnbrook (1996: 53) writes, there are further limitations to the basic word frequency list related to the word forms as well as the frequencies. Inflected forms of words are not counted together, but word forms with two (or more) POS tags or meanings are counted together. This can be partially solved by annotating the text with POS tags. We used the CLAWS tagger described in section 3.2 to assign word-class codes to the Labour and LibDem data.

Once the data has been tagged we have access to what Francis and Kučera (1982) call grammatical words, i.e. words and their associated parts of speech. According to CLAWS, the Labour and LibDem data contain no words that are ambiguous by POS. This means that each word in the data appears only within one part of speech, although in a much larger corpus (or corpus from another domain), you could find both noun and verb usage of the word *will* for example. We can compare the two files for their relative use of grammatical categories using the Matrix method applied at the POS level. For 1 d.f., at $p < 0.01$ the cut-off of 6.63 would indicate that there are 30 POS tags significantly overused or underused between the Labour and LibDem data. At the 99.99% level ($p < 0.0001$), there are 17 significant POS tags. The top 20 tags (with the largest LL values) in this set are shown in Table 4.4.

Table 4.4 Top 20 most significant differences at POS level between Labour and LibDem manifestos

		LibDem manifesto		Labour manifesto			
	POS	**Frequency**	**Rel. freq.**	**Frequency**	**Rel. freq.**	**O/U-use**	**LL**
1	MC	124	0.61	587	2.09	-	197.39
2	RT	13	0.06	105	0.37	-	55.26
3	VBZ	119	0.58	334	1.19	-	48.96
4	MD	22	0.11	122	0.43	-	48.15
5	NN2	1999	9.80	2271	8.08	+	39.30
6	DDQ	98	0.48	47	0.17	+	38.37
7	APPGE	199	0.98	438	1.56	-	31.61
8	VM	637	3.12	650	2.31	+	28.85
9	VV0	646	3.17	662	2.36	+	28.49
10	RR	379	1.86	368	1.31	+	22.77
11	GE	39	0.19	119	0.42	-	20.85
12	VH0	73	0.36	184	0.65	-	20.56
13	NNO	0	0.00	17	0.06	-	18.55
14	II21	68	0.33	41	0.15	+	18.19
15	IW	119	0.58	258	0.92	-	17.58
16	VVN	346	1.70	624	2.22	-	16.52
17	CSW	0	0.00	15	0.05	-	16.37
18	IO	633	3.10	718	2.55	+	12.64
19	NPM1	0	0.00	11	0.04	-	12.01
20	VVG	433	2.12	476	1.69	+	11.49

The most significant difference at the POS level is for the tag *MC* that marks cardinal numbers. The Labour manifesto includes more than 3 times as many cardinal numbers as in the LibDem one. This is largely the year 1997 as highlighted by the comparison at the word level. Also, we find *three*, *2004*, *2010* occurring quite frequently. Looking at concordances for these items, we observe that the Labour manifesto includes a large number of pledges for completion over the next *three* years, by *2004*, or by *2010*.

The second most significant difference at POS level is for POS tag *RT* (time adverb) that includes occurrences of *now* and *today* more frequently in the Labour manifesto. We have already commented on the keyword *now* above, but the keyword *today* also seems to act as a marker for mentions of Labour's achievements since the previous election.

The key POS tag *APPGE* (pre-nominal possessive pronoun) is overused in the Labour text (LL value of 31.61) and this is mostly due to the preference in the Labour manifesto for the keyword *our* as described above.

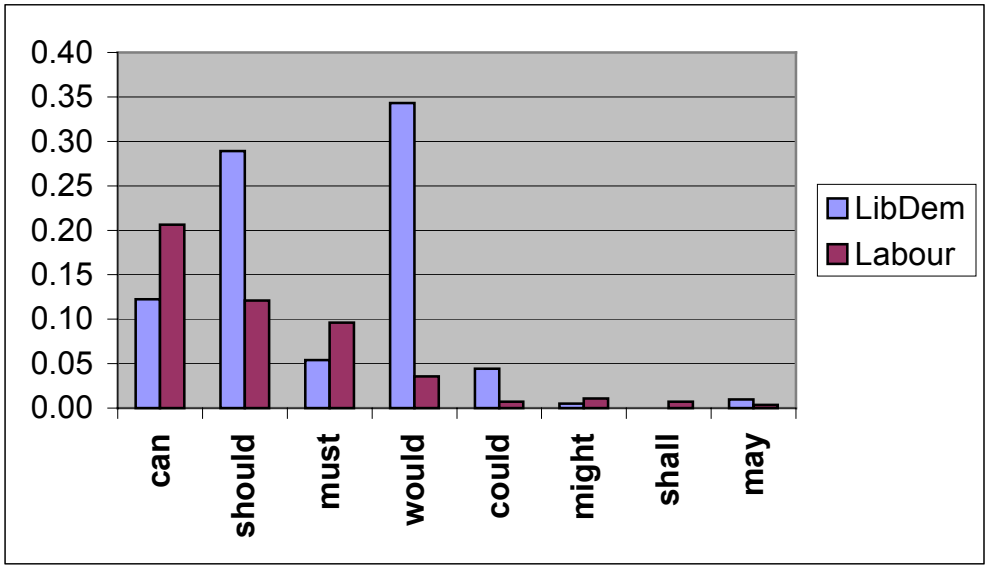


Figure 4.5 Relative use of modal verbs in LibDem and Labour manifestos

With a LL value of 28.85, we find that the LibDem manifesto overuses *modal verbs* (VM). This word class includes *will* and *would* and we have already discussed these keywords from the word level comparison. However, we can now compare the relative frequency of use of the other modal verbs. This is illustrated by Figure 4.5, which omits *will* as the highest frequency modal in order to obtain a reasonable scale for the chart.

4.4.3 Comparison at the semantic tag level

We used the USAS tagger described in section 3.2 to assign semantic field codes to the Labour and LibDem data. We can then compare the two files for their relative use of USAS categories using the Matrix method applied at the semantic level. For 1 d.f., at $p < 0.01$ the cut-off of 6.63 would indicate that there are 60 USAS tags significantly overused or underused between the Labour and LibDem data. At the $p < 0.0001$ level the critical value is 15.13, giving 18 significant USAS tags. The top 20 tags (with the largest LL values) in this set are shown in Table 4.5.

The most significant difference (LL value 141.97) in the semantic comparison is for the tag N1 representing the semantic field *numbers*. This is largely due to words with POS tag MC (as highlighted by the POS level comparison) being overused in the Labour manifesto.

The next most significant difference (LL value 72.72) indicates the overuse of the concept of *permission* (S7.4+) in the LibDem manifesto. Upon examining the concordance for this tag (part of which is shown in Figure 4.6), we find that 47 of the entries are the word *liberal*, and 44 of these refer to the *Liberal Democrat(s)*. In fact, these items are mistagged by the automatic semantic tagger and should obtain the G1.2 tag indicating the *political* semantic field. If we recalculate by omitting the 44 mistakes, we see that the relative frequencies are 0.43% in the LibDem document compared to 0.17% in the Labour one, and this still results in a significant LL value of 28.36. Looking at the other terms in this field such as *allow*, *right*, and *entitled*, we might form the hypothesis that the LibDem manifesto focuses more on personal freedoms than the Labour text, and study this in more detail. This hypothesis is corroborated by evidence from the sixth most significant difference, which is the concept of *constraint* (A1.7-) overused in the LibDem manifesto (0.38% compared to 0.12%). The minus sign at the end of the tag indicates the negative concept and the words we find within this category are *freedom(s)* and *liberties*.

Table 4.5 Top 20 most significant differences at semantic level between Labour and LibDem manifestos

		LibDem manifesto		Labour manifesto				
	Semantic tag	**Freq.**	**Rel. freq.**	**Freq.**	**Rel. freq.**	**O/U-use**	**LL**	**Semantic category**
1	N1	142	0.70	547	1.95	-	141.97	Numbers
2	S7.4+	131	0.64	47	0.17	+	72.72	Permission
3	T3-	139	0.68	375	1.33	-	50.05	Time: new and young
4	G1.1	362	1.77	293	1.04	+	46.13	Government etc.
5	I3.1	170	0.83	413	1.47	-	41.49	Work and employment
6	A1.7-	77	0.38	33	0.12	+	35.01	Constraint
7	M3	141	0.69	92	0.33	+	32.03	Vehicles and transport on land
8	A3+	236	1.16	490	1.74	-	27.95	Being
9	O4.3	30	0.15	6	0.02	+	26.08	Colour and colour patterns
10	N5	76	0.37	198	0.70	-	24.19	Quantities
11	A6.1-	99	0.49	63	0.22	+	23.74	Comparing: different
12	X2.4	93	0.46	59	0.21	+	22.45	Investigate, examine, test, search
13	W5	27	0.13	7	0.02	+	19.84	Green issues
14	T2++	38	0.19	114	0.41	-	19.30	Time: Continuing
15	T2-	58	0.28	32	0.11	+	18.25	Time: Stopping
16	A2.1+	156	0.76	321	1.14	-	17.60	Affect: Modify, change
17	N4	43	0.21	119	0.42	-	16.88	Linear order
18	O1	30	0.15	11	0.04	+	16.29	Substances and materials
19	N5-	110	0.54	88	0.31	+	14.56	Quantities
20	S4	40	0.20	108	0.38	-	14.44	Kin

The third most significant category is *Time: new and young* (T3-) which is overused in the Labour manifesto (1.33%) relative to LibDem (0.68%). This category marks the words *new*, *child(ren)*, *young*, and *modern* amongst others. The keyword *new* has already been identified by the word level comparison. The young/family terms relate to the family policy area mentioned below. A related category at position sixteen is that of *affect: modify, change* (A2.1+), which is overused in the Labour document

(1.14%) compared to the LibDem text (0.76%). This category contains words such as *reform(s), develop(ment)* and *change*. Fairclough (2000: 18) links *reform* to the sense of political renewal conveyed by Labour indicated by keywords such as *new*.

n: yes"> We will also	allow	people to stand for elected of
"> We will extend the	right	to vote by post and investigat
wers of Select Committees and	allow	more pre-legislative scrutiny
s more say over the budget by	allowing	them to propose spending amend
te from the Finance Bill , to	allow	for greater consultation on ta
acerun: yes"> We will	allow	the Welsh Assembly the right t
allow the Welsh Assembly the	right	to pass primary legislation an
cerun: yes"> We would	allow	further devolution of powers a
span> They are essential to a	liberal	society in which people are en
black;layout-grid-mode:line'>	Liberal	Democrats will : <o:p> </o:p>
trong framework of individual	rights	, extending the protection alr
by European law , so that the	rights	of the individual outweigh the
d personal relationship legal	rights	, such as next-of-kin arrangem
span style='color:black'> The	Right	to Know and the Right to Priva
k'> The Right to Know and the	Right	to Privacy <o:p> </o:p> </span
e individuals should have the	right	to know as much as possible ab
eplace the system of warrants	approved	by Ministers with a system of
by Ministers with a system of	approval	by judges to remove any confli
r:black;font-style:normal'> A	Right	to Environmental Information ,
:normal'> We will protect the	right	to legal and peaceful protest
e that farm animals should be	entitled	to high welfare standards . <s

Figure 4.6 Concordance of key concept *permission* from LibDem manifesto

At eleventh position we see the concept *comparing: different* (A6.1-) which is used to a greater extent in the LibDem manifesto (0.49% compared to 0.22%). This includes words such as *other(s)*, *discrimination*, *different*, *separate*, and *conflict*. The reasons behind this difference are not clear and require further investigation. We might hypothesise that the lower count in the Labour text stems from the 'one-nation politics' of Labour whose discourse is inclusive and consensual, and would de-emphasize such words with negative connotations.

The fourteenth and fifteenth entries for *time: continuing* (T2++) and *time: stopping* (T2-) can be examined together. Continuity concepts occur more frequently in the Labour document and the reverse is true for concepts of ending/stopping. From the concordances of these concepts they seem to mark government policies that Labour would continue pursuing if they were to stay in power and that the Liberal Democrats would end if they were elected.

At eighteenth position in the table we have *substances and materials* (O1) used to a greater extent in the LibDem manifesto (0.15% relative to 0.04%). This category includes words such as *fuel(s)*, *air*, *water*, *gas*, and *petrol*. Partly this seems to be

related to the LibDem textual focus on environmental issues mentioning fuel taxation policy and conservation of resources.

Emerging from the comparison at the semantic or conceptual level, we can see relative differences in the prominence of party policy areas. Labour's document focuses more on *work and employment* (USAS tag I3.1), and *kin* (S4) representing family issues. The LibDem manifesto devotes more of its content to *vehicles and transport* (M3) reflecting transport policy, and to *green issues* (W5) and *colour* (O4.3) indicating green/environmental policy. This is also shown at the word level in Table 4.3 with the key words *green* and *environmental* showing increased use in the LibDem document, but the comparison at semantic level provides more reliable evidence of the trend since several key words and phrases contribute and confirm the trend, e.g. *pollution* and *environmentally friendly*.

4.4.4 Conclusion to the worked example

This ability to extract key concepts and create or suggest hypotheses about major trends from the two documents demonstrates clearly the advantages of the comparison at the semantic level in addition to the (stylistic comparison) at the word and POS levels. We have to examine a smaller number of key items in the semantic comparison than we otherwise would for the word level[65]. The same words representing the concepts are available at the word level. However, the word level profile is in the region of 4,000 lines and the words forming key concepts are spread throughout the profile, so overall trends are more difficult to identify. Furthermore, it is not possible to identify some of the significant concepts at the word or POS level. Consider, *work and employment* (I3.1) mentioned above. Of the words in this category such as *work(ing)*, *staff*, *(un)employment*, *job(s)*, and *employees*, only the word *work* (LL 10.60) is significant in the word level comparison. Collecting together words into their semantic fields allows us to see trends that are invisible at the word level. Henry

[65] Nevill-Manning et al (1999) similarly report the speed improvements for finding useful information in large collections (digital libraries) using a hierarchical structure of phrases.

and Roseberry (2001: 101) also report a similar finding where an important semantic class groups together low frequency words that would otherwise have been missed.

Two further advantages of the comparison at the semantic level are that multi-word-units are counted together and variants within a lemma are usually grouped together. Multi-word-units are identified by the list of templates associated with the USAS system described in section 3.2.2. In the LibDem data for example, the following terms are identified: *local authorities*, *public transport*, *human rights*, *United Kingdom*, *league tables*.

In this section, we have looked at the language used in the United Kingdom General Election manifestos of the Labour and Liberal Democratic (LibDem) parties from the June 7th 2001 election. The initial results have suggested numerous avenues for further investigation to pursue a type III (data-driven, see section 2.3) study, ranging from lexical studies, through grammatical variation to analyses of party political differences (political linguistics). Some of these are summarised here:

1. An investigation of the inclusive language of Labour, indicated by their manifesto having greater use of the word *our*
2. An investigation into the differing use of modal verbs between the LibDem and Labour manifestos, signposted by the overuse of *would* in the LibDem manifesto
3. An investigation into the relative use of *permission* and *freedom* concepts, highlighted by significantly greater use of these concepts in the LibDem manifesto
4. An investigation into the political renewal senses conveyed by overuse of terms such as *new*, *modern*, *reform*, and *change* in the Labour manifesto
5. An investigation into party policy differences between LibDem and Labour indicated by significant differences in the relative use of concepts related to environmental issues, family issues, work and employment, and transport

4.5 The Matrix tool

In this section, we will focus on the software tool that implements the Matrix method. We will describe the user interface and the component architecture.

4.5.1 The user interface

The user interface to the Matrix tool has changed significantly over time. In this section we give an overview of the interfaces and discuss the reasons for the changes.

Early versions of Matrix were developed in the C programming language to run on UNIX Solaris systems using the *curses* library, see Strang (1986). The curses library supplies a terminal-independent screen-painting and keyboard-handling facility for text-based terminals, such as VT100s, the Linux console, and the simulated terminal provided by X11 programs such as xterm. Display terminals support various control codes to perform common operations such as moving the cursor, scrolling the screen, and erasing areas. The curses library hides all the details of different terminals. Figure 4.7 shows a screenshot of the main menu provided by Tmatrix (the T indicating *terminal* version).

A semantically tagged file is loaded by typing ‘l’ and then entering the filename. Files in other formats with varying degrees of annotation can be loaded using other key presses. For this example, we have loaded the tagged LibDem data described in section 4.4. Once a file has loaded, the frequency profile is displayed by typing ‘f’ from the main menu. Figure 4.8 shows the profile as displayed in Tmatrix showing actual frequencies (column headed *1*) and relative frequencies (column headed *1%*).

Concordances are then produced by typing ‘a’ and then the line number of the item. Figure 4.9 shows the Tmatrix display for the concordance of the word *our* from the LibDem data. Significantly, Tmatrix can show concordances for POS and semantic tags by selecting these items from POS or semantic tag frequency lists. This allows us to examine key grammatical items and concepts from the annotated data.

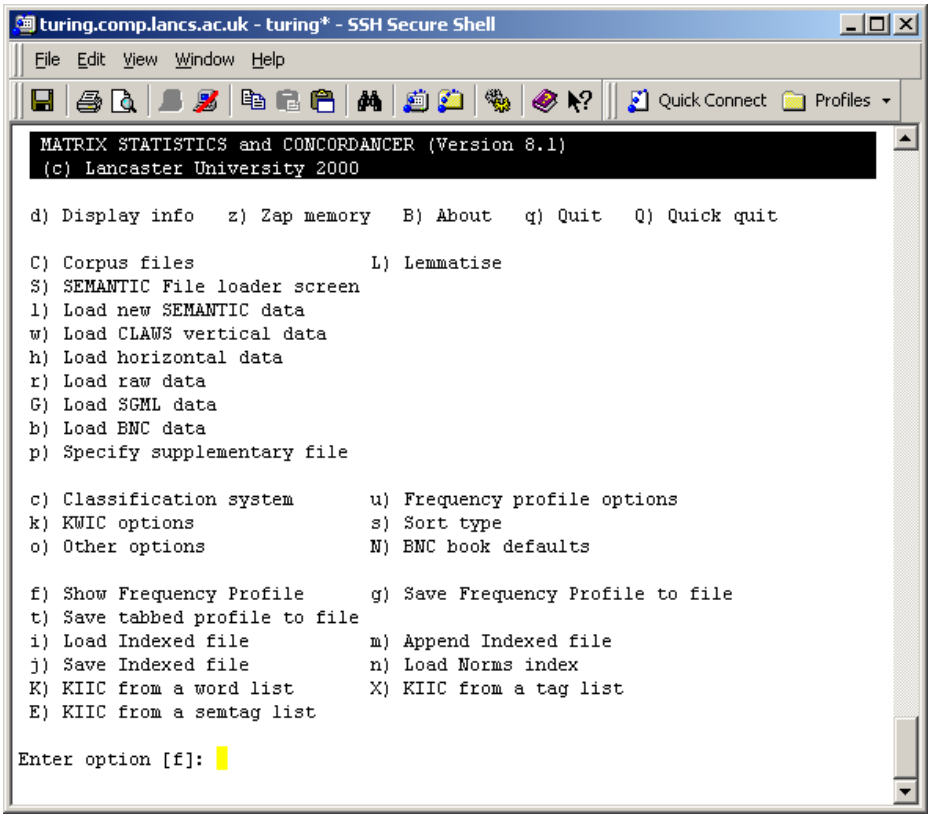


Figure 4.7 Screenshot of Tmatrix menu

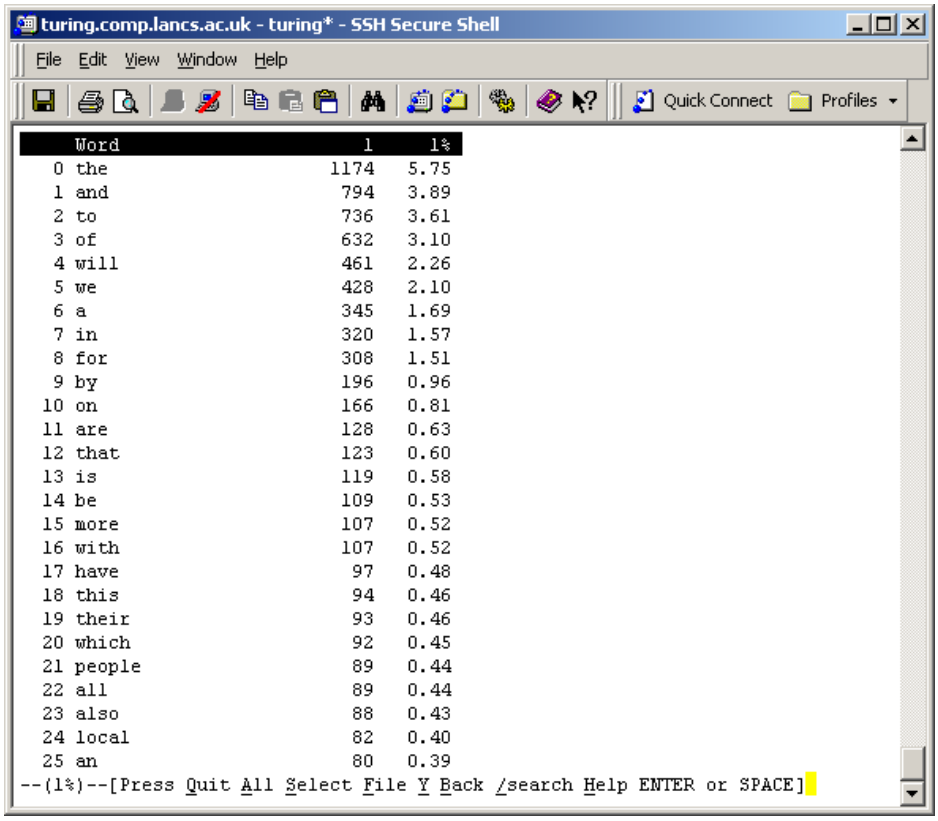


Figure 4.8 Tmatrix frequency profile for the LibDem data

Figure 4.9 Tmatrix concordance for the word 'our'

Other option screens are available to specify additional levels of annotation (POS or semantic tagging) to be displayed in the frequency profile and concordance, but those are not shown here. An indexed file can be saved from the main menu by typing ‘j’. This stores the internal Matrix index (described in section 4.2) in a persistent file which can be reloaded later to save time. The comparison of frequency profiles is achieved by loading in a previously saved index as a *text norm* by typing ‘n’ from the main menu. The text norm’s frequencies are then compared to any file loaded in the tool, whether as an index or directly. Figure 4.10 shows the semantic frequency profile comparison of the LibDem (column headed *1*) and Labour (column headed *Norm*) manifestos. This corresponds to the data shown in Table 4.5.

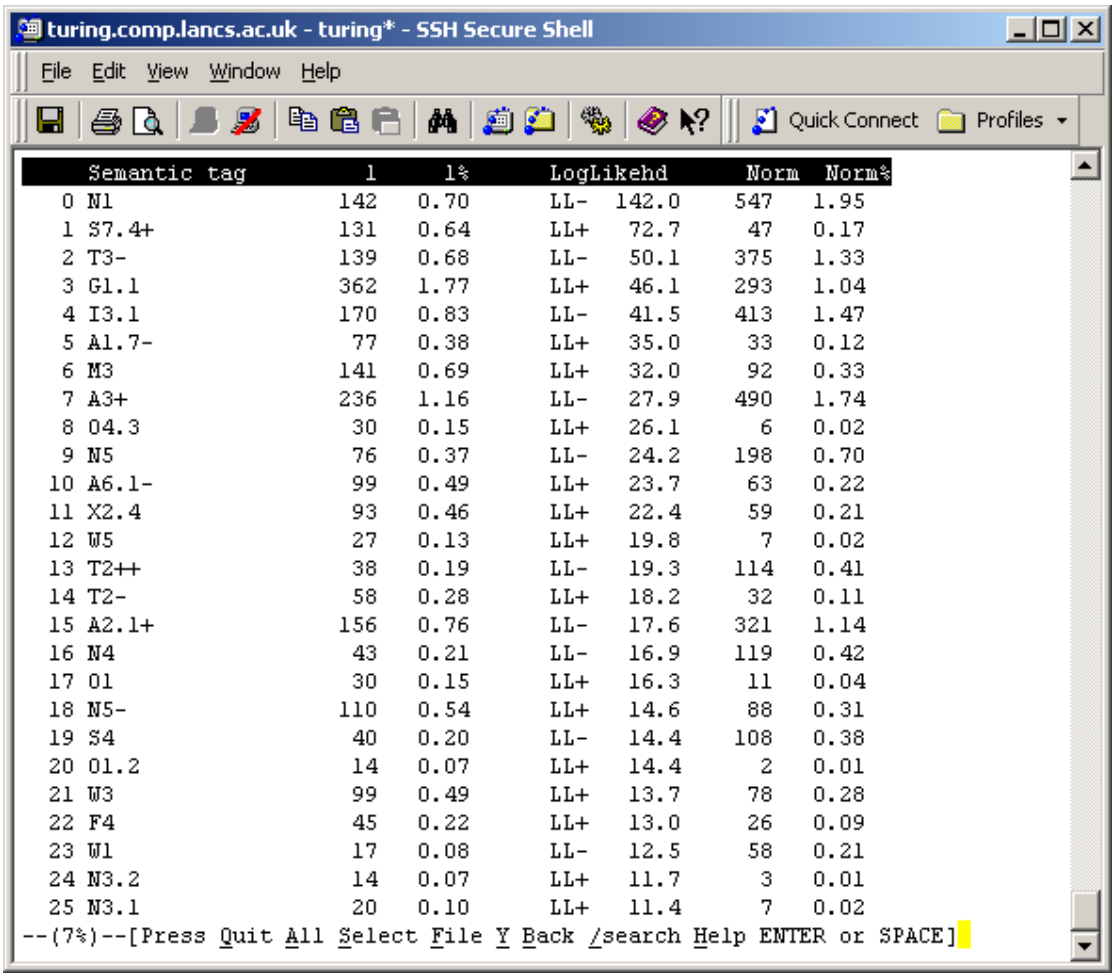


Figure 4.10 Tmatrix screenshot showing frequency profile comparison

The next stage in development of the Matrix tool was to change the interface to a graphical one, intended to be easier to learn for new users familiar with multiple windows and pull-down menus. We developed a graphical interface to be used for the X Window System with the OSF/Motif toolkit release 1.2 for X11R5 (see Heller et al, 1994). The initial graphical interface was built using Moguig, a Motif user interface builder developed at Lancaster University by Andy Colebourne (Colebourne et al, 1993). Figure 4.11 shows a screenshot of the X windows user interface (named Xmatrix). Operations are driven by menus and buttons, and they are controlled by mouse clicks. The main part of the window includes space for the frequency list. The one shown is for a text unrelated to the LibDem manifesto. Figure 4.12 shows a concordance for the POS tag VVD. Concordances are produced by double clicking on frequency lines in the main Xmatrix window.

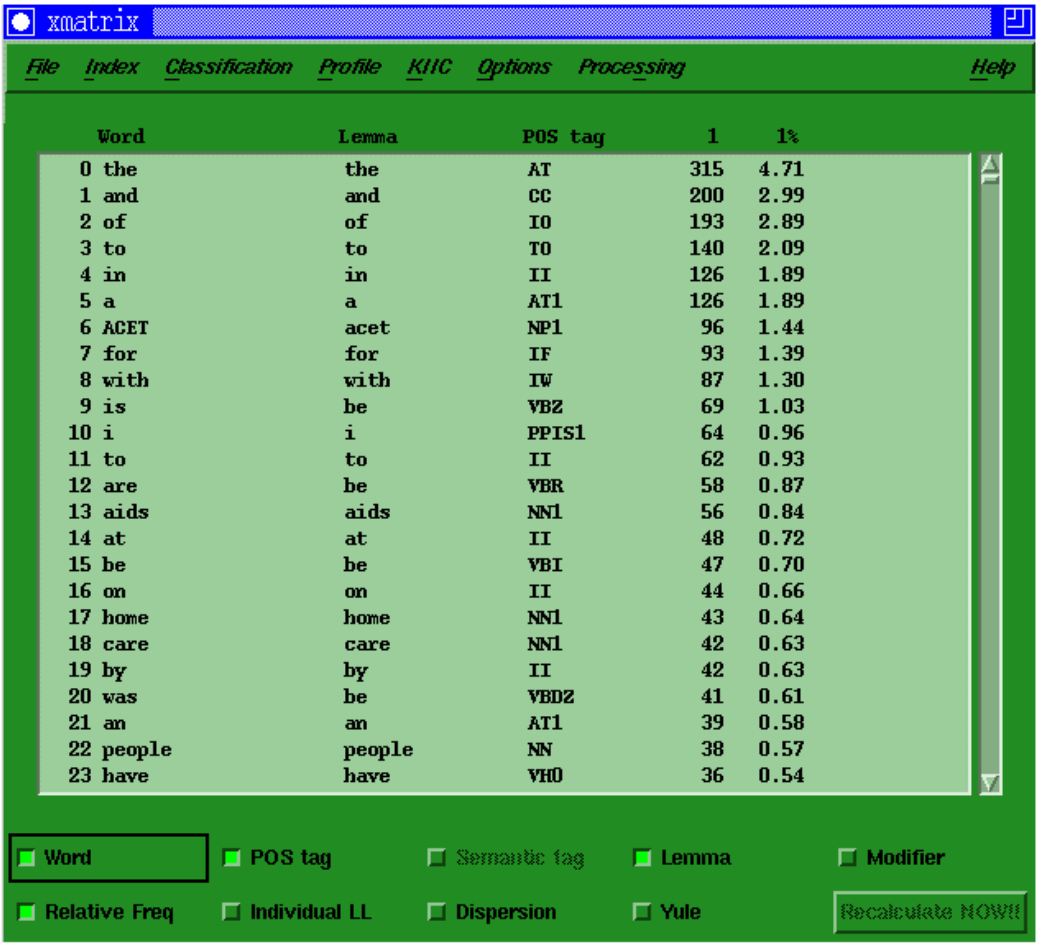


Figure 4.11 Xmatrix screenshot of main window

Figure 4.12 Xmatrix screenshot of the concordance window

The next and most recent major change in the user interface of the Matrix tool was driven by a change in the intended users of the system. Previously only experienced corpus researchers were expected to use the tool. We now intended to expand the use of the system so it could be used by systems engineers. The results of this work are described further in section 5.3.3. The new users would be unfamiliar with corpus linguistic tools and methods. It was also intended that the technicalities of the corpus annotation tools should be hidden from view in order to make the tool simple to use in extracting information from the texts being studied. We also wanted to make the resulting tool available to users who were unfamiliar with UNIX and the command line interface it provides. A very familiar interface to most users is the web browser and so we decided to build a web front-end to Matrix, called Wmatrix. All processing is done on the remote web server so users can gain access from any platform that provides a browser such as Netscape or Microsoft Internet Explorer.

A user of Wmatrix begins by uploading their corpus file to the web server via the web browser. Annotation tools available in Wmatrix include CLAWS (part-of-speech tagger), USAS (semantic field tagger) and LEMMINGS (a lemmatiser). Wmatrix provides the functions of Tmatrix and Xmatrix with production of frequency lists, statistical comparison of those lists, and KWIC concordances. The annotated output can be presented in a web browser from different viewpoints depending on the role taken by the user of the system, but the examples shown here will be from the *corpus linguist* viewpoint. This viewpoint presents the results as file icons in the browser and allows the most flexibility in the way the user operates on the data. The annotated file is presented to the user in a workarea along with word, POS and semantic tag frequency lists prepared by Wmatrix. These can be downloaded but can also be browsed using the web browser application. The user can click on a word or tag from the frequency lists and see a standard key word in context concordance for that item. This is prepared on the fly from the corpus on the web server.

Other (non-linguist) viewpoints produced for Wmatrix include *Revere*, *Summary*, and *Quality*. These were developed during the Revere project described in section 5.3.3. They provide immediate access to key word and key concept lists. They also include user defined filters to quantify relevant concepts. For example, in the *quality*

viewpoint to be used to examine standards documents, we included filters for modal verbs, power terms, optionality and definiteness. In the *Revere* viewpoint, intended for use with system requirements specifications, we defined filters for modal verbs, obligation and necessity concepts, and candidate roles taken by actors in the domain.

Figure 4.13 shows the web front end which we developed. The main area of the screen shows the existing workareas as folder icons. The bar with a shaded background (shown in light blue on screen) at the top of the window contains the menus listing the main operations provided in Wmatrix. The list of possible viewpoints is shown as a set of check boxes. Each viewpoint shows a different way of presenting the same data. The *tag wizard* function in the REVERE menu automates the annotation process using CLAWS and USAS, although each level of the annotation process can be activated manually as well.

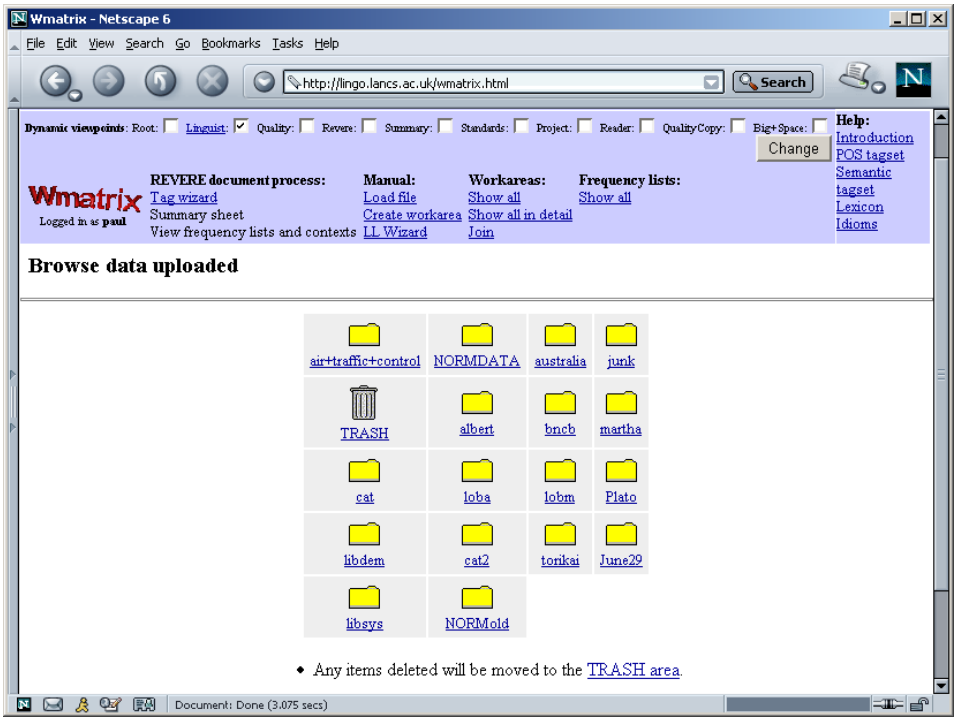


Figure 4.13 Wmatrix screenshot of the workareas

Figure 4.14 Wmatrix screenshot of libdem workarea

Figure 4.14 shows the user's view of the LibDem data once the tag wizard has been run. The contents of the workarea are presented with icons representing each file alongside the set of possible operations that can be carried out on each one. Frequency lists can be viewed by clicking on the *list* options. Figure 4.15 show the frequency list containing words and their associated POS tags. Concordances can be produced by clicking on the *context by* option in the workarea view or by clicking on the *context* options alongside the frequency list display. Figure 4.16 shows the concordance produced for the semantic tag *B3* (medicines and medical treatment). Larger amounts of context for each line can be seen by clicking on the *more* and *full* links.

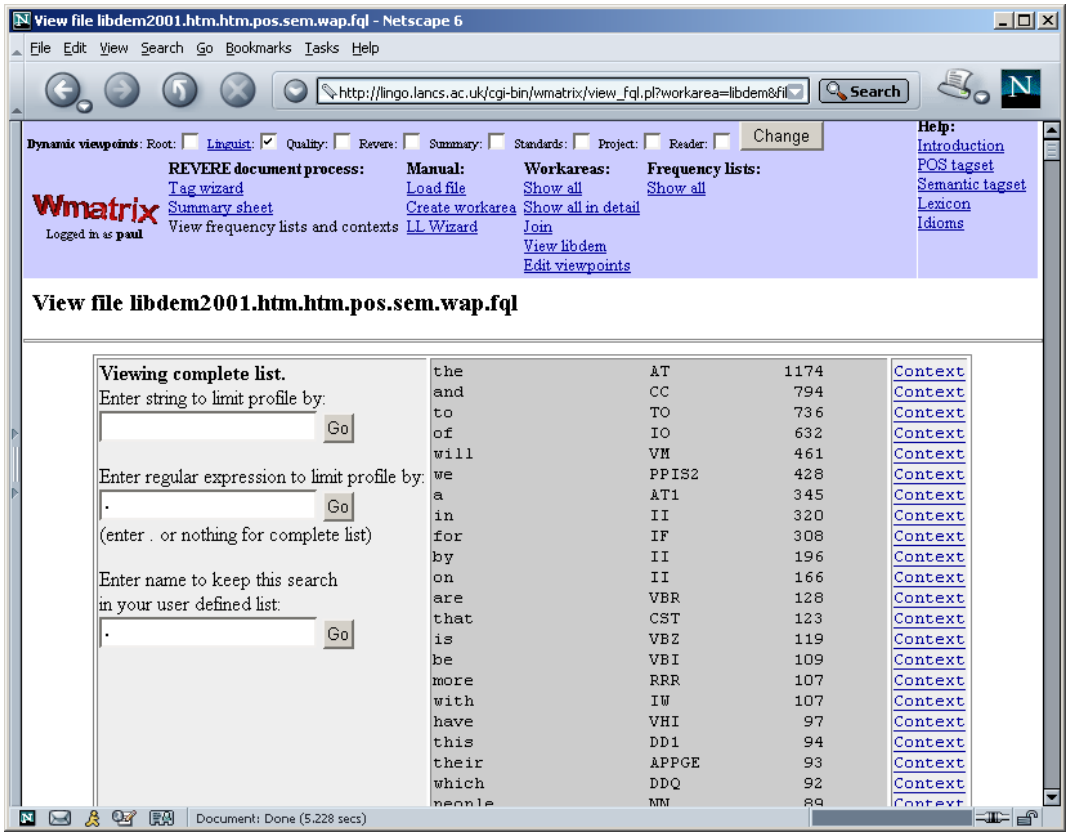


Figure 4.15 Wmatrix screenshot showing LibDem frequency list

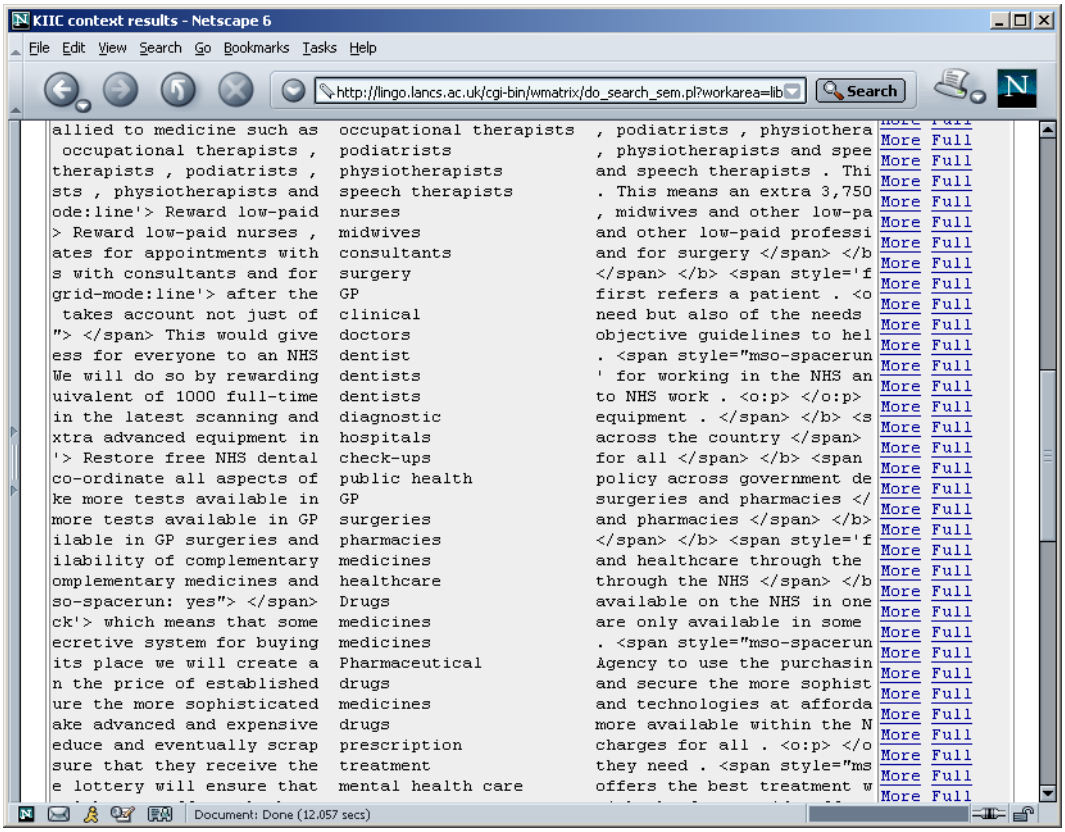


Figure 4.16 Wmatrix screenshot showing LibDem concordance

Comparison of frequency lists is achieved by clicking on the *compare to* options in the workarea view or selecting two frequency lists from the list produced from the *frequency list show all* menu item in the menu bar. Figure 4.17 shows the display when comparing the LibDem and Labour manifestos as before.

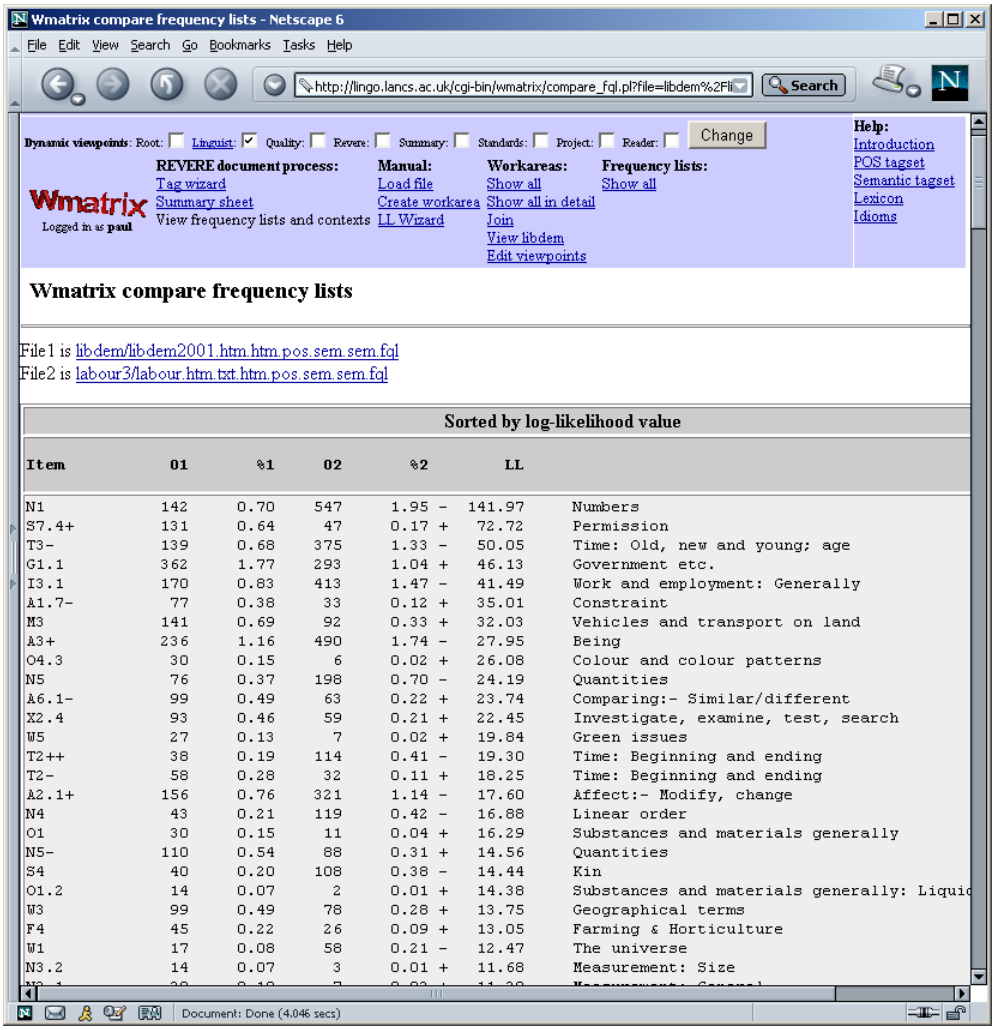


Figure 4.17 Wmatrix screenshot showing comparison of LibDem and Labour manifestos at the semantic level

4.5.2 Architecture of Wmatrix

Much of the program code has been reused from Tmatrix and Xmatrix in the creation of Wmatrix. These modules remain in the C programming language, but the web interface is built using HTML and the Perl programming language (version 5), (Wall et al, 1996). The Perl scripts reside on a Sun UNIX server running Solaris and the

Apache web server, and make use of the CGI (common gateway interface), (Gundavaram, 1996) to create web pages dynamically for the end user to browse.

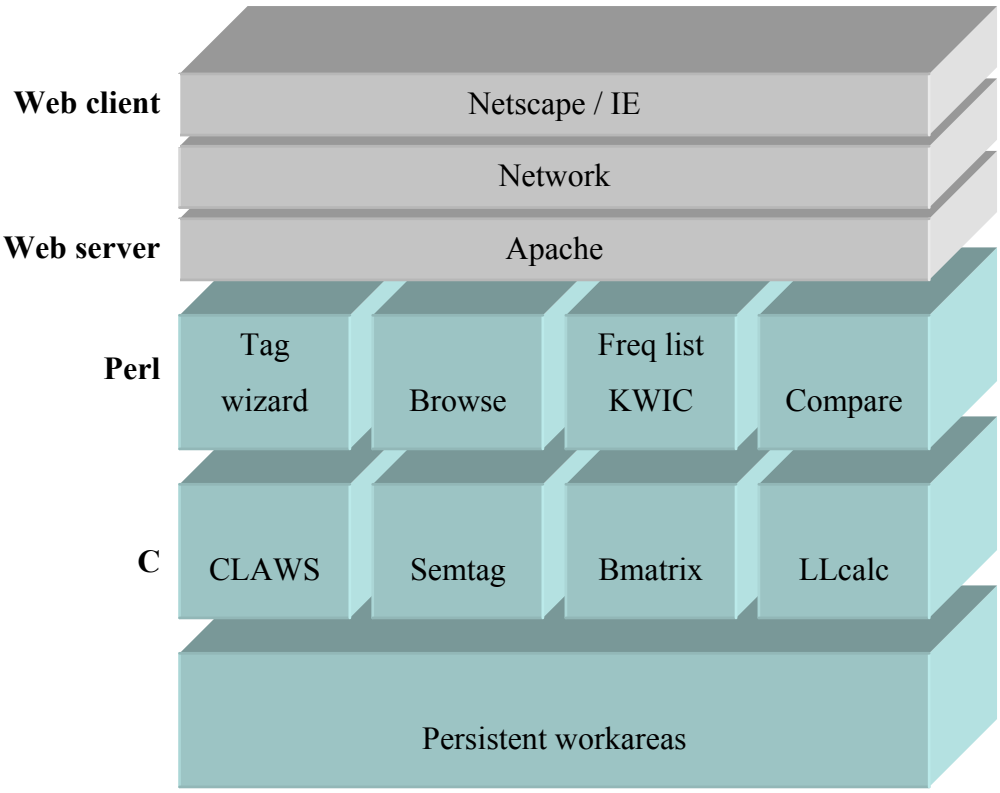


Figure 4.18 Architecture of Wmatrix

Figure 4.18 shows the major components of Wmatrix. Represented in the bottom three layers of the diagram, Wmatrix resides on the web server which the web browser client accesses via the network in the usual way. One of the C components of Wmatrix is shown as Bmatrix. This is intended to represent the component which encapsulates the frequency list production and concordancing functionality of Tmatrix and Xmatrix. Bmatrix runs as a UNIX command-line executable; the B stands for *batch* mode. Files stored in the users' workareas are pre-processed by the Perl scripts before being presented to the user in a web page. To record the user's identity, a small Javascript module runs during the login process to create a cookie stored in the web browser. Figure 4.19 shows the sequence of actions of the user and operations of Wmatrix when two word frequency lists are to be compared.

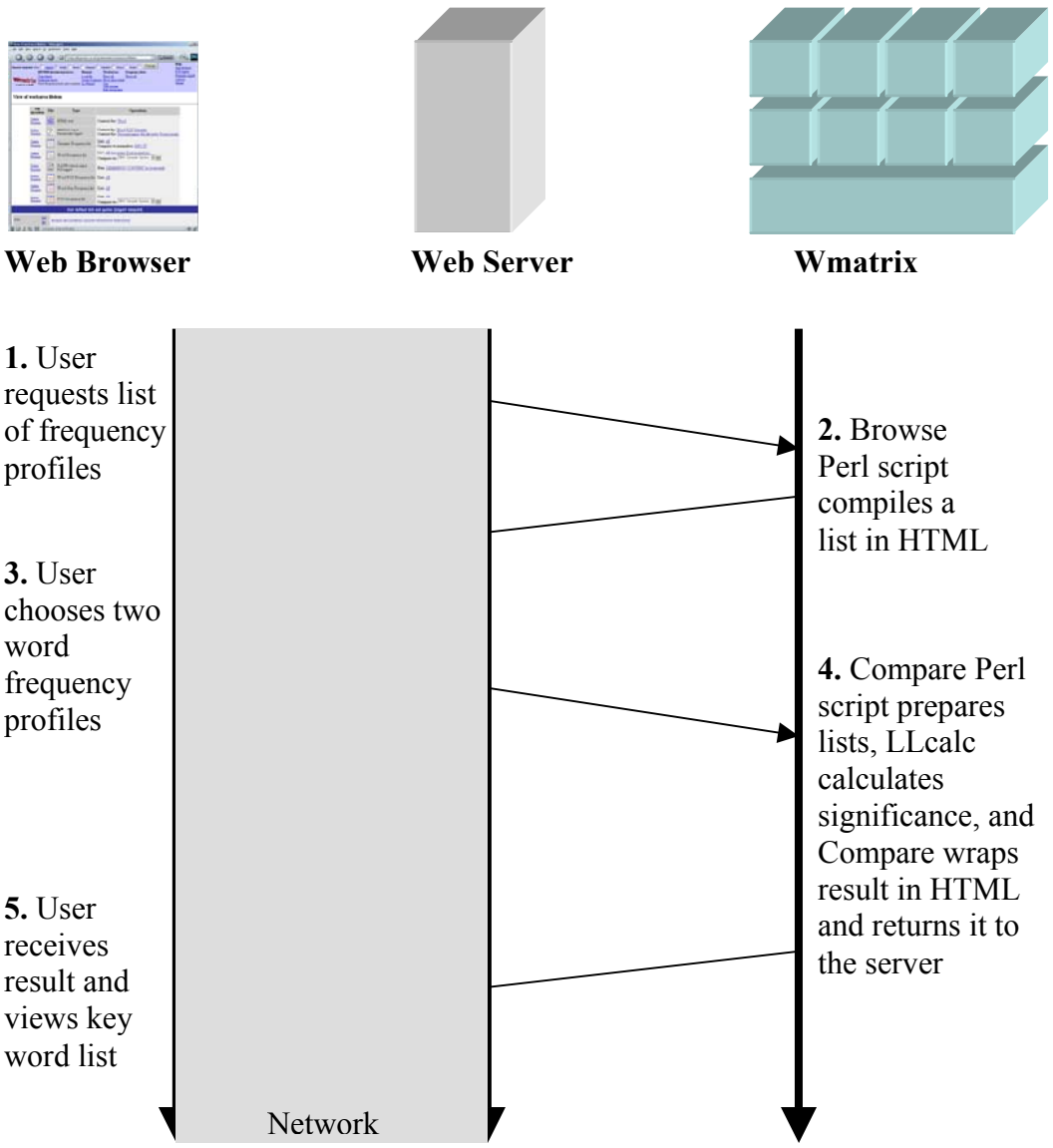


Figure 4.19 Wmatrix flow of operations when comparing frequency profiles

4.5.3 Further functions of Matrix

Although not central to the Matrix method described in this thesis, the Matrix tool[66] is capable of comparing annotated corpora in other ways. The application of the tool in these cases is in improving the quality of the computational annotation tools. This additional functionality can be described as parallel concordancing. This term usually applies to the concordancing of parallel corpora, which contain translation equivalents

[66] In this case, we refer to the X-windows version of the tool

in one or more languages of a collection of texts. In the Matrix tool, we can examine parallel concordances of the same text with different versions of the annotations. We present here three possible uses of this technique:

1. *Error rate calculation for a POS tagger*. Parallel versions of the same corpus, one POS tagged manually and one POS tagged automatically, to identify where the automatic tagging fails to match with the manual tagging.
2. *Training of an annotation tool*. Parallel versions of the same corpus, one automatically tagged as a benchmark ('correct') version and one automatically tagged with the latest version of a tagger (Smith, 1997: 146).
3. *Tagset and tagger comparison*. Parallel versions of the same corpus, one tagged with a rule-based POS tagger and one tagged with a probabilistic POS tagger, to enable mapping of one tagset to another and comparison of the performance of two POS taggers (Belmore, 1997).

4.6 Summary

This chapter has described in detail the Matrix method and the software tool that has been implemented to carry out frequency profiling of corpora, and comparison of those profiles across corpora. In order to suggest possible research questions for further investigation, the Matrix method uses the log-likelihood ratio statistic to compare frequencies and then rank them in terms of significant difference.

A worked example illustrated the method with two corpora consisting of UK 2001 General Election manifestos. We have seen that the method extends the keywords approach to key grammatical classes and key concepts. Key grammatical categories and semantic classes are used to group together lower frequency words and those words which would, by themselves, not be identified as key, and would otherwise be overlooked. Comparison at the POS and semantic levels reduces the number of key categories that the researcher should examine. The method can replace stage one of Leech and Fallon's (1992) process (quantitative extraction), and it assists in their stage two (qualitative examination).

5. Evaluation of Matrix

"There are three kinds of lies: lies, damn lies, and statistics"
(Attributed to Benjamin Disraeli and Samuel Clemens)

5.1 Introduction

In this chapter, we first comparatively evaluate the method from a statistical point of view. However, the evaluation is not sufficient by statistics alone. The application and results of the method and software tool are just as important. This method has already been applied to study social differentiation in the use of English vocabulary, profiling of learner English and semantic analysis of technical documents from the software engineering domain. Results of these applications are shown as case studies.

5.2 Statistical and comparative evaluation of the method

In section 2.7 we examined various statistics used to compare frequency data in corpus linguistics and other domains. We concluded that most of the formulae were not suitable for our purpose of comparing complete frequency lists. First we will compare two results from Francis and Kucera (1982: 544) using their normalised ratio (NR) value. In Table 5.1 we show that NR values cannot be compared in the same way as chi-squared and log-likelihood values can. The two example values for NR quoted in Francis and Kucera (1982: 544) are 1.36 and 1.32 which are quite similar relative to the range of other NR values. We might infer that the relative difference between 'being' in the informative and imaginative sub-corpora was of roughly the same interest to the corpus user as the difference between common noun frequencies in the same two sub-corpora.

Table 5.1 Comparison of NR and LL values

Tag	Informative corpus	Imaginative corpus	NR	Info.%	Imag.%	LL
common noun	182466	45332	1.36	24.0	17.6	3686.9
being form of verb be	546	140	1.32	0.07	0.05	9.1
Corpus size	761138	257613				

As we can see the NR value treats the relative frequency difference between common noun and being as the same since the ratio of values is roughly equivalent. We calculated the LL statistic for the same two examples as shown in the table. The LL statistic is over 3000 (3686.9) for the common noun comparison, and less than 10 (9.1) for the 'being' comparison. LL places more emphasis on the common noun difference of 6.4% (24.0 – 17.6) compared to 0.02% (0.07 – 0.05). This is a more useful result as the 6.4% accounts for a far greater portion of the difference between the two sub-corpora.

The second part of this section considers our application of frequency list comparison and looks at cases where the contingency table may become skewed. From section 2.7 there remained a question over the specific use of the chi-squared and log-likelihood tests in frequency profiling, although in general the log-likelihood test was preferred. In order to test the reliability of the chi-squared and the log-likelihood statistics we carried out a large number of *simulation* experiments (i.e. with simulated contingency tables).[67] For each experiment, there were three experimental conditions which determined the characteristics of the comparison between the corpora. These were:

1. *The ratio of the sizes of the two corpora*. We assume that one corpus is the normative corpus, and therefore the comparison corpus is less than or equal in size

[67] The GLIM4 statistical software package was used to do this (Francis, Green and Payne, 1993) but the simulations can be coded in any programming language with access to a reliable random-number generator.

to the normative corpus. Seven different ratio values were used, of 1:1, 2:1, 5:1, 10:1, 20:1, 100:1 and 200:1.

2. *The probability of the word occurring in text*. This probability was allowed to vary from 1 in 500 to 1 in 1,000,000 to reflect a good range observed in the BNC. To give context to these values, the probability of 'the' occurring in standard British English is 1/16, the probability of 'and' is 1/37 and the probability of 'reliable' is 1/44823 (estimated from the BNC). Ten different probability values were used; these are 1/500, 1/1000, 1/2000, 1/4000, 1/8000, 1/16000, 1/32000, 1/64000, 1/100000 and 1/1000000. We assumed that the probability of the target word was the same in the normative text and in the comparative text, as we wished to test the distribution of the test statistics X^2 and G^2 under the null hypothesis of no difference in probabilities.
3. *The size of the normative corpus*. This was allowed to vary from 100000, 500000, 1 million, 5 million, 10 million, 20 million, 50 million and 100 million words, taking eight different values. The BNC is at the top end of this range, with 100 million words.

We should note that our simulations could equally apply to frequency comparison of POS and semantic tags since the range of probabilities in our experiment adequately covers the range of values observed for tag frequencies (see Table 2.3 in section 2.6.2).

We assumed that the probability of the target word was the same in the normative text and in the comparative text. For each experiment, the size of the normative corpus C, the ratio of the sizes of the two corpora R, and the probability of word occurrence p were all fixed, and 10,000 simulated 2×2 tables of word frequencies were generated. Table 5.2 below shows the expected values of the 2×2 table. For each simulated table, the chi-squared test statistic X^2 and the log-likelihood test statistic G^2 were calculated, and at the end of each experiment the X^2 statistics were ranked in ascending order; the G^2 statistics were similarly ranked. The 95^{th}, 99^{th}, 99.9^{th} and 99.99^{th} percentiles were then determined. The experiment was replicated one hundred times to obtain empirical estimates of the standard deviation of the percentiles.

Table 5.2 Expected values of the 2×2 tables considered in the experiment

	Normative corpus	Comparative corpus	Total
No. of target words	pC	pRC	p(1+R)C
Number of non-target words	(1-p)C	(1-p)RC	(1-p)(1+R)C
Total number of words	C	RC	(1+R)C

Finally, for each experiment, we determined whether the usual Cochran (1954) rule would determine if the simulated values were unreliable. As the rule is based on expected values, and not actual simulated values, this could be determined once at the beginning of each experiment. From Table 5.2, the smallest expected value is PRC (as $P < 1$ and $R \leq 1$ by design). Thus, if PRC was less than 5, the Cochran rule was held to be true.

A total of 560 (7×10×8) experiments were carried out, covering each possible combination of the three experimental factors. With one hundred replicates of each experiment, the entire study generated 560,000,000 (10,000×100×560) simulated 2×2 tables.

5.2.1 Results and discussion.

Table 5.3 shows some typical output from the simulation experiments, giving the results for the 95th percentile of the X^2 test statistic where p=1/16000. For each combination of normative corpus size C and ratio R, the table entries give the mean and standard deviation (in brackets) of the 100 simulated values of the test statistic. We define the test statistic to be accurately represented by the chi-squared distribution if the 95th percentile of the chi-squared distribution on 1 degree of freedom (3.840) is contained in the interval defined by the mean plus or minus two standard deviations. Cells of the table in bold font show where this condition does not hold. It can immediately be observed that the test statistic is accurate in most of the table, but not

in the top left hand corner. This corner is characterised by small expected values in the 2×2 table, and examination of Table 5.4, which gives the smallest expected value, indicates that the Cochran condition provides a good guide to accuracy in this case. Cells of the table in bold font show where the Cochran rule is true.

Table 5.3 Means and standard deviations of the 95th simulated percentile of the chi-squared test statistic under independence for various 2×2 tables with p=1/16000[68]

p=1/16000	Ratio R						
Corpus size C	**1 / 200**	**1 / 100**	**1 / 20**	**1 / 10**	**1 / 5**	**1 / 2**	**1 / 1**
100,000	**0.060**	**8.772**	3.718	4.068	**3.410**	**3.528**	3.752
	(0.000)	**(0.664)**	(0.206)	(0.192)	**(0.098)**	**(0.069)**	(0.062)
500,000	**5.0945**	**2.615**	**3.400**	**3.605**	3.763	3.836	3.831
	(0.070)	**(0.100)**	**(0.119)**	**(0.037)**	(0.074)	(0.081)	(0.070)
1,000,000	**2.253**	**3.505**	**3.523**	3.736	3.805	3.828	3.823
	(0.069)	**(0.064)**	**(0.055)**	(0.059)	(0.070)	(0.081)	(0.074)
5,000,000	**3.707**	**3.324**	3.801	3.811	3.831	3.850	3.845
	(0.048)	**(0.034)**	(0.075)	(0.075)	(0.068)	(0.076)	(0.084)
10,000,000	**3.244**	3.739	3.813	3.834	3.840	3.842	3.837
	(0.019)	(0.101)	(0.067)	(0.065)	(0.079)	(0.065)	(0.082)
20,000,000	3.720	3.752	3.813	3.842	3.851	3.847	3.854
	(0.073)	(0.077)	(0.073)	(0.076)	(0.070)	(0.070)	(0.078)
50,000,000	3.773	3.809	3.829	3.833	3.849	3.836	3.845
	(0.047)	(0.071)	(0.075)	(0.072)	(0.081)	(0.074)	(0.080)
100,000,000	3.820	3.826	3.835	3.843	3.844	3.845	3.831
	(0.068)	(0.072)	(0.071)	(0.067)	(0.074)	(0.071)	(0.077)

[68] Standard deviations are in parentheses. Cells where the 95% critical value of the chi-squared distribution on 1 degree of freedom (3.84) lies outside interval defined by the mean plus or minus two standard deviations are defined as inaccurate and shown in bold.

Of course, Table 5.3 is a small portion of the output generated. Similar tables were generated for G^2 and X^2; for the 99th, 99.9th and 99.99th percentiles as well as the 95th percentile, and for each of the ten values of the proportion p. Table 5.5 summarises the results into a single display. The table consists of an array of ten rows and five columns. The first four columns show the accuracy of the tests at the 5%, 1%, 0.1% and 0.01% significance level, and the final column the standard Cochran rule. Each of the ten rows represents a different proportion. Within a cell of the array, an 8×7 grid shows the simulated accuracy. Rows of each grid represent the eight corpus sizes in ascending order, and the columns of the grid represent the seven ratios, again in ascending order. The symbols used are: ■ both tests inaccurate; ◘ chi-squared test inaccurate; ▣ likelihood ratio test inaccurate; □ both tests accurate. A ✖ indicates that the smallest expected value of the generated table is less than 5. The critical values we used were 3.84 (5% level), 6.63 (1% level), 10.83 (0.1% level) and 15.13 (0.01% level which is not usually listed in published tables).

Table 5.4 Smallest expected values in the 2×2 tables when p=1/16000

p=1/16000							
Corpus size	**Ratio R**						
C	**1 / 200**	**1 / 100**	**1 / 20**	**1 / 10**	**1 / 5**	**1 / 2**	**1 / 1**
100,000	**0.031**	**0.063**	**0.313**	**0.625**	**1.250**	**3.125**	6.250
500,000	**0.156**	**0.313**	**1.563**	**3.125**	6.250	15.625	31.250
1,000,000	**0.313**	**0.625**	**3.125**	6.250	12.500	31.250	62.500
5,000,000	**1.563**	**3.125**	15.625	31.250	62.500	156.250	312.500
10,000,000	**3.125**	6.250	31.250	62.500	125.000	312.500	625.000
20,000,000	6.250	12.500	62.500	125.000	250.000	625.000	1250.000
50,000,000	15.625	31.250	156.250	312.500	625.000	1562.500	3125.000
100,000,000	31.250	62.500	312.500	625.000	1250.000	3125.000	6250.000

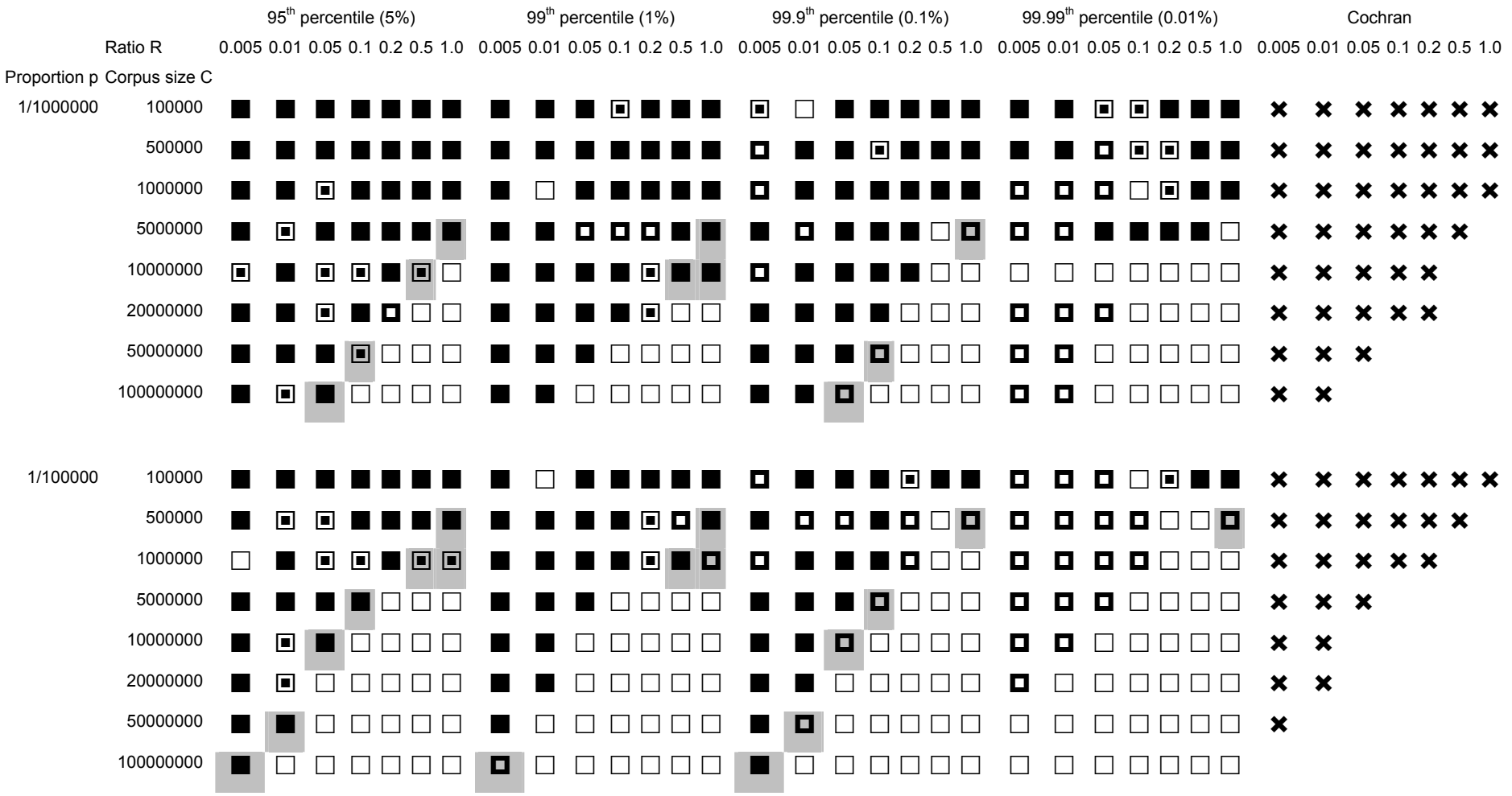
It can be seen from Table 5.5 that the overall pattern in each 8×7 grid in similar to that in Table 5.3. The chi-squared and likelihood ratio test statistics are accurate in much of each grid apart from the top left hand corners. There is also an observable trend as we move down the ten rows from the top to the bottom of the overall table. The trend

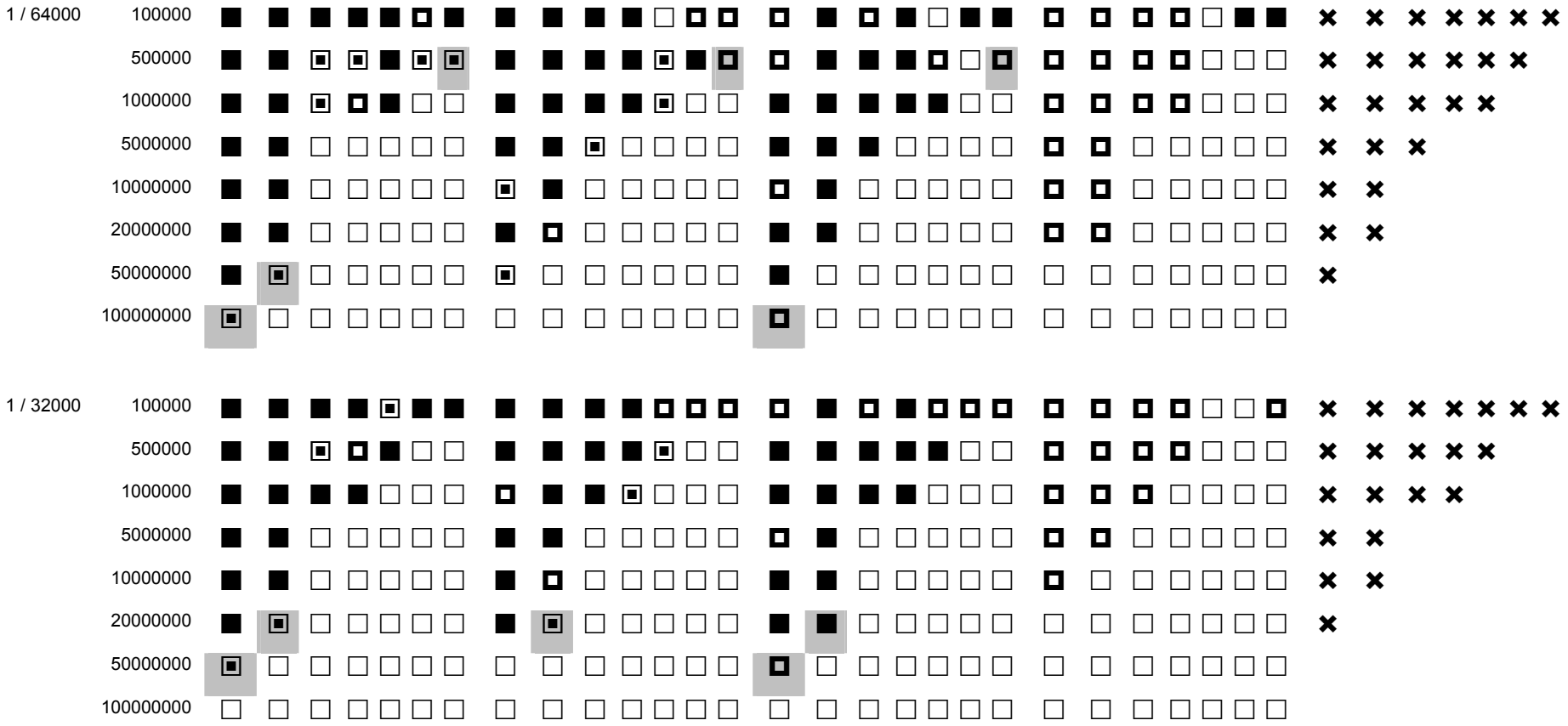
moving down the overall table is for fewer cells in each 8×7 grid to be marked indicating fewer inaccurate tests as the proportion increases. A trend moving from left to right (from 5%, 1% to 0.1% and 0.01%) in the overall table is less easy to detect. For all the proportions, there are slightly fewer inaccurate tests observed at the 99.99th percentile (0.01%) than at the 95th percentile (5%) level.

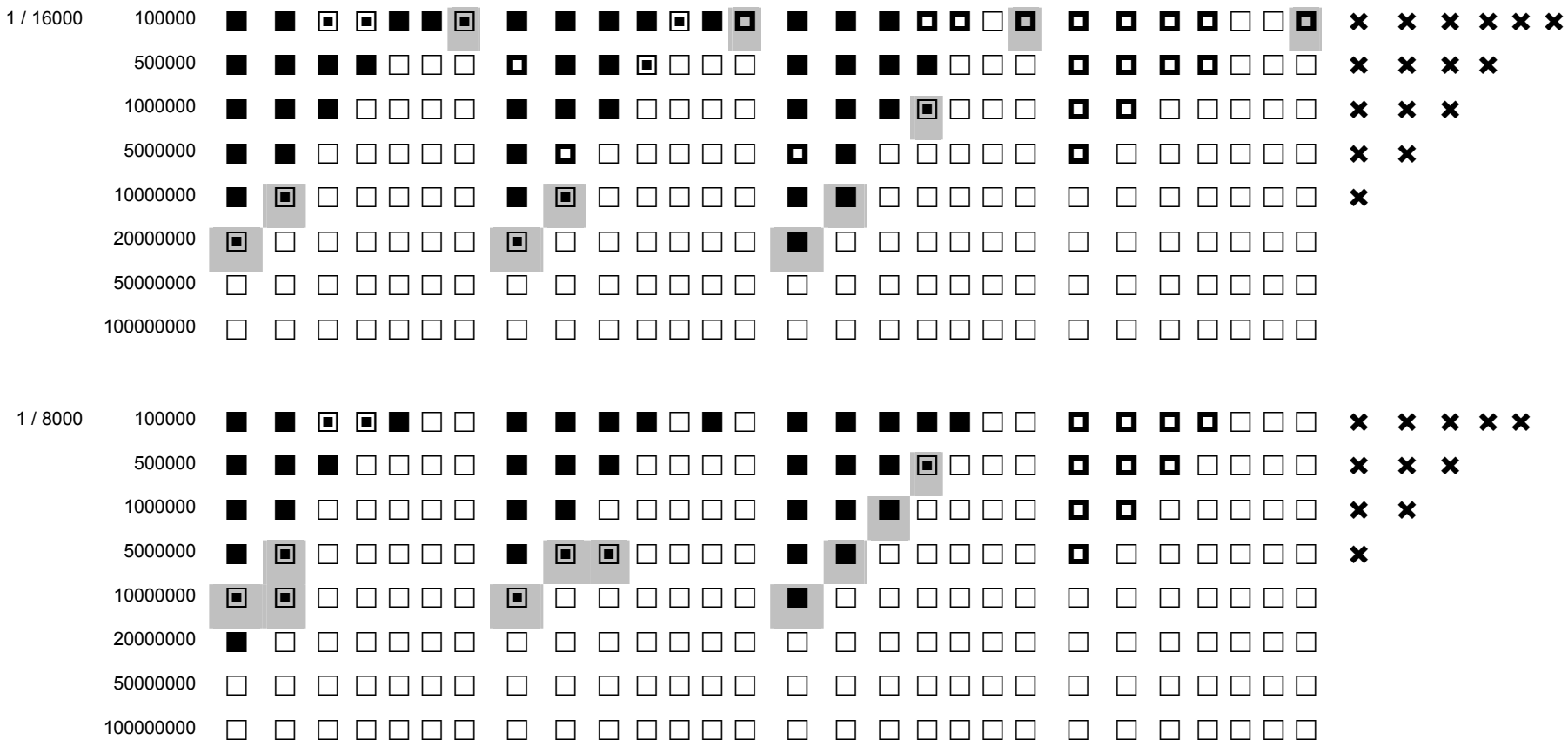
Let us now contrast the accuracy of the two tests. When one test rather than both is inaccurate, and this occurs in 246 grid cells, the likelihood ratio test is inaccurate 81 times, and the chi-squared test is inaccurate 165 times. There is an overall trend from left to right which distinguishes the two tests. It is most clear at the 0.01% level in which, with the exception of 25 cells, the likelihood ratio is accurate in all the cells. This compares with 120 cells where the chi-squared test is inaccurate. Examining the data underlying the results at the 0.01% level in particular, the simulations for the likelihood ratio showed much smaller standard deviations than for the chi-squared test, as well as the mean statistic for the chi-squared test being skewed. The likelihood ratio test remains accurate for very heavily skewed tables at the 0.01% level, even for example, in one of the cells in the first grid at the 0.01% level represents a simulation of a word occurring once in a 1,000,000 word corpus as compared to the word occurring 100 times in a 100 million word corpus. The cells are the top right of the grids at the 0.01% level are false positives since the simulated critical values fall below the listed one.

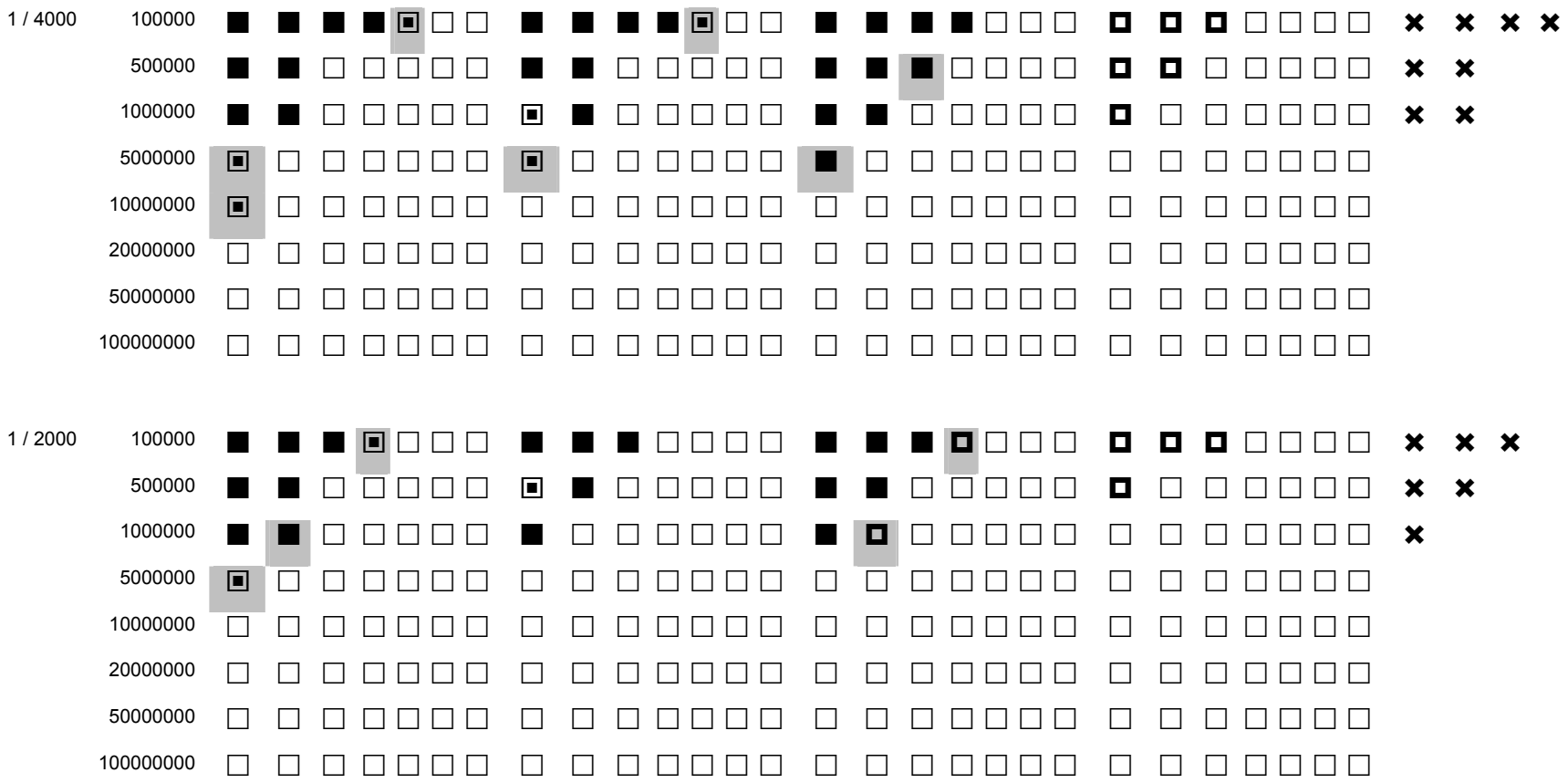
If for each row we compare the four grids showing the accuracy of the tests to the grid showing the Cochran rule, we can see that the Cochran rule provides a good guide to accuracy in most cases. However, there are some cases (marked with a shaded background) which show inaccurate tests that are not covered by the Cochran rule. These are cases where the smallest expected value in the generated table is 5 or greater. The cells in question show inaccuracy in both chi-squared and likelihood ratio tests. At the 5% level the largest expected value showing an inaccurate test is 12.5. At the 1% level we observe a value of 10, at the 0.1% level it is 7.8125 and at the 0.01% level the largest expected value coinciding with an inaccurate test is 6.25. This suggests that the Cochran rule needs to be extended.

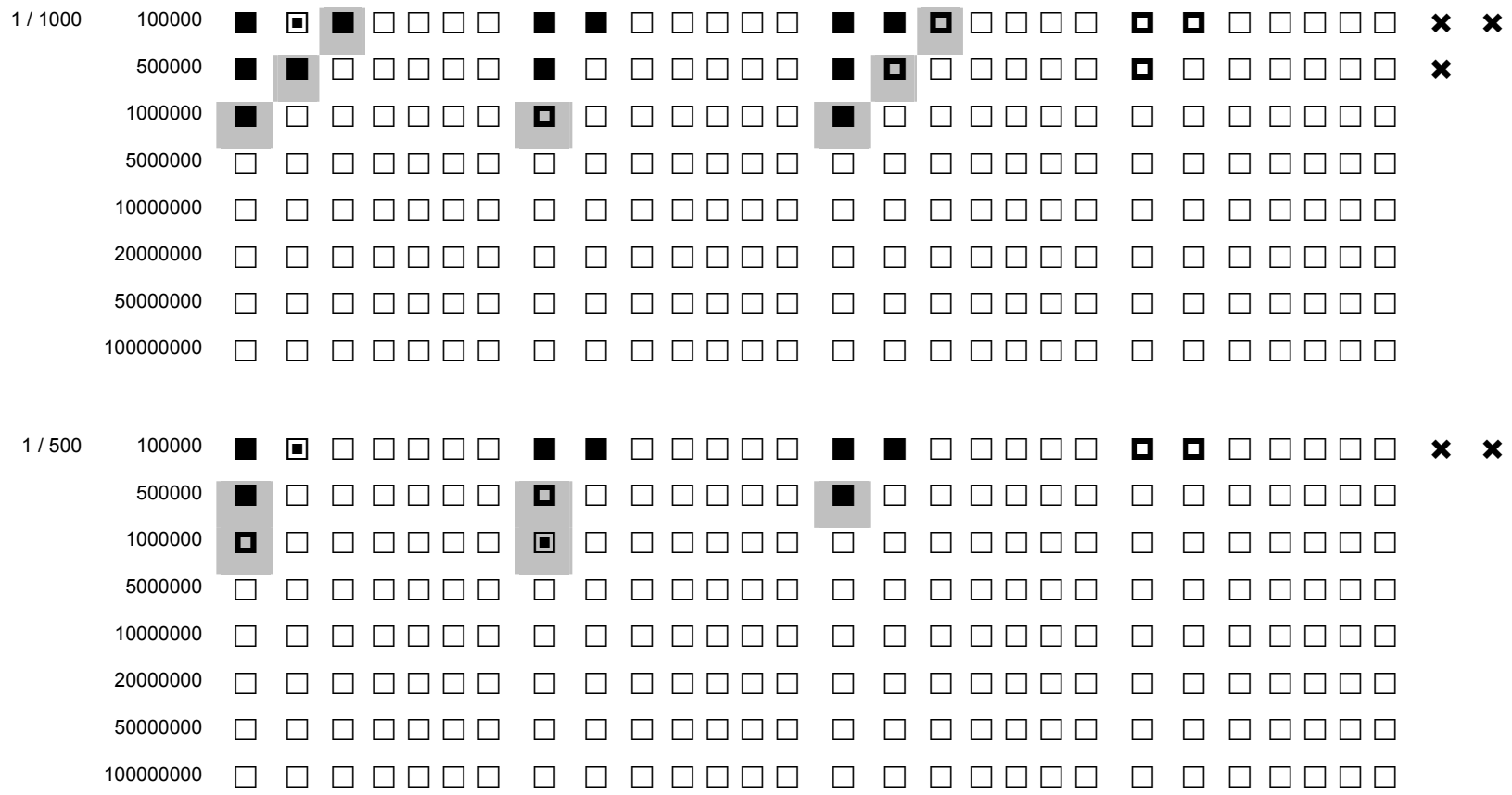
Table 5.5 Accuracy of chi-squared and likelihood tests at 5%, 1%, 0.1 and 0.01% levels, for various values of size, proportion and ratio




1 / 64000
100000
500000
1000000
5000000
10000000
20000000
50000000
100000000
1 / 32000
100000
500000
1000000
5000000
10000000
20000000
50000000
100000000


1 / 16000
100000
500000
1000000
5000000
10000000
20000000
50000000
100000000
1 / 8000
100000
500000
1000000
5000000
10000000
20000000
50000000
100000000


1 / 4000
100000
500000
1000000
5000000
10000000
20000000
50000000
100000000
1 / 2000
100000
500000
1000000
5000000
10000000
20000000
50000000
100000000


1 / 1000
100000
500000
1000000
5000000
10000000
20000000
50000000
100000000
1 / 500
100000
500000
1000000
5000000
10000000
20000000
50000000
100000000

The statistics have more evidence from larger corpora and can therefore detect smaller differences in frequency. It is a feature of the tests that values are greater in larger corpora. So, for example, a X^2 value of 500 obtained from a 1 million versus 9 million word normative test is not comparable with the same X^2 value of 500 obtained from a 10,000 versus 90,000 word comparison. This does not mean the tests are flawed, but that we must be careful when considering their results.

5.2.2 Conclusion to the statistical comparison

There are several conclusions we can draw from our experiments. The statistical tests are accurate for the most part with various combinations of corpus size, word probability and ratio of corpora. From the point of view of both of the statistical tests there are no problems comparing unbalanced sized corpora as long as we avoid low expected values in the contingency table. At the 5% level, the Cochran rule should be extended to ensure expected values are 13 or more. At the 1% level, the Cochran rule should be extended to ensure expected values are 11 or more. At the 0.1% level, the Cochran rule should be extended to ensure expected values are 8 or more. The usual Cochran rule is sufficient at the 0.01% level for the chi-squared test. However (ignoring false positives), we can safely lower the Cochran rule at the 0.01% level for the log-likelihood test to expected values of 1 or more. The trade-off is that the critical value is higher than at the usual 5% level at 15.13.

There is a difference between establishing statistical significance and practical significance. In carrying out tests of significance we should always bear in mind the other issues relevant to corpus comparison as listed in section 2.7 and reiterated in our method section 4.3 which are equally important as determining significance.

5.3 Evaluation of the Matrix tool in practice

The evaluation of the tool in practice will be shown by three case studies. In each of these studies the Matrix tool was used to highlight key items for further study. The first one shows the Matrix tool being used to investigate social differentiation via vocabulary studies. The second section shows the technique being extended to

grammatical analysis to perform a contrastive analysis of native and non-native speaker language. Thirdly, the Matrix tool is applied at a second level of annotation, namely that of semantic tagging. The application in this last case is to the semantic analysis of software engineering requirements documents.

5.3.1 Case study one: Vocabulary studies

This section describes a formative evaluation of the Matrix tool at the word level. It shows an early version of the method in which we used the chi-squared test rather than the log-likelihood value.

In Rayson et al (1997), selective quantitative analyses of the demographically sampled spoken English component of the BNC were carried out. This is a sub-corpus of about four and a half million words, in which speakers and respondents are identified by such factors as gender, age, social group and geographical region. Using the Matrix method, a comparison was performed of the vocabulary of speakers, highlighting those differences that are marked by a very high value of significant difference between different sectors of the corpus according to gender, age and social group.

Let us focus in this section on gender variation. Using the whole of the demographic sub-corpus material for which gender of speaker is indicated, we found that female speakers have a larger share of the corpus than male speakers according to a number of different measures. Firstly, there is small built-in bias in the corpus, in that 75 female respondents but only 73 male respondents were enlisted as volunteers to participate in the collection of data. In addition, female speakers overall took a larger share of the language collected, as shown by these figures in Table 5.6.

This leads to larger female than male representation in the corpus. For every 100 word tokens spoken by men in the demographic corpus, 151 were spoken by women. This illustrates the need to normalise frequency data when comparing corpora. The Matrix method does take sub-corpus size into account when performing its calculations. Using the method, we calculated the 25 most significant words showing overuse in

male and female speech in the demographic corpus, and these are shown in Table 5.7 and Table 5.8.

Table 5.6 Distribution of the BNC demographic subcorpus between female and male speakers

	Female Speakers	**Male Speakers**
Number of speakers	561	536
Number of turns	250,955	179,844
Number of words spoken	2,593,452	1,714,443
Number of turns per speaker	447.33	335.53
Number of words per turn	10.33	9.53

Table 5.7 Words most characteristic of male speech

Word	**Males**	**M %**	**Females**	**F %**	χ^2
fucking	1401	0.08	325	0.01	1233.1
er	9589	0.56	9307	0.36	945.4
the	44617	2.60	57128	2.20	698.0
yeah	22050	1.29	28485	1.10	310.3
aye	1214	0.07	876	0.03	291.8
right	6163	0.36	6945	0.27	276.0
hundred	1488	0.09	1234	0.05	251.1
fuck	335	0.02	107	0.00	239.0
is	13608	0.79	17283	0.67	233.3
of	13907	0.81	17907	0.69	203.6
two	4347	0.25	5022	0.19	170.3
three	2753	0.16	2959	0.11	168.2
a	28818	1.68	39631	1.53	151.6
four	2160	0.13	2279	0.09	145.5
ah	2395	0.14	2583	0.10	143.6
no	14942	0.87	19880	0.77	140.8
number	615	0.04	463	0.02	133.9
quid	484	0.03	339	0.01	124.2

one	9915	0.58	12932	0.50	123.6
mate	262	0.02	129	0.00	120.8
which	1477	0.09	1498	0.06	120.5
okay	1313	0.08	1298	0.05	119.9
that	31014	1.81	43331	1.67	114.2
guy	211	0.01	95	0.00	108.6
da	459	0.03	338	0.01	105.3
yes	7102	0.41	9167	0.35	101.0

Table 5.8 Words most characteristic of female speech

Word	**Males**	**M %**	**Females**	**F %**	χ^2
she	7134	0.42	22623	0.87	3109.7
her	2333	0.14	7275	0.28	965.4
said	4965	0.29	12280	0.47	872.0
n't	24653	1.44	44087	1.70	443.9
I	55516	3.24	92945	3.58	357.9
and	29677	1.73	50342	1.94	245.3
to	23467	1.37	39861	1.54	198.6
cos	3369	0.20	6829	0.26	194.6
oh	13378	0.78	23310	0.90	170.2
Christmas	288	0.02	1001	0.04	163.9
thought	1573	0.09	3485	0.13	159.7
lovely	414	0.02	1214	0.05	140.3
nice	1279	0.07	2851	0.11	134.4
mm	7189	0.42	12891	0.50	133.8
had	4040	0.24	7600	0.29	125.9
did	6415	0.37	11424	0.44	109.6
going	3139	0.18	5974	0.23	109.0
because	1919	0.11	3861	0.15	105.0
him	2710	0.16	5188	0.20	99.2
really	2646	0.15	5070	0.20	97.6
school	501	0.03	1265	0.05	96.3
he	15993	0.93	26607	1.03	90.4
think	4980	0.29	8899	0.34	88.8
home	734	0.04	1662	0.06	84.0
me	5182	0.30	9186	0.35	83.5

Perhaps the most notable (though predictable) finding illustrated in these tables is the tendency for swear words to be more characteristic of male speech than female speech. Pronouns were investigated further as a result of the comparison at POS tag level, as shown in Table 5.9. Females make significantly greater use of the feminine pronoun *she/her/hers* and also of the first-person pronoun *I/me/my/mine*.

Table 5.9 POS as percentage of word tokens

	Males %	**Females %**	**$\chi 2$**
Pronouns	13.37	14.55	1016.27
Verbs	20.30	21.52	721.51
Common Nouns	8.49	7.93	395.18
Proper Nouns	1.44	1.64	257.78

After further investigations, it was notable that the distribution of taboo vocabulary was highly significant along all three dimensions of gender, age and social group. We found that the archetypal user of swearwords is to be found among male speakers in the social range C2/D/E under the age of 35. Our first finding was rather unsurprising, but it did confirm that the Matrix method was capable of being used in studying lexical variation across sub-corpora of the BNC.

Kilgarriff (2001) repeated our experiment using the Mann-Whitney test and noted that our results showed a bias towards high frequency words. Such a bias is not surprising since we are focussing on overused items in each case. It would be possible to reconfirm our results using the LL statistic although we would expect little change in the selection of words in the above tables since the chi-squared and LL statistics are similar for high frequency words. As noted in section 2.7.1, there are disadvantages with using the Mann-Whitney test. Due to problems of too many zeros in the Mann-Whitney test, Kilgarriff (2001) reports that his technique omits words with less than 20 occurrences in the combined (male-female) corpus. This amounts to ignoring 83% of the word types in the demographic corpus. Low frequency words are worth investigating (usually as underused items) if they are shown by the Matrix method to be significant, as well as high frequency words.

5.3.2 Case study two: Grammatical analysis of learner corpora

In Granger and Rayson (1998), two similar-sized corpora of native and non-native writing were first compared at the lexical level using Matrix. The corpora were analysed by the CLAWS part-of-speech tagger (described in section 3.2.1), and this permitted a Matrix comparison at the major word-class level.

The non-native speaker corpus is taken from the International Corpus of Learner English (ICLE) database (Granger, 1998). It consists of argumentative essay writing by advanced French-speaking learners of English. The control corpus of similar writing is taken from the Louvain Corpus of Native English Essays (LOCNESS) database.[69] For this study, some of the POS tags were grouped together, generally using the first letter of the CLAWS tag that represents the major word class.

Figure 5.1 displays the distribution of the nine major word categories in the native and non-native corpora. Three categories prove to have similar frequencies in the two corpora: articles (AT), adjectives (J) and verbs (V). But the non-native speaker (NNS) writers overused three categories significantly: determiners (D), pronouns (P) and adverbs (R), and also significantly underused three: conjunctions (C), prepositions (I) and nouns (N).

Not unexpectedly, this type of profile raises more questions than it answers. Aside from the question of whether overall similarity of frequency may conceal individual differences, there are questions relating to the over- and underused groups. To answer these questions, it is necessary to look both at the grammatical subcategories and the lexical items they contain. In order to determine significant patterns of over- and underuse, we produced profiles for lemmas in each major word category and sub-category and sorted them in decreasing order of significance. The most significant findings resulting from the comparison of word categories and lemmas in the two

[69] The non-native speaker corpus consists of c. 280,000 words of formal writing (both argumentative essays on general topics and literature exam papers) by advanced EFL university students of French mother-tongue background. The native speaker corpus consists of c. 230,000 words of similar writing by British and American university students.

corpora are summarised in Table 5.10. The table contains only items that are either significantly over- or underused, not those with similar frequencies.

Table 5.10 Patterns of over- and useruse in the NNS corpus

	Overuse	Underuse
AT	*a*	*the*
D	**most indefinite determiners**	
	all, some, each, a few, another	*many*
P	**most indefinite pronouns**	
	everybody, nobody, one, oneself, something, everything, a bit, a lot, lots	*no-one, no, anyone, everyone someone*
	first and second personal pronouns	
CC	*but, or*	*and*
CS	**some complex subordinators**	**most subordinators**
	as far as, as soon as, even if	*until, after, before, when, (al)though, while, whilst, whether (or not)*
I		**most prepositions**
	between, towards, without, above, during, of, on, about, before, among	*for, over, throughout, upon, into, along, out, despite, regarding,*
	in spite of, in front of, thanks to, by means of, till	*per, including, by, off, after, to, amongst, until, up, than*
RP		**most adverbial particles**
RR	**short adverbs of native origin (especially place and time)**	**-ly adverbs**
N		**overall underuse of nouns**
V	**auxiliaries**	*-ing* and *-ed* **participles**
	infinitives	

In the French learner corpus, the indefinite article *a* is overused and the definite article *the* underused. This proportionally higher use of indefinites by the NNS writers suggests that they are conforming less to the norms of formal writing. In his analysis of word frequencies in the LOB corpus, Johansson (1985: 30) notes that 'category J (learned texts), which has the highest frequency of the definite article, has the lowest

frequency of the indefinite article'. These results also demonstrate that an analysis based on major word categories alone, such as that represented in Figure 5.1, can be very misleading since in the case of articles, it showed no difference between the native and non-native corpus. This shows that we should avoid conflating categories if possible as Everitt (1992: 41) warns.

The French learners significantly overuse most indefinite determiners and pronouns. A high frequency of such words has been found to be favoured in speech and disfavoured in formal writing. Devito (1966, 1967) notes that speech has more indefinite quantifying words, while Johansson (1978: 11, 27) points at the low frequency of indefinite pronouns ending in *–thing/-one/-body* in academic English.

There is also a very significant overuse in the learner corpus of the first and second personal pronouns. Variability studies associate this feature with the involved nature of speech and point to the low frequency of indices of personal reference in academic writing (Biber 1988, Rayson et al. 1997).

The general underuse of conjunctions brought out by Figure 5.1 conceals a complex situation. It is striking to note that concessive subordinators (e.g. *(al)though*, *while, whereas*) that, according to Altenberg (1986: 18) are more prevalent in writing, are significantly underused by learners. It is also noteworthy that the two subordinators which are usually associated with speech, namely *if* and *because,* are not underused by learners, unlike most of the other subordinators.

The learner writers underuse the category of prepositions. According to Rayson et al. (1997) use of prepositions differs more than for most other categories between speech and writing. A high proportion of prepositions is associated with the informative and nominal tendency of written language.

Johansson (1985: 30) contrasts the nominal style of informative prose with the verbal style of imaginative prose. Svartvik and Ekedahl's (1995) study equally links up a lower density of nouns with the category of imaginative texts and conversations. The overall underuse of nouns that characterises French learner argumentative writing is thus clearly a further sign of a tendency towards oral style.

Though the overall frequency of verbs is similar in learner and native texts, there are considerable differences in the verbal forms used. The most striking feature is the overuse of auxiliaries, a characteristic of conversational English.

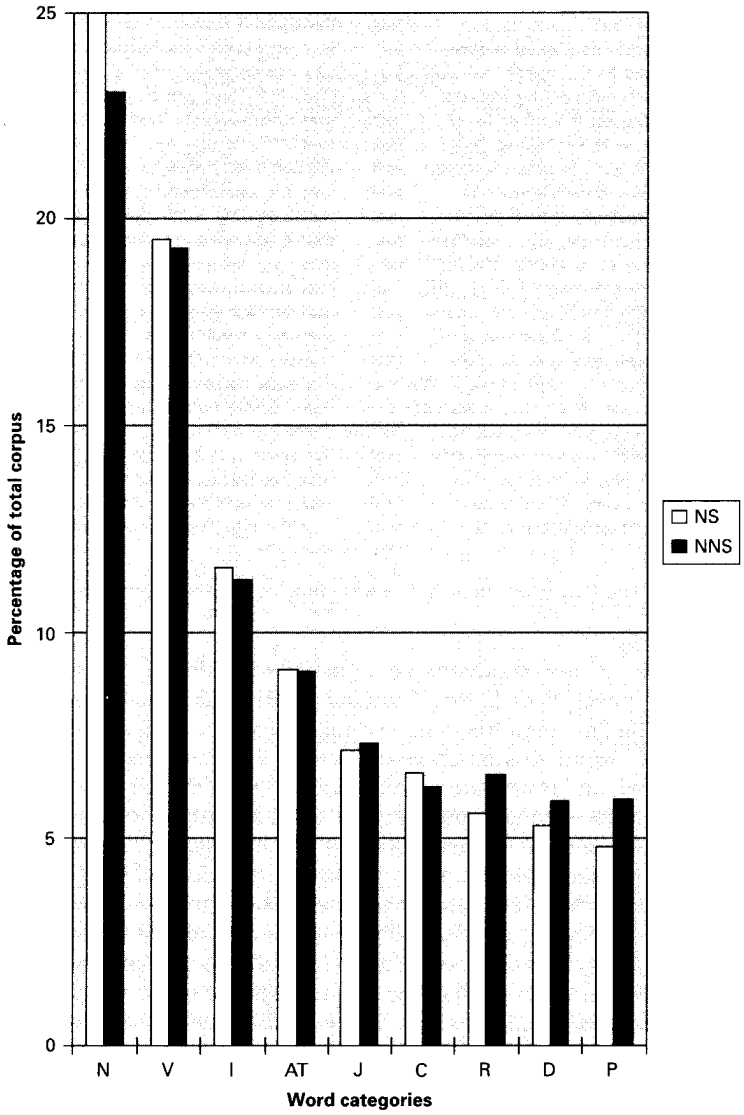


Figure 5.1 Major word category breakdown in NS and NNS corpora

In conclusion, these patterns of significant overuse and underuse for POS categories demonstrate that the learner data displayed many of the stylistic features of spoken rather than written English. This contributes, along with other studies, to a better understanding of learner grammar and lexis and helps to inform teachers in their ELT course design.

5.3.3 Case study three: Semantic analysis and information extraction

The Matrix technique has more recently been applied to compare corpora analysed at the semantic level in a systems engineering domain. The motivation for this work is that despite natural language's well-documented shortcomings in the requirements engineering literature as a medium for precise technical description, its use in software-intensive systems engineering remains inescapable. This poses many problems for engineers who must derive problem understanding and synthesise precise solution descriptions from free text.

This is true both for the largely unstructured textual descriptions from which system requirements are derived, and for more formal documents, such as standards, which impose requirements on system development processes. We describe an experiment that has been carried out in the REVERE project (Rayson et al, 2000) to investigate the use of probabilistic natural language processing techniques to provide systems engineering support.

The target documents are field reports of a series of ethnographic studies at an air traffic control (ATC) centre. This formed part of a study of ATC as an example of a system that supports collaborative user tasks (Bentley et al, 1992). The documents consist of both the verbatim transcripts of the ethnographer’s observations and interviews with controllers, and of reports compiled by the ethnographer for later analysis by a multi-disciplinary team of social scientists and systems engineers. The field reports form an interesting study because they exhibit many characteristics typical of information available in this domain. The volume of the information is fairly high (103 pages) and the documents are not structured in a way (say around business processes or system architecture) designed to help the extraction of requirements. The text was analysed by the part-of-speech tagger, CLAWS (see section 3.2.1), and the semantic field tagger (see section 3.2.2), in preparation for the application of the Matrix tool.

The normative corpus that we used was a 2.3 million-word subset of the BNC derived from the transcripts of spoken English. Using this corpus, the most over-represented semantic categories in the ATC field reports are shown in Table 5.11. The log-likelihood test is applied as described in section 4.3 and represents the semantic tag's frequency deviation from the normative corpus. The higher the figure, the greater the deviation.

Table 5.11 Over-represented categories in ATC field reports

Log-likelihood	Semantic tag	Semantic field (examples from the text)
3366	S7.1	power, organising (‘controller’, ‘chief’)
2578	M5	flying (‘plane’, ‘flight’, ‘airport’)
988	O2	general objects (‘strip’, ‘holder’, ‘rack’)
643	O3	electrical equipment (‘radar’, ‘blip’)
535	Y1	science and technology (‘PH’)
449	W3	geographical terms (‘Pole Hill’, ‘Dish Sea’)
432	Q1.2	paper documents and writing (‘writing’, ‘written’, ‘notes’)
372	N3.7	measurement (‘length’, ‘height’, ‘distance’, ‘levels’, ‘1000ft’)
318	L1	life and living things (‘live’)
310	A10	indicating actions (‘pointing’, ‘indicating’, ‘display’)
306	X4.2	mental objects (‘systems’, ‘approach’, ‘mode’, ‘tactical’, ‘procedure’)
290	A4.1	kinds, groups (‘sector’, ‘sectors’)

With the exception of *Y1* (an anomaly caused by an interviewee’s initials being mistaken for the PH unit of acidity), all of these semantic categories include important objects, roles, functions, etc. in the ATC domain. The frequencies with which some of

these occur, such as *M5* (flying), are unsurprising. Others are more revealing about the domain of ATC. Figure 5.2 shows some of the occurrences of the semantic category *O2* (general objects). The important information revealed here is the importance of 'strips' (formally, 'flight strips'). These are small pieces of cardboard with printed flight details that are the most fundamental artefact used by the air traffic controller to manage their air space. Examination of other words in this category also reveals that flight strips are held in 'racks' to organise them according to (for example) aircraft time-of-arrival.

to write ' l260L'i n red on a	strip	, whilst at thesame time instru
he Isle of Man ... " This	strip	was towards ' the bottom of one
cated by the beacon printed in	box	' B ' of the strip (second lef
on printed in box ' B ' of the	strip	(second left) Stripsseemed br
arrival time over thatbeacon (	box	' A ') This was obviously only
viously only approximate- some	strips	were out ofposition , and I got
al line near the callsign on a	strip	to indicate anunusual speed . <
med much busier . Therewere 16	strips	in one of his racks . A ;
rewere 16 strips in one of his	racks	. A ; ' c<Tide &gt
sy , that talking and using an	input	device might also be , but that
hat talking and using an input	device	might also be , but that the pr
: " the nice thing about	strips	is their flexibility . " a

Figure 5.2 Browsing the semantic category O2

Similarly, browsing the context for *Q1.2* (paper documents and writing) would reveal that controllers annotate flight strips to record deviations from flight plans, and *L1* (life, living things) would reveal that some strips are 'live', that is, they refer to aircraft currently traversing the controller's sector. Notice also that the semantic categories' deviation from the normative corpus can also be expected to reveal roles. In this example, the frequency of *S7.1* (power, organising) confirms the importance of the roles of 'controllers' and 'chiefs', identified by the role analysis described above.

Using the frequency profiling method does not automate the task of identifying abstractions, much less does it produce fully formed requirements that can be pasted into a specification document. Instead, it helps the engineer quickly isolate potentially significant domain abstractions that require closer analysis.

5.4 Summary

The first part of this chapter was used to describe a comparative evaluation of the choice of the log-likelihood statistic used by the Matrix method over the chi-squared statistic. Even beyond the limits of the Cochran rule at the 0.01% level, the log-likelihood test has been to shown to be reliable for the comparison of frequencies between two corpora when the contingency table becomes skewed. Without the hypothesis testing link, it is of use for comparison even when the data become skewed.

In this chapter using three case studies, we have also shown the results of the Matrix method and tool at the lexical, word-class and semantic levels. The Matrix technique has been shown to have applications for the investigation of social differentiation via vocabulary studies, contrastive analysis of native and non-native speaker language, and to the semantic analysis of software engineering requirements documents. In addition, it has also been used as follows:

1. Distinctiveness lists contrasting speech and writing, conversational and task-oriented speech, imaginative and informative writing, in British English (Leech, Rayson and Wilson, 2001).
2. Grammatical word class variation within the British National Corpus Sampler. (Rayson, Wilson and Leech, 2002).
3. Content analysis of cancer-care doctor-patient interactions (Thomas and Wilson, 1996).

6. Conclusions

"And in the end …" (Lennon and McCartney, Abbey Road, 1969).

6.1 Summary of the work

In chapter two, we surveyed the field of corpus linguistics and described the existing process model of 'question – build – annotate – retrieve – interpret'. We saw that most studies chose in advance which linguistic features to examine, even when examining whole texts or varieties of language. We looked at the practice of corpus annotation and saw the multiple levels at which it can be applied. We surveyed the various statistical techniques used to compare frequencies and frequency profiles across corpora. We have seen that keywords can be extracted statistically and manually. The advantages of the log-likelihood ratio over the other measures were summarised at the end of chapter two. We considered various alternatives such as Fisher's Exact test, the chi-squared test, McNemar's chi-squared test, the Mann-Whitney test, normalised ratios and a group of measures suggested by Berg.

Although in our survey, log-likelihood was shown to be better 'in general' than the chi-squared test, there remained a question over its specific use in the comparison of frequency profiles.

In chapter three, the tools needed for the creation and exploitation of (annotated) corpora were categorised into three major groups: corpus development, corpus editing, and information extraction. We looked at features implemented in the software and gave examples of software in each of the three groups, focussing particularly on software falling into the third category. Implementing corpus software for web access was seen as advantageous. We summarised the inclusion or exclusion of twelve important features in nine of the most widely cited retrieval software

packages in corpus linguistics. Only one (WordSmith) was shown to be capable of statistical comparison of word frequency lists. None of the tools combined the annotation-awareness capability with the comparison of frequency lists.

In chapter four, a worked example compared two corpora consisting of UK 2001 General Election manifestos. We extended the keywords approach to key grammatical classes and key concepts. The combination of annotation-awareness and comparison of frequency profiles proved to be of use in a practical data-driven approach as discussed in section 1.1.

The first part of chapter five described a comparative evaluation of the log-likelihood statistic against the chi-squared statistic specifically in the area of comparison of frequency profiles. Even beyond the limits of the Cochran rule at the 0.01% level, the log-likelihood test has been to shown to be reliable for the comparison of frequencies between two corpora when the contingency table becomes skewed.

In the second part of chapter five we showed three case studies containing applications of the Matrix method and tool at the lexical, word-class and semantic levels. The application domains were for the investigation of social differentiation via vocabulary studies, contrastive analysis of native and non-native speaker language, and the semantic analysis of software engineering requirements documents.

6.2 The method proposed

Chapter four described in detail the Matrix method and the software tool that has been implemented to carry out frequency profiling of corpora, and comparison of those profiles across corpora. The Matrix method uses the log-likelihood ratio statistic to compare frequencies and then rank them in terms of significant difference. We have shown that the Matrix method can be used in both types of corpus comparison (A and B, as described in section 2.7).

The Matrix method and tool assists corpus investigation by statistical comparison of frequency profiles at the lexical level and extends this to the word-class and semantic

field levels. Key grammatical categories and semantic classes are used to group together lower frequency words and those words which would, by themselves, not be identified as key, and would otherwise be overlooked. Comparison at the POS and semantic levels reduces the number of key categories that the researcher should examine. Multi-word-units and lemma variants are also grouped together during the annotation process, making key concepts easier to identify. The Matrix method extends the whole text-focussed approach by informing the researcher as to specific linguistic features that should be studied further. This method is described in section 4.3 and evaluated in chapter 5. The Matrix tool acting in a filtering manner allows the researcher to examine more data in a shorter period of time that would otherwise be possible.

The method has been shown to discover key items in the corpora which differentiate one corpus from another. It can be used as a quick way in to find the differences between the corpora and is shown to have applications in the study of social differentiation in the use of English vocabulary, profiling of learner English and document analysis in the software engineering process.

We do not propose a completely automated approach. The tool suggests a group of key items by decreasing order of significance that distinguish one corpus from another. It is then that the researcher should investigate occurrences of the significant items in the corpora using standard corpus techniques such as KWIC concordancing, by which the reasons behind their significance can be discovered and explanations sought for the patterns displayed. By this process, we can compare the corpora under investigation and make hypotheses about the language use that they represent. When applying the Matrix method it is vitally important to take into account the issues relevant to comparison of corpora as discussed in section 2.7; representativeness, homogeneity and comparability. The fourth issue, related to choice and reliability of statistical tests, was addressed directly in this thesis. The Matrix method itself may assist in assessing representativeness, homogeneity and comparability of corpora, but in this case the researcher using the method should keep these issues in mind when interpreting the results.

As stated in our statistical evaluation of Matrix (section 5.2.2), there is a difference between establishing statistical significance and practical significance. We are not proposing complete reliance on the results of the log-likelihood test, or the comparison of all LL values to the listed critical values (5%, 1% and so on). If a statistically significant result is required for a particular item, we can rely on testing LL values at the 0.01% level with a critical value of 15.13 (as seen in section 5.2). In other disciplines such as sociology and psychology, much has been written about the practice of carrying out significance tests. In order to raise critical awareness of the tests, Morrison and Henkel (1970) collected together publications in (amongst others) the areas of (i) the difference between statistical inference (from significance testing) and scientific inference (the general process by which scientific knowledge is generated) (ii) the over-reliance on conventional levels of significance (0.05, 0.01 etc) and (iii) the methodological issues regarding the null hypothesis. We agree that corpus linguists need to heed the same warnings.

6.3 Limitations and future work

Turning first to technical aspects of the Matrix tool, currently there are limitations on the language and size of corpus that can be processed. We will describe later in this section the possibilities of substituting other corpus annotation tools into Matrix. This would be required since CLAWS and USAS are applicable only to English. Other languages using different encoding systems, such as the Unicode standard[70], would require reimplementation of parts of the tool (for production of frequency lists and concordances) in Java, for example, which supports Unicode. The Matrix tool can already process large corpora, such as the BNC, but the processing time involved in dynamically indexing the corpus prohibits its use interactively via the web interface. Database solutions, such as that used by Davies (2002), have already been noted in section 3.2.

From the point of view of corpus linguists, one question that needs further investigation is whether the Matrix technique is suitable for use in corpus studies of

[70] See the website at http://www.unicode.org/ for a description of the Unicode standard

languages other than English. The Matrix tool relies on POS and semantic taggers being available in the language of the corpus to be examined, and large reference corpora if we wish to carry out a type A study as described in section 2.7. POS taggers are available for the annotation of many different languages, but this is not the case for semantic taggers. WordNet databases are available in a number of European languages so this may be of assistance in future development of semantic taggers for these languages. Additionally, a growing number of languages have a national corpus project or similar under way or completed. The keywords technique itself does seem to be applicable to other languages, for example Polish (Uzar and Waliński, 2000).

We must also investigate how closely the Matrix technique is tied to the two levels of linguistic annotation. If we wish to extend the method to include syntactic phenomena, for example, then using total word count as the basis for comparison is not suitable. Ball (1994) points out that syntactic constructions such as the cleft sentence are clauses, and their relative frequency should be measured in terms of the number of clauses, and not words.

The Matrix method and tool is closely tied to the corpus annotation tools, CLAWS and USAS. These tools do not attain 100% accuracy, hence researchers using the Matrix method need to keep in mind the possibility of errors in tagging. Other POS taggers are available, as noted in section 3.2.1, applying similar sets of POS tags. Semantic taggers differ widely in their application to all-words or open-class words, and tagsets are far from standardised even at the first level of detail when they are arranged in a hierarchical structure. Also, the USAS tagger relies on the CLAWS tagset during disambiguation. We would expect that a large investment of time would be needed to test the Matrix method with a substitute semantic tagger.

In section 2.7 we mentioned that in a different application area, extracting groups of associated words, Weeber et al (2000), used a combination of the log-likelihood ratio and Fisher's Exact test for the full word frequency range. We might attempt to use this combination to see what affect it has on the results from the Matrix method. We would also need to assess the statistical and practical complexity of comparing the log-likelihood statistic with Fisher's Exact test, to enable us to include the hypothesis testing aspect in the method.

6.4 Objectives and novel contributions

Our main objective in this study was to investigate possibilities for automation in selecting words for further study, and checking concordance lines for the correct sense of a word and the distribution of a word. In this study, we have shown that Matrix is suitable for these tasks. The sub-objectives as listed in section 1.5 have been met and we claim that this thesis makes the following novel contributions to the field:

1. a data-driven method for corpus comparison has been developed which uses macroscopic analysis (the study of the characteristics of whole texts or varieties of language) to inform the microscopic level (focussing on the use of a particular linguistic feature) by suggesting linguistic features to be further investigated
2. the method integrates the comparison of corpora with word-class tagging and lexical semantic tagging, it extends the keywords procedure to key grammatical categories and key concepts
3. the method can be used for comparison of differently sized corpora, not just equal-sized pairs
4. the method can be applied to the full frequency profile without requiring a lower frequency boundary
5. a comparison of the reliability of the log-likelihood and chi-squared statistics with various combinations of corpus size, word probability and ratio of corpora
6. an annotation-aware software tool implementing the method has been developed
7. the application of the software tool to political linguistics, vocabulary studies, learner corpora and information extraction has been shown

7. References

Aarts, J. (1991). Intuition-based and observation-based grammars. In Aijmer, K. and Altenberg, B. (eds.), *English Corpus Linguistics: Studies in honour of Jan Svartvik*. Longman, London, pp. 44 – 62.

Agresti, A. (1990). *Categorical data analysis*. Wiley, New York.

Aitchison, J. (1991). *Language change: progress or decay (2nd edition)*. Cambridge University Press, Cambridge.

Alexandersson, J., Buschbeck-Wolf, B., Fujinami, T., Kipp, M., Koch, S., Maier, S., Reithinger, N., Schmitz, B., Siegel, M. (1998). *Dialogue Acts in VERBMOBIL-2 -- Second Edition*. Verbmobil-Report 226. DFKI GmbH, Saarbrucken, Germany.

Altenberg, B. (1986). Contrastive linking in spoken and written English. In Tottie, G. and Bäcklund, I. (eds.) *English in Speech and Writing: a symposium*. Almqvist and Wiksell, Stockholm, pp. 13 – 40.

Altenberg, B. (1989). Review of D. Biber (1988) Variation across speech and writing. *Studia Linguistica* 43 (2), pp. 167 – 174.

Aston, G. and Burnard, L. (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*, Edinburgh University Press, Edinburgh.

Atwell, E., Leech, G. and Garside, R. (1984). Analysis of the LOB corpus: progress and prospects. In Aarts, J. and Meijs, W. (eds.), *Corpus Linguistics*. Amsterdam: Rodopi, pp. 41 – 52.

Atwell, E. and Elliott, S. (1987). Dealing with Ill-formed English Text. In Garside, R., Leech, G., and Sampson, G. (eds.), *The Computational Analysis of English: A Corpus-based Approach*. Longman, London , pp. 120 – 138.

Baayen, R. H. (1993). Statistical models for word frequency distributions: a linguistic evaluation. *Computers and the Humanities*, Kluwer, The Netherlands, 26, pp. 347 – 363.

Baayen, R. H. (1997). Review of D. Biber (1995) Dimensions of register variation: a cross-linguistic comparison. *Literary and Linguistic Computing*, 12 (1), Oxford University Press, pp. 65 – 67.

Baayen, R. H. (2001). *Word frequency distributions*. Kluwer, The Netherlands.

Ball, C. N. (1994). Automated text analysis: cautionary tales. *Literary and Linguistic Computing*, Vol. 9, no. 4, Oxford University Press, Oxford, pp. 295 – 302.

Barnbrook, G. (1996). *Language and Computers: a practical introduction to the computer analysis of language*. Edinburgh University Press, Edinburgh.

Barnett, S. and Cronin, T. M. (1986). *Mathematical formulae for engineering and science students, fourth edition*. Longman, London.

Bateman, J., Forrest, J. and Willis, T. (1997). The use of syntactic annotation tools: partial and full parsing. In Garside, R., Leech, G., and McEnery, A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London. pp 166 – 178.

Beale, A. D. (1985a). A Probabilistic Approach to Grammatical Analysis of Written English. *Proceedings of the second conference of the European Chapter of the Association of Computational Linguistics*, 27-29 March 1985, Geneva, pp. 159 – 165.

Beale, A. D. (1985b). Grammatical Analysis by Computer of the Lancaster-Oslo/Bergen Corpus of British English Texts. *Proceedings of the 23rd Annual Meeting of the Association of Computational Linguistics*, 8-12 July 1985, Illinois, pp. 293 – 298.

Beale, A. D. (1987). Towards a distributional lexicon. In Garside, R., Leech, G. and Sampson, G. (eds.), *The Computational Analysis of English: a corpus-based approach*. Longman, London, pp. 149 – 162.

Beale, A. D. (1989). *The Development of a Distributional Lexicon: A Contribution to Computational Lexicography*. Unpublished Ph.D. thesis, Lancaster University.

Belmore, N. (1997). Comparing tagging systems. In Ljung, M. (ed.) *Corpus-based studies in English: papers from the seventeenth International Conference on English language research on computerized corpora (ICAME 17), Stockholm, May 15-19, 1996*, Rodopi, Amsterdam, pp. 331 – 338.

Bentley, R., Rodden, T., Sawyer, P., Sommerville, I, Hughes, J., Randall, D., Shapiro, D. (1992). Ethnographically-informed systems design for air traffic control, In *Proceedings of CSCW '92*, Toronto, November 1992.

Berber Sardinha, T. (1999). Using KeyWords in text analysis: Practical aspects. *DIRECT Papers* 42, São Paulo and Liverpool.

Berg, C. N. (1997). *Developing a corpus specific stop-list using quantitative comparison*. MSc thesis AFIT/GIR/LAL/97D-2. Graduate School of Logistics and Acquisition Management, Air Force Institute of Technology, Wright-Patterson Air Force Base, Ohio.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press, Cambridge.

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8 (4), Oxford University Press, Oxford, pp. 243 – 257.

Biber, D. (1995). *Dimensions of register variation : a cross-linguistic comparison*. Cambridge University Press, Cambridge.

Biber, D. and Finegan, E. (1989). Drift and the evolution of English style: a history of three genres. *Language* 65, pp. 487 – 517.

Biber, D., Conrad, S., and Reppen, R. (1998). *Corpus Linguistics: investigating language structure and use.* Cambridge University Press, Cambridge.

Bird, S. and Liberman, M. (1998). Towards a formal framework for linguistic annotations, In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'98)*, Sydney, December 1998.

Bird, S. and Liberman, M. (2001). A formal framework for linguistic annotation, *Speech Communication* 33 (1,2), Elsevier, pp. 23 – 60.

Bird, S., Maeda, K., Ma, X., Lee, H., Randall, B., and Zayat, S. (2002). TableTrans, MultiTrans, InterTrans and TreeTrans: Diverse Tools Built on the Annotation Graph Toolkit, *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands, Spain, 29 – 31 May 2002, pp. 364 – 370.

Boguraev, B., Garigliano, R., and Tait, J. (1995). Editorial. *Natural Language Engineering*. 1 (1), Cambridge University Press, pp. 1 – 7.

Booth, B. (1987). Text input and pre-processing: dealing with the orthographic form of texts. In Garside, R., Leech, G., and Sampson, G. (eds.) *The Computational analysis of English: a corpus-based approach*. Longman, London, pp. 97 – 109.

Bradley, J. and Rockwell, G. (1995). TACTweb: Argument and Evidence on the Internet. In *proceedings of Joint International Conference of the Association*

for Computers and the Humanities and the Association for Literary & Linguistic Computing (ACH/ALLC 1995), July 11-15, 1995, University of California, Santa Barbara, California, pp. 11 – 13. Online at http://pigeon.cch.kcl.ac.uk/docs/papers/tactweb.html

Brill, E. (1992). A simple rule-based part-of-speech tagger, In *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, Italy.

Brodda, B. (1991) Doing corpus work with PC Beta; or, how to be your own computational linguist. In Johansson, S. and Stenström, A.-B. (eds.) *English computer corpora: Selected papers and research guide*. Mouton de Gruyter, Berlin, pp. 259 – 282..

Burnage, G. and Dunlop, D. (1993) Encoding the British National Corpus. In Aarts, J., de Haan, P., and Oostdijk, N. (eds.) *English language corpora: Design, analysis and exploitation*. Rodopi, Amsterdam, pp. 79 – 95.

Burnard, L. (1992). Tools and techniques for computer-assisted text processing. In C. S. Butler (ed.) *Computers and written texts*. Blackwell, Oxford.

Burnard, L. (ed.) (1995). *Users reference guide for the British National Corpus. Version 1.0.* Oxford University Computing Services, Oxford.

Butler, C. (1985). *Statistics in linguistics*. Blackwell, Oxford.

Carroll, J. B., Davies, P., and Richman, B. (1971). *The American Heritage Word Frequency Book*. Houghton Mifflin, Boston.

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. *Proceedings of COMPLEX'94: 3rd Conference on Computational Lexicography and Text Research (Budapest, July 7-10 1994).* Budapest, Hungary, pp. 23 – 32.

Church, K. W. and Gale, W. A. (1995). Poisson mixtures. *Natural Language Engineering*, 1(2), Cambridge University Press, pp. 163 – 190.

Clear, J. (1992). Corpus sampling. In G. Leitner (ed.) *New directions in English language corpora*. Mouton-de-Gruyter, Berlin, pp. 21 – 31.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics*, 10, pp. 417 – 451.

Colebourne, A., Sawyer, P., and Sommerville, I. (1993). MOG user interface builder: a mechanism for integrating application and user interface. *Interacting with Computers*, volume 5, number 3, Elsevier, pp. 315 – 331.

Collins, H. and Scott, M. (1997). Lexical Landscaping in Business Meetings, in Bargiela F. and Harris, S. (eds.) *The Languages of Business: an international perspective*, Edinburgh University Press, Edinburgh, pp. 183 – 208.

Copeck, T., Barker, K., Delisle, S., and Szpakowicz, S. (1999). More Alike than not - An Analysis of Word Frequencies in Four General-purpose Text Corpora. *Proceedings of the 1999 Pacific Association for Computational Linguistics Conference (PACLING 99), Waterloo (Ontario, Canada), 25-28 August 1999*, pp. 282 – 287.

Cowden-Clarke, M. V. (1881). *The complete concordance to Shakespeare: being a verbal index to all the passages in the dramatic works of the poet, new and revised edition*. Bickers & Son, London.

Cressie, N. and Read, T. R. C. (1984) Multinomial Goodness-of-Fit Tests. *Journal of the Royal Statistical Society. Series B (Methodological),* Vol. 46, No. 3, pp. 440 – 464.

Cressie, N. and Read, T. R. C. (1989). Pearson's X^2 and the Loglikelihood Ratio Statistic G^2: A comparative review. *International Statistical Review*, **57**, 1, Belfast University Press, N.I., pp. 19 – 43.

Cunningham, H. (1999). A definition and short history of Language Engineering. *Natural Language Engineering*. **5** (1), Cambridge University Press, Cambridge, pp. 1 – 16..

Cunningham, H., Wilks, Y., and Gaizauskas, R. (1996) GATE – a General Architecture for Text Engineering. In Proceedings of the 16th Conference on Computational Linguistics (COLING-96), Copenhagen, Aug, 1996.

Cunningham, H., Bontcheva, K., Tablan, V., and Wilks, Y. (2000). Software Infrastructure for Language Resources: a Taxonomy of Previous Work and a Requirements Analysis. In *Proceedings of the Second Conference on Language Resources Evaluation (LREC), Athens*.

Cutting, D., Kupiec, J., Pederson, J., Sibun, P. (1992). A practical part-of-speech tagger, In *proceedings of Third Conference on Applied Natural Language Processing*, Trento, Italy, pp. 133 – 140.

Daelemans, W., van den Bosch, A., Zavrel, J., Veenstra, J., Buchholz, S., Busser, B. (1998) Rapid development of NLP modules with memory-based learning, In *Proceedings of ELSNET in Wonderland (ELSNET98),* Utrecht, pp. 105 – 113.

Dahl, H. (1979) *Word frequencies of spoken American English*. Verbatim, Michigan.

Davies, M. (2002). Using Relational Databases to Create Unlimited and User-Defined Annotation on Large Corpora: A 100 Million Word Corpus of Historical and Modern Spanish. Presented at the *5th Annual CLUK (Computational Linguistics in the UK) Research Colloquium.* Leeds, England.

De Cock, S. (1998). A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics*, Vol. 3 (1), John Benjamins, pp. 59 – 80.

Devito, J. (1966). Psychogrammatical factors in oral and written discourse by skilled communicators, *Speech Monographs* 33, pp. 73 – 76.

Devito, J. (1967). Levels of abstraction in spoken and written language, *Journal of Communication* 17, pp. 354 – 361.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7), pp. 1895 – 1923.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19, 1, March 1993, MIT Press, pp. 61 – 74.

Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the Association for Computing Machinery*, 16 (2), pp. 264 – 285.

Edwards, J. A. (1995) Principles and alternative systems in the transcription, coding and mark-up of spoken discourse. In Leech, G. N., Myers, G., and Thomas, J. (eds.) *Spoken English on computer: Transcription, mark-up and application.* Longman, London, pp. 19 – 34.

Everitt, B. S. (1992). *The analysis of contingency tables, 2nd edition*. Chapman and Hall, London.

Fairclough, N. (2000). *New Labour, New Language?* Routledge, London.

Firth, J. R. (1935). The technique of semantics. *Transactions of the Philological Society*, David Nutt, London, pp. 36 – 72.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930 – 1955. In *Studies in Linguistic analysis*, Special volume, Philological Society, pp. 1 – 32.

Fligelstone, S. (1992). Developing a scheme for annotating text to show anaphoric relations. In Leitner, G. (ed.), *New directions in corpus linguistics*. Mouton de Gruyter, Berlin, pp. 153 – 170.

Fligelstone, S. (1995). *Jaws: using lemmatisation rules and contextual disambiguation rules to enhance CLAWS output*. Project report, Linguistics Department, Lancaster University.

Fligelstone, S., Rayson, P., and Smith, N. (1996). Template analysis: bridging the gap between grammar and the lexicon. In Thomas, J. and Short, M. (eds.), *Using corpora for language research: studies in the honour of Geoffrey Leech*. Longman, London, pp. 181 – 207.

Fligelstone, S., Pacey, M., and Rayson, P. (1997). How to generalise the task of annotation. In Garside, R., Leech, G., and McEnery, A. (eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora.* Longman, London, pp. 122 – 136.

Foley, W. A. (1997). *Anthropological Linguistics*. Blackwell, Oxford.

Francis, B., Green, M. and Payne, **C.** (1993) *The GLIM4 system*. Oxford University Press, Oxford.

Francis, G. (1993). A corpus-driven approach to grammar: principles, methods and examples. In Baker, M., Francis, G., and Tognini-Bonelli, E. (eds.) *Text and technology: in honour of John Sinclair*. Benjamins, The Netherlands, pp. 137 – 156.

Francis, W. N. (1992). Language corpora B. C. In Svartvik, J. (ed.) *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4 – 8 August 1991*. Mouton de Gruyter, Berlin, pp. 17 – 32.

Francis, W. N. and Kučera, H. (1964). *Manual of information to accompany a standard sample of present-day edited American English, for use with digital computers.* Department of Linguistics, Brown University, Providence.

Francis, W. N. and Kučera, H. (1982). *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin, Boston.

Frank, E., Paynter, G.W., Witten, I. H., Gutwin, C. and Nevill-Manning, C. G. (1999) Domain-specific keyphrase extraction. In proceedings of *Sixteenth International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, Morgan Kaufmann Publishers, San Francisco, CA, pp. 668 – 673.

Fries, C. C. and Traver, A. A. (1950). *English word lists: a study of their adaptability for instruction.* George Wahr Publishing Company, Ann Arbor, Michigan.

Gale, W. A., Church, K. W., and Yarowsky, D. (1992) One sense per discourse. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, pp. 233 – 237.

Gaizauskas, R. and Robertson, A. M. (1997). Coupling Information Retrieval and Information Extraction: A New Text Technology for Gathering Information from the Web. In *Proceedings of RIAO 97: Computer-Assisted Information Searching on the Internet, Montreal, Canada*, pp. 356 – 370.

Garside, R. (1987) The CLAWS Word-tagging System. In Garside, R, Leech, G. and Sampson, G. (eds.), *The Computational Analysis of English: A Corpus-based Approach*. Longman, London, pp. 30 – 41.

Garside, R. (1995). Grammatical tagging of the spoken part of the British National Corpus: a progress report. In Leech, G., Myers, G., and Thomas, J. (eds.) *Spoken English on Computer*, Longman, London, pp. 161 – 167.

Garside, R. and Leech, F. A. (1985). A Probabilistic Parser. *Proceedings of the second conference of the European Chapter of the Association of Computational Linguistics*, 27-29 March 1985, Geneva, pp. 166 – 170.

Garside, R. and Smith, N. (1997). A Hybrid Grammatical Tagger: CLAWS4, in Garside, R., Leech, G., and McEnery, A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*, Longman, London, pp. 102 – 121.

Garside, R., and Rayson, P. (1997). Higher-level annotation tools. In Garside, R., Leech, G., and McEnery, A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London. pp 179 – 193.

Garside, R., Leech, G., and Sampson, G. (eds.) (1987) *The Computational Analysis of English: A Corpus-based Approach*. Longman, London.

Garside, R., Leech, G., and McEnery, A. (eds.) (1997) *Corpus Annotation: Linguistic Information from Computer Text Corpora.* Longman, London.

Ghadessy, M., Henry, A. and Roseberry, R. L. (eds.) (2001). *Small corpus studies and ELT: theory and practice*, John Benjamins, Amsterdam.

Granger, S. (1993). International Corpus of Learner English. In Aarts, J., de Haan, P., and Oostdijk, N. (eds.) *English language corpora: Design, analysis and exploitation*. Rodopi, Amsterdam, pp. 57 – 71.

Granger, S. (1998). The computer learner corpus: a versatile new source of data for SLA research. In S. Granger (ed.) *Learner English on Computer*. Longman, London, pp. 3 – 18.

Granger, S. (1999). Use of tenses by advanced EFL learners: evidence from an error-tagged computer corpus. In Hasselgård, H. and Oksefjell, S. (eds.) *Out of corpora: studies in honour of Stig Johansson.* Rodopi, Amsterdam. pp. 191 – 202.

Granger, S. and Rayson, P. (1998). Automatic profiling of learner texts. In S. Granger (ed.) *Learner English on Computer*. Longman, London, pp. 119 – 131.

Greenbaum, S. (ed.) (1996) *Comparing English Worldwide: The International Corpus of English*. Clarendon Press, Oxford.

Grefenstette, G. (1999). Tokenization. In van Halteren, H, (ed.) *Syntactic wordclass tagging*, Kluwer, The Netherlands, pp. 117 – 133.

Grefenstette, G. and Tapanainen, P. (1994) What is a Word, What is a Sentence? Problems of Tokenization. In *Proceedings of 3rd conference on Computational Lexicography and Text Research (COMPLEX'94)*, Budapest, July 7-10, 1994, pp. 79 – 87.

Greene, B. B. and Rubin, G. M. (1971). *Automatic Grammatical Tagging of English*. Providence RI: Department of Linguistics, Brown University.

Grishman, R. (1986). *Computational linguistics: an introduction*. Cambridge University Press.

Guarino, N. (1995) Formal ontology, conceptual analysis and knowledge representation. *International Journal of Human and Computer Studies*, 43 (5/6), pp. 625 – 640.

Gundavaram, S. (1996). *CGI Programming on the World Wide Web*. O'Reilly & Associates, Sebastopol, California.

van Halteren, H. and Oostdijk, N. (1993). Towards a syntactic database: The TOSCA analysis system. In Aarts, J., de Haan, P., and Oostdijk, N. (eds.) *English language corpora: Design, analysis and exploitation*. Rodopi, Amsterdam, pp. 145 – 161.

van Halteren, H., Zavrel, J., and Daelemans, W. (1998). Improving data driven wordclass tagging by system combination. In *Proceedings of COLING-ACL'98* August 10-14, Montreal, Canada, pp. 491 – 497.

Heller, D., Ferguson, P., and Brennan, D. (1994). *Volume 6A: Motif Programming Manual (2nd edition).* O'Reilly & Associates, USA.

Henry, A. and Roseberry, R. L. (2001). Using a small corpus to obtain data for teaching a genre. In Ghadessy, M., Henry, A. and Roseberry, R. L. (eds.) *Small corpus studies and ELT: theory and practice*, John Benjamins, Amsterdam, pp. 93 – 133.

van Herwijnen, E. (1994). *Practical SGML, 2nd Edition*. Kluwer, Boston.

Hickey, R. (1993a) *Lexa: Corpus processing software*. 3 vols. Volume 1: Lexical analysis. Volume 2: Database and corpus management. Volume 3: Utility library. Norwegian Computing Centre for the Humanities, Bergen.

Hickey, R. (1993b) Corpus data processing with Lexa, *ICAME Journal* 17, pp. 73 – 95.

Hisamitsu, T. and Niwa, Y. (2001). Extracting useful terms from parenthetical expressions by combining simple rules and statistical measures: a comparative evaluation of bigram statistics. In Bourigault, D., Jacquemin, C. and L'Homme, M-C. (eds.) *Recent advances in computational terminology*. Benjamins, The Netherlands, pp. 209 – 224.

Hockey, S. (2000). Electronic texts in the humanities. Oxford University Press, Oxford.

Hockey, S. (2001). Concordance programs for corpus linguistics. In Simpson, R. C. and Swales, J. M. (eds.) *Corpus linguistics in North America: selections from the 1999 symposium*, University of Michigan Press, Ann Arbor, pp. 76 – 97.

Hockey, S. and Martin, J. (1987). The Oxford Concordance Program Version 2. *Literary and Linguistic Computing*, 2, Issue 2, Oxford University Press, Oxford, pp. 125 – 131.

Hoffmann, S. and Lehmann, H. M. (2000). Collocational evidence from the British National Corpus. In Kirk, J. M. (ed.) *Corpora galore: analyses and techniques in describing English.* Rodopi, Amsterdam, pp. 17 – 32.

Hofland, K. (1991) Concordance programs for personal computers. In Johansson, S. and Stenström, A.-B. (eds.) *English computer corpora: Selected papers and research guide*. Mouton de Gruyter., Berlin, pp. 283 – 306.

Hofland, K. and Johansson, S. (1982). *Word frequencies in British and American English*. The Norwegian Computing Centre for the Humanities, Bergen, Norway.

Hughes, L. and Lee, S. (1994) (eds.) *CTI Centre for Textual Studies resources guide (1994)*, CTI Centre for Textual Studies, Oxford.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press, Cambridge.

Ide, N. (1996) *Corpus Encoding Standard. MULTEXT EAGLES - Document CES 1*. Version 1.4. Nancy Ide, Coordinator.

Ide, N. (1998). Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. *Proceedings of the First International Language Resources and Evaluation Conference*, Granada, Spain, pp. 463 – 470.

James, G., Davison, R., Cheung, A. H. Y. and Deerwester, S. (eds.) (1994). *English in computer science: A corpus-based lexical analysi*s. Language Centre, Hong Kong University of Science and Technology.

Jelinek, F. (1990) Self-organised language modeling for speech recognition, In Waibel, A, and Lee, K-F. (eds.) *Readings in speech recognition*, Morgan Kaufman, pp. 450 – 506.

Johansson, S. (1978). *Some aspects of the vocabulary of learned and scientific English.* Acta Universitatis Gothoburgensis, Göteburg.

Johansson, S. (1985) Word frequency and text type: Some observations based on the LOB corpus of British English texts. *Computing and the Humanities*, 19: pp. 23 – 36.

Johansson, S. (1994). Continuity and change in encoding of computer corpora. In Oostdijk, N. and de Haan, P. (eds.) *Corpus-based research into language: in honour of Jan Aarts*. Rodopi, Amsterdam, pp. 13 – 31.

Johansson, S., Leech, G. and Goodluck, H. (1978). *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers*. Department of English, University of Oslo.

Johansson, S., Atwell, E., Garside, R., and Leech, G. (1986). *The tagged LOB corpus: users' manual*. Norwegian Computing Centre for the Humanities, Bergen.

Johansson, S. and Hofland, K. (1989). *Frequency analysis of English vocabulary and grammar: based on the LOB Corpus*, 2 volumes. Clarendon Press, Oxford.

Jones, R. L. (1987). Accessing the Brown Corpus using an IBM PC, *ICAME Journal* 11, Norwegian Computing Centre for the Humanities, pp. 44 – 47.

Jones, R. L. (1997). Creating and using a corpus of spoken German. In Wichmann, A., Fligelstone, S., McEnery, T. and Knowles, G. (eds.) *Teaching and language corpora*. Longman, London and New York, pp. 146 – 156.

Jones, S. and Sinclair, J. (1974). English lexical collocations: A study in computational linguistics, *Cahiers de Lexicologie* 24: pp. 15 – 61.

Juilland, A. and Chang-Rodriguez, E. (1964). *Frequency dictionary of Spanish words*. Mouton & Co., The Hague.

Juilland, A., Edwards, P. M. H. and Juilland, I. (1965). *Frequency dictionary of Rumanian words*. Mouton & Co., The Hague.

Juilland, A., Brodin, D., and Davidovitch, C. (1970). *Frequency dictionary of French words*. Mouton & Co., Paris.

Kabán, A. and Girolami, M. (2000). *Unsupervised topic separation and keyword identification: A Projection Approach*, Computing and Information Systems Technical report. No. 10, University of Paisley.

Kageura, K. (1999). Bigram Statistics Revisited: A Comparative Examination of Some Statistical Measures in Morphological Analysis of Japanese Kanji Sequences. *Journal of Quantitative Linguistics*, Vol. 6, No. 2, pp. 149 – 166.

Kahrel, P., Barnett, R. and Leech, G. (1997). Towards cross-linguistic standards or guidelines for the annotation of corpora. In Garside, R., Leech, G., and McEnery, A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London. pp 231 – 242.

Karlsson, F. (1994). Robust parsing of unconstrained text. In Oostdijk, N. and de Haan, P. (eds.) *Corpus-based research into language: in honour of Jan Aarts*. Rodopi, Amsterdam, pp. 121 – 142.

Karlsson, F., Voutilainen, A., Heikkilä, J., and Anttila, A. (1995) (eds.) *Constraint Grammar, a language-independent system for parsing unrestricted text*. Mouton de Gruyter, Berlin and New York.

Kessler, B. (2001). *The significance of word lists*. Center for the study of language and information, Stanford, California.

Kilgarriff, A. (1996a). Which words are particularly characteristic of a text? A survey of statistical approaches. In Evett, L. J., and Rose, T. G. (eds.) *Language Engineering for Document Analysis and Recognition (LEDAR), AISB96 Workshop proceedings*, Brighton, England. Faculty of Engineering and Computing, Nottingham Trent University, UK, pp. 33 – 40.

Kilgarriff, A. (1996b) Why chi-square doesn't work, and an improved LOB-Brown comparison. *ALLC-ACH Conference*, June 1996, Bergen, Norway.

Kilgarriff, A. (1997). Using word frequency lists to measure corpus homogeneity and similarity between corpora. *Proceedings 5th ACL workshop on very large corpora.* Beijing and Hong Kong, pp. 231 – 245.

Kilgarriff, A. (2001). Comparing Corpora. *International Journal of Corpus Linguistics*, John Benjamins, Amsterdam, Vol. 6, no. 1, pp. 97 – 133.

Kilgarriff, A. and Rose, T. (1998). Measures for corpus similarity and homogeneity. In *proceedings of the 3rd conference on Empirical Methods in Natural Language Processing*, Granada, Spain, pp. 46 – 52.

Kilgarriff, A. and Tugwell, D. (2002). Sketching words. In Corréard, M-H. (ed.) *Lexicography and natural language processing: a festschrift in honour of B. T. S. Atkins*, Euralex, pp. 125 – 137.

Kirk, J. M. (1994). Taking a byte at corpus linguistics. In Flowerdew, L. and Tong, A. K. K. (eds.) *Entering Text*. Language Centre, Hong Kong University of Science and Technology, Hong Kong, pp. 18 – 49.

Kirk, J. M. (ed.) (2000). *Corpora galore: analyses and techniques in describing English.* Rodopi, Amsterdam.

Knowles, G. (1993). The machine-readable Spoken English Corpus. In Aarts, J., de Haan, P., and Oostdijk, N. (eds.) *English language corpora: design, analysis and exploitation*. Rodopi, Amsterdam, pp. 107 – 119.

Knowles, G. (1995) Converting a corpus into a relational database: SEC becomes MARSEC. In Leech, G. N., Myers, G., and Thomas, J. (eds.) *Spoken English on computer: Transcription, mark-up and application.* Longman, London, pp. 208 – 219.

Krenn, B. and Samuelsson, C. (1997). *The Linguist's Guide to Statistics: Don't Panic.* (Version dated December 19, 1997). Available from http://www.coli.uni-sb.de/~krenn/stat_nlp.ps.gz

Kretzschmar, W. A., Meyer, C. F., and Ingegneri, D. (1997). Uses of inferential statistics in corpus studies. In Ljung, M. (ed.) *Corpus-based studies in English: papers from the seventeenth International Conference on English language research on computerized corpora (ICAME 17), Stockholm, May 15-19, 1996*, Rodopi, Amsterdam, pp. 167 – 177.

Kučera, H. and Francis, W. N. (1967). *Computational analysis of present-day American English*. Brown University Press, Providence.

Lancashire, I. (1991) (ed.) *The Humanities Computing Yearbook (1989-90).* Oxford University Press, Oxford.

Lancashire, I., Bradley, J., McCarty, W., Stairs, M., and Wooldridge, T. R. (1996). *Using TACT with Electronic Texts: a guide to text-analysis computing tools, version 2.1 for MS-DOS and PC DOS*. Modern Language Association of America, New York.

Lebart, L., Salem, A., and Berry, L. (1998). *Exploring textual data*. Kluwer, The Netherlands.

Lee, D. Y. W. (2000). Modelling variation in spoken and written language: the multi-dimensional approach revisited. Unpublished PhD thesis, Linguistics Department, Lancaster University.

Lee, D. Y. W. and Rayson, P. (2000). *Xkwic: a powerful concordancer for research*. Handout from workshop presented at Teaching and Language Corpora conference (TALC2000), 19 – 23 July 2000, Graz, Austria.

Leech, F. A. (1987). *An Approach to Probabilistic Parsing*. Unpublished M.Phil. thesis, Lancaster University.

Leech, G. (1986). Automatic grammatical analysis and its educational applications. In Leech, G., and Candlin, C. N. (eds.) *Computers in English language teaching and research*. Longman, London and New York, pp. 205 – 214.

Leech, G. (1987). General introduction. In Garside R., Leech G. and Sampson G. (eds.) *The Computational Analysis of English: A Corpus-based Approach*. Longman, London, pp. 1 – 15.

Leech, G. (1991). The state of the art in corpus linguistics. In Aijmer, K. and Altenberg, B. (eds.), *English Corpus Linguistics: Studies in honour of Jan Svartvik*. Longman, London, pp. 8 – 29.

Leech, G. (1992). Corpus linguistics and theories of linguistic performance. In Svartvik, J. (ed.) *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4 – 8 August 1991*. Mouton de Gruyter, Berlin, pp. 105 – 122.

Leech, G. (1993). 100 million words of English: a description of the background, nature and prospects of the British National Corpus project. *English Today* 33, Vol. 9, No. 1, Cambridge University Press.

Leech, G. (1997) Introducing Corpus Annotation. In Garside, R., Leech, G., and McEnery, A. (eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora.* Longman, London, pp. 1 – 18.

Leech, G. and Garside, R. (1991). Running a grammar factory: The production of syntactically analysed corpora or 'treebanks'. In Johansson, S. and Stenström, A.-B. (eds.), English Computer Corpora: Selected Papers and Research Guide. Mouton de Gruyter, Berlin, pp. 15 – 32.

Leech, G. and Fallon, R. (1992). Computer corpora – what do they tell us about culture? *ICAME Journal*, 16, Norwegian Computing Centre for the Humanities, Bergen, Norway, pp. 29 – 50.

Leech, G. and Fligelstone, S. (1992). Computers and corpus analysis. In Butler, C. S. (ed.) *Computers and written texts*. Blackwell, Oxford, pp. 115 – 140.

Leech, G., Garside, R., and Bryant, M. (1994a). The large-scale grammatical tagging of text: Experience with the British National Corpus. In Oostdijk, N., and de Haan, P. (eds.) *Corpus-based research into language: in honour of Jan Aarts*. Rodopi, Amsterdam, pp. 47 – 64.

Leech, G., Garside, R., and Bryant, M. (1994b) CLAWS4: The tagging of the British National Corpus. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)* Kyoto, Japan, pp. 622 – 628.

Leech, G. and Eyes, E. (1997) Syntactic annotation: treebanks. In Garside, R., Leech, G., and McEnery, A. (eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora.* Longman, London, pp. 34 – 52.

Leech, G. and Smith, N. (1999). The use of tagging. In van Halteren, H, (ed.) *Syntactic wordclass tagging*, Kluwer, The Netherlands, pp. 23 – 36.

Leech, G. and Wilson, A. (1999). Standards for tagsets. In van Halteren, H, (ed.) *Syntactic wordclass tagging*, Kluwer, The Netherlands, pp. 55 – 80.

Leech, G., Rayson, P., and Wilson, A. (2001). *Word frequencies in written and spoken English: based on the British National Corpus*. Longman, London.

Lehmann, H. M., Schneider, P., and Hoffmann, S. (2000). BNCweb. In Kirk, J. M. (ed.) *Corpora galore*. Rodopi, Amsterdam, pp. 259 – 266.

Ljung, M. (1974). *A frequency dictionary of English morphemes*. AWE/Gebers, Stockholm.

Lorge, I. (1949). *Semantic count of the 570 commonest English word*s. Columbia University Press, New York.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, International Business Machines Corporation, 2 (2), pp. 159 – 165.

Lyne, A. A. (1985). *The vocabulary of French business correspondence*. Slatkine, Geneva.

Lyne, A. A. (1986). In praise of Juilland's D: a contribution to the empirical evaluation of various measures of dispersion applied to word frequencies. In *proceedings of the Colloque International CNRS, Université de Nice, 5-8 June 1985, Méthodes quantitatives et informatiques dans l'étude des textes en hommage à Charles Muller*. Slatkine-Champion, Geneva, pp. 589 – 597.

McEnery, A. and Daille, B. (1993). Database Design for Corpus Storage: The ET10-63 Data Model. *UCREL Technical Papers* no. 1, Lancaster University.

McEnery, T., and Wilson, A. (1996) *Corpus Linguistics*. Edinburgh University Press, Edinburgh.

McEnery, A. and Rayson, P. (1997) A corpus/annotation toolbox. In Garside, R., Leech, G., and McEnery, A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London, pp. 194 – 208.

McEnery, A.M., Wilson, A, Sanchez-Leon, F. and Nieto-Serano, A. (1997). Multilingual Resources for European Languages: Contributions of the Crater Project, *Literary and Linguistic Computing*, Volume 12, Issue 4, OUP, Oxford, pp. 219 – 226.

McEnery, T., Piao, S., and Xin, X. (2000). Parallel alignment in English and Chinese. In Botley, S., McEnery, A., and Wilson, A. (eds.) *Multilingual corpora in teaching and research*. Rodopi, Amsterdam, pp. 177 – 191.

McEnery, T., Baker, P., and Hardie, A. (2000). Swearing and abuse in modern British English. In Lewandowska-Tomaszczyk, B. and Melia, P. J. (eds.) *PALC'99: Practical applications in language corpora: papers from the International conference at the University of Łódź, 15 – 18 April 1999*. Peter Lang, Frankfurt, pp. 37 – 48.

MacWhinney, B. (1991) *The CHILDES project: tools for analysing talk*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

MacWhinney, B. (1995) *The CHILDES project: tools for analysing talk*. Second edition. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

MacWhinney, B. and Snow, C. (1990). The Child Language Data Exchange System, *ICAME Journal* 14, pp. 3 – 25.

Mair, C. (1997). Parallel corpora: a real-time approach to the study of language change in progress. In Ljung, M. (ed.) *Corpus-based studies in English: papers from the seventeenth International Conference on English language research on computerized corpora (ICAME 17), Stockholm, May 15-19, 1996*, Rodopi, Amsterdam, pp. 195 – 209.

Marcus, M., Santorini, B., and Marcinkiewicz, M. (1993). Building a large annotated corpus of English: the Penn Treebank, *Computational Linguistics* 19(2), pp. 313 – 330.

Marshall, I. (1983). Choice of grammatical word-class without global syntactic analysis: tagging words in the LOB corpus. *Computers and the Humanities* 17, pp.139 – 150.

Mason, O. (2000). *Programming for corpus linguistics: how to do text analysis with Java*. Edinburgh University Press, Edinburgh.

Mehta, C. R. and Patel, N. R. (1983). A Network Algorithm for Performing Fisher's Exact Test in r x c Contingency Tables. *Journal of the American Statistical Association*, Vol. 78, No. 382. (Jun., 1983), pp. 427 – 434.

Mengel, A. and Lezius, W. (2000). An XML-based encoding format for syntactically annotated corpora. In *Proceedings of the Second International Conference on Language Resources and Engineering (LREC 2000)*, Athens, pp. 121 – 126.

Meunier, F. (1998). Computer tools for the analysis of learner corpora. In Granger, S. (ed.) *Learner English on computer*. Longman, London, pp. 19 – 37.

Meyer, C. F. (1991). A corpus-based study of apposition in English. In Aijmer, K. and Altenberg, B. (eds.), *English Corpus Linguistics: Studies in honour of Jan Svartvik*. Longman, London, pp. 166 – 181.

Milton, J. (1999). Lexical thickets and electronic gateways: making text accessible by novice writers. In Candlin, C. and Hyland, K. (eds.) *Writing: Texts, processes and practices*, Longman, London, pp. 221 – 243.

de Mönnink, I. (1997). Using corpus evidence and experimental data: a multimethod approach. In Ljung, M. (ed.) *Corpus-based studies in English: papers from the seventeenth International Conference on English language research on computerized corpora (ICAME 17), Stockholm, May 15-19, 1996*, Rodopi, Amsterdam, pp. 227 – 244.

Morrison, D. E. and Henkel, R. E. (1970). *The significance test controversy: a reader*. Butterworths, London.

Mosteller, F. and Rourke, R. E. K. (1973). *Sturdy statistics*. Reading, Massachusetts, Addison-Wesley.

Nelson, G., Wallis, S. and Aarts, B. (2002). *Exploring Natural Language: Working with the British Component of the International Corpus of English.* John Benjamins, Amsterdam.

Nevill-Manning, C. G., Witten, I. H., and Paynter, G. W. (1999). Lexically-Generated Subject Hierarchies for Browsing Large Collections. *International Journal of Digital Libraries*, 2 (2/3), pp. 111 – 123.

Noreen, E. W. (1989). *Computer-intensive methods for testing hypotheses: an introduction*. John Wiley and Sons, New York.

Oakes, M. P. (1998). *Statistics for corpus linguistics.* Edinburgh University Press, Edinburgh.

Ooi, V. B. Y. (2000). Asian or Western realities? Collocations in Singaporean-Malaysian English. In Kirk, J. M. (ed.) *Corpora galore: analyses and techniques in describing English.* Rodopi, Amsterdam, pp. 73 – 89.

Oostdijk, N. (1991). *Corpus linguistics and the automatic analysis of English*. Rodopi, Amsterdam.

Oostdijk, N. and de Haan, P. (eds.) (1994). *Corpus-based research into language: in honour of Jan Aarts*. Rodopi, Amsterdam.

Paice, C. D. (1977). *Information retrieval and the computer*. Macdonald and Jane's, London.

Pearson, K. (1904). On the theory of contingency and its relation to association and normal correlation. *Biometric Series* No. 1, Drapers' Co. Memoirs, London.

Pedersen, T. (1996). Fishing for exactness. In *Proceedings of the South-Central SAS Users Group Conference, Austin, TX*, pages 188 – 200.

Pedersen, T., Kayaalp, M. and Bruce, R. (1996). Significant lexical relationships. In *Proceedings of the 13th National Conference on Artificial Intelligence*. AAAI Press/The MIT Press, Menlo Park, CA, pp. 455 – 460.

Peppé, S. (1995). The Survey of English Usage and the London-Lund corpus: computerising manual prosodic transcription. In Leech, G., Myers, G., and Thomas, J. (eds.) *Spoken English on computer: transcription, mark-up and application.* Longman, London, pp. 187 – 202.

Pęzik, P. (forthcoming). Automated analysis of linguistic and ontological units. In Lewandowska-Tomaszczyk, B. (ed.) *Proceedings of PALC 2001*, Peter Lang, Frankfurt.

Popping, R. (1997). Computer programs for the analysis of texts and transcripts. In Roberts, C. W. (ed.) *Text analysis for the social sciences: methods for drawing statistical inferences from texts and transcripts*. Lawrence Erlbaum, New Jersey, pp. 209 – 221.

Quinn, A. (1993) An object-oriented design for a Corpus Utility Program. In Aarts, J., de Haan, P., and Oostdijk, N. (eds.) *English language corpora: Design, analysis and exploitation*. Rodopi, Amsterdam., pp. 215 – 225.

Rayson, P. and Wilson, A. (1996). The ACAMRIT semantic tagging system: progress report. In Evett, L. J., and Rose, T. G. (eds.) *Language Engineering for Document Analysis and Recognition (LEDAR), AISB96 Workshop proceedings*, Brighton, England. Faculty of Engineering and Computing, Nottingham Trent University, UK, pp. 13 – 20.

Rayson, P., Leech, G., and Hodges, M. (1997). Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*. 2 (1). John Benjamins, Amsterdam/Philadelphia, pp. 133 – 152.

Rayson, P., Garside, R., and Sawyer, P. (2000). Assisting requirements engineering with semantic document analysis. In *Proceedings of RIAO 2000 (Recherche d'Informations Assistie par Ordinateur, Computer-Assisted Information Retrieval) International Conference,* Collège de France, Paris, France, April 12-14, 2000. C.I.D., Paris, pp. 1363 – 1371.

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. In *proceedings of the workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000)*. 1-8 October 2000, Hong Kong, pp. 1 – 6.

Rayson, P., Wilson, A. and Leech, G. (2002) Grammatical word class variation within the British National Corpus sampler. In Peters, P., Collins, P., and Smith, A. (eds.) *New frontiers of corpus research: Papers from the Twenty First International Conference on English Language Research on Computerized Corpora, Sydney 2000*. Rodopi, Amsterdam, pp. 295 – 306.

Read, T. R. C. and Cressie, N. A. C. (1988). Goodness-of-fit statistics for discrete multivariate data. *Springer series in statistics*. Springer-Verlag, New York.

Renouf, A. (1987). Corpus development. In Sinclair, J. M. (ed.) *Looking up: an account of the COBUILD project in lexical computing*. Collins, London, pp. 1 – 40.

Renouf, A. (ed.) (1998). *Explorations in corpus linguistics*. Rodopi, Amsterdam.

Reppen, R. (2001). Review of MonoConc Pro and WordSmith Tools. *Language Learning & Technology*, Vol. 5, No. 3, May 2001, published electronically at http://llt.msu.edu/, pp. 32 – 36.

Ringbom, H. (1998). High-frequency verbs in the ICLE corpus. In Renouf, A. (ed.) *Explorations in corpus linguistics*. Rodopi, Amsterdam, pp. 191 – 200.

Rockwell, G., Passmore, G. and Bradley, J. (1997). TACTweb: The Intersection of Text-Analysis and Hypertext, *Journal of Educational Computing Research*, Vol. 17, no. 3, Baywood Publishing Company, Inc., New York, pp. 217 – 230.

Roland, D., Jurafsky, D., Menn, L., Gahl, S., Elder, E. and Riddoch, C. (2000). Verb Subcategorization Frequency Differences between Business-News and Balanced Corpora: the role of verb sense. In *proceedings of the workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000)*. 1-8 October 2000, Hong Kong, pp. 28 – 34.

Salton, G. and McGill, M. J. (1983). *Introduction to modern information retrieval*. McGraw-Hill, New York.

Sampson, G. (1987a). Probabilistic Models of Analysis. In Garside, R., Leech, G. and Sampson, G. (eds.) *The Computational Analysis of English: A Corpus-based Approach*. Longman, London, pp. 16 – 29.

Sampson, G. (1987b). The grammatical database and parsing scheme. In Garside, R., Leech, G. and Sampson, G. (eds.) *The Computational Analysis of English: A Corpus-based Approach*. Longman, London, pp. 82 – 96.

Sampson, G. (1995). *English for the computer: The SUSANNE corpus and analytic scheme*. Clarendon Press, Oxford.

Schmitt, N. and McCarthy, M. (eds.) (1997). *Vocabulary: description, acquisition and pedagogy*. Cambridge University Press, Cambridge.

Scott, M. (1996-99). *WordSmith: software tools for Windows, versions 1 – 3*. Oxford University Press, Oxford.

Scott, M. (1997). PC analysis of key words – and key key words. *System* 25 (2), Elsevier, pp. 233 – 245.

Scott, M. (2000a). Reverberations of an Echo. In Lewandowska-Tomaszczyk, B. and Melia, P. J. (eds.) *PALC'99: Practical applications in language corpora: papers from the International conference at the University of Łódź, 15 – 18 April 1999*. Peter Lang, Frankfurt, pp. 49 – 65.

Scott, M. (2000b). Focusing on the text and its key words. In Burnard, L. and McEnery, T. (eds.) *Rethinking language pedagogy from a corpus perspective: papers from the third international conference on teaching and language corpora.* Peter Lang, Frankfurt, pp. 104 – 121.

Scott, M. (2001a). Comparing corpora and identifying key words, collocations, and frequency distributions through the WordSmith Tools suite of computer programs, in Ghadessy, M., Henry, A. and Roseberry, R. L. (eds.) *Small corpus studies and ELT: theory and practice*. John Benjamins, Amsterdam, pp. 47 – 67.

Scott, M. (2001b). Mapping key words to *problem* and *solution*. In Scott, M. and Thompson, G. (eds.) *Patterns of Text: in honour of Michael Hoey*, Benjamins, Amsterdam, pp. 109 – 127.

Scott, M. and Johns, T. (1993). *MicroConcord software*. Oxford University Press, Oxford.

Short, M., Semino, E., and Culpeper, J. (1996). Using a corpus for stylistics research: speech and thought presentation. In Thomas, J. and Short, M. (eds.) *Using corpora for language research*. Longman, London, pp. 110 – 131.

Sin, K. K. and Roebuck, D. (1996). Language Engineering for Legal Transplantation: Conceptual Problems in Creating Common Law Chinese. *Language and Communication,* Vol. 16, No. 3, Elsevier, pp. 235 – 254.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Sinclair, J. (1995). Corpus typology – a framework for classification. In Melchers, G. and Warren, B. (eds.) *Studies in Anglistics*. Almqvist & Wiksell, Stockholm, pp. 17 – 33.

Sinclair, J. (1996). *EAGLES Preliminary recommendations on Corpus Typology* EAG--TCWG--CTYP/P Version of May, 1996, ILC-CNR, Pisa.

Sinclair, J. (1999). A way with common words. In Hasselgård, H. and Oksefjell, S. (eds.) *Out of corpora: studies in honour of Stig Johansson*. Rodopi, Amsterdam. pp. 157 – 179.

Smith, N. (1997). Improving a Tagger, in Garside, R., Leech, G., and McEnery, A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*, Longman, London, pp. 137 – 150.

Sparck Jones, K. (1971). *Automatic keyword classification for information retrieval*. Butterworths, London.

Sperberg-McQueen, C. and Burnard, L. (1990). *Guidelines for the encoding and interchange of machine-readable texts.* Draft version 1.0. Association for Computers and the Humanities, Association for Computational Linguistics, Association for Literary and Linguistic Computing, Chicago and Oxford.

Sperberg-McQueen, C. M. and Burnard, L. (eds.) (2002). *Guidelines for Text Encoding and Interchange (edition P4)*. Published for the TEI Consortium by the Humanities Computing Unit, University of Oxford. Distributed by University Press of Virginia.

Stiles, W. B. (1992). *Describing talk: a taxonomy of verbal response modes*. Sage, Beverly Hills.

Strang, J. (1986). *Programming with curses*. O'Reilly & Associates, Sebastopol, California.

Stubbs, M. (1993). British traditions in text analysis: from Firth to Sinclair. In Baker, M., Francis, G., and Tognini-Bonelli, E. (eds.) *Text and technology: in honour of John Sinclair*. Benjamins, The Netherlands, pp. 1 – 33.

Stubbs, M. (1996). *Text and corpus analysis: computer-assisted studies of language and culture*. Blackwell, Oxford.

Svartvik, J. and Ekedahl, O. (1995). Verbs in public and private speaking. In Aarts, B. and Meyer, C. (eds.) *The verb in contemporary English*. Cambridge University Press, Cambridge, pp. 273 – 289.

Taylor, L.J. and Knowles, G. (1988). *Manual of Information to Accompany the SEC Corpus*. UCREL, Lancaster University.

Teubert, W. (2001). A province of a federal superstate, ruled by an unelected bureaucracy – keywords of the Euro-sceptic discourse in Britain. In Musolff, A., Good, C., Points, P., Wittlinger, R. (eds.) *Attitudes towards Europe: language in the unification process*. Ashgate, Aldershot, pp. 45 – 86.

Thomas, J. and Wilson, A. (1996). Methodologies for studying a corpus of doctor-patient interaction. In J. Thomas and M. Short (eds.) *Using corpora for language research*. Longman, London, pp 92 – 109.

Thompson, H. and McKelvie, D. (1996) A Software Architecture for SGML Annotation. *Proceedings of SGMLEurope 96, Oxford*. SoftQuad, pp. 172 – 177..

Thorndike, E. L. (1921). *Teacher's word boo*k, Columbia Teachers College, New York.

Thorndike, E. L. (1932). *A teacher's word book of 20,000 word*s, Columbia Teachers College, New York.

Thorndike, E. L. and Lorge, I. (1944). *The teacher's word book of 30,000 word*s. Columbia University Press, New York.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Benjamins, The Netherlands.

Tribble, C. and Jones, G. (1997). *Concordances in the classroom*. Athelstan, Houston, Texas.

Tribble, C. (2000). Genres, keywords, teaching: towards a pedagogic account of the language of project proposals. In Burnard, L. and McEnery, T. (eds.) *Rethinking language pedagogy from a corpus perspective: papers from the third international conference on teaching and language corpora.* Peter Lang, Frankfurt, pp. 75 – 90.

Tribble, C. (2001). Small corpora and teaching writing. In Ghadessy, M., Henry, A. and Roseberry, R. L. (eds.) *Small corpus studies and ELT: theory and practice*, John Benjamins, Amsterdam, pp. 381 – 408.

Tsujii, J. (2000). Generic NLP technologies: language, knowledge and information extraction. In Proceedings of 38th Annual meeting of the Association for Computational Linguistics (ACL 2000), 1-8th October 2000, Hong Kong, pp. 11 – 18.

Uzar, R. and Waliński, J. (2000). A comparability toolkit: some practical issues for terminology extraction. In Lewandowska-Tomaszczyk, B. and Melia, P. J. (eds.) *PALC'99: Practical applications in language corpora: papers from the International conference at the University of Łódź, 15 – 18 April 1999*. Peter Lang, Frankfurt, pp. 445 – 457.

Virtanen, T. (1997). The progressive in NS and NNS student compositions: evidence from the International Corpus of Learner English. In Ljung, M. (ed.) *Corpus-*

based studies in English: papers from the seventeenth International Conference on English language research on computerized corpora (ICAME 17), Stockholm, May 15-19, 1996, Rodopi, Amsterdam, pp. 299 – 309.

Voutilainen, A. (1999). A short history of tagging. In van Halteren, H. (ed.) *Syntactic wordclass tagging*, Kluwer, Netherlands, pp. 9 – 21.

Wall, L., Christiansen, T., and Schwartz, R.L. (1996). *Programming Perl 2nd Edition.* O'Reilly & Associates, Sebastopol, California.

Weeber, M, Vos, R., and Baayen, R. H. (2000) Extracting the Lowest Frequency Words: Pitfalls and Possibilities. *Computational Linguistics* 26(3), MIT Press, pp. 301 – 317.

West, M. (1953). *A general service list of English word*s. Longman, London.

Wichmann, A. (1991). *Beginnings, Middles and Ends: A Study of Initiality and Finality in the Spoken English Corpus*. Unpublished Ph.D. thesis, Lancaster University.

Wikberg, K. (1997). On the study of discourse and style using the techniques of corpus linguistics. In Ljung, M. (ed.) *Corpus-based studies in English: papers from the seventeenth International Conference on English language research on computerized corpora (ICAME 17), Stockholm, May 15-19, 1996*, Rodopi, Amsterdam, pp. 311 – 327.

Wikberg, K. (1999). The style marker *as if (though)*: a corpus study. In Hasselgård, H. and Oksefjell, S. (eds.) *Out of corpora: studies in honour of Stig Johansson*. Rodopi, Amsterdam. pp. 93 – 105.

Williams, K. (1976). The failure of Pearson's goodness of fit statistic. *The Statistician*, Vol. 25, No. 1, Royal Statistical Society, pp. 49.

Williams, R. (1983). *Keywords: a vocabulary of culture and society, 2nd edition*, Fontana Press, London.

Wilson, A. (1991). No, not, and never: negation in a corpus of spoken interview transcripts. *Lancaster papers in Linguistics*, number 73. Linguistics Department, Lancaster University.

Wilson, A. (1993). Towards an integration of content analysis and discourse analysis: the automatic linkage of key relations in text. *UCREL technical paper 3*, Lancaster University.

Wilson, A. (1997). Conceptual analysis of later Latin texts: a conceptual glossary and index to the Latin Vulgate translation of the Gospel of John. PhD thesis, Lancaster University.

Wilson, A. and Rayson, P. (1993). Automatic content analysis of spoken discourse: a report on work in progress. In Souter, C. and Atwell, E. (eds*.), Corpus based computational linguistics*. Rodopi, Amsterdam, pp. 215 – 226.

Winograd, T. (1972) *Understanding natural language*. Edinburgh University Press, Edinburgh.

Woods, A., Fletcher, P., and Hughes, A. (1986). *Statistics in language studies*. Cambridge University Press, Cambridge.

Xunfeng, X. and Kawecki, R. (2001). *The Use of an On-line Trilingual Corpus for the Teaching of Reading Comprehension in French*. Presented at The 6th TELRI Seminar 'Multilingual Corpus Research', Bansko, Bulgaria, 9 – 11 November 2001.

Yates, F. (1934). Contingency tables involving small numbers and the chi-squared test. *Journal of the Royal Statistical Society Supplement*. Volume 1, pp. 217 – 235.

Yates, F. (1984). Tests of significance for 2 × 2 contingency tables. *Journal of the Royal Statistical Society, A*, 147, part 3, pp. 426 – 463.

Yeh, A. (2000). More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, Saarbrücken, Germany, July 2000, pp. 947 – 953.

Yule, G. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge University Press, Cambridge.

Zhou, J. (1999). Phrasal terms in real-world IR applications. In Strzalkowski (ed.) *Natural language information retrieval*, Kluwer, The Netherlands, pp. 215 – 259.

Zipf, G. K. (1935). *The Psychobiology of language*. Houghton Mifflin, New York.

Zipf, G. K. (1949). *Human behaviour and the principle of least effort*. Addison-Wesley, Reading.

Appendix A. CLAWS C7 tagset

APPGE	possessive pronoun, pre-nominal (e.g. my, your, our)
AT	article (e.g. the, no)
AT1	singular article (e.g. a, an, every)
BCL	before-clause marker (e.g. in order (that),in order (to))
CC	coordinating conjunction (e.g. and, or)
CCB	adversative coordinating conjunction (but)
CS	subordinating conjunction (e.g. if, because, unless, so, for)
CSA	as (as conjunction)
CSN	than (as conjunction)
CST	that (as conjunction)
CSW	whether (as conjunction)
DA	after-determiner or post-determiner capable of pronominal function (e.g. such, former, same)
DA1	singular after-determiner (e.g. little, much)
DA2	plural after-determiner (e.g. few, several, many)
DAR	comparative after-determiner (e.g. more, less, fewer)
DAT	superlative after-determiner (e.g. most, least, fewest)
DB	before determiner or pre-determiner capable of pronominal function (all, half)
DB2	plural before-determiner (both)
DD	determiner (capable of pronominal function) (e.g any, some)
DD1	singular determiner (e.g. this, that, another)
DD2	plural determiner (these, those)
DDQ	wh-determiner (which, what)
DDQGE	wh-determiner, genitive (whose)
DDQV	wh-ever determiner, (whichever, whatever)
EX	existential there
FO	formula

FU	unclassified word
FW	foreign word
GE	Germanic genitive marker - (' or's)
IF	for (as preposition)
II	general preposition
IO	of (as preposition)
IW	with, without (as prepositions)
JJ	general adjective
JJR	general comparative adjective (e.g. older, better, stronger)
JJT	general superlative adjective (e.g. oldest, best, strongest)
JK	catenative adjective (able in be able to, willing in be willing to)
MC	cardinal number, neutral for number (two, three..)
MC1	singular cardinal number (one)
MC2	plural cardinal number (e.g. sixes, sevens)
MCGE	genitive cardinal number, neutral for number (two's, 100's)
MCMC	hyphenated number (40-50, 1770-1827)
MD	ordinal number (e.g. first, second, next, last)
MF	fraction, neutral for number (e.g. quarters, two-thirds)
ND1	singular noun of direction (e.g. north, southeast)
NN	common noun, neutral for number (e.g. sheep, cod, headquarters)
NN1	singular common noun (e.g. book, girl)
NN2	plural common noun (e.g. books, girls)
NNA	following noun of title (e.g. M.A.)
NNB	preceding noun of title (e.g. Mr., Prof.)
NNL1	singular locative noun (e.g. Island, Street)
NNL2	plural locative noun (e.g. Islands, Streets)
NNO	numeral noun, neutral for number (e.g. dozen, hundred)
NNO2	numeral noun, plural (e.g. hundreds, thousands)
NNT1	temporal noun, singular (e.g. day, week, year)
NNT2	temporal noun, plural (e.g. days, weeks, years)
NNU	unit of measurement, neutral for number (e.g. in, cc)
NNU1	singular unit of measurement (e.g. inch, centimetre)
NNU2	plural unit of measurement (e.g. ins., feet)
NP	proper noun, neutral for number (e.g. IBM, Andes)

NP1	singular proper noun (e.g. London, Jane, Frederick)
NP2	plural proper noun (e.g. Browns, Reagans, Koreas)
NPD1	singular weekday noun (e.g. Sunday)
NPD2	plural weekday noun (e.g. Sundays)
NPM1	singular month noun (e.g. October)
NPM2	plural month noun (e.g. Octobers)
PN	indefinite pronoun, neutral for number (none)
PN1	indefinite pronoun, singular (e.g. anyone, everything, nobody, one)
PNQO	objective wh-pronoun (whom)
PNQS	subjective wh-pronoun (who)
PNQV	wh-ever pronoun (whoever)
PNX1	reflexive indefinite pronoun (oneself)
PPGE	nominal possessive personal pronoun (e.g. mine, yours)
PPH1	3rd person sing. neuter personal pronoun (it)
PPHO1	3rd person sing. objective personal pronoun (him, her)
PPHO2	3rd person plural objective personal pronoun (them)
PPHS1	3rd person sing. subjective personal pronoun (he, she)
PPHS2	3rd person plural subjective personal pronoun (they)
PPIO1	1st person sing. objective personal pronoun (me)
PPIO2	1st person plural objective personal pronoun (us)
PPIS1	1st person sing. subjective personal pronoun (I)
PPIS2	1st person plural subjective personal pronoun (we)
PPX1	singular reflexive personal pronoun (e.g. yourself, itself)
PPX2	plural reflexive personal pronoun (e.g. yourselves, themselves)
PPY	2nd person personal pronoun (you)
RA	adverb, after nominal head (e.g. else, galore)
REX	adverb introducing appositional constructions (namely, e.g.)
RG	degree adverb (very, so, too) RGQ wh- degree adverb (how)
RGQV	wh-ever degree adverb (however)
RGR	comparative degree adverb (more, less)
RGT	superlative degree adverb (most, least)
RL	locative adverb (e.g. alongside, forward)
RP	prep. adverb, particle (e.g about, in)
RPK	prep. adv., catenative (about in be about to)

RR	general adverb
RRQ	wh- general adverb (where, when, why, how)
RRQV	wh-ever general adverb (wherever, whenever)
RRR	comparative general adverb (e.g. better, longer)
RRT	superlative general adverb (e.g. best, longest)
RT	quasi-nominal adverb of time (e.g. now, tomorrow)
TO	infinitive marker (to)
UH	interjection (e.g. oh, yes, um)
VB0	be, base form (finite i.e. imperative, subjunctive)
VBDR	were
VBDZ	was
VBG	being
VBI	be, infinitive (To be or not... It will be ..)
VBM	am
VBN	been
VBR	are
VBZ	is
VD0	do, base form (finite)
VDD	did
VDG	doing
VDI	do, infinitive (I may do... To do...)
VDN	done
VDZ	does
VH0	have, base form (finite)
VHD	had (past tense)
VHG	having
VHI	have, infinitive
VHN	had (past participle)
VHZ	has
VM	modal auxiliary (can, will, would, etc.)
VMK	modal catenative (ought, used)
VV0	base form of lexical verb (e.g. give, work)
VVD	past tense of lexical verb (e.g. gave, worked)
VVG	-ing participle of lexical verb (e.g. giving, working)

VVGK	-ing participle catenative (going in be going to)
VVI	infinitive (e.g. to give... It will work...)
VVN	past participle of lexical verb (e.g. given, worked)
VVNK	past participle catenative (e.g. bound in be bound to)
VVZ	-s form of lexical verb (e.g. gives, works)
XX	not, n't
ZZ1	singular letter of the alphabet (e.g. A,b)
ZZ2	plural letter of the alphabet (e.g. A's, b's)

Appendix B. USAS tagset

A1	General and abstract terms
A1.1.1	General actions, making etc.
A1.1.2	Damaging and destroying
A1.2	Suitability
A1.3	Caution
A1.4	Chance, luck
A1.5	Use
A1.5.1	Using
A1.5.2	Usefulness
A1.6	Physical/mental
A1.7	Constraint
A1.8	Inclusion/Exclusion
A1.9	Avoiding
A2	Affect
A2.1	Affect:- Modify, change
A2.2	Affect:- Cause/Connected
A3	Being
A4	Classification
A4.1	Generally kinds, groups, examples
A4.2	Particular/general; detail
A5	Evaluation
A5.1	Evaluation:- Good/bad
A5.2	Evaluation:- True/false
A5.3	Evaluation:- Accuracy
A5.4	Evaluation:- Authenticity
A6	Comparing
A6.1	Comparing:- Similar/different
A6.2	Comparing:- Usual/unusual

A6.3	Comparing:- Variety
A7	Definite (+ modals)
A8	Seem
A9	Getting and giving; possession
A10	Open/closed; Hiding/Hidden; Finding; Showing
A11	Importance
A11.1	Importance: Important
A11.2	Importance: Noticeability
A12	Easy/difficult
A13	Degree
A13.1	Degree: Non-specific
A13.2	Degree: Maximizers
A13.3	Degree: Boosters
A13.4	Degree: Approximators
A13.5	Degree: Compromisers
A13.6	Degree: Diminishers
A13.7	Degree: Minimizers
A14	Exclusivizers/particularizers
A15	Safety/Danger
B1	Anatomy and physiology
B2	Health and disease
B3	medicines and medical treatment
B4	Cleaning and personal care
B5	Clothes and personal belongings
C1	Arts and crafts
E1	General emotional actions, states and processes
E2	Liking
E3	Calm/Violent/Angry
E4	Happy/sad
E4.1	Happy/sad: Happy
E4.2	Happy/sad: Contentment
E5	Fear/bravery/shock
E6	Worry, concern, confident
F1	Food

F2	Drinks
F3	Cigarettes and drugs
F4	Farming & Horticulture
G1	Government, Politics and elections
G1.1	Government etc.
G1.2	Politics
G2	Crime, law and order
G2.1	Crime, law and order: Law and order
G2.2	General ethics
G3	Warfare, defence and the army; weapons
H1	Architecture and kinds of houses and buildings
H2	Parts of buildings
H3	Areas around or near houses
H4	Residence
H5	Furniture and household fittings
I1	Money generally
I1.1	Money: Affluence
I1.2	Money: Debts
I1.3	Money: Price
I2	Business
I2.1	Business: Generally
I2.2	Business: Selling
I3	Work and employment
I3.1	Work and employment: Generally
I3.2	Work and employment: Professionalism
I4	Industry
K1	Entertainment generally
K2	Music and related activities
K3	Recorded sound etc.
K4	Drama, the theatre and show business
K5	Sports and games generally
K5.1	Sports
K5.2	Games
K6	Children’s games and toys

L1	Life and living things
L2	Living creatures generally
L3	Plants
M1	Moving, coming and going
M2	Putting, taking, pulling, pushing, transporting &c.
M3	Vehicles and transport on land
M4	Shipping, swimming etc.
M5	Aircraft and flying
M6	Location and direction
M7	Places
M8	Remaining/stationary
N1	Numbers
N2	Mathematics
N3	Measurement
N3.1	Measurement: General
N3.2	Measurement: Size
N3.3	Measurement: Distance
N3.4	Measurement: Volume
N3.5	Measurement: Weight
N3.6	Measurement: Area
N3.7	Measurement: Length & height
N3.8	Measurement: Speed
N4	Linear order
N5	Quantities
N5.1	Entirety; maximum
N5.2	Exceeding; waste
N6	Frequency etc.
O1	Substances and materials generally
O1.1	Substances and materials generally: Solid
O1.2	Substances and materials generally: Liquid
O1.3	Substances and materials generally: Gas
O2	Objects generally
O3	Electricity and electrical equipment
O4	Physical attributes

O4.1	General appearance and physical properties
O4.2	Judgement of appearance (pretty etc.)
O4.3	Colour and colour patterns
O4.4	Shape
O4.5	Texture
O4.6	Temperature
P1	Education in general
Q1	Linguistic actions, states and processes; communication
Q1.1	Linguistic actions, states and processes; communication
Q1.2	Paper documents and writing
Q1.3	Telecommunications
Q2	Speech acts
Q2.1	Speech etc:- Communicative
Q2.2	Speech acts
Q3	Language, speech and grammar
Q4	The Media
Q4.1	The Media:- Books
Q4.2	The Media:- Newspapers etc.
Q4.3	The Media:- TV, Radio and Cinema
S1	Social actions, states and processes
S1.1	Social actions, states and processes
S1.1.1	Social actions, states and processes
S1.1.2	Reciprocity
S1.1.3	Participation
S1.1.4	Deserve etc.
S1.2	Personality traits
S1.2.1	Approachability and Friendliness
S1.2.2	Avarice
S1.2.3	Egoism
S1.2.4	Politeness
S1.2.5	Toughness; strong/weak
S1.2.6	Sensible
S2	People
S2.1	People:- Female

S2.2	People:- Male
S3	Relationship
S3.1	Relationship: General
S3.2	Relationship: Intimate/sexual
S4	Kin
S5	Groups and affiliation
S6	Obligation and necessity
S7	Power relationship
S7.1	Power, organizing
S7.2	Respect
S7.3	Competition
S7.4	Permission
S8	Helping/hindering
S9	Religion and the supernatural
T1	Time
T1.1	Time: General
T1.1.1	Time: General: Past
T1.1.2	Time: General: Present; simultaneous
T1.1.3	Time: General: Future
T1.2	Time: Momentary
T1.3	Time: Period
T2	Time: Beginning and ending
T3	Time: Old, new and young; age
T4	Time: Early/late
W1	The universe
W2	Light
W3	Geographical terms
W4	Weather
W5	Green issues
X1	Psychological actions, states and processes
X2	Mental actions and processes
X2.1	Thought, belief
X2.2	Knowledge
X2.3	Learn

X2.4	Investigate, examine, test, search
X2.5	Understand
X2.6	Expect
X3	Sensory
X3.1	Sensory:- Taste
X3.2	Sensory:- Sound
X3.3	Sensory:- Touch
X3.4	Sensory:- Sight
X3.5	Sensory:- Smell
X4	Mental object
X4.1	Mental object:- Conceptual object
X4.2	Mental object:- Means, method
X5	Attention
X5.1	Attention
X5.2	Interest/boredom/excited/energetic
X6	Deciding
X7	Wanting; planning; choosing
X8	Trying
X9	Ability
X9.1	Ability:- Ability, intelligence
X9.2	Ability:- Success and failure
Y1	Science and technology in general
Y2	Information technology and computing
Z0	Unmatched proper noun
Z1	Personal names
Z2	Geographical names
Z3	Other proper names
Z4	Discourse Bin
Z5	Grammatical bin
Z6	Negative
Z7	If
Z8	Pronouns etc.
Z9	Trash can
Z99	Unmatched