CORPUS LINGUISTICS FOR HISTORY:

THE METHODOLOGY OF INVESTIGATING PLACE-NAME DISCOURSES IN DIGITISED NINETEENTH-CENTURY NEWSPAPERS

AMELIA T. JOULAIN-JAY

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

LANCASTER UNIVERSITY

OCTOBER 2017

ABSTRACT

The increasing availability of historical sources in a digital form has led to calls for new forms of reading in history. This thesis responds to these calls by exploring the potential of approaches from the field of corpus linguistics to be useful to historical research. Specifically, two sets of methodological issues are considered that arise when corpus linguistic methods are used on digitised historical sources.

The first set of issues surrounds optical character recognition (OCR), computerised text transcription based on image reproduction of the original printed source. This process is error-prone, which leads to potentially unreliable word-counts. I find that OCR errors are very varied, and more different from their corrections than natural spelling variation from a standard form. As a result of OCR errors, the test OCR corpus examined has a slightly inflated overall token count (as compared to a hand-corrected *gold standard*), and a vastly inflated type count. Not all spurious types are infrequent: around 7% of types occurring at least 10 times in my test OCR corpus are spurious. I also find evidence that real-word errors occur.

Assessing the impact of OCR errors on two common collocation statistics, Mutual Information (MI) and Log-Likelihood (LL), I find that both are affected by OCR errors. This analysis also provides evidence that OCR errors are *not* homogenously distributed throughout the corpus. Nevertheless, for small collocation spans, MI rankings are broadly reliable in OCR data, especially when used in combination with an LL threshold. Large spans are best avoided, as both statistics become increasingly less reliable in OCR data, when used with larger spans. Both statistics attract non-negligible rates of false positives. Using a frequency floor will eliminate many OCR errors, but does not reduce the rates of MI and LL false positives.

Assessing the potential of two post-OCR correction methods, I find that VARD, a program designed to standardise natural spelling variation, proves unpromising for dealing with OCR errors. By contrast, Overproof, a commercial system designed for OCR errors, *is* effective, and its application leads to substantial improvements in the reliability of MI and LL, particularly for large spans.

The second set of issues relate to the effectiveness of approaches to analysing the discourses surrounding place-names in digitised nineteenth-century newspapers. I single out three approaches to identifying place-names mentioned in large amounts of text without the need for a geo-parser system. The first involves relying on USAS, a semantic tagger, which has a 'Z2' tag for geographic names. This approach cannot identify multi-word place-names, but is scalable. A difficulty is that frequency counts of place-names do not account for their possible polysemy; I suggest a procedure involving reading a random sample of concordance lines for each place-name, in order to obtain an estimate of the actual number of mentions of that place-name in reference to a specific place. This method is best used to identify the most frequent place-names. A second, related, approach is to automatically compare a list of words tagged 'Z2' with a *gazetteer*, a reference list of place-names. This method, however, suffers from the same difficulties as the previous one, and is best used when accurate frequency counts are not required. A third approach involves starting from a principled, text-external, list of place-names, such as a population table, then attempting to locate each place in the set of texts. The scalability of this method depends on the length of the list of place-names, but it can accommodate any quantity of text. Its advantage over the two other methods is that it helps to contextualise the findings and can help identify place-names which are *not* mentioned in the texts.

Finally, I consider two approaches to investigating the discourses surrounding place-names in large quantities of text. Both are scalable operationalisations of proximity-based collocation. The first approach starts with the whole corpus, searching for the place-name of interest and generating a list of statistical collocates of the place-name; these collocates can then be further categorised and analysed via concordance analysis. The second approach starts with small samples of concordance lines for the place-name of interest, and involves analysing these concordance lines to develop a framework for description of the phraseologies within which place-names are mentioned. Both methods are useful and scalable; the findings they yield are, to some extent, overlapping, but also complementary. This suggests that both methods may be fruitfully used together, albeit neither is ideally-suited for comparing results across corpora. Both approaches are well-suited for exploratory research.

ACKNOWLEDGEMENTS

I wish to acknowledge the large number of people who have contributed to making this thesis possible. One of its conclusions is that collaboration is invaluable for progress; the length of this list may serve to demonstrate this. Although any errors remain mine, I am indebted to the people mentioned here.

From an institutional point of view, I am of course indebted to Lancaster University, and more specifically to the departments of History, and of English Language and Linguistics, who have supported me in both logistic and personal ways namely through respectively, the *Spatial Humanities: Texts, GIS, Places* project, which funded this thesis, and the Centre for Corpus Approaches to Social Science, who hosted me for the duration of my research. The support of these two last entities was particularly abundant and varied. My supervisors, of course, merit particular mention: Andrew Hardie, whose thoroughness was matched only by his patience and humanity, and Ian Gregory, whose unfaltering enthusiasm and creativity helped me find direction where I saw none. Of the others associated with these entities, I wish to mention particularly Carmen Dayrell and Dana Gablasova, who shared their office with me and oftentimes a hearing ear and useful advice; Ruth Avon, my colleague, and especially, my friend; Paul Atkinson, Cat Porter, Paul Rayson, CJ Rupp, Andrew Moore, Daniel Kershaw, Chris Donaldson, and Paty Murrieta-Flores, who often provided crucial insights or support and whose work as part of the *Spatial Humanities* project is closely related to mine; the gang of fellow CASS PhD students whose friendliness and warmth are second to none, Rosie Knight, Ruth Byrne, Robbie Love, Jen Hughes, Gill Smith and Abi Hawtin; and finally, Amanda Potts, who taught me corpus linguistics in the first place, and to whom I owe my landing there.

In the History department, James Taylor, Stephen Pumfrey, and Michael Hughes merit particular thanks for the interest they took in my work and their insights and encouragement, as well as Deborah Sutton and Mercedes Camino, who were my directors of studies and constantly held ready a smile and time for a chat. I also want to celebrate the welcoming History department family of PhD students, who I was honoured to join. In the Linguistics department, Chris Hart was always up for an interesting discussion and gave me a book on contemporary news coverage of Russia, Silke Brandt supported me in my first attempts at giving undergraduate seminars, and Vicki Haslam, who has been immensely helpful ever since I first stepped onto campus for a visit-day, granted me extra time to mark essays at the expense of her week-end when my uncle unexpectedly passed away.

Also at Lancaster, Alistair Baron graciously allowed me to use VARD and DICER and found time in his busy schedule to answer my many questions about them. Jake Arkinstall programmed from scratch a tailor-made collocation calculator for me in Python, just for fun, and not only over-night, but in clear and well-commented code. And Joanne Wood was a wonderful boss and colleague who valued my work as a Writing Mentor.

I could hardly forget Laurence Anthony, who spent hours drilling me on how to locate the correct form hidden by an OCR error, and whose enthusiasm is contagious. Thanks are also due to Sebastian Hoffman, who took time over Skype to discuss the approach he adopted to correct OCR errors in his collection of Singaporean newspapers. The department of English at Uppsala University, Sweden, has earned my gratitude for a warm welcome. Particular thanks go to Erik Smitterberg who allowed me to use his hard work, in the form of the Corpus of Nineteenth Century Newspaper English, for my research on OCR errors.

I am indebted to the Research Society for Victorian Periodicals (RSVP), whose members' encouragement and interest shaped the direction of my work; the RSVP granted me the 2016 Gale Dissertation Research Fellowship in Nineteenth-Century Media – I hope this thesis lives up to it. Special thanks are owed to Bob Nicholson, who allowed me to consult his thesis before I had begun any work of my own. Thanks must also go to the British Library staff, who were in contact with us throughout the project, and to Seth Cayley from Gale/Cengage for his enthusiasm and explanations.

No acknowledgement would be complete without naming my husband, Michael Jay, who not only proved an unending source of support but also showed genuine interest in the research, has probably read several versions of every passage in this thesis, and wrote the Matlab code which produced the beautiful histograms in chapters 4 and 5. My parents and sisters also deserve mention for showing support from a distance, even when they received only radio silence on my part.

Finally, I must also mention the administrators, the cleaners, the librarians, who make the world go round and yet are often forgotten, and the musicians (Richard Leigh, Roseanna Lea, Colin Steel, Deedee Bridgewater, Dianne Reeves, Tierney Sutton, Kate Rusby, Mark Tanner, ...) who make even the tedious enjoyable.

TABLE OF CONTENTS

ABBREVIATIONS

BL	British Library
c19th	nineteenth-century
CADS	corpus-assisted discourse studies
CASS	Centre for Corpus Approaches to Social Sciences
CDA	critical discourse analysis
CLAWS	Constituent Likelihood Automatic Word-tagging System
CNNE	Corpus of Nineteenth-Century Newspapers
CQP	Corpus Workbench Corpus Query Processor
DICER	Discovery and Investigation of Character Edit Rules
EEBO	Early English Books Online
EMEMT	Early Modern English Medical Texts
EN	English
ER	Error Rate
ERLN	*The Era*
ESRC	Economic and Social Research Council (UK)
FR	French
GIS	geographic(al) information system(s)
HPTE	*Hampshire/Portsmouth Telegraph*
KWIC	keyword-in-context
LL	Log Likelihood statistic
LR	Log Ratio statistic
MI	Mutual Information statistic
MT	machine translation
NINES	Networked Infrastructure for Nineteenth-Century Electronic Scholarship
NYT	*New York Times*
OCR	optical character recognition
PC	political correctness
PMGZ	*Pall Mall Gazette*
pmw	per million words
POS	part(s)-of-speech
RASIM	refugees, asylum seekers, immigrants and migrants
RDNP	*Reynold's Weekly Newspaper*
UCREL	University Centre for Computer Corpus Research on Language (Lancaster University)
UK	United Kingdom
US(A)	United States (of America)
USAS	UCREL Semantic Analysis System
VARD	VARiant Detector
XML	eXtensible Markup Language

Other 4 letter abbreviations refer to newspapers from the *19th Century British newspapers* collection, see full list in appendix 10.1. Capital letters are used throughout to distinguish references to lemmas and semantic fields from references to word-tokens or word-types.

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

1.1 GENERAL GOALS

This thesis is situated, both materially and intellectually, at the intersection of two projects based at Lancaster University (UK). The *Spatial-Humanities: Texts, GIS, Places* project, funded by the European Research Council[1] for 2012-2016, is a multi-disciplinary project committed to developing ways of investigating spatial patterns in qualitative data, especially through the use of Geographical Information Systems (GIS)[2], for the benefit of disciplines in the Humanities; a studentship from the project funded this thesis. Second, the *ESRC Centre for Corpus Approaches to Social Science* (CASS, 2013-2018) is a multi-disciplinary research centre funded by the UK Economic and Social Research Council[3], committed to applying corpus linguistic methods[4] to questions in the Social Sciences; CASS hosted this project.

Both projects are thus dedicated to applying computerized methodologies beyond their original disciplinary boundaries. In this spirit, the goal of this thesis is, first and foremost, to contribute to the methodology of History by clearing the methodological ground for the use of corpus linguistic methods in History. A second goal, which relates more specifically to the *Spatial Humanities*' research agenda, is to contribute to a methodology for investigating spatial patterns in large amounts of text. This thesis will hence prove relevant not just to Historians, but also to a wide range of scholars, particularly in the Humanities and Social Sciences, interested in this goal.

The aims of this thesis have also been crucially influenced by the nature of the dataset at my disposal: the British Library's *19th Century British Newspapers (part 1)* collection, which

[1] Funded under the European Union's Seventh Framework Programme (FP7/2007-2013), agreement number 283850.
[2] A methodology which makes uses of software to allow for the analysis and representation of large amounts of spatial data.
[3] Grant reference: ES/K002155/1.
[4] A methodology which makes uses of software to allow for the quantitative and qualitative analysis of large amounts of textual data, see also section 2.2.1.

consists of electronic texts generated using optical character recognition technologies (OCR)[5] with varying degrees of accuracy. Corpus linguistic methods applied to large amounts of textual data rely sooner or later on word-counts; OCR errors, then, since they affect the reliability of word-counts, constitute a major theoretical issue for the application of corpus linguistic methods to OCR data. This thesis will hence prove relevant not only to Historians, but also to scholars in a range of other disciplines interested in analysing large amounts of OCR data.

The project driving this thesis is hence inter-disciplinary in nature. Although it focuses on nineteenth century data, its methodological character will make it relevant to a wide range of scholars, particularly those in the Humanities and Social Sciences who will benefit from being able to analyse large amounts of textual data in both quantitative and qualitative ways. Since different readers may be expected to have groundings in varied disciplines, I have made few assumptions about their prior knowledge and have attempted to clarify discipline-specific terms when first used. Formatting and referencing conventions are also necessary, even if the choice of one discipline's practices rather than another's may be ultimately arbitrary; I have chosen to follow those most common in Linguistics.

A word of caution is necessary regarding the spelling of the examples from nineteenth century newspapers in this thesis. Throughout, when I give examples of OCR data, they have in general not been corrected and hence contain OCR errors, including misspellings, omissions and/or additions (as compared to the original sources). Providing the unmodified OCR text has the double advantage that it helps the reader assess the state of this OCR data, as well as making it easier for the reader to locate the examples in the source data[6]. On the other hand, it can make it harder for the reader to understand the examples. For this reason, where the state of the OCR is irrelevant to the discussion, I sometimes apply some manual correction to improve readability.

[5] A computerized way of turning images into text, see also section 3.2.1.

[6] Most users access the British Library's *19th Century Newspapers* collection via the Gale/Cengage web-portal (gale.cengage.co.uk/british-library-newspapers/19th-century-british-library-newspapers-part-i.aspx) which displays images searchable using an underlying layer of linked OCR data. As of this writing, that layer of OCR data should be identical to that used in this thesis.

1.2 RESEARCH AIMS

This study has two major research aims. Both of them are part of the overarching goal of broadening understanding of the strengths and limitations of using digital resources and methods in History. The first major research aim focuses on a particularly problematic issue: OCR errors and their impact on corpus linguistic methods. This aim encompasses the following research questions:

1. What is, in theory, the impact of OCR errors on the statistics of collocation[7]?
2. In practice, what are OCR errors like and how do they impact frequency figures?
3. In practice, how do OCR errors impact on two common collocation statistics?
4. How effective are two existing automated OCR correction techniques?
5. How much of an impact does the most promising correction technique have on the two collocation statistics?

The second aim is to establish and evaluate a methodology for investigating spatial patterns in large amounts of text. This aim encompasses devising and testing approaches to investigating questions such as 'what places are mentioned in this corpus?' and 'what is said about this place?' as well as evaluating approaches with regards to the following questions:

6. How feasible is the approach? (i.e. How resource-intensive is it? : how time-consuming, how computationally intensive, etc.)
7. How scalable is the approach? (i.e. Is the approach feasible with large amounts of data?)
8. What degree of granularity does the approach allow for? (i.e. Is the approach able to facilitate a certain level of qualitative detail when working with larger amounts of data? For example, can it facilitate comparisons across newspapers? Across genres? Over time?)

[7] Collocation analysis is the study of patterns of word co-occurrence, see section 2.3.2.3.

1.3 Organisation of the thesis

This thesis is divided into four parts. Part 1 consists of chapter 2, the literature review, which covers relevant work in History and Linguistics. On the History side, I review work in History which has used corpus linguistic methods[8], and demonstrate its sparseness to date. Moreover, I consider the increasingly frequent debates on the use by Historians and other humanists of digitised historical sources. I also review some methodologically exploratory work in this field that is directed towards tapping into the research potential of such digitised sources. On the Linguistics side, I review work which uses corpus linguistic methods to analyse newspaper texts, describing the variety of theoretical and methodological approaches that such work may adopt. I focus on studies which are not linguistically-motivated (e.g. having a focus on grammar), but which instead investigate language as a way of gaining insight into contemporary or historical social issues. This chapter includes an introduction to basic corpus linguistic methods and concepts, which will be relevant throughout the thesis. Readers who are not familiar with corpus linguistics may find it handy to refer back periodically to sections 2.2.1 (which introduces corpus linguistics in general) and 2.3.2 (which introduces basic corpus linguistic methods and concepts).

Part 2 consists of three chapters dealing with the thesis's first research aim: the issue of OCR errors and their impact on digital text analysis methods. Chapter 3 explores the issue of OCR errors, and outlines the *theoretical* implications of OCR errors for the computation of two common collocation statistics. Chapter 4 investigates the *empirical* impact of OCR errors on these two statistics, and formulates recommendations for working with OCR data using corpus linguistic approaches. Chapter 5 evaluates the effectiveness of two existing pieces of software for correcting OCR errors. It also discusses the impact of OCR error-corrections on the two collocation statistics when computed on OCR data processed using the most promising corrective approach.

[8] Corpus linguistics is a set of methods and tools which facilitate the qualitative and quantitative analysis of large amounts of digital texts, see section 2.2.1.

Part 3 consists of two chapters dealing with analytical approaches to investigating spatial patterns in nineteenth-century newspapers. Chapter 6 focuses on the question ‘what places are mentioned in these texts?’. It discusses three feasible approaches to answering this question without recourse to geo-parsing (an automated method of tagging place-names, see section 6.3.1.3), and evaluates these approaches in terms of their scalability and granularity of analysis. To illustrate the three approaches, this chapter goes on to compare patterns of mentions of British cities in three nineteenth century newspapers, with particular attention to the relationship between article genres and these patterns of mentions. Chapter 7 focuses on the question ‘what is said about this place in these texts?’, using France and Russia as case studies. It discusses, illustrates and evaluates two approaches to investigating discourses surrounding France and Russia in three nineteenth century newspapers.

Part 4 consists of the final, concluding chapter, which summarises the findings in the other chapters, discusses their broader implications, and outlines avenues for further research.

1.4 Chapter summary

In this introductory chapter, I have outlined the context and aims of my thesis, as well as the research questions associated with my two primary aims. I have also briefly summarised the structure of the thesis. In the next chapter, I move on to my review of relevant literature across the two disciplines that this thesis brings together.

PART 1: LITERATURE REVIEW

2 LITERATURE REVIEW

2.1 INTRODUCTION

Since this thesis aims to contribute to the methodology of History by exploring the application of corpus linguistic methods to the analysis of large amounts of historical newspapers, it is relevant to review, first, work done in History using corpus linguistics (this is done in section 2.2), and second, work in Linguistics which uses corpus linguistic methods to investigate newspapers (this is done in section 2.3).

2.2 WORK USING CORPUS LINGUISTICS WITHIN THE FIELD OF HISTORY

This section addresses the question: how and to what extent has corpus linguistics been used to benefit historical scholarship? Corpus linguistics is briefly introduced in section 2.2.1. Section 2.2.2 discusses the sparseness of historical work which has drawn on corpus linguistics. Section 2.2.3 explores salient aspects of the debate over the use of digitised source material within historical research that are relevant to the use of digital approaches (including corpus linguistics) in History. Section 2.2.4 gives an overview of work which, although not drawing on corpus linguistics directly, is methodologically exploratory and can be seen as an intermediate step between the traditional practices of historical scholarship and more methodologically innovative historical work that draws on corpus linguistic methods and concepts.

2.2.1 INTRODUCING CORPUS LINGUISTICS

Corpus linguistics is a now-established method of computerized textual analysis which allows for the processing of large bodies of texts called *corpora* (singular *corpus*) (see, for instance, McEnery and Hardie 2012). A corpus is essentially a digital collection of texts, but it differs from a generic electronic archive in that it is designed with a specific research purpose or purposes in mind. Corpus linguistic software can manipulate such a corpus in order to facilitate non-linear (i.e. other than beginning-to-end) readings, as an aid to both quantitative and qualitative analysis (Hunston 2002: 2).

Critically, corpus linguistics offers both software and conceptual tools which can assist a researcher in identifying linguistic trends across vast quantities of texts, trends which can then become the focus of qualitative analysis. Mautner (2007) exemplifies this clearly in her work on attitudes towards the elderly. She explains that quantitative and qualitative approaches are combined in her research, with early steps involving the use of corpus linguistic software to generate quantitative information about the large quantity of text she is investigating, and later steps involving a qualitative focus on 'particularly promising entry points into the data' which have been highlighted by the quantitative analyses (Mautner 2007: 55).

Corpus linguistics is defined in various ways (see Taylor 2008 for a survey of different definitions). In particular, there is an important conceptual distinction between views of corpus linguistics as a subfield of linguistics devoted to producing new theories of language based as much as possible exclusively on evidence from a corpus, and views of corpus linguistics as a set of methods for analysing corpora which can potentially be applied to a great number of problems within various areas of linguistics and beyond.

This distinction is discussed by Hardie and McEnery (2010), who use the name 'neo-Firthian' for the school of corpus linguistics associated with the first set of views, and the name 'methodologist' for the school associated with the latter. (The name 'neo-Firthian' stems from the theoretical foundation of much work produced by this school on ideas developed by the linguist J. R. Firth, in particular the concept of collocation, which will be elaborated on in section 2.3.2.3; but see also McEnery and Hardie 2012: 122-32.) This distinction also sometimes appears in the literature as one between 'corpus-driven' and 'corpus-based' research (terms originally suggested by Tognini-Bonelli 2001). Unfortunately, these terms have not always been used consistently, and as a result have become somewhat confusing. A critique of these terms and the original distinction drawn by Tognini-Bonelli can be found in Hardie and McEnery (2010) as well as McEnery and Hardie (2012: 147-52). The approach adopted in this thesis can be described as 'methodologist': here, I consider corpus linguistics to be a set of methodological tools and concepts which can be drawn on to investigate research questions both within and beyond traditional linguistic concerns. Nevertheless, as Hardie and McEnery (2010) emphasise, the divide between the two schools can easily be overstated, and in practice, scholars from both schools frequently interact and collaborate in a variety of academic contexts.

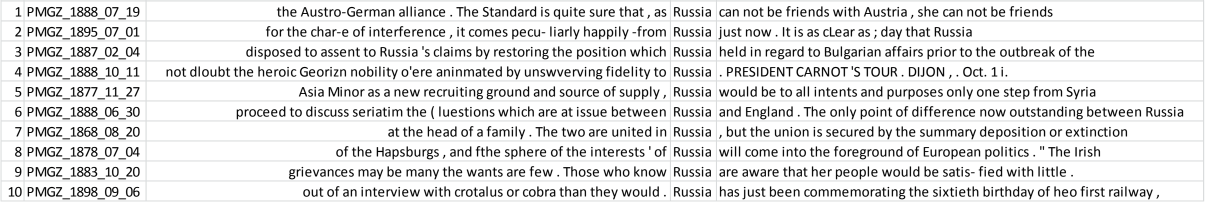
Corpus linguistics has, to date, seldom been referred to in the historiographical literature, as will be shown in the following sections. *Text mining*, on the other hand, is more often referred to and, so far, seems to have become familiar to more historians than has corpus linguistics. Since both involve using computers to process large quantities of text, the reader familiar with text mining but not corpus linguistics may be forgiven for confusing them. Indeed, the distinction is somewhat fuzzy. A simple distinction might be that corpus linguistics falls broadly within the scope of linguistics, whereas text mining is a subfield of the area of computer science known as *natural language processing*. This distinction is complicated, however, by the fact that corpus linguistics has developed alongside natural language processing (also known as *computational linguistics*).

The histories of both fields have been covered elsewhere (see Hirst 2013 for a history of computational linguistics; and McEnery and Hardie 2013 for a history of corpus linguistics). Nevertheless, a brief account of these histories may be helpful here.

McEnery and Hardie (2013), in their survey of the history of corpus linguistics, suggest that the field has involved, since its early days in the late 1950s, an approach to language analysis committed to looking at actual examples of language in use. This empirical approach to language was not new, they note, but the development of computer technology was required to allow for the development of the field (McEnery and Hardie 2013: 728).

Hence work of a pioneering nature could occur before the advent of computers, drawing on methodological concepts and approaches which anticipated those exploited in modern corpus linguistics. Early examples of concordances (which I come back to in section 2.3.2.1, but see Figure 2.1 for an example) were thus produced several centuries before the advent of computers. As Hunston (2012: 1366) notes, manual concordances of the Christian Bible, produced since the thirteenth century, 'are a direct precursor of the concordancing programs used in corpus linguistics'. Frequency lists (described in section 2.3.2.2) were also used to aid the analysis of extensive amounts of text as early as the late nineteenth century, as pointed out by McEnery and Hardie (2013: 728), who cite the example of Käding's (1897) work on a corpus of German.

Figure 2.1 Example of a concordance: 10 random instances of 'Russia' in the Pall Mall Gazette (PMGZ)



1	PMGZ_1888_07_19	the Austro-German alliance . The Standard is quite sure that , as	Russia	can not be friends with Austria , she can not be friends
2	PMGZ_1895_07_01	for the char-e of interference , it comes pecu- liarly happily -from	Russia	just now . It is as cLear as ; day that Russia
3	PMGZ_1887_02_04	disposed to assent to Russia 's claims by restoring the position which	Russia	held in regard to Bulgarian affairs prior to the outbreak of the
4	PMGZ_1888_10_11	not dloubt the heroic Georizn nobility o'ere aninmated by unswverving fidelity to	Russia	. PRESIDENT CARNOT 'S TOUR . DIJON , . Oct. 1 i.
5	PMGZ_1877_11_27	Asia Minor as a new recruiting ground and source of supply ,	Russia	would be to all intents and purposes only one step from Syria
6	PMGZ_1888_06_30	proceed to discuss seriatim the (luestions which are at issue between	Russia	and England . The only point of difference now outstanding between Russia
7	PMGZ_1868_08_20	at the head of a family . The two are united in	Russia	, but the union is secured by the summary deposition or extinction
8	PMGZ_1878_07_04	of the Hapsburgs , and fthe sphere of the interests ' of	Russia	will come into the foreground of European politics . " The Irish
9	PMGZ_1883_10_20	grievances may be many the wants are few . Those who know	Russia	are aware that her people would be satis- fied with little .
10	PMGZ_1898_09_06	out of an interview with crotalus or cobra than they would .	Russia	has just been commemorating the sixtieth birthday of heo first railway ,

Nevertheless, modern corpus linguistics could only take off with the advent of computers in the 1950s and 1960s, since this technology fulfilled the key requirement of corpus linguistics, which is the ability to store and manipulate huge quantities of textual data (McEnery and Hardie 2013: 728). McEnery and Hardie note that the move to using computers happened almost as soon as the technology was available, with Busa's pioneering work in the early 1950s even relying on the precursor technology of punched-card machines to generate concordances of Thomas Aquinas' poetry (McEnery and Hardie 2013: 728-29).

Important corpus linguistic work was hence undertaken from the late 1950s onwards, including landmark works such as that produced by Randolph Quirk's Survey of English Usage research unit, founded in 1959, and Francis and Kucera's Brown Corpus, published in 1964 (McEnery and Hardie 2013). Until around 1990, however, corpus linguistic work remained somewhat peripheral to the field of linguistics in general, and was essentially led by a small number of specific dedicated research centres. This is at least in part, McEnery and Hardie (2013) suggest, because in the 1960s and 1970s, the renowned linguist Noam Chomsky's ideas were extremely influential, and these ideas acted as an important impediment to the spread of corpus linguistic methods.

Why this impediment? In linguistics, there is a long-standing distinction between *language performance* and *language competence*. The former refers to the way in which individuals use language in context, whereas the latter refers to the abstract knowledge which those individuals draw upon when using ('performing') language in a given context. Chomksy's strongly argued position was that it is competence, not performance, which should be studied by linguists. The goal of linguistic theory is then to describe the native speaker's intuition (or abstract knowledge) of their language. Chomsky further argued that performance data cannot help linguistic theory, because any such data would necessarily consist of a skewed subset of the infinity of possible sentences of a language, a subset which moreover would contain 'performance errors and ungrammatical forms that do not adequately reflect the competence of the speakers that produced them' (McEnery and Hardie 2013: 730).

Hence, for Chomsky, linguistics should adopt a *rationalist* approach in order to produce linguistic theory, one in which 'explanations of the workings of the language system are arrived at via a native speaker linguist reflecting on their own knowledge of language and giving grammaticality judgements on artificially concocted sentences' (McEnery and Hardie 2013: 730). This view is very much opposed to the *empiricist* approach that motivated linguists to analyse corpora (which are collections of examples of language performance).

During the period in which Chomskyan ideas dominated, corpus linguistic research therefore remained peripheral. In contrast, McEnery and Hardie (2013: 741) suggest that a 'shift in the status of corpus linguistics' has taken place since 1990, with corpus linguistic methods now being adopted in most areas of linguistics by researchers who would not necessarily describe themselves as 'corpus linguists'.

Turning to computational linguistics (which includes, but is not limited to, text mining), Hirst (2013: 707) suggests that the field originated from the field of machine translation (MT) in the 1940s. The idea of machine translation was to produce software capable of translating any text from one language to another without additional human input. Unfortunately, much of this early research involved relatively naïve conceptions about the nature of language and translation, and 'by the early 1960s, it became clear that current approaches to MT were not successful' (Hirst 2013: 709). Although until then MT research had been well-funded, a committee established by the US National Academy of Sciences published a damning report into

MT in 1966, which led to drastic funding cuts and the termination of many MT research projects (Hirst 2013: 709).

The field of computational linguistics progressively 'disentangled' itself from the now tainted field of MT, and became viewed as a 'facet of the then-glamorous field' of artificial intelligence (Hirst 2013: 710). The main concern of computational linguistics at this stage was to produce computerized systems capable of 'understanding' (i.e. responding appropriately to) 'natural' language, or language as it is produced by human beings in normal circumstances. The idea was that such systems would be able to respond appropriately to a human user 'conversing' with these systems 'in order to instruct them or to seek information or advice from them' (Hirst 2013: 714). Work in this period drew inspiration from Chomskyan linguistics and hence had little interest for models of language based on extensive observation of actual instances of language use. Neither did it interact much with the work being carried out by corpus linguists. This changed in the early 1990s, when technological developments made the sharing and processing of corpora less costly and time-consuming. Hirst says that at this stage computational linguistics evolved 'from a rationalist enterprise inspired by artificial intelligence and armchair linguistics to an empiricist undertaking based on corpora and statistics' (Hirst 2013: 716).

For a period, the interests of corpus linguistics and computational linguistics hence converged around technical issues related to processing and analysing corpora. For computational linguists, the aim was to use corpora as a means to develop probabilistic models of language better equipped to deal with 'natural' language. For corpus linguists, the interest was to develop ways of manipulating corpora in order to facilitate analysis of the language contained in the corpus itself. Hence, although the fields seemed to converge for a while, they 'remained largely separate research fields with different motivations and methods of analysis' (Hirst 2013: 716).

This view is supported by McEnery and Hardie (2012: 228), who suggest that the fields 'converged greatly' in the 1980s and 1990s, but that 'this period of intersection seems to have passed, to some degree, except perhaps in the relatively narrow areas' related to forms of automated tagging (i.e. enriching the corpora with annotations to facilitate its indexing, searching and/or analysis). They hence draw a distinction between the motivations of corpus linguistics and computational linguistics, stating that 'corpus linguistics is ultimately about *finding out about the nature and usage of language*', whereas computational linguistics may incidentally 'be concerned with modelling the nature of language computationally' but focuses more specifically 'on *solving technical problems involving language'* (McEnery and Hardie 2012: 228). Hence, text mining, as an area of computational linguistics, and like other forms of data mining, is chiefly concerned with automatically extracting information from large datasets. In this, it is unlike corpus linguistics, which in the methodologist conception primarily aims to provide researchers with tools to examine, manipulate, and analyse the language contained in large datasets.

In summary, corpus linguistics is a set of computer-assisted methods developed to facilitate the quantitative and qualitative analysis of large amounts of text. Corpus linguistics can be defined in different ways. One school of corpus linguistics, the neo-Firthian school, defines it as a subarea of linguistics devoted to the development of novel theories of language relying on evidence from corpora. This thesis, along with the methodologist school, defines corpus linguistics as a set of methods rather than as a particular theoretical perspective. Nevertheless, some of the conceptual frameworks developed by the neo-Firthian school have wide currency among corpus linguists and will also be exploited here. Corpus linguistics has historically interacted with text mining and other aspects of computational linguistics, but, unlike text mining, has an important qualitative focus on the characteristics of language in particular contexts and is not conceived as a press-of-the-button way of extracting information from large amounts of text.

2.2.2 HISTORICAL WORK USING CORPUS LINGUISTICS

As we have seen, corpus linguistics is a set of methods and tools (including software and theoretical concepts) for analysing large quantities of text. In the context of the 'digital revolution' – which has suddenly made overwhelming quantities of text available to scholars – advances in corpus linguistics would be expected to benefit any scholar who is faced with a need to deal with large quantities of textual evidence. This is often the case for historians, particularly those using newspapers either as object of study *per se* or as a window onto some other object of study. However, although there has been an occasional article attempting to promote the use of corpus linguistics in History (e.g. Mahlberg 2010; Welling 2001), to date remarkably little work published in historical journals uses corpus linguistics. Among such work, Colella (2013) uses the ProQuest archive of British Periodicals (collection II) to uncover patterns of evaluation associated with three expressions ('man of business', 'business habits' and 'business life') in order to comment on changing attitudes towards business values in the nineteenth century; Pionke (2014) uses the ProQuest archive of British Periodicals (collections I and II) as well as AntConc[1] to explore Britain's role in Cuba's Victorian history, and Liddle (2015) uses AntConc to explore the development of the genre of 'leading articles'. Such work is pioneering in its use of corpus linguistics for History, but nevertheless has yet to exploit the full potential of corpus linguistic methods. Thus, much scope remains for bringing more of the affordances of corpus linguistics to the field of History.

Beyond historical journals, work using corpus linguistics on historical material is often published in linguistic or interdisciplinary journals, but does not seem to be making much impact within the field of History. Examples include Pumfrey et al. (2012), who trace changes in the use of the word 'experiment' in the 17th century in order to explore changing conceptions of science (a study published in the journal *Literary and Linguistics Computing*); and Prentice and Hardie (2009) and Bos (2012), studies which explore historical newspapers using corpus

[1] One of the most commonly used pieces of corpus linguistics software, freely available at http://www.laurenceanthony.net/software/antconc/ .

linguistic methods and software, but are integrated within sub-areas of linguistics such as 'historical pragmatics' or 'historical discourse analysis', rather than within the field of History.

This is not to say that the potential of digital technologies to transform historical scholarship has gone unremarked. On the contrary, historians have been pondering the implications of digital technologies for their work for several decades already, and these debates have become increasingly prominent within some journals and at certain conferences. The next section reviews some of these debates. Section 2.2.4 will review some of the methodologically exploratory work which *has* been published in historical journals.

2.2.3 Debates surrounding digitisation in historical journals

Since the 1990s, journals such as the *Journal of Victorian Studies,* the *Journal of Victorian Culture* and the *Victorian Periodical Review* have hosted increasingly frequent contributions on the topics of scholars' visions and concerns regarding advances in digital technologies. At the turn of the century, those scholars who were engaged in discussions about digital developments seemed mostly concerned with understanding the implications of the digitisation of historical material. One digitisation project which thus attracted comment was the *Rossetti Archive*, the first instalment of which was made available in Spring 2000, and which aims to facilitate the 'scholarly study' of Dante Gabriel Rossetti (a well-known nineteenth-century author and artist) by providing a digital edited collection of texts and images related to his work (see McGann n.d.). The project sparked debate about the usefulness and perils of working with digital rather than paper archives; see for example Potter's (1998) criticism of the project and McGann's (1998) response.

A recurrent concern voiced in the debates surrounding the *Rossetti Archive* and digitisation in general is that digital versions of printed historical material lack the materiality of the originals, thus leading to a loss of historically relevant context such as textures, smells and visual elements (see for example Towheed 2010, quoted below). In this regard, Mussell (2009: 93) suggests that historians faced with the digitisation of historical material have often placed emphasis ‘on what is lost’. Another concern is that the use of digital resources might displace the use of originals, leading to at best a bias towards digitised material and at worse a loss of skills in handling originals, or loss of the originals themselves. Leary (2005) famously referred to this bias towards digitised material as provoking an ‘offline penumbra’, since its consequence is the casting into obscurity of material unavailable in digital format irrespective of its historical significance. In the course of lamenting the phenomenon of libraries disposing of their physical archives, Towheed expresses a concern that ‘with the removal of this material archive, we lose valuable additional evidence of nineteenth-century reading practice, as well as significantly reduce the familiarity of contemporary readers with the size, shape, format, smell and texture of these books’ (2010: 141).

These concerns have been taken seriously by digital advocates. Nicholson – a historian and digital advocate who goes by the name ‘the digital Victorianist’ in his actively maintained online presence – summarises their source as relating to the distance between originals and digital versions: ‘by the time we access them, many digital newspapers have been remediated three times (single issue<bound volume<microfilm<digitisation); each step serves to distance us from the original text’ (Nicholson 2013: 61). To address this concern, digital advocates often emphasise the argument that digital versions should not be considered to be ‘surrogates’ of the originals; rather, they should be thought of as other ‘editions’ with their own editorial characteristics, requiring critical awareness of the way in which they are produced and how they differ from other existing editions. An excellent example in this respect is Fyfe (2016), who provides a critical history of the British Library's *19th Century Newspapers* collection. Hence Mussell, editor of the Digital Forum in the *Victorian Periodical Review* and author of *The Nineteenth-Century Press in the Digital Age* (Mussell 2012a) emphasises the point that ‘digital resources should not replace the material in the archive but instead complement it, providing another way to approach whatever is being studied’ (Mussell 2012b). Likewise, he stresses elsewhere that ‘it is important that we recognise the editorial work that goes into producing

digital resources and that we think seriously about the various transformations that material must undergo in order to be delivered on screen' (Mussell 2008: 98).

As Mussell's comment illustrates, the debate extends to another issue, namely the extent to which new skills and methods will need to be developed and integrated into the practices of historical scholarship in order to make use of digitised historical material for scholarly purposes. On the optimistic side, some, like Mussell, see such skills and methods as an extension and continuation of interpretative practices already developed in the field. Mussell (2012c) suggests that scholars (and historians in particular) interested in the study of the press are naturally concerned with mediation and its implications, and may thus be particularly well placed to apply their critical skills to reflecting about the editorial process of creating such digital 'editions'. As he puts it,

> the study of newspapers and periodicals has always turned on the question of mediation: how publications present texts; how different forms of publications represent other, absent forms; and how the fragmented print archive represents an absent, thriving print culture. The future of nineteenth-century newspapers and periodicals depends upon how they are interpreted by a new media, but it is a media that we are well-placed to use, critique, and appropriate. (Mussell 2012c)

On the more pessimistic side, others, such as Pearson, warn that familiarity with the process of analysing editorial characteristics of a print version may be distinct from familiarity with the process required for digital versions: 'The vast majority of scholars know how to read the signs of a publishing history in a text when we have it as a published book on our desk; but far fewer know how to determine the provenance and status of an e-text' (Pearson 2008: 88).

This debate has also sometimes revolved around the adequacy of the concept of 'digital natives', usually traced back to Prensky (2001). This notion refers to the idea that young scholars who have grown up with digital technologies are somehow naturally skilled at exploiting them, in contrast to 'digital immigrants', who have come to use digital technologies later in life and maintain a certain clumsiness or 'accent' as a result. Mussell (2012c) provides an example of criticism of this idea. He emphasises the difference between the superficial competence which may come naturally to so-called 'digital natives', and the 'deeper, critical proficiency' which is useful for scholarly purposes but requires more advanced knowledge of the relevant technologies (Mussell 2012c: 203).

Until the late 2000s, these debates on digital resources tended to focus on issues of access to online material or on the differences between print and digitised material. They have not, however, tended to engage with the methodological implications of having access to material in a new (digital) medium. This state of affairs prompted Mussell to point out that

> as all those who produce digital resources know, digitisation is a transformation, yet what often remains neglected in the production of surrogates is that these digital objects behave in different ways: they have become data and are available to be processed (Mussell 2010: 280).

Most scholars contributing to these discussions have been openly enthusiastic about the convenience of accessing material online, with benefits such as not needing to work with fragile originals (e.g. Turner 2006: 309) and having an easier time locating and accessing historical material (e.g. Brake 2001: 127), as well as searching through it (e.g. Sanders 2009: 303). But in spite of this, further concerns have been raised about what may be *lost* when working with digital material.

For instance, some scholars have suggested that working with digital rather than print material may eliminate a certain form of serendipity useful to research, by eliminating the immediate context within which – say – a particular newspaper article may be found. For example, Sanders (2009: 304) argues that working with print material makes it impossible to completely ignore what he calls 'intrinsic context' – the immediate context constituted by what surrounds a passage in its original publication. This context can be hard to retrieve online where search results are often presented in fragmentary form. Although Leary (2005) agrees that the kind of 'browsing' done when holding a print version of a historical newspaper can be discouraged by the type of searching done using digital search tools, he argues that another kind of serendipity emerges when exploring digital versions of historical material. Commenting on his experience of 'Googling the Victorians', Leary (2005: 76) suggests that 'such experiences reinforce the conviction that the very randomness with which much online material has been placed there, and the undiscriminating quality of the search procedure itself, gives it an advantage denied to more focused research'. Leary's (2005) research is relatively open-ended and exploratory, but even in studies involving a more thoroughly defined methodology, similar comments can be found. An example of this is Gibbs and Cohen (2011: 76), who make a similar point when commenting on their experiences of using Google N-grams[2] to search Google's collection of Victorian novels.

We thus see that, progressively, the discussion in the pages of historical journals has shifted from an emphasis on the value and perils of using digitised historical material, to the methodological implications of doing so. It has been noted that the change brought about to the discipline of History by virtue of the introduction of digital methods and resources, 'has thus far been incompletely theorized' with 'its practical consequences… still emerging' (Stauffer 2011: 63). Hitchock has famously made this point in particularly strong terms:

> History as a discipline, largely uninvolved in the production of digital resources and apparently uninterested in changing how it illustrates its scholarship to accommodate the digital, has put its head in the sand and tried to ignore the whole issue. (Hitchcock 2013: 12)

It has also been noted that the field has as yet engaged insufficiently with the methodological potential associated with the digitisation of historical material (Nicholson 2013: 61). Nevertheless, an increasing number of articles are appearing in historical journals with the aim of encouraging scholars to explore ways of drawing scholarly benefit from digitised historical material. Some such articles are the topic of the following section.

[2] Google's Ngram Viewer allows users to browse through n-grams (i.e. sequences of n words, see also footnote 13 in section 2.3.3.2) derived from their collection of textual sources from the 1500s to today.

2.2.4 'DISTANT READING' AND METHODOLOGICALLY EXPLORATORY HISTORICAL RESEARCH

Among historical scholars keen to engage with the new methodological possibilities of digital resources, Moretti's concept of *distant reading* has proven popular. Moretti introduces 'distant reading' in an article (Moretti 2000) in which he was concerned with what he called the 'problem' of world literature. Basically, he suggests that close reading is inadequate to tackle this problem, because it will inevitably (for practical reasons) result in focusing on a canon which cannot be considered sufficiently broad to yield a greater understanding of world literature. To study world literature, then, would require a change in methods, towards more collaborative work, essentially building on other people's analysis of national literatures and looking for overarching patterns. The result of this may be a study of literature involving virtually no reading of the actual texts ('without a single direct textual reading'; Moretti 2000: 57), but only at this price is it possible to tackle the problem of world literature. This, then, is what Moretti called 'distant reading': the endeavour to build on other people's detailed work (itself mostly dependent on close reading) in order to provide a synthesis that allows one to observe patterns at a higher order of abstraction. This endeavour depends on distancing oneself from the actual texts; as Moretti puts it, distance from the text '*is a condition of knowledge* [italics in original]' (2000: 56).

The concept of distant reading seems to have struck a chord with Victorian scholars, some of whom share a similar concern for getting beyond the study of a fairly restricted nineteenth-century canon. Turner (2006), for example, laments what he calls a 'smash-and-grab' approach to the study of historical newspapers, basically consisting of the close study of a randomly selected amount of material without consideration for how this material fits within the overall context of the 'periodical-ness' of the press. Fyfe (2009) also reports that scholars working during the Victorian era encountered difficulties in dealing with the volumes of cheap literature produced in the period and laments their conclusion that only 'arbitrary principles' could guide their selection of material to study.

However, the concept of distant reading appears to have been used more loosely as a prompt to carry out methodological explorations of various kinds, usually involving an attempt to analyse historical material without carrying out close reading but relying instead on other properties of the material. Liddle (2012: 234), for example, shows how a preliminary analysis of the word-count of leading articles challenges the consensus among scholars that the leading article originated from a sole inventor, following 'a single developmental path' and taking 'its Victorian form fairly early'. Gibbs and Cohen (2011) show how Google's n-gram viewer allows them to follow the decline of religious themes appearing in the titles of Victorian novels after 1850. Nicholson (2012b: 79-80; see also Nicholson 2013: 69) adopts a number of creative methodologies. For instance, he searches for references to America, Germany and France within 10 words of the word 'competition' to describe the changing attitude to the United States expressed in the British press. Other examples of methodologically innovative work in the field of Victorian studies that make reference to Moretti's concept are Deswarte (2010); Stauffer (2011) and Heuser and Le-Khac (2011) (working under the direction of Moretti himself); and Vuohelainen (2013).

The work in this thesis can be situated in the continuity of this methodologically innovative tradition. As noted in chapter 1, chapters 6 and 7 will outline approaches to analysing historical sources in a way other than through direct reading. These approaches' innovative character lies in the use of corpus linguistic software, tools and techniques, and thus this thesis responds to at least two important issues which have attracted comment in the debates summarised above. First, since corpus linguistic methods facilitate the exploration of large amounts of data, they help address one of the principal concerns behind 'distant reading': that of going beyond a small subset of the texts relevant to one's research question. Second, they provide a set of techniques which can help the researcher analyse large amounts of text with a degree of robustness and systematicity which is not currently afforded by most existing interfaces for access to collections of digitised historical resources. Indeed, it has been repeatedly noted that the existing interfaces to historical material (and especially newspapers) have been designed with little scholarly input (see for example McGann 2008, for whom this realization was an important motivation for developing his own collaborative platform, NINES) and 'with close rather than distant forms of reading in mind' (Nicholson 2012a: 242). As a result, users constantly grapple with problems. One issue which has attracted comment is inconsistencies in results produced by keyword searching. Bingham (2010) gives an excellent summary of the issues associated with keyword searching, using the Nexis platform[3] as an illustration. Likewise Chase (2009) points out the inconsistencies between his search results in two similar collections presented via two different platforms, the *Nineteenth Century Serials Edition* and the *19th Century British Newspapers*. Having to contend with such low-level issues is hardly a sound basis for more complex forms of analysis.

This situation prompted Nicholson (2012a: 242) to suggest that there are, in the current state of affairs, three possible solutions. The first is to abandon the study of newspapers in favour of literature (which is available in more distant-reading-friendly forms), allowing for more flexible and innovative methods. The second is to build one's own Victorian periodicals archive platform allowing for new methods, and the third is to adjust one's methods to the available platforms' affordances. Although Nicholson suggests that the second option would be preferable, he also immediately dismisses it as relying on resources not usually available to Victorian historians. However, creating such archives in a form well-suited to quantitative and qualitative research has been precisely the concern of much early work in computational linguistics and corpus linguistics. Nowadays, interfaces such as CQPweb exist which allow users to explore datasets in a variety of ways (Hardie 2012). The CQPweb server at Lancaster University in fact already contains various historical datasets such as EEBO-TCP[4]. This thesis, by using corpus linguistic methods, thus provides an example of Nicholson's (2012a) preferred, yet assumed impractical, option.

[3] Nexis is an online platform which provides access to news and business information published since the 1980s to today.

[4] A collection of over 125,000 books published in English between 1475 and 1700, see http://www.textcreationpartnership.org/tcp-eebo/

In sum, corpus linguistics, an established subfield of linguistics, has developed technical and conceptual tools to assist researchers in analysing vast quantities of text. However, the full potential of these tools for the benefit of historical scholarship is far from being realised. Although work is being done using corpus linguistics to analyse historical material, this work is mostly being published outside of historical journals, and to date does not seem to be making much of an impact on historical scholarship. Nevertheless, historians have in parallel been grappling with the implications of technological advances for over two decades. In the area of nineteenth-century history, prominent journals have, since the 1990s, hosted increasingly frequent contributions on the topics of scholars' visions and concerns related to advances in digital technologies. Early papers of this type focused on the value or perils of working with digitised historical material, whereas later contributions have progressively shifted towards assessing the potential of new methods more suited to benefiting from the available digitised material. At this stage, however, most of these methodological discussions have remained exploratory or have drawn on text mining rather than corpus linguistics (see section 2.2.1). Partly this has been due to a lack of suitable interfaces for exploring textual datasets in helpful ways. But corpus linguistics is equipped with exactly this set of interfaces. Hence, investigating the full potential of corpus linguistics to benefit historical scholarship is a logical next step.

2.3 Work using corpus linguistics to analyse newspapers

As the previous section established, the potential of corpus linguistic methods has not yet been fully exploited in historical research. Corpus linguistics is too broad a field to review here. In order to illustrate some of the major corpus linguistic techniques, I will focus on reviewing the main methods which have been used to analyse newspaper data[5]. Even within this seemingly narrow focus, studies using corpus linguistics to analyse newspaper data are too numerous to be exhaustively reviewed here. I will merely draw on these numerous studies to help depict the breadth of corpus linguistics' methodological landscape. Therefore, this section

[5] Of course, corpus linguistic methods are used with a variety of texts, not just newspaper data. The choice to focus here on newspaper data is purely to maximise relevance to this thesis.

will be of most benefit to readers with little familiarity with corpus linguistics. The major questions I will address are 'how are corpus linguistic methods used to facilitate analysis of newspaper data?' and 'what kinds of questions are explored through the corpus linguistic analysis of newspaper data?'

No individual subfield of linguistics or corpus linguistics is solely devoted to the study of newspapers. Even the emerging field of media linguistics (see Dobrosklonskaya 2013; Perrin 2013) focuses on a broad range of media types, not just newspapers. Thus, section 2.3.1 briefly introduces the main subfields of linguistics which have produced analyses of newspaper data using corpus linguistic tools. Section 2.3.2 introduces some basic conceptual and practical corpus linguistic tools, which will be relevant to various discussions throughout the thesis. Section 2.3.3 illustrates how these tools are combined within methodological approaches.

2.3.1 Theoretical approaches

Since the 1980s, corpus linguistic methods have been increasingly incorporated into the ‘methodological toolbox’ of the various subfields of linguistics (McEnery and Hardie 2012: 226). As a result, researchers drawing on corpus linguistic methods may work within a range of theoretical traditions, both established and emerging. This section describes the main theoretical approaches which have produced, or are beginning to produce, studies of newspaper data drawing on the corpus linguistic methodological apparatus. In each section, a brief description of the original field or theoretical framework is provided, followed by an introduction to the subfield within which research drawing on corpus linguistic tools is embedded. The discussion also explores the reasons why newspaper data is a focus of interest for these subfields.

2.3.1.1 Discourse analysis

Discourse analysis is a broad area of linguistics which, as the name suggests, focuses on understanding discourse. Defining the field further can be problematic, because the term *discourse* itself has a wide range of definitions, which have strong disciplinary bases but vary even within the disciplines themselves (Mills 2004: 1). Reviews and discussions of the range of these definitions exist, for example Mills (2004). Within linguistics, Bloor and Bloor (2007) list six common ways in which *discourse* is defined, ranging from 'a discourse' referring to a specific verbal address (oral or written), usually fairly long, about a specific topic, to 'all the phenomena of symbolic interaction and communication between people, usually through spoken or written language or visual representation' (Bloor and Bloor 2007: 6-7). Nevertheless, common to all these definitions is the scale at which analysis is performed. Discourse analysts may be interested in language for different ultimate aims, such as understanding more about how a particular argument is logically constructed or how a particular linguistic exchange relates to issues of sociocultural power. But their analyses will always involve at some stage attempts to describe language at a level *above the sentence level*. This observation distinguishes discourse analysis from areas of linguistics which operate below the sentence level, such as syntax or morphology, but is not sufficient to distinguish the field from other areas such as pragmatics or stylistics. I come back to this observation below.

Discourse analysis encompasses a broad range of interests and forms of analysis. In terms of what discourse analysts actually do, Paltridge suggests a distinction between 'textually oriented' approaches and more 'socially oriented ones'. The former 'concentrate mostly on language features of texts' whereas the latter 'consider what the text is doing in the social and cultural setting in which it occurs' (Paltridge 2012: 2). Approaches to discourse analysis also differ in their purpose. Gee (2011) helpfully distinguishes between 'descriptive' and 'critical' approaches. *Descriptive approaches* aim to describe language for its own sake, in order to explain how language works and why it works the way it does. In contrast, *critical approaches*, in addition to this descriptive aim, 'also want to speak to and, perhaps, intervene in, social or political issues, problems, and controversies in the world' (Gee 2011: 10). Here, we are interested only in those approaches which have produced work which analyses newspaper data using corpus linguistic methods. Two major approaches have produced such work: corpus-based critical discourse analysis (corpus-based CDA) and corpus-assisted discourse studies (CADS).

2.3.1.1.1 Corpus-based Critical Discourse Analysis

Critical discourse analysis (CDA) is an approach to analysing language which, like most linguistic approaches to discourse analysis, is interested in providing 'accounts of the production, internal structure, and overall organization of texts' (Kress 1990: 84). Nevertheless, it is situated on the 'socially oriented' side of the scale because, unlike other kinds of discourse analysis, it is focused on the way in which these linguistic structures are related to 'structures of power and domination' (Kress 1990: 85).

CDA is a 'critical' approach in that it is not interested in language for its own sake but is instead 'interested in the way in which language and discourse are used to achieve social goals and in the part this use plays in social maintenance and change' (Bloor and Bloor 2007: 2). It is also associated with certain particular political commitments (Kress 1990: 85) which are made explicit in many critical discourse analysts' writings; such scholars thus do not aim for 'the type of objectivity that is sometimes claimed by scientists or linguists' (Bloor and Bloor 2007: 5).

Although linguistic categories are usually central to a CDA analysis, the extent to which they are drawn upon and the range of categories mobilized in any single study can vary significantly (Wodak and Meyer 2009: 21). The selection of the relevant categories depends on the research question as well as, to some extent, the theoretical framework adopted (see Wodak and Meyer 2009 for a good overview). Nevertheless, these methods often involve intensive reading and analysis and tend to be very time-consuming. In consequence, most CDA studies rely on analysing a narrow set of material, 'typical texts' in the words of Wodak and Meyer (2009: 23). This makes the findings of individual studies hard to generalize (Paltridge 2012: 144), and also open to the criticism that the material has been selected according to biased procedures in order to support the researchers' political agenda (an observation formulated by, among others, Stubbs 1997).

These observations have prompted research that draws on corpus linguistic methods in association with a CDA framework of interpretation (Paltridge 2012: 144-68 provides an overview of this intersection). Newspaper data is often the focus of corpus-assisted CDA research (see O'Halloran 2010 for an overview). Indeed, the frequency and regularity of newspaper publishing and the wide distribution of its products across a particular population make newspapers an ideal means to study what Baker (2006: 13) has referred to as 'the incremental effect of discourse'. This formulation captures the assumption that repeated exposure to particular patterns of language can have powerful effects, such as reinforcing cultural stereotypes, on social actors – and, through them, on the distribution of power in society. Work analysing newspapers within a corpus-based CDA framework includes Orpin (2005), Baker et al. (2008), Rasinger (2010), Caldas-Coulthard and Moon (2010), Baker et al. (2013), and Cheng and Lam (2013).

2.3.1.1.2 Corpus-assisted discourse studies

The approach called 'corpus-assisted discourse studies' has much in common with corpus-based CDA. Like corpus-assisted CDA, CADS uses corpus data, as well as other sources external to the corpus, to assist its investigations of discourse as a means of communication embedded in social life. Nevertheless, CADS researchers are keen to distance themselves from CDA. Partington et al. thus state:

> It must also be emphasised that CADS is not tied to any particular school of discourse analysis, certainly not, for instance to critical discourse analysis (CDA). Unlike CDA, it has no overarching political agenda and has very different attitudes to and traditions of how language data should be managed. (Partington et al. 2013: 10)

CADS clearly has a social orientation (as opposed to a purely textual orientation), like CDA. Whether it also has a 'critical' orientation, in the sense described by Gee (2011), cited above, is debatable and depends, to some extent, on one's precise definition of *critical*. But probably the most important difference between corpus-based CDA and CADS is simply the aforesaid difference in political orientation. Indeed, as emphasized by Partington (2013, quoted above), CADS aims not to have a defined political orientation, in contrast to CDA which is more or less by definition left-wing, and sometimes specifically Marxist, as in the case of leading critical discourse analyst Norman Fairclough (who considers Marx to have been 'a discourse analyst "avant la lettre"', Fairclough and Graham 2002: 187).

CADS scholars also study newspaper data. According to Partington et al. (2013: 11), this is simply because a core group of Italian scholars engaged in CADS research are operating within Political Science faculties and so tend to focus on politically involved texts. Examples of studies which investigate newspaper articles include all the papers presented in the special edition of *Corpora* on CADS (vol 5., issue 2, 2010) (Clark 2010; Duguid 2010b, 2010a; Marchi 2010; Partington 2010; Taylor 2010); as well Fusari (2010), Freake et al. (2011), Partington (2012), and case-studies in Partington et al. (2013).

2.3.1.2 Pragmatics

Pragmatics is 'the scientific study of all aspects of linguistic behaviour' defined by its focus on the relationship between language as an abstract system and language used in specific contexts (Bublitz et al. 2010: v). Its emphasis is hence on language as *performed*. Definitions of pragmatics can be narrower or broader, encompassing more or less focus on the social context of language use (Taavitsainen and Jucker 2010: 5). Pragmatics traditionally focuses on spoken data and is of little relevance here. *Historical* pragmatics, however, *is* interesting for our purposes. Indeed, since spoken data is rarely available for historical periods, historical pragmatics typically has to rely on written data (Taavitsainen and Jucker 2010: 7).

The distinction between what is covered by the terms 'historical pragmatics', 'historical discourse analysis' and 'historical dialogue analysis' is somewhat fuzzy (see Taavitsainen and Jucker 2010: 6 for a brief discussion). In any case, historical pragmatics encompasses at least three areas of interest: 'the language use in earlier periods, the development of language use and the principles of such developments' (Taavitsainen and Jucker 2010: 6). Within historical pragmatics, examination of written data is sometimes undertaken, with apologies, by looking at written data which is as close to spoken language as possible (drama or trial proceedings for example, see Culpeper and Kytö 2010, a monograph on the subject). At other times, however, the focus on spoken language is simply left to one side, and written texts are considered in and of themselves, in terms of their nature as part of an interactive process of communication (Taavitsainen and Jucker 2010: 9). This is the case for historical pragmatic research which focuses on (historical) news discourse (see Claridge 2010 for a survey of the field). In such work, newspaper articles are seen as an instantiation of 'a form of public communication' (Claridge 2010: 588).

Not all historical pragmatics work relies on electronic corpora. Nevertheless, Kytö (2010: 33) notes that the widening availability of computerized tools to store and analyse language have been important in the development of the field. As Kytö (2010: 52-53) points out, the goal of (historical) pragmatics is to investigate the relationship between the forms and functions of language. Electronic methods are well-suited to provide the researcher with information about the forms of language, but not its functions. The kind of historical pragmatic questions which corpus linguistics is good at answering thus relate to 'the frequency and distribution of particular linguistic features across periods, genres and language user groups' (Kytö 2010: 52). Hence, evidence derived from corpus linguistic methods 'can be used to support claims about linguistic stability or change and about the factors that have promoted or retarded developments' (Kytö 2010: 52). All that said, work in historical pragmatics using corpus linguistics to look at historical newspapers is hard to find, partly because the field is sometimes hard to distinguish from the related field of discourse analysis. Examples include, arguably, Prentice and Hardie (2009)[6] and Bos (2012).

[6] Although the authors published the paper in the *Journal of Historical Pragmatics*, they considered the work to be an example of corpus-based CDA (Andrew Hardie, personal communication, 10/01/2016).

2.3.1.3 Stylistics

Stylistics is ‘a sub-discipline of linguistics that is concerned with the systematic analysis of style in language and how this can vary according to such factors as, for example, genre, context, historical period and author’ (Jeffries and McIntyre 2010: 1). It seeks to provide a formal framework to explain the effects upon readers of particular texts (Jeffries and McIntyre 2010: 4). Hence, like pragmatics, stylistics is interested in the relationship between form and function. Unlike pragmatics, however, the focus is mostly only on one direction – explaining the effect (function) by describing the form. Another distinction from pragmatics is that where pragmatics has tended to focus on spoken language, stylistics has been historically predominantly concerned with written texts (Jeffries and McIntyre 2010: 4), particularly literary publications. Nevertheless, the distinction between stylistics, pragmatics, sociolinguistics, and other areas of linguistics can sometimes be hard to capture since there is a degree of overlap (Jeffries and McIntyre 2010: 4).

Work in stylistics which uses corpus linguistic methods has become known as *corpus stylistics*. An important advocate of corpus stylistics is Michaela Malhberg, who demonstrates the usefulness of corpus linguistics for stylistics in her work on Dickens (Mahlberg 2007, 2012, 2013). She suggests that corpus linguistics offers several advantages to stylistics, including the ability to trace patterns ‘systematically throughout the text’ and to complement the researcher’s intuition by providing additional perspectives on a text (Mahlberg 2007: 31).

As noted above, work in stylistics has been traditionally interested in ‘literary’ texts; Toolan (1998: ix) defines stylistics simply as ‘the study of language in literature’. News discourse is hence not a traditional focus of stylistics. But news discourse has been a natural focus for *critical stylistics* which, like CDA, is interested in the relationship between power and language. Critical stylistics is still a relatively new subfield of stylistics and has as yet produced little work (Jeffries 2009 is the first monograph on the subject). Work in critical stylistics which focuses on newspaper language includes Jeffries and Walker (2012) and Tabbert (2013).

2.3.1.4 Sociolinguistics

Sociolinguistics is the area of linguistics interested in the relationship between language use and social patterns. As Baker observes, ‘sociolinguists are therefore often interested in identifying how the identity of a person or social group relates to the way that they use language’ (Baker 2010a: 3). Sociolinguistics developed as a field in the 1960s, in opposition to Chomskyan linguistics, which focused on language as a system used by ‘idealized speaker-hearers’, and which discarded the relevance of context to understanding language (Wodak et al. 2010: 3; see also section 2.2.1 on Chomsky). In contrast, the field of sociolinguistics is premised on the idea that language cannot be understood ‘without taking many layers of social context into account, be it the situational context of utterances, the geographical origin of the speakers, their age, gender, social class, ethnicity, and so forth’ (Wodak et al. 2010: 1-2).

Early sociolinguistic work drawing on corpus linguistics dates back at least three decades. Bauer (2002) provides an overview of sociolinguistic research using corpora, and Baker (2010a) details in depth how corpus linguistic methods can be mobilized in sociolinguistic research. Sociolinguistics tends to be interested in spoken data, but, as with pragmatics, problems of access to data mean that *historical* sociolinguistics has to rely on written data (see Nevalainen 2010 for an overview of the field of historical sociolinguistics). Unsurprisingly then, historical sociolinguistics has also made use of corpus linguistics (see Cantos 2012 for an overview).

Historical sociolinguistics is interested in newspapers for various reasons, including as a source of insight into the development of the language and form of newspapers, as a source of insight into broader patterns of language variation and change, and as a source of insight into historical attitudes towards language (Percy 2012). Percy (2012) provides an excellent review of historical sociolinguistic research into early newspapers and advertisements. Work in historical sociolinguistics which uses corpus linguistics to analyse newspapers includes Bauer (1994) and Fitzmaurice (2010).

2.3.1.5 Summary

Several major areas of linguistics including pragmatics, stylistics, sociolinguistics and discourse analysis have produced work analysing newspapers using corpus linguistic methods. Discourse analysis in particular has a natural interest in newspapers and has produced a vast amount of such work, under the label of 'corpus-based CDA' and, more recently, under the label 'corpus-assisted discourse studies'. Generally, discourse analysis is interested in newspapers as a repository or instantiation of discourse(s). CDA has a more specific interest in newspapers as a form (or container) of discourse(s) which reflects and may intervene in the distribution of power in society. Areas such as pragmatics, stylistics and sociolinguistics have a less natural interest in newspapers, but their subareas of historical pragmatics, critical stylistics and historical sociolinguistics have each developed an interest in newspapers for various reasons. Pragmatics and stylistics are interested in the relationship between form and function in the language of newspapers. Pragmatics focuses on newspapers as a form of communicative performance, whereas stylistics is interested in newspapers as a literary form. Sociolinguistics is interested in newspapers as a repository of social patterns of language variation. Although it is possible to distinguish the interests of these fields, as I have done above, in practice they overlap substantially.

The focus in historical sociolinguistics, historical pragmatics and CADS is very strongly diachronic. But this is not true of corpus-based CDA or critical stylistics, which may be interested in texts for their synchronic characteristics rather than for how they point to change over time. In all cases, corpus linguistic methods are welcomed for their ability to manipulate a great amount of data, supporting both quantitative and qualitative forms of analysis. This facility is particularly crucial when studying newspaper data, since its abundance means that studies aiming to be representative (or to make generalizable statements) need a broader pool of material than studies based on other sources of data. Since the focus of each field is somewhat different, it makes sense that each would tend to develop methodological preferences. Nevertheless, all these studies draw on the same pool of fundamental corpus linguistic techniques; describing all of them is beyond the scope of this thesis, but those which are relevant to the discussion in this thesis are introduced in the next section.

2.3.2 Basic methodological tools and concepts relevant to this thesis

This section introduces basic corpus linguistic tools and concepts which will be relevant to subsequent parts of this thesis. It is not, and makes no attempt to be, an exhaustive survey of corpus linguistic methodology. In particular, issues related to corpus building and annotation are not treated here. In-depth treatments of these points can be found elsewhere, e.g. Garside et al. (1997) for issues related to corpus annotation, and Wynne (2005) for issues related to corpus construction. The discussion assumes no prior knowledge of corpus linguistics; corpus linguists may want to skip this section.

2.3.2.1 Concordances

A concordance is a listing of the occurrences of a given word[7] in a given corpus (e.g. Baker 2006: 71; Hunston 2002: 39). Extensive discussions of the use of concordances exist in the literature; Baker (2006: 71-86) is a good introduction. Concordances are typically presented in what is known as the *key-word-in-context* format (KWIC), which places the word of interest (sometimes called *node*, Hunston 2002: 39) at the centre of the concordance line, with a limited number of words appearing to the left and right of it (see Figure 2.1). Most modern corpus analysis software allows users to toggle easily between a concordance and specific locations in the corpus; in this way, the researcher can move from a concordance line to a larger extract of the original text than is visible in the concordance (e.g. from 20 words around the node to 50 or more words around the node).

As mentioned above (see 2.3.2.1), concordances date back to at least medieval times and do not, in theory, require computers. In practice, computers speed up the process of concordancing by many orders of magnitude, thus making the use of concordances feasible in

[7] Concordances can be drawn not only for single words, but also for other kinds of units, such as word segments (e.g. suffixes), multi-word expressions or categories present in the annotation, see McEnery and Hardie (2012:35).

short time-spans. In fact, provided the text has previously been digitised, drawing up a full concordance for a given word will take only seconds, even if a corpus contains several thousand books. A comparable manual concordance might take several lifetimes. This ease of manipulation of a large body of evidence is one of the main appeals of using corpora: 'the act of evidence gathering is made simple, freeing the researcher's effort for the act of interpretation' (Hunston 2002: 214). But the automation of the concordancing process does not imply an automation of the process of analysis. Hunston clarifies this point:

> Concordance lines present information; they do not interpret it. Interpretation requires the insight and intuition of the observer. (Hunston 2002: 65)

Although concordances can be extremely helpful in gathering many different occurrences of a given word or phrase, 'their use is limited by the ability of the human observer to process information' (Hunston 2002: 67). Corpus linguists hence often make use of further tools designed to help them with 'assessments of frequency and significance' which can be difficult to make 'impressionistically' on the basis of concordances alone (Hunston 2002: 67). Some of these further tools are described in the following sections.

2.3.2.2 Frequency lists

The expression *word list* or *frequency list* is used in corpus linguistics to refer specifically to 'a list of all the words in a corpus along with their frequencies' (Baker 2006: 51). In the technical terminology of corpus analysis, a frequency list enumerates all the *word-types* (distinct specific sequences of characters, that is distinct word-forms) together with counts of how many tokens there are of each word-type in the corpus (where a *word-token*, or just *token*, is a single particular instance of a *word-type* at one particular point in the running text[8]). A frequency list can be presented in various orders, whether by order of (first) occurrence in the corpus, frequency or alphabetical order (Hunston 2002: 67). Corpus linguistic software usually allows

[8] In other words, a *token count* will include all words in a text, whereas a *type count* includes only *distinct* words. For example in the phrase 'the apples in the tree', there are 5 tokens, but only 4 types (since the word-type 'the' occurs twice).

the user to move from a frequency list to a concordance by clicking on a word-type in the frequency list.

Frequency lists can be used at various stages in research, for example as an entry point into the data – when frequent words can be identified as foci for analysis (see section 2.3.3.1) – or later on, as a way of quantifying effects which have been observed by concordance analysis. Sometimes frequency lists are used as an aid to comparing different texts or groups of texts. In this case, *raw frequencies* (the number of times a word-form occurs in a corpus) are not necessarily all that useful and usually need to be supplemented by *relative frequencies* (raw frequencies divided by the overall size of the corpus, often expressed per hundred, thousand or million words). Relative frequencies are preferable for comparisons because they take into account the size of the groups of texts under comparison (Baker 2006: 51; McEnery and Hardie 2012: 50).

Baker notes that frequency lists should normally be used in association with qualitative methods of analysis, in particular concordance analysis:

> Frequency lists can be helpful in determining the focus of a text, but care must be taken not to make presuppositions about the ways that words are actually used within it. This is where taking an approach which combines quantitative and qualitative analysis will be more productive than simply relying on quantitative methods alone. A concordance analysis is one of the most effective techniques which allows researchers to carry out this sort of close examination. (Baker 2006: 71)

2.3.2.3 Collocation and related concepts

2.3.2.3.1 Collocation

The concept of *collocation* appears recurrently in the linguistic literature, but can have a variety of definitions. McEnery and Hardie (2012: 123) note that beyond ‘basic generalities’ about collocation, ‘a great multitude of definitions’ become apparent as soon as one attempts ‘to pin down collocation either operationally or conceptually’. These various definitions will not be reviewed here; see McEnery and Hardie (2012: 122-33) for an overview of some definitions and operationalisations. In this thesis, I will adopt a proximity-based definition of collocation, that is, collocation as ‘the phenomena of certain words frequently occurring next to or near each other’ (Baker 2006: 96).

Analysing collocation in this sense relies on the analyst being able to determine how frequently words occur, as well as how near to each other they occur. Determining these factors can be done manually or computationally. The manual approach – which involves simply reading through a concordance – is favoured by many scholars in the neo-Firthian tradition (see 2.2.1); McEnery and Hardie (2012: 126-27) name it the ‘collocation-via-concordance’ approach. In support of the computational approach – dubbed ‘collocation-via-significance’ by McEnery and Hardie (2012: 126-27), Hunston (2002: 12) explains that ‘more information can be processed more accurately by the statistical operations of the computer than can be dealt with by the human observer’. More will be said about the statistics of collocation in part 2 of this thesis; both approaches to collocation will be implemented to some extent in chapter 7.

Use of the concept of collocation in studies of newspaper language is motivated by the idea that there is something noteworthy about a tendency for two or more words to co-occur. This idea involves a chain of related assumptions. (1) The sequencing of words in language is not random. (2) Sequencing is determined by language-internal constraints as well as by choices of the language producer. (3) Sequencing which cannot be explained by language-internal constraints betrays something about the thought processes of the language producer. (4) Recurrent sequencing choices in instances of language produced by a great number of different language producers betray thought processes which are shared by at least some of these language producers. (5) The recurrent expression, through language, of these thought processes may have a social effect.

Assumptions (1) to (3) are explained by Baker (2006: 47-48) in his discussion of frequency. Frequency is of interest to discourse analysts, he argues, ‘because language is not a random affair’ (Baker 2006: 47) (assumption 1); further, although language follows certain rules, language users still have choices to make (assumption 2). The relationship between these two assumptions and assumption 3 is clarified in his statement:

> It is the tension between these two states – language as a set of rules vs. language as free choice – that makes the concept of frequency so important. If people speak or write in an unexpected way, or make one linguistic choice over another, more obvious one, then that reveals something about their intentions, whether conscious or not. (Baker 2006: 48)

Assumption (4) follows from assumption (3), since if one instance of a language choice reveals something about the intentions (or thought processes) of the language producer, then many instances of language choices made by many different language producers must reveal something about the intentions of these different language producers. Assumption (5) is harder to justify, since it relies on an (inevitably) complex theory of the relationship between language, cognition and social change. Nevertheless, it is central in discourse analysis research as Fowler, among others, clarifies:

> Critical linguistics looks at language, not as a system on its own but as something that 'intervenes' in the social world, largely by perpetuating the assumptions and values of that world. (Fowler 1987: 482-83; cited in Hunston 2002: 109)

The precise mechanism by which language intervenes in the social world is still the object of research. One important source of theories attempting to capture this mechanism is the field of *media effects*. From that perspective, Perse (2000) and Bryant and Oliver (2009) review some of the main theories which attempt to describe the ways in which the content of media (including language) may have an impact on people's beliefs and behaviours. One such theory, for example, is the *agenda setting* theory:

> Agenda setting is the process of the mass media presenting certain issues frequently and prominently with the result that large segments of the public come to perceive those issues as more important than others. Simply put, the more coverage an issue receives, the more important it is to people. (Coleman et al. 2009: 147)

Outside of media effect theories, Baker's (2006) expression 'the incremental effect of discourse', mentioned in section 2.2.1, also attempts to capture the mechanism by which language (and repetition in particular) affects the social world.

A final point on assumption (3). For a given research purpose, not all examples of sequencing which cannot be explained in terms of language-internal constraints will attract further attention. Hunston (2002: 68) distinguishes between *motivated* and *unmotivated* collocations. *Motivated* collocations have an immediately apparent logical explanation, such as the observation that *toys* co-occurs with *children* more often than with *women* or *men* (Hunston 2002: 68). In contrast, *unmotivated* collocations are more likely to be of interest to the researcher. This distinction, it is important to clarify, will necessarily be subjective and highly dependent on the precise interest of the researcher. In research on social issues, for instance, a finding that *sexy* collocates with human-related nouns (like *woman* or *outfit*) more than with other types of nouns (like *stone* or *continent*) might be considered uninteresting (because we know that sex is relevant to people but not to stones or continents). But a finding that *sexy* collocates with words that refer to women consistently more than with words that refer to men *would* be noteworthy. Caldas-Coulthard and Moon (2010), in fact, report this finding (see also section 2.3.3.5), and it leads them to question whether the pattern makes sense or whether it points to a potentially problematic broader social pattern of differential attitudes towards men and women.

2.3.2.3.2 Syntactic collocation

In the previous section, I mentioned that this thesis adopts a proximity-based definition of collocation. A common operationalisation of the statistical approach to this type of collocation involves software testing 'the significance of the co-occurrence frequency of that word and everything that appears near it once or more in the corpus' (McEnery and Hardie 2012: 52). However, as Hunston (2002: 71) notes, such tests assume that words occur in random sequences, which is not the case. Instead, certain sequences are possible (syntactically) whilst others are not. For this reason, some corpus linguists take syntax into account, considering collocation as 'mediated by larger syntactic units' (Evert 2005: 19)[9], and defining it more narrowly as the frequent co-occurrence of words *occurring in given slots of a specific syntactic structure*. For example, Pearce (2008) compares uses of MAN[10] and WOMAN in terms of how they occur in specific grammatical and lexical relationships. Among other differences, he finds that, in his corpus, MAN, but not WOMAN, is the subject of (and ergo collocates with) verbs like *swear* and *curse*, whereas WOMAN is subject of (and ergo collocates with) verbs like *nag* and *berate* (Pearce 2008: 13).

Another illustration of syntactic collocation comes from Kilgarriff et al. (2004), who discuss *word sketches*, an operationalisation of syntactic collocation which will be mentioned in section 2.3.3.3. Word sketches are 'one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour' (Kilgarriff et al. 2004: 116). For instance, a word sketch of the verb PRAY may show that it is associated with modifiers such as *silently* or *together*, and that it tends to have as subject pronouns such as *we* or *I* (Kilgarriff et al. 2004: 119). Word sketches will not be used in this thesis, but the concept of syntactic collocation is relevant to the *sampling approach* presented in chapter 7.

2.3.2.4 Keyness

In section 2.3.2.2, it was mentioned that frequency lists are sometimes used to compare groups of texts. Baker (2006) notes, however, that comparing the most frequent items in word lists may not be that interesting. One reason is that the most frequent items are likely to be grammatical words (Baker 2006: 123) since 'with few exceptions, almost all forms of language have a high proportion of grammatical words' (Baker 2006: 53). But even a focus on only the

[9] Evert (2005) actually defines 'collocation' in an entirely different way. What is called 'collocation' in this thesis corresponds to what Evert calls 'co-occurrence'. The proximity-based definition of collocation given in 2.5.2.3.1 corresponds roughly to what he calls 'positional co-occurrence' (2005:18), whereas the syntactic definition of collocation presented in 2.5.2.3.2 corresponds roughly to his 'relational co-occurrence' (Evert 2005:19). The terms 'proximity-based collocation' and 'syntactic collocation' which I use here were suggested by Andrew Hardie, personal communication, 15/05/2014.

[10] As pointed out on p.xi, capitals will be used throughout this thesis to refer to lemmas as opposed to individual word-forms. Defined in intuitive terms, a *lemma* is a set of word-forms which are very closely related in meaning: if a researcher is interested in the word *man*, they are probably also interested in the word *men*. The lemma MAN encompasses both of these forms – *man* and *men* (but not *woman*, *mankind*, *human*, etc.).

most frequent *lexical* items may only 'confirm expectations surrounding the genre of the text' rather than pointing to differences between two groups of texts (Baker 2006: 124). In Baker's (2006) example of comparing the two sides of parliamentary debates on banning fox-hunting, the lexical items found at the top of frequency lists for the anti-hunting and pro-hunting side of the debates were often common to both. Words such as 'hunting', 'hon.' or 'mr.' appeared at the top of both lists and related to 'the context of parliament' or 'the subject under discussion' rather than the differences in position between both groups (Baker 2006: 124).

For this reason, a measure of *saliency* rather than *frequency* would be helpful for comparison of texts; *keyness* is precisely such a measure (Baker 2006: 125). *Keywords* are 'words which are significantly more frequent in one corpus than another' (Hunston 2002: 68). Crucially, this statistical sense of the word *keyword* is to be distinguished from another sense of *keyword* as referring simply to 'the word that is currently under examination', as in the expression 'key word in context' (KWIC), sometimes used to refer to a concordance (Baker 2006: 71, see also 2.5.2.1). Keyness analysis can be applied not only to words, but also to categories (whether semantic or grammatical). Categories which are overrepresented in one (sub)corpus compared to another are then called *key categories*.

Keywords are thus words which are unusually frequent in one corpus *compared* to another corpus. Keyword analysis is thus useful for contrasting two groups of texts, where this is a research goal in and of itself, but it can also be helpful for commenting on a single corpus of interest. In such a case, the other corpus is chosen carefully to be representative of language in general – such a corpus is known as a *general corpus*, or a *reference corpus* (see for example McEnery et al. 2006: 59-60).

Although keywords and collocates are in some ways conceptually very different, both are statistical abstractions – McEnery and Hardie (2012: 41) comment that 'keywords are a statistical abstraction from frequency lists and collocations are a statistical abstraction from a concordance'. Hardie (forthcoming) argues that the same statistic can be effectively used to measure both. Interestingly, in one of its operationalisations, the calculation of collocation borrows from keyword analysis. This operationalisation makes use of what Taylor (2010) and Partington (2012) call a *concordance-corpus*.

A *concordance-corpus* is essentially an excerpt of a bigger corpus which is formed by extracting a specified amount of context around a given word. (Hence the name *concordance-corpus*, since the new corpus contains *only* the material available in the concordance of a given word.) Such a concordance-corpus thus contains all the words which co-occur (within a specified distance, or *span*) with the word under investigation. Applying keyness analysis to a pair of such concordance-corpora will produce what is essentially a list of collocates of the word under investigation, although only collocates which co-occur with the word under investigation statistically-significantly more often in one of these concordance-corpora than in the other will appear in this list[11]. This is why studies such as Partington (2012) and Taylor (2010), which apply a keyness analysis to such concordance-corpora, are treated no differently in the discussion below from studies which simply use collocation analysis. Keyness analysis is a major technique in corpus linguistics which would almost certainly prove very useful for historical research, but which it has not been possible to investigate in detail in this thesis.

2.3.3 *METHODOLOGICAL PROCEDURES*

This section reviews methodological procedures adopted in studies of newspaper data which rely at least partly on corpus linguistic techniques. The discussion focuses on studies which are not linguistically-motivated, but instead investigate language as a way of gaining insight into contemporary or historical social issues/conditions. Studies have been excluded if they are geared exclusively towards describing aspects of grammar or genres, such as Bauer (1994), who tracks grammatical change in twentieth-century English; Westin and Giesler (2002), who undertake a multi-dimensional diachronic study of editorials; Albakry (2007), who counts prescribed and proscribed grammatical constructions; Clark (2010), who investigates

[11] Andrew Hardie, personal communication (11/04/2014) suggests that a more precise term for this could be *contrastive collocation*.

changing linguistic patterns of evidentiality; Fitzmaurice (2010) and Brownlees (2012), (who investigate expressions of authorial identity in historical newspapers; and Duguid (2010a) and Bos (2012), who investigate respectively the informalisation and popularization of newspaper discourse.

The discussion is organized around recurring methodological patterns characterized by certain sequences in the use of tools, and certain ways of operationalising a research question. The names given to these patterns are for purposes of reference in this thesis and do not constitute a standard typology. The categorisation of studies that I adopt is certainly debatable; there is a high degree of variety in the methodologies in the literature and any kind of categorisation would capture certain likenesses but not others across these methodologies. Nevertheless, since space is lacking to review in detail all the relevant studies, the categorisation has been adopted as an organizing principle for the discussion. Each section describes one approach and discusses, for purposes of illustration, one or two studies categorised under that approach.

2.3.3.1 The focus-on-frequency approach

The focus-on-frequency approach is useful when a researcher begins without pre-selected expressions. The questions asked are 'what words are frequent in this corpus?' and 'what patterns surround the use of these frequent words?'. The wider implications of such findings depend on the choice of corpus. A good example of this approach is Hakam (2009), who investigates coverage of the cartoon controversy (of 2005) in English-speaking versus Arabic-speaking newspapers. She uses her investigation of frequent words in these newspapers to point out strategies used by Arabic-speaking newspapers to signal their identity, even when reproducing Western stories. Although this is the only study I am aware of that uses this approach on newspapers specifically, it is a standard approach in corpus linguistics generally and is for example suggested by Mahlberg (2007: 22).

Hakam (2009)'s data consists of 187,000 words from 442 articles published in 19 English-language Arab newspapers based in 12 different Arab countries. These selected texts, published between late 2005 and Spring 2006, were harvested from the newspapers' websites and were selected because they were about the ‘cartoons controversy’. Selected texts are divided into (a) articles reproducing international (but Western-based) press agency briefs (with or without minor alterations) and (b) articles generated in the Arab world. A control corpus is also collected, consisting of 10,700 words from 31 texts generated by the Associated Press, which was the source for many of the articles in the Arab corpus.

First, Hakam generates frequency lists. Then frequent words which appear relevant to the study are further investigated in terms of their collocates. Hakam finds that some words are collocates in both Western and Arab accounts, but that others, for example *blasphemous* as a collocate of *cartoons*, are only collocates in the Arab accounts. She also uses the concept of *collocational incongruity* – a pattern of collocation which points to an association which is unexpected in a given community[12] – to show that Arab newspapers were editing the press agency briefs to intentionally signal an affiliation to the Arab world. This was done, for example by adding the honorific ‘peace be upon Him’ after mentions of the Prophet Muhammad, even if the brief was quoting somebody who would not have actually used such an honorific (such as the Austrian Foreign Minister).

In sum, the focus-on-frequency approach can accommodate relatively open-ended research questions. It does not involve determining *a priori* what linguistic expressions are of interest. Its starting-point is instead the generation of frequency lists in order to select frequent words to investigate further. Patterns of use of these frequent words are then further explored using collocation and concordance analysis.

[12] This concept depends on the idea that collocation provides insight into thought patterns that are salient within a given discourse community, see also the discussion on the rationale for studying patterns of collocation in relation to social issues in 2.3.2.3.1.

2.3.3.2 The contrasting-corpora approach

The contrasting-corpora approach, like the focus-on-frequency approach, does not require (and in fact could not accommodate) pre-selected expressions. Instead, it is used by researchers interested in comparing corpora overall; it requires that the two or more corpora under comparison be carefully designed to represent some meaningful distinction, whether of time-period, location, or genre. The main question here is 'what distinguishes this corpus from that corpus?', and it is most often answered by investigating words identified as being unusually frequent in one corpus compared to another (i.e. *keywords*; see section 2.3.2.4). Two examples are discussed below. The first is Gabrielatos and Baker (2008), who contrast a corpus of British tabloids to a corpus of British broadsheets, both constructed to contain articles referring to refugees, asylum seekers, immigrants and/or migrants, in order to comment on differences in how the quality and popular press represent these social groups. The second is Cheng and Lam (2013), who contrast two pairs of corpora containing Western newspapers on one side, and Eastern newspapers on the other, constructed to contain references to Hong Kong, in order to investigate changing representations of Hong Kong in the East and West before and after its handover in 1997. Other research in this category includes O'Halloran (2010), whose case-study contrasts a purpose-built corpus of articles from the British tabloid *The Sun* about the European Union expansion of 2004 to a reference corpus, in order to investigate discourses surrounding immigration from the new European Union countries; Fitzsimmons-Doolan (2009), who searches for similarities between a corpus of discourse about language policies and a corpus of discourse about immigration to test whether the two discourses overlap; Baker (2010b), who contrasts the representation of Islam in British tabloid and broadsheet newspapers; Fusari (2010), who contrasts the metaphors used in Italian and Anglo-Saxon newspapers' coverage of the Alitalia crisis (2008-2009); and Aull and Brown (2013), who contrast a corpus of writing about a male sport event and one about a female sport event to investigate the treatment of gender in the coverage of sports in the American press.

Gabrielatos and Baker (2008) are interested in how refugees, asylum seekers, immigrants and migrants (referred to collectively as RASIM) are 'linguistically defined and constructed' in the British press (Gabrielatos and Baker 2008: 8). They also ask: 'what are the frequent topics of, or issues discussed in, news articles relating to RASIM?'; 'what attitudes toward RASIM emerge from the body of UK newspapers seen as a whole?'; and 'are conventional distinctions between broadsheets and tabloids reflected in their stance toward (issues relating to) RASIM?' (Gabrielatos and Baker 2008: 8). They explore a 140-million corpus of 175,000 UK press articles published between 1996 and 2005, selected from nineteen newspapers including tabloids and broadsheets based on a complex query in Nexis intended to yield articles mentioning RASIM.

First, the coverage of RASIM in broadsheet and tabloid newspapers are contrasted by examining a list of keywords generated by comparing the broadsheet papers against the tabloids. Concordances are then used to explore the uses of the RASIM expressions (*refugees*, *asylum seekers*, *immigrants* and *migrants*) as well as the keywords. The discursive construction of RASIM is further investigated by drawing up collocates of the RASIM expressions.

Gabrielatos and Baker also used the technique of c-collocates (see section 2.3.3.3) to identify collocates which maintain their association with RASIM over time. A further analytical step is the exploration of semantic prosodies to uncover socio-political choices in the representation of RASIM. *Semantic prosody* is the arbitrary investing of meaning to a word as a consequence of the repeated use of particular shades of meaning in conjunction with that particular word. For example, in this case, the repeated use of words such as *swarm*, *flood* or *gang* with RASIM expressions invest RASIM with negative connotations which they need not have otherwise. Gabrielatos and Baker suggest, for example, that their findings point to more in-depth treatment of RASIM issues in broadsheet than in tabloid newspapers, and that there seems to be in broadsheets newspapers an 'overall more positive, or less negative, stance' towards RASIM (Gabrielatos and Baker 2008: 30).

Cheng and Lam (2013) investigate the changing representations of Hong Kong in Western and Chinese media since its handover in 1997, asking: ‘how have the Western perceptions of Hong Kong changed over the intervening decade [since Hong Kong’s handover], when compared with the Chinese, and what are the possible reasons for the changes?’ (Cheng and Lam 2013: 178). Their corpus contains 1,686,424 words from 2,427 newspaper articles and reports published between 1996 and 1998 and between 2006 and 2008 in British and American newspapers and international organizations, and in newspapers from China, Hong Kong and Taiwan. These texts were identified by searching for *Hong Kong* and *handover* together in several online news databases and were then used to form four corpora, two for the 1996-1998 period and two for 2006-2008.

Cheng and Lam’s first step is to identify the most frequent two-word concgrams[13] that occurred in all corpora. Afterwards, they identify key semantic categories (e.g. 'politics', 'vehicles and transport on land') in each corpus by comparing the most recent Western corpus to the other three corpora, in order to answer the question of how Western recent accounts of Hong Kong differed from their Eastern counterparts and from earlier Western accounts. Then, they identify two-word concgrams which contain words frequent in the key semantic categories they identified (e.g. the concgrams *political party* and *political system* in the 'politics' semantic category). Finally, they look at concordances of these two-word concgrams and analyse them in terms of their semantic prosody (see explanation above). They found that Eastern perceptions had changed little over time. In contrast, Western perceptions showed change over a ten year period – with, for example, more attention being given to political issues in the later than in the earlier Western corpus.

[13] *Concgrams*, like *n-grams*, are operationalisations of the concept of collocation. Where n-grams are sequences of words which occur together repeatedly *in the exact same sequence* within a corpus (these are sometimes also described as *word clusters*, *lexical clusters* or *lexical bundles*), concgrams are sequences of words *in whatever order* and potentially non-contiguously. Cheng et al. (2013:414) define them as ‘all the permutations of constituency variation and positional variation generated by the association of two or more words’. Concgrams for two words A and B would hence include AB, BA (exhibiting positional variation) and ACB (displaying *constituency variation*).

These examples demonstrate that the contrasting-corpora approach, like the focus-on-frequency approach, is useful for relatively open-ended research questions. Unlike the focus-on-frequency approach, it does require the careful assembling of corpora which can be compared on a meaningful basis using keyness analysis. Keywords are then investigated further using concordance and collocation analysis. In Gabrielatos and Baker (2008), the corpora being compared differ in genre (broadsheet versus tabloids), whereas in Cheng and Lam (2013), the corpora differ in terms of time-period (before and after the 1997 handover) and cultural context of publishing (East and West).

2.3.3.3 The expression-intensive approach

In contrast to the two previous approaches, the expression-intensive approach is used in studies which are interested in investigating one (or several) pre-selected expressions. I will describe two variants of this approach. Variant 1 involves focusing on describing a single expression. Here, the question being asked is simply 'how is this particular expression used in this corpus?' or 'what is the range of meanings which this expression can adopt?', and the choice of expression being investigated is what makes the study relevant beyond linguistics. The example discussed at length below is Baker et al. (2013), who focus on describing uses of the word *Muslim* in the British press as a means of gaining insight into how the Muslim community is represented by the British media. Other examples of this variant are Kim (2014), who explores occurrences of *North Korea* in order to investigate the representation of North Korea in American newspapers; Marchi (2010), who explores occurrences of words such as *moral* to explore changing conceptions of morality in the British media; and Taylor (2010), who explores occurrences of expressions such as *science* or *the experts* to explore changing conceptions of scientific authority in the British media.

Variant 2 involves contrasting uses of two paradigmatically-related expressions. (The term paradigmatically-related describes linguistic expressions which could fit in the same slot in a sentence, for example *bought*, *purchased* or *acquired* in the incomplete sentence 'I just ____ some bread'; see Saussure [1916] 1986: 123.)[14] Here, the question asked is 'what are the similarities or differences in how these two expressions are being used?'. Again, the relevance beyond linguistics comes from the choice of expressions. The example discussed below is Taylor (2013) who is interested in how BOY and GIRL are used in the British press. Her research has wider implications in terms of tracking the representation of genders and the sexualisation of children. Another example of this variant is Vessey (2014), who contrasts how two pairs of words, *national* (EN) – *nationale* (FR) and *Canadian* (EN) – *Canadien* (FR), are used in the Canadian press, depending on whether they are used in the native language or in the other language (as 'borrowed words'). She uses this analysis to explore nationalism and representations of the French-speaking and English-speaking communities of Canada.

Baker et al. (2013) are interested in the representation of Islam and Muslims in the British press, and focus on exploring one specific word (*Muslim*). They use a corpus containing 143 million words from articles published in various daily and Sunday national British newspapers over the period 1998 to 2009. The articles included were selected from the Nexis platform (UK section) using a query constructed to yield articles mentioning terms related to discussions of Islam (such as *Islam*, *Koran* or *imam*).

First, they draw up lists of syntactic collocates of *Muslim* (see 2.3.2.3.2). Next, they explore the noun collocates of *Muslim* by manually examining concordances for each collocate, then classifying the collocates into 'thematic categories'. These categories are then quantified in terms of their relative frequencies and lexical richness. Hence at this point, the researchers can make claims such as that the word *Muslim* as an adjective occurs more often (in 37.6% of cases) in their corpus as a marker of ethnicity or nationality than in reference to religion (8.7%). They also point out that their category 'ethnicity or nationality' is realized by fewer different words than is their category 'conflict', which they suggest means the latter topic is more salient in their corpus.

[14] The distinction between *syntagmatic* and *paradigmatic* relations is usually traced back to Saussure's *Course on General Linguistics*, although he actually calls the latter *associative relations*.

In a second phase of their study, Baker et al. expand their core *Muslim* to the two frequent two-word clusters *Muslim world* and *Muslim community*. Those clusters are chosen because *world* and *community* were 'the two most frequent immediate right-hand noun collocates' of *Muslim* (Baker et al. 2013: 268). They explore these by drawing up concordances for these clusters and manually scanning them in order to determine the typical contexts in which they are used. Based on their observations, they then generate new concordances for related expressions which allow them to verify or specify these observations. For example, they look at a concordance for *Muslim communities* in order to investigate further the degree to which *Muslim* is used to refer people in a way which minimizes their diversity. They also concordance *Muslim world and* in order to investigate what kind of other expressions are typically put on par with the idea of a *Muslim world*. In cases where more detailed analysis is required (for example in order to categorize line by line who *Muslim community* is being used to refer to), they use a procedure known as *down-sampling* which reduces a vast sample to a more manageable one for manual analysis. This involves asking the software to produce a random sample of a specified number (in this case 100) of concordances lines for manual analysis. Their investigations allow them to produce conclusions such as that the expressions *Muslim world* and *Muslim community* 'help to create the idea of Muslims as belonging to a distinct and separate 'imagined community' at both the global and national level, and, thus, contribute towards a process of "othering"' (Baker et al. 2013: 275).

Taylor (2013) focuses on exploring the similarities in usage over time of the two lemmas, BOY and GIRL, in British broadsheets. She uses three annual corpora (known as SiBol 93, SiBol 05 and Port 2010) each containing the entire production of the *Guardian*, *Times* and *Telegraph* for their respective years (1993, 2005, 2010). She begins at the level of 'basic data analysis', producing frequency counts for each corpus for the word-forms *girl*, *girls*, *boy* and *boys*. She finds that 'the frequencies of BOY and GIRL seem to be relatively stable over the three years' (Taylor 2013: 97). Next, she investigates which age-groups are being referred to with BOY and GIRL by generating a random sample of 100 concordance lines for each lemma from each corpus. She finds that 'GIRL is more likely to refer to adults than BOY', a pattern consistent over time (Taylor 2013: 98). She also reports that 'it appears that over time GIRL is less frequently used to refer to adult women, while the reverse trend is seen for BOY' (Taylor 2013: 98).

These basic steps complete, Taylor undertakes to investigate collocates in two different ways. The first technique involves generating collocate lists for each lemma for each year. She then manually compares these lists to identify collocates (for each lemma) that occur in all years; such diachronically unchanging collocates have been dubbed *consistent collocates*, or *c-collocates* (Baker et al. 2008). 31% of the c-collocates identified for each lemma are shared by BOY and GIRL. Next, she groups the c-collocates into thematic sets, a step that she notes is 'researcher-driven and subjective', in contrast to the earlier steps (Taylor 2013: 99). She comments that the 'c-collocates refer to people and relationships', with the largest set consisting 'of descriptions referring to age, physical appearance, character, and so on' (Taylor 2013: 99). One of her findings at this stage is a 'consistent association' between both lemmas and terms related to aspects of sexual relationships (such as *rape, molest, love* and *marry*), an association which remains stable over time (Taylor 2013: 99). However, she also finds that 'items referring to sexual activity are more frequent amongst the GIRL collocates'. This observation, she points out, challenges the idea of a 'sexualisation of female children', since 'the central term for referring to these individuals, GIRL, displays a stable history of association with sex in the broadsheet newspapers from 1993 to 2010' process' (Taylor 2013: 99). The second technique she uses to investigate collocates involves generating a word sketch (see section 2.3.2.3.2) for each lemma for each year. These word sketches are then manually compared (for each lemma) 'in order to identify the items which appeared in the Word Sketches for all three years' process'. This procedure also allows Taylor to identify c-collocates (Taylor 2013: 102). She finds similar results with this method as with the previous one, although she notes that 'this second method is much quicker, but the researcher has less control over the process' (Taylor 2013: 102).

Next, Taylor uses *distributional thesauri*, which identify ‘words which behave in similar ways to the search term’ (Taylor 2013: 102). Again, she generates these for each lemma for each year and manually compares them. She finds that ‘the lemma which behaves most like the search term is its gender equivalent, that is BOY for GIRL and *vice versa*’ and further reports that ‘the thesaurus candidates are similar over the three years and for the two items’ (Taylor 2013: 105). She also notes that WOMAN, but not MAN, behaves in similar ways to both BOY and GIRL.

Next, Taylor briefly describes her investigation of recurrent sequences of words (or *word clusters*), focusing on those word clusters which occur in all three years (i.e. *c-clusters*). Looking at frequent four-word-clusters containing BOY or GIRL, she then groups clusters according to their function. For example, one grouping, which includes clusters such as *a group of girls* and *a pack of boys*, is identified as having a *counting* function; this function is verified by reading concordance lines containing the clusters in that grouping. Each grouping is then analysed in quantitative and qualitative ways. For *counting* four-word-clusters, she thus finds a peak in the number of occurrences for both BOY and GIRL clusters in the year 2005, but does not provide an interpretation for this, noting instead that more corpora would be needed to determine whether this pattern can be found in these other corpora, or whether it is simply an artefact of the sampling. On the qualitative side, Taylor notes that different *counting* expressions are used for BOY and GIRL, and that GIRL expressions are more diverse. Animal metaphors, for example, are used for both GIRL and BOY, but *flock (of girls)* and *breed (of girls)* occur in c-clusters for GIRL but not BOY, whereas *pack (of boy(s))* occurs in c-clusters for BOY but not GIRL. Sometimes, she notes, counting nouns typically associated with GIRL are used for BOY; in such cases, there is a tendency for these boys to be ‘seen as sharing “feminine” characteristics in some way’ (Taylor 2013: 107).

In sum, the expression-intensive approach involves focusing in more detail on one or more pre-selected linguistic expressions. The approach centrally involves a detailed analysis of the uses of specific linguistic expressions by generating lists of collocates and further investigating the collocation patterns through the use of concordances. Studies in this category are characterized by a great diversity of methodological trajectories compared to studies drawing on the approaches described in the other sections. This is partly because, when researchers begin investigating one pre-selected linguistic expression, they end up noticing other interesting expressions, which they often go on to investigate as well. Hence, although research in this category typically *begins* by focusing on just one expression, as the investigation proceeds, the focus often expands or shifts to cover other single- or multi-word expressions.

2.3.3.4 The selection-via-concordance approach

The selection-via-concordance approach is used in studies interested in paradigmatically-related expressions (see section 2.3.3.3). I will describe two variants of this approach. Variant 1 involves contrasting words *within a set* of paradigmatically-related expressions. Here, preliminary questions are: 'what are the possible ways of referring to concept x?' and 'how are they different?'; and further questions are: 'in which situations are some ways of referring to concept x preferred?' and 'what does that tell us about the way in which language-users conceive of these situations?'. An example of variant 1 is Orpin (2005), who observes that corruption-related incidents are spoken about in different terms depending on their location, and wants to know what that tells us about how they are perceived.

Variant 2 involves contrasting *sets* of paradigmatically-related words, typically describing groups of social actors. Here, preliminary questions are: 'what are the possible ways of referring to groups A and B?' and 'what patterns surround these ways of referring to groups A and B?'; and further questions are: 'what is the difference in how groups A and B are represented?' and 'what does that tell us about how groups A and B are conceived of by a given language community?'. An example of variant 2 is Prentice and Hardie (2009), who are interested in how both sides of the Glencairn uprising are represented in the London press of the 1650s; similar studies include García-Marrugo (2013), who is interested in how both sides of the Colombian civil war are represented in the Colombian press.

Orpin (2005) is interested in the representation of corruption-related incidents in the British press, and how it varies according to the location (within or outside of the UK) of the incident described. She explores this within the 4 newspaper sub-corpora of the *Bank of English* corpus which, together, contain over 800 texts published between 1990 and 1996 in the *Guardian,* the *Independent*, the *Times* and *Today*. Orpin also uses the entirety of the *Bank of English* (at the time 323 million words) as a reference corpus.

First, Orpin identifies the set of lexical choices available to someone writing about corruption. She uses a CDA framework (see section 2.3.1.1.1), within which the choice between possible lexical options is seen as ideologically meaningful. Orpin proceeds to identify these choices by using a thesaurus (in order to find synonyms and near-synonyms of *corruption*) and also by drawing up a list of significant collocates of *corruption* in the newspaper corpus and manually selecting the relevant words. She then further restricts her results to words frequent enough in her newspaper corpus to be amenable to productive study (she sets the bar at a minimum of 15 occurrences). She ends up with 8 selected nouns: *bribery, corruption, cronyism, graft, impropriety/ies, malpractice(s), nepotism* and *sleaze*.

Next, Orpin endeavours to build an overall profile for each selected word. This is done in three stages. First, the overall frequency of each word in the reference corpus is determined. Then, change over time is explored by comparing the overall frequency of the word in the *Bank of English* to its frequency in a comparable earlier corpus (the *Birmingham Collection of English Texts*, a corpus containing 18 million words produced before 1985). Finally, distribution figures are drawn up for the word in the *Bank of English* in order to further explore which kinds of contexts it tends to occur in. (Orpin relies on the built-in genre categorisation of the *Bank of English*). She finds for example that *sleaze* is absent in the *Birmingham Collection of English Texts* but is the second most frequent item of her selected set in the *Bank of English;* that *malpractice(s)* occurs around 300% more often in the *Bank of English* than would be expected from the *Birmingham Collection of English Texts*; and that *graft* is far more frequent in American than in British books in the *Bank of English*.

The next step involves exploring more detailed usage-profiles for each word of interest. Orpin explores 'typical contexts' of use by manually scanning a concordance from the newspaper corpus for each word. Then the word's frequent associations are investigated by drawing up lists of significant collocates in the *Bank of English*. Finally, the newspaper concordances are manually scanned anew in order to verify the conclusions formulated by looking at the concordances and significant collocates. She finds, for example, that *graft* seems to be 'strongly connected with Italy' whereas *sleaze* 'is particular associated with British politics' (Orpin 2005: 47).

After all this, Orpin finds that the words chosen to describe corruption-related incidents in foreign countries have worse connotations than those chosen to describe incidents in Britain, but that this difference diminishes over time. Orpin's interpretation of the difference is that such incidents are perceived as worse abroad. But she suggests that the diminished difference over time reflects a growth in awareness within Britain of the gravity of the corruption-related incidents taking place in Britain itself.

Prentice & Hardie (2009) are interested in how the London press of the 1650s represents in- and out-groups in the Royalist rebellion against Cromwell which took place in 1653-1654 in Scotland and became known as the Glencairn Uprising. They focus particularly on the representation of the leader of the Uprising, the Earl of Glencairn. They use a part of the Lancaster Newsbook Corpus, which contains 1 million words from newsbooks (i.e. early newspapers) published in the 1650s. The part selected contains 'a complete collection of every newsbook published in London between the middle of December 1653 and the end of May 1654' (Prentice and Hardie 2009: 30).

First, they identify different ways of referring to one or the other side of the Uprising. They do this manually, both relying on their prior historical knowledge, and identifying relevant terms appearing in concordances of pre-selected terms. (For example, they notice by drawing up a concordance for *Glencairn* that he is often mentioned alongside Kenmore, Middleton and Athol, so they include the names of these people in the Scottish-side list.) Next, they draw up concordances for each of the items on both lists. They then manually categorize each concordance line, using a qualitative procedure drawing on a CDA framework. Each category's occurrences are then counted.

They find that the English side is generally associated with positive semantic categories such as *honourable* or *loyal* whereas the opposite is true for the Scottish side (which are associated with categories such as *criminality* or *violence*). However, they point out that more fine-grained analysis reveals that both sides are actually portrayed in somewhat contradictory ways. They illustrate this by focusing on the concordance lines for Glencairn and showing how he is represented both as failing in his endeavours but also (rarely but nevertheless more often than Morgan, his English counterpart) as worthy of admiration.

In conclusion, the selection-via-concordance approach is one used by researchers who have determined one or more linguistic expressions to focus on before using any corpus linguistic tool, but who begin their research by expanding their pool of expressions using corpus linguistic tools (possibly alongside other tools such as a thesaurus). The first step involves using concordances to identify expressions paradigmatically related to the pre-selected one(s). The next step then involves analysing each expression in order to contrast usages either between two sets of paradigmatically related expressions (in variant 2), or between items within one set (in variant 1).

The selection-via-concordance approach is similar to the expression-intensive approach in that, once expressions to be explored in the selection-via-concordance approach have been identified, these expressions may be investigated using similar tools to those used in the expression-intensive approach. Moreover, both approaches investigate pre-selected expressions. Although the boundary between the two is therefore somewhat fluid, the main distinguishing characteristic is that in the expression-intensive approach, the number of expressions to be investigated is usually small and pre-determined, which allows for a relatively detailed exploration of the patterns related to a single expression. By contrast, in the selection-via-concordance approach, although the number of (initially) pre-selected expressions is also small, the number of expressions to be explored is quickly expanded in the first steps of the approach, and the resulting number of expressions of interest to the researchers may preclude such detailed explorations.

2.3.3.5 The tracking-expressions approach

The tracking-expressions approach does something similar to the contrasting-corpora approach except that it involves focusing on one (or more) linguistic expressions, contrasting their uses in several corpora. Hence, like the contrasting-corpora approach, it allows differences or similarities to be identified between corpora which may represent historical, generic or geographical comparison-points; unlike that approach, however, it focuses primarily on one single concept (though this can be done repeatedly to incorporate several concepts). The technique could be considered a hybrid between the expression-intensive and the contrasting-corpora approaches since it involves both focus on a single expression, as well as a focus on comparing two corpora. The question being asked here is thus 'how do patterns associated with the use of this linguistic expression compare in both corpora?'; this consequently leads to the sociologically more interesting question 'how does the representation of this concept or group compare between these corpora?'. Depending on the differences between the corpora, the answer to this latter question may allow the researcher to comment on historical changes in representation, or on geographical or social variation in representation. The examples reviewed below are Johnson et al. (2003), who track how the concept of 'political correctness' has changed use over time in the British press, and Caldas-Coulthard and Moon (2010), who compare how specific adjectives are used to describe genders in broadsheet versus tabloid newspapers. Other studies in this category include Partington (2012), who tracks how discourses on anti-Semitism have changed in the UK press between 1993 and 2009; and

Jaworska and Krishnamurthy (2012), who compare uses of the word *feminism* in the British and German press. Pumfrey et al. (2012) (see section 2.2.2), who investigate early uses of the word *experimental* in order to trace changing conceptions of science, adopt a similar approach (although not with newspaper data).

Johnson et al. (2003) are interested in how the expression 'political correctness' has been utilised in the British press. They use a purpose-built corpus containing approximately four million words from articles from *The Times*, *The Independent* and *The Guardian* published during 1994, 1999 and an 'interim period from mid-1996 to mid-1997' (Johnson et al. 2003: 31); to be included in the corpus, articles had to contain one or more versions of the terms *political(ly) (in)correct(ness)* or *PC* (referring to *political correctness* – the abbreviations were manually disambiguated). They also use a written general corpus (which contains newspapers as well as other texts) as a reference corpus.

First, they plot the frequencies of the political correctness terms in each sub-corpus (each sub-corpus corresponding to one time-period). This step allows quantification of the increase or decrease of incidences of use of particular terms over time. They find for example that 'PC-related terms' are more frequent in *The Times* than in the other newspapers, and that 'PC-related terms' decrease in frequency over the 1994-1999 period.

Next, Johnson et al. compare these corpora to their reference corpus using keyness analysis (see 2.3.2.4). This allows words to be identified which are significantly more prominent in the target[15] than in the reference corpus. These words can be analysed to provide insight into what the corpus is *about*. Contrasting keyword lists for different time-periods allows changes in how the terms are used over time to be tracked. To yield a strong interpretation, such changes have to be investigated by going back and forth from the keyword lists to concordances, in order to ascertain how the keywords under study are actually being used.

[15] In keyness analysis, see section 2.3.2.4, the *target corpus* is the corpus which the researchers are interested in analysing and the *reference corpus* is the corpus used as a benchmark.

By this procedure, Johnson et al. (2003) are able to document, for example, how use of terms related to political correctness decreased in *The Times* and *The Guardian* between 1994 and 1999, whereas the same trend of decrease in *The Independent* ends in 1997 and is then followed by an increase between 1997 and 1999. They are further able to suggest that this overall decrease was accompanied by a shift in the way in which terms related to political correctness were being used. For instance, in 1994, *The Times*, which supported the Conservative government, used terms associated with political correctness to undermine Labour's traditional commitments. On the other hand, *The Guardian*, which supported Labour, used these terms to present Labour's traditional commitments as under threat from the Conservatives. By 1999, with Labour now in power, The Guardian was no longer using these terms to refer to Labour's upholding of its traditional commitments, but was instead using the term as an attack on Labour's failing to uphold those traditional values (Johnson et al. 2003: 37). So, Johnson et al. argue, being *politically correct* went from being a good Labour-thing to do, to being a Labour-way of hiding a failure to do the right thing.

Caldas-Coulthard and Moon (2010) explore the representation of genders in tabloid versus broadsheet British newspapers. They use data from the *Bank of English* which contained at the time five million words from 5 British tabloids and an unspecified number of words from broadsheet newspapers. They use frequency lists to compare the usage of pre-selected adjectives in both types of papers. They also look at how these adjectives are being used in context by exploring collocates. They suggest that these methods are an effective way of revealing dominant ideological patterns, with women being 'constantly judged in terms of social and aesthetic esteem, especially, but not exclusively in the tabloid press' which places them in a less powerful position compared to men who are 'evaluated in terms of their function and status in society' (Caldas-Coulthard and Moon 2010: 124).

In sum, the tracking-expressions approach focuses on comparing how one (or more) linguistic expressions are being used in several corpora, often corpora representing sequential time periods. One way of doing this is to compare the frequency and collocation patterns of these expressions in each corpora, as done by Caldas-Coulthard and Moon (2010); in this case, the corpora studied do not have to be thematic[16]. An alternative technique adopted by Johnson et al. (2003), involves constructing thematic corpora, which contain a pre-determined amount of context surrounding the target linguistic expression(s) (corresponding to what Partington 2012 and Taylor 2010 call a concordance-corpus, see section 2.3.2.4), then contrasting these corpora to one another or to a common reference corpora using keyword analysis. Keywords are then further investigated using collocation and concordance analysis. Although this approach seems similar to the expression-intensive approach, in that both approaches may often involve focusing on just one linguistic expression, the expression-intensive approach is interested in describing patterns related to the use of one or more linguistic expressions *in a single corpus* whereas the tracking-expressions approach involves comparing the use of one or more linguistic expressions across *several* corpora.

2.3.3.6 Methodological choices

As this admittedly cursory review has shown, corpus linguistic methods can be combined in a variety of ways, as part of a methodology geared towards studying newspaper data. Methodological choices must be made and will depend on the nature and purpose of the study at hand. Some of the factors influencing methodological decisions are described in this section. (I will not here address issues related to corpus building, assuming instead that corpora suitable for a given research purpose have already been assembled.)

One of the strengths of corpus linguistic approaches is that they can accommodate relatively open-ended research, in the sense that the precise focus of the research can be allowed to emerge from the data under investigation, rather than being entirely determined *a priori*. Hence, in the contrasting-corpora and the focus-on-frequency approaches, the research

[16] By *thematic*, I mean here a corpus composed of texts selected for their relevance to a particular theme, such as the corpus of texts about fox-hunting used in Baker (2006) (see section 2.3.2.4) or Hakam (2009)'s corpus about the cartoon controversy (see section 2.3.3.1), as opposed to a more general corpus such as the *Bank of English* which contains texts selected without regard for the topics they cover.

can begin with a question such as 'what is characteristic of this group of texts (possibly as compared to another group of texts)?'. It can then be allowed to gain a more specific focus according to what is identified as being (unusually) frequent in the target corpus, as well as what piques the researcher's curiosity.

Nevertheless, whatever the ultimate aim, research involving corpus linguistic tools must, at some point, involve attention to one or more specific linguistic expressions. These expressions do not need to be selected in advance if a contrasting-corpora or focus-on-frequency approach is to be used, but they can also be entirely determined in advance, in which case approaches such as the expression-intensive or the tracking-expressions approaches will be favoured. Sometimes, the researcher will have some idea of what linguistic expression may be appropriate to investigate, given their research question, but still wish to finalize their decision based on what is present in the data. In such cases, an approach such as the selection-via-concordance approach will be ideal in allowing the researcher to identify in their data a range of linguistic expressions of potential interest to them.

Once the linguistic expression(s) have been selected, the variability of methods increases, and depends largely on the purpose of the analysis. Some factors are: whether the study focuses on a single, or several, linguistic expressions; whether the study focuses on patterns of use of one or more linguistic expressions or whether it focuses on comparing and/or contrasting (sub)corpora; whether the study has a diachronic component; whether the study focuses on identifying similarities, differences, or both (whether between patterns of use of linguistic expressions, or between (sub)corpora) and, of course, whether the study also relies on methods and theories from beyond corpus linguistics. It would be too lengthy to summarize the outcomes of all possible combinations of such choices here; the discussion in this section has provided ample illustration of possible outcomes.

In any case, whatever the approach, the historical or sociological implications of a given study depend on the choice of corpus and/or linguistic expression under investigation. The examples provided in this section have illustrated the point that research using corpus linguistic tools to investigate newspaper data can allow researchers to comment on such topics as how a concept may be conceptualized differently over time, how social groups may be represented differently in different newspapers or differently from one another, or how events may be perceived differently by different cultural communities.

2.4 Conclusion

Corpus linguistics is an established area of linguistics which has produced methods that can facilitate both quantitative and qualitative forms of analysis on large amounts of text. Although this a crucial ability for the Humanities in an age of digital textual abundance, these methods have not yet had much impact on historiographical scholarship. Nevertheless, the field of History has been engaging with the implications of the increasing availability of digitized sources, and discussions are increasingly centring around the need for new methods well-suited to their nature. Corpus linguistic methods address this need, and the remainder of this thesis will hence make a major contribution to the methodology of History by exploring some of the issues which historians will encounter when attempting to use corpus linguistic methods with digitized sources.

To illustrate the breadth of existing corpus linguistic approaches, I briefly reviewed approaches to investigating newspaper data to answer research questions which were not purely linguistically-motivated, but instead touched on historical and/or social issues. Relevant studies are embedded within different established and emerging subfields of linguistics including critical discourse analysis, corpus-assisted discourse studies, historical pragmatics, critical stylistics and historical sociolinguistics. These studies are methodologically diverse; some of the common approaches were outlined in the chapter, including approaches drawing on frequency lists, concordance analysis, and collocation analysis, techniques which will be used in chapters 6 and 7. In the next chapter, I will introduce one of the major issues facing scholars working with digitized texts, the issue of Optical Character Recognition errors, and discuss its implications for corpus linguistic methods, with a special focus on collocation analysis.

PART 2: DEALING WITH OPTICAL CHARACTER RECOGNITION ERRORS

3 Assessing the theoretical impact of OCR errors on collocation statistics

3.1 Introduction

A number of remediations[1] separate the original nineteenth-century material artefact which the Victorian reader would have held in their hands and the 'data' which I explore in the rest of this thesis. The purpose of this chapter is to draw attention to these remediations and explore some related issues, with a particular focus on optical character recognition (OCR) errors, which are introduced during the process of digitization. Section 3.2 introduces the data and its remediations, as well as OCR itself. Section 3.3 explores the theoretical impact of OCR errors on the statistics of collocation.

3.2 The *19th Century British Newspapers (part 1)* collection

3.2.1 Remediations of the 19th Century British Newspapers collection

The data source used in this thesis is the British Library's *19th Century British Newspapers,* a collection of Victorian newspapers selected by the British Library in consultation with an academic panel (Shaw 2007). The selection 'includes 17 national and 29 regional newspapers'[2], totalling 2.2 million pages (Shaw 2007). The collection was digitized in partnership with the commercial company Gale (a division of Cengage Learning) and is available to subscribing institutional libraries. More information on the collection is provided via the Gale/Cengage portal (including King 2007; and Shaw 2007), but see also King (2005),

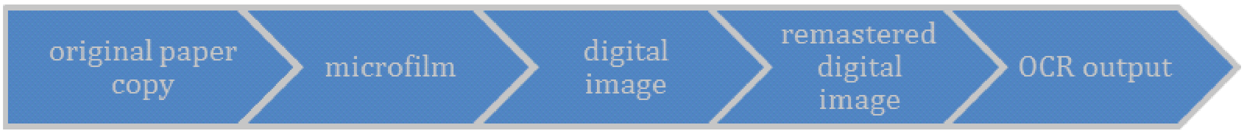
[1] I will use the term *remediation* to refer to the rendering of content originating in one media (e.g. a print newspaper sheet) into another media (e.g. a digital document).

[2] I am aware that this totals 46, that the figure usually cited (including in the Gale/Cengage documentation) is 48, and that there are 49 titles listed in the title list provided by Gale (http://solutions.cengage.com/Gale/Database-Title-Lists/bl_ncnp.html). I do not know the reason for this discrepancy. Nevertheless, 46 is the number of titles received by the Spatial Humanities project, though data is incomplete for at least some of the titles. The full list of 46 titles is included in appendix 1.

Fleming and King (2009) and Tanner et al. (2009). Fyfe (2016) provides an excellent history of the collection's creation and conservation since the nineteenth century.

In order to make the digital version of the Victorian pages available to the online reader, the British Library had to undertake several remediations of the original copies. Typically, microfilms of the originals were scanned, rather than the actual printed issues. The resulting digital images were then manipulated to increase the success of the subsequent OCR, for example by increasing contrast or correcting slant. At this point, the pages were ready for the OCR process (see below). Hence, before reaching Gale/Cengage or Lancaster University, the Victorian pages had already undergone several remediations (see Figure 3.1).

Figure 3.1. Typical remediations separating the Victorian artefact and the OCR output



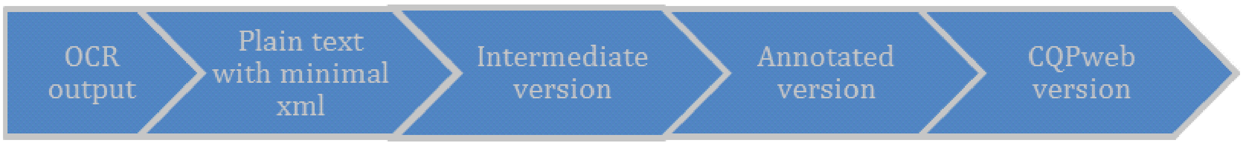
OCR is essentially a computer's best guess at what writing is on a given image. I will not go into detail here about how this is done; for an overview of OCR technologies, see Baird and Tombre (2014) or Bennamoun and Mamic (2012). There are currently only two options for digitizing material which is not born-digital: (i) rekeying, and (ii) scanning then OCR'ing (see for example Cohen and Rosenzweig 2006; or Holley 2009). The advantage of OCR is that it saves time compared to rekeying, especially for large collections. But its success rate is very variable and depends on the material being digitized as well as the technique used; see for example Blanke et al. (2012) for a comparison of several OCR programs on a range of historical text-types.

The data received by Lancaster University from the British Library was in the form of OCR output files[3]. These are XML files containing headers with metadata about the newspapers, as well as a list of all the words 'read' by the OCR software and their locations on the original

[3] Presumably produced using ABBYY FineReader; see Fyfe (2016: 566).

page. In order to be usable with corpus linguistic software, these files need to be manipulated. Figure 3.2 shows the different versions produced through successive processing phases.

Figure 3.2 Successive versions of OCR data produced through processing



Going from OCR output to plain text is relatively unproblematic. It involves removing the XML headers and position tags, determining line breaks (using changes in the *y* coordinates), removing end-of-line hyphens and rejoining split words, and tokenizing. The target of the processing phase is the annotated version, which is annotated for parts-of-speech (i.e. grammatical nature of words, such as *verbs* or *nouns*) with CLAWS 4 (Garside and Smith 1997) and for semantic tags (i.e. tags describing the category of meaning of a word, such as *health and disease* or *crime, law and order*) with the USAS tagger (Wilson and Thomas 1997). This is the version which serves as input for the corpus linguistic tool – in this case CQPweb, which requires its own separate pre-processing (Hardie 2012).

Passages with poor OCR, however, may be problematic for the annotation phase: OCR errors can lead to 'error cascades' (Alex and Burns 2014: 97), where each successive layer of annotation encounters problems caused by the problems at the previous step (e.g. an OCR error causes a tokenization problem which then causes a part-of-speech tagging problem; see Lopresti 2009 for a discussion of the impact of OCR errors on NLP processing). The cost-benefit ratio for such passages is thus high, with computational power expended to solve these problems but still producing mostly unusable output. Hence, an intermediate step may be useful to mediate between the plain text version containing low quality OCR passages and the version which goes through the annotation process.

Two solutions to the problem of low-quality passages were explored by our project. One solution is to (attempt to) correct the OCR errors. This solution will be discussed in chapter 5.

The other is to remove these passages; this involves defining and identifying 'low-quality' passages (for a similar approach, see for example Alex and Burns 2014). The appeal of this solution is that it produces a 'cleaner' version which will be less computationally taxing for the next processing stages. The problem with this solution is that removing portions of a corpus after its construction renders its design more opaque, which can cause interpretative difficulties later on. The argument against this solution can be formulated as follows.

Ensuring that a corpus is appropriate to the research questions which it will be used to explore is a critical issue in corpus linguistics (e.g. McEnery and Hardie 2012: 6). *Design* is the term used to refer to choice of the texts which make up a corpus, and corpus design is crucial for ensuring the appropriate matching of corpus and research questions. The clearer the corpus design, the easier it is for the researcher to draw appropriate conclusions and generalisations from the corpus. In the present case, if the researcher starts with a corpus of historical material which is made up of complete units – full runs of a given newspaper, or a set of complete issues from a journal – the design is fairly transparent, and the researcher can simply research that newspaper, or bear in mind the dates of the issues included in the corpus, and draw appropriate generalisations from the analysis in the context of the design. Removing portions of the corpus here and there, however, leaves a 'Swiss cheese' corpus, whose design has become opaque. It will be tempting for the researcher to conceive of this corpus as still made up of complete issues, but the corpus is in fact made up of *partial* issues, and the implications of this for the conclusions which the researcher may draw are difficult to establish. Arguably, it is not the removal of poor OCR passages which raises these implications, but rather the presence of poor OCR in the first place. However, the presence of poor OCR in the corpus is arguably more likely to alert the researcher to this issue, and help the researcher to become aware of the limitations of the corpus (for example, if the researcher is led to assess the volume of missing content). In contrast, removing the passages obscures the issue and encourages the researcher to overlook these implications – out of sight, out of mind. One way to address this when removing poor OCR passages is to leave in their place a marker of the amount of content removed, which can at least provide the researcher with an indication of how much content is missing.

Despite these limitations, a procedure for identifying low-quality passages was developed by the *Spatial Humanities* project (specifically by CJ Rupp) to help prepare a version for testing purposes. All tokens were compared with a lexicon (which included named-entities); any token not found in the lexicon was marked as an error. Next, each line received a score according to the proportion of words marked as 'errors'. If this score was higher than 30%, the line was marked as 'bad'; otherwise it was marked as 'good'. Any group of 5 consecutive 'bad' lines was then marked as 'bad'; all subsequent lines were included in the 'bad' block until 5 consecutive lines were marked as 'good'. Finally, 'bad' blocks were removed. This procedure generates a 'line-filtered' intermediate version which can then go through the next processing steps. In this thesis, unless clearly indicated, results were never generated from the line-filtered version.

3.2.2 Quality assessments of the OCR in the 19th Century Newspapers collection

OCR is reportedly very successful on contemporary material but struggles with historical material (see for example Tanner et al. 2009). Factors affecting the quality of the OCR include factors related to the conservation of the originals – e.g. conservation of the original page, conservation of the ink (including bleed-through and smudging) – the quality of the microfilms, the quality of the scanned images (including resolution, colour and contrast); but also the layout and typographical features of the original – with complex layouts being particularly error-prone, certain fonts having higher error-rates than others, and changes in fonts being harder to process for the OCR software (e. g. Balk and Conteh 2011: 156-7; Holley 2009: 2-4; King 2005: 168). Other factors also enter into the equation, such as the nature and content of the lexicon being used; these factors also interact, so that, for instance, proper names are more error-prone in part because of the capital-letter start (Tanner et al. 2009) and in part because they are often absent from lexicons (Alex and Burns 2014: 98).

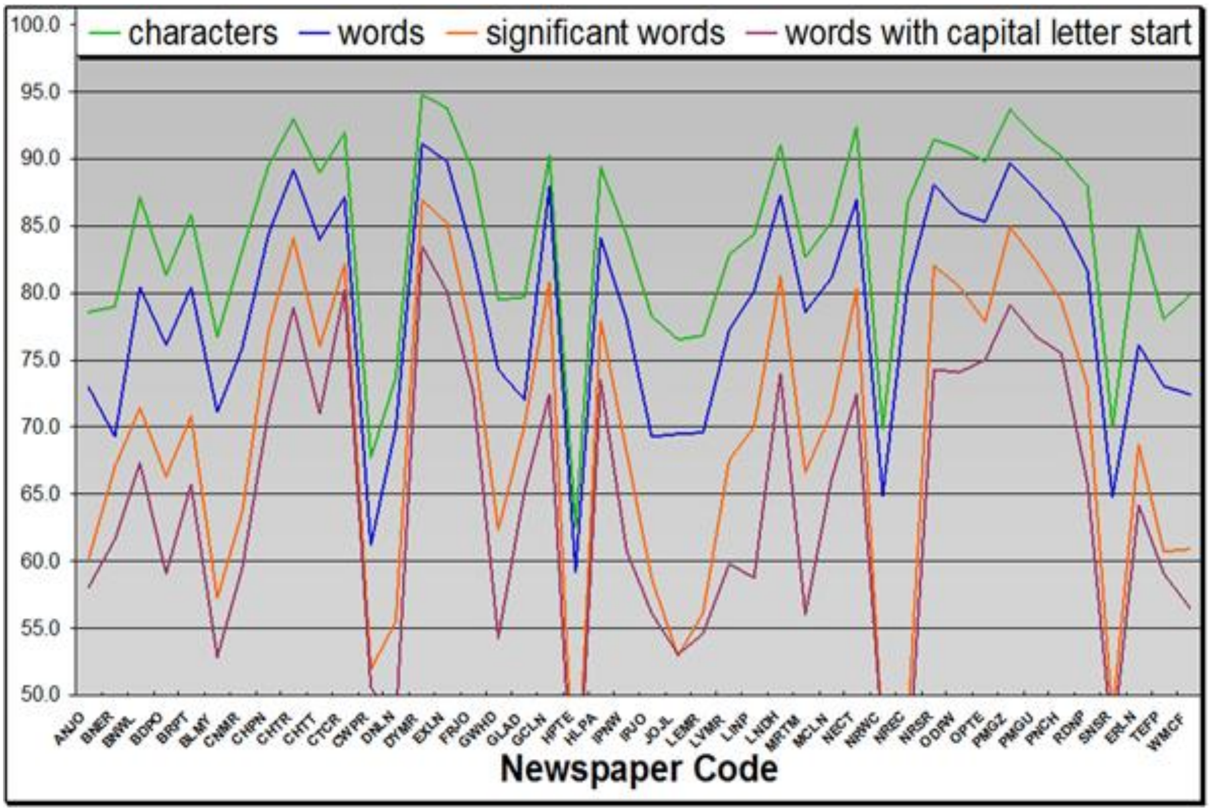
Tanner et al. (2009) have provided an assessment of the quality of the OCR in the *19th Century British Newspapers* collection. They selected around 1% of the pages in the collection, then identified two different 'zones' (portions of a column of text) on each page which were included in their test sample. These zones were selected 'on the basis of being among the clearest on the page image' (2009, section 5.1). For all included text in the sample, the relevant OCR output text was isolated, and a version was also re-keyed based on the page image. The OCR output and re-keyed versions were then automatically compared to produced OCR accuracy measures. Since the best portions of the image were selected for inclusion, the measures provide 'an assessment of the maximum likely performance' rather than a description of average performance (2009, section 5). They calculate 5 measures: 'character accuracy', 'word accuracy', 'significant word accuracy[4]' (where 'significant word' is defined as 'content words for which users might be interested in searching, not the very common function words such as "the", "he", "it", etc.', 2009, section 5.1), 'words with capital letter start' and 'number group accuracy'. Figure 3.3 shows the results for the 4 first measures. A key to the abbreviations used for the newspapers can be found in appendix 10.1. The averages reported for the collection are 83.6% character accuracy, 78% word accuracy, 68.4% significant word accuracy, and 63.4% words with capital letter start accuracy (2009, section 6).

But how good *should* OCR be? A decision as to how good is 'good enough' must be to some extent arbitrary, and dependent on the intended uses of the data. Holley (2009), who carried out an assessment of OCR quality on a collection of historical newspapers (1803-1954) held by the National Library of Australia, reported that talking to other libraries and OCR contractors had brought up a figure of 90% accuracy rates as an upper threshold for 'poor OCR accuracy' (2009: 5). However, they noted that it was unclear whether this figure was a character

[4] The concept of *significant word* as defined by Tanner et al. (2009) is of very limited use to linguists, who may often be interested in *both* content words and function words.

or word accuracy measure, and concluded that they were not yet in a position to be able to determine a baseline of 'acceptable' OCR accuracy (Holley 2009: 5, 12). Tanner et al. (2009, section 6.1) reported that with a word accuracy of greater than 80%, 'most fuzzy search engines will be able to sufficiently fill in the gaps or find related words such that a high search accuracy (>95-98%) would still be possible from newspaper content because of repeated significant words'. They further point out that given that the average accuracy rates they measured for the *19th Century British Newspapers* collection fall below this threshold, 'searching the resource will not be as satisfactory for the end user as might be desired' (Tanner et al. 2009, section 6.1). However, Tanner et al. seem to be assuming that researchers are interested in locating relevant articles, for which it is enough to retrieve one out of several relevant words in a given article. For research using corpus linguistic methods, it is likely that a higher accuracy will be necessary, since the researcher will generally be interested in retrieving most or all instances of a word or sequence of words rather than retrieving entire units of texts such as articles.

Figure 3.3. Tanner et al.'s (2009) assessment of OCR quality in the *19th Century British Newspapers* (reproduced from Tanner et al. 2009)



What is clear is that the current reported OCR accuracy for historical digitized sources such as the *19th Century British Newspapers collection (part 1)* may not be quite 'good enough' even for simple fuzzy searching. In his incendiary article, Hitchcock (2013), echoing Leary (2005)'s comments about an 'offline penumbra' (see section 2.2.3) hence complains (with respect to fuzzy searching OCR material) that

> *52 per cent of the Burney Collection and a similar proportion of other resources are entirely unfindable, and as importantly it will always be the same 52 per cent, determined by typeface, layout, bleed through and a host of other factors no one has thoroughly investigated.* (Hitchcock 2013: 13-14)

And Prescott (2014: 336), in his response to Hitchcock (2013), calls the OCR issues 'a major issue' for users of the *Eighteenth Century Collection Online*. Assessing the impact of OCR errors on corpus linguistic methods, then, appears essential if these are to become part of the historian's toolkit.

3.3 ESTIMATING THE IMPACT OF OCR ERRORS ON COLLOCATION STATISTICS

The study of collocation patterns (see section 2.3.2.3.1) is central in corpus linguistics. In the collocation-via-significance approach (see 2.3.2.3.1), collocation patterns are detected using statistics such as Log-Likelihood (LL) and Mutual Information (MI). No value of collocation statistics is enough to attribute *importance* to any finding; only a human, subjective, judgement can do this. Nevertheless, collocation statistics help us describe objectively the evidence which we possess, and in this way, can help strengthen and/or evaluate our subjective conclusions.

There are reasons to be concerned that errors in the corpus may affect collocation statistics. All collocation statistics rely at root on frequency counts – and yet in a corpus with OCR errors, frequency counts are unreliable. If frequency counts are unreliable, can statistics based on them be reliable? Conceptually, it is possible to argue that OCR errors should *not* be a concern for collocation statistics. First, a corpus is originally conceived as a sample intended to be representative of a wider population. If the distribution of errors is homogenous across the corpus, then excluding them from frequency counts (in effect from the corpus) should simply lead to the counts reflecting of a random subset of the original corpus, and the results should still be equally representative of the whole population. However, it is reasonable to doubt that OCR errors are homogenously distributed across the corpus. Indeed, there are known factors which affect the quality of the OCR, such as the conservation of the originals, the font and layout of the original, and so on (see previous section), which operate in non-random ways. Hence, *a priori*, all we can say is that OCR errors *may* affect collocation statistics, but that this effect *might* be small; empirical testing is required to clarify this.

Section 3.3.1 discusses the formulas for two common collocation measures, LL and MI, seeking to determine, from a theoretical perspective, which factors related to the distribution of OCR errors may affect collocation statistics and how. Section 3.3.2 then presents the results of empirical testing.

3.3.1 Formulas

MI and LL statistics are measures which describe the association between two words, a *node* and a *collocate*. To calculate these statistics, one picks a *node* (i.e. a word of interest) and a *span* (i.e. the number of words around the node within which *collocates* – words which occur in proximity to the node – will be sought). In this section, I describe how these statistics are computed. In this chapter, I will be following the convention set out in Evert (2005), where the contingency table (which contains the frequency counts used in the calculation of MI and LL statistics) for collocation is expressed as in Table 3.1. Note that 'span' will be taken to refer to the number of words considered right and left of the node (e.g. a span of 3 consists of 3 words to the right and left of the node, i.e. a total of 6 words around the node to be considered), whereas the 'window' will refer to the part of the corpus constituted by the sum total of words which fall within the determined span of each instance of the node in the corpus (e.g. if the span is 3, the window will be a sub-corpus containing a number of words equal to 2x3 multiplied by the number of times the node occurs). Also note that ‘corpus 1’ in the contingency tables refers to the window surrounding instances of a given node, whereas ‘corpus 2’ refers to all parts of the corpus other than the window.

Table 3.1. Contingency table for collocation statistics (observed values)

	Occurrences of collocate	Occurrences of all other words	Totals
in corpus 1	O_{11}	O_{12}	R_1
in corpus 2	O_{21}	O_{22}	R_2
in total	C_1	C_2	N

Collocation statistics compare observed and expected values. Table 3.2 shows the contingency table for expected values, as calculated based on the observed values.

Table 3.2. Contingency table for collocation statistics (expected values)

	Occurrences of collocate	Occurrences of all other words
in corpus 1	$E_{11} = \frac{R_1 C_1}{N}$	$E_{12} = \frac{R_1 C_2}{N}$
in corpus 2	$E_{21} = \frac{R_2 C_1}{N}$	$E_{22} = \frac{R_2 C_2}{N}$

For purposes of implementation, only 3 values appearing in these contingency tables will typically be known: O_{11} (the number of times the node and collocate co-occur within the window), C_1 (the number of times the collocate occurs in the whole corpus) and N (the number of word-tokens in the whole corpus). Two additional values will also typically be known: S_1 (the number of times the node occurs in the whole corpus), and S_2 (the span). All the other values in the table can be calculated based on these known values. The formulas are as follows:

O_{12}	(the frequency of all other words inside the collocation window)	$= R_1 - O_{11}$
O_{21}	(the number of times the collocate occurs outside the window)	$= C_1 - O_{11}$
O_{22}	(the frequency of all words other than the collocate outside the window)	$= R_2 - O_{21}$
R_1	(the total words in the collocation window)	$= (2 \times S_2) \times S_1$
R_2	(all words outside the window)	$= N - R_1$
C_2	(all words other than the collocate in the whole corpus)	$= N - C_1$

Since all other values can be expressed in terms of these 5 known variables (O_{11}, C_1, N, S_1 and S_2), in the discussion that follows, I will focus on how these 5 variables may be affected by OCR errors and how these in turn will impact the statistics.

3.3.1.1 Log-Likelihood

Log-Likelihood is a significance statistic which is often used to help assess whether there is enough evidence in the data to support a conclusion about a given collocation pattern. A high LL value suggests that there is enough evidence to discard the idea that the observed pattern is a fluke.

The formula for LL is twice the sum of the natural logs of the observed values divided by the expected values, each natural log having been multiplied by the observed value (Oakes 1998: 42)[5]:

$$LL = 2\sum_{ij} O_{ij} \ln\left(\frac{O_{ij}}{E_{ij}}\right)$$

3.3.1.2 Mutual Information

Mutual Information is an effect size statistic which is often used to describe the strength of the association between words. A high MI value indicates that two words are strongly associated, i.e. that they co-occur more often than would be expected under the assumption that all words are distributed homogeneously in a given corpus. MI also takes into account the size of

[5] Oakes (1998) writes the LL formula with logarithms in the form ln(A) – ln(B); since ln(A) – ln(B) is equivalent to ln(A/B), his formula is equivalent to the one I give. However, due to the way computers store numbers, implementing the formula in the two different forms can be expected to give rise to very small differences (beyond 3 decimal points). Hence, not much should be made of very small differences between LL values calculated by different software.

the corpus. Infrequent words can attract high MI values; for this reason, it is good practice to use MI in combination with a significance statistic such as LL, to avoid placing too much emphasis on results based on a small number of observations (e.g. Hardie forthcoming).

The formula for MI is the binary log of the observed number of co-occurrences, divided by the expected number of co-occurrences (Oakes 1998: 63)[6]:

$$MI \ = \log_2 \left(\frac{O_{11}}{E_{11}}\right)$$

3.3.2 *In theory: how are OCR errors expected to impact on collocation statistics?*

To illustrate the effect of OCR errors on collocation statistics, let us start by considering the situation where errors are distributed homogenously, i.e. where errors affect the node and collocate equally, affect them equally in all regions of the corpus (including whether or not the node and collocate are co-occurring), and only affect the frequency of the node and of the collocate (but have no effect on the total wordcount). In such a situation, we would expect that the LL statistic derived from the hypothetical OCR data would be smaller than the corresponding LL statistic derived from its hypothetical *gold standard* counterpart (i.e. its error-free counterpart; I will be henceforth referring to this as simply *gold*), because less evidence would be available in the OCR data than in the gold data, some evidence having being made inaccessible by the errors. On the other hand, we would expect the OCR-derived MI statistic to be identical to its gold-derived statistic because the strength of association between the two words should be (virtually) identical in any corpus and its random subsets, and because if the errors are distributed homogenously, our hypothetical OCR data *would*, in effect, be a random subset of the gold data.

[6] Oakes(1998) provides this formula in a different format, in terms of dependent probabilities, taking the binary log of P(collocate|node)/P(collocate). To relate the two, note that P(collocate) = C1/N and P(collocate|node) = O11/R1.

This situation is illustrated in Table 3.3 and Table 3.4, where our hypothetical gold standard data contains 1 million words and 200 occurrences of the node and collocate each, with 50 co-occurrences between the node and collocate within a span of 5 words to the left and right of the node. Table 3.3 illustrates the situation where both node and collocate attract the same error-rate (25%); Table 3.4 shows that our statistics still respond in the expected way, even if we relax one of the assumptions (the assumption that the node and collocate will be affected equally) by giving the node an error rate of 10% and the collocate and error of 50%.

To calculate the number of co-occurrences in the OCR data, given certain error rates, *when it is assumed that errors are distributed homogenously*, I use the following formula, which uses the same notation as in section 3.3.1, but with *new* referring to the value in the hypothetical OCR data, and *gold* referring to the value in the hypothetical gold standard data:

$$new\ O_{11} = new\ C_1 *\ new\ S_1\ (\frac{gold\ O_{11}}{gold\ C_1 *\ gold\ S_1})$$

Table 3.3. LL and MI scores for 25% error rate for both node and collocate, assuming all other variables unchanged

	Corpus size	Span	Frequency of the node	Frequency of the collocate	Co-occurrences	LL	MI
Gold standard	1 million	5	200	200	50	398.37	6.96
OCR data	1 million	5	150	150	28.125	221.86	6.96

Table 3.4 LL and MI scores for 10% error rate for node and 50% error rate for collocate, assuming all other variables unchanged

	Corpus size	Span	Frequency of the node	Frequency of the collocate	Co-occurrences	LL	MI
Gold standard	1 million	5	200	200	50	398.37	6.96
OCR data	1 million	5	180	100	22.5	178.32	6.96

These hypothetical examples, however, assume what we must find out empirically – that the distribution of errors is homogenous. Henceforth, my discussion will avoid this assumption. In the following section, I will attempt to isolate the effect of several variables and discuss whether alterations to that variable could happen empirically.

3.3.2.1 The number of co-occurrences

If only the number of co-occurrences is affected, with all other values unchanged, both statistics are lower in the OCR version than in the gold standard (see Table 3.5). This is the expected behaviour of the statistics, since fewer co-occurrences, all other things being equal, means a weaker relationship between the words, as well as less evidence for the observed pattern. However, this situation will arise spuriously if the OCR errors disproportionately affect the node and/or the collocate when they occur together as opposed to when they occur apart.

Table 3.5LL and MI scores when the number of co-occurrences drops, all other values unchanged

	Corpus size	Span	Frequency of the node	Frequency of the collocate	Co-occurrences	LL	MI
Gold standard	1 million	5	200	200	50	398.37	6.96
OCR data	1 million	5	200	200	25	161.02	5.96

3.3.2.2 The overall number of the collocate, or of the node

If only the overall frequency of the collocate is affected, all other values unchanged, both scores are higher in the OCR version than in the gold standard (see Table 3.6). This is the expected behaviour of the statistics, since a lower collocate frequency without a corresponding drop in co-occurrences means a relatively stronger relationships between the words, as well as relatively more evidence for the observed pattern. The same is true of a higher overall frequency of the node, all other things being equal (see Table 3.7). This will happen spuriously if the errors affect the collocate or node disproportionately when they occur together as opposed to apart.

Table 3.6 LL and MI scores for a lower overall frequency of the collocate, all other things being equal

	Corpus size	Span	Frequency of the node	Frequency of the collocate	Co-occurrences	LL	MI
Gold standard	1 million	5	200	200	50	398.37	6.96
OCR data	1 million	5	200	100	25	484.28	7.96

Table 3.7 LL and MI scores for a lower overall frequency of the node, all other things being equal

	Corpus size	Span	Frequency of the node	Frequency of the collocate	Co-occurrences	LL	MI
Gold standard	1 million	5	200	200	50	398.37	6.96
OCR data	1 million	5	100	200	25	468.67	7.96

3.3.2.3 The span

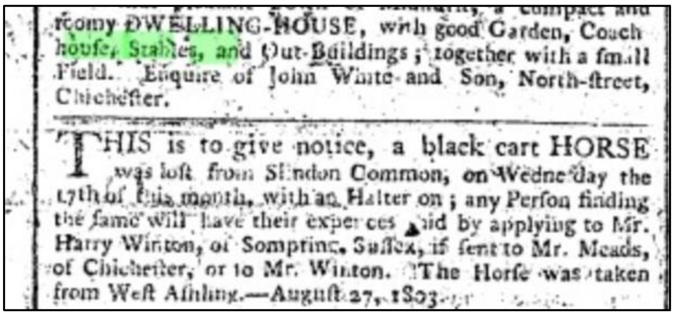
If the span is widened, everything else being equal, the MI and LL scores in the OCR version will be lower than in the gold standard (see Table 3.8). This is the expected behaviour of the statistics, since a wider span should normally lead to a higher number of co-occurrences, hence a widening of the span without a corresponding increase in the number of captured co-occurrences leads to a weaker pattern, with less evidence for it. This situation could arise spuriously if spurious characters and spaces disproportionately inflate the wordcount around the node (i.e. inside the collocation window) as compared to further away from the node.

Table 3.8 LL and MI scores for a wider span, everything else being equal

	Corpus size	Span	Frequency of the node	Frequency of the collocate	Co-occurrences	LL	MI
Gold standard	1 million	5	200	200	50	398.37	6.96
OCR data	1 million	10	200	200	50	329.02	5.96

This is not as far-fetched as it may seem. In the example provided in Figure 3.4, the 4 words 'houses, stables and out-buildings' in the image of the original (on the left) are replaced by 8 words 'a hoof Stah az ad O urt etldings' in the OCR transcription (on the right). Hence, whereas in the original, 'houses' and 'stables' would both be included in the window of 'coach' (immediately preceding 'houses') with a span of 2, in the OCR, only 'hoof' would fall in the window with a span of 2, and 'Stah' would require a span of 3 or more to fall within the window (though without much benefit in this case, since both words are misspelled in the OCR),

Figure 3.4 Original and OCR transcription (from HPTE 29/08/1803, p.2)



compact and roomy DWELLING-HOUSE, with good Garden, Coach houſe, Stables, and Out-Buildings; together with a ſmall Field. Enquire of John White and Son, North-ſtreet, Chicheſter.

THIS is to give notice, a black cart HORSE was loſt from Slindon Common, on Wedneſday the 17th of this month, with an Halter on; any Perſon finding the ſame will have their expences paid by applying to Mr. Harry Winton, of Sompting, Suſſex, if ſent to Mr. Meads, of Chicheſter, or to Mr. Winton. The Horſe was taken from Weſt Aſhling.—Auguſt 27, 1803.

```
WEL IN 14tSE vWi II goodGarden Coach a
hoof Stah az ad O urt etldings togaset willafmd
li dltLT loh and Son
Cildch er
isto nrive 17ORSE
E w iLi MnIV C oimon the
17 t t d ia njtt ar a I it r aon any co fiding
tee at iil avetheir exnpei ces id b plyn t r
ufS fi ftmto Mr lilaids j
ofCii eier ortoMrW tn The was takein
20 CiceiirTHAr Mr Wtt t ton I p
```

3.3.2.4 The corpus size

If the corpus size is inflated, all other values remaining equal, both statistics will be higher in the OCR data than in the gold standard (see Table 3.9). For LL, this makes intuitive sense: a larger corpus means a greater evidence base. For MI, however, this may appear paradoxical at first sight. It can be glossed as follows: if the corpus size is larger but the occurrences and co-occurrences have remained unchanged, then in relative terms we are now dealing with *less frequent* nodes and collocates. If the node and collocate are less frequent, then they should be expected to co-occur less often, but the number of co-occurrences is also unchanged, hence the greater effect size statistic. This situation might arise spuriously if the word count is inflated, and unevenly so (e.g. by spurious characters and spaces occurring disproportionately outside the collocation window as opposed to inside it). (If the wordcount was inflated proportionally inside and outside of the window, then the number of co-occurrences should be proportionally reduced as some of the co-occurrences would now fall outside of the window; see section 3.3.2.1).

Table 3.9 MI and LL scores for a larger corpus size, everything else unchanged

	Corpus size	Span	Frequency of the node	Frequency of the collocate	Co-occurrences	LL	MI
Gold standard	1 million	5	200	200	50	398.37	6.96
OCR data	2 million	5	200	200	50	467.39	7.96

3.4 SUMMARY

The dataset used in this thesis is the *19th Century British Newspapers* collection (part 1) owned by the British Library. The data was provided to the *Spatial Humanities* project by the British Library in the form of OCR output. OCR, however, is of variable effectiveness on historical material. Tanner et al. (2009) hence report an average word accuracy rate of 78% for the collection. The central question asked in this chapter is: how will the OCR errors impact the statistics used in corpus linguistic analysis? The chapter focuses on two common collocation statistics, Mutual Information (MI) and Log Likelihood (LL). It is reasonable to expect that these statistics will be affected by OCR errors, since they depend on frequency counts which are inevitably affected by OCR errors. In theory, the statistics will be affected by the distribution of errors across word-types, the distribution of errors across instances of a single word-type, the distribution of errors within a corpus, and the existence and distribution of spurious characters or spaces affecting wordcounts.

To summarize, both statistics will be spuriously *low* if the errors disproportionally affect the node or collocate when they occur together (as opposed to apart); and if spurious characters or spaces are more likely to occur around the node than elsewhere in the corpus, causing a disproportional inflation of the wordcount inside the collocation window (as opposed to outside of it). Conversely, both statistics will be spuriously *high* if the errors disproportionally affect the node or collocate when they occur apart and if the wordcount is disproportionally inflated outside of the window than inside it (i.e. if spurious characters or spaces are less likely around the node than farther away from it). Moreover, both statistics are sensitive to corpus size, which means that using either statistic to compare words across (sub)corpora of different sizes (or with different error rates) may be problematic.

Table 3.10 summarizes the relationship between MI and LL scores and the 5 variables typically used to calculate them (corpus size, span, frequency of the node, frequency of the collocate, number of co-occurrences). The table shows the direction of change for MI and LL scores when one of these 5 variables decreases, assuming that all other variables remain unchanged.

Table 3.10 Relationship between MI and LL scores and 5 variables

	MI	LL
decrease in corpus size	decrease	decrease
decrease in span	increase	increase
decrease in frequency of the node	increase	increase
decrease in frequency of the collocate	increase	increase
decrease in number of co-occurrences	decrease	decrease

In addition, the following factors related to OCR errors have been identified as being potential disruptors of MI and LL scores:

1. the distribution of errors within portions of a corpus (in particular if they affect the distribution of errors across instances of a single word-type)
2. the existence of spurious characters or spaces affecting wordcounts, and their distribution in the corpus

In the next chapter, I will examine the empirical impact of OCR errors on MI and LL statistics by calculating equivalent statistics in a hand-corrected sample and in a paired uncorrected OCR sample.

4 Assessing the empirical impact of OCR errors on frequency counts and collocation statistics

4.1 Introduction

The previous chapter explored the theoretical impact of OCR errors on MI and LL statistics. This chapter will explore their impact using observational data, the CNNE matching corpus, which is introduced in section 4.2. Comparative measures used in this chapter and the next are introduced in section 4.3. General observations of the data are presented in section 4.4. Finally section 4.5 considers the empirical impact of OCR errors on MI and LL.

4.2 Assembling the CNNE matching corpus

The CNNE matching corpus is a set of text samples which I assembled for the purposes of this thesis. It contains 9 versions of the same source material, articles from the *19th Century British Newspapers collection* (part 1) (see section 3.2). One version, the *gold standard* (which will also be referred to as the *gold sample*), consists of a near-perfect[1] rendition of the source material. One version, the *uncorrected sample*, consists of uncorrected OCR text from the British Library OCR for the *19th Century British Newspapers* collection. The remaining 7 versions consist of OCR samples which have been automatically corrected using two different programs; these versions will be introduced and discussed in chapter 5. I will use the term *OCR sample* to refer to any of the 8 samples *other than* the gold sample.

The starting point for assembling the CNNE matching corpus was the set of files making up the Corpus of Nineteenth-Century Newspaper English (CNNE). CNNE was created by Erik Smitterberg at Uppsala University for the purposes of investigating diachronic changes in historical English. When I was granted[2] access to the corpus in June 2015, the corpus contained

[1] The gold standard is assumed to be perfect, but human analysts do make mistakes. Since no elaborate cross-checking procedures were used, the gold standard is described as merely *near-perfect* (rather than *perfect*).

[2] Permission for me to use this corpus for the purposes of assessing the effectiveness of OCR post-correction procedures and the impact of OCR errors on corpus linguistic analysis was granted by Smitterberg himself.

200 articles published during two periods, 1830-1850 and 1875-1895. The articles were sourced from English provincial and metropolitan newspapers; together, they amounted to over 300,000 words (English Department Uppsala University undated; Smitterberg 2014; Varieng 2014). The CNNE files were composed using images of c19th newspapers available online via the diffusion portals of the British Library's *19th Century Newspapers online*. Smitterberg downloaded the images of individual newspaper pages, OCR'ed them himself using commercial software, then proof-read them twice (once within the OCR software and once more after exporting the output)[3].

Not all of the CNNE files came from newspaper titles available as part of the British Library's *19th Century British Newspapers (part 1)* collection for which I had the OCR data. The first task for assembling the CNNE matching corpus was therefore to identify the overlap between the CNNE files and the British Library OCR data in my possession; there were 110 overlapping files. The next step involved precisely matching the text in these CNNE files to the OCR data. Since the OCR was not always clear enough to unambiguously match it to the correct text without referring to the original sources, this process involved me consulting the images of the original sources on Gale/Cengage's online portal for the *19th Century British Newspapers collection (part 1)*. I was able to match the OCR text to the CNNE text with a high degree of confidence for 107 files. This OCR text thus formed the uncorrected sample. To assemble the corresponding 107 gold sample files, I removed the minimal annotation which was present in the original CNNE files and trimmed them to the precise extent of the OCR text that I had been able to match. The articles included in these final samples come from 12 different publications (see Figure 4.1) and 17 years between 1830 and 1892 (see Figure 4.2). The full list of sources is available in appendix 10.4. The gold sample contains 160,616 word-tokens, the uncorrected sample 162,617 (see also Table 4.1).

[3] Erik Smitterberg, personal communication, 09/06/2015.

Figure 4.1. Files per publication in the CNNE matching corpus

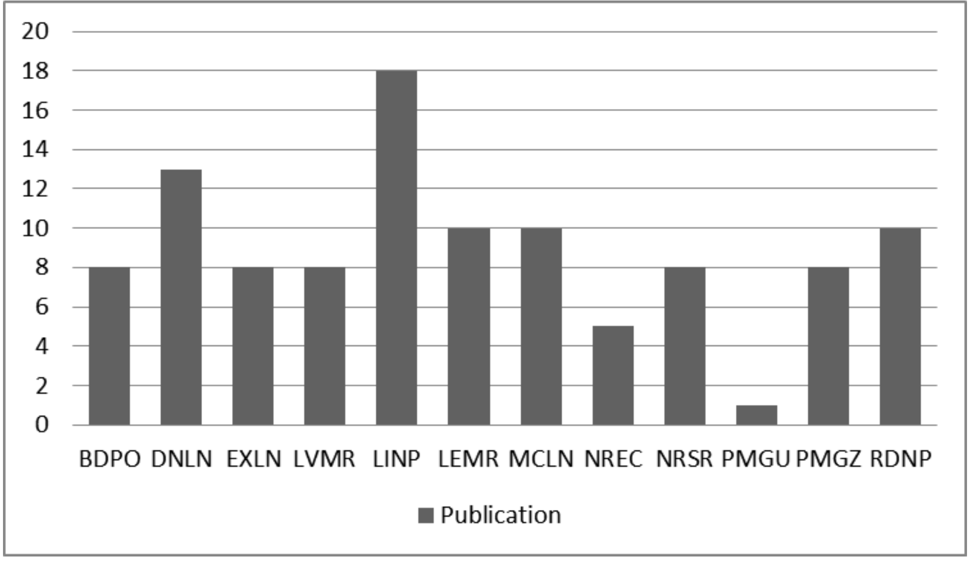
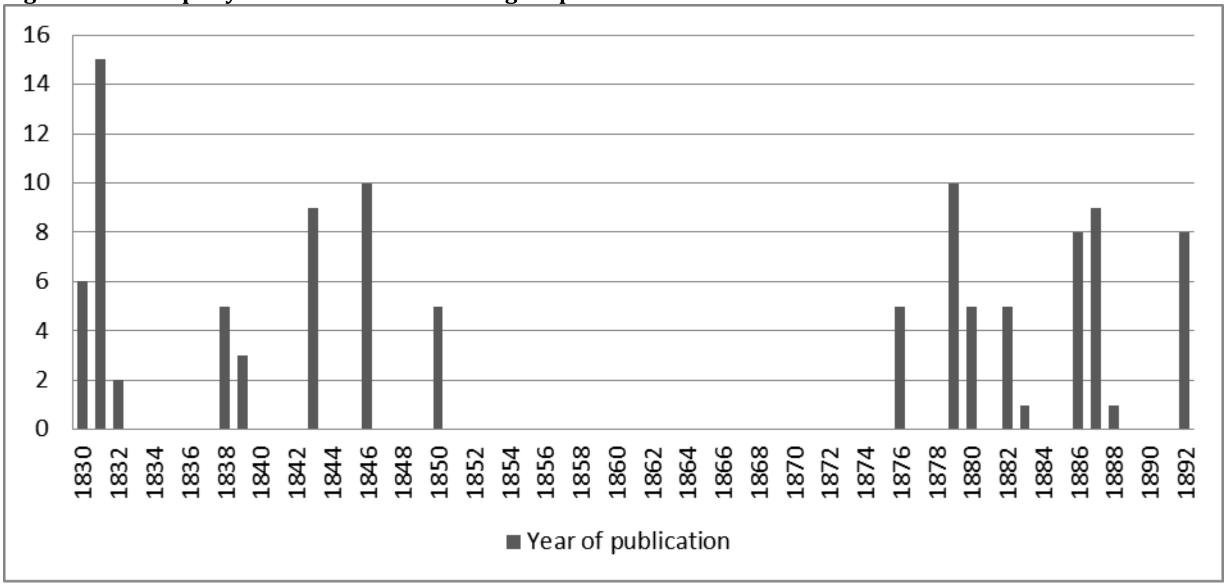


Figure 4.2. Files per year in the CNNE matching corpus



4.3 Methods and measures of comparison

This chapter and the next draw on various methods and measures of comparison which are introduced in this section. Section 4.3.1 introduces *file edit-distance*, a unit of measure which describes the difference between two files, so called in reference to *edit-distance*, a common measure for comparing two strings[4]. Subsequently, section 4.3.2 defines the terms *recall*,

[4] *String* is a computational term which refers to a sequence of characters (as opposed to other types of data such as numbers). *This* is a string, as is *dlkjf.3f*.

precision, *false positive* and *false negative*, which will be used to describe and evaluate the success of corrective procedures, as well as the impact of OCR errors on collocation statistics.

4.3.1 FILE EDIT-DISTANCE

To help compare files in different samples of the CNNE matching corpus, I used the OCReval tool (Carrasco 2014), created and made available as part of the IMPACT project (Balk 2009; Balk and Conteh 2011). The tool automatically aligns sets of files and generates various so-called *error rates*[5]. The rate of interest to us is what Carrasco calls the *word error rate*, but which I will call *file edit-distance*. This represents the number of words needing to be changed (either deleted, substituted for another word, or inserted) in one set of files (deemed 'bad') in order to obtain the other set of files (deemed 'good'). This number is normalized per 100 words *in the version deemed good*; this means the figure can reach more than 100% if more than 100 words in the version deemed bad need changing in order to obtain 100 words in the version deemed good.

Generally speaking, *edit-distance* is a standard method for comparing the similarity between two strings. An edit-distance quantifies the number of operations required to change one string into another. Operations can be deletions, insertions or substitutions. At an edit-distance of 1 from *tree*, for example, we find strings such as *thee*, *free*, and so on. The most common operationalisation of edit-distance is known as ‘Levenshtein distance’ (Levenshtein 1966). The OCReval interface labels its output measure an ‘error-rate’, but because of the way it is computed, it resembles more closely a standard edit-distance. However, unlike the usual form of edit-distance, which counts operations on *characters* within a *string*, the measure produced by OCReval counts operations on *words* within a *file*, in order to quantify the similarity between two files. This is why I dub this *file* edit-distance.

[5] OCReval is programmed to compare two sets of files as a whole: running OCReval on two folders, each containing the files from one sample, returns a single report. For the Spatial Humanities project, Andrew Moore implemented a modified version of OCReval which allows the user to make file-by-file comparisons. This modification allowed me to assess the variability of the OCR quality from one file to another, and from one group of files (e.g. grouped by year or publication) to another.

4.3.2 Recall, precision, false positives, false negatives

The terms *recall*, *precision*, *false positives* and *false negatives* refer to standard statistical concepts used across disciplines including computer science, medicine and psychology. They describe the success of any retrieval procedure, whether it be a search-engine query or a significance test. *Recall* and *false negatives* are two sides of the same coin: *recall* refers to the proportion of desired results which were retrieved, with *false negatives* referring to those desired results which failed to be retrieved. Likewise, *precision* and *false positives* are the two sides of another, related coin: *precision* refers to the proportion of retrieved results which were desired, with *false positives* referring to those retrieved results which were not desired.

In this thesis, *false positives* and *false negatives* are used in the context of statistical procedures. In section 4.5.5, for example, the *false positives* are those node-collocate pairs which are above a statistical threshold in the uncorrected sample but below it in the gold sample, and the *false negatives* are those node-collocate pairs which are below a statistical threshold in the uncorrected sample but above it in the gold sample.

In chapter 5, I will often use *recall* and *precision* in the context of evaluating corrective procedures. In this context, *recall* represents the number of changes which have been made by a corrective procedure, as a proportion of the changes which are needed to make the text completely correct. *Precision* refers to the proportion of changes actually made which were correct, as opposed to those which introduced a new error (by substituting an error with another error, or by substituting a correct word with an incorrect one).

It is possible to produce an exact measure of recall and precision for corrective procedures (by manually verifying each change) but in this chapter and the next, I will use an estimate calculated on the basis of the file edit-distances provided by OCReval, since this is much faster to generate than using manual analysis. The figures will be estimates, because they will be calculated based on *average* file edit-distances per file, which implies that smaller files will be accorded a greater weighting than their size would warrant. Another source of imprecision is that the difference in overall corpus size between the different corpus versions is not taken into account; however, since this difference is not too great (see Table 4.1 and Table 5.7), it will not much distort the figures. The recall and precision figures will be calculated using the following formulas:

$$recall \ = \frac{changes\ made}{total\ errors}$$

where *changes made* is taken to be the average file edit-distance between the uncorrected and corrected versions, and *total errors* is taken to be the average file edit-distance between the uncorrected and gold versions.

Similarly,

$$precision \ = 1 - \frac{\frac{1}{2}(changes\ made - improvement)}{changes\ made}$$

where *changes made* is as above, and *improvement* is taken to be the difference between how bad the uncorrected files were (i.e. the file edit-distances between the uncorrected and gold versions) and how bad the corrected versions are (i.e. the file edit-distances between the corrected and gold versions) (see also section 4.3.1)[6]. It is possible to show that this formula equals the definition of precision provided above:

$$precision = \frac{good\ changes}{changes\ made} \qquad \text{(by definition)}$$

$$precision = \frac{changes\ made\ -\ bad\ changes}{changes\ made}$$

$$precision = \ 1 - \frac{bad\ changes}{changes\ made}$$

[6] At first glance, it may seem logical to consider *improvement* to be equivalent to the average file edit-distance between the uncorrected and corrected files. However, this comparison would not provide a correct value for the improvement, because *all* differences between uncorrected and corrected files would count as improvements, whereas in fact some of the changes may consist of a correct word incorrectly changed, or an incorrect word changed to another incorrect word. The calculation of *improvement* provided above avoids this pitfall.

The quantity $bad\ changes$ is equal to $\frac{1}{2}(changes\ made - improvement)$ since $improvement$ is equal to $good\ changes - bad\ changes$ (and $changes\ made$ equals $good\ changes + bad\ changes$).

4.4 RESULTS: OCR QUALITY

Before looking at the impact of OCR errors on collocation statistics, it is interesting to answer some general questions about the quality of the OCR in the CNNE matching corpus. Without a gold standard, it is difficult to answer questions such as how many errors are present? Do errors tend to be frequent? Do real word errors occur often? This section takes advantage of having a gold standard at hand to make observations related to these questions. Section 4.4.1 shares some observations based simply on comparing overall type and token counts in the gold and uncorrected CNNE samples. Section 4.4.2 shares some observations based on file edit-distances (generated with OCReval, see 4.3.1) between the gold and uncorrected CNNE samples.

4.4.1 USING SIMPLE FREQUENCY COMPARISONS

What are OCR type and token counts like? It might be expected that token counts in OCR data would be larger than they should be, because of problems of segmentation (i.e. placement of spaces) and stray characters (see section 3.3.2.4). At the same time, token counts might be expected to be smaller if the OCR software simply missed out some words or portions of text. Hence a first interesting observation is that there is only a very small difference in wordcount between the gold and OCR versions in the CNNE matching corpus, the OCR version containing only around 1.25% more words (see Table 4.1).

Table 4.1 Type and token counts in the uncorrected and gold CNNE matching corpus

	Gold corpus	OCR corpus
N (tokens)	160616	162617
N (types)	13831	26954
Type/token ratio	8.60%	16.57%

How about the type count? We would expect an OCR type count to be higher than it should be, since OCR texts will include non-existent words (due to errors), but it is difficult to predict the magnitude of this effect. The second interesting observation is hence that the uncorrected type count is roughly double the gold type count, and that the type/token ratio is likewise about twice as high in the uncorrected sample as in the gold sample (see Table 4.1).

Since there is such a difference between uncorrected and gold type counts, it is interesting to take a closer look at the types present in the OCR data. Are all types in the gold data present in the OCR data, or are some of them missing? What proportion of types in the OCR data are absent from the gold data (and hence definitely errors)? The third interesting observation is hence that most but not all (86%) of the types occurring in the gold corpus also occur in the OCR corpus, but that only 44% of types in the OCR corpus also occur in the gold corpus; the remaining 56% of OCR types are hence errors (first line of Table 4.2).

56% is a high count of errors. A reasonable assumption is that most errors will occur only very infrequently, so can many of these errors be eliminated by using a *frequency floor* (i.e. a frequency filter which excludes words which occur less often than a set frequency value)? The third line of Table 4.2 shows that of the types which occur at least 10 times in the OCR corpus, 93% also occur in the gold corpus. This suggests that using a frequency floor when working with OCR data can be useful, since most of the remaining types are correct. Nevertheless, using a frequency floor does not eliminate all errors since 7% of types occurring more than 10 times in the OCR corpus do not occur in the gold corpus; this also means that some errors are not, as may be expected, rare.

Table 4.2 Relationship between the types occurring in the uncorrected and gold samples

Types occurring in…	Count	% of types in the OCR corpus (occurring at least 10 times)
both OCR and gold samples	12005	44.54
OCR sample, and at least 10 times in the gold sample	1865	6.92
at least 10 times in the OCR sample, and gold sample	1594	(93.38)

Beyond overall type and token counts, how reliable are token counts for individual types, or, in other words, how reliable is the frequency count for a given word of interest? A natural expectation would be that the OCR frequency count of types which occur in the gold corpus should always be the same or smaller than the frequency count of that type in the gold corpus. However, this would be failing to take into account *real-word errors* (an error which happens to coincide with an existing word, e.g. *Prussia* without its *P* spells *Russia*, which is still a real word), which could inflate the frequency count of a particular type.

Table 4.3 shows that 57% of types in the OCR corpus occur more often in the OCR corpus than in the gold corpus; however, this number falls to only 15% when taking into account only types occurring at least 10 times in the OCR corpus. This is a huge proportion, but the explanation for this is straightforward: 55% of types in the OCR corpus do not occur in the gold corpus (and these count towards the types which occur more often in the OCR corpus than in the gold corpus). So in fact, only 2% of OCR types which occur in the gold version attract inflated frequency counts (which presumably involve real-word errors). Note, however, that when considering only words which occur at least 10 times in the OCR corpus and *do* occur in the gold corpus, the proportion rises to 8% of types attracting inflated frequency counts. This implies that real-word errors may be encountered more often when working with a frequency floor than when working without. Whilst this is an interesting observation, it is also a logical one: most types occurring less than 10 times in the OCR corpus are errors and hence cannot be confused with real-word errors, so real-word errors will be diluted when all types in the OCR corpus are considered (in contrast to when only the types occurring at least 10 times are considered).

However, to say that 8% of types are over-reported due to real-word errors is not the same as saying that 8% of tokens are real-word errors. From these figures, it is impossible to tell what proportion of tokens may be described as real-word errors in the OCR data. Both OCR frequency counts which happen to match their gold counterpart and deflated OCR frequency counts might also include real-word errors, since some of the instances of a given type may be incorrect or missing, offsetting the impact of the real-word errors on the frequency count.

So, in the final analysis, is it indeed the case – as would be expected – that most words are under-reported in the OCR version? Yes: most types (75%) which occur more than 10 times in the OCR corpus are under-reported compared to the gold data. However, because many OCR types do not occur in the gold version, the figure without a frequency floor is actually as low as 16%: only 16% of OCR types overall occur less often in the OCR data than in the gold data.

Table 4.3 Over- and under-estimates for uncorrected type frequencies compared to gold type frequencies

	Count of types (out of all types in the OCR corpus)	**Count of types (out of all types occurring at least 10 times in the OCR corpus)**
Occur more often in the OCR sample	15492	263
%	57.48	15.41
… and don't occur in the gold (suspected non-dictionary words)	*14949*	*113*
%	*55.46*	*6.62*
…and do occur in the gold (involving suspected real-word errors)	*543*	*150*
%	*2.01*	*8.79*
Occur as often in the OCR and gold sample	7117	157
%	26.4	9.2
Occur less often in the OCR sample (and do occur in the gold sample)	4345	1287
%	16.12	75.4

4.4.2 Using OCReval

Looking at overall type and token counts is insightful, but does not provide us with an estimate of the proportion of individual tokens which are actually errors in the OCR corpus, nor does it tell us much about how variable OCR quality may be from one section of the corpus to another. Here is where OCReval (see section 4.3.1) is useful; comparing the files in the uncorrected sample one-by-one to their gold counterparts allows us to assess whether the errors affect different parts of the sample to the same extent.

Table 4.4 Quality of OCR in original files, file edit-distance between raw and gold versions

Min	3.63
Median	15.75
Max	97.43
Mean	22.23
N (files)	107

The files composing CNNE (see also section 4.2) were selected from only certain types of articles, and only from images which were deemed easily readable (to a human). Hence, one would expect that the average quality of the OCR in the CNNE matching corpus would be better than that in the whole *19th Century Newspapers* (part 1) collection. If this is found to be indeed the case, this would suggest that there is a relationship between what is deemed ‘readable’ by a human and the performance of OCR software.

Tanner et al.'s (2009) word accuracy rate measures the percentage of words in their OCR samples which correctly match the words in their re-keyed gold standard. Although this figure is not directly equivalent to my file edit-distance, their accuracy rate can be considered to correspond to 1 minus my file edit-distance, simplifying somewhat. Surprisingly, 1 minus the average file edit-distance (see Table 4.4) for the CNNE matching corpus (22.23%) is virtually the same as the word accuracy rate of 78% for the *19th Century Newspapers* collection reported by Tanner et al. (2009) (see section 3.2.2).

This seems to suggest that human image readability is weakly related, if at all, to OCR-software image readability. This would possibly in turn suggest that the quality of the image (which is related to the state of conservation of the original material) is not such an important factor in OCR-software performance. Indeed, Tanner et al. do mention that the OCR errors are due not just to deterioration of originals, but also to problems with the software. If this is true, it suggests that there is scope for improving the quality of the digitized material by improving OCR technology, or by using human post-editing[7].

However, the comparison between the overall file edit-distance for the CNNE gold and OCR versions and the error rates reported by Tanner et al. is somewhat misleading, because the figures provided by Tanner et al. are *upper thresholds*, or estimates of the best quality achieved in the data that they tested (see section 3.3.2). So if the figure for CNNE is comparable to that of Tanner et al., that suggests that overall OCR quality in CNNE may well be higher than in the collection tested by Tanner et al. However, another misleading factor in the comparison is that Tanner et al. determined what constituted the 'best' parts of the tested collections by selecting the best portions of particular images. In other words, they started off by assuming that image readability by humans is related to OCR-software performance. This means no conclusion can be drawn about this relationship from comparing the CNNE figures to Tanner et al.'s figures.

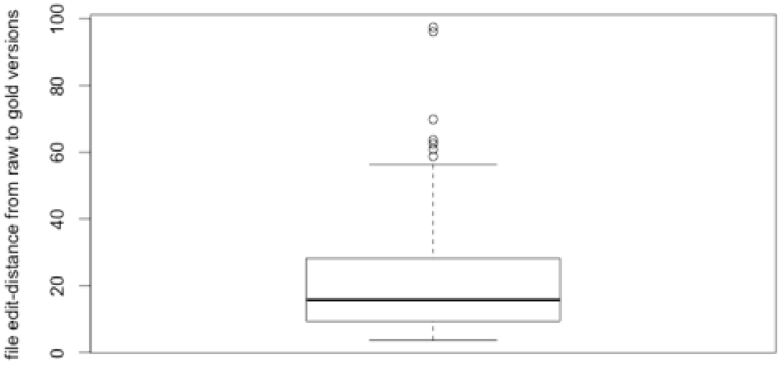
Beyond the overall figures, how variable is OCR quality in CNNE? One might expect only fairly good OCR quality (since the images were selected as among the most readable to a human). Instead, there appears to be quite a spread in the quality of the OCR from one file to another in the CNNE matching corpus, as shown in Figure 4.3. Figure 4.3 is a box plot; hence the values are plotted on the y axis (the width is irrelevant). The thickest horizontal line shows the median[8] value. The bottom-most and top-most horizontal lines (at the end of the 'whiskers') show the range of the data (excluding outliers, which I will come back to in a moment). The

[7] Laurence Anthony points out to me that the *micro job* model (e.g. Amazon's *Mechanical Turk*) can achieve fast and reliable results at relatively low cost.

[8] The median is the middle value when the values in a set are arranged from smallest to largest (e.g. the 5th of 9 values).

horizontal lines forming the 'box' indicate the range within which falls 50% of the data, known as the *interquartile range*[9]. Hence, the 'whiskers' show the extent of the remaining 25% of the data on either side (excluding outliers). *Outliers* are values which are exceptionally far from the median; by definition, 'exceptionally far' is farther than 1.5 times the interquartile range from the box edges (which mark the lower and upper quartiles[10]). If there are outliers, the whisker extends only as far as the last value which is not an outlier. Outliers are then shown as individual dots. Although most of the files have file edit-distances between 10 and 30%, the full spread of file edit-distances ranges from 3.63% to 97.43%. For my purposes, this is good news, since it allows us to evaluate the corrective procedure on a range of file-qualities, not just on relatively good OCR. This result also suggests that there may, indeed, be a weak relationship, or none at all, between human image readability and OCR-software readability.

Figure 4.3 Variation in quality of OCR in original files



Beyond this, it is interesting to examine some key variables in the CNNE matching corpus. Figure 4.4 and Figure 4.5 show how the file edit-distances vary by year and publication. Although both figures show extensive variation, suggesting that the spread and average quality of the OCR may be expected to vary considerably depending on the year and publication of the sample considered, two caveats should be born in mind.

[9] The range between the values situated at 25% of the data and 75% of the data (e.g. the 10th and 30th in a series of 40 values).

[10] The *lower quartile* is the value situated at 25% of the data (e.g. the 10th in a series of 40 values); the *upper quartile* is the value situated at 75% of the data (e.g. the 30th in a series of 40 values). These are shown by the horizontal lines delimiting the box.

First, for some years and publications, less than 5 files are available in the sample; the average and spreads for these years and publications are hence at best only very broadly indicative of the 'true' values for those years and publications. PMGU has less than 5 files; NREC has exactly 5 files; the other publications have between 8 and 13 files (see Figure 4.1). Only the following years have more than 5 files (between 8 and 15): 1831, 1843, 1846, 1879, 1886, 1887 and 1892 (see Figure 4.2).

Figure 4.4 Quality of OCR in original files, per year

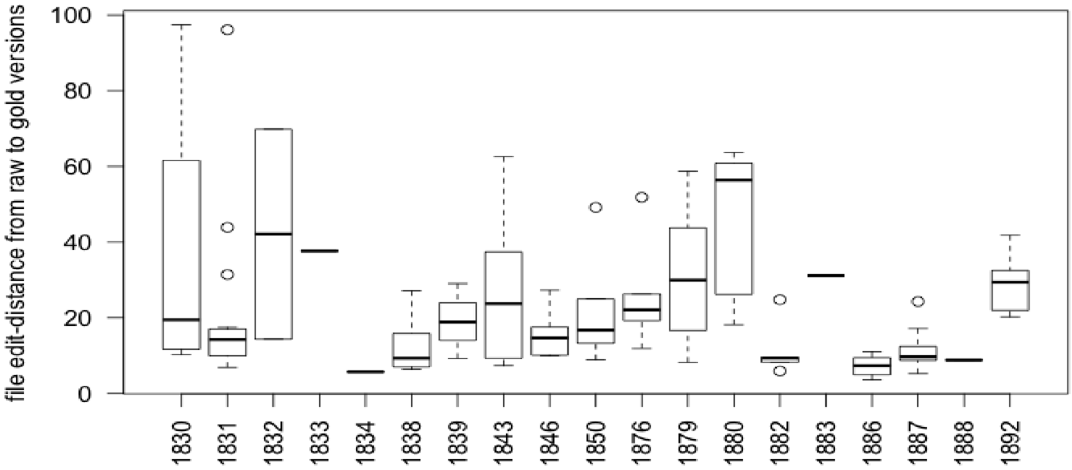
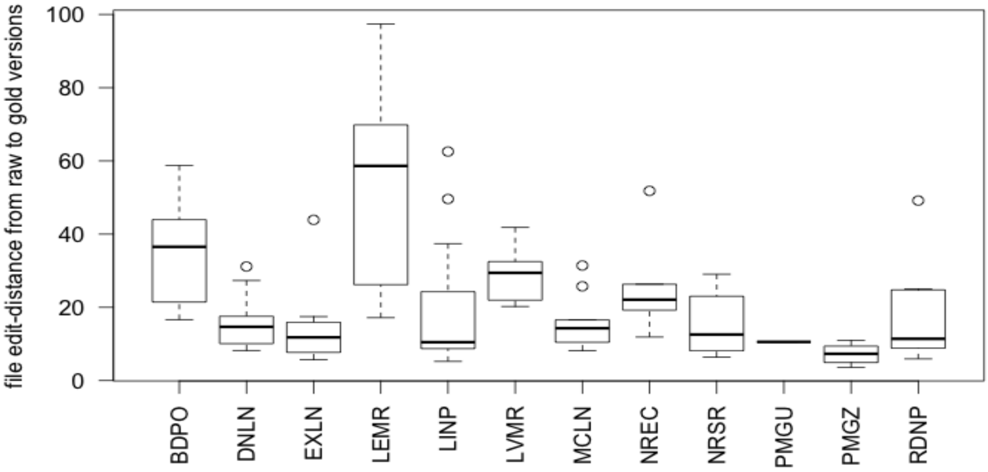


Figure 4.5 Quality of OCR in original files, per publication



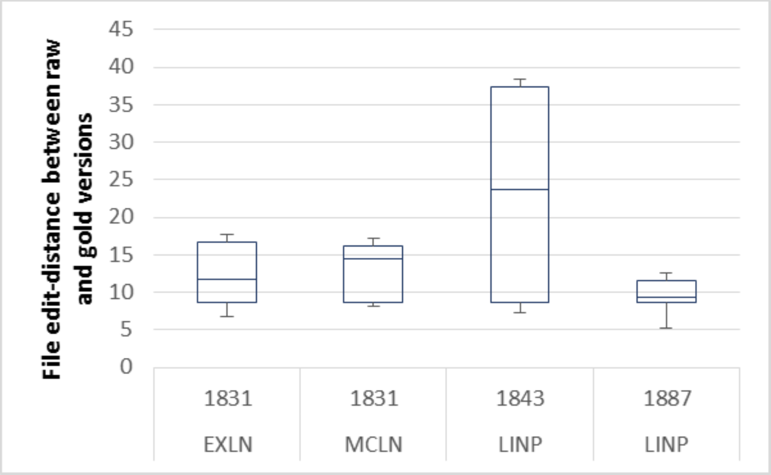
Second, year and publication are not independent variables in the CNNE matching corpus. No sampled year contains files from more than 4 different publications represented in the sample, and no sampled publication contains files from more than 6 different years represented in the sample (see Table 4.5). Given this situation, it would be interesting to ask whether there is a notable difference in the average and spread of OCR quality between different publications in the same year and the same publication in different years. However, there is only one year in which there are two publications with more than 5 files for that year-publication pair; likewise, there is only one publication for which there are two years in which those year-publication pairs have more than 5 files (see Table 4.5).

Table 4.5 Interaction between publication and year in the CNNE matching corpus

	1830	1831	1832	1833	1834	1838	1839	1843	1846	1850	1876	1879	1880	1882	1883	1886	1887	1888	1892	TOTAL
BDPO												8								8
DNLN									10			2			1					13
EXLN		6	1		1															8
LVMR																			8	8
LINP								9									8	1		18
LEMR	1	1	1	1									5				1			10
MCLN	3	7																		10
NREC											5									5
NRSR						5	3													8
PMGU		1																		1
PMGZ																8				8
RDNP										5				5						10
TOTAL	4	15	2	1	1	5	3	9	10	5	5	10	5	5	1	8	9	1	8	107

The figures presented in Figure 4.6 can hence only be considered anecdotal. Although they seem to suggest that year is the more important factor, this impression is confounded by the observation (see Figure 4.5) that EXLN and MCLN have similar means and spreads in the first place. In fact, the close interaction between year and publication in this sample make the interpretation of Figure 4.4 and Figure 4.5 particularly perilous. LEMR, for example, is the publication with the greatest spread in file edit-distances; however it is also the publication with the greatest spread in terms of years represented in the CNNE matching corpus, so it is difficult to ascertain whether either year or publication is the determining factor. Nevertheless, there are clear differences across publications: BDPO and PMGZ both occur in only one year, yet have very different spreads. However, the difference between these publications may be an artefact of the sampling, so even here, no firm conclusion can be drawn from the observation of these figures.

Figure 4.6 Effect of year and publication on quality of OCR



4.5 RESULTS: COLLOCATION STATISTICS

Section 4.4 has shown that OCR errors do have an impact on the wordcounts in the uncorrected CNNE matching sample. This section explores the impact of OCR errors on the collocation statistics themselves. To do this, I will compare collocation statistics calculated from the gold and uncorrected CNNE matching samples. I also calculated collocation statistics for one of the corrected samples; the comparison with those is discussed in section 5.4.4.

4.5.1 CALCULATING THE COLLOCATION STATISTICS

The collocation statistics were calculated as follows. I chose 140 nodes from across the whole range of frequencies in the gold sample (these are listed in Table 4.6, Table 4.7 and see also appendix 10.2 for the full list alongside their frequencies in the three samples for which collocation statistics were calculated). For each node, I then identified all the words which co-occurred with that node within several spans: I considered spans of 3, 4, 5, 10, 20 and 50 words to the left and right of the node. For each node/collocate pair, I then calculated the MI and LL statistics using the formulas given in section 3.3.1.

Table 4.6 Test nodes and their frequencies in the CNNE gold and uncorrected matching samples (ordered by descending gold frequency) (1/2)

	Word	Frequency (gold)	Frequency (OCR)
1	the	14559	12838
2	of	7298	6989
3	to	4646	4500
4	that	1878	1727
5	were	1013	858
6	who	529	455
7	time	270	244
8	great	222	204
9	after	207	191
10	police	158	111
11	place	153	127
12	government	137	97
13	london	107	85
14	law	105	90
15	liverpool	79	61
16	railway	72	59
17	power	65	59
18	england	60	42
19	english	58	45
20	heard	58	45
21	building	52	43
22	murder	52	45
23	afternoon	47	36
24	manchester	44	25
25	although	40	29
26	weather	35	27
27	serious	31	24
28	used	28	26
29	bradford	25	11
30	fall	25	31
31	medical	25	20
32	bristol	24	22
33	somewhat	23	19
34	beyond	20	15
35	engines	19	18
36	mansion	18	7
37	ministry	17	16
38	war	17	15
39	birmingham	16	9
40	raised	16	15
41	plain	15	12
42	powers	15	13
43	france	14	9
44	owners	14	12
45	paris	14	14
46	daily	13	12
47	settlement	13	7
48	fresh	12	11
49	bearing	11	8
50	religion	11	8
51	render	11	7
52	reports	11	10
53	attacked	10	7
54	battalion	10	10
55	disease	10	10
56	empire	10	9
57	europe	10	5
58	merchant	10	7
59	stop	10	10
60	agricultural	9	8
61	battle	9	6
62	brussels	9	1
63	creatures	9	6
64	indian	9	8
65	loud	9	10
66	thinking	9	8
67	dublin	8	7
68	efficient	8	8
69	lancashire	8	5
70	meredith	8	5
71	sitting	8	6
72	americans	7	4
73	banstead	7	7
74	belgian	7	3
75	belgium	7	1
76	cambridge	7	7
77	enjoy	7	8
78	netherlands	7	2
79	russia	7	6
80	russian	7	4

Table 4.7 Test nodes and their frequencies in the CNNE gold and uncorrected matching samples (ordered by descending gold frequency) (2/2)

	Word	Frequency (gold)	Frequency (OCR)
81	spent	7	7
82	asylum	6	5
83	declaration	6	4
84	grave	6	6
85	offences	6	4
86	rod	6	5
87	science	6	5
88	sewage	6	6
89	sheffield	6	2
90	wanting	6	4
91	blockade	5	5
92	creature	5	3
93	flag	5	1
94	knots	5	5
95	perform	5	4
96	rothschild	5	3
97	unconscious	5	5
98	arthur	4	4
99	cheap	4	4
100	deed	4	4
101	energetic	4	3
102	englishman	4	2
103	englishmen	4	4
104	european	4	3
105	handkerchief	4	3
106	kitchen	4	3
107	lancaster	4	3
108	outer	4	3
109	reckon	4	3
110	sheet	4	4
111	tickets	4	3
112	yarmouth	4	2
113	animated	3	3
114	austria	3	2
115	beautifully	3	3
116	cash	3	3
117	constructing	3	3
118	despotism	3	2
119	emigrate	3	3
120	females	3	2
121	hailes	3	1
122	inquired	3	3
123	lunatic	3	3
124	noticeable	3	3
125	pinioned	3	1
126	radicals	3	1
127	rushing	3	3
128	songs	3	2
129	spaces	3	3
130	tearing	3	3
131	using	3	3
132	vaccination	3	2
133	altercation	2	1
134	audacious	2	2
135	betwixt	2	1
136	brown	2	2
137	bulgarians	2	1
138	cholera	2	1
139	statesmanship	1	1
140	valentia	1	1

Some of the resulting scores had to be excluded; this happened in two cases. One case is when the number of co-occurrences multiplied by the span yielded a number greater than the total corpus size. This produces a negative R_2 (using the notation introduced in section 3.3.1); since it is not possible to take the log of a negative number, a negative R_2 causes problems for the computation of LL. This situation arises only for large spans combined with frequent words. The other case is when the number of co-occurrences was greater than the number of occurrences of the collocate. This produces a negative O_{21}; since it is not possible to take the log of a negative number, a negative O_{21} causes problems for the computation of LL. This situation is especially common for smaller collocate frequencies and for larger spans. It arises from the presence of overlapping spans (when two nodes occur within a range smaller than the chosen span): when spans overlap, the same occurrence of the collocate is counted several times (as many times as the number of spans around the node occurrences in which it is captured). In these situations, the statistics cannot be meaningfully computed using the formulas provided in section 3.1.1.

Table 4.8 shows the number of included and excluded node/collocate pairs at each span. 'Above floor' denotes node/collocate pairs which both occur at least 10 times in the sample; 'below floor' denotes node/collocate pairs for which one or both occurs less than 10 times in the sample. It should be noted that the number of excluded cases is always greater below than above the floor, that it increases with the span, and that it is greatest in the uncorrected sample and smallest in the gold sample; this last point highlights another manifestation of the impact of OCR errors on OCR-derived figures.

Table 4.8 Number of statistics calculated per span in the CNNE gold, uncorrected and Overproof-corrected matching samples

		(span)	3	4	5	10	20	50
UNCORRECTED	Included	above floor	5720	7208	8499	13547	19938	29236
UNCORRECTED	Excluded	above floor	11	24	32	112	380	1737
UNCORRECTED	Included	below floor	37561	43573	48004	50351	44017	69722
UNCORRECTED	Excluded	below floor	2237	4421	7110	30791	69563	98892
CORRECTED	Included	above floor	6650	8377	9891	15703	22966	33851
CORRECTED	Excluded	above floor	1	5	8	58	255	1313
CORRECTED	Included	below floor	31040	35483	38500	41018	39460	63372
CORRECTED	Excluded	below floor	1609	3344	5503	21826	46772	65125
GOLD	Included	above floor	7152	9015	10686	17029	25079	37083
GOLD	Excluded	above floor	1	5	8	49	234	1280
GOLD	Included	below floor	28492	32233	34592	37121	36473	58070
GOLD	Excluded	below floor	1573	3192	5287	18525	39078	53961

One limitation of this study is presumably that the small sample size (only around 160,000 words, see Table 4.1) precludes the obtaining of large values for LL (since overall amount of data is one of the determining factors for LL). Nevertheless, the use of nodes of varying frequencies will help to establish the interaction between node/collocate frequency and LL; in addition, as will be seen below, a range of MI and LL values are obtained from the CNNE matching samples, so that despite that limitation, this study still constitutes a useful starting-point for the investigation of the impact of OCR errors on collocation statistics derived from OCR data.

The remainder of section 4.5 is organised as follows. Section 4.5.2 discusses the overall difference between the gold and uncorrected statistics. Section 4.5.3 considers the variation in average differences across word-types. Section 4.5.4 considers whether the ranking of collocations is conserved between the gold and uncorrected data. Finally, section 4.5.5 discusses the reliability of the uncorrected statistics in terms of false positives (uncorrected statistics which appear to be significant when the corresponding gold statistic is not) and false negatives (uncorrected statistics which appear not to be significant when the corresponding gold statistic *does* appear to be significant) (see also section 4.3.2).

Throughout the discussion in this chapter (and the next), the difference between gold and uncorrected statistics will be described in terms of 'points of difference', which is simply the default unit for that statistic. For MI, since MI is measured on a log base 2 scale, a difference of 1 point between A and B means that the effect associated with A is twice as large as that associated with B; a difference of 4 points would mean that the effect associated with A would be 16 times larger than that associated with B. LL points of difference are less intuitively interpretable, but since LL involves a natural log scale, the amount of confidence represented by one LL point of difference will become greater when dealing with differences between larger values; so a difference of 1 between LL values of 2 and 1 represents a smaller difference in confidence than a difference between LL values of 22 and 21.

In addition, differences will always be expressed as the difference between the gold statistic and the (uncorrected or corrected) OCR statistic (in that order), which implies that *positive* difference values will be statistics which are *smaller* in the OCR data than in the gold data (i.e. they are *under*-estimated in the OCR data). Conversely, *negative* difference values will be statistics which are *over*-estimated in the OCR data.

Given that *points of difference* is not a usual measure for collocation statistics, it may be useful to mention some common benchmarks to help gauge the importance of a given points of difference figure. I am *not* advocating that node/collocate pairs attracting MI or LL statistics above a certain value should be considered *important* (see also section 3.3), but bearing these values in mind will help evaluate the results of comparing the statistics derived from OCR and gold standard data.

In the corpus linguistic literature, 3 is sometimes suggested as a cut-off point for MI scores that are of interest (Hunston 2002: 71-72) so may be used to help gauge what may constitute a ‘large’ number of points of difference for MI statistics. When such a cut-off point is used, a difference of 3 points for MI scores would mean that a collocation-pair with a gold MI score of just above 0 would attract an MI score of just above 3 and be interpreted as meaningful (false positive); a difference of -3 points would mean a collocation-pair with a MI gold score of just below 6 would attract an MI score of just below 3 and be considered as not meaningful (false negative). For LL, at $p < 0.001$, LL normally needs to be equal or greater than 10.83 (Oakes 1998: 266)[11], a figure which can be used to help gauge what may constitute a 'large' number of points of difference for LL statistics. A difference of +/-3 points for LL would mean that a collocation-pair with a gold statistic of just above 7.83 would end up as a false negative in the OCR data and that a collocation-pair with a gold statistic of just below 13.83 would end up as a false positive in the OCR data.

A recommendation made by Hardie (forthcoming) is to use an effect size statistic such as MI as a ranking statistic, in combination with an LL cut-off point. The impact of OCR errors on the statistics under this set-up will be examined in this chapter and the next. The recommendation will be tested using an LL cut-off point of 10.83. Another possible step which will be examined is that of using a frequency floor. Section 4.4.1 showed that using a frequency floor of 10 would eliminate most – though not all – OCR errors. In this context, and since corpus linguists are normally more interested in the *more* frequent patterns, it makes sense to consider whether the impact of OCR errors will be the same at different points in the frequency range, and, in particular, whether the trend is substantially different for only those words which occur at least 10 times in the OCR corpus. *Node/collocate pairs above the frequency floor* will hence refer, below, to node/collocate pairs for which *both* the node and collocate occur at least 10 times in the OCR corpus.

4.5.2 OVERALL VARIATION BETWEEN GOLD AND UNCORRECTED STATISTICS

What is the magnitude of the difference between gold and uncorrected statistics? And how likely are we to encounter an uncorrected statistic which is very different from its corresponding gold statistic? The histograms in this chapter address these questions: they plot the probability that each uncorrected statistic is within a certain range of distance (or *bin*) from

[11] Note that the table provided in Oakes (1998) is for the chi-squared distribution, but this is the same distribution as the LL distribution, see Cressie and Read (1984).

the corresponding gold statistic. Except for Figure 4.9, the ticks along the x axes indicate the centres of the bins. So, for instance, in Figure 4.7, the ticks being placed at 0.8 axis scale points one from another indicates that the bin size is 0.8; for instance, the height of the bars placed above the '0' tick captures the probability that a given MI uncorrected statistic is situated within -0.4 to 0.4 points of difference from its gold counterpart. Some of the histograms show the full extent of the data; others are *trimmed*: only a subset of the data is shown, for a limited range in the x axis, and with a smaller bin-width, to reveal more detail. For the histograms showing the full extent of LL differences in this chapter and the next, it will not be possible to show the bin centres as this would result in so many ticks as to render the x axis unreadable; these graphs will hence have to be understood impressionistically. The reader is invited to refer to the trimmed LL histograms provided for more detailed information about the spread of LL differences. Note also the logarithmic scale on the y axis of the histograms.

Figure 4.7 (along with Figure 4.8, which represents the same data) shows that most (between 50% and 70% of) uncorrected MI statistics are no more than 0.4 points of difference greater or smaller than their gold counterpart for every span considered. Between 20 and 40% of further uncorrected MI statistics are between 0.4 and 1.2 points of difference *greater* (i.e. the difference is negative) than their gold counterparts. Around 7% of uncorrected MI statistics are within 0.4 and 1.2 points of difference *smaller* than their gold counterparts. The remaining statistics are up to 5.2 points of difference smaller or greater than their gold counterparts; however, less than 2% of statistics exhibit such extreme differences, with the proportion of statistics reaching each more extreme value decreasing exponentially.

As was observed in section 3.3.2, if the OCR errors were distributed homogenously, we would expect corresponding pairs of MI statistics to be *identical* in the OCR and gold samples. Finding that most OCR statistics have close to 0 points of difference from their gold counterparts is hence expected. It is also reassuring, since it suggests that, for the most part, the impact of OCR errors on MI statistics is negligible. On the other hand, the fact that *not all* OCR statistics have 0 points of difference in fact provides evidence that the OCR errors are *not* homogenously distributed.

The effect of span is difficult to examine on the histograms, but is clearer on the boxplots. Figure 4.8 shows that, at all spans considered, most MI statistics are *over*-estimates of their gold counterparts, but that 50% of the statistics remain within 0.5 points of difference of their gold counterparts. The figure further shows an effect of span, with wider spans attracting a greater spread of differences, though the effect is not of a large magnitude.

Figure 4.7 Probability that any given uncorrected MI statistic will be situated at a given distance of the corresponding gold statistic (full extent)

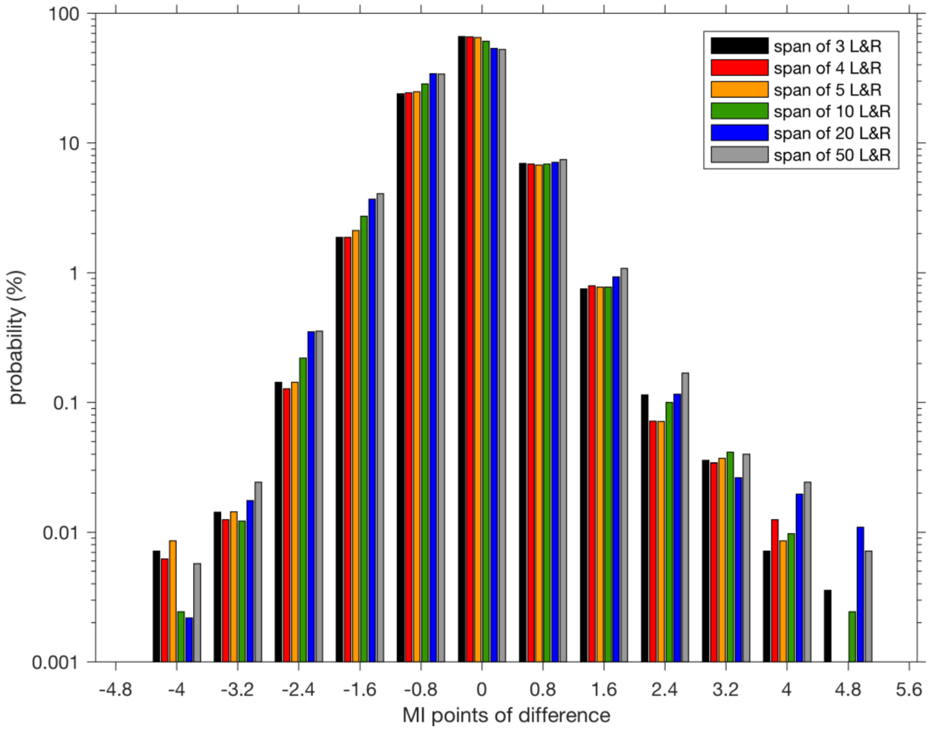
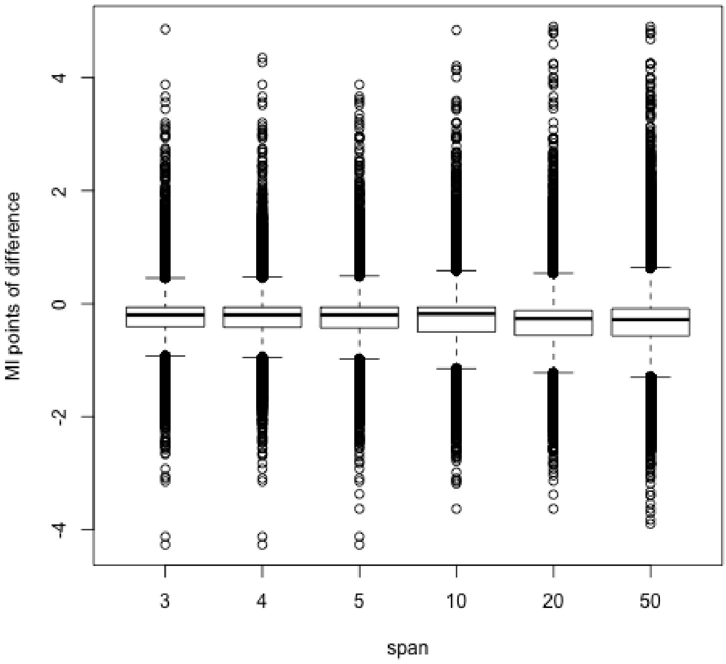
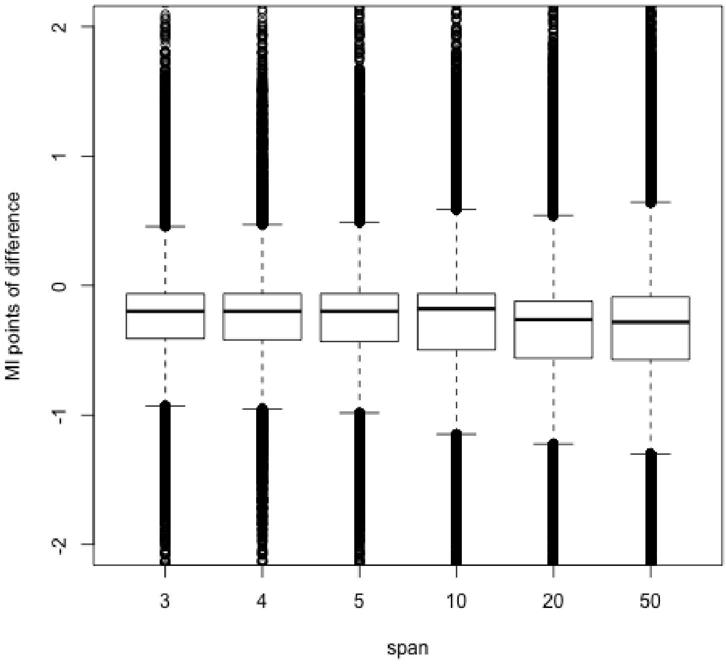


Figure 4.8 Spread of differences between uncorrected MI statistics and gold MI statistics (full extent followed by close-up)





For LL, Figure 4.9 (as well as its trimmed versions, Figure 4.10 and Figure 4.11) shows that, as for MI, the probability of encountering a statistic at a given distance from its gold counterpart becomes exponentially smaller as the distance increases. However, the spread of LL values is a lot broader. Only 30-40% are situated within 0.4 points of difference of their gold counterpart. Around 20%-30% of statistics are situated within 0.4 to 1.2 points of difference *smaller* than their gold counterpart at every span, and around 8% at every span are situated within 0.4 to 1.2 points of difference *greater* than their gold counterpart. Although over 90% of statistics are situated within 5 points of difference than their counterparts, there are still 2-3% of statistics situated within 5 to 15 points of difference *smaller*, and around 1% of statistics situated within 5 to 15 points of difference *greater* than their gold counterpart. Beyond this, the range of LL differences extends to a hundred points of difference *greater* and a thousand points of difference *smaller* than their gold counterparts. Figure 4.12 shows that there is also a weak effect of span on the LL, with more extreme values and a greater spread occurring with greater spans.

In contrast to MI, OCR LL statistics would be expected to be *smaller* than their gold counterparts if the OCR errors were distributed homogenously (see section 3.3.2). Unfortunately, it is difficult to assess *how much smaller* they should be. Nevertheless, finding that LL statistics are smaller in the OCR sample is neither surprising nor problematic. In contrast, finding OCR LL statistics which are *greater* than their gold counterparts is problematic, and provides evidence that OCR errors are not distributed homogeneously.

Figure 4.9 Probability that any given uncorrected LL statistic will be situated at a given distance of the corresponding gold statistic (full extent)

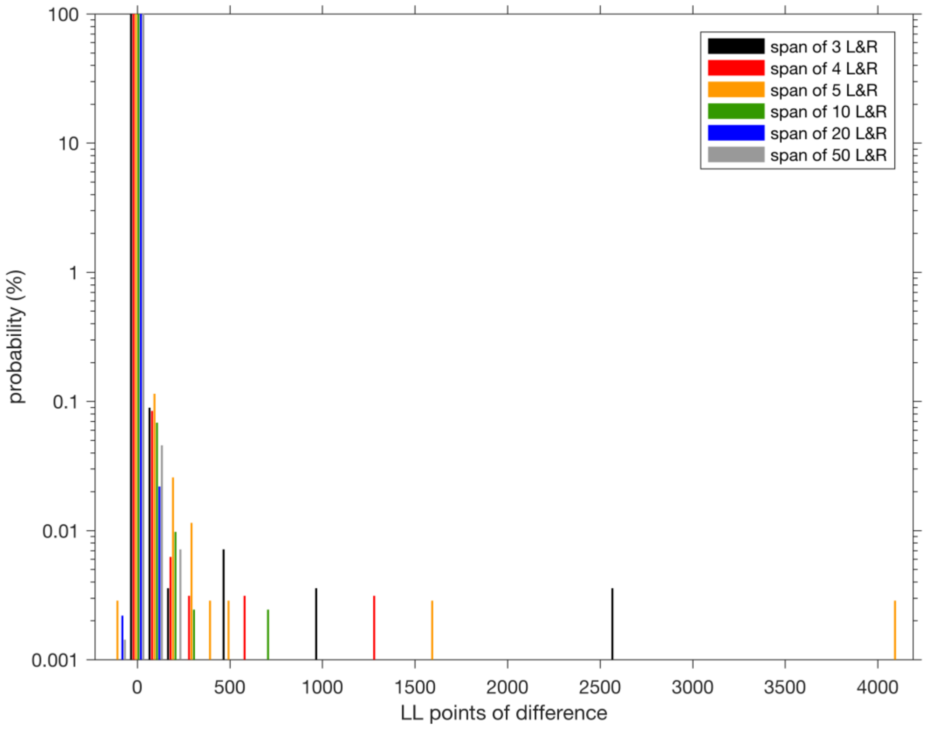


Figure 4.10 Probability that any given uncorrected LL statistic will be situated at a given distance of the corresponding gold statistic (trimmed to -100, 100)

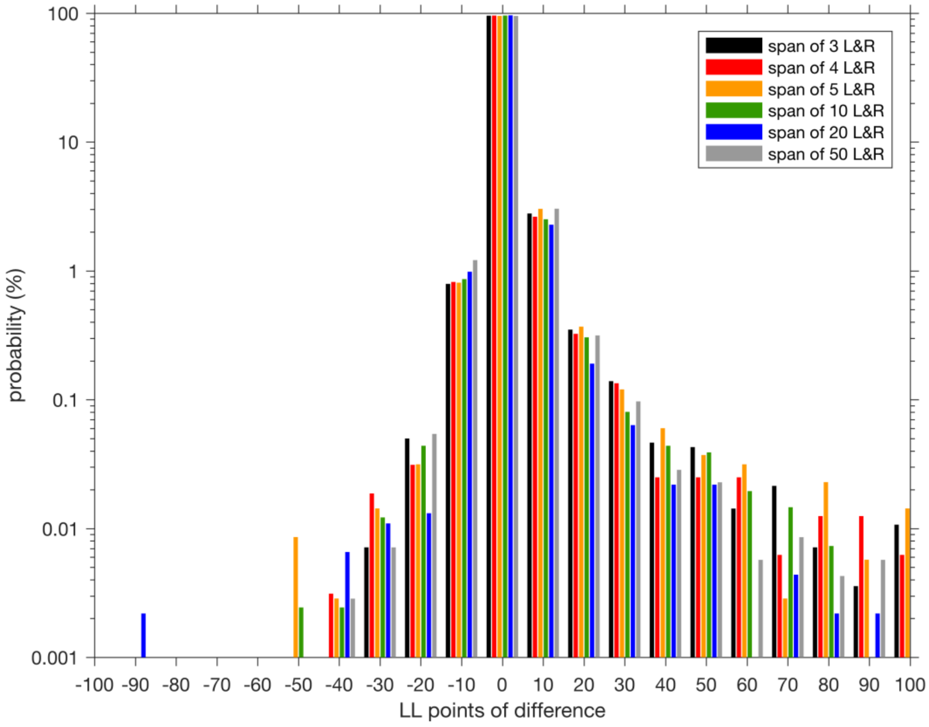


Figure 4.11 Probability that any given uncorrected statistic will be situated at a given distance of the corresponding gold statistic (trimmed to -5, 5)

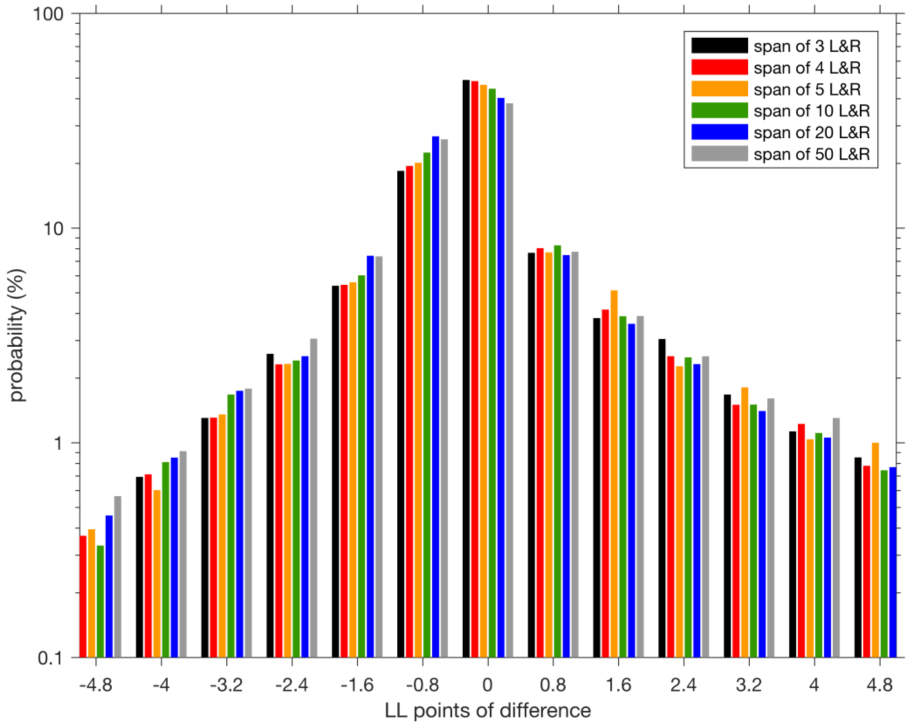
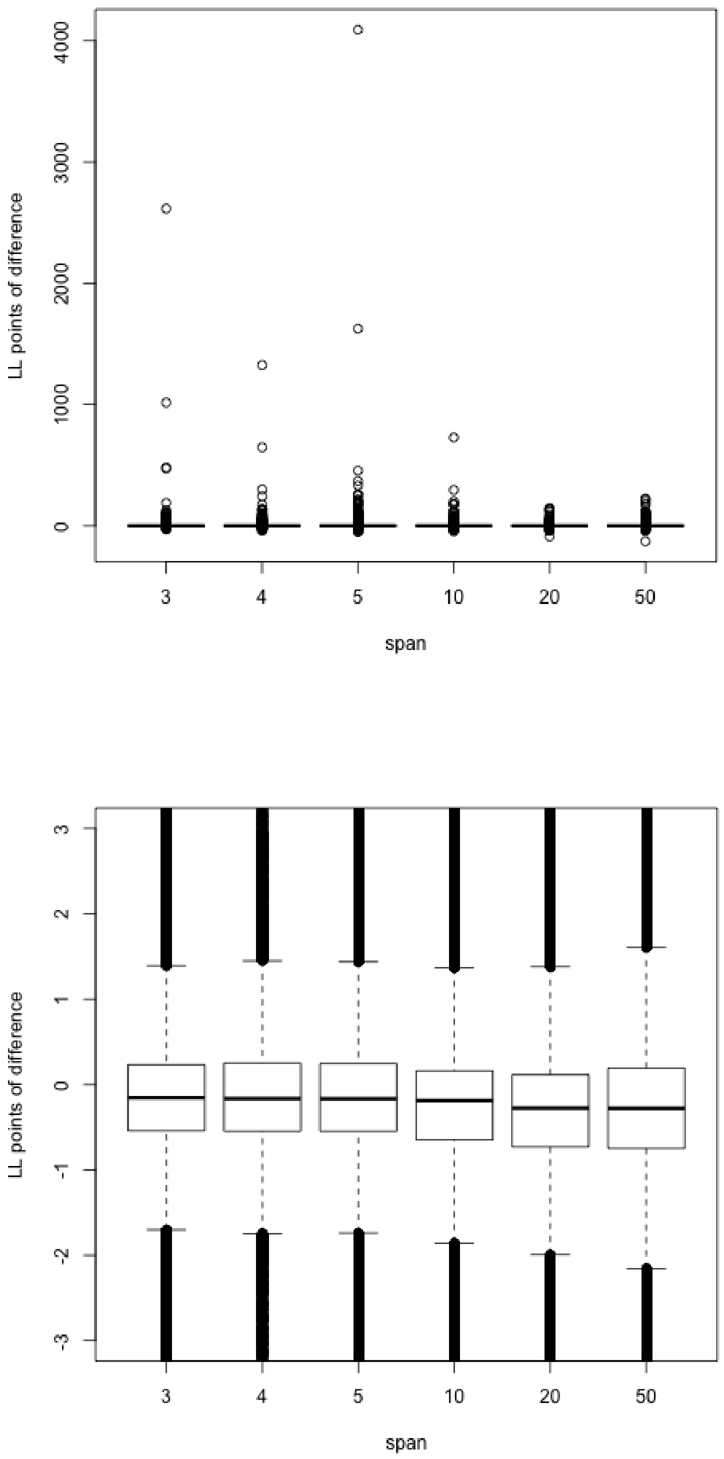


Figure 4.12 Spread of differences between uncorrected LL statistics and gold LL statistics (full extent followed by close-up)



For MI values in combination with an LL threshold of 10.83 (Figure 4.13 and Figure 4.14), the distribution changes shape compared to that without an LL threshold: more values are *over*-estimated (i.e. negative differences), and fewer values are *under*-estimated (i.e. positive differences), a situation which is potentially problematic, since this may lead to an observer making unwarranted conclusions.

Figure 4.13 Probability that any given uncorrected MI statistic will be situated at a given distance of the corresponding gold statistic, when considering only MI statistics for which the corresponding LL statistic is at least 10.83 (full extent)

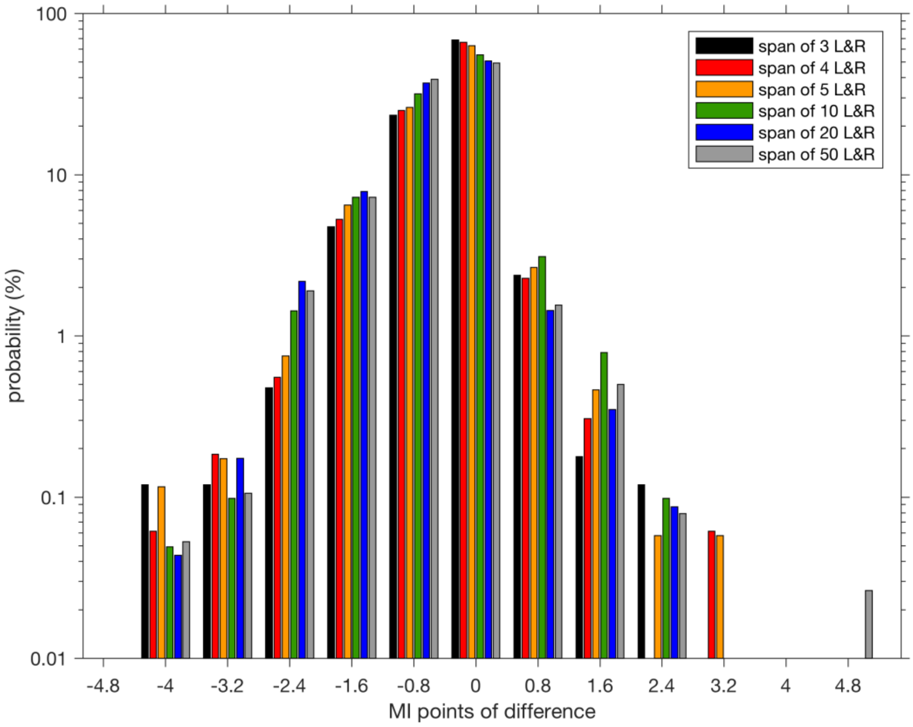
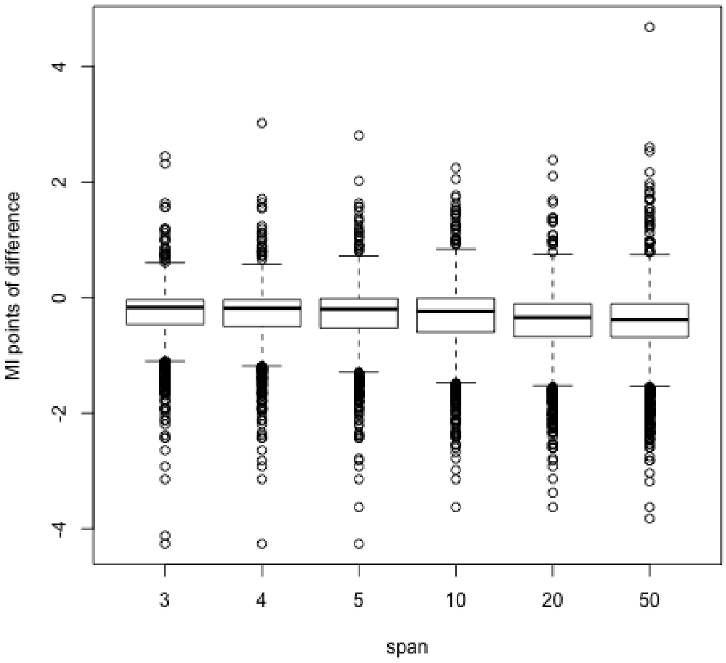
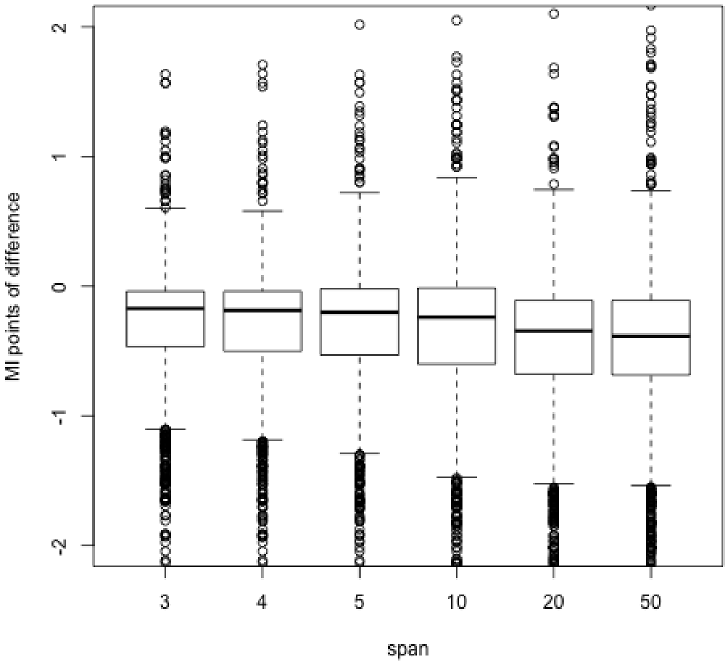


Figure 4.14 Spread of differences between uncorrected and gold MI statistics, with an LL threshold of at least 10.83 (full extent followed by close-up)





Do the statistics for low-frequency node/collocate pairs behave differently from other statistics? Does using a frequency floor, which, as seen in section 4.4.1, removes most erroneous types, much improve the reliability of the statistics? Excluding node/collocate pairs for which the node or collocate occurs less than 10 times in the uncorrected corpus (Figure 4.15 and Figure 4.16) shows a different spread of results: fewer extreme *over*-estimates are obtained, but MI statistics are less likely to be very close to 0 compared to Figure 4.7 and Figure 4.8, and more likely to be *over*-estimates. Likewise, for LL, as shown on Figure 4.17, Figure 4.18, Figure 4.19 and Figure 4.20, the spread of values narrows, but LL statistics are also less likely to be within the 0.4 points of difference and a little more likely to be *over*-estimates compared to the overall trend shown in Figure 4.9, Figure 4.10, Figure 4.11 and Figure 4.12. Finally, for MI with an LL threshold (see Figure 4.21 and Figure 4.22), the effect of the frequency floor is particularly stark, with many of the more extreme values on either side disappearing. This last result is in fact very good, with most differences very close to 0 and the spread of differences fairly narrow, although the tendency is still for most MI statistics to be *over*-estimates relative to their gold counterpart. Nevertheless, this result suggests that using the statistics in this set-up – MI with a LL threshold and a frequency floor – may be considered reliable in OCR data. The overall results also show that there is indeed an interaction between frequency of node/collocate and OCR errors, and that using a frequency floor *does not* solve all problems; in fact, *over*-estimates are more likely for node/collocate pairs above the frequency floor.

Figure 4.15 Probability that any given uncorrected MI statistic will be situated at a given distance of the corresponding gold statistic, for node/collocate pairs above the frequency floor (full extent)

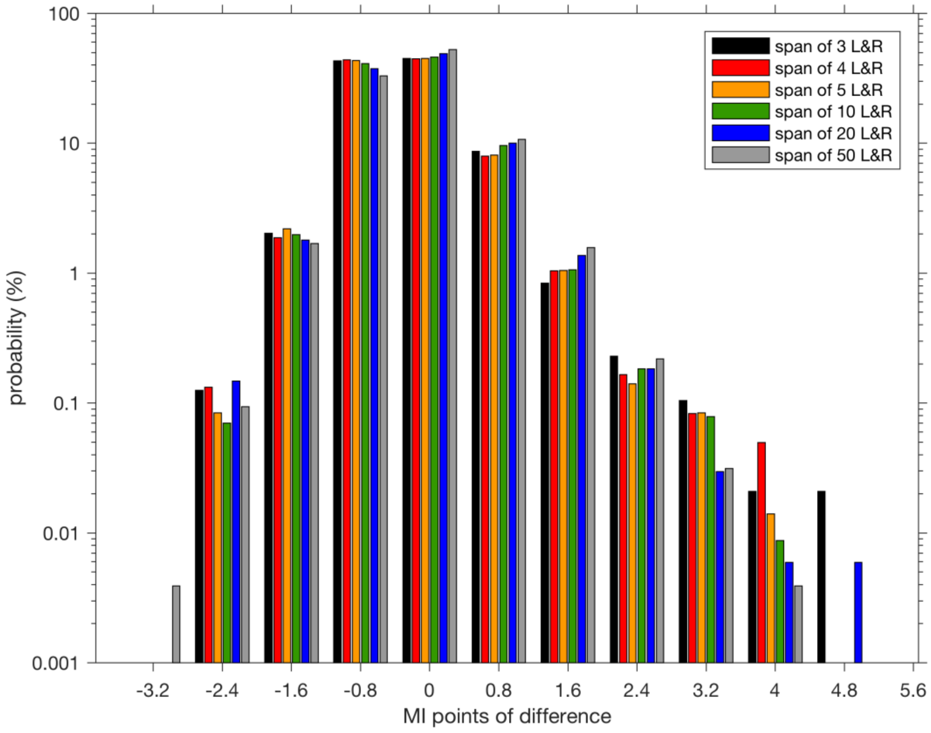
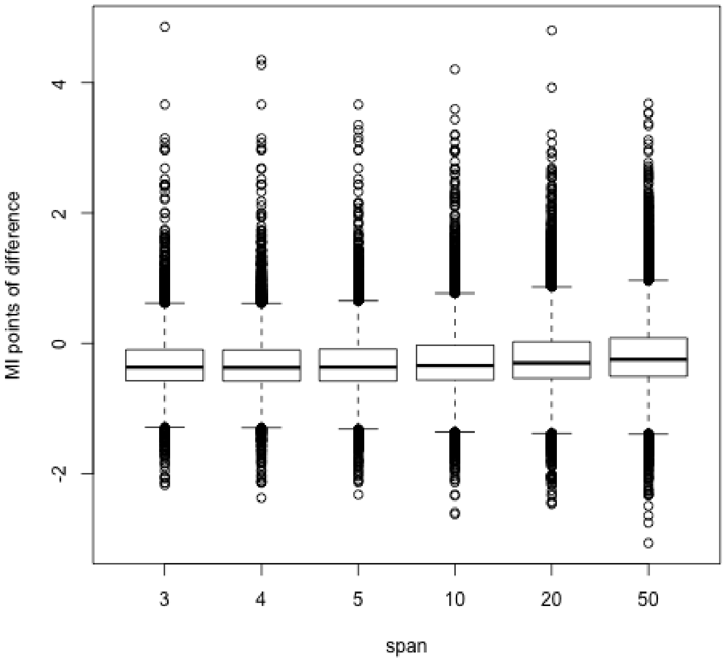


Figure 4.16 Spread of differences between uncorrected and gold MI statistics, for node/collocate pairs above the frequency floor (full extent followed by close-up)



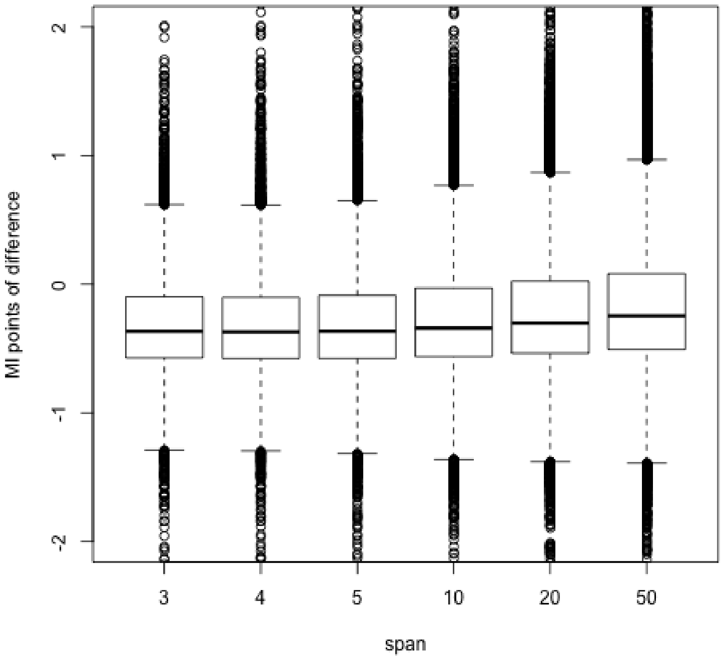


Figure 4.17 Probability that any given uncorrected LL statistic will be situated at a given distance of the corresponding gold statistic, for node/collocate pairs above the frequency floor (full extent)

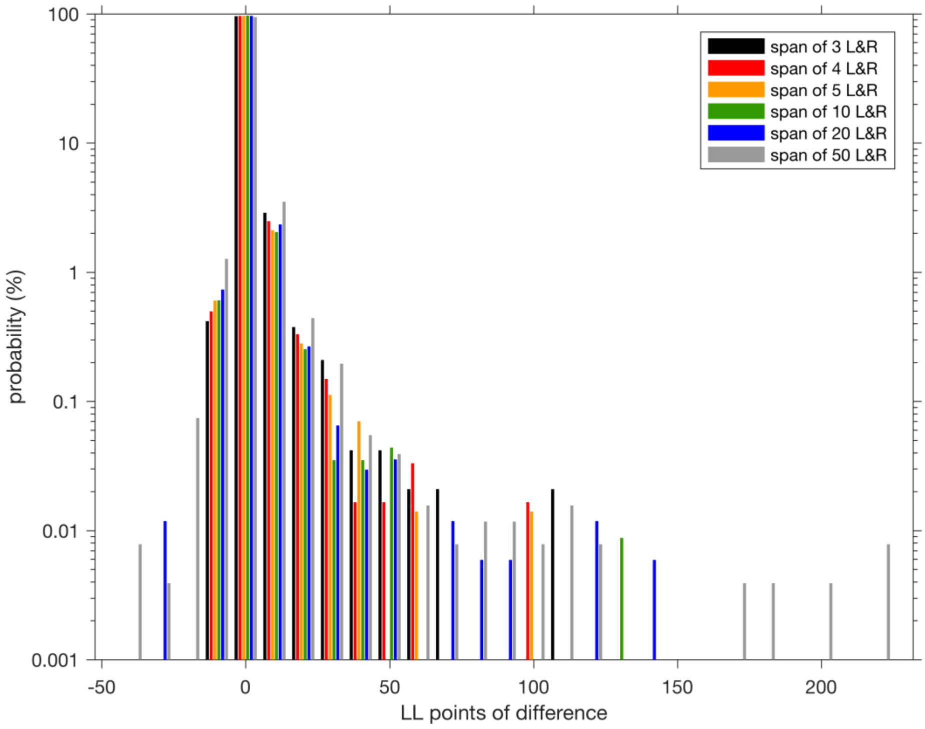


Figure 4.18 Probability that any given uncorrected LL statistic will be situated at a given distance of the corresponding gold statistic, for node/collocate pairs above the frequency floor (trimmed to -100, 100)

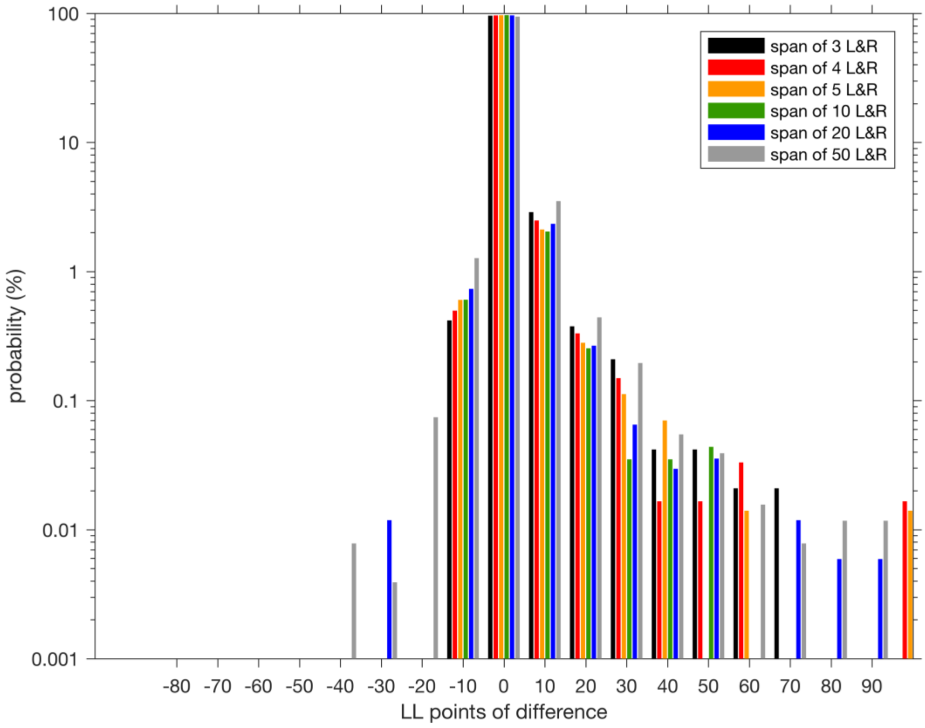


Figure 4.19 Probability that any given uncorrected LL statistic will be situated at a given distance of the corresponding gold statistic, for node/collocate pairs above the frequency floor (trimmed to -5, 5)

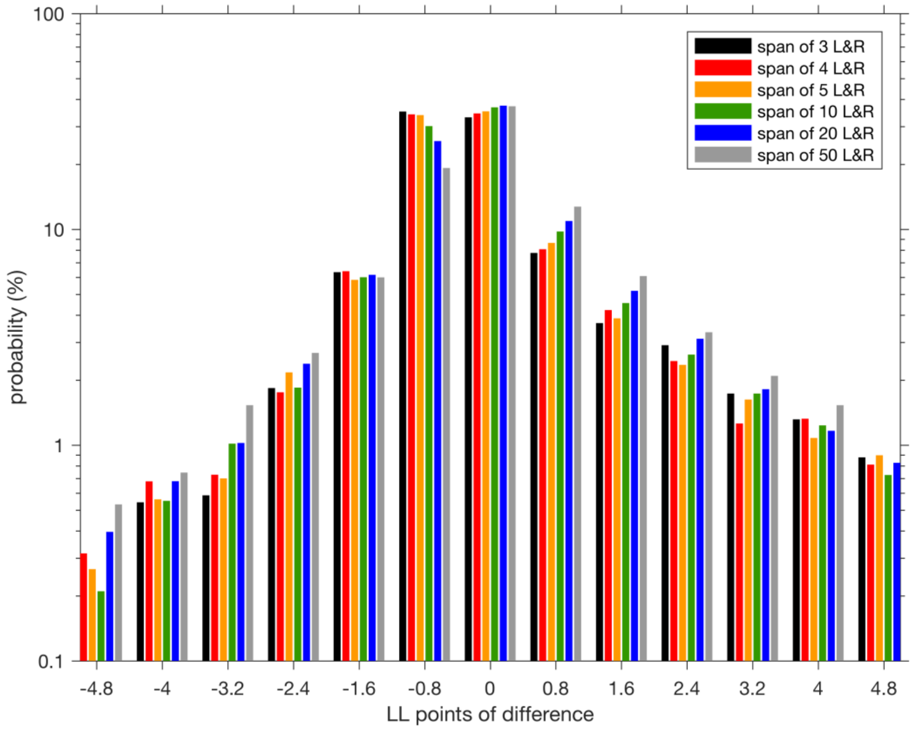
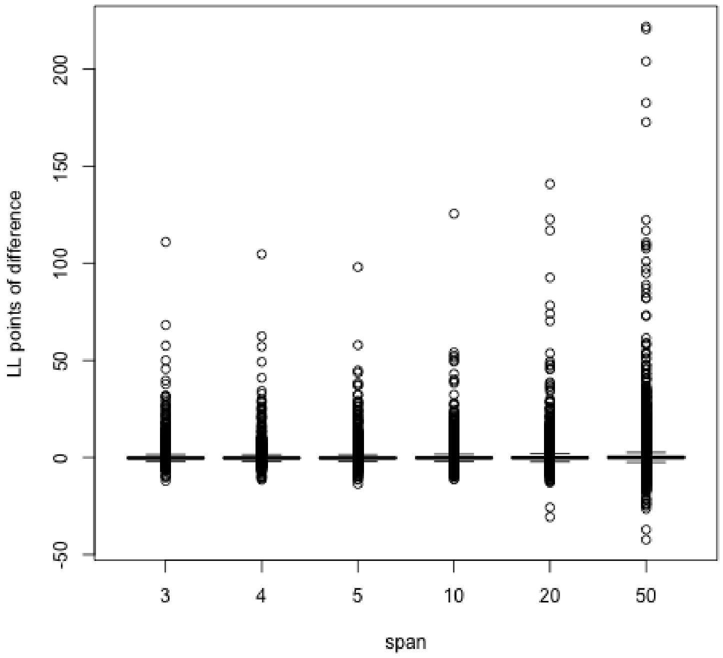


Figure 4.20 Spread of differences between uncorrected and gold LL statistics, for node/collocate pairs above the frequency floor (full extent followed by close-up)



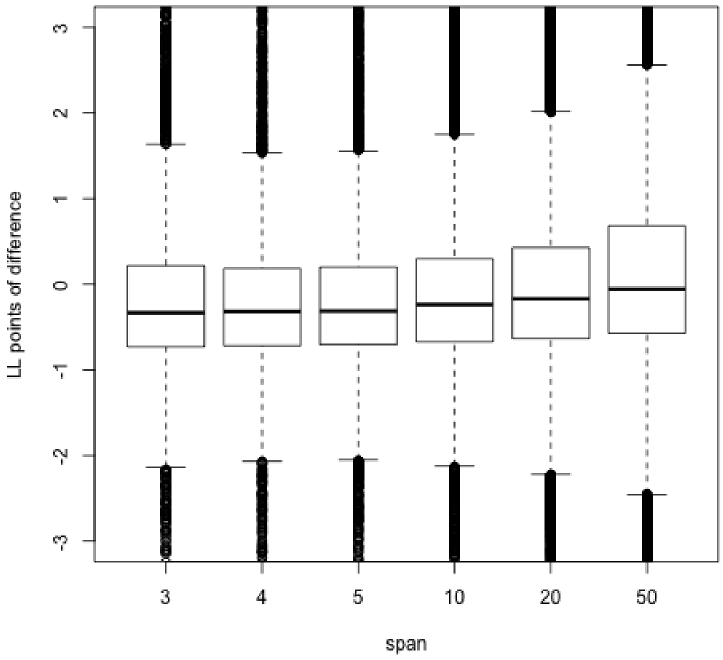


Figure 4.21 Probability that any given uncorrected MI statistic will be situated at a given distance of the corresponding gold statistic, for node/collocate pairs above the frequency floor, with an LL threshold of at least 10.83 (full extent)

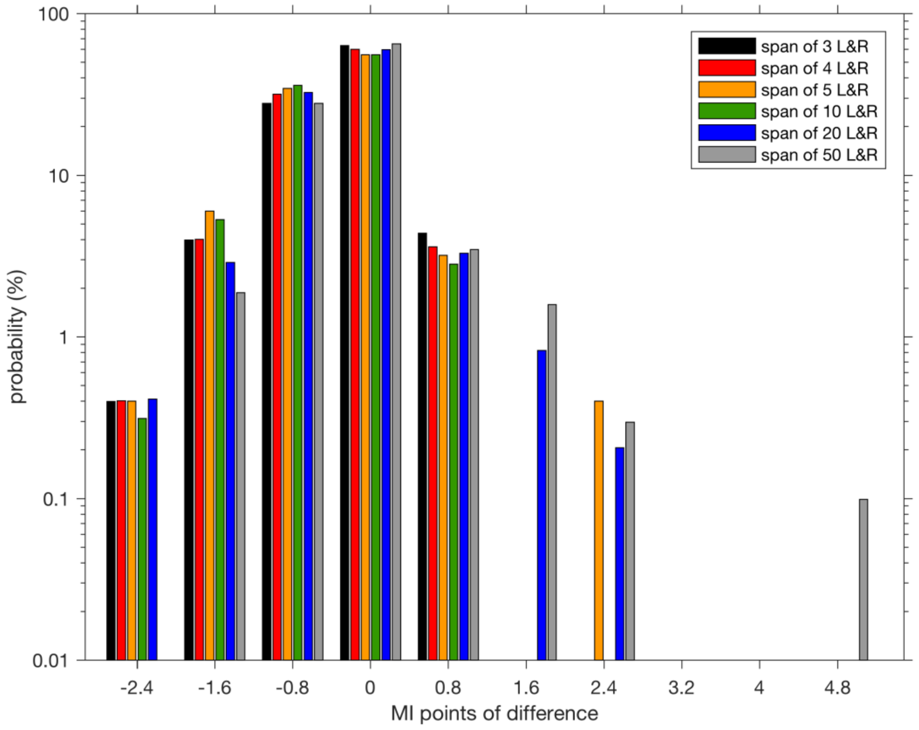
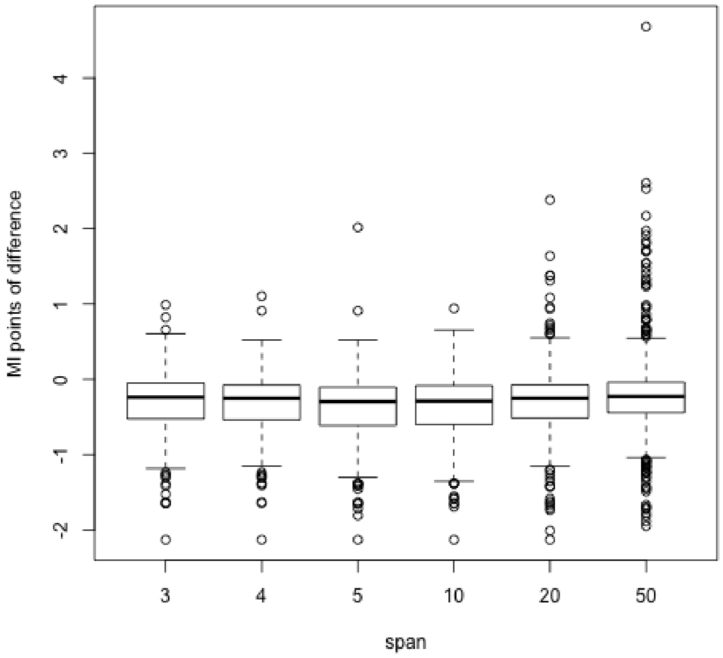
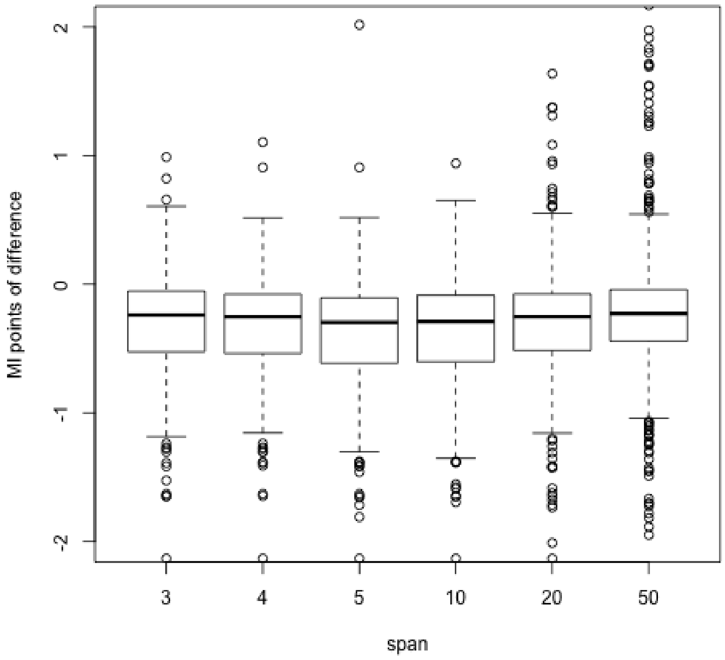


Figure 4.22 Spread of differences between uncorrected and gold MI statistics, for node/collocate pairs above the frequency floor, with an LL threshold of at least 10.83 (full extent followed by close-up)





4.5.3 Variation in average differences across word-types

In the previous section, very different trends were observed when considering all node/collocate pairs as opposed to only those above the frequency floor. This suggests that token frequencies may be a factor in the impact of OCR errors on collocation statistics. The figures in this section show the relationship between the frequency of the node in the gold sample and the *average* difference between the gold and uncorrected statistic. Each graph plots the average difference in statistics across the node/collocate pairs for a single node, with each span shown in a different colour. Since an average of a set of differences containing both positive and negative differences would not be very informative, I have plotted the average of the absolute values of the differences in half the graphs, while in the other half I have plotted the average difference calculated separately for positive and negative values.

Figure 4.23 and Figure 4.24 show that there is indeed an effect of frequency, with, broadly speaking, smaller frequencies attracting larger average differences. However, the frequency range which attracts the broadest average differences is the range between 10 and 100. This is an unfortunate but important result, because words in this range (in a corpus of around the size considered here) will often be of interest to corpus linguists, since here is where many of the content words will be found. This result also provides additional evidence that the impact of OCR errors will vary across word-types. Similar observations apply to LL (see Figures 4.23-4.25), with the most extreme average differences occurring within the 10-100 frequency bracket, although the overall effect of frequency is less strong than for MI, and slightly larger average differences also occur on the high end of the frequency range.

Figure 4.23 Average distance between gold and uncorrected MI statistic (absolute values), by frequency of node in gold corpus

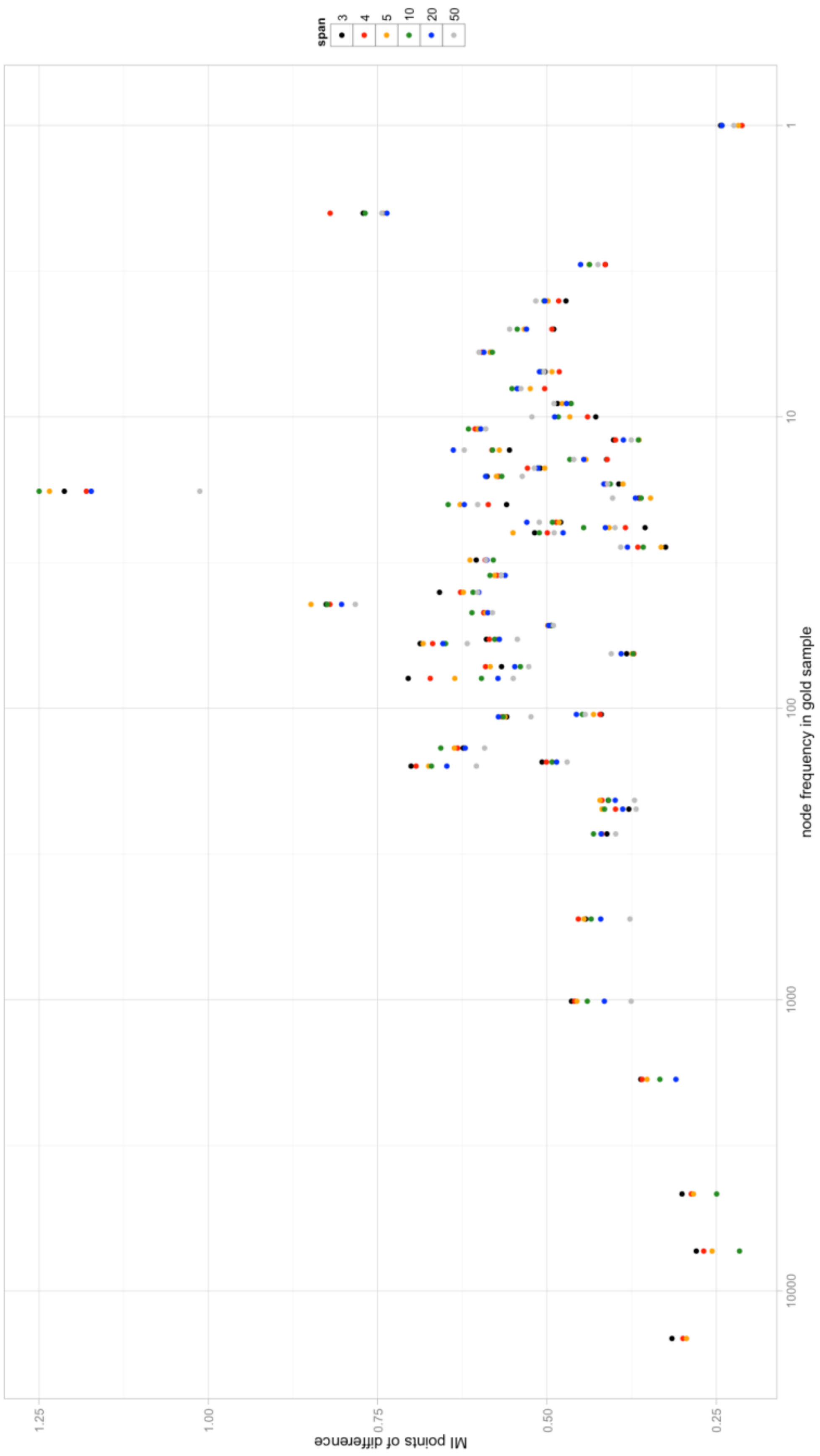


Figure 4.24 Average distance between gold and uncorrected MI statistic (for positive values only), by frequency of node in gold corpus

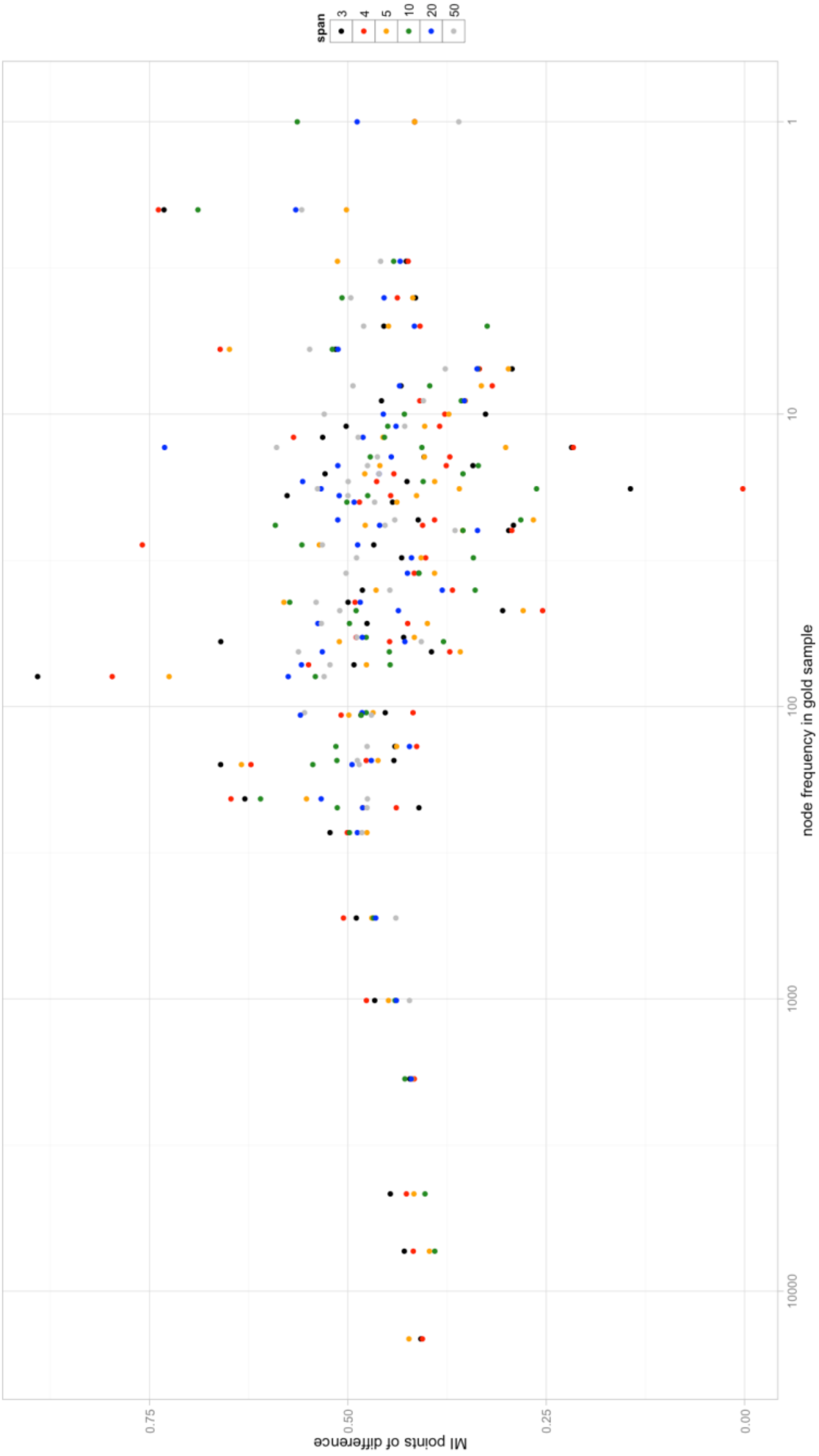


Figure 4.25 Average distance between gold and uncorrected MI statistic (for negative values only), by frequency of node in gold corpus

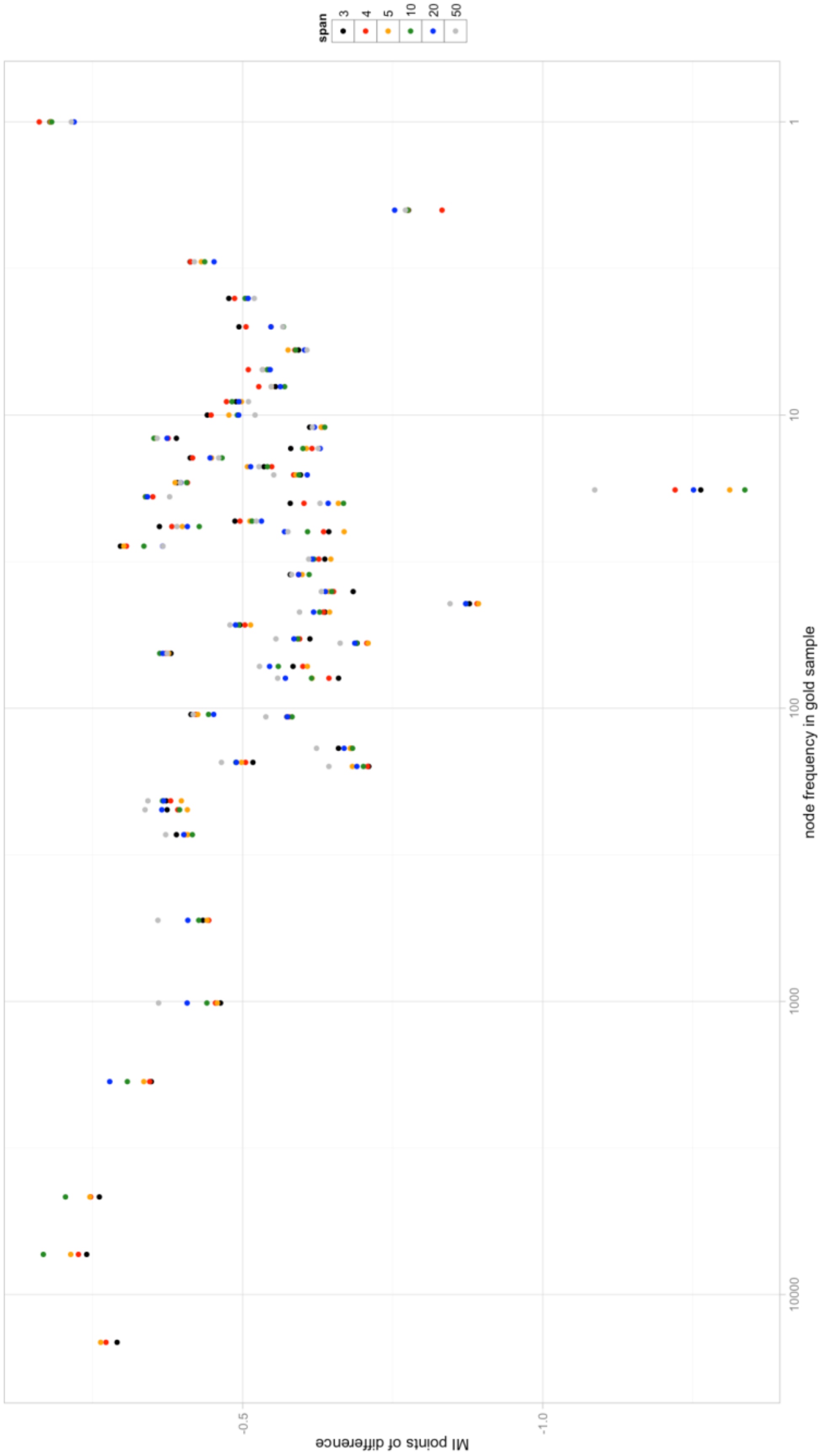


Figure 4.26 Average distance between gold and uncorrected LL statistic (absolute values), by frequency of node in gold corpus

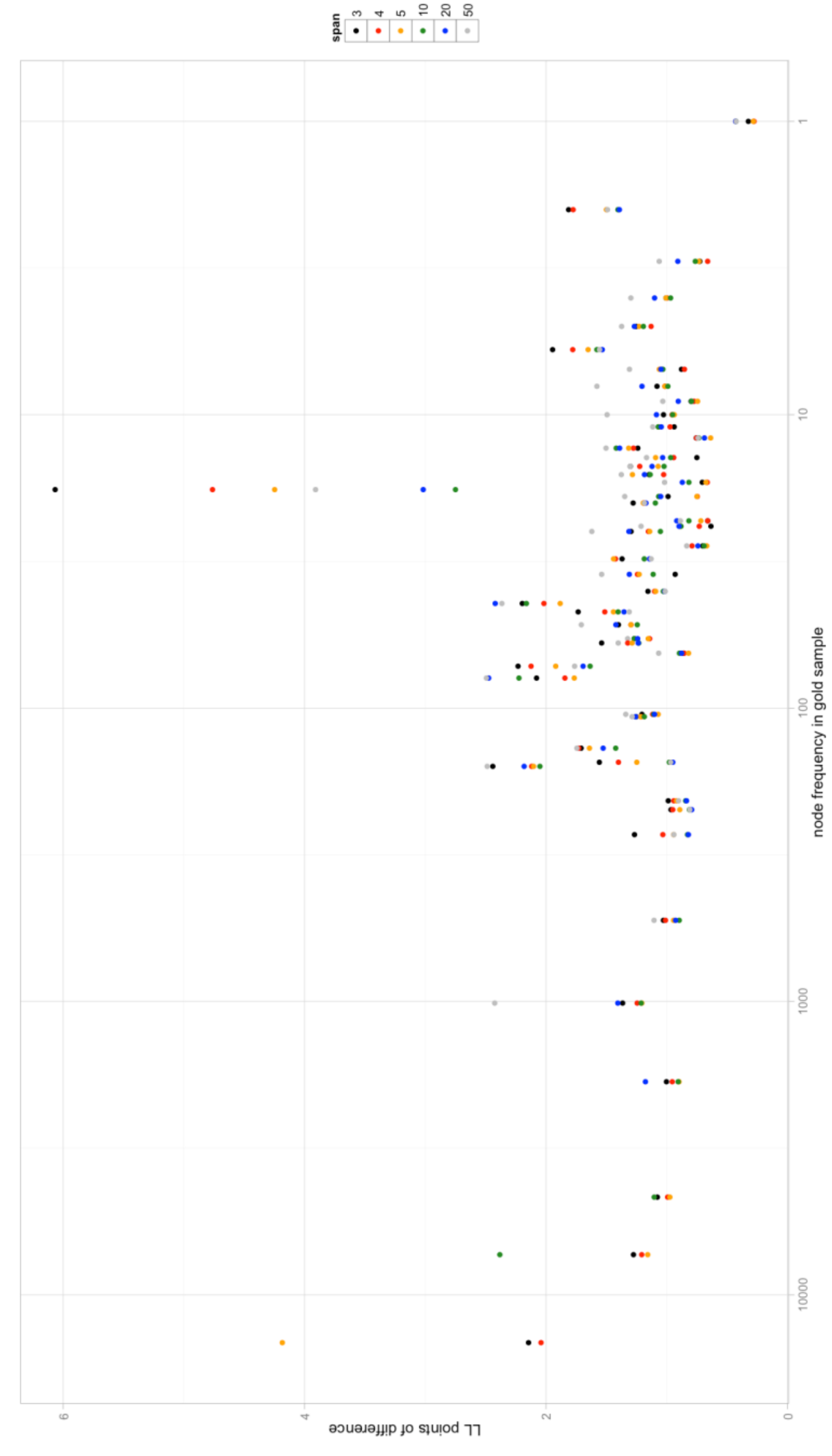


Figure 4.27 Average distance between gold and uncorrected LL statistic (for positive values only), by frequency of node in gold corpus

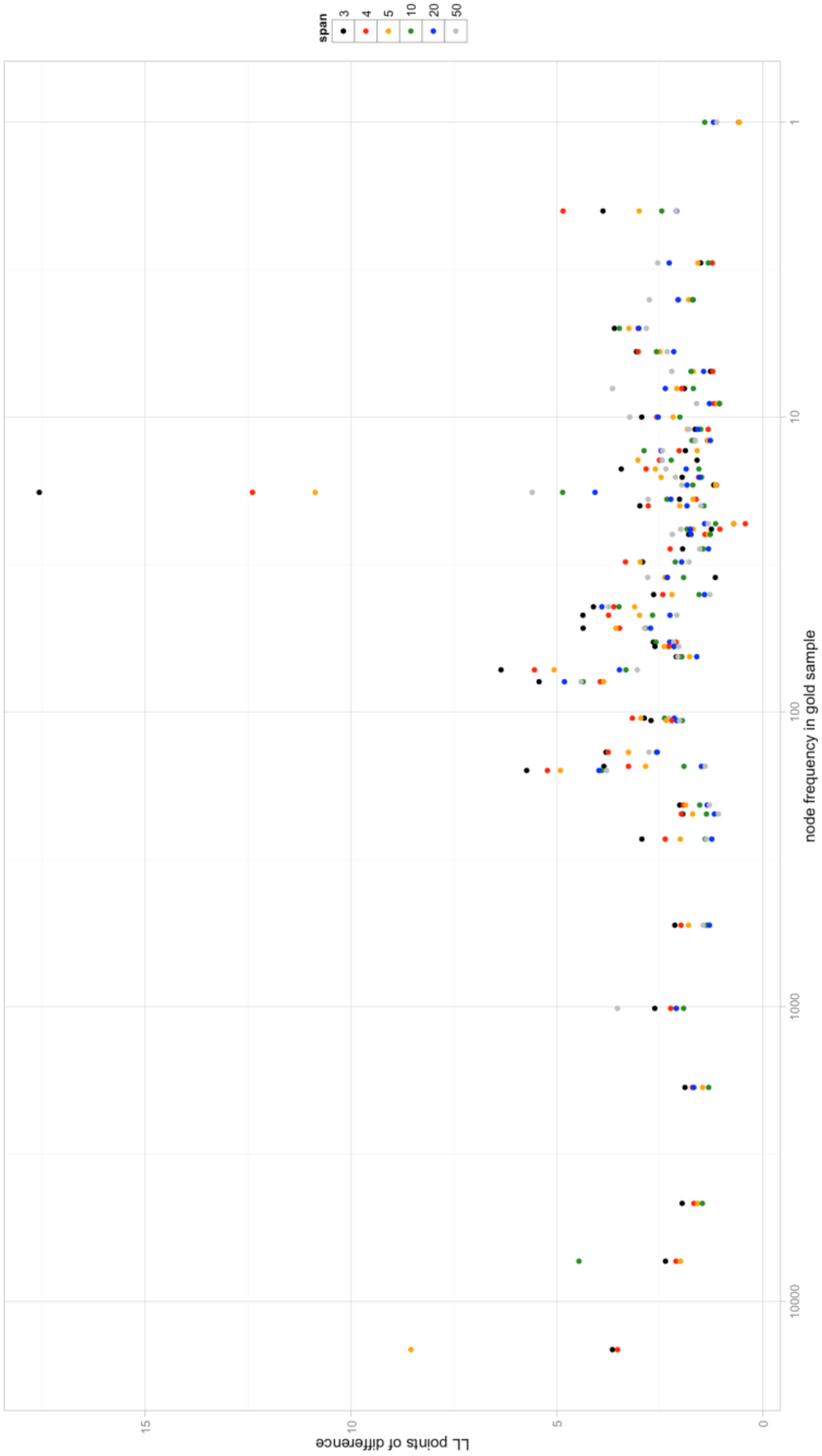
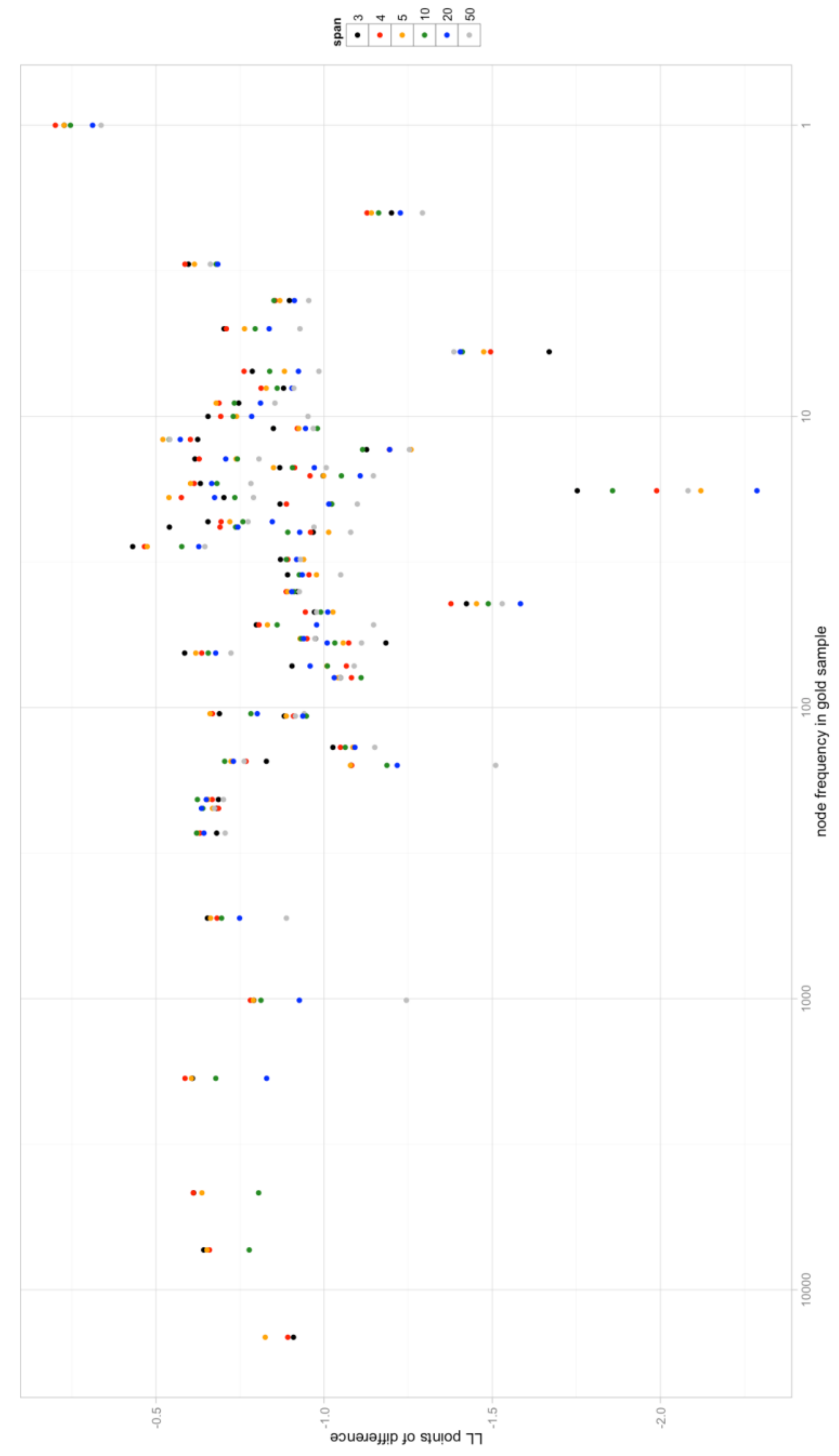


Figure 4.28 Average distance between gold and uncorrected LL statistic (for negative values only), by frequency of node in gold corpus



4.5.4 CONSERVATION OF RANKING

How distant the uncorrected statistics are from their gold counterparts is not the only concern. As we have seen, average distances can vary starkly across word-types, and the spread of differences exhibited by the statistics can be wide indeed, especially for LL. What we also want to know, then, is whether the *ranking* of the statistics is conserved. What I mean by ranking here corresponds to the order in which the statistics associated with node-collocate pairs appear when sorted by ascending or descending value. Hence, I am asking whether, if statistic A is greater than statistic B in the uncorrected sample, we can be confident that this is also the direction of the difference between statistics A and B in the gold sample? This question is especially important for MI, which *should* ideally be used as a ranking statistic.

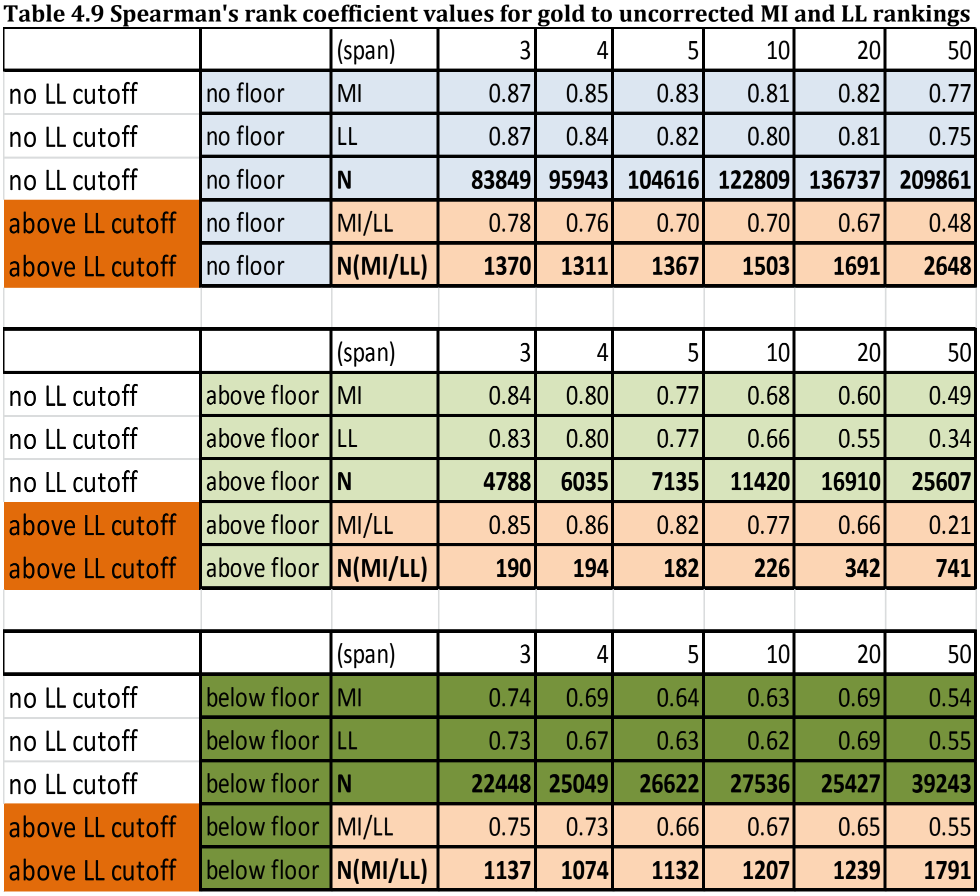
Table 4.9 shows a measure of the similarity of the rankings of MI and LL statistics in the uncorrected and gold CNNE matching corpus. The measure used is *Spearman's rank coefficient*, which is useful in this case because it does not require that the data follow a normal distribution (a special kind of distribution which is not exhibited by my data). Spearman's coefficient varies between -1 and 1, with 0 indicating the complete independence of the two variables, 1 indicating that the two variables vary in perfect proportion and in the same direction, and -1 indicating that the two variables vary in perfect proportion but in the opposite direction (one rises as the other falls). Often, Spearman's rank coefficient is used with the null hypothesis that there is no relationship between the variables considered. In this case, however, the two rankings should be identical (i.e. Spearman's rank coefficient should be 1) *if there is no effect of OCR on the statistics*. Any deviation from 1, then, indicates an effect of OCR on the statistics. In Table 4.9, 'floor' refers to the frequency floor of at least 10 occurrences for the node *and* collocate in the uncorrected corpus[12]. 'LL cutoff' refers to the LL threshold of at least 10.83. 'MI/LL' refers to rankings *by MI values* for only those pairs of node/collocate which attract a LL

[12] Note that cases where the node and collocate are situated on *different* sides of the frequency floor will not be captured in either the 'above floor' or the 'below floor' figures; this is why the number of pairings above and below the floor *do not* add up to the total number of pairings without a frequency floor.

value above the LL threshold. 'N' refers to the number of pairings considered (i.e. the number of statistics ranked in each sample).

Table 4.9 shows a clear interaction of span (with OCR errors) for both MI and LL, with the uncorrected rankings becoming more and more distant from the gold rankings as the span increases. It is also clear that there is also an interaction of the frequency of the node and collocate for both MI and LL statistics, with the rank coefficients above the floor being better than those below the floor for small spans (though not for the largest spans).

Table 4.9 Spearman's rank coefficient values for gold to uncorrected MI and LL rankings



		(span)	3	4	5	10	20	50
no LL cutoff	no floor	MI	0.87	0.85	0.83	0.81	0.82	0.77
no LL cutoff	no floor	LL	0.87	0.84	0.82	0.80	0.81	0.75
no LL cutoff	no floor	**N**	**83849**	**95943**	**104616**	**122809**	**136737**	**209861**
above LL cutoff	no floor	MI/LL	0.78	0.76	0.70	0.70	0.67	0.48
above LL cutoff	no floor	**N(MI/LL)**	**1370**	**1311**	**1367**	**1503**	**1691**	**2648**

		(span)	3	4	5	10	20	50
no LL cutoff	above floor	MI	0.84	0.80	0.77	0.68	0.60	0.49
no LL cutoff	above floor	LL	0.83	0.80	0.77	0.66	0.55	0.34
no LL cutoff	above floor	**N**	**4788**	**6035**	**7135**	**11420**	**16910**	**25607**
above LL cutoff	above floor	MI/LL	0.85	0.86	0.82	0.77	0.66	0.21
above LL cutoff	above floor	**N(MI/LL)**	**190**	**194**	**182**	**226**	**342**	**741**

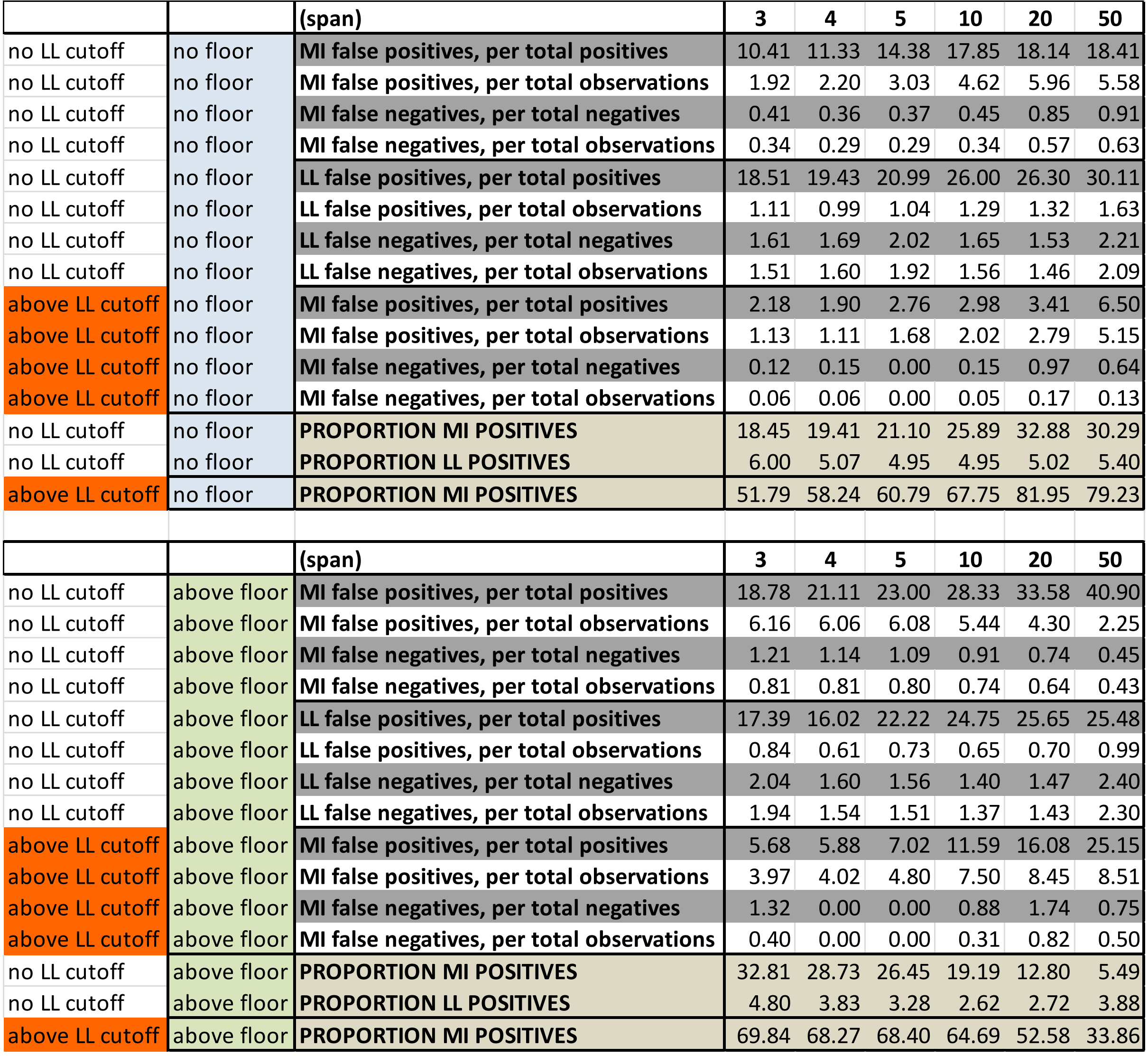
		(span)	3	4	5	10	20	50
no LL cutoff	below floor	MI	0.74	0.69	0.64	0.63	0.69	0.54
no LL cutoff	below floor	LL	0.73	0.67	0.63	0.62	0.69	0.55
no LL cutoff	below floor	**N**	**22448**	**25049**	**26622**	**27536**	**25427**	**39243**
above LL cutoff	below floor	MI/LL	0.75	0.73	0.66	0.67	0.65	0.55
above LL cutoff	below floor	**N(MI/LL)**	**1137**	**1074**	**1132**	**1207**	**1239**	**1791**

4.5.5 *Thinking in terms of false positives and false negatives*

An intuitive way of thinking about the magnitude of the impact caused by OCR errors to the statistics is to think in terms of false positive and negatives. This will answer questions such as 'if I am looking at a positive result in OCR data, how likely is it that this positive result should *not*, in fact, be positive?' Table 4.10 shows the rates of false positives and negatives (in percentages) out of all positive observations, and out of all observations. For reference, the last three lines in each table (labelled 'proportion of MI/LL positives') show the percentage of total results in each condition which are above the threshold for the result to qualify as 'positive'. The thresholds used (as discussed in 4.5) are MI=3 and LL=10.83.

In accordance with the observations in the previous sections, rates of false negatives are very small for both MI and LL. Rates of false positives, however, are relatively high overall (between 10% and 20% for MI and between 18% and 30% for LL, relative to all positive observations), with greater spans attracting worse rates. These rates are worryingly high. When considering only statistics for node/collocate pairs above the frequency floor, the rates of false negatives and false positives do not change substantially for LL. For MI, the rates of false negatives increase, though they remain small, whilst the rates of false positives increase to between 20% and 40%. MI with a LL threshold clearly remains the preferable set-up, with the rates of false positives being considerably smaller for all spans when looking at all the statistics, and considerably smaller for the smaller spans only when looking at node/collocate pairs above the frequency floor.

Table 4.10 Percentage rates of false positives and false negatives in the uncorrected sample



		(span)	3	4	5	10	20	50
no LL cutoff	no floor	**MI false positives, per total positives**	10.41	11.33	14.38	17.85	18.14	18.41
no LL cutoff	no floor	**MI false positives, per total observations**	1.92	2.20	3.03	4.62	5.96	5.58
no LL cutoff	no floor	**MI false negatives, per total negatives**	0.41	0.36	0.37	0.45	0.85	0.91
no LL cutoff	no floor	**MI false negatives, per total observations**	0.34	0.29	0.29	0.34	0.57	0.63
no LL cutoff	no floor	**LL false positives, per total positives**	18.51	19.43	20.99	26.00	26.30	30.11
no LL cutoff	no floor	**LL false positives, per total observations**	1.11	0.99	1.04	1.29	1.32	1.63
no LL cutoff	no floor	**LL false negatives, per total negatives**	1.61	1.69	2.02	1.65	1.53	2.21
no LL cutoff	no floor	**LL false negatives, per total observations**	1.51	1.60	1.92	1.56	1.46	2.09
above LL cutoff	no floor	**MI false positives, per total positives**	2.18	1.90	2.76	2.98	3.41	6.50
above LL cutoff	no floor	**MI false positives, per total observations**	1.13	1.11	1.68	2.02	2.79	5.15
above LL cutoff	no floor	**MI false negatives, per total negatives**	0.12	0.15	0.00	0.15	0.97	0.64
above LL cutoff	no floor	**MI false negatives, per total observations**	0.06	0.06	0.00	0.05	0.17	0.13
no LL cutoff	no floor	**PROPORTION MI POSITIVES**	18.45	19.41	21.10	25.89	32.88	30.29
no LL cutoff	no floor	**PROPORTION LL POSITIVES**	6.00	5.07	4.95	4.95	5.02	5.40
above LL cutoff	no floor	**PROPORTION MI POSITIVES**	51.79	58.24	60.79	67.75	81.95	79.23

		(span)	3	4	5	10	20	50
no LL cutoff	above floor	**MI false positives, per total positives**	18.78	21.11	23.00	28.33	33.58	40.90
no LL cutoff	above floor	**MI false positives, per total observations**	6.16	6.06	6.08	5.44	4.30	2.25
no LL cutoff	above floor	**MI false negatives, per total negatives**	1.21	1.14	1.09	0.91	0.74	0.45
no LL cutoff	above floor	**MI false negatives, per total observations**	0.81	0.81	0.80	0.74	0.64	0.43
no LL cutoff	above floor	**LL false positives, per total positives**	17.39	16.02	22.22	24.75	25.65	25.48
no LL cutoff	above floor	**LL false positives, per total observations**	0.84	0.61	0.73	0.65	0.70	0.99
no LL cutoff	above floor	**LL false negatives, per total negatives**	2.04	1.60	1.56	1.40	1.47	2.40
no LL cutoff	above floor	**LL false negatives, per total observations**	1.94	1.54	1.51	1.37	1.43	2.30
above LL cutoff	above floor	**MI false positives, per total positives**	5.68	5.88	7.02	11.59	16.08	25.15
above LL cutoff	above floor	**MI false positives, per total observations**	3.97	4.02	4.80	7.50	8.45	8.51
above LL cutoff	above floor	**MI false negatives, per total negatives**	1.32	0.00	0.00	0.88	1.74	0.75
above LL cutoff	above floor	**MI false negatives, per total observations**	0.40	0.00	0.00	0.31	0.82	0.50
no LL cutoff	above floor	**PROPORTION MI POSITIVES**	32.81	28.73	26.45	19.19	12.80	5.49
no LL cutoff	above floor	**PROPORTION LL POSITIVES**	4.80	3.83	3.28	2.62	2.72	3.88
above LL cutoff	above floor	**PROPORTION MI POSITIVES**	69.84	68.27	68.40	64.69	52.58	33.86

4.6 Summary

This chapter investigated the impact of OCR errors on frequency counts and collocation statistics using the CNNE matching corpus, a set of parallel corpora corresponding to the same source texts having undergone OCR, one a hand-corrected 'gold' version, one an uncorrected version, and several automatically corrected versions which will be discussed in the next chapter. A clear impact of OCR errors was found on both frequency counts and collocation statistics. The uncorrected version contained a greater number of tokens and a substantially greater number of types then the gold version. Most uncorrected types in fact do not occur in the gold corpus. Moreover, these erroneous types are not, as might be assumed, mostly infrequent: 14% of types occurring at least 10 times in the uncorrected corpus do not occur in the gold corpus, suggesting that although a frequency floor *will* help by eliminating most errors, it will not eliminate all errors. Another finding was that 8% of types occurring at least 10 times in the uncorrected corpus occur more often in the uncorrected corpus than they do in the gold corpus, revealing the presence of real-word errors. However, as would be expected, most types which occur more than 10 times in the uncorrected corpus occur less often in the uncorrected corpus than in the gold corpus. Although the overall wordcounts differ little, with the uncorrected OCR portion of the CNNE matching corpus containing 1.25% more words than the gold corpus, comparing overall wordcounts per file revealed extensive variation in the difference between gold and uncorrected wordcounts from file to file, which may be due to an uneven distribution in spurious characters and spaces, suggesting that future work may usefully pursue this matter further.

Using OCReval showed that the quality of the OCR varied much across files in the collection. The quality of the OCR also varied across year of publication and publication titles; however, these two variables are not independent in the CNNE matching corpus, so no conclusion could be drawn regarding which variable had a more substantial impact on OCR quality.

Comparing MI and LL statistics in the gold and uncorrected CNNE matching samples revealed an impact of OCR errors. Throughout, there was evidence for an effect of span, with more extreme distances and a greater spread of distances associated with larger spans, leading to the recommendation to avoid working with very large spans in OCR data. The overall results suggested that for most statistics, the impact of OCR errors remained small: 50-70% (depending on span) of MI uncorrected statistics were within 0.4 points of distance greater or smaller than their gold counterparts and 100% within 5 points of distance; and 30-40% of LL statistics were within 0.4 points of distance, with 90% within 5 points of distance. Using MI with an LL threshold resulted in a similar proportion of statistics within 0.4 points of distance from their gold counterparts, but more *over*-estimates and fewer *under*-estimates than when considering all MI statistics regardless of the LL statistic attracted by the same node/collocate pair.

Looking at the interaction between impact on collocation statistics and frequency of node showed that, broadly speaking, more frequent nodes attracted smaller differences between uncorrected and gold statistics. *However*, nodes which occurred between 10 and 100 times had the most extreme average distances, a problematic result, since nodes in this frequency band are likely to be of interest to researchers in a corpus of the size considered for this study. This explains why using a frequency floor of 10 did not lead to as much improvement in the uncorrected statistics as may have been expected. For both MI and LL, the spread of distances narrows, but the probability of encountering a statistic within 0.4 points of difference greater or smaller than their gold counterpart decreases, and the probability of encountering an *over*-estimation increases somewhat. Using MI in combination with an LL threshold as well as a frequency floor of 10 gives excellent results, however, leading to the recommendation to use this combination in OCR data.

The impact of OCR was also observed on the ranking of the statistics, though the uncorrected rankings for both MI and LL may still be considered reasonably reliable, with rank coefficients between .77 and .84 at spans of 3, 4 and 5 for node/collocate pairs above the frequency floor. More worrying, though only a small proportion of the negative observations were false negatives, a large proportion of both the MI and the LL positive results were false positives.

In conclusion, a clear impact of OCR errors was found on both frequency counts and collocation statistics. Unfortunately, OCR errors may lead to *over*-estimated collocation statistics and thus to false positives. Best practice therefore involves using a frequency floor, avoiding large spans, and using MI as a ranking statistic, in combination with LL as a significance threshold.

5 CORRECTING OCR ERRORS

5.1 INTRODUCTION

The purpose of this chapter is to introduce and evaluate two OCR post-correction solutions. In section 5.2, I briefly introduce OCR post-correction and explain the choice of VARD and Overproof as the solutions to be tested here. Sections 5.3 and 5.4 then describe and evaluate, respectively, VARD, and Overproof. Their evaluation relies on the CNNE matching corpus and the comparative methods introduced in the previous chapter; see in particular section 4.2 for an introduction to the CNNE matching corpus, and section 4.3 for an introduction to the comparative methods and measures including the measure *file edit-distance*.

5.2 INTRODUCING OCR POST-CORRECTION

OCR post-correction, the process of correcting OCR output, is a focus of much present research, and numerous recent papers report on endeavours to implement more effective and practical solutions (e.g. Daðason et al. 2014; Reynaert 2008; Volk et al. 2011; Wick et al. 2007). Solutions for correcting OCR hence exist, but none can guarantee 100% correctness, and there is always a trade-off between quality gain and time spent. Indeed, post-correcting OCR output can be so time-consuming that in some cases, simply typing up the original text ('re-keying') can end up being faster than undertaking the OCR and correction process (e.g. Cohen and Rosenzweig 2006). Commercial solutions reporting high effectiveness do exist but can be prohibitively priced. Evershed and Fitch (2014), for example, offer a state-of-the-art procedure (called Overproof) reported to be highly successful. The price to use this commercial solution on all of the *19th Century British Newspapers* (part 1) data, however, would rise to tens of thousands of dollars. As a result, I considered first a non-commercial alternative, VARD, which, although free of monetary cost, is also relatively time-consuming because it requires a gold standard and a training phase (see below). The intent of this chapter is hence to assist readers in evaluating the cost/benefit ratio (in terms of time and/or money) of these techniques for a given dataset, by exemplifying this procedure for my own dataset.

Before describing how VARD and Overproof work, a general description of OCR post-correction procedures is necessary; for a more in-depth treatment of post-correction procedures, see Kukich (1992). OCR post-correction involves three principal stages. The first consists of identifying words which are likely to be OCR errors, and which require correction. This is typically done by comparing each tokenized item to an external lexicon. Any item not found in that lexicon is flagged up as an 'error' needing correction. Both the recall and precision (see 4.3.2) are imperfect at this stage: some errors *(*real word errors) will not be found because they happen to be existing words (false negatives), e.g. *sigh* to *sign*. In addition, some correct words will be falsely identified as errors because they are not in the lexicon (false positives). Words may be omitted from the lexicon for a number of reasons, including if the word has fallen out of usage or is simply infrequent. The second stage involves generating candidates for replacements. This stage is operationalized very differently from one method to another and leads to different results. Broadly, though, the post-correction program will generate some suggestions (or *candidates*) which are found in its lexicon and which resemble in some defined way the item identified as an error. It will also rate each of the candidates using defined criteria. In the final stage, the software decides whether or not to perform a correction, and which candidate to adopt if it is making a correction. The criteria which guide these decisions differ from one method to another.

VARD[1] is a piece of software developed by Baron to normalize historical texts containing 'natural' spelling variation, i.e. spelling variation occurring in the original texts for reasons other than OCR errors, such as that occurring in Early Modern English texts prior to the standardization of English spelling (Baron 2011). A distinctive feature of this procedure is the use of letter replacement rules (see e.g. Robertson and Willett 1992 for an early description of

[1] Note that in this thesis, references to VARD are actually references to VARD 2 (developed by Alistair Baron). For references to the original VARD, see e.g. Rayson et al. (2005).

this approach to dealing with historical spelling variation). These rules describe the way in which one letter, or crucially, a cluster of letters, may be systematically used instead of another letter or cluster of letters. The rules are generated by a preliminary training phase involving manually correcting a portion of the data within VARD; this manually corrected sample can then be exported into DICER, a companion piece of software which analyses the corrected sample and generates a list of rules. A description of how VARD and DICER operate is given in the next section.

VARD had not, prior to this thesis, been tested on OCR data. The solution seemed promising to me[2] because of its reliance on letter replacement rules. It seemed intuitive that errors produced by OCR software would manifest at least some regularities: some characters would, for example, be systematically confused for other characters. It seemed logical that letter replacement rules might be able to capture these regularities. Moreover, a preliminary examination of OCR errors present in the *19th Century British Newspapers* revealed that many of them occurred due to substitutions (e.g. *collecrisely* for *collectively*), precisely the kind of pattern which should be responsive to letter replacement rules. For instance, in one page from the collection containing a total of 9,837 characters, analysis with OCReval (see 4.3.1) revealed that 72 characters were spurious (insertions), 177 were confused (substitutions) and 272 were lost (deletions).

Another reason why VARD seemed promising is its treatment of edit-distance. Most OCR post-correction solutions use some form of edit-distance (see section 4.3.1), especially at the second stage. VARD also incorporates edit-distance, but the importance of edit-distance is determined during the training phase (see next section); this seemed a useful feature since it can theoretically mitigate some of the problems associated with using edit-distance for dealing with OCR errors. In their discussion of edit-distance, Evershed and Fitch (2014: section 5.3) note that although 90% of error-to-correct-forms pairings were within an edit-distance of 3, the

[2] The idea of using VARD on OCR was initially suggested to me by Alistair Baron.

remaining 10% cases were not 'hopeless' cases, but instead simply often longer words, with solutions which their software was able to identify at edit-distances of up to 8. Solutions which use edit-distance to rank candidates hence risk discarding the correct solution when it is at a greater edit-distance than other, incorrect, candidates. (Results obtained using DICER further reinforce this point; see section 5.3.1.) In short, edit-distance should not be a primary measure for determining candidates for correcting OCR errors. That VARD determines the importance accorded to edit-distance during training hence seemed promising.

Some OCR post-correction solutions make use of character confidence rates – ratings generated by the OCR software to indicate a degree of 'confidence' that the character chosen is indeed the character present in the source (e.g. Holley 2009: 2). VARD is unable to take these into account, but I did not consider this a disadvantage. Indeed, among others, Evershed and Fitch (2014, section 5) report having used character confidence rates without much success.

Upon finding that VARD in fact lacked promise for OCR post-correction, I decided also to test Overproof, since it had been reported by Evershed and Fitch (2014) to be highly successful, and because it incorporated a very different procedure from that used by VARD. This method, referred to by Evershed and Fitch as 'reverse OCR', is to compare images of the errors to bad-quality images of words in the lexicon; this sounded promising to me.

5.3 VARD

5.3.1 INTRODUCING VARD AND DICER

The operation of VARD (which stands for ‘variant detector’) and DICER (which stands for ‘discovery and investigation of character edit rules’) are described extensively by Baron (2011). Initially, VARD simply compares each word to a lexicon; words not in the lexicon are flagged as *errors*. To generate candidates (second stage), VARD uses several tools. The first, the *known variants list*, is a list of mappings between errors and corrections which can be added to manually and/or automatically during the training phase; for Early Modern English, this approach is likely to achieve high precision but low recall (Baron 2011: 96). The second, *phonetic matching*, uses various algorithms and rules to map words to others which contain similar sounds (see Baron 2011: 96-99 for more detail); for Early Modern English, this is expected to result in high recall but low precision. The third, the *character edit rules* (see Baron 2011: 99-109), encompasses what is called elsewhere in this thesis 'letter replacement rules'. These rules describe likely correspondences between one or more characters in an error and one or more characters in the correction. Insertions (nothing replaced by one or more characters) and deletions (one or more characters replaced by nothing) are also allowed for. These letter replacement rules are used to generate suggestions by being applied one by one to the erroneous word; the resulting strings are then compared to the lexicon and the list of known variants, and any successful match is added to the list of candidates. The recall and precision of this method depend in part on the number of rules: more rules will mean more candidates can be found, which will increase the recall but reduce the precision; conversely, fewer rules will mean higher precision but lower recall (Baron 2011: 107).

Once the candidates have been generated, they are ranked by *confidence score*. This confidence score is the outcome of four heuristics. The first is edit-distance (which was not used to generate candidates) (Baron 2011: 112): the error is compared to each of the candidates. The Levenshtein distance (a standard measure of edit-distance, see section 4.3.1) is calculated and then normalized by the length of the strings (since a given edit-distance is more important for a shorter word than for a longer one). Finally, a similarity score is returned which ranges between 0 and 1: 1 for exact similarity, 0 for entirely dissimilar strings. The three other heuristics are the methods described above – known variants list, phonetic matching and letter replacement rules. For each of these methods, a score ranging between 0 and 1 is given to each candidate: 1 if the candidate is predicted by that method, and 0 if it is not. Penalties are also applied to take into account candidates which are obtained via the use of more than one heuristic, for example if phonetic matching is used to match an item in the known variants list (Baron 2011: 113). These scores, however, do not constitute the final confidence score, but are instead called the *predicted* *recall* score: they describe the probability of the candidate being correct (Baron 2011: 115). These recall scores are then combined with a *predicted precision* score for each method, which is generated as a function inversely proportional to the number of candidates generated by that method (for formula, see Baron 2011: 117); indeed the more numerous the candidates generated by a given method, the less likely each candidate is to be the right one. The contribution of the recall and precision scores towards the final confidence score depends on a configurable precision/recall weighting called the *F-score*.

Finally, once the candidates have been generated and scored, the correction is made *if and only if* the highest ranked candidate has a confidence score above a configurable threshold. Various such thresholds will be tested below.

VARD hence requires training to be effective. The training serves to determine the predicted recall and precision of each method *on a particular dataset*. This involves manually correcting a sample of the data inside VARD. As the training proceeds, a tally is kept of the effectiveness of each method in predicting the candidate which is chosen as the correct one by the user; this tally is then summarized into a weighting which determines the precise calculation of the confidence score for each candidate (see Baron 2011: 122-26 for more detail on how this is done). Besides this, the training can also be used to enhance the known variants list, and to identify useful letter replacement rules.

Identifying useful letter replacement rules is currently[3] done by exporting the training files from VARD and importing them into DICER for analysis. DICER is a tool which examines pairs of errors and their associated corrections (as determined during the manual training stage). For each pair, DICER produces a list of rules describing the transformations which may be used to turn the error into the correct form (see Baron 2011: 134-40 for details on how this works). The tool then presents these rules, along with other useful statistics such as how often

[3] In the latest VARD release. VARD 2.5.4.

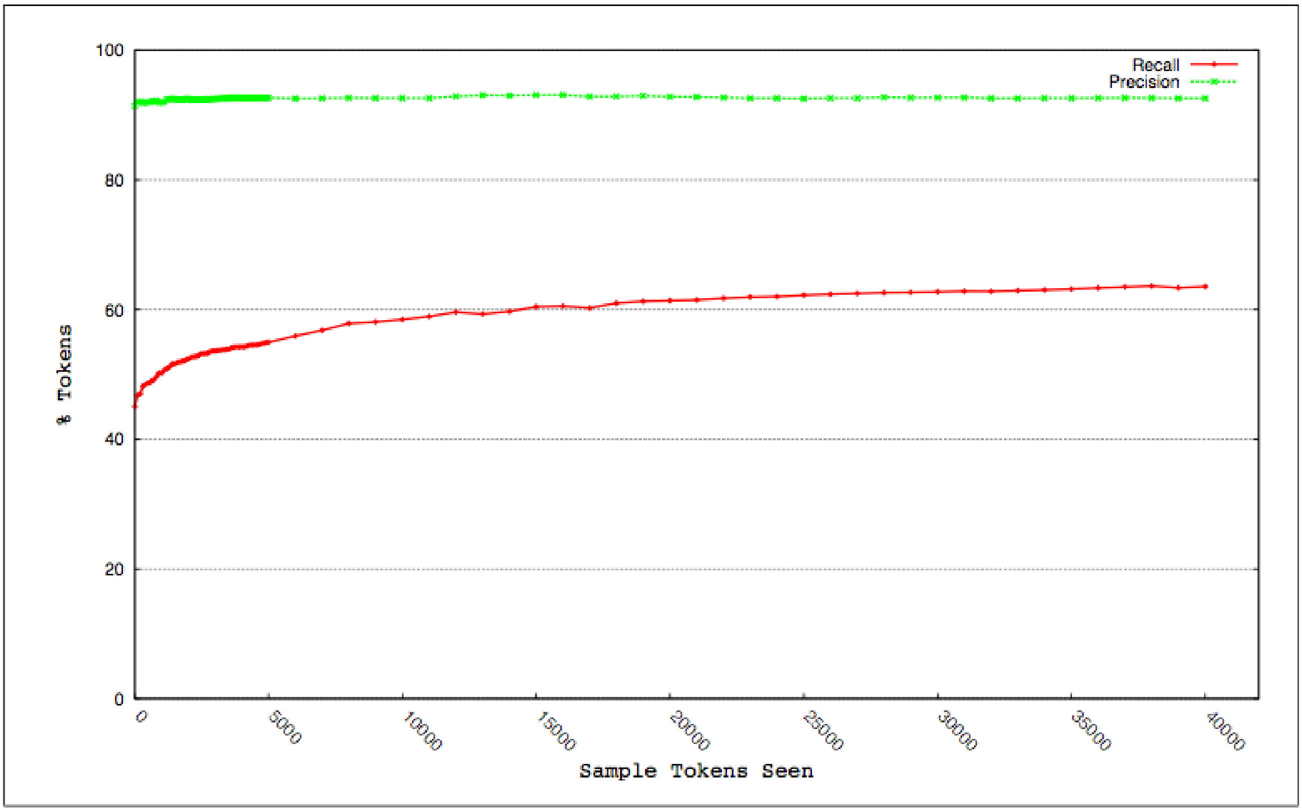
each rule is applied, in a form which is practical for further investigation of the kinds and distribution of errors in the training sample.

5.3.2 Training VARD

When using Vard, training is required in order to set appropriate weightings for the calculations of confidence scores, and to determine letter replacement rules (see previous section). DICER can help identify useful letter replacement rules, but it can also help identify 'regions' of the corpus which may behave differently (i.e. parts of the corpus which have errors which are corrected according to different letter replacement rules than other parts of the corpus).

Several questions need to be considered when assembling a training sample. First, how much training is required, or, put differently, how large should the training sample be? Baron and Rayson (2009) found that the amount of training required depends on the type of data, but that generally speaking a steep improvement in the recall is obtained with the first few thousand tokens of training, after which only marginal improvement is obtained. Improvements in precision are marginal throughout (see Figure 5.1).

Figure 5.1. Improvements in precision and recall of VARD 'corrections' as a function of number of tokens in the training sample (reproduced from Baron and Rayson 2009: 14)

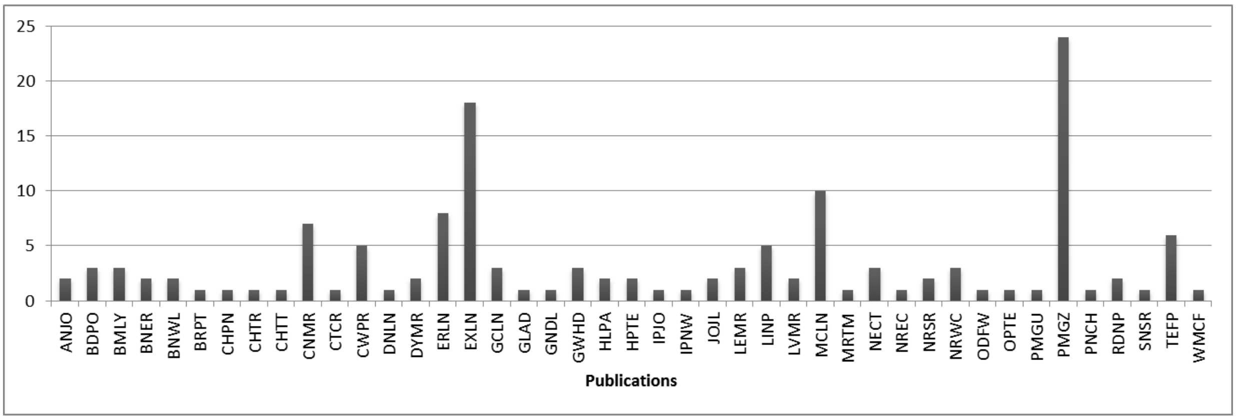


Second, what should the training sample include? I am not aware of any research investigating possible consequences of the training sample's content on the effectiveness of VARD corrections. Theoretically, it may seem that the ideal training sample would reflect perfectly both the balance and nature of types of errors in the target corpus. Assembling such a sample, however, is not possible, because which factors affect OCR, and how, is not fully known. The next best practice, then, is to guarantee that a spread of data is available in the training sample which reflects the variation in factors suspected to have an impact on OCR errors in the target corpus; in this case, the factors which I could control were year of publication and publication title.

A final, related, question is whether the training sample should attempt to reflect the proportion of types of errors in the target corpus. Again, this is not possible, because factors affecting OCR in our data are unknown. In our case, the target corpus features unequal proportions of data from different years of publication and publication titles, so reflecting those proportions in the training sample would lead to over-represented publication titles and over-represented years in the target corpus dwarfing or eliminating entirely under-represented titles and years from the training sample. The decision was hence made to favour variation over proportional representation. This means no inferences can be made from the sample about the proportion of error-types in the whole collection, but it has the advantage that, potentially, 'different' portions of the corpus can be identified by DICER after the training.

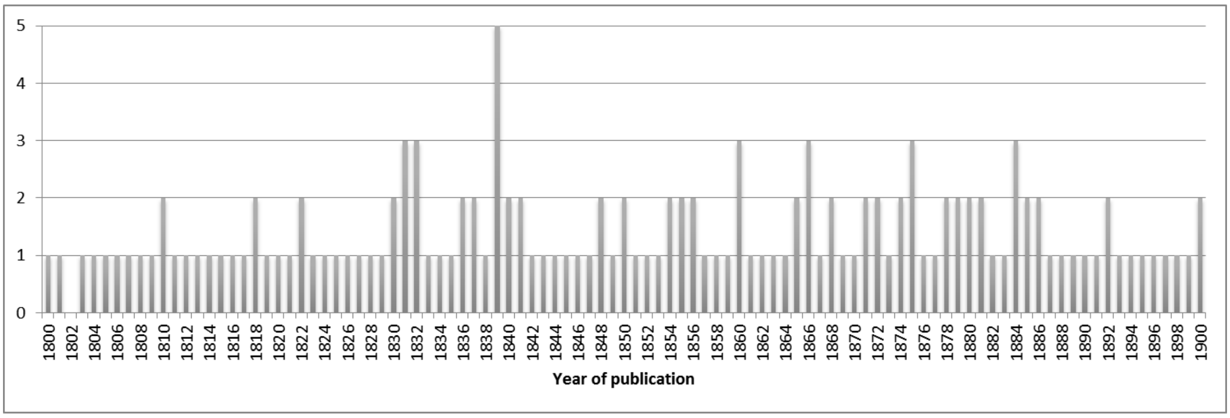
An initial sample was hence collated by selecting one random (without replacement) page per publication and one random (without replacement) page per year from the British Library's manifest[4] for the collection. This initial selection totals 142 pages amounting to 860,883 words from 43 publications and 100 years; see appendix 10.3 for a full list. Figure 5.2 and Figure 5.3 show the distribution of issues across the publications and years represented (see appendix 10.1 for the titles corresponding to the 4-letter publication codes). This initial selection was then further down-sized to an extent which would be practicable for manual correction; the down-sampling was done using VARD's in-built random partitioner. The selection was divided into 2212 partitions of between 300 and 500 words.

Figure 5.2. Pages per publication in the VARD training sample



[4] The term *manifest* designates a file which lists other files and provides information about them, i.e. a kind of index, or metadata file, for a series of data files.

Figure 5.3. Pages per year in the VARD training sample



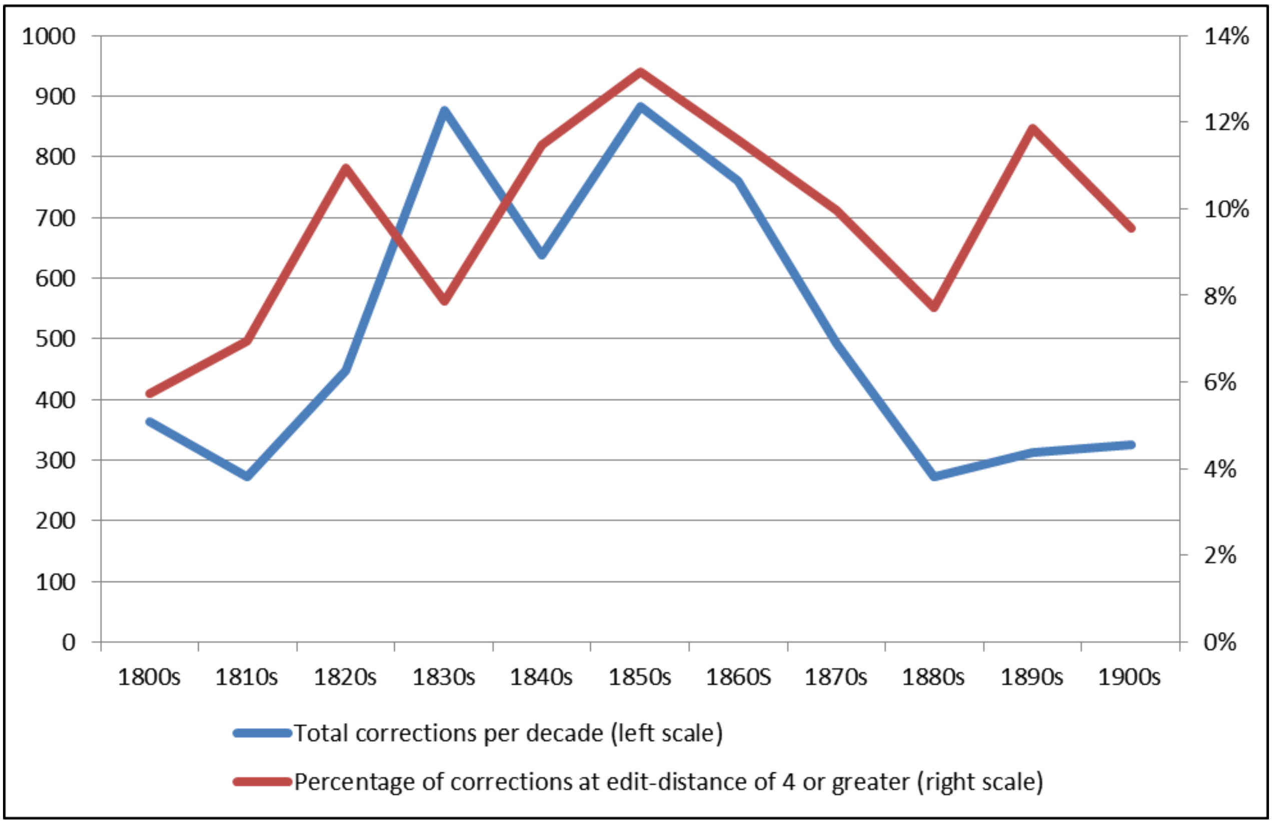
40 hours were spent on the manual training. This gave time for 5,640 corrections, from around 140 partitions, amounting to around 56,400 words seen (or around 6.5% of the training sample). In terms of edit-distance, as reported by DICER, 51% of errors were at an edit-distance of 1, 26% at an edit-distance of 2, 13% at an edit-distance of 3, and 10% at edit-distances of 4 or greater. These figures, however, underestimate the spread of edit-distances for the OCR errors in the collection. This is because only the most straightforward errors could be corrected during the manual training process. Indeed, VARD is unable to take into account errors which straddle spaces: a split word cannot (usually) be rejoined, a missing word cannot be added, a group of words stuck together by missing spaces cannot be split again, and stray character clusters which form spurious words of their own cannot be deleted. Hence, for an error to be amenable to correction during the training phase in VARD, it must be part of a word which can be mapped strictly one-to-one with a word in the original source.

Even though the edit-distance figures reported are highly likely to underestimate the *true* values for the OCR errors in the collection, they are still *worse* than those reported by Baron (2011) for natural spelling variation. In his research into historical spelling variation, Baron reports that in the Innsbruck corpus (a corpus of re-keyed letters from the 14th to the 17th century, see Baron 2011: 51) 58% of changes were at edit-distance 1, with a further 30% at edit-distance 2 and just below 9% of errors at edit-distance 3, leaving around 3% of errors at edit-distance 4 or greater (2011: 73). The figures were very similar for Baron's EMEMT samples (re-keyed medical texts from the 16th and 17th centuries, see Baron 2011: 51). This suggests that the task of automatically correcting OCR errors presents additional challenges compared to correcting 'natural' spelling variation. This impression is reinforced by the high number of letter replacement rules suggested by DICER for the OCR data: for 5,640 changes involving 7,952 operations[5], DICER provides 2,152 rules, most of which (1,613 rules, or 74%) apply only once. This high number of rarely-operative rules reveals a high degree of variation in the types of errors present in the OCR data.

In terms of variation over time, more changes were made to the data from the middle of the century (see Figure 5.4). This could be interpreted in two ways. On the one hand, it might suggest that the data from the middle of the century is worse off to begin with. This is supported by the number of changes involving high edit-distances; the 1850s, for example, exhibit 13.15% of changes at edit-distances of 4 or greater compared to 'only' 5.75% of changes at edit-distance of 4 or greater in 1800s. On the other hand, these figures could indicate that the errors in the middle of the century are more 'predictable' (i.e. easier for VARD to correct after the training process) than those in the rest of the period. This is a likely scenario, since there is also more training data for the middle of the century (see Figure 5.3), meaning that VARD will have become better 'attuned' to that period, relative to other periods. In this case, the pattern shown in Figure 5.4 is simply an artefact of the selection and correction decisions. This also implies that the number of changes made by VARD to different portions of a corpus is not a reliable diagnostic tool: for the reasons just discussed, it would be hasty to conclude that a portion of the corpus with more VARD corrections has a worse OCR quality than a portion of the corpus with fewer VARD corrections.

[5] One correction can often involve several operations, e.g. 'ckapet' -> 'chapel' (ANJO 03/07/1805) involves two operations: 'substitute k for h in second position' and 'substitute t for l in final position'. Operations are not to be conflated with edit-distance: one operation can bridge more than one edit-distance (e.g. the rule 'substitute dl for ch' operates at an edit-distance of 2).

Figure 5.4. Total changes in the VARD training sample, and percentage of changes at edit-distance of 4 or greater, per decade



In terms of types of errors/changes (see Figure 5.5), in all decades, 60-80% of corrective operations are substitutions (the average is 64.49% for the whole sample). Deletions vary between 11% and 30% (24.99% for the whole sample), and insertions between 5% and 23% (10.53% for the whole sample). There is hence a noticeable amount of variation in the types of errors occurring in each decade, but it is hard to know what is going on just from looking at the error-types. I will therefore look at the rules more specifically.

Figure 5.5. Percentage of changes involving deletions, insertions or substitutions in the VARD training sample, per decade

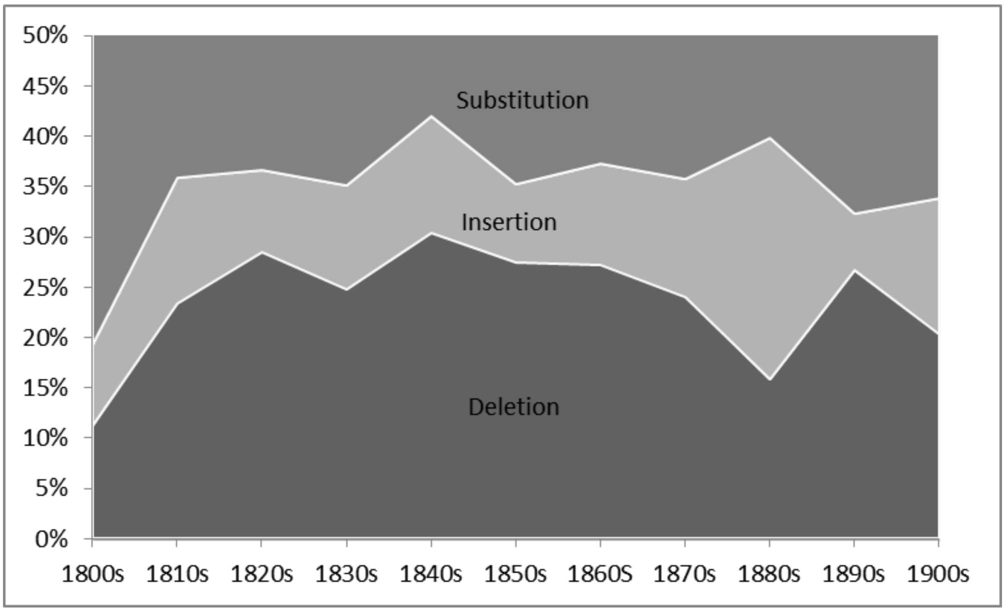


Table 5.1. Most important letter replacement rules overall (top 20) and per decade (top 5) in the VARD training sample

N°	Rule	1800s	1810s	1820s	1830s	1840s	1850s	1860s	1870s	1880s	1890s	1900s	TOP 5
1	**Delete I**	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	10
2	**Delete L**	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	9
3	**Insert space**	Y	Y	Y	Y	Y	N	Y	N	Y	N	Y	8
4	**Delete T**	N	Y	N	Y	Y	Y	Y	N	N	Y	Y	7
5	**Delete N**	N	N	N	Y	Y	N	N	Y	N	N	Y	4
6	**Substitute F > S**	Y	N	N	N	N	N	N	N	N	N	N	1
7	**Delete S**	N	N	N	N	N	Y	N	N	N	N	N	1
8	**Delete A**	N	N	N	N	N	Y	N	N	N	N	N	1
9	**Delete R**	N	N	Y	N	N	N	N	N	N	N	N	1
10	**Delete E**	N	N	N	N	N	N	N	N	N	N	N	0
11	**Delete V**	N	N	N	N	N	N	N	N	N	Y	N	1
12	**Substitute B > H**	N	Y	N	N	N	N	N	N	N	N	N	1
13	**Insert E**	N	N	N	N	N	N	N	N	Y	N	N	1
14	**Substitute O > E**	N	N	N	N	N	N	N	N	N	N	N	0
15	**Substitute C > E**	N	N	Y	N	N	N	N	N	N	N	N	1
16	**Substitute E > S**	N	N	N	N	N	N	Y	Y	N	N	N	2
17	**Substitute II > I**	N	N	N	N	N	N	N	N	N	N	N	0
18	**Delete O**	N	N	N	N	N	N	N	N	N	N	N	0
19	**Substitute U > N**	N	N	N	N	N	N	N	N	N	N	N	0
20	**Substitute EE > E**	N	N	N	N	N	N	N	N	N	N	N	0
21	**Insert S**	N	N	N	N	N	N	N	N	Y	N	N	1
25	**Insert C**	N	N	N	N	N	N	N	Y	N	N	N	1
32	**Substitute TI > N**	N	N	N	N	N	N	N	N	N	Y	N	1
57	**Substitute I > S**	Y	N	N	N	N	N	N	N	N	N	N	1
99	**Substitute L > S**	Y	N	N	N	N	N	N	N	N	N	N	1

Table 5.1 shows the 20 most important rules for the entire sample, as well as any other rules in the top 5 for an individual decade. The first column (N°) shows the ranking in the rule list for the entire sample (highlighted if not in the top 20); the last column shows in how many decades the rule features in the top 5 (highlighted if in at least 2 decades). The central columns show whether or not a rule appears in the top 5 for that decade (highlighted if not).

A few things are immediately apparent: first, some rules are important in all decades, whereas others are more important in some decades than others. Second, although substitutions are by far the most frequent operation (constituting 64.49% of operations overall), they are more varied than deletions, which constitute 24.99% of operations overall yet make up a majority of the most widely used rules both for individual decades and for the whole sample. It thus seems that the pool of likely spurious characters (i.e. characters that need to be deleted) is relatively small, compared to the pool of possible substitutions. This suggests that an approach centred around 'letter replacement' rules is unlikely to be very successful, because the list of rules needed would be very long, as well as conflicting: in the top 30 rules overall, rules appear for E>S, E>C and E>O (several of which may apply in the same situation, e.g. the error 'elot' could just as well be corrected to 'slot' as to 'clot'). How can the software know which replacement of E to choose? Clearly letter replacement rules by themselves will not suffice.

In terms of identifying different portions of the corpus which behave differently – as identified from the divergent applicability of letter replacement rules – no significant effects of either year or newspaper were identified from the DICER results. The only major difference found pertained to the 'f->s' rule, which only applies at the beginning of the corpus. This is an unsurprising result given the long 's' is known to have all but fallen out of usage by the early nineteenth century (Attar 2010). Since this rule is unlikely to cause problems in other parts of the corpus where it does not apply, it was deemed unnecessary to train VARD separately on various parts of the dataset, and the tests below were hence done on the training sample as a whole.

To go from the training phase to the correction phase, a list of rules needs to be imported into VARD. In the current version of VARD (version 2.5.4), this list needs to be manually composed from the rules identified by DICER and manually re-imported. The list of rules provided by DICER is too lengthy to be re-imported in its entirety into VARD. Indeed, since more rules lead to more potential candidates (see section 5.3.1), doing so will lead to lower precision scores, which will hence lead to less changes at a given threshold. Hence the full list of rules could not be imported. Instead, two lists of rules were composed.

The first list (from now on referred to as the 'long list') was assembled from the 75 rules which applied most frequently (20 times or more), covering 4,166 out of 7,952 cases (or 52% of cases). Some of these rules were further split into several rules (as long as the separate rules would still apply 20 or more times); e.g. rule #5 is 'delete N' (and it applies in 129 cases), but studying the list of cases in which it applies revealed that in 64 cases, the more specific 'change MN to M' rule actually applied, so this transformation rule was added to the list. This lengthened the list of 75 rules to 88. The next step was to remove insertion rules, which are too computationally onerous to run. No insertion rule could be replaced by a substitution rule which applied more than 20 times; hence insertion rules were removed without replacement. This led to a final list of 79 rules.

A shorter list of rules (from now on referred to as the 'short list') was also composed, to test whether more or fewer rules generated better results during the correction phase. For this shorter list, the 20 most frequently applied rules provided by DICER provided the starting point; these rules applied in 60 or more cases, covering together 2,425 cases out of 7,952 (or 30%). The list was reduced to 18 rules after removing one insertion rule and one rule which did not adequately describe the changes it arose from.

Once the lists of rules are provided (see appendix 10.5), the correction setup needs to be defined. VARD uses a configurable *weighting threshold* which determines how many changes VARD attempts: only candidate changes which have attracted a weighting higher than the threshold will be applied. Weightings are assigned to correction candidates according to how likely to be correct VARD deems them to be. Using a higher threshold hence leads to fewer changes being applied, but with these changes being, at least in theory, *better* ones.

In practice, Baron's (2011: 151) investigations suggested that the recall (i.e. the number of changes attempted out of changes needing to be made) varies more dramatically than the precision (i.e. the number of correct changes out of changes made), so that a lower weighting threshold is often more effective, as long as it is above a certain value. Since the nature of our data is so different from the data that Baron (2011) was looking at, it is difficult to predict what threshold might be most effective. 4 threshold values were hence tested: 70, 50, 30 and 10. At a threshold of 70, no changes were made. The difference between thresholds of 50, 30 and 10 will be described in the next section.

VARD uses 4 systems to assign weightings to candidates: edit-distance, phonetic matching, previous experience and letter replacement rules (see section 5.3.1). A high weighting means a proposed correction is considered to be more likely to be correct. The relative importance VARD accords to these systems for calculating the weighting of each candidate is determined by the training phase. In my data, the training phase revealed that two of these systems, edit-distance and phonetic matching, were not particularly helpful. Edit-distance was not very helpful because the right correction was often at a higher edit-distance from the OCR error than other candidate corrections (see also Table 5.4). Phonetic matching was also unhelpful, and unsurprisingly so since OCR errors (unlike historical spelling variants) are not related to phonetics. Since these two systems did not prove helpful during the training phase, VARD therefore gave them low importance for future corrections. In practice then, VARD judged candidates predominantly based on the two remaining systems, previous experience and letter replacement rules. However, since most errors are unique, previous experience will, in general, not have been of relevance to the assessment of a particular set of candidate changes.

Hence, letter replacement rules is ultimately the single system on which VARD relied most to assess candidate changes.

5.3.3 Assessing changes made by VARD

In order to evaluate the effectiveness of the changes introduced by VARD, the uncorrected sample of the CNNE matching corpus (see section 4.2) was corrected using 8 different setups: 2 at a weighting threshold of 10, 2 at a threshold of 30, 2 at a threshold of 50, and 2 at a threshold of 70. At each threshold, one correction setup used the short list of rules, the other the long list of rules. The threshold-of-70 versions did not need to be analysed further since under these setups, no changes were introduced by VARD. Throughout, the CNNE matching samples corrected by VARD will be referred to as *short* or *long* (in reference to which list of rules was used in the correction setup) and as *10*, *30* or *50* (in reference to the weighting threshold used). Section 5.3.3.1 compares the uncorrected and VARD-corrected samples in order to describe the nature and quantity of the changes introduced by VARD. Section 5.3.3.2 then compares the uncorrected, VARD-corrected and gold samples in order to evaluate the effectiveness of the changes introduced by VARD.

5.3.3.1 Comparing the uncorrected and VARD-corrected versions in the CNNE matching corpus: what changes were introduced by VARD?

Before considering how useful VARD proved as an OCR-correction solution, this section looks at how much change VARD introduced. Figure 5.6 (and Table 5.2) compare the corrected versions against the uncorrected version (using OCReval, see 4.3.1); the box plots (see 4.4.2) show the spread of file edit-distances (defined in section 4.3.1) for a given correction setup. Figure 5.6 reveals that there is quite a spread in how many changes each file gets, although the spread is smaller in setups involving the long list of rules then in setups involving the short list. Naturally, higher thresholds lead to less changes (see section 5.3.2). Overall, setups involving the short list of rules also produce more changes than setups involving the long list, which fits in with what was expected (see section 5.3.2). It would be natural to expect that files which are worse off in the first place attract more changes than files which start off with better quality OCR; Figure 5.7 suggests this expectation is roughly borne out by the data.

Figure 5.6. Distance between uncorrected OCR files and corrected versions

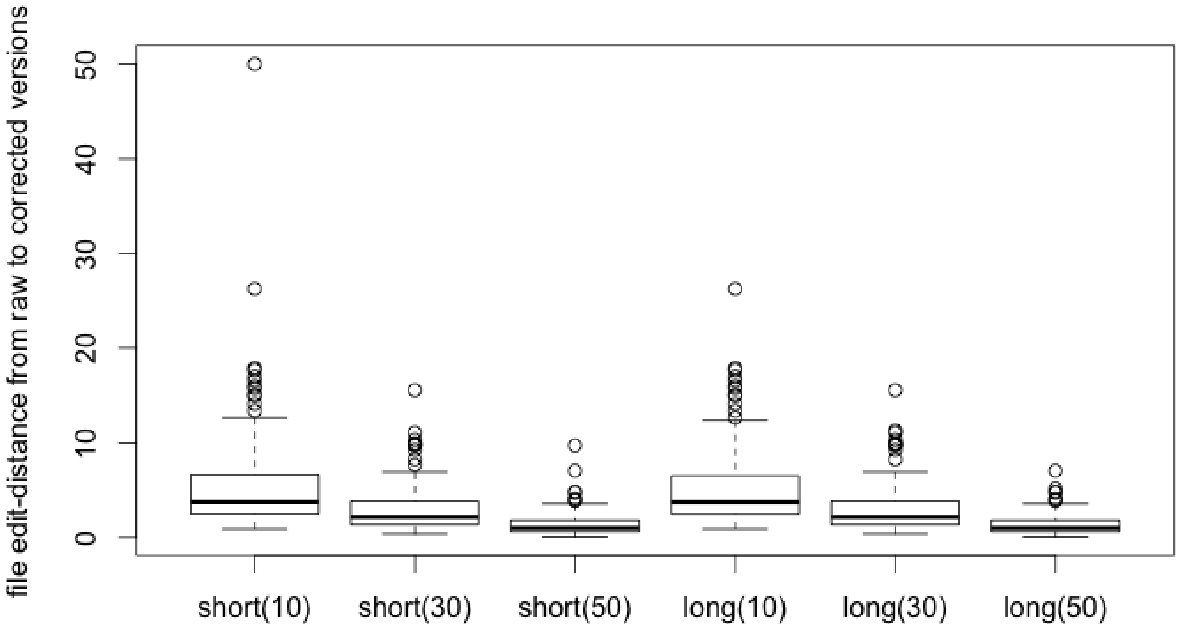
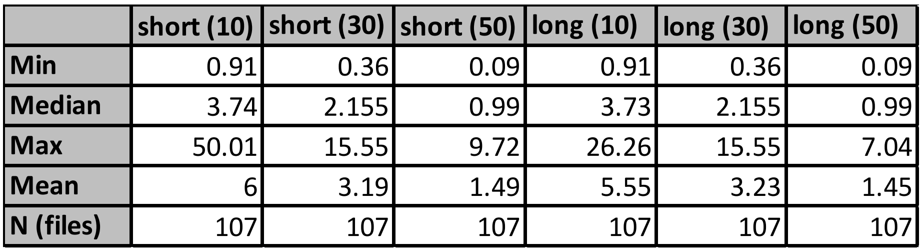
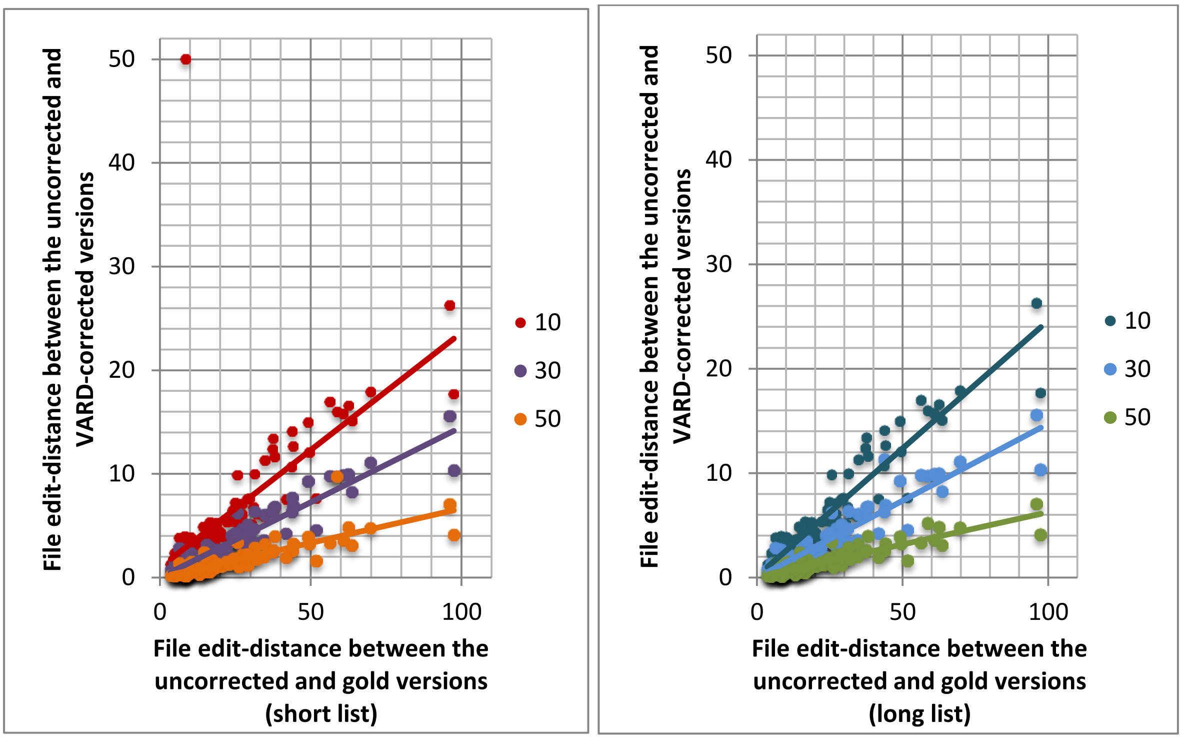


Table 5.2. Distance between uncorrected OCR files and corrected versions (file edit-distance)



	short (10)	short (30)	short (50)	long (10)	long (30)	long (50)
Min	0.91	0.36	0.09	0.91	0.36	0.09
Median	3.74	2.155	0.99	3.73	2.155	0.99
Max	50.01	15.55	9.72	26.26	15.55	7.04
Mean	6	3.19	1.49	5.55	3.23	1.45
N (files)	107	107	107	107	107	107

Figure 5.7. Relationship between original quality of file (x axis) and number of changes made during the corrective procedure (y axis)



5.3.3.2 Comparing all parts of the CNNE matching corpus: how effective were the changes introduced by VARD?

We have seen that there is some difference between the correction setups in terms of how much they change the original files, but is there a difference in terms of how much improvement they yield? Figure 5.8 (and Table 5.3) suggest that overall there is virtually no difference between the different corrected versions in terms of their distance to the gold standard, and that the corrected versions are of about the same quality as the uncorrected data (i.e. there are roughly the same number of errors in the 'corrected' versions as in the uncorrected version).

Figure 5.8. Comparison between uncorrected and corrected versions and gold standard

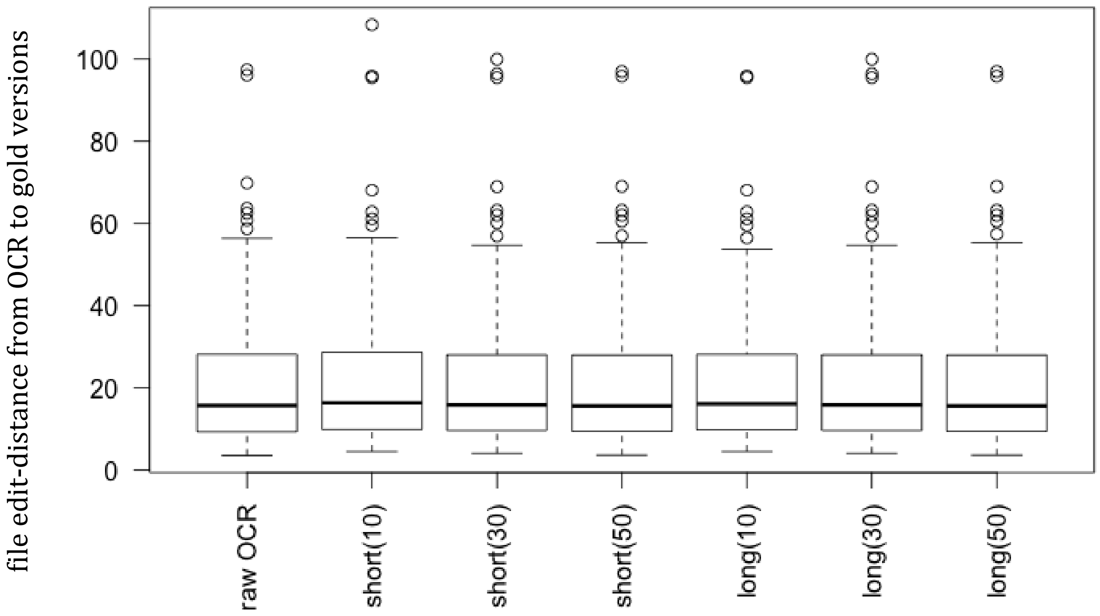
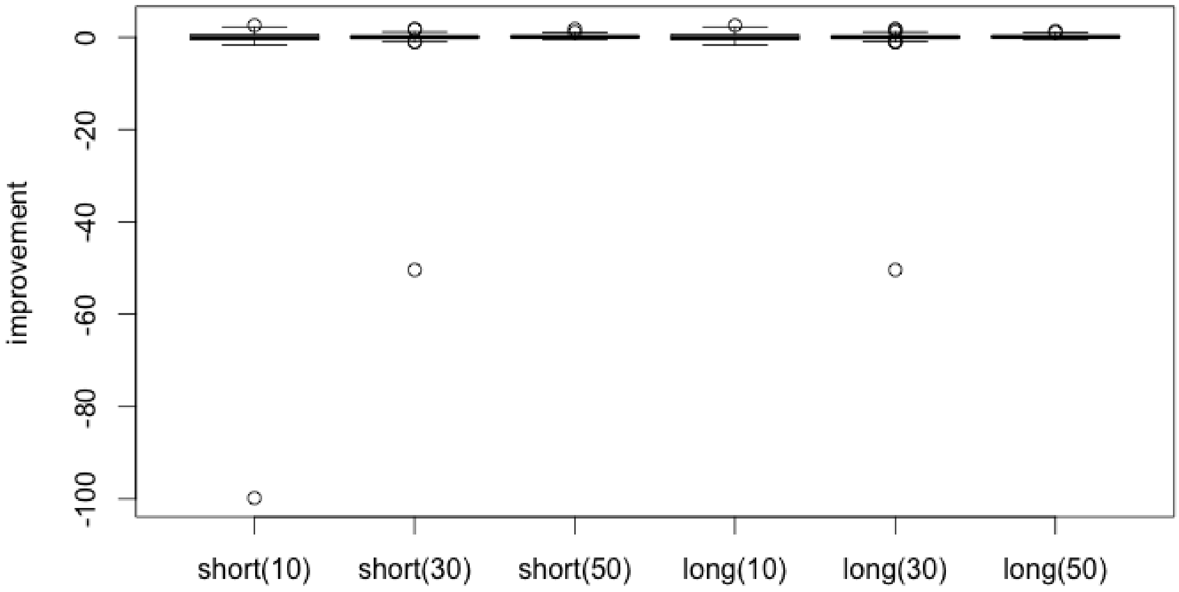


Table 5.3. Comparison between uncorrected and corrected versions and gold standard

	raw	short (10)	short (30)	short (50)	long (10)	long (30)	long (50)
Min	3.63	4.56	4.1	3.71	4.56	4.1	3.71
Median	15.75	16.39	15.91	15.64	16.15	15.91	15.64
Max	97.43	108.37	100	97	95.93	100	97
Mean	22.23	23.14	22.63	22.09	22.21	22.64	22.09
N (files)	107	107	107	107	107	107	107

This result is surprising, showing virtually no overall difference between the different corrective setups, even though these setups had been found to yield a different number of changes (see previous section). Figure 5.8 and Table 5.3, whilst showing the overall outcome of the corrective procedure, obscure any variation that exists between files in terms of how much they are improved by the corrective procedure. This variation can be teased out by taking the difference between how bad the original OCR files were (i.e. the file edit-distances between the OCR and the gold parts of the CNNE matching corpus) and how bad the 'corrected' versions are (i.e. the file edit-distances between the 'corrected' and the gold parts of the CNNE matching corpus). This is what is shown on Figure 5.9 and Table 5.4. Here, a positive figure indicates an overall improvement from the corrective procedure for that file, whereas a negative figure indicates that the corrective procedure actually leaves the file worse off. The figures show that although some improvements are made to some files in all corrective setups, for most files in all corrective setups the changes amount to virtually no difference in the quality of the output, and for some files, the output is actually much worse than the original OCR. More could be said about what is going on with a manual assessment of 'good' and 'bad' changes occurring each corrective setup, but such an extensive qualitative investigation is beyond the scope of the present analysis.

Figure 5.9. Improvement brought by corrective procedure (difference between distance from original files to gold files and corrected files to gold files) (full extent followed by close-up)



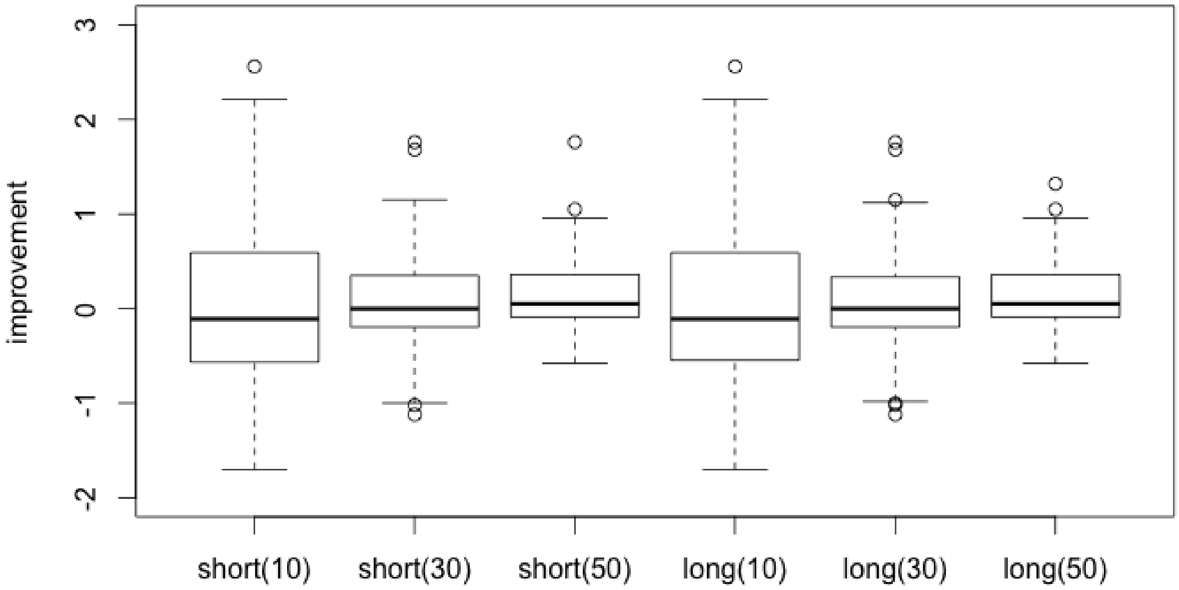


Table 5.4. Improvement brought by corrective procedure (difference between distance from original files to gold files and corrected files to gold files)

	short (10)	short (30)	short (50)	long (10)	long (30)	long (50)
Min	-99.83	-50.4	-0.58	-1.7	-50.4	-0.58
Median	-0.11	0	0.05	-0.11	0	0.05
Max	2.56	1.76	1.76	2.56	1.76	1.32
Mean	-0.9	-0.4	0.14	0.03	-0.41	0.14
Mean (magnitude)	1.6	0.87	0.27	0.67	0.86	0.26
N (files)	107	107	107	107	107	107

Table 5.5. Recall and precision figures for the different corrective setups

	short (10)	short (30)	short (50)	long (10)	long (30)	long (50)
Recall	23%	14%	6%	23%	14%	6%
Precision	48%	50%	52%	48%	50%	52%

It is clear from these figures that VARD has not proved an effective corrective solution. For the sake of completeness, Table 5.5 shows the recall and precision measures calculated using the approximation method outlined in section 4.3.2. They suggest that VARD has changed very little, and that of the changes that it made, only about half were improvements. The averages are no different between the two list of rules. There is also virtually no difference in the precision between the different thresholds, but there is, as expected, a notable impact on recall.

These figures lead to a clear conclusion – that this approach to correcting OCR errors will not be effective without far more manual work than the 40 hours I was able to invest (there is no 'low-hanging fruit'). The process of using VARD to correct OCR errors is time-consuming, since it involves both manual training and creating a gold standard (for purposes of assessing the trained system), but does not pay off. The great number of rules produced by DICER and the high edit-distance separating the OCR errors from their corrections suggest that the variety and nature of OCR errors poses challenges for automatic correction, and that an approach based on letter replacement rules is unlikely to prove effective on a large amount of mixed data such as the British Library's *19th Century Newspapers*.

5.4 OVERPROOF

5.4.1 *INTRODUCING OVERPROOF*

Overproof is an elaborate, and expensive[6], OCR post-correction solution, which Evershed and Fitch (2014) describe as ‘a fully functional end to end batch OCR corrector delivering corrected texts at a high rate on a standard commercial “cloud” server’. It uses a combination of a number of tools and heuristics which rely both on knowledge about how language works (a *language model*) and knowledge about what OCR errors can look like (an *error model*). Evershed and Fitch (2014) do not go into much detail about how errors are first identified, noting simply that they use 5-gram (i.e. n-grams[7] containing 5 words) data 'to quickly triage the input, making uncontentious and simple corrections' and then passing 'blocks containing suspect text', along with information about the context of these blocks, to the part of their software which generates candidates and make the corrections. Both the language model and error model intervene in the generation and scoring of candidates. The language model includes n-gram frequency data, from which a lexicon is created, containing words which are likely to be correct; the data also serves to generate probabilities that certain words will appear in the texts to be corrected. Also included in the language model is knowledge about the context of specific errors: Evershed and Fitch (2014) mention using 'topic words', but without providing a precise definition. The error model includes two heuristics which help determine the probability that a given lexicon word may be corrupted into a given, observed, error. One of these, the *confusion matrix*, is akin to, though more complex than, VARD's letter replacement rules. This confusion matrix is obtained from initially training Overproof on test data, and it contains information about the transformations which particular characters and clusters of characters are likely to undergo. Hence, in addition to accounting for operations which affect a single group of characters, as VARD does – substitution, deletion and insertion – Overproof can also account for operations

[6] With the current estimated size of the total c19th newspapers (part 1) around 30 billion words, the budget required for using Overproof on the collection would be counted in the tens of thousands of dollars, at the advertised monthly price of $5420 + $4.80 per million words, when submitting at least 1 billion words.

[7] *N-grams* are sequences of words which co-occur in a specific order, see also section 2.3.3.2.

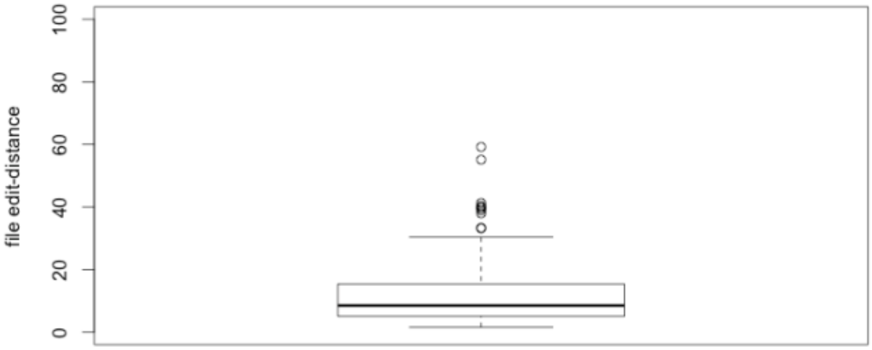
which affect clusters of characters – split, pair and join. The other heuristic, *reverse OCR*, is a technique based on comparing bad-quality bitmaps (i.e. images) of observed OCR errors and of lexicon words. The system does not generate all possible candidates using these heuristics, but instead generates candidates which are highly probable first, stopping once a sufficient number (between 1 and 20) of candidates have been generated. Each candidate is then evaluated in turn, and the candidate which is associated with the lowest transformational cost, given the language and error models, is accepted as the correction.

5.4.2 Evaluating Overproof

5.4.2.1 Changes introduced by Overproof

Kent Fitch graciously agreed to correct a sample of my data using Overproof for free. I sent him the uncorrected CNNE sample. Fitch reported that the processing stream encountered 164,367 tokens, of which it changed 18,698. Figure 5.10 (summary values in Table 5.6) shows the amount of change introduced by Overproof, as revealed by comparing the Overproof-corrected sample to the uncorrected sample (compare to Figure 5.6 for the corresponding VARD figures). Both the median and mean are roughly twice as high for Overproof as they are for VARD: Overproof has clearly introduced more changes than VARD (in any setup) did.

Figure 5.10. Distance between uncorrected files and Overproof-corrected ones (full extent followed by close--up)



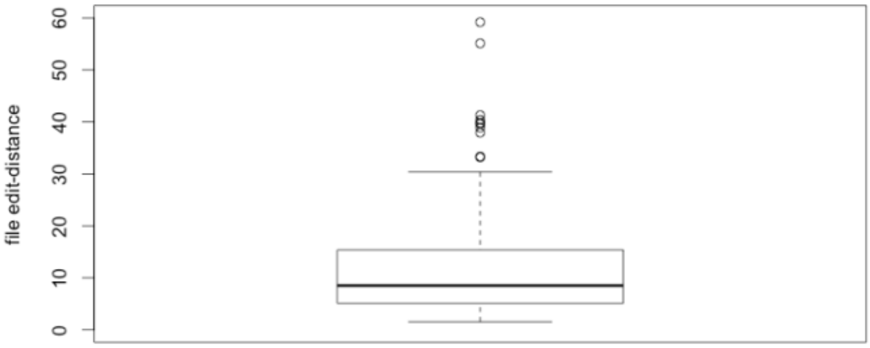
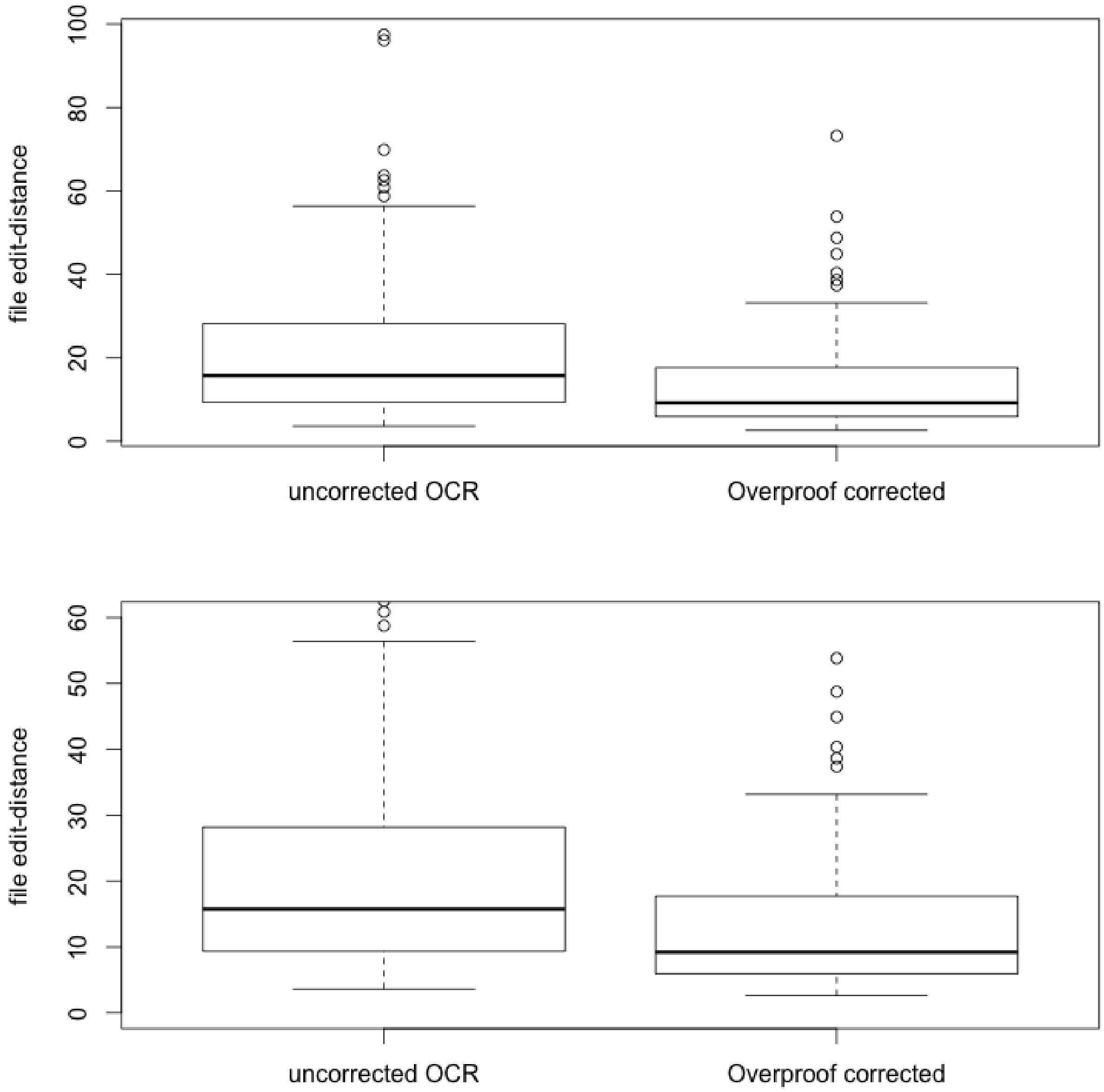


Figure 5.11. Comparison between uncorrected and corrected versions and gold standard (full extent followed by close-up)



However, we also want to know if these changes are helpful or not. Figure 5.11 compares the distances between the uncorrected and gold files to the distances between the Overproof-corrected and gold files (compare to Figure 5.8 for VARD). Whereas for VARD, there was virtually no difference between the distance separating the uncorrected version to the gold standard and that separating the corrected versions to the gold standard, here the distances separating the corrected version are generally smaller than those separating the uncorrected version from the gold standard. This is quite promising. The difference between these two figures gives us a measure of how much improvement has been brought about by the corrections; this is shown in Figure 5.12 (compare Figure 5.9 for VARD). We see that the improvement varies extensively from one file to another; nevertheless, all files improve at least a little (the minimum improvement is 0.46; see Table 5.6, which provides summary values for Figure 5.10, Figure 5.11 and Figure 5.12).

Figure 5.12. Improvement brought by corrective procedure (difference between distance from original files to gold files and corrected files to gold files) (full extent followed by close-up)

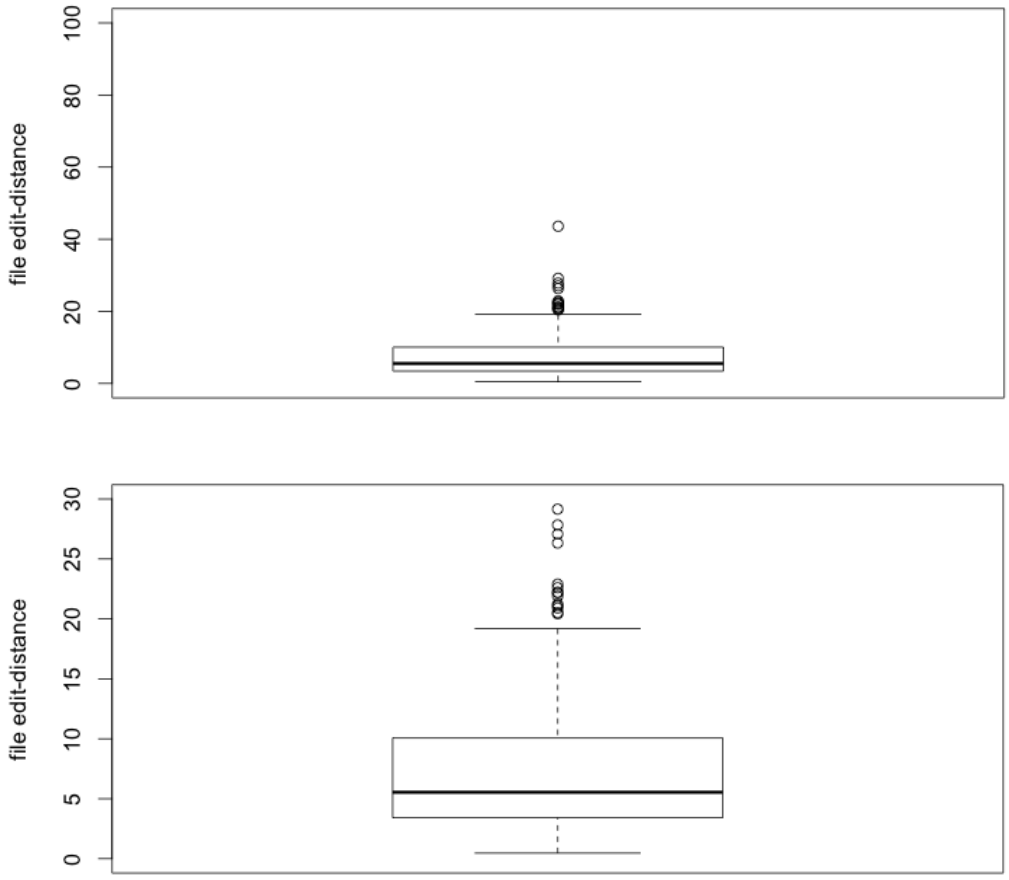


Table 5.6. Summary values for Figure 5.10, Figure 5.11 and Figure 5.12

	uncorrected to Overproof	Overproof to gold	uncorrected to gold	improvement from uncorrected to Overproof
Min	1.55	2.62	3.63	0.46
Median	8.51	9.21	15.75	5.54
Max	59.2	73.2	97.43	43.6
Mean	12.645	13.778	22.233	8.455
N (types)	107	107	107	107
See figure	5.10	5.11	5.11	5.12

In terms of precision and recall (see section 4.3.2), the average precision is 1 – [½ (12.64 – 8.45)]/12.64 which is around 83%, and the average recall is 12.64/22.23

which is around 56%. These are promising figures, especially compared to the best VARD figures (23% recall with a threshold of 10 and 52% precision with a threshold of 50). In practice, however, the question that matters is whether the changes made by Overproof translate into a higher reliability of the statistics derived from the data. It is hence worth taking a closer look at some of the characteristics of the Overproof-corrected version.

5.4.3 IMPACT ON TYPE AND TOKEN COUNTS

Before looking at the impact on collocation statistics in the next section, this section considers how type and token counts in the Overproof sample differ from those in the uncorrected sample. The figures cited here are hence a counterpart to those provided in section 4.4.1.

Table 5.7 provides overall type and token counts for the uncorrected, Overproof and gold versions of the CNNE matching corpus. Unsurprisingly, the Overproof version is situated between the uncorrected and gold versions for both type and token counts. Its type/token ratio, however, is much more similar to the gold corpus, with the number of types considerably closer to the gold size. Hence even in these general respects, the Overproof corrections yield considerable improvement. This is further visible simply by looking at the number of hapaxes (words which occur only once, most of which are errors in the OCR corpus): the number of hapaxes halves from the uncorrected to the Overproof version, coming from 3x more to only 1.5 more than in the gold corpus.

Table 5.7. Type and token counts in the uncorrected, Overproof-corrected and gold versions of the CNNE matching sample

	uncorrected corpus	Overproof corpus	gold corpus
N (tokens)	162617	161983	160616
N (types)	26954	16944	13831
Type/token ratio	16.58%	10.46%	8.61%
Hapaxes	18750	9078	6114
Hapaxes as percentage of types	69.56	53.58	44.21

Although the Overproof type and token counts are closer to the gold version than the uncorrected version, there are still many more types in the Overproof version than in the gold. The proportion of gold types represented in the Overproof version is only very slightly higher than the uncorrected equivalent, with around 3% more gold types occurring in the Overproof-corrected than in the uncorrected (see Table 5.7). Overproof fares considerably better in terms of the proportion of types in the corrected corpus which are actually present in the gold corpus: whereas only 44% of types in the uncorrected sample occur in the gold, 74% of types in the Overproof sample also occur in the gold (see Table 5.8, which is the counterpart of Table 4.2). This indicates that Overproof has indeed succeeded in removing many errors.

The proportion of types in the Overproof corpus which occur at least 10 times in the gold corpus is also higher than the equivalent uncorrected proportion: whereas 7% of uncorrected types occurred at least 10 times in the gold corpus, 11% of those in the Overproof corpus do.

Since it had been suggested (in section 4.5) that working with a frequency floor may be useful with OCR data, it is also interesting to look at what happens when a frequency floor is introduced in the Overproof version. With a frequency floor of 10 in the Overproof version, almost 99% of the types retrieved also occur in the gold corpus. This is an improvement from around 93% in the uncorrected version. It is also a remarkable figure: it suggests that using a frequency floor of 10 in the Overproof corpus excludes virtually all non-real-word OCR errors from the results.

Table 5.8. Relationship between the types occurring in the Overproof and gold samples

Types occurring in …	Count	% of types in the OCR corpus (occurring at least 10 times)	% of types in the gold corpus (occurring at least 10 times)
both OCR and gold samples	12480	73.65	90.23
OCR sample, and at least 10 times in the gold sample	1866	11.01	(99.89)
at least 10 times in the OCR sample, and gold sample	1778	(98.67)	12.86

Let us now consider the reliability of frequency counts for individual types (see Table 5.9 and its counterpart for the uncorrected to gold comparison, Table 4.3). Whilst the uncorrected version had 57% of types which were over-estimated, the Overproof-corrected version has only 34% of types which occur more than they should. On the other hand, whereas the figure was only 15% among types occurring at least 10 times in the OCR corpus, for the Overproof version the figure is still almost 27% of types over-estimated among types occurring at least 10 times in the Overproof version. And where for the uncorrected corpus, most of these over-estimated types (55 out of the 57%) were simply types which did not occur in the gold version, in the Overproof version only 26 out of the 34% are types which do not occur in the gold. In fact 25% of Overproof types which occur at least 10 times are over-estimates of types which do occur in the gold, from around 9% in the uncorrected version. This suggests that real-word errors may well be more of a problem in the Overproof version than in the uncorrected version. Finally, there are overall very slightly more underestimations in the Overproof version than in the uncorrected version (17% compared to 16%), but much fewer when considering only types which occur at least 10 times (54% compared to 75%).

Table 5.9. Over- and under-estimates for Overproof type frequencies compared to gold type frequencies

	Count of types (out of all types in the OCR corpus)	Count of types (out of all types occurring at least 10 times in the OCR corpus)
Occur more often in the OCR sample	5764	482
%	34.02	26.75
... and don't occur in the gold (suspected non-dictionary words)	*4464*	*24*
%	*26.35*	*1.33*
...and do occur in the gold (involving suspected real-word errors)	*1300*	*458*
%	*7.67*	*25.42*
Occur as often in the OCR and gold sample	8243	340
%	48.65	18.87
Occur less often in the OCR sample (and do occur in the gold sample)	2937	980
%	17.33	54.38

Together, these results suggest that figures from the Overproof version will be less conservative than those from the uncorrected version, attracting in particular more real-word errors. This is a real disadvantage, because real-word errors can be more problematic than other types of errors. First, other types of errors usually only affect counts once (they are not counted where they should be, but cause no other problems in the counts). In contrast, real-word errors affect counts *twice*: they are not counted where they should be, and they are also counted where they should not be. Moreover, errors which are not real words are likely to be noticeable to the user. Real-word errors, in contrast, may mislead the user, who may not recognise them as errors when they encounter one. However, the benefit of increased word accuracy and type representation in the Overproof version may well offset this disadvantage. Comparing the collocation statistics derived from the Overproof version to those of the uncorrected and gold versions may help further evaluate the usefulness of Overproof's correction.

5.4.4 *IMPACT ON COLLOCATION STATISTICS*

So far, the advantages of using Overproof seem mitigated by the increased risk of encountering real-word errors. An important question for the analyst interested in collocation patterns is whether the corrections effected by Overproof substantially impact the reliability of the collocation statistics. This section is the counterpart of section 4.5, and will compare the differences between gold and Overproof statistics to the differences between gold and uncorrected statistics, all of which are derived from the corresponding CNNE matching sample (see section 4.2). The number of node/collocate pairs from the Overproof sample included and excluded from the analysis is shown (for each span) in Table 4.8.

5.4.4.1 Overall variation

Do the Overproof-corrections improve the collocation statistics obtained from OCR statistics? Figure 5.13 and Figure 5.14 (as compared to Figure 4.7 and Figure 4.8) for MI and Figures 5.15-5.18 (as compared to Figures 4.9-4.12) for LL show that although the spread of MI and LL distances are similar in the Overproof and in the uncorrected samples, improvement is visible in the Overproof sample: indeed, in the Overproof sample, the spread of values is somewhat narrower, and the likelihood of encountering statistics with very small differences from their gold counterpart is greater, than in the uncorrected sample. For MI with a LL threshold (see Figure 5.19 and Figure 5.20, as compared to Figure 4.13 and Figure 4.14), the impact of Overproof-correction is more dramatic, but of a similar nature, to the improvement accrued for MI and LL statistics when considered separately: this combination then seems especially felicitous in Overproof-corrected data. The same observations also hold for MI above a frequency floor (see Figure 5.21 and Figure 5.22 as compared to Figure 4.15 and Figure 4.16) and LL above a frequency floor (see Figure 5.23 to Figure 5.26 as compared to Figures 4.17 to 4.20). For MI with an LL threshold above a frequency floor (see Figure 5.27 and Figure 5.28, as compared to Figure 4.21 and Figure 4.22), the impact is again similar, but more dramatic, suggesting, again, that using MI with an LL threshold is particularly useful in Overproof-corrected data.

Figure 5.13 Probability that any given Overproof MI statistic will be situated at a given distance of the corresponding gold statistic (full extent)

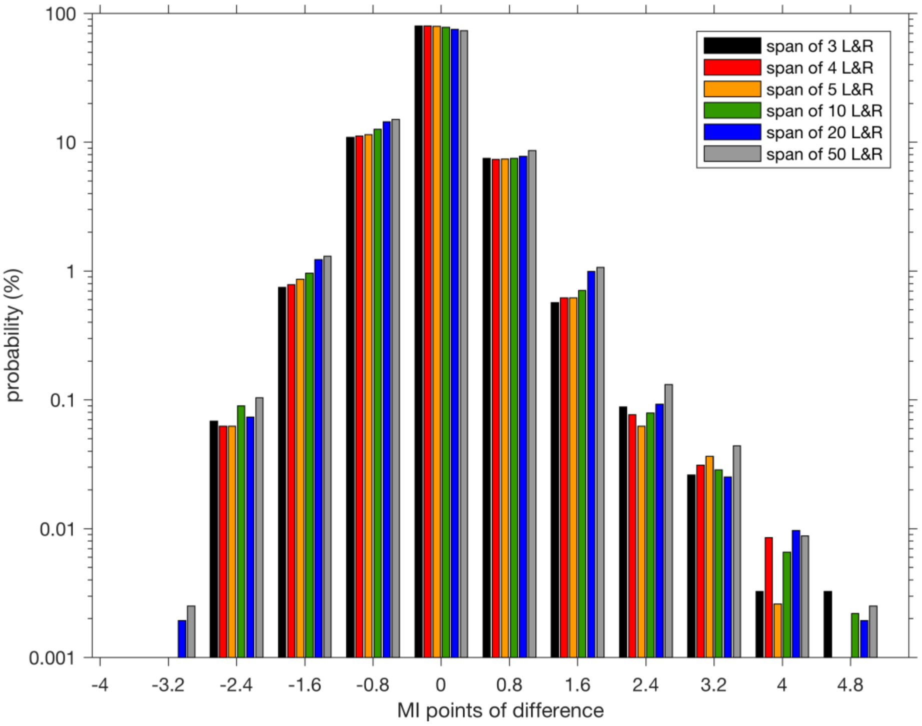
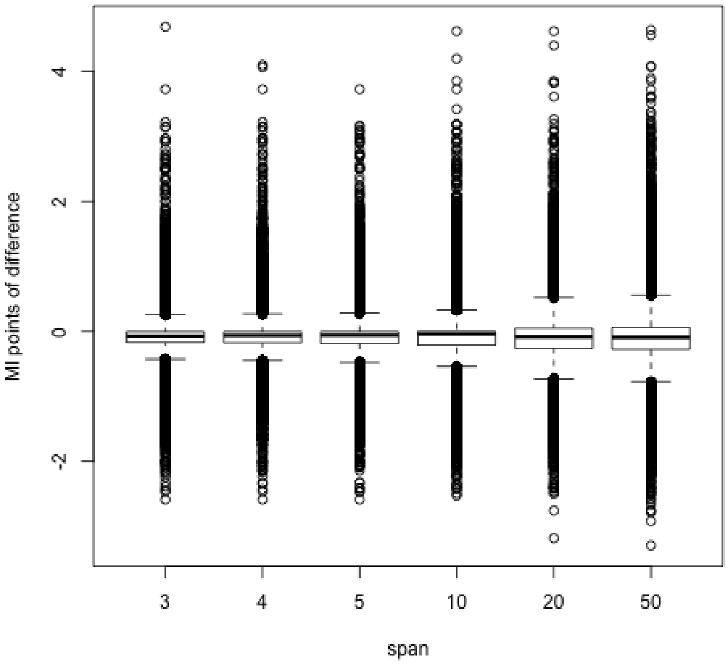


Figure 5.14 Spread of differences between Overproof and gold MI statistics



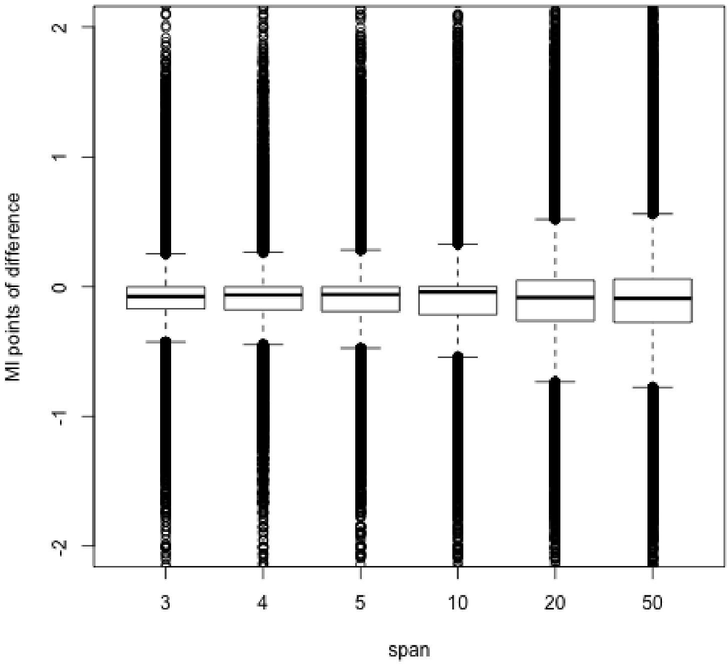


Figure 5.15 Probability that any given Overproof LL statistic will be situated at a given distance of the corresponding gold statistic (full extent)

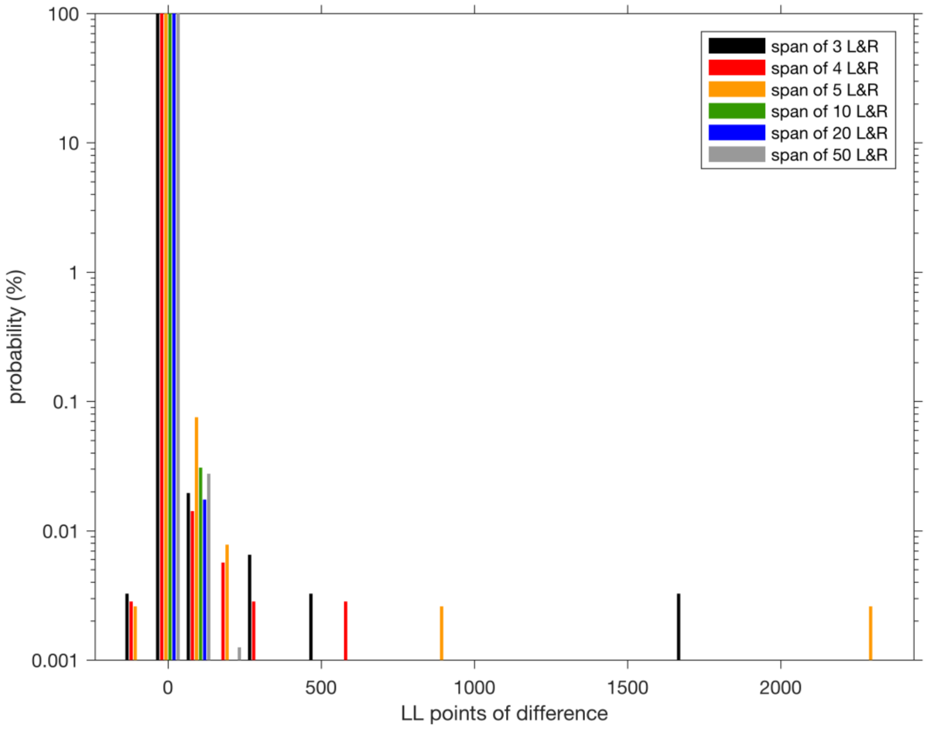


Figure 5.16 Probability that any given Overproof LL statistic will be situated at a given distance of the corresponding gold statistic (trimmed to -100, 100)

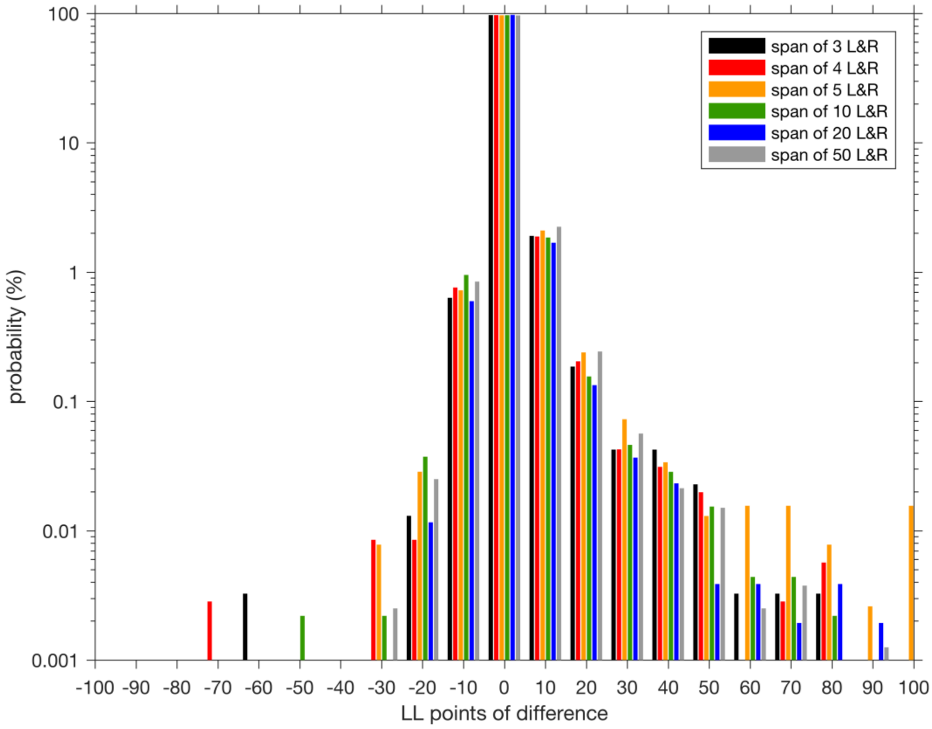


Figure 5.17 Probability that any given Overproof LL statistic will be situated at a given distance of the corresponding gold statistic (trimmed to -5, 5)

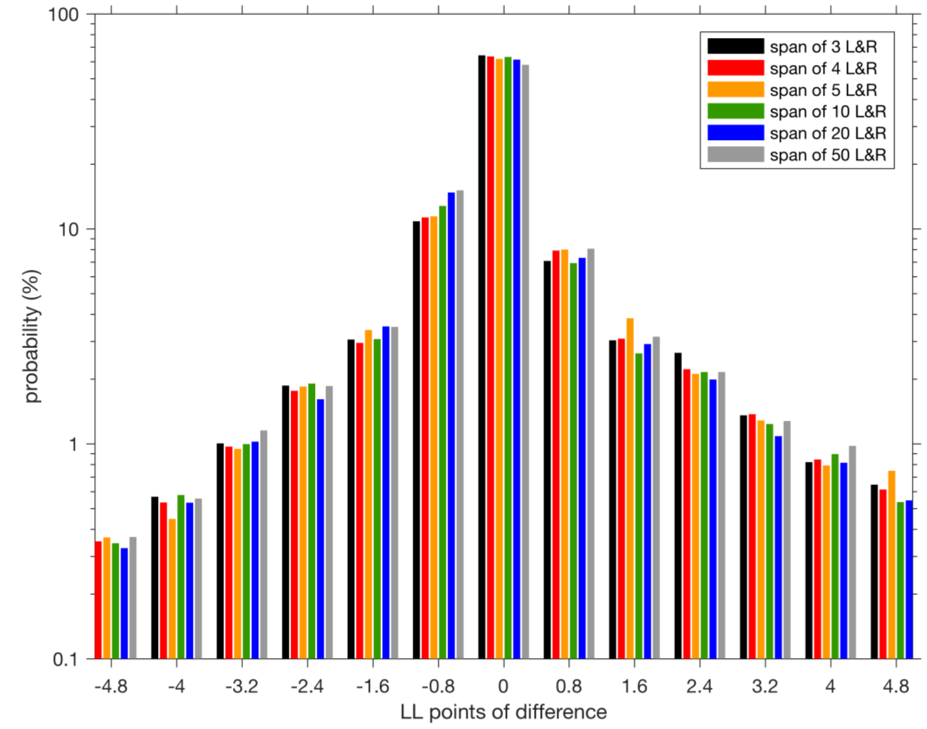
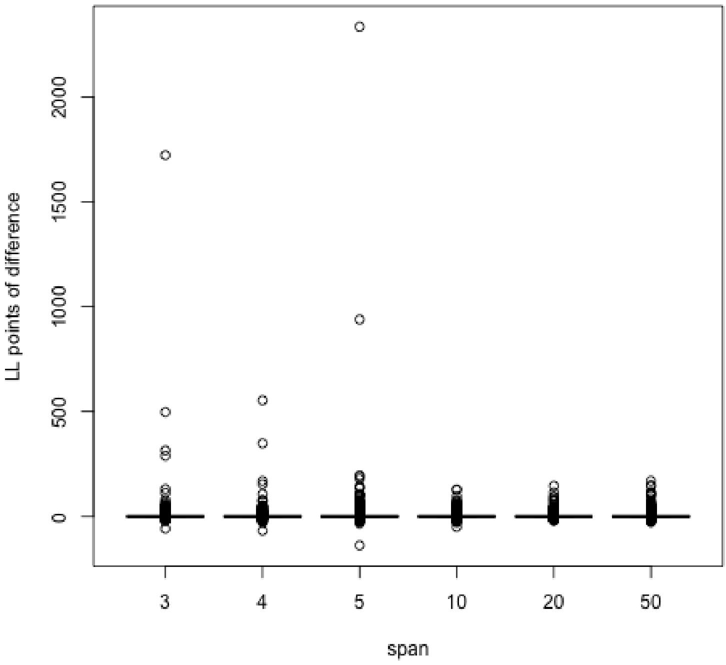


Figure 5.18 Spread of differences between Overproof LL statistics and gold LL statistics



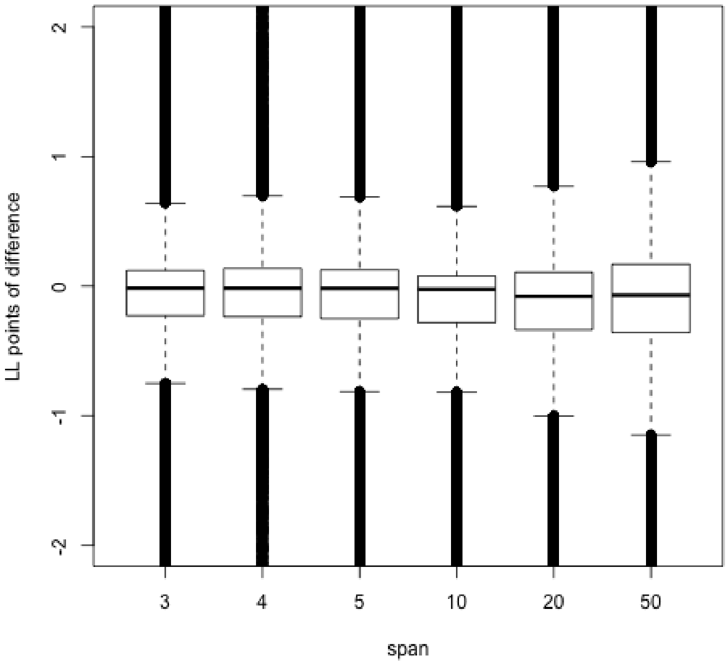


Figure 5.19 Probability that any given Overproof MI statistic will be situated at a given distance of the corresponding gold statistic, when considering only MI statistics for which the corresponding LL statistic is at least 10.83 (full extent)

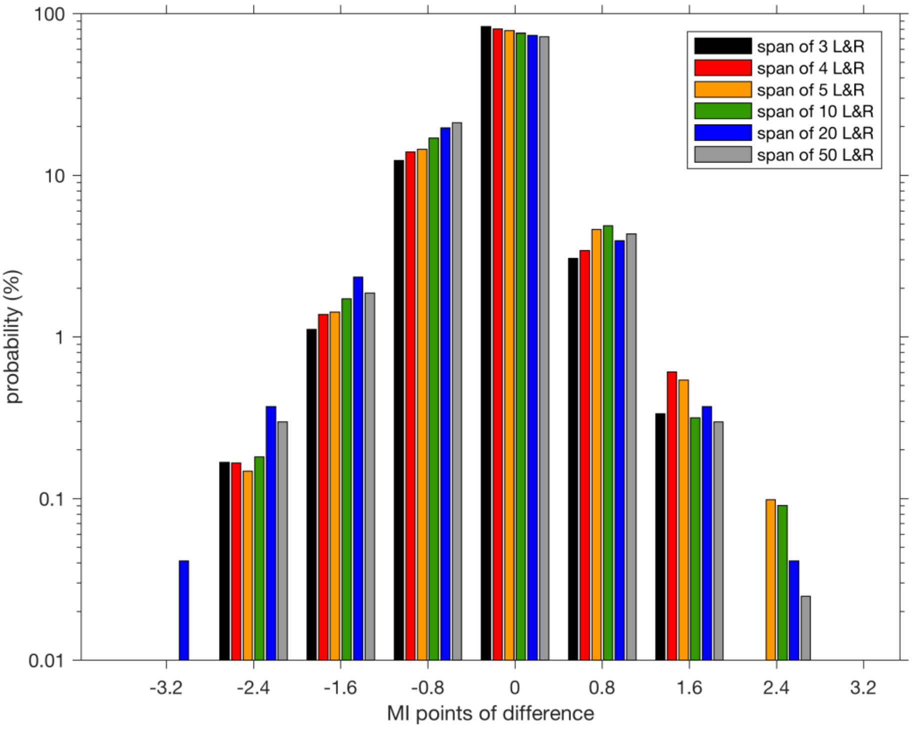
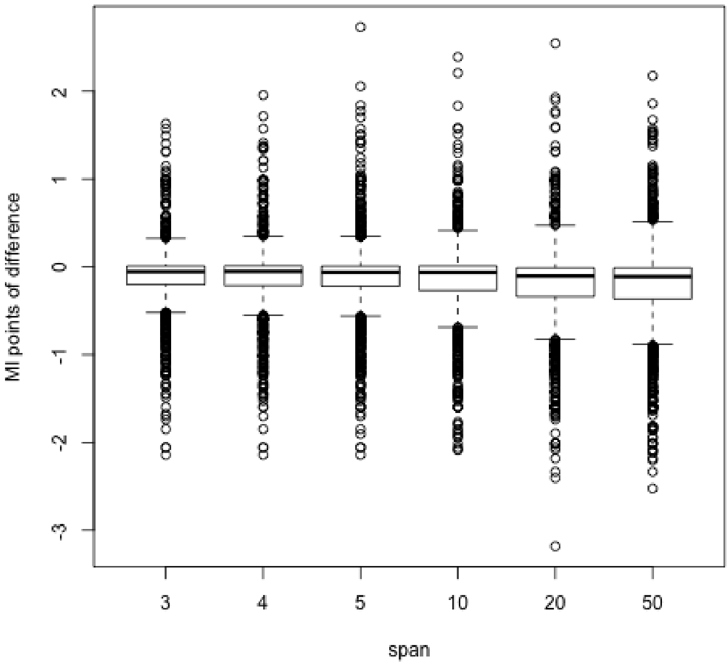


Figure 5.20 Spread of differences between Overproof and gold MI statistics, with an LL threshold of at least 10.83



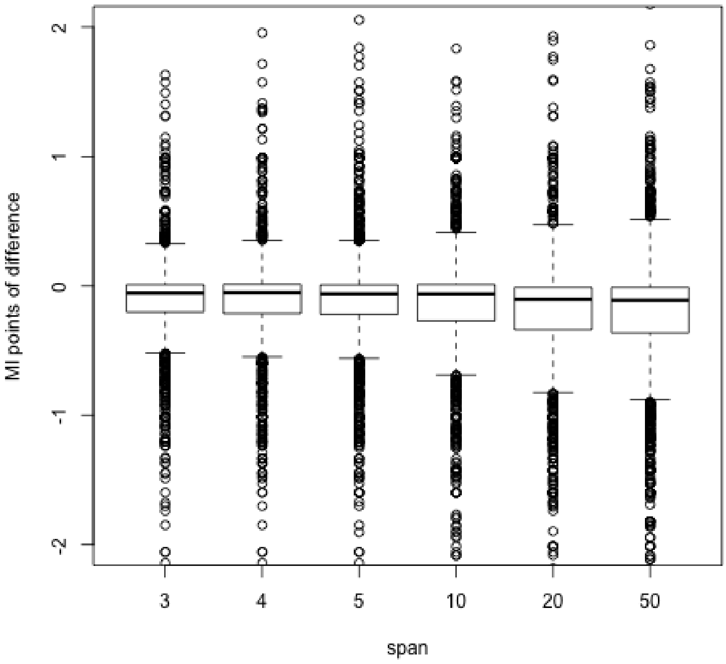


Figure 5.21 Probability that any given Overproof MI statistic will be situated at a given distance of the corresponding gold statistic, for node/collocate pairs above the frequency floor (full extent)

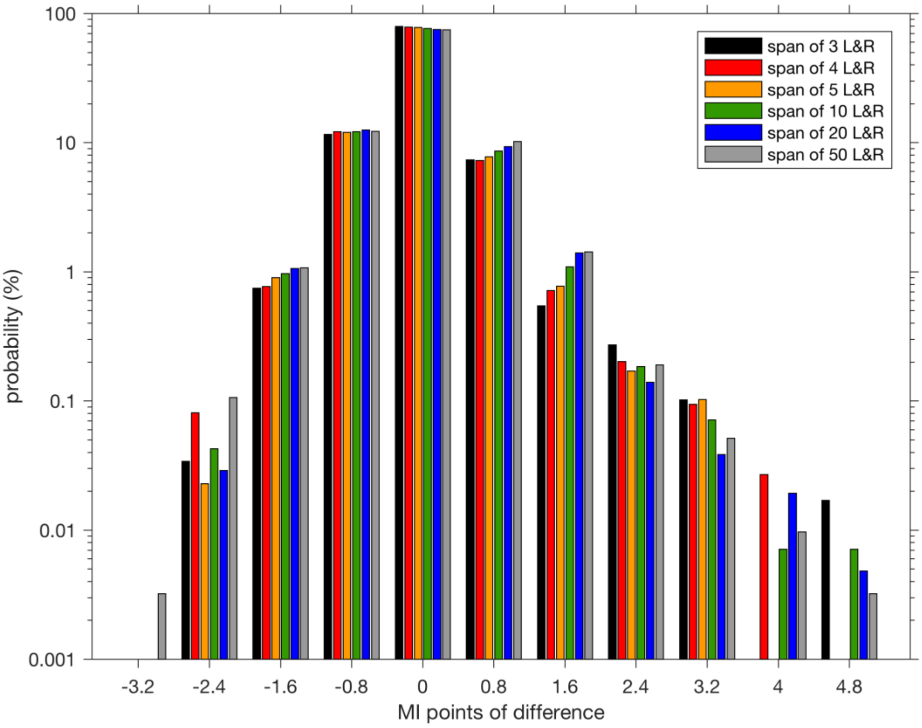
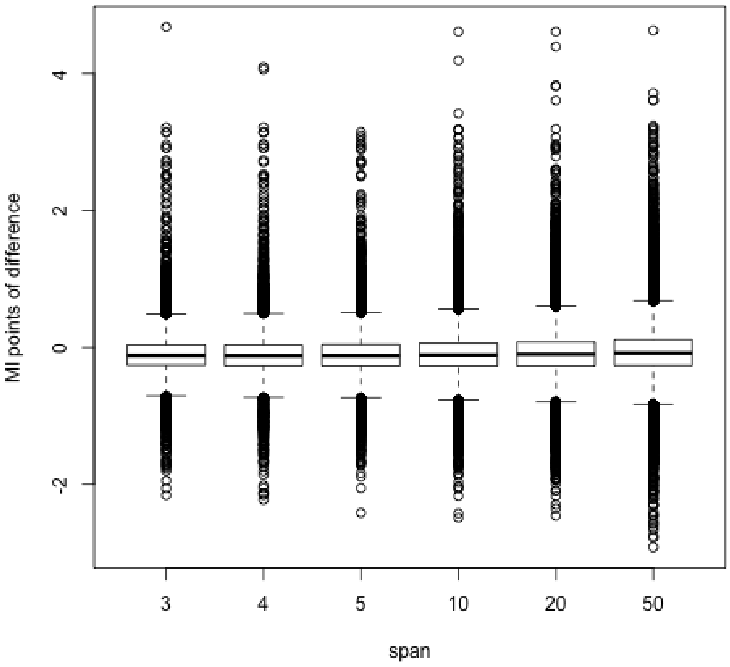


Figure 5.22 Spread of differences between Overproof and gold MI statistics, for node/collocate pairs above the frequency floor



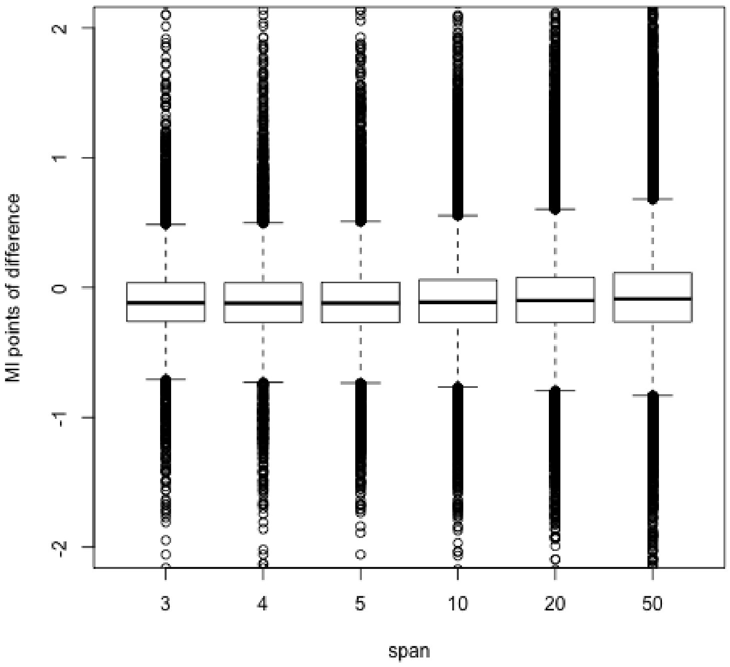


Figure 5.23 Probability that any given Overproof LL statistic will be situated at a given distance of the corresponding gold statistic, for node/collocate pairs above the frequency floor (full extent)

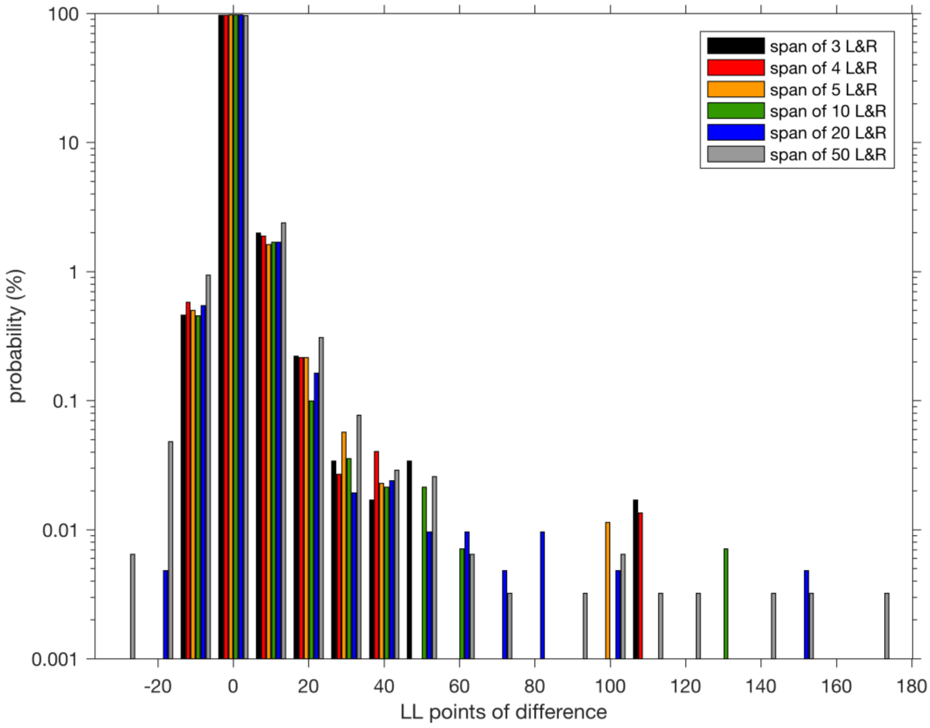


Figure 5.24 Probability that any given Overproof LL statistic will be situated at a given distance of the corresponding gold statistic, for node/collocate pairs above the frequency floor (trimmed to -100, 100)

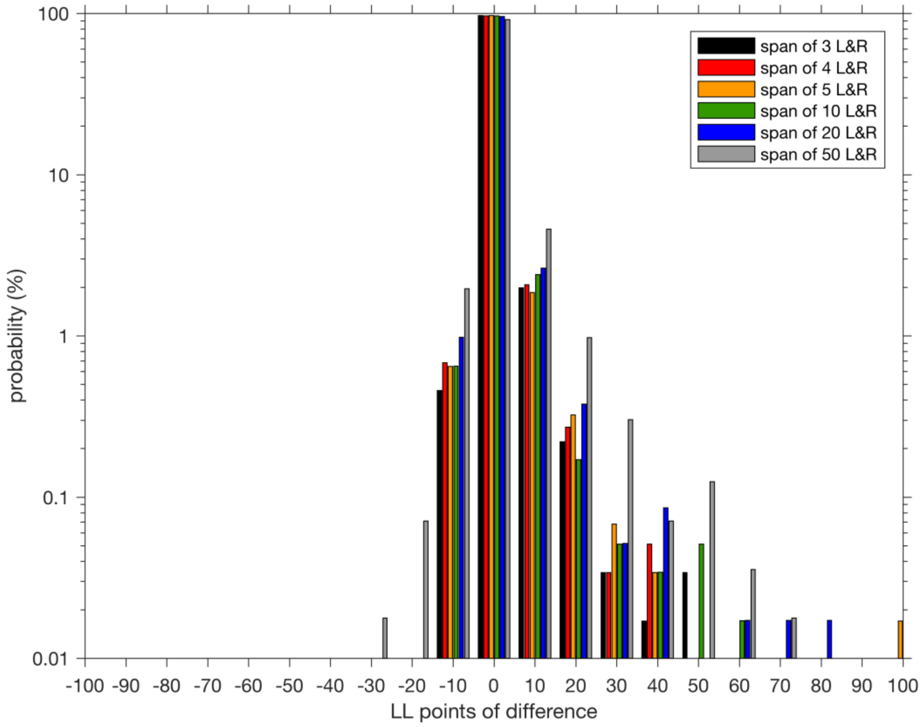


Figure 5.25 Probability that any given Overproof LL statistic will be situated at a given distance of the corresponding gold statistic, for node/collocate pairs above the frequency floor (trimmed to -5, 5)

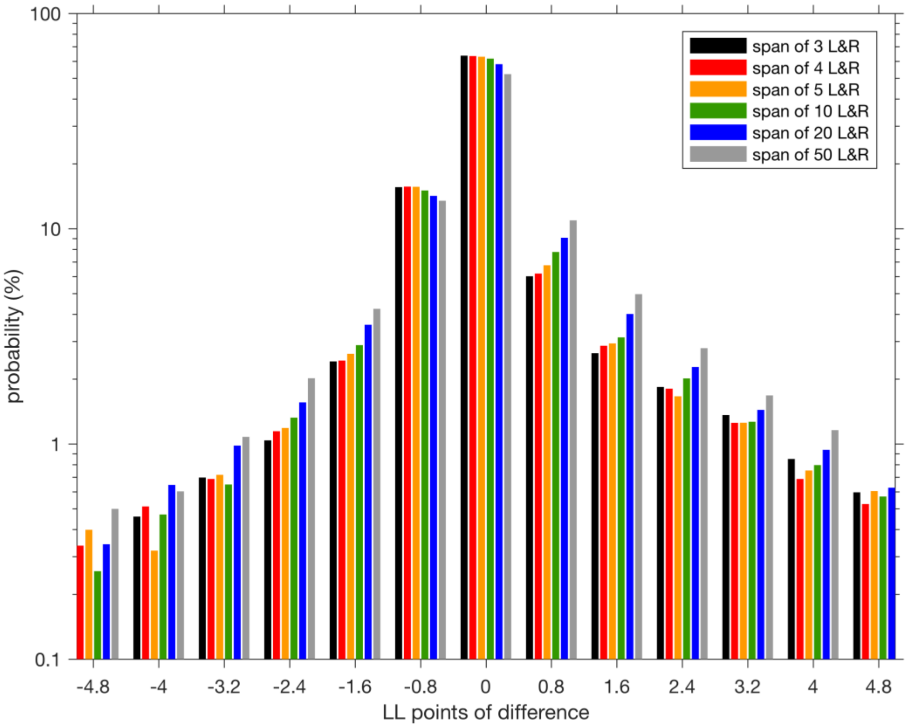
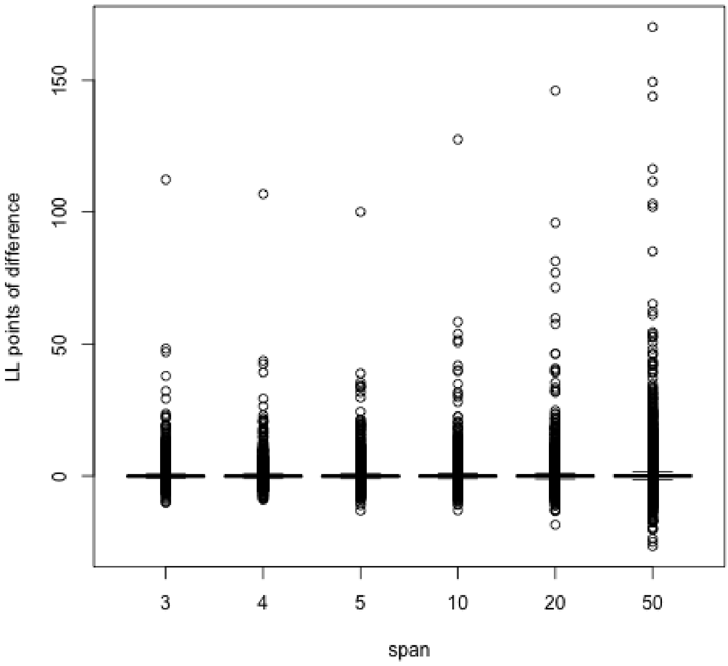


Figure 5.26 Spread of differences between Overproof and gold LL statistics, for node/collocate pairs above the frequency floor



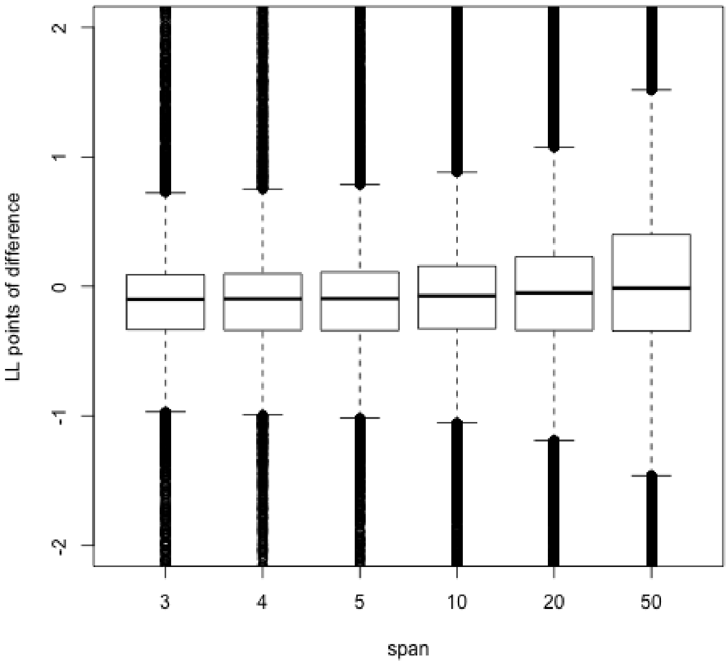


Figure 5.27 Probability that any given Overproof MI statistic will be situated at a given distance of the corresponding gold statistic, for node/collocate pairs above the frequency floor, with an LL threshold of at least 10.83 (full extent)

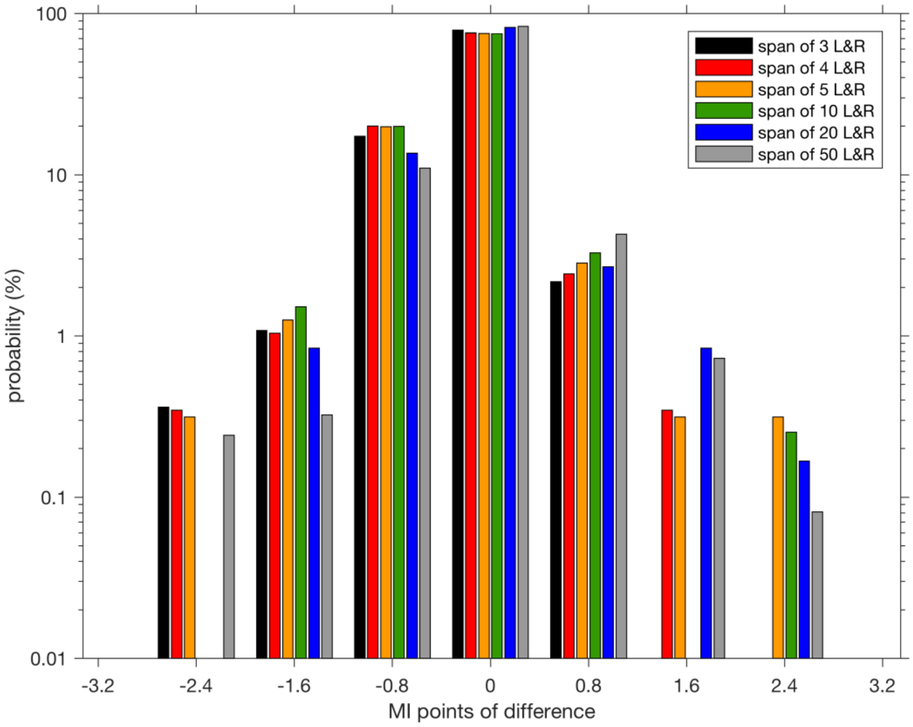
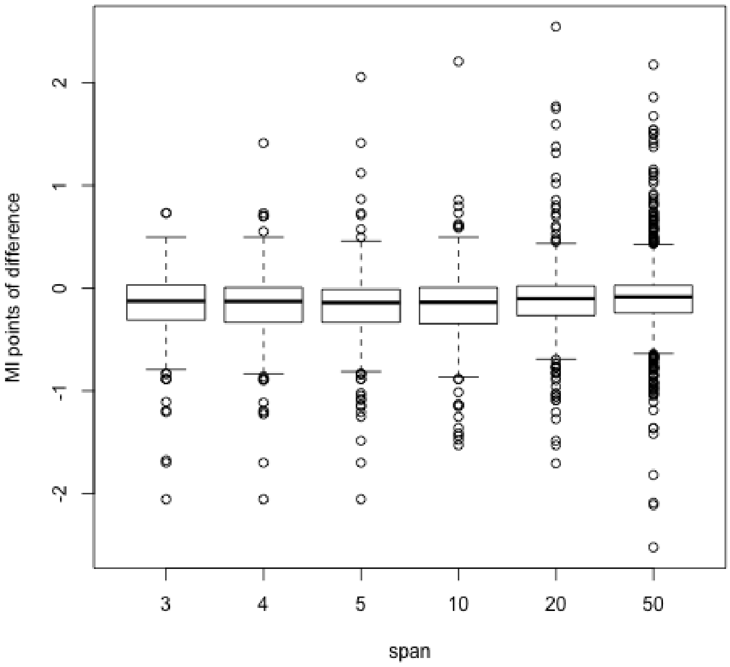
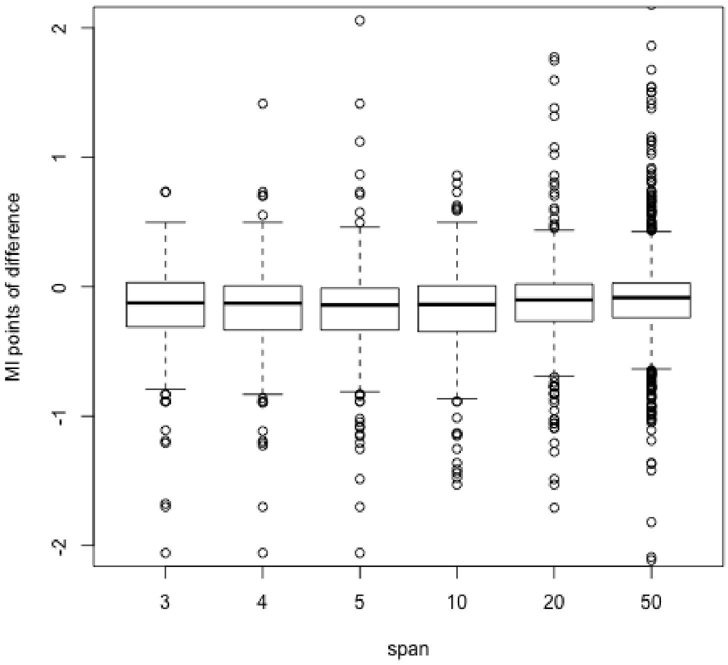


Figure 5.28 Spread of differences between Overproof and gold MI statistics, for node/collocate pairs above the frequency floor, with an LL threshold of at least 10.83





5.4.4.2 Average differences across word-types

For both MI and LL,

Figure 5.29 to 5.34 (as compared to Figures 4.23 to 4.28) show very similar trends overall. Nevertheless, the effect of Overproof correction is also visible here, with the spread of average values becoming somewhat narrower on all figures. Hence, in Overproof-corrected data, it remains true that the nodes in the frequency band 10-100 attract the most extreme average distances for both MI and LL.

Figure 5.29. Average distance between gold and Overproof MI statistic (absolute values), by frequency of node in gold corpus

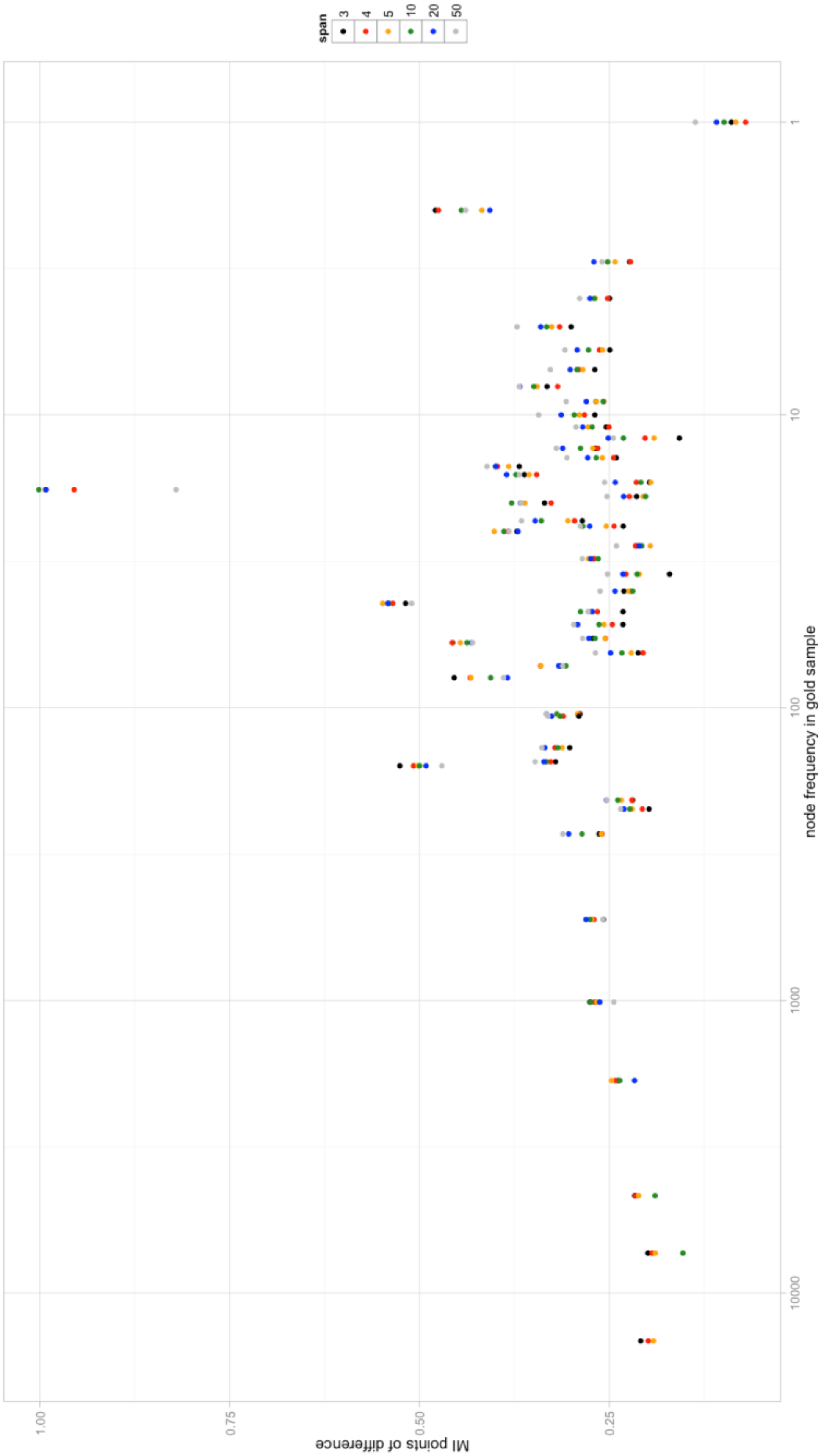


Figure 5.30. Average distance between gold and Overpoof MI statistic (for positive values only), by frequency of node in gold corpus

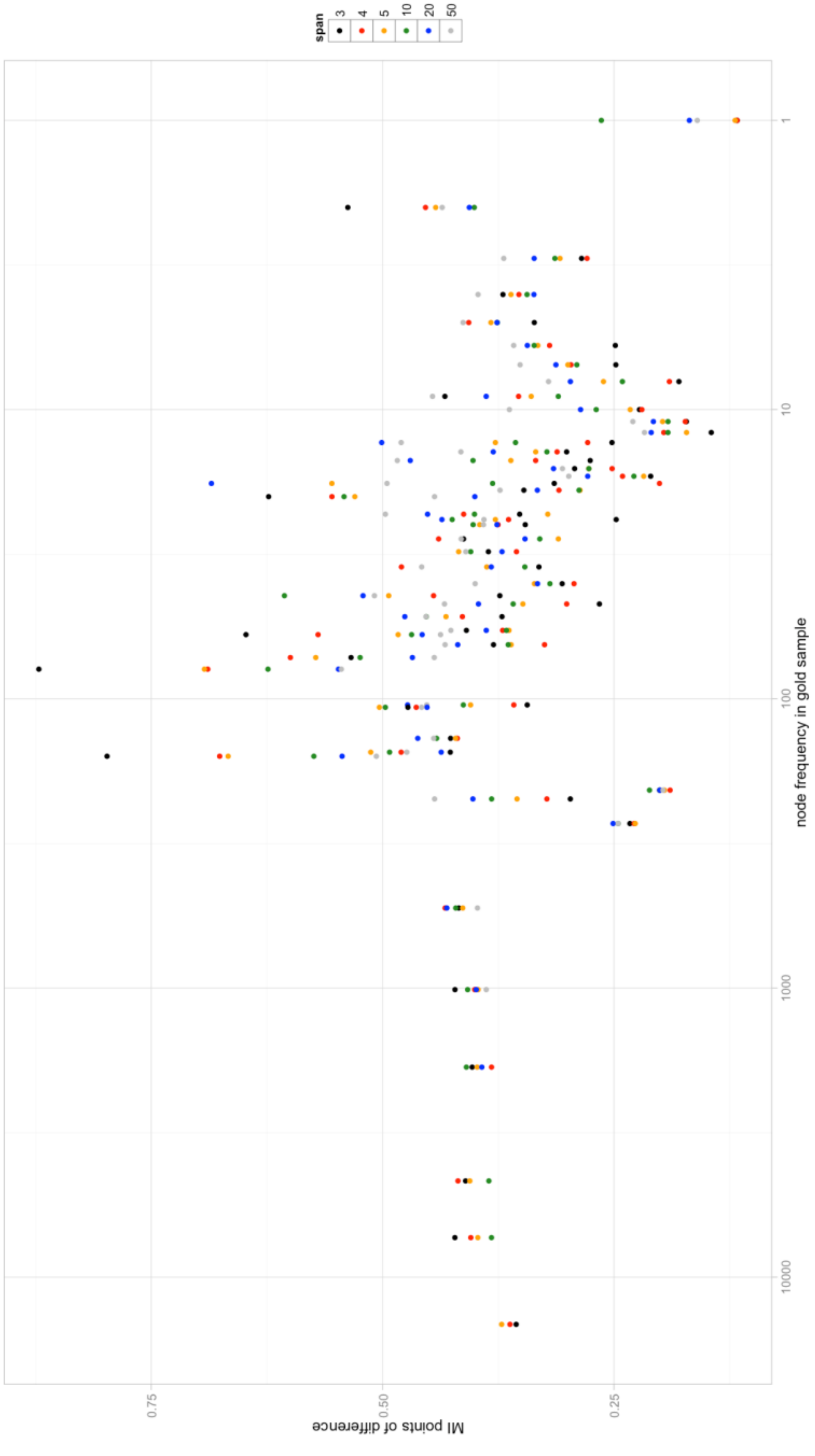


Figure 5.31. Average distance between gold and Overpoof MI statistic (for negative values only), by frequency of node in gold corpus

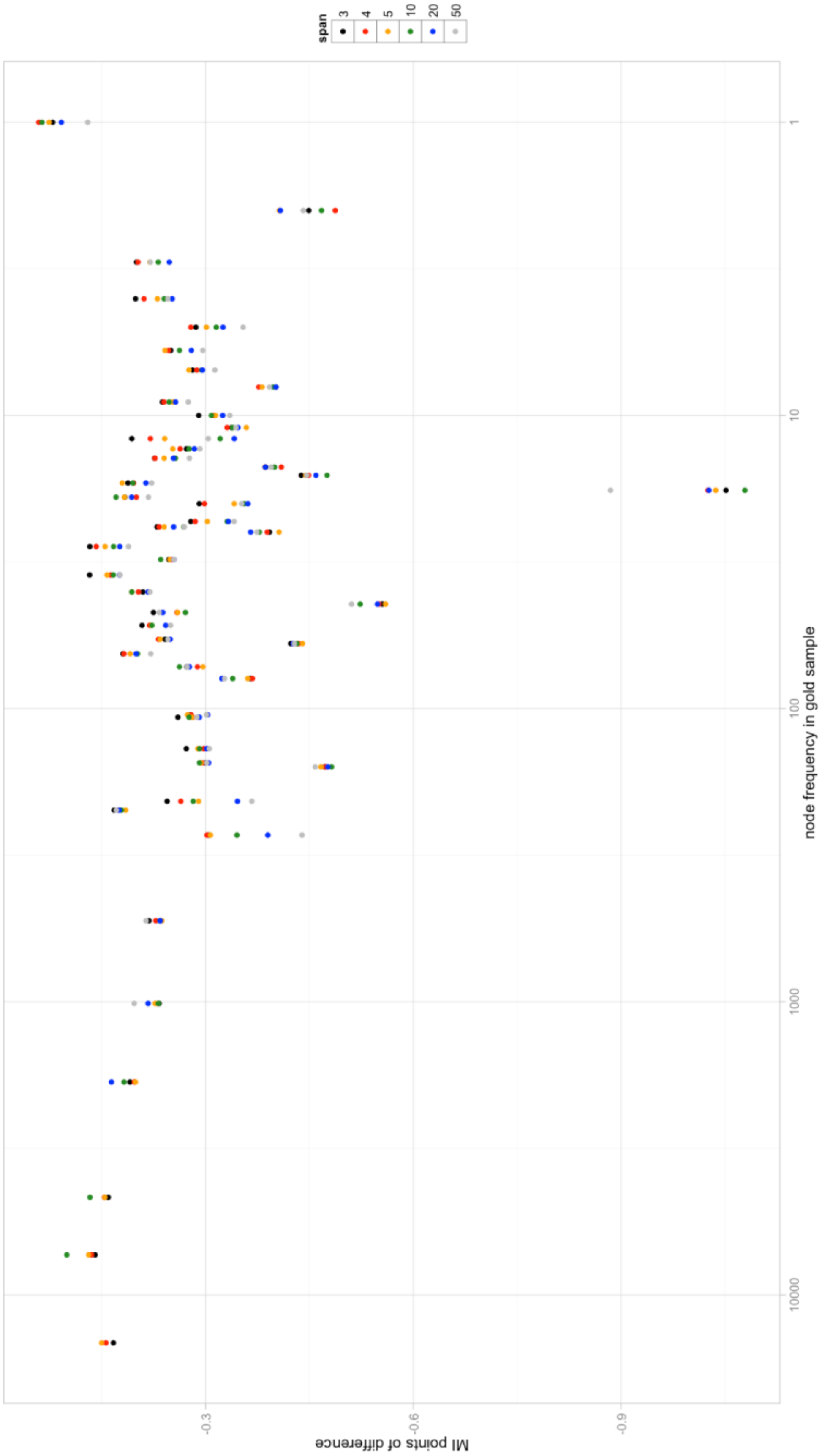


Figure 5.32. Average distance between gold and Overproof LL statistic (absolute values), by frequency of node in gold corpus

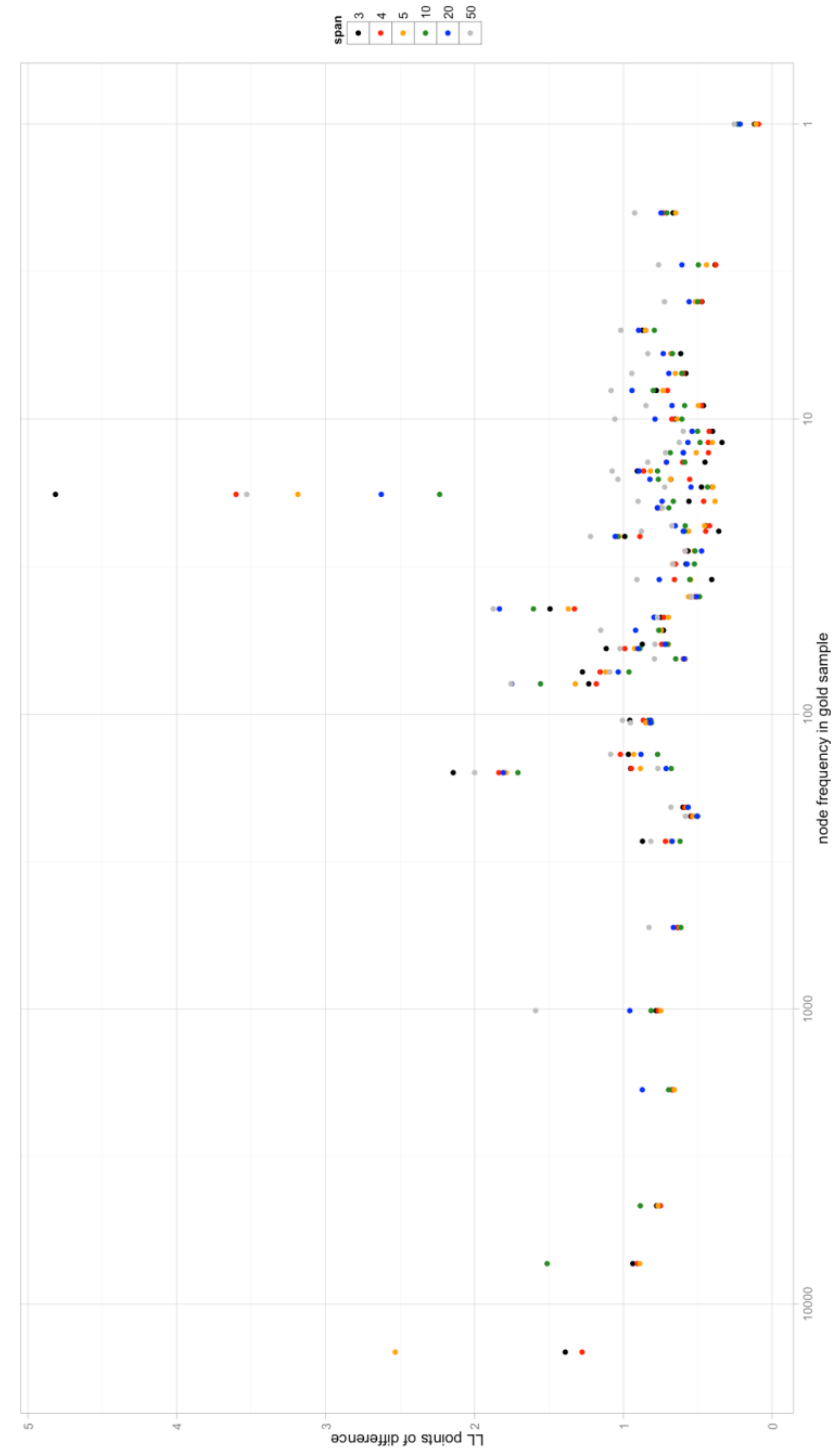


Figure 5.33. Average distance between gold and Overpoof LL statistic (for positive values only), by frequency of node in gold corpus

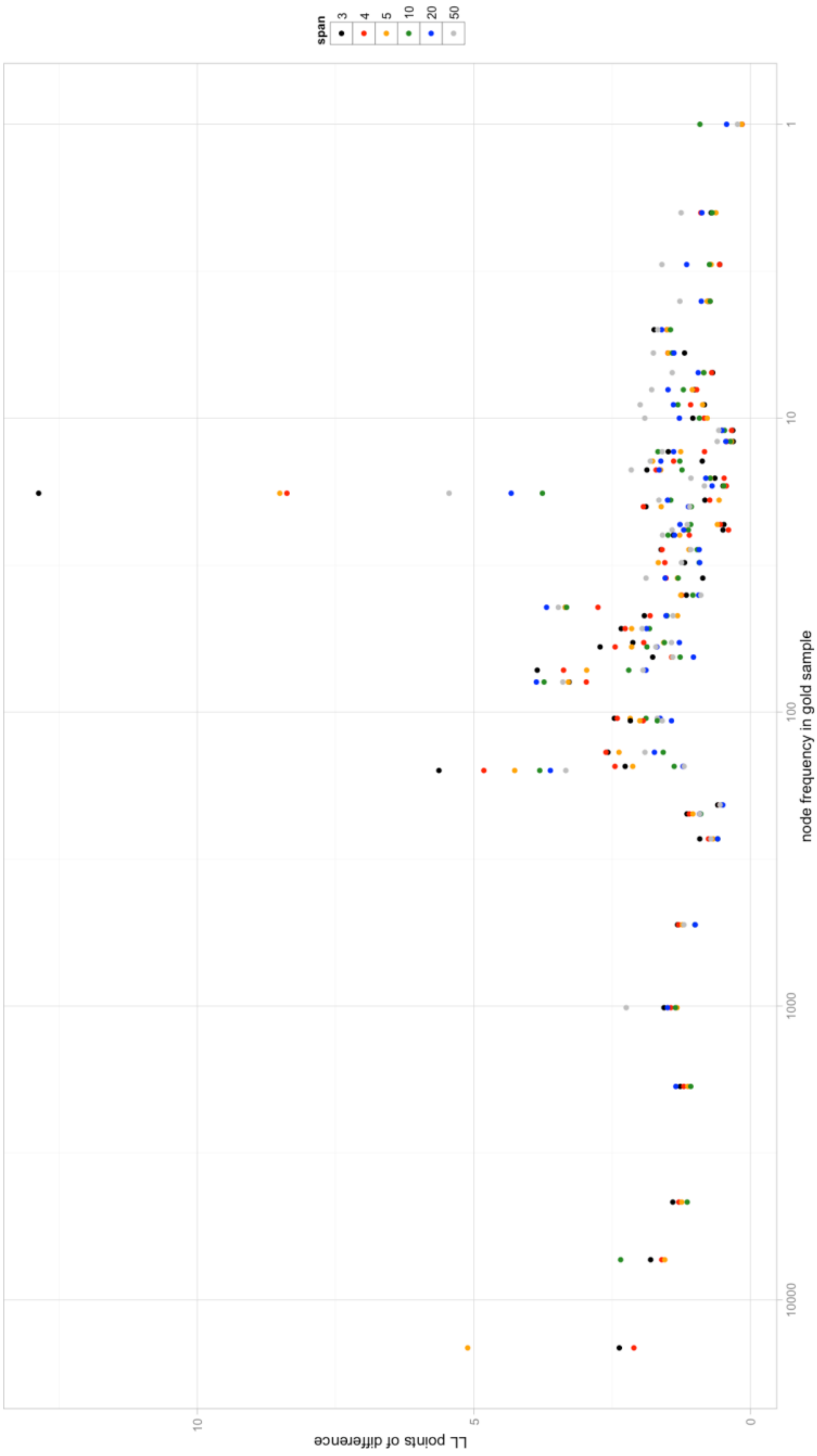
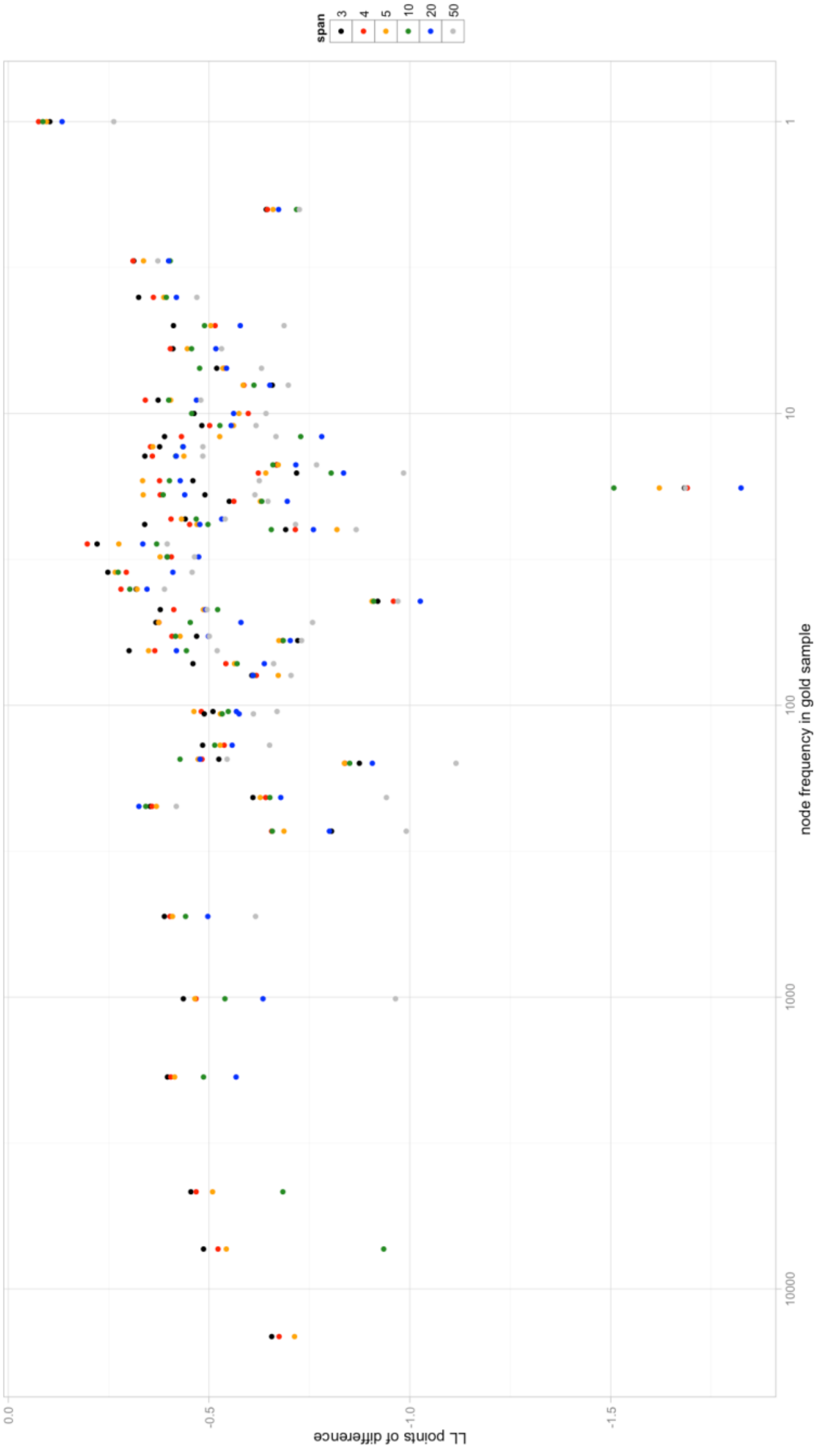


Figure 5.34. Average distance between gold and Overpoof LL statistic (for negative values only), by frequency of node in gold corpus



5.4.4.3 Conservation of ranking

In rankings too, we see a clear effect from the Overproof correction, with all coefficients increasing from their uncorrected OCR values (see Table 5.10, as compared to Table 4.9). The improvement is the most dramatic for the larger spans, however, which suggests that Overproof may be especially helpful for working with larger spans.

Table 5.10. Spearman's rank coefficient values for gold to Overproof MI and LL rankings

		(span)	3	4	5	10	20	50
no LL cutoff	no floor	MI	0.90	0.88	0.87	0.86	0.87	0.85
no LL cutoff	no floor	LL	0.90	0.88	0.86	0.85	0.87	0.83
no LL cutoff	no floor	**N**	**90906**	**104320**	**113669**	**133937**	**150628**	**229346**
above LL cutoff	no floor	MI/LL	0.85	0.84	0.79	0.79	0.78	0.66
above LL cutoff	no floor	**N(MI/LL)**	**1577**	**1569**	**1699**	**1829**	**2023**	**3309**
		(span)	3	4	5	10	20	50
no LL cutoff	above floor	MI	0.88	0.86	0.84	0.78	0.73	0.65
no LL cutoff	above floor	LL	0.88	0.85	0.83	0.77	0.69	0.52
no LL cutoff	above floor	**N**	**5748**	**7252**	**8568**	**13712**	**20296**	**30542**
above LL cutoff	above floor	MI/LL	0.89	0.88	0.84	0.81	0.73	0.41
above LL cutoff	above floor	**N(MI/LL)**	**235**	**238**	**245**	**313**	**466**	**996**
		(span)	3	4	5	10	20	50
no LL cutoff	below floor	MI	0.87	0.84	0.82	0.81	0.82	0.74
no LL cutoff	below floor	LL	0.86	0.83	0.81	0.80	0.82	0.74
no LL cutoff	below floor	**N**	**24445**	**27381**	**29123**	**30462**	**29232**	**46069**
above LL cutoff	below floor	MI/LL	0.83	0.82	0.77	0.77	0.78	0.71
above LL cutoff	below floor	**N(MI/LL)**	**1308**	**1298**	**1414**	**1465**	**1483**	**2240**

5.4.4.4 Rates of false positives and negatives

Table 5.11 (as compared to Table 4.10) shows that Overproof also leads to improvements in the rates of false positives and false negatives. The rate of MI false positives approximately halves under all conditions, both in relation to positive observations and to overall observations. For LL, the reduction in false positives is less dramatic, but nevertheless impressive; furthermore, the reduction is more dramatic for the larger spans, almost halving at a span of 50. This suggests again (see also section 5.4.4.3) that using Overproof for OCR post-correction may be especially helpful for researchers wishing to work with large spans. The rates of false negatives for MI increase slightly under all conditions, but the increase is very slight, and the rates remain very small (usually less than 1% both relative to negative observations and to all observations) in all conditions. The rates of false negatives for LL improve somewhat, but the effect is very small, and the rates are small to begin with, so remain of little concern.

Table 5.11 Percentage rates of false positives and false negatives in the Overproof-corrected sample

		(span)	3	4	5	10	20	50
no LL cutoff	no floor	**MI false positives, per total positives**	4.66	6.25	9.08	11.66	11.19	8.60
no LL cutoff	no floor	**MI false positives, per total observations**	0.86	1.23	1.93	2.98	3.53	2.41
no LL cutoff	no floor	**MI false negatives, per total negatives**	0.49	0.48	0.45	0.69	1.18	1.40
no LL cutoff	no floor	**MI false negatives, per total observations**	0.40	0.38	0.36	0.52	0.81	1.01
no LL cutoff	no floor	**LL false positives, per total positives**	12.15	13.51	16.60	17.36	16.93	17.63
no LL cutoff	no floor	**LL false positives, per total observations**	0.71	0.69	0.88	0.84	0.80	0.89
no LL cutoff	no floor	**LL false negatives, per total negatives**	1.19	1.22	1.37	1.24	1.26	1.70
no LL cutoff	no floor	**LL false negatives, per total observations**	1.12	1.16	1.30	1.18	1.20	1.62
above LL cutoff	no floor	**MI false positives, per total positives**	1.34	1.50	1.77	2.55	2.35	3.33
above LL cutoff	no floor	**MI false positives, per total observations**	0.67	0.83	0.98	1.63	1.89	2.54
above LL cutoff	no floor	**MI false negatives, per total negatives**	0.56	0.37	0.11	0.63	1.47	1.26
above LL cutoff	no floor	**MI false negatives, per total observations**	0.28	0.17	0.05	0.23	0.29	0.30
no LL cutoff	no floor	**PROPORTION MI POSITIVES**	18.54	19.72	21.26	25.53	31.58	28.04
no LL cutoff	no floor	**PROPORTION LL POSITIVES**	5.85	5.14	5.29	4.86	4.71	5.03
above LL cutoff	no floor	**PROPORTION MI POSITIVES**	50.00	55.27	55.45	63.88	80.48	76.20

		(span)	3	4	5	10	20	50
no LL cutoff	above floor	**MI false positives, per total positives**	7.58	10.07	11.52	14.55	16.87	19.93
no LL cutoff	above floor	**MI false positives, per total observations**	2.45	2.94	3.08	2.78	2.15	1.11
no LL cutoff	above floor	**MI false negatives, per total negatives**	1.61	1.46	1.21	1.24	1.08	0.95
no LL cutoff	above floor	**MI false negatives, per total observations**	1.09	1.04	0.89	1.00	0.94	0.90
no LL cutoff	above floor	**LL false positives, per total positives**	11.91	15.22	19.81	17.68	18.79	18.58
no LL cutoff	above floor	**LL false positives, per total observations**	0.56	0.59	0.72	0.50	0.54	0.74
no LL cutoff	above floor	**LL false negatives, per total negatives**	1.45	1.28	1.15	1.00	1.07	1.65
no LL cutoff	above floor	**LL false negatives, per total observations**	1.38	1.23	1.11	0.97	1.03	1.59
above LL cutoff	above floor	**MI false positives, per total positives**	3.66	5.50	5.12	7.51	10.34	12.44
above LL cutoff	above floor	**MI false positives, per total observations**	2.53	3.81	3.46	4.80	5.54	4.04
above LL cutoff	above floor	**MI false negatives, per total negatives**	3.49	2.25	0.97	3.50	2.17	1.20
above LL cutoff	above floor	**MI false negatives, per total observations**	1.08	0.69	0.31	1.26	1.01	0.81
no LL cutoff	above floor	**PROPORTION MI POSITIVES**	32.30	29.14	26.71	19.07	12.75	5.55
no LL cutoff	above floor	**PROPORTION LL POSITIVES**	4.71	3.89	3.62	2.82	2.87	3.98
above LL cutoff	above floor	**PROPORTION MI POSITIVES**	68.95	69.20	67.61	63.89	53.52	32.47

5.5 Summary and Conclusion

The dataset used in this thesis is the *19th Century British newspapers* collection (part 1) owned by the British Library. The data was provided to the *Spatial Humanities* project by the

British Library in the form of OCR output. OCR, however, is of variable effectiveness on historical material. Tanner et al. (2009) report an average word accuracy rate of 78%[8] for this collection. The central question that was asked in this chapter is: is it possible for the project to correct these errors? OCR post-correction methods exist, but levels of effectiveness vary. First, I experimented with VARD, software developed to deal with spelling variation in Early Modern English. This option appeared promising because it involved letter replacement rules being derived from the manual correction of a sample of data from the collection. Letter replacement rules describe regularities in the corrections which have been made to a training sample; they identify single characters or clusters of characters in the original dataset and the single character or cluster of characters which tend to replace them in the manually corrected sample. Since OCR is an automated process, it would be expected to generate errors according to predictable patterns, patterns which, it was hoped, might be captured by these letter replacement rules.

A preliminary sample of the newspaper collection was constructed by selecting a random page per year and a random page per publication; the final training sample consisted of a random subset of this preliminary sample and amounted to around 56,400 words yielding 5,640 corrections. Although only the most straightforward errors could be corrected during the manual training process, analysis of the edit-distance separating the errors in the original from their manual corrections revealed that OCR errors are more complex than the natural historical spelling variation discussed by Baron (2011). In addition, a high number of letter replacement rules were abstracted from the training phase, most of which (74%) applied only once, revealing a high degree of variation in the types of errors present in the data. Together, these findings reveal that the task of automatically correcting OCR errors is a challenging one, and that an approach based on letter replacement rules alone will not be effective.

[8] Note that Tanner et al. (2009)'s figure is an assessment of the *maximal* quality of OCR in the collection and is not intended to be understood as representative of the average quality of the collection, see sections 3.3.2 and 4.4.2.

To assess the effectiveness of the corrective procedure involving a trained version of VARD, a gold sample of around 160,000 words was put together constituted by articles from CNNE which could be matched to the OCR data in my possession; the matched portions of our data were then corrected using different setups in VARD. These different versions of the same articles - collectively, the CNNE matching corpus - were compared using the OCReval tool. Most uncorrected OCR files in the CNNE matching corpus were found to have file edit-distances between 10% and 30%, although the files ranged from 3.63% to 97.43%; this was an adequate range to assess the effectiveness of the corrective procedure. Incidentally, the range of OCR quality exhibited by the files in the CNNE matching corpus also suggested that images 'readable' to a human (which had thus 'made it' into the CNNE corpus) may still attract a broad range of variation of OCR quality.

As expected, the different corrective setups used in VARD produced different numbers of changes, with the longer list of rules producing fewer changes, and likewise the higher confidence thresholds. In terms of the effectiveness of the changes made to the uncorrected portion of the CNNE matching corpus, virtually no difference was found between the different corrective setups. The corrective procedure basically yielded no improvement overall, with most files attracting minimal improvement or deterioration, and some exceptional files attracting considerable deterioration. From this, it is clear that a corrective approach involving VARD - or letter replacement rules on their own - will not be effective for a large amount of mixed OCR data, at least without a substantial investment in time spent training VARD.

Since VARD proved unpromising, I tested Overproof, a state-of-the-art commercial system which, among other heuristics not used by VARD, uses *reverse OCR* – a technique consisting in comparing bad quality images of errors and words in the lexicon. Overproof was tested on the same dataset as VARD. Overproof was clearly successful in improving the quality of the files, with all corrected files ending up more similar to the gold files compared to the uncorrected files, and with the edit-distance between the corrected and gold files being on average 8.45 file edit-distance closer than the distance between the uncorrected and gold files. The Overproof corrected type/token ratio was also substantially closer to the gold one than the uncorrected one, with the number of hapaxes halving in the Overproof corrected version. However, little improvement was shown in the proportion of gold types occurring at least 10 times in the Overproof corpus, suggesting that even in this corrective setup, the loss in lexical diversity in OCR data is important. And whilst the Overproof corrected version has a smaller proportion of types with over-estimated frequencies than the uncorrected version, this difference is especially due to improvements in the low frequency items. Moreover the figures suggested that real-word errors could be expected to be more of a problem in the Overproof corrected version than in the uncorrected version. If the Overproof-corrected version contains more real-word errors than the uncorrected version, considering the impact of the corrections on the reliability of the collocation statistics becomes crucial. Comparing the difference between the gold and Overproof statistics and the gold and uncorrected OCR statistics reveals that Overproof corrections have a substantial positive impact on MI and LL, with improvements across the board, most dramatically for large spans and smaller frequencies (where, admittedly, there is most room for improvement). In fact, the improvement to *over*-estimates is strong for both MI and LL, as was shown by all the methods used: MI and LL statistics are substantially more reliable both in ranking terms, and in terms of false positives, in Overproof-corrected OCR data than in uncorrected data.

PART 3: INVESTIGATING DISCOURSES SURROUNDING PLACE-NAMES

6 FINDING OUT WHAT PLACES ARE MENTIONED IN A CORPUS

6.1 INTRODUCTION

One of the major aims of this thesis is to explore potential methodologies for the investigation of discourses surrounding place-names in large collections of text. The discussion centres around two main questions: 'what places are mentioned in this corpus?' and 'what is said about these places?'. This chapter deals with the former question; the next chapter deals with the latter. In both chapters, various approaches to answering the question will be explored and their strengths and weaknesses discussed, with a particular focus on their potential to be used comparatively (comparing time-periods and genres) and their scalability (potential to be used with large amounts of text).

Section 6.2 briefly introduces the three newspapers chosen as case-studies for this chapter and the next. Section 6.3 discusses approaches towards answering the question 'what places are mentioned in this corpus?'. Section 6.4 presents some findings obtained using the approaches described in section 6.3. The discussion in this chapter will focus on names of cities; in the next chapter, the focus will shift to names of countries.

6.2 THREE NEWSPAPERS

The discussions in this chapter and the next revolve around three newspapers chosen as case-studies. Their choice was motivated by several factors but is ultimately arbitrary: a choice had to be made, since working with the entirety of the *19th Century British Newspapers* collection was not feasible for this thesis. The first newspaper was chosen purely for pragmatic reasons: *The Era* (ERLN) was chosen because it initially appeared to have a high OCR accuracy. Unfortunately, this decision was based on a word recognition rate calculated for the year 1900 only. However, the rate of OCR errors changes over the time-period, and it changes differently for each of the newspapers in the BL collection. Hence, although ERLN had the highest word recognition rate of the collection for 1900 according to our calculations, it does not actually have the best overall OCR quality of the collection (see Tanner et al. 2009 for some average figures for the newspapers in the BL collection).

I subsequently chose two further newspapers for their suitability as comparison-points with ERLN, for their historical interest, and for their relatively good OCR rates as determined based on the figures provided in Tanner et al. (2009). The two newspapers chosen, *The Pall Mall Gazette* (PMGZ) and *Reynold's Weekly Newspaper* (RDNP) are already relatively well-known in the community of historical scholars. I considered this an advantage; since the methods explored here are new in the field of history, working with sources which are already well-known could help highlight the novel aspects of these methods.

This trio of newspapers are all national papers published in the South of England. They were all published contemporaneously for at least the last three decades of the nineteenth century. Moreover they provide an interesting social contrast: PMGZ describes itself as an 'intellectual' paper, whereas RDNP attracts a 'working-class' readership. ERLN can be considered a 'trades' paper given its association with the Licensed Victuallers' Association (see sections 6.2.1 to 6.2.3).

The data for all three newspapers was acquired from Gale/Cengage. This version of the data includes a layer of annotation which separates portions of texts into 'articles' and classifies these articles into article genres. There is no published documentation on how this classification was devised and applied, or commentary on its reliability. Although the classification hence needs to be taken with a pinch of salt, its categories ('adverts', 'news', 'commerce', 'crime', 'arts', 'birth, death, etc.', 'sports' and 'illustrations') seem general enough to provide some useful and reasonably reliable indication of content. All three newspapers were further annotated with POS tags and semantic tags and indexed for use in CQPweb; see also section 3.2. In the next subsections, I provide some basic information on each of the newspapers.

6.2.1 THE PALL MALL GAZETTE (PMGZ)

The *Pall Mall Gazette* (PMGZ), which we have from its first publication in 1865 until 1900[1], was 'a bold attempt to realize Thackeray's fancy of a paper "written by gentlemen for gentlemen" ' (Bourne, H. R. Fox, cited in British Library n.d.: 274). A daily evening paper, it originally contained 8 pages of 2 columns and later expanded to 10 pages of 3 columns. The paper started off with Tory allegiances and became increasingly so aligned under its first editor, Frederic Greenwood. It later became 'a leading liberal paper' when the ownership passed to Henry Yates Thompson in the 1880s (Kent 2009). The paper was said to have had 'an influence out of all proportion to its modest circulation' (British Library n.d.) with Cranfield citing a regular circulation figure of 8,360 per day (Cranfield, cited in British Library n.d.).

The nearly 35 years of data amount to over 468 million words from 10,620 issues. Issues contained mostly 'News' (making up on average 52% of an issue's wordcount) and 'Adverts' (25%), followed by 'Commerce' and 'Arts' (8% each) (Figure 6.1). PMGZ's overall word-recognition score as given by Tanner et al. (2009) is ~94% character accuracy rate, ~90% word accuracy rate, ~85% significant word accuracy rate and ~79% capitalized word accuracy rate. Figure 6.4 shows the first page of the very first issue; Figure 6.1 and Figure 6.2 provide visualizations of the generic make-up of PMGZ over time; Figure 6.3 shows the average wordcount per issue over time.

[1] Except for the year 1872, which was missing from the electronic archive received from the British Library and Gale/Cengage.

Figure 6.1 Generic make-up of PMGZ (proportion of overall wordcount per year in each article category)

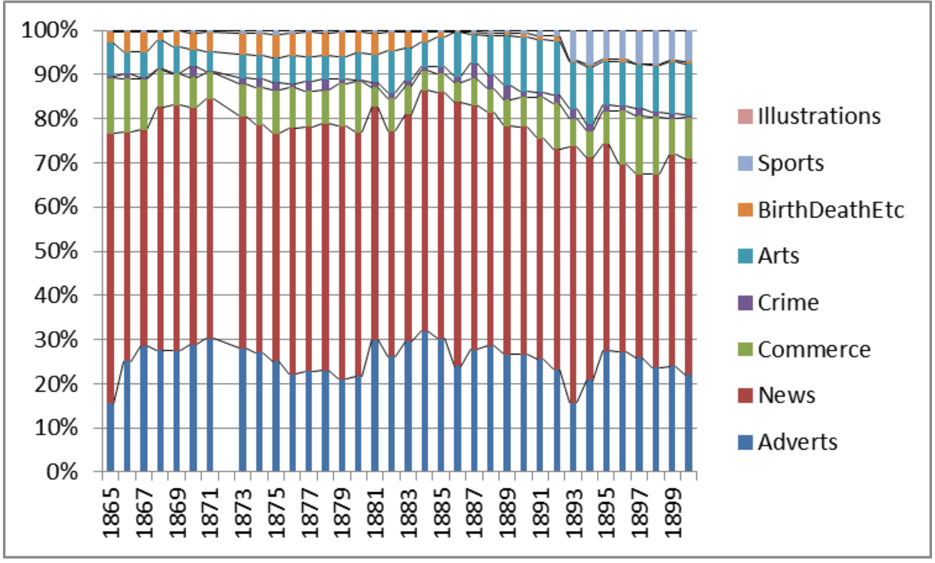


Figure 6.2 Generic make-up of PMGZ (raw wordcount in each article category per year)

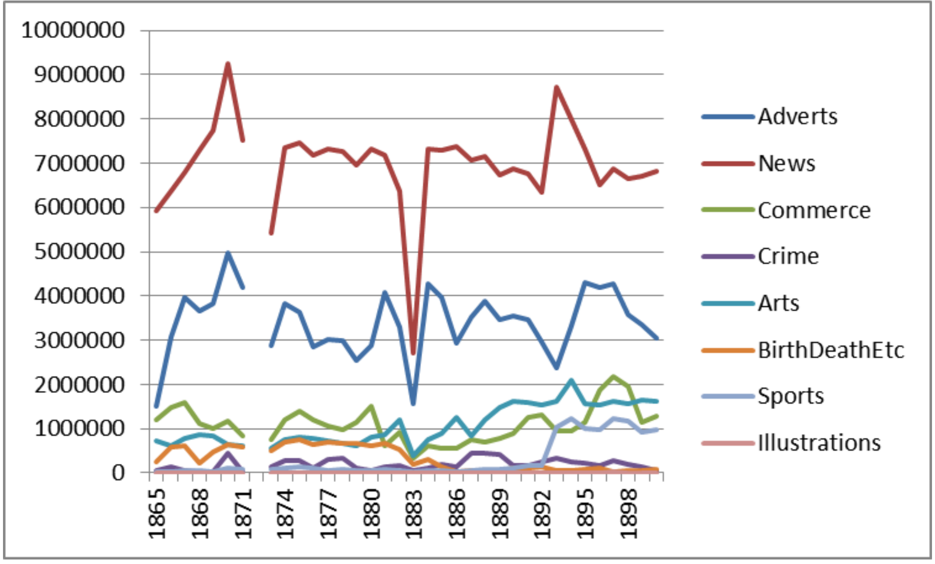


Figure 6.3 Size of PMGZ issues over time

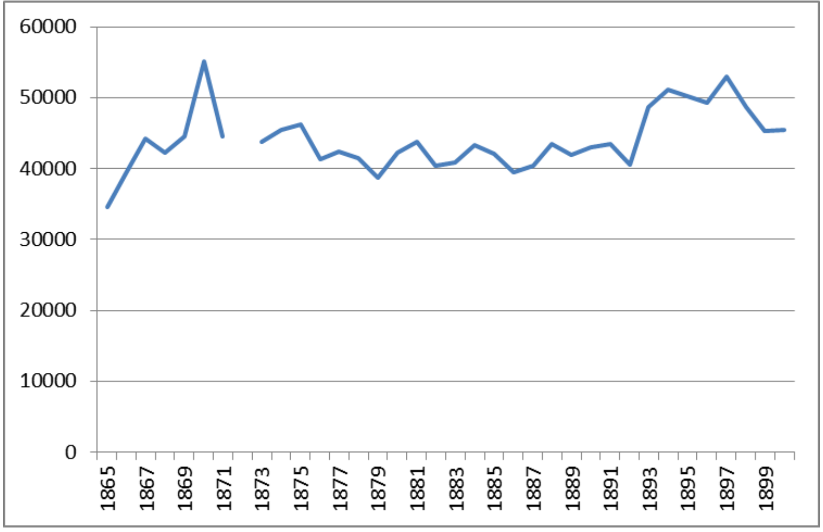
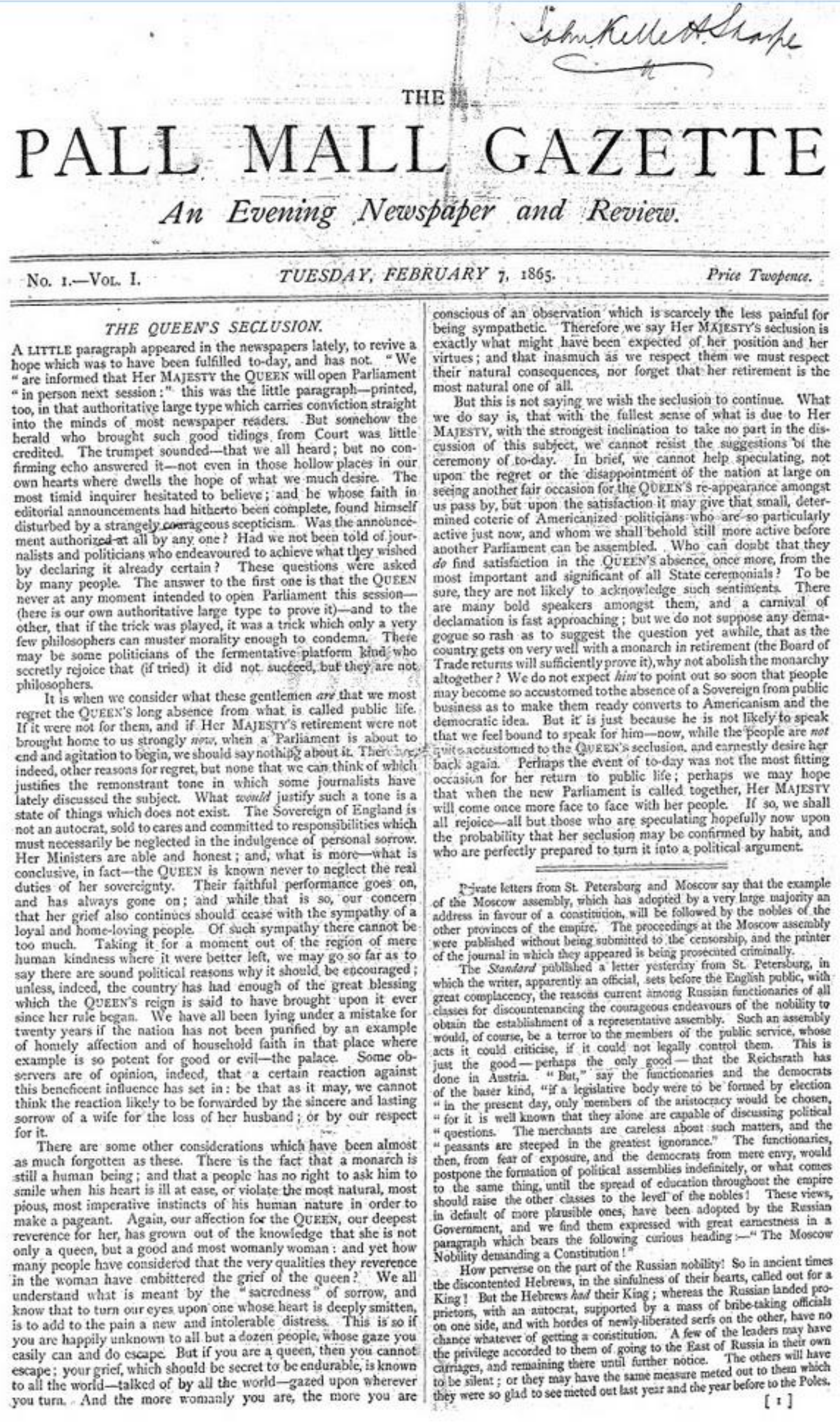


Figure 6.4 First page of first issue of PMGZ



John Kelle A Sharpe

THE

PALL MALL GAZETTE

An Evening Newspaper and Review.

No. 1.—Vol. I. *TUESDAY, FEBRUARY 7, 1865.* *Price Twopence.*

THE QUEEN'S SECLUSION.

A LITTLE paragraph appeared in the newspapers lately, to revive a hope which was to have been fulfilled to-day, and has not. "We "are informed that Her MAJESTY the QUEEN will open Parliament "in person next session:" this was the little paragraph—printed, too, in that authoritative large type which carries conviction straight into the minds of most newspaper readers. But somehow the herald who brought such good tidings from Court was little credited. The trumpet sounded—that we all heard; but no confirming echo answered it—not even in those hollow places in our own hearts where dwells the hope of what we much desire. The most timid inquirer hesitated to believe; and he whose faith in editorial announcements had hitherto been complete, found himself disturbed by a strangely courageous scepticism. Was the announcement authorized at all by any one? Had we not been told of journalists and politicians who endeavoured to achieve what they wished by declaring it already certain? These questions were asked by many people. The answer to the first one is that the QUEEN never at any moment intended to open Parliament this session—(here is our own authoritative large type to prove it)—and to the other, that if the trick was played, it was a trick which only a very few philosophers can muster morality enough to condemn. There may be some politicians of the fermentative platform kind who secretly rejoice that (if tried) it did not succeed, but they are not philosophers.

It is when we consider what these gentlemen *are* that we most regret the QUEEN'S long absence from what is called public life. If it were not for them, and if Her MAJESTY'S retirement were not brought home to us strongly *now*, when a Parliament is about to end and agitation to begin, we should say nothing about it. There are, indeed, other reasons for regret, but none that we can think of which justifies the remonstrant tone in which some journalists have lately discussed the subject. What *would* justify such a tone is a state of things which does not exist. The Sovereign of England is not an autocrat, sold to cares and committed to responsibilities which must necessarily be neglected in the indulgence of personal sorrow. Her Ministers are able and honest; and, what is more—what is conclusive, in fact—the QUEEN is known never to neglect the real duties of her sovereignty. Their faithful performance goes on, and has always gone on; and while that is so, our concern that her grief also continues should cease with the sympathy of a loyal and home-loving people. Of such sympathy there cannot be too much. Taking it for a moment out of the region of mere human kindness where it were better left, we may go so far as to say there are sound political reasons why it should be encouraged; unless, indeed, the country has had enough of the great blessing which the QUEEN'S reign is said to have brought upon it ever since her rule began. We have all been lying under a mistake for twenty years if the nation has not been purified by an example of homely affection and of household faith in that place where example is so potent for good or evil—the palace. Some observers are of opinion, indeed, that a certain reaction against this beneficent influence has set in: be that as it may, we cannot think the reaction likely to be forwarded by the sincere and lasting sorrow of a wife for the loss of her husband; or by our respect for it.

There are some other considerations which have been almost as much forgotten as these. There is the fact that a monarch is still a human being; and that a people has no right to ask him to smile when his heart is ill at ease, or violate the most natural, most pious, most imperative instincts of his human nature in order to make a pageant. Again, our affection for the QUEEN, our deepest reverence for her, has grown out of the knowledge that she is not only a queen, but a good and most womanly woman: and yet how many people have considered that the very qualities they reverence in the woman have embittered the grief of the queen? We all understand what is meant by the "sacredness" of sorrow, and know that to turn our eyes upon one whose heart is deeply smitten, is to add to the pain a new and intolerable distress. This is so if you are happily unknown to all but a dozen people, whose gaze you easily can and do escape. But if you are a queen, then you cannot escape; your grief, which should be secret to be endurable, is known to all the world—talked of by all the world—gazed upon wherever you turn. And the more womanly you are, the more you are conscious of an observation which is scarcely the less painful for being sympathetic. Therefore we say Her MAJESTY'S seclusion is exactly what might have been expected of her position and her virtues; and that inasmuch as we respect them we must respect their natural consequences, nor forget that her retirement is the most natural one of all.

But this is not saying we wish the seclusion to continue. What we do say is, that with the fullest sense of what is due to Her MAJESTY, with the strongest inclination to take no part in the discussion of this subject, we cannot resist the suggestions of the ceremony of to-day. In brief, we cannot help speculating, not upon the regret or the disappointment of the nation at large on seeing another fair occasion for the QUEEN'S re-appearance amongst us pass by, but upon the satisfaction it may give that small, determined coterie of Americanized politicians who are so particularly active just now, and whom we shall behold still more active before another Parliament can be assembled. Who can doubt that they *do* find satisfaction in the QUEEN'S absence, once more, from the most important and significant of all State ceremonials? To be sure, they are not likely to acknowledge such sentiments. There are many bold speakers amongst them, and a carnival of declamation is fast approaching; but we do not suppose any demagogue so rash as to suggest the question yet awhile, that as the country gets on very well with a monarch in retirement (the Board of Trade returns will sufficiently prove it), why not abolish the monarchy altogether? We do not expect *him* to point out so soon that people may become so accustomed to the absence of a Sovereign from public business as to make them ready converts to Americanism and the democratic idea. But it is just because he is not likely to speak that we feel bound to speak for him—now, while the people are *not* quite accustomed to the QUEEN'S seclusion, and earnestly desire her back again. Perhaps the event of to-day was not the most fitting occasion for her return to public life; perhaps we may hope that when the new Parliament is called together, Her MAJESTY will come once more face to face with her people. If so, we shall all rejoice—all but those who are speculating hopefully now upon the probability that her seclusion may be confirmed by habit, and who are perfectly prepared to turn it into a political argument.

Private letters from St. Petersburg and Moscow say that the example of the Moscow assembly, which has adopted by a very large majority an address in favour of a constitution, will be followed by the nobles of the other provinces of the empire. The proceedings at the Moscow assembly were published without being submitted to the censorship, and the printer of the journal in which they appeared is being prosecuted criminally.

The *Standard* published a letter yesterday from St. Petersburg, in which the writer, apparently an official, sets before the English public, with great complacency, the reasons current among Russian functionaries of all classes for discountenancing the courageous endeavours of the nobility to obtain the establishment of a representative assembly. Such an assembly would, of course, be a terror to the members of the public service, whose acts it could criticise, if it could not legally control them. This is just the good—perhaps the only good—that the Reichsrath has done in Austria. "But," say the functionaries and the democrats of the baser kind, "if a legislative body were to be formed by election "in the present day, only members of the aristocracy would be chosen, "for it is well known that they alone are capable of discussing political "questions. The merchants are careless about such matters, and the "peasants are steeped in the greatest ignorance." The functionaries, then, from fear of exposure, and the democrats from mere envy, would postpone the formation of political assemblies indefinitely, or what comes to the same thing, until the spread of education throughout the empire should raise the other classes to the level of the nobles! These views, in default of more plausible ones, have been adopted by the Russian Government, and we find them expressed with great earnestness in a paragraph which bears the following curious heading:—"The Moscow Nobility demanding a Constitution!"

How perverse on the part of the Russian nobility! So in ancient times the discontented Hebrews, in the sinfulness of their hearts, called out for a King! But the Hebrews *had* their King; whereas the Russian landed proprietors, with an autocrat, supported by a mass of bribe-taking officials on one side, and with hordes of newly-liberated serfs on the other, have no chance whatever of getting a constitution. A few of the leaders may have the privilege accorded to them of going to the East of Russia in their own carriages, and remaining there until further notice. The others will have to be silent; or they may have the same measure meted out to them which they were so glad to see meted out last year and the year before to the Poles.

[1]

6.2.2 *REYNOLD'S WEEKLY NEWSPAPER (RDNP)*

Reynold's Newspaper, which we have from its first publication under the name *Reynolds's Weekly Newspaper* on 5 May 1850 until 1900, is described as 'the most popular post-Chartist radical newspaper until at least the twentieth century' (Shirley 2009) and 'the most outspokenly radical paper of the day', which 'appealed to the lower to lower-middle classes, politically democratic and radical, of low educational standard' (Ellegard, cited in British Library n.d.: 7). Consistently leftist, it was published on Sunday in London, with 3 earlier editions for distribution beyond London but carrying the Sunday date. An issue contained 16 pages of 4 columns at the beginning of the period, and 8 pages of 8 columns from 1861.

Its circulation was considerable: 'upwards of 350,000 copies weekly', 'a very large circulation in London, and yet more in the north of England, where Chartist opinions held their ground' according to Fox Bourne (Fox Bourne, cited in British Library n.d.: 348); it 'became the most widely read paper of Victorian England' according to Engel (Engel, cited in British Library n.d.: 28).

The nearly 51 years of data amount to almost 290 million words from 2,635 issues. Issues contained mostly 'News' (making up on average 50% of an issue's wordcount) and 'Adverts' (18%), followed by 'Arts' (6%) (Figure 6.5). RDNP's overall word-recognition score as given by Tanner et al. (2009) is ~87% character accuracy rate, ~82% word accuracy rate, ~74% significant word accuracy rate and ~65% capitalized word accuracy rate. Figure 6.8 shows the first page of the very first issue; Figure 6.5 and Figure 6.6 provide visualizations of the generic make-up of RDNP over time; Figure 6.7 shows the average wordcount per issue over time.

Figure 6.5 Generic make-up of RDNP (proportion of overall wordcount per year in each article category)

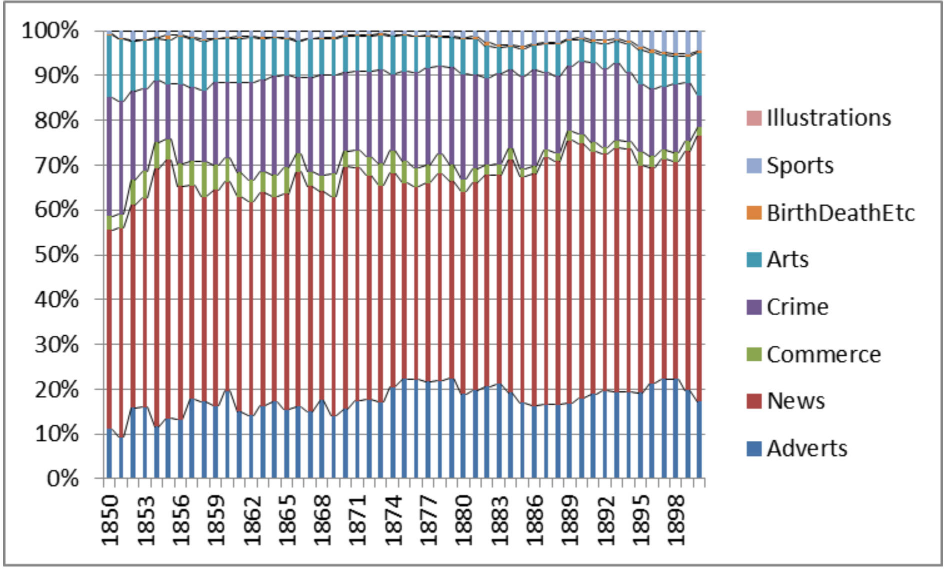


Figure 6.6 Generic make-up of RDNP (raw wordcount in each article category per year)

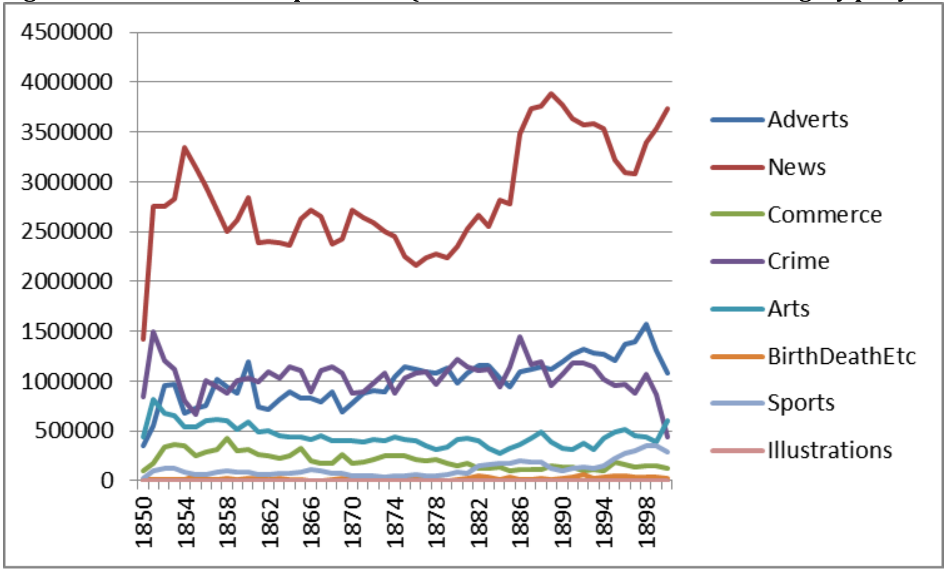


Figure 6.7 Size of RDNP issues over time

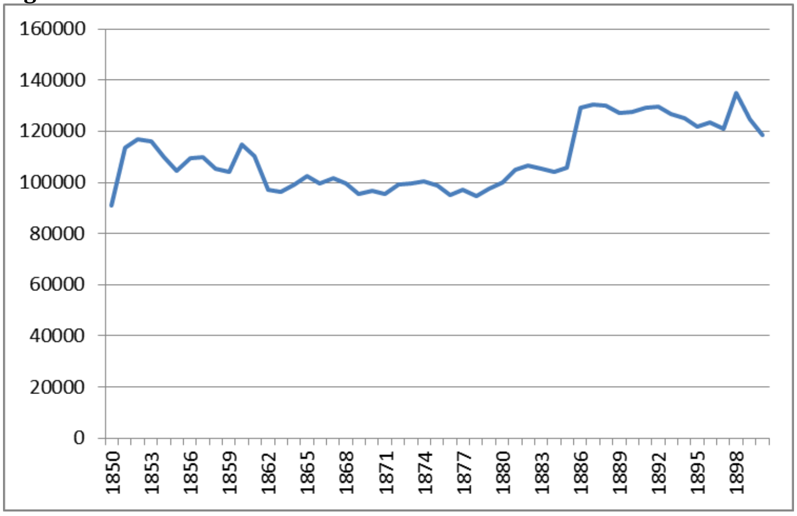
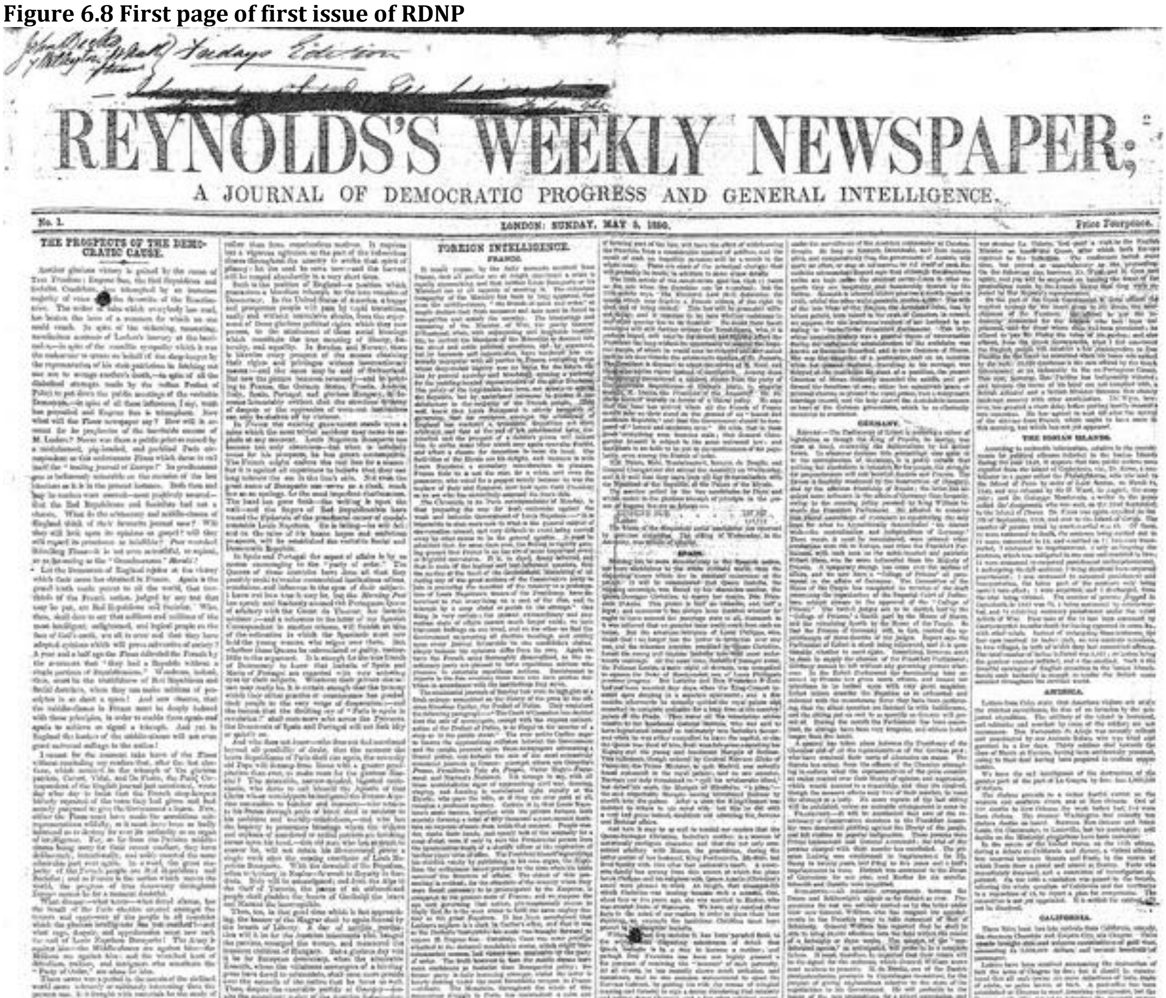


Figure 6.8 First page of first issue of RDNP



John Dicks
7 Wellington Street
Strand

Sundays Edition

REYNOLDS'S WEEKLY NEWSPAPER;

A JOURNAL OF DEMOCRATIC PROGRESS AND GENERAL INTELLIGENCE.

No. 1. LONDON: SUNDAY, MAY 5, 1850. Price Fourpence.

THE PROSPECTS OF THE DEMOCRATIC CAUSE.

GEORGE W. M. REYNOLDS.

FOREIGN INTELLIGENCE.

FRANCE.

GERMANY.

SPAIN.

ROME.

ITALY.

GREECE.

AUSTRIA.

THE IONIAN ISLANDS.

AMERICA.

CALIFORNIA.

6.2.3 THE ERA (ERLN)

The Era was a weekly newspaper (usually published on Sunday), which we have from its first publication on 30 September 1838 to the end of 1900[2], for a total of 3,096 issues and nearly 380 million words. According to James (2009), the *Era* was 'the leading theatrical journal of the Victorian period', originally published by the Licensed Victuallers' Association 'to represent the interests of those in the catering trade'. The paper started off broadly Liberal under Leitch Ritchie as an editor, but became more Conservative under Frederick Ledger's editorship and definitely by 1846; however politics was never a central concern of the paper (British Library n.d.). Circulation was deemed to be 'good', at upwards of 5,000 per week (British Library n.d.).

Overall, the paper contains mostly 'Adverts' (48%) and 'Arts' (38%), followed by 'News' (10%), although of the three papers considered here, this is the only paper to see a considerable change in its generic make-up over time. The period up to 1860 is characterized by a roughly equal proportion of News, Adverts and Sports, as well as non-negligible portions of Commerce, Crime and Arts; the period from 1860 sees a progressive near-disappearance of Crime and Commerce, as well as drastic diminutions of the proportions of Sports and News, in favour of increasing space devoted to Adverts and Arts (see Figure 6.9 and Figure 6.10).

Figure 6.11 shows how the average word-count per issue varies over time: in general terms, early issues contained 16 pages presented in 4 columns per page, for a total of around 80,000 words per issue. At the end of the period, issues contain 32 pages presented in 5 columns per page, for a total of around 190,000 words per issue. *The Era*'s overall word-recognition score as given by Tanner et al. (2009) is ~85% character accuracy rate, ~76% word accuracy rate, ~68% significant word accuracy rate and ~64% capitalized word accuracy rate. Figure 6.12 shows the first page of the very first issue of ERLN.

[2] Except for the years 1848, 1849 and 1850, which were missing from the electronic archive received from the British Library and Gale/Cengage.

Figure 6.9 Generic make-up of ERLN (proportion of overall wordcount per year in each article category)

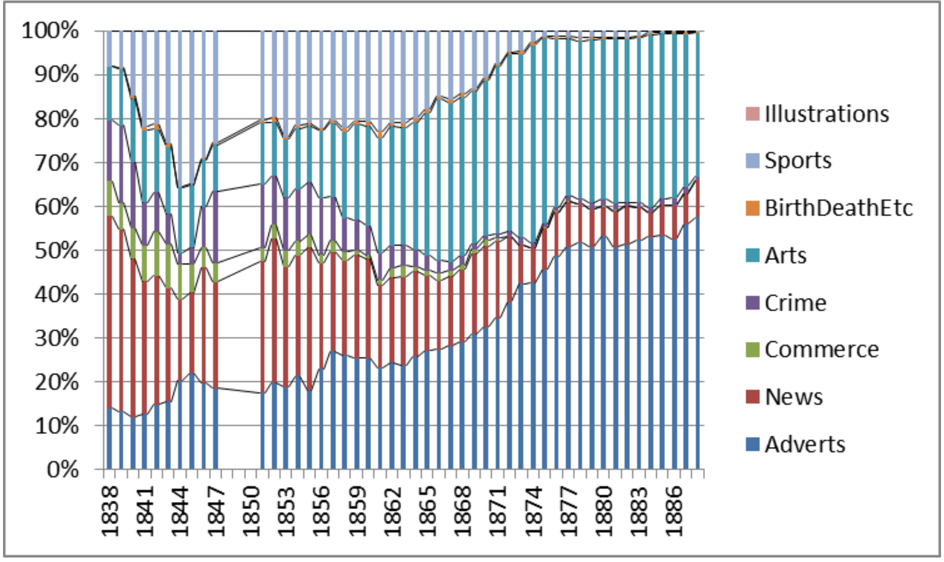


Figure 6.10 Generic make-up of ERLN (raw wordcount in each article category)

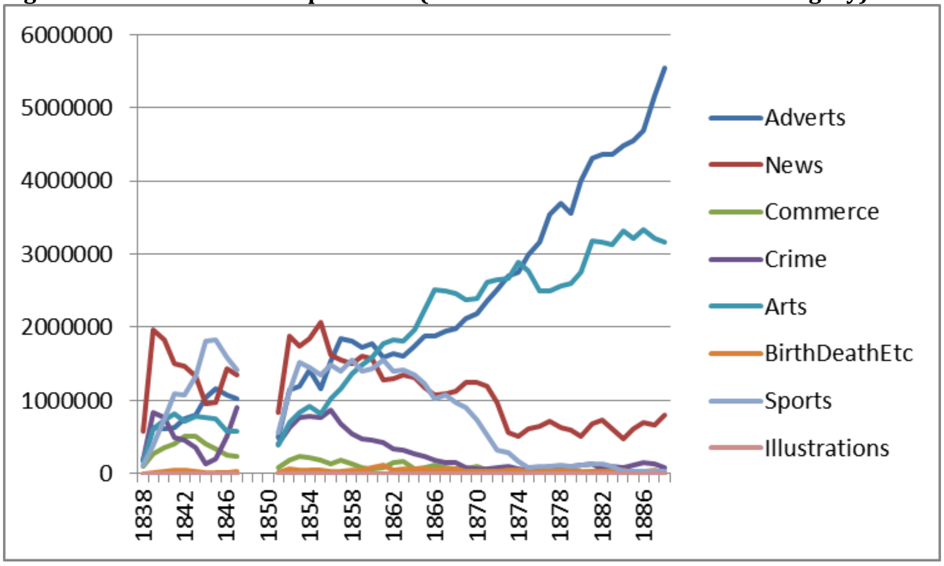


Figure 6.11 Size of ERLN issues over time

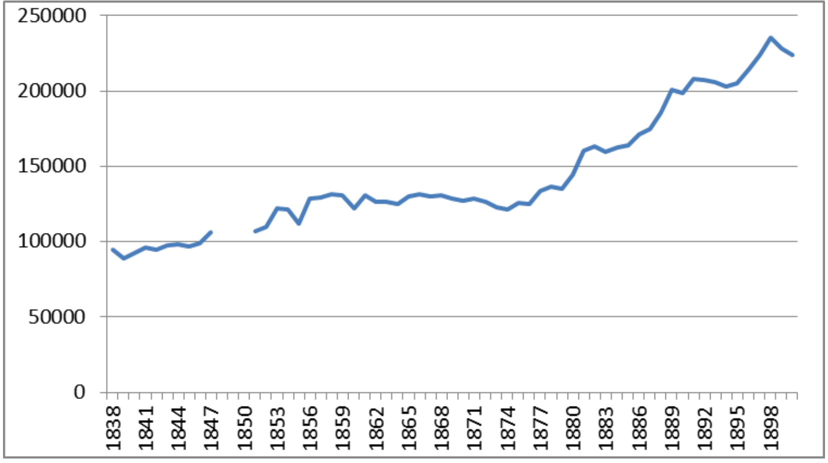
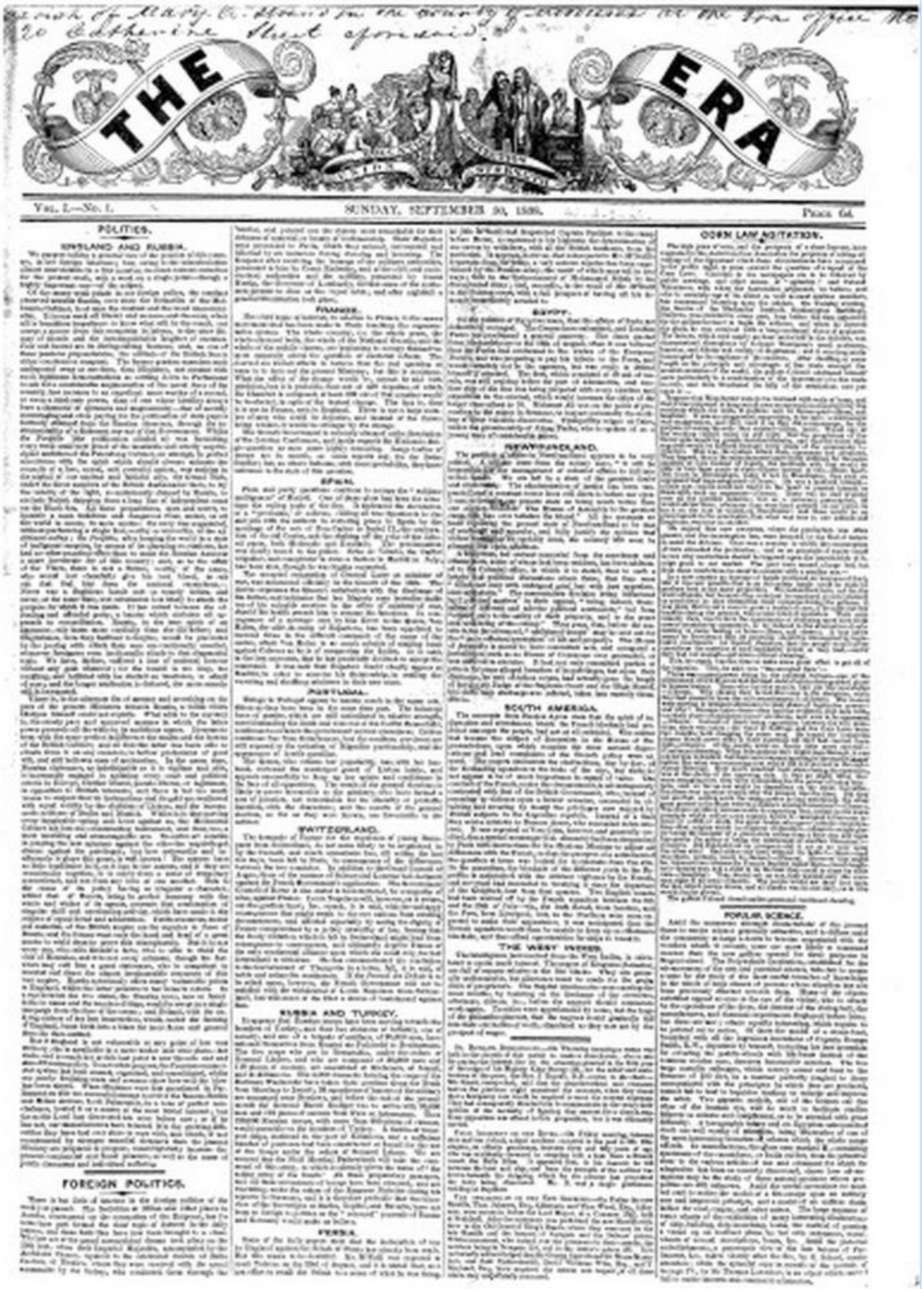


Figure 6.12 First page of first issue of ERLN



…rish of Mary le Strand in the County of Middlesex at the Era Office No 20 Catherine Street aforesaid.

THE ERA

Vol. I.—No. 1. SUNDAY, SEPTEMBER 30, 1838. Price 6d.

POLITICS.

ENGLAND AND RUSSIA.

[illegible]

FOREIGN POLITICS.

[illegible]

FRANCE.

[illegible]

SPAIN.

[illegible]

PORTUGAL.

[illegible]

SWITZERLAND.

[illegible]

RUSSIA AND TURKEY.

[illegible]

PERSIA.

[illegible]

EGYPT.

[illegible]

NEWFOUNDLAND.

[illegible]

SOUTH AMERICA.

[illegible]

THE WEST INDIES.

[illegible]

CORN LAW AGITATION.

[illegible]

POPULAR SCIENCE.

[illegible]

6.2.4 COMPARATIVE SUMMARY OF THE 3 PAPERS

Table 6.1 summarizes the main features of the newspapers for ease of reference. Table 6.2 also provides overall type and token counts, and overall number of issues.

Table 6.1 Main features of PMGZ, RDNP and ERLN

Abbreviation	PMGZ	RDNP	ERLN
Full name	*The Pall Mall Gazette*	*Reynold's Weekly Newspaper*	*The Era*
Type	Daily evening paper	Sunday paper	Sunday paper
Dates	1865-1900	1850-1900	1838-1900
Audience	Intellectuals	Working-class	Publicans
Circulation (per publication day)	8,360	350,000	5000
Political leanings	Started off right, but then left (from 1880s)	Always left	Started off left, but then right (from mid-1840s)
Issue size	40,000, then 50,000 words per issue	100,000, then 120,000 words per issue	100,000, rising to 200,000 words per issue
Issue format	8 pages x 2 columns then 10 pages x 3 columns	16 pages x 4 columns, then 8 pages x 8 columns	16 pages x 4 columns, then 32 pages x 5 columns
Total word-count	468 million words	290 million words	449 million words
OCR quality	90% word accuracy rate	82% word accuracy rate	76% word accuracy rate
Main article genres	News (52%), Adverts (25%), followed by Commerce (8%) and Arts (8%)	News (50%), Crime (18%) and Adverts (18%), followed by Arts (6%)	Adverts (48%), and Arts (38%), followed by News'(10%)

Table 6.2 Main figures for PMGZ, RDNP and ERLN

Abbreviation	PMGZ	RDNP	ERLN
Word-count	468,324,154 tokens	289,617,880 tokens	448,894,847 tokens
Word-count	23,386,491 types	22,153,306 types	24,333,352 types
Type/token ratio	0.0499 types per token	0.0765 types per token	0.0542 types per token
Number of issues	10,620 issues	2,635 issues	3,096 issues

6.3 FINDING OUT 'WHAT PLACES ARE MENTIONED'

This section describes and evaluates approaches to answering the question 'which places are mentioned in this body of texts?'. The first subsection looks at approaches which start from the corpus, whilst the second subsection considers starting from a source external to the corpus. The last subsection provides a summary table of the approaches considered.

6.3.1 Starting from the corpus

6.3.1.1 Frequency list

The most conceptually straightforward way to identify places mentioned in a corpus is to read the texts from beginning to end; but this is an impossibly lengthy task for hundreds of millions of words. The next simplest approach is to read through a list of all words which occur in the texts: a frequency list (introduced in section 2.3.2.2). This will already greatly reduce the count of words to be read: to illustrate, there are 448,894,847 word-tokens in *The Era* but these are instantiations of 'only' 24,333,352 word-types (see Table 6.2). Nevertheless, 20 million words is still an overwhelming figure, and it would not be feasible to simply read through such a list, especially multiple times if one intends to compare several publications.

Part of the problem is that the number of word-types in the data is inflated by the occurrence of words containing OCR errors. It is difficult to determine precisely the magnitude of this effect[3], but it is easy to demonstrate that this effect is important by looking at *hapaxes* (words which occur only once in a corpus – they typically make up a majority of types in a corpus). In *The Era*, there are 20,561,761 hapaxes, which account for 84.5% of all types. Many of these hapaxes are likely to be due to OCR errors. It makes sense to assume that OCR errors will always be over-represented in a list of hapaxes from a corpus which contains OCR errors, because OCR errors are very likely (and much more likely than 'real' words) to occur only once.

Table 6.3 illustrates this effect. It shows 20 hapaxes taken at random from *The Era;* of these, only 1 (*Claudio-*) may be an accurate representation of the text contained in the original source. In fact, out of 100 hapaxes taken at random from *The Era*, I found none which were definitely 'real' words', only 4 which could possibly be accurate representations of words in the sources though they are likely also errors ('Milandel', 'Llandfairelydlogan', 'Eshward' and 'Dustenan'), and a further 6 which were quite likely to be accurate representations of words in

[3] Difficult, that is, without comparing the data to a correct version. In section 4.4.1, I compare the type counts in the uncorrected and gold versions of the CNNE matching corpus and find that the type-count doubles in the uncorrected version compared to the gold one. Even knowing a *token* error-rate, it is difficult to estimate from there how many *types* may be incorrect: though Tanner et al. (2009) report a 76% word-accuracy figure for ERLN (see also section 3.2.2), which implies that 24% of *tokens* are incorrect, it is hard to draw out the implications in terms of *types*.

the sources but which were incorrectly tokenized due to some problem with punctuation (e.g. ‘sad-they’ and ‘SUPERSTITIONS.’). The remaining 90 hapaxes definitely contained OCR errors. This means it is reasonable to expect that a vast majority of the hapaxes in *The Era* contain OCR errors. Let us say that 90% of the 20 million hapaxes in ERLN are errors; removing these from the list of types would bring the number of types down from over 24 million to only around 4.5 million. This is a considerable reduction, but the list is still too long to go through, especially if this is to be done for several publications.

Table 6.3 20 random hapaxes from ERLN

Lambert-rca'
Liverpelw
ARlANI:EDl
berneavetmetnt
reextleinthe
deliglitfuil
Grteshnamn-itrect
alusceay
Etlerems
-'iib-rin
'>8110W
24Chmiad
Johntss
Waetcoaf
Claudio-
Rluf
Swoeepstalses
Chureh.row
tri7Ze
z[ULUNh’E

Even leaving these considerations aside, the frequency list approach is not ideal. In theory, looking through a frequency list for the whole corpus should allow me to identify every single place-name mentioned in the corpus. But in practice this is not the case: only single-word place-names will be identifiable from the list, excluding multi-word place-names such as ‘United Kingdom’. Moreover only place-names *known* to the researcher will be recognized, since frequency lists provide no context for their entries. In addition, as said above, frequency lists will also have very poor precision since only a small number of items on the list will likely be place-names; in fact only 24 out of the first 1000 entries in a frequency list for ERLN were identified as potential place-names. Finally, the frequency provided for an identified place-name may not be very accurate: aside from the imprecisions related to OCR errors (see chapters in part 2), the counts will not distinguish between place-names and their homonyms (such as 'Derby' the city and 'Derby' the sports event). This difficulty is not unique to working with simple frequency lists and will be discussed further in the next section.

6.3.1.2 Annotation

A more sophisticated way of identifying place-names mentioned in the corpus is to use the available annotation. Two main systems of annotation have been applied to the newspapers: CLAWS 6 for parts-of-speech and USAS for semantic categories (see also section 3.2). In CLAWS 6, a relevant tag is 'NP1' which stands for singular proper nouns, and which should in theory capture all place-names, as well as other proper nouns such as names of people and organizations. In USAS, a relevant tag is 'Z2' which stands for geographical names and which could be expected to capture all place-names, and perhaps only place-names.

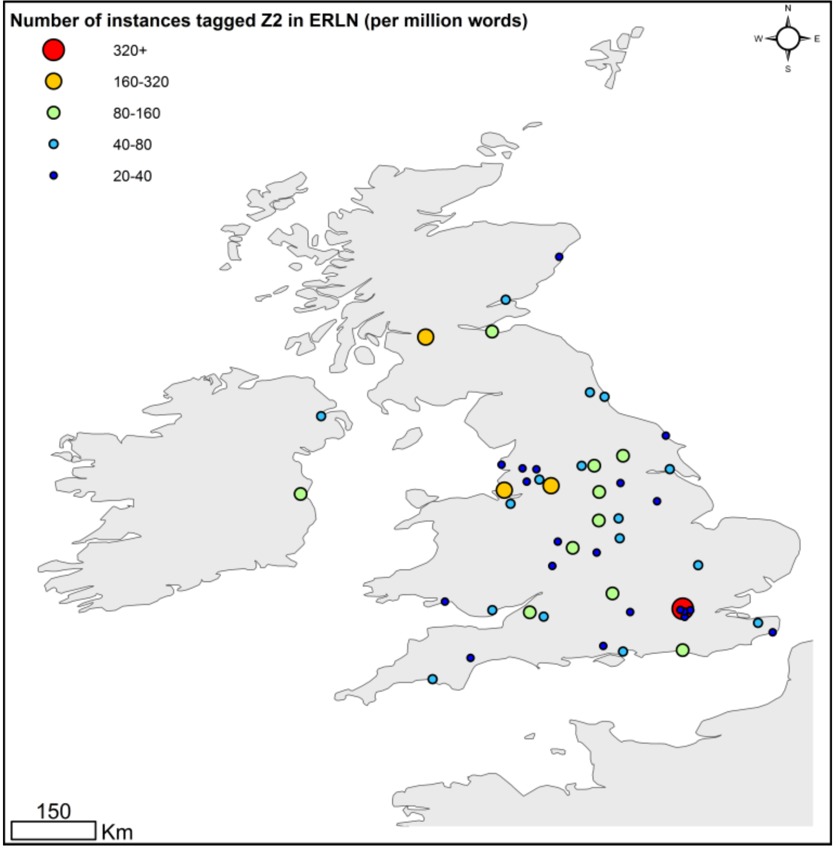
For instance, there are 49,859,843 occurrences of the tag 'NP1' in ERLN, which break down into 7,311,790 different word-types. For 'Z2', the figures are 8,521,813 total instances, from 394,996 different word-types. Although this is less lengthy than the frequency list, the lists are still too long to be read through in their entirety, especially if this is to be done for several corpora. Nevertheless, such a list may provide a good starting-point, for example for identifying the most frequently mentioned places in a given corpus.

Starting from the top of the list, then, how many items in the list would one need to go through in order to find a given number of place-names? 262 out of the first 1000 items in the NP1 frequency list were identified as potential place-names. 586 out of the first 1000 items in the Z2 frequency list were identified as potential place-names. Included in the count of potential place-names were names of continents, countries, cities and neighbourhoods, as well as their demonym forms (e.g. 'European', 'English', 'Londoners'), and names of specific geographic features (e.g. 'Thames', 'Alps'). Excluded from the count were generic words such as 'theatre' or 'square', generic geographical features (such as 'coast' or 'river)', people's names (such as 'James') which are not also names of cities, directions (such as 'East' or 'Northern'), ways of referring to time (such as 'Saturday'), common words (such as 'artillery' or 'cooking'), and specific buildings/infrastructure (e.g. 'london-bridge'). I also did not include in my counts abbreviations (e.g. 'st'), numbers (e.g. '21') or errors (e.g. 'L.ondon'). In any case, the figures provided should be considered as indicative only – there is no guarantee that the proportion would be similar lower down in the list, and I have not systematically checked my assessments using concordances, so I might have missed words that were actually place-names, and very probably included a number of words which were actually never used as place-names in those sources.

Between the 'NP1' and 'Z2' tag, then, the latter is definitely more precise (i.e. a higher proportion of its results are likely to actually be place-names), although neither is 100% precise. From here on, then, I will only be testing the 'Z2' tag. It is clear that even with the 'Z2' list, more entries will need to be read than the number of place-names one is searching for. Nevertheless, the numbers are such that it is certainly feasible to identify, say, the 50 or 100 British cities most frequently mentioned in a newspaper corpus of this size. Of course, the less (relatively) frequent an item is, the further down the list it will be. Hence, more entries will need to be read if one is interested in countries in ERLN than if one is interested in cities, since countries are generally much less frequent than cities in this newspaper.

What about recall? Are all place-names in the corpus correctly tagged 'Z2'? To what extent can the frequency reported in the Z2 list be considered reliable? Table 6.4 shows the 50 most frequent UK cities in the Z2 list (these are mapped in Figure 6.13). The first two columns show the rank of the city when the list is sorted by descending frequency in the whole corpus; the first column shows this rank when all hits are considered, and the second column shows the ranks when only Z2 hits are considered. It is interesting to notice that the same cities are most frequent whether we are considering all mentions in the corpus, or only those tagged 'Z2', but that their rankings can be very different. On average, the Z2 hits capture just over 90% of all mentions of the place-names (see last column of Table 6.4), but in one dramatic case ('Reading'), the proportion is only 43.7%, and in four cases ('Coventry', 'Wigan', 'Exeter' and 'Cambridge'), the proportion is in the range 70-79%.

Figure 6.13 50 British cities with most Z2 tags in ERLN



Since the figures can be very different, which figure is more correct, the total number of mentions of a place-name, or the number of times it is labelled 'Z2'? Is the USAS tagging system working correctly and excluding the 10%-50% of cases where the place-name is not in fact being used to refer to a place, or is it simply failing to capture all of the relevant instances? Answering this question involves looking at what is being tagged as 'Z2'. Methods for exploring the way in which place-names are used in context are the subject of chapter 7. Here, I will simply look at a random sample of 100 concordance lines for 'London', 'Derby', 'Oxford' and

‘Nottingham’ first in the whole corpus, and then within the pool of instances tagged as Z2[4]. The question that I ask is simply ‘in how many of these concordance lines is the place-name used to refer to the British city?’

Table 6.5 shows the number of concordance lines in each sample in which the place-name was used to refer to the British city. Included in the count are:

- cases where the place-name is part of an address (e.g. ‘21, Lemon street, London’) and
- cases where the place-name refers to a person’s origin (e.g. ‘re-engagement by well-known London conductor’).

Excluded are cases where the place-name is part of the name of a building, business or organization (e.g. ‘at the London Music Hall, Sheffield’), since in such cases the presence of the place-name is not an indication that the surrounding discourse refers to the city’s location.

Included in brackets are cases which could arguably be either included or excluded:

- cases where the place-name is part of a telegraphic address (e.g. “Cleaning, London”) – in my sample, this only occurs for London; and
- cases where the place-name is part of the title of a work of fiction, presumably as a reference to the city (e.g. the play title ‘Dangers of London’). (Cases where the place-name is part of the title of a work of fiction but clearly does not refer to the city are excluded.)

[4] Note that these 100 random concordance lines were taken from the line-filtered version of ERLN, see section 3.2.1. Since the figures are only indicative anyway, this should not alter the validity of the conclusions.

Table 6.4 50 British cities with most Z2 tags in ERLN

Ranked by all hits	Ranked by Z2 hits	Place-name	Hits in whole corpus	Per cent of whole corpus hits tagged as Z2
1	1	london	583610	90.9
2	2	liverpool	14035	95.4
3	3	manchester	99656	96.5
4	5	york	85132	81
5	4	glasgow	80814	93.9
6	6	birmingham	70232	96.6
7	7	leeds	62694	96
8	8	brighton	52685	89.1
9	14	derby	45744	84.5
10	9	dublin	45134	96.4
11	12	oxford	44299	89
12	10	sheffield	43353	95
13	11	edinburgh	42983	92.4
14	13	bristol	41028	95
15	15	bradford	36623	93.2
16	16	hull	35018	95.5
17	17	leicester	34332	87
18	19	newcastle	30230	92.4
19	18	nottingham	30177	96.3
20	22	cambridge	30078	79.5
21	20	belfast	26810	97
22	24	bath	26706	80.8
23	23	chester	26677	81.5
24	21	cardiff	26441	96.5
25	48	reading	25467	43.7
26	27	bolton	23060	84
27	31	preston	22133	80.2
28	25	portsmouth	21170	95.1
29	26	sunderland	20565	95.6
30	29	canterbury	20472	90.5
31	28	dundee	19416	95.7
32	30	plymouth	18992	96.6
33	32	southamptor	18327	95.3
34	37	exeter	18254	77.1
35	33	aberdeen	17737	92.4
36	35	lincoln	17044	84.1
37	39	wigan	16480	75.5
38	34	greenwich	15981	93.7
39	46	coventry	15427	73
40	36	croydon	14811	95.2
41	38	blackburn	14517	94.6
42	42	dover	12916	91.9
43	40	blackpool	12854	95.8
44	44	worcester	12591	92.6
45	43	scarborough	12535	94.6
46	41	wolverhampt	12342	97.1
47	47	chelsea	12005	93.3
48	45	brixton	11927	96.4
49	50	doncaster	11586	93
50	49	swansea	11415	95.2

Table 6.5 In ERLN, concordance lines (out of 100) in which word is used to designate British city

N	Place-name	Designating UK town/city in Z2 random sample (telegraphic address, work of fiction)	Designating UK town/city in random sample from whole corpus (telegraphic address, work of fiction)
1	London	86 (7, 2)	84 (1, 7)
2	Oxford	36 (0, 1)	45 (0, 0)
3	Derby	35 (0, 1)	36 (0, 0)
4	Nottingham	91 (0, 0)	94 (0, 1)

From Table 6.5, we see that the proportion of instances where the place-name seems to be referring to the city is not much different in the samples from the whole corpus and the samples from among words tagged ‘Z2’. This means that the instances excluded from the ‘Z2’ count are more likely to be genuine cases than not, so that in most cases the overall frequency of mentions of the city-name when used as reference to that city is actually going to be closer to the whole-corpus value than the frequency of instances tagged ‘Z2’. Another implication of this finding is that infrequent places may be altogether missing from a ‘Z2’ list. This should not matter as long as we are only working with place-names which are mentioned frequently, but if exhaustivity is a concern, this may be something worth considering further.

Beyond the (non-)usefulness of Z2 counts, we also see from Table 6.5 that there is a noticeable difference between place-names in terms of the proportion of instances in which they are being used in a context in which they do not refer to the British city. In my examples above, ‘London’ and ‘Nottingham’ almost always refer to the cities but ‘Oxford’ and ‘Derby’ more often than not refer to something other than the British city. For ‘Oxford’, the most common sense is ‘oxford street’ (the street in London), which occurs in 42 cases out of 100 in the sample[5]. For ‘Derby’ the most common sense is sports-related (e.g. ‘the Derby winner’). This sports-related sense is hard to quantify for the whole corpus, but in my sample the expression

[5] Incidentally, the case of 'oxford street' provides an illustration of the impact of corpus preparation techniques: in the line-filtered version, hyphens were removed, leading to many cases of oxford-street becoming *oxford street*. Hence, in the line-filtered version, over 20,000, or 34% of total instances, of oxford were in fact *oxford street*. In the non line-filtered version, however, hyphens were not systematically removed, as a result of which only 700 instances of *oxford street* are retrieved.

'the Derby' (with an article) never occurs to refer to city of 'Derby'; in the entire ERLN corpus, there are 14,816 occurrences of 'the Derby', which constitutes over 32% of all occurrences of 'Derby'.

The issue of polysemy which has just been identified is a problem which goes beyond the use of annotation. All frequency figures – not just those associated with specific tags – are vulnerable to this problem. One solution may be to generate an estimate for each place-name of the proportion of instances which are used to refer to the desired place. This estimate could then be used to 'adjust' the frequency figures downwards where necessary. I come back to this issue and this suggestion in section 6.4.

A suggested procedure for using the 'Z2' tag which takes into account such an 'adjustment' could be the following:

- generate a list of words tagged 'Z2'
- order it by descending frequency
- identify the type of place-names one is interested in (e.g. British cities) and decide on a number of place-names sought for (e.g. 50)
- read through the list, identifying candidates until the desired number has been identified
- for each place-name identified, obtain the overall number of mentions
- for each place-name identified, take a random sample of 100 concordance lines (out of *all* concordances lines, including not only ones where the place-name has been tagged Z2 but also those where it has not) and count the number of concordance lines in which the place-name is being used to refer to the place
- use the proportion found in the previous step to adjust the overall number of mentions downwards, where necessary

At this end of this procedure, one will then have a list of the N most frequently mentioned place-names in a given corpus, along with an estimate of the number of times the place-name is used to refer to the place of interest (see also section 6.4.2).

To summarize, an advantage of using annotation queries over a simple frequency list is that the precision is greater (i.e. more of the items on the list are actually place-names). Nevertheless, the method also inherits some of the difficulties associated with working with frequency lists: 'Z2' lists will still omit multi-word place-names and rely on the researcher's knowledge of place-names (given the lack of context). In addition, although the Z2 list is shorter than the frequency list, it still contains over 390 thousand items. Reading through entire Z2 lists is still not feasible, so in practice they will probably be most helpful to identify only the most frequent places mentioned. In addition, it is preferable to work with an overall count of mentions rather than a count of mentions tagged 'Z2' since the 'Z2' tags do not capture all mentions of a given place in the corpus, and since in most cases the instances left out are more likely than not to be genuine cases where the place-name is being used to refer to a place.

6.3.1.3 Geo-parsing

6.3.1.3.1 The whole corpus

Geo-parsing is a convenient technique for automatically identifying place-names in a text or corpus. It involves making use of a gazetteer – a list of place-names with their alternative spellings, geographical coordinates and optional characteristics (e.g. population count, region) – to automatically locate and tag all instances of place-names referenced in the gazetteer. This technique would hence be expected to greatly improve the precision of results compared to the previous procedures described, as well as not having to rely on the researcher's prior knowledge of place-names. However, it does not solve all problems. One problem which remains unsolved is polysemy: although geoparsers incorporate solutions to address this, some degree of manual disambiguation is always required (see e.g. Gregory and Hardie 2011: 302). Another problem is that the method, to date, is very computationally intensive and requires much time and resources to be applied to a large amount of data. It hence remains difficult to reliably geo-parse very large quantities of text; the data to which I have access for this thesis has not been geo-parsed in full. Are there workable alternatives to geo-parsing a whole corpus? The following sub-sections explore some potential ones.

6.3.1.3.2 A part of the corpus

An alternative to geo-parsing a whole body of text is to simply identify a portion of text which is relevant to a given research question, and geo-parse only that. This is the approach taken by Rupp et al. (2014), who start off with a search query, and geo-parse the retrieved concordance lines. This is a promising approach, which allows for answering the question 'what places are mentioned in this subcorpus?' – a question which may, for example, come down to answering a question of the type 'what spatial patterns are associated with this theme?'.

This approach has the advantage that it is technically easier to achieve than geo-parsing a whole corpus, whilst still having the advantages of geo-parsing (such as high recall, and the possibility to investigate the discourses surrounding place-names since some context is retained). On the other hand, it also retains the disadvantages of geo-parsing, namely the need for some manual disambiguation of polysemic place-names. Beyond this, however, the approach is actually unable to answer fully the question 'what places are mentioned in this corpus?' since it only allows for identifying the places mentioned in a part of the corpus. Since this approach has been discussed in published research, I will not test it further here.

6.3.1.3.3 A list of words

Another alternative to geo-parsing a whole corpus is to geo-parse a list of words – a frequency list or list of words tagged 'Z2', for example. This is a promising solution, which combines some of the advantages of both methods: geo-parsing removes the need to rely on the researcher's knowledge of place-names and promises high recall, whilst working with a frequency list relieves some of the computational intensity required. Unfortunately, the approach also combines some of the weaknesses of both approaches: multi-word place-names will still be non-retrievable, and polysemy remains an issue. Furthermore, a shortcoming of the approach, as compared to geo-parsing the whole corpus, is that whilst it allows for answering the question 'what places are mentioned in this corpus?', it makes the transition to answering questions about 'what is said about these places?' difficult. This is because the association between the place-name and its context is broken. Bridging this gap may be manageable in two ways: either by re-processing the original data to add the additional information generated by the geo-parser (but this is ultimately not very different from whole-corpus geo-parsing in the first place – which may then be preferable); or by searching for the identified place-names one by one in the corpus. This latter method will likely prove time-consuming and not very scalable, but it should allow further investigation of a relatively small selection of place-names; see section 6.4.4 for an illustration of this approach.

6.3.2 Starting from outside the corpus

A different approach from the previous ones is to start from outside the corpus altogether. The idea here is to start off from a principled, prior list of place-names, and then search for those in the corpus. An example of such a list of place-names is a population table; this will be illustrated in section 6.4.2. Whilst searching several corpora or sections of a corpus for each place from a list one-by-one may be time-consuming, it provides a principled starting-point for exploration, helps provide context, and can also help identify absences – something which is very difficult to do using the approaches considered previously.

6.3.3 Summary

The various approaches and their strengths and weaknesses are summarized in Table 6.6. The problem of 'incorrect matching' refers to the pairing of a place-name with the wrong set of coordinates.

Table 6.6 Summary of relative strengths and weakness of various approaches to finding out 'what places are mentioned in this corpus'

Strengths	Weaknesses
Close-reading the whole corpus	
- (Exhaustivity)	- Not scalable
- Less dependence on prior knowledge	
- Provides context	
Frequency list	
- (More efficient than close-reading the whole corpus)	- Not scalable
	- Depends on researcher's prior knowledge
	- Excludes multi-word place-names
Annotation: NP1	
	- Not very precise
	- Problem of polysemy
	- Depends on researcher's prior knowledge
	- Unlikely to allow for reaching more than 'top-of-the-list'
	- Not total recall
	- Excludes multi-word place-names
Annotation: Z2	
- More precise than NP1	- Problem of polysemy
	- Depends on researcher's prior knowledge
	- Unlikely to allow for reaching more than 'top-of-the-list'
	- Not total recall
	- Excludes multi-word place-names
Geo-parsing: whole corpus	
- Exhaustivity	- Still technically challenging
- Allows for answering 'what places are mentioned in relation to theme X'	- Problem of polysemy
- Allows for geographical grouping of search terms, e.g. 'what is said about places in the North'	- Problem of incorrect matching
- Can capture multi-word place-names	
Geo-parsing: a section of a corpus	
- Allows for answering 'what places are mentioned in relation to theme X'	- Problem of polysemy
- Can capture multi-word place-names	- Problem of incorrect matching
	- Doesn't allow for answering 'what places are mentioned (overall) in this corpus'
Geo-parsing: a word-list	
- May come close to exhaustivity	- Problem of polysemy
	- Problem of incorrect matching
	- Not total recall
	- Difficulty in bridging gap to question 'what is said about these places'
	- Excludes multi-word place-names
	- Poor precision
	- Unreliable counts
Starting from an external source	
- May provide context	- May pose problems of scalability
- Allows for identifying absences	

6.4 PLACES MENTIONED IN ERLN, PMGZ AND RDNP

Having considered different approaches to finding out which places are mentioned in a large set of digital texts, this section will present some findings based on using the approaches identified as promising alternatives to geo-parsing a whole corpus. These are looking at place-names chosen because they attract the most 'Z2' tags (as per section 6.3.1.2); looking at place-names listed in an external source (here a population table) (as per section 6.3.2); and geo-parsing a list of words tagged 'Z2' (as per section 6.3.1.3.3). The discussion will focus on names of British cities, but similar methods could be applied at other scales – to investigate country-names or names of neighbourhoods, for example.

One caveat applies to the figures below: all counts of mentions provided below contain several layers of error. The first source of error is OCR errors. Results from previous chapters suggest that most frequency counts are likely to be under-estimates, although the magnitude of the effect may differ from one word to another and from one part of a corpus to another. Nevertheless, in some cases the count may actually present a small over-estimation due to real-word errors. The other source of error is polysemy. This applies in two ways. Place-names may be homonyms of common or proper nouns (e.g. 'Derby' the city and 'Derby' the sports event), and they may also simply refer to different places (e.g. 'Newcastle', which may refer to either Newcastle-under-Lyme, Staffordshire, or Newcastle-upon-Tyne, Tyne and Wear). The 'adjustments' (introduced in section 6.3.1.2) are intended to correct for polysemy, but while the first type of polysemy is fairly easy to spot, the latter is much harder; adjustment factors hence come with their own layer of error (beyond the statistical one associated with the sampling from which they have originated). Given that adjustment factors need to be produced for each place-name and body of text considered, adjustments have not been applied to all figures; adjusted figures are clearly indicated as such.

6.4.1 British cities with the most mentions

Which British cities are mentioned most in each newspaper? This question can be answered using the 'Z2' list approach (introduced in section 6.3.1.2). Table 6.7 shows the 20 most mentioned cities in each newspaper, by order of their appearance in the Z2 list (ordered by descending frequency); it also provides the (non-adjusted) overall number of mentions (in both raw and relative form). Highlighted in bold are cities which appear lower than their overall number of mentions would seem to warrant – the discrepancy is due to the difference in proportions of overall mentions which have been tagged as 'Z2' from one place-name to the other (see section 6.3.1.2). Figure 6.14, Figure 6.15 and Figure 6.16 map the 50 most mentioned British cities in each of the 3 newspapers.

Table 6.7 The twenty British cities most mentioned in ERLN, PMGZ and RDNP

PMGZ		
City	N° of mentions (overall, non-adjusted)	N° of mentions per million words
london	608347	1298.99
york	79579	169.92
liverpool	50828	108.53
oxford	44415	94.84
brighton	39766	84.91
dublin	37422	79.91
manchester	36276	77.46
cambridge	32210	68.78
edinburgh	31333	66.9
dover	24221	51.72
glasgow	22499	48.04
sheffield	18996	40.56
chatham	16080	34.34
birmingham	16560	35.36
southampton	14916	31.85
bristol	15983	34.13
exeter	15079	32.2
derby	22339	47.7
leeds	13932	29.75
reading	40618	86.73

ERLN		
City	N° of mentions (overall, non-adjusted)	N° of mentions per million words
london	583610	1300.1
liverpool	143035	318.64
manchester	99656	222
glasgow	80814	180.03
york	85132	189.65
birmingham	70232	156.46
leeds	62694	139.66
brighton	52685	117.37
dublin	45134	100.54
sheffield	43353	96.58
edinburgh	42983	95.75
oxford	44299	98.68
bristol	41028	91.4
derby	45744	101.9
bradford	36623	81.58
hull	35018	78.01
leicester	34332	76.48
nottingham	30177	67.23
newcastle	30230	67.34
belfast	26810	59.72

RDNP		
City	N° of mentions (overall, non-adjusted)	N° of mentions per million words
london	242177	836.19
liverpool	34318	118.49
york	37043	127.9
manchester	18218	62.9
dublin	14785	51.05
bristol	14012	48.38
birmingham	13545	46.77
glasgow	13171	45.48
leeds	11599	40.05
oxford	11800	40.74
brighton	11063	38.2
derby	14793	51.08
portsmouth	10291	35.53
sheffield	8335	28.78
edinburgh	9701	33.5
dover	7719	26.65
chester	7290	25.17
chelsea	6959	24.03
hull	6840	23.62
cork	7542	26.04

From Table 6.7, Figure 6.14, Figure 6.15 and Figure 6.16, the following observations can be made:

- Overall, ERLN seems to mention all British cities more often (relative to the content of its issues) than the other two newspapers. This may be related to differences in generic make-up; perhaps places are relatively more frequent in advertisements than in other genres, for example, and ERLN has more advertisement content than the other two papers. This hypothesis is explored in section 6.4.3.
- Although only the most frequently mentioned British cities are captured on the maps, it is clear that all three newspapers mention places all around the UK, or at least all those parts of the UK where important centres of population can be found. Perhaps there is a simple relationship between mentions of places and their population, with more populous places mentioned more often in the newspapers? This hypothesis is explored in section 6.4.2.

From these maps and tables, it is difficult to identify absences; it is also difficult to identify a clear difference in overall spatial trends between newspapers. Perhaps this is because there is no overall difference, or perhaps this approach does not allow us to see these differences very clearly.

Figure 6.14 50 British cities with most mentions in ERLN

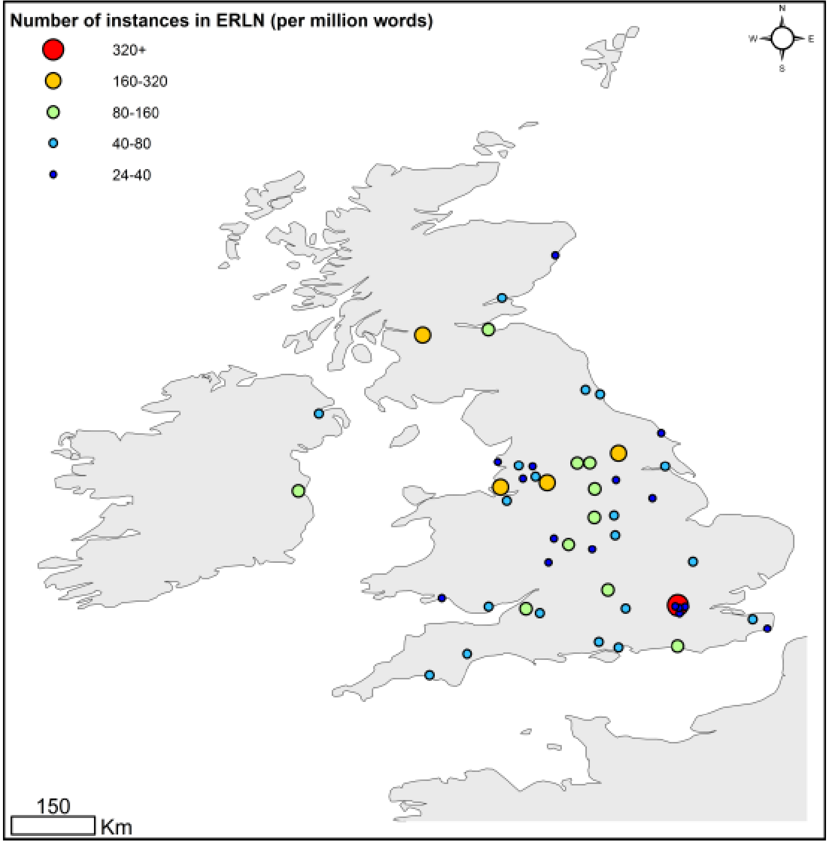


Figure 6.15 50 British cities with most mentions in PMGZ

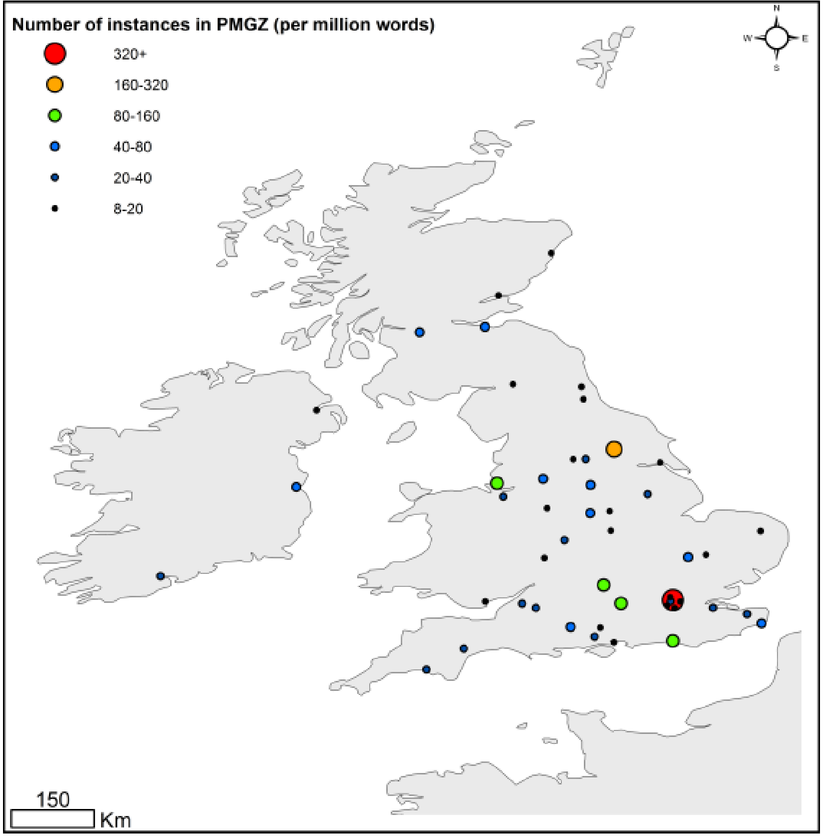
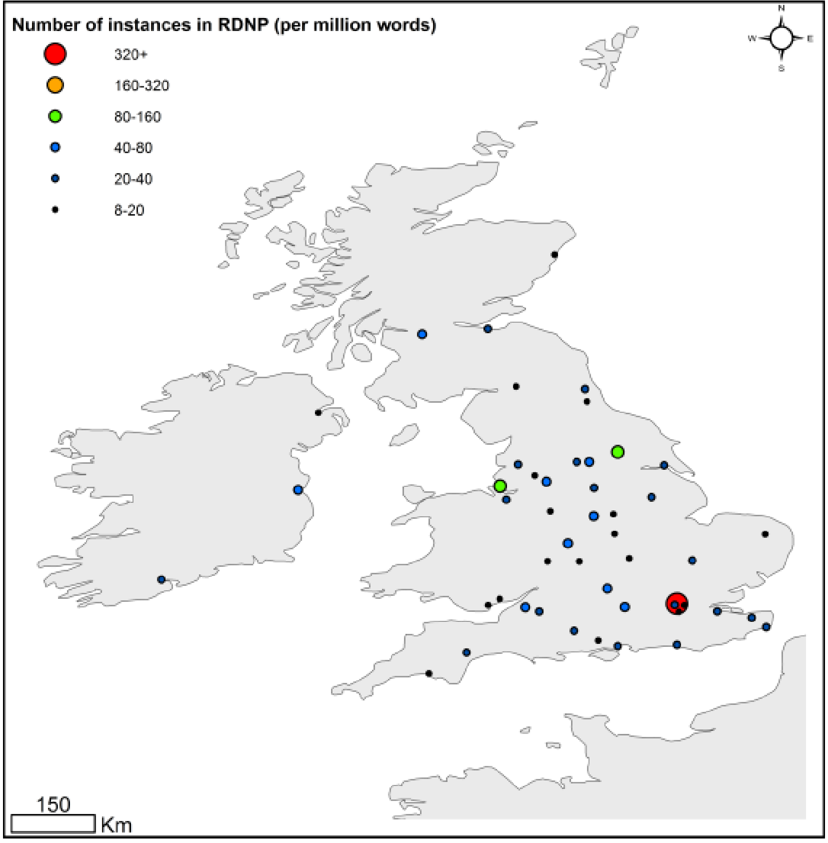


Figure 6.16 50 British cities with most mentions in RDNP



6.4.2 RELATIONSHIP BETWEEN MENTIONS AND POPULATION

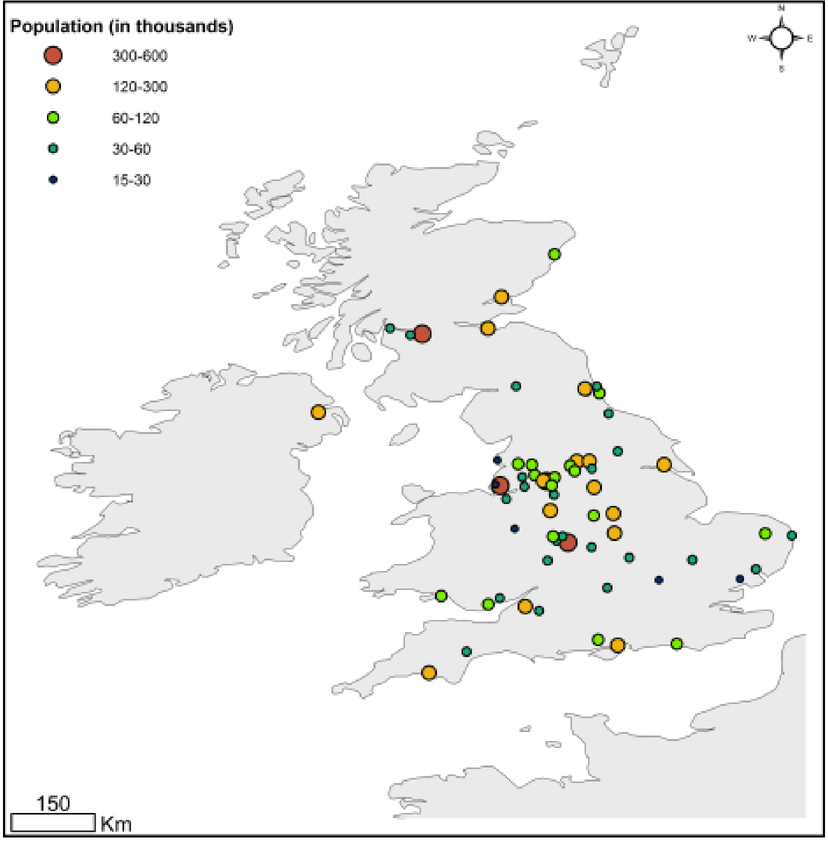
In the previous section, it was suggested that there may be a relationship between a city's population and the number of mentions it attracts in the newspapers. The simplest way to explore this relationship is to use the approach involving starting off from an external source (as introduced in section 6.3.2). Indeed, considering only cities which are mentioned frequently may lead to circularity: if there is a relationship between mentions and populations, then cities which are mentioned frequently are likely also more populous. Hence using an external source listing cities which are both more and less populous is a preferable option to using the list of most frequently mentioned cities obtained in the previous subsection. Since exploring the relationship between mentions and population requires data on population, an obvious external source to turn to is a population table. The population table I have used is from Mitchell and Deane (1971) who very usefully summarize the statistics from various censuses collected by the Registrar General from first taking office in 1837 through to the twentieth century; although the table understandably focuses on population centres (i.e. it does not report on small villages), it includes cities with populations ranging from a few thousand to a few hundred thousand inhabitants.

Hence, in this subsection, the cities considered are different from those in the previous subsection. One notable omission is London, which does not figure in Mitchell and Deane's (1971) population table. It would have been possible to find population figures for London from another source, but certain pragmatic considerations led to accepting its omission. First, the boundaries of London are notoriously difficult to define; finding appropriate and comparable population figures is hence not a straightforward task. Second, both London's population and its number of mentions in the newspapers are of such a greater magnitude than all other British cities that including it in visual representations dwarfs all other patterns, even when using log scales.

Mitchell and Deane (1971)'s population table mentions 71 cities. 6 of these had to be excluded. Reading was excluded because a majority of mentions of 'reading' do not refer to the city (see also section 6.3.1.2). 'Bournemouth' and 'Southend-on-sea' were excluded because they did not figure in the 1851 census. 'King's Lynn', 'South Shields' and 'St Helens' were excluded because, being multi-word place-names, their frequency of mentions could not be simply derived from a frequency list. This left 65 remaining cities, all of which figure in censuses for the years 1851, 1861, 1871, 1881, 1891 and 1901.

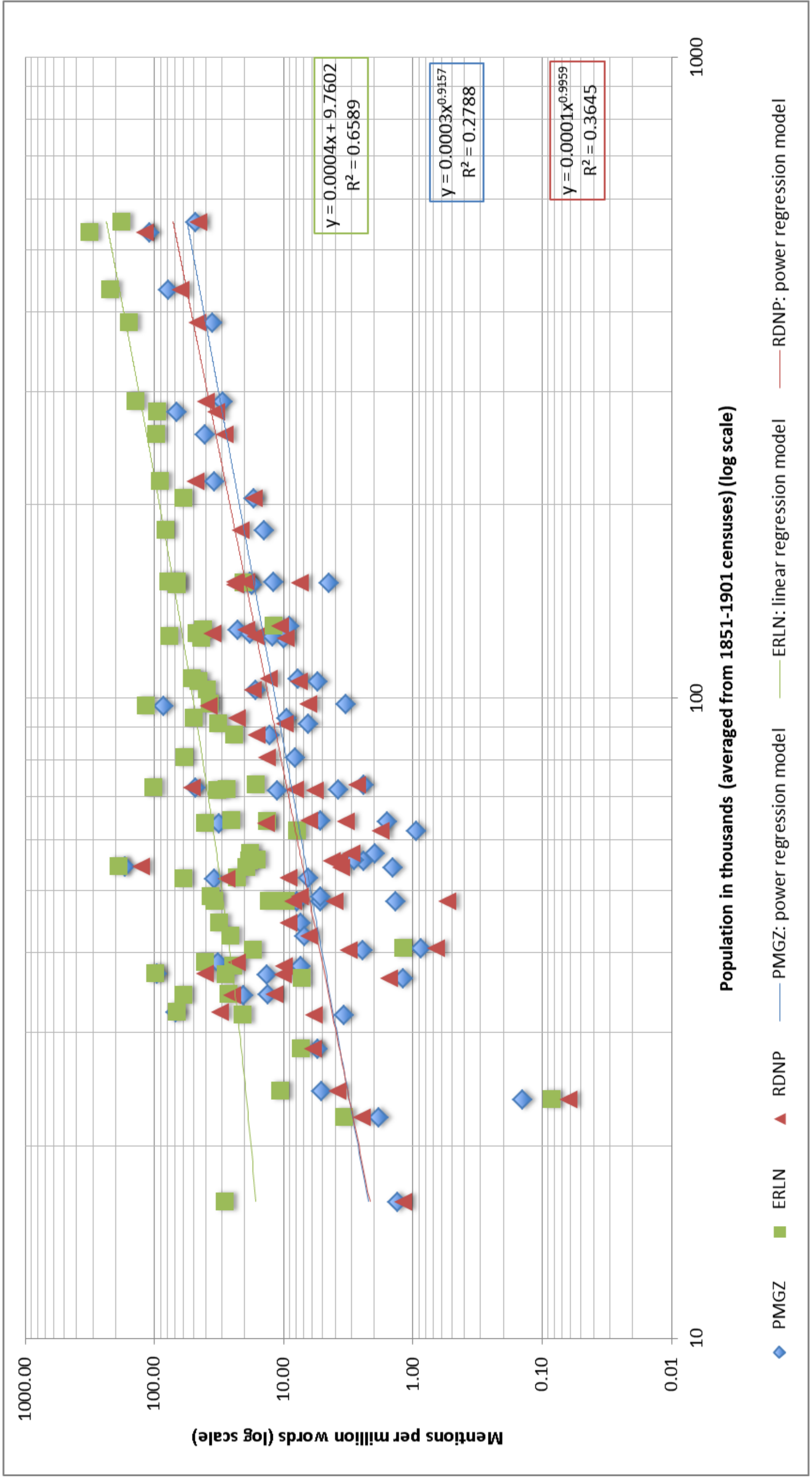
What would be an appropriate population figure to compare to an overall number of mentions over a period of decades in several newspapers? I chose to use an average of the population figures provided in each of the 6 censuses between 1851 and 1901; this seemed the best equivalent to overall number of mentions over a large period of time. Of course the newspapers themselves cover different periods: all end in 1900, but RDNP starts in 1850, ERLN in 1838, and PMGZ in 1865. Nevertheless, I chose to keep all the information available, rather than trimming the analyses to the common period 1865-1900. Figure 6.17 hence shows the averaged population figures from the censuses between 1851-1901 for 65 British cities.

Figure 6.17 Population in British cities (averaged from 6 censuses between 1851 and 1901)



So, is there a relationship between the number of mentions a city gets in the newspapers and its population? Figure 6.18 plots the number of mentions per million words for each newspaper against the averaged population figures over the 1851-1901 period; log scales are used for both axes since both the population and frequency data are heavily skewed with many observations with small values and few with large values.

Figure 6.18 Relationship between population and mentions of British cities in ERLN, PMGZ and RDNP overall



A quick look at Figure 6.18 suggests that there is a broad relationship between population and mentions, with more populous cities tending to attract more mentions than less populous cities. To explore this relationship, regression lines were fitted. Linear, power, logarithmic and exponential models were tested for all three newspapers and the best fitting model is the one shown in the figure.

For PMGZ and RDNP, the best fitting models were power models with R^2 values of 0.28 and 0.36 respectively, as compared to values of 0.17 and 0.35 for the linear models. For ERLN, the linear model was a better fit, with a R^2 value of 0.66 as opposed to 0.39 for the power model. R^2 values are indicative of the 'goodness of fit' of a regression model; they are a measure of the proportion of variance which can be explained by the model. Hence an R^2 of 0.1 suggests that only 10% of the variation in the data can be explained by the model. Here, the R^2 values suggest that the model fits relatively well, although it does not explain all of the variance in the data. This in turn suggests that population is quite likely to be a factor in the difference in the mentions of the places in these 3 newspapers – although much more so for ERLN than for the 2 other newspapers – but that it cannot explain all the variation: some other factors must also play a part.

Hence the relationship between population and mentions is broadly present in all 3 newspapers, but is strongest in ERLN and weakest in PMGZ. It is tempting to come up with hypotheses as to why this might be. For example, one might posit that the relationship is stronger in ERLN because the newspaper offers more reports on (and adverts for) artistic and sporting events whose frequency in a given location is likely to correlate with population; in contrast PMGZ being a more 'high-brow' magazine would be more likely to focus on places associated with higher social classes – and the social 'importance' of such places does not necessarily correlate with population. Nevertheless, whilst it is easy to formulate such hypotheses, it is far harder to verify them.

Figure 6.18 also suggests that the relationship between population and mentions fits best for more populous cities: on the figure, these cities lie closer to the lines of the regression models. In contrast, the least populous cities in all three newspapers tend to be mentioned less than their population would predict, and cities with middle-range populations are often mentioned more than expected given their population.

Since some cities are mentioned more or less than expected given their population, it is interesting to ask which cities these are. Of course, some deviation from the prediction is expected, so cities have to be far enough from the model for their deviation to be considered 'surprising'. The average distance between the expected and observed mentions of places for the three newspapers are 10.84 (RDNP), 13.45 (PMGZ) and 21.34 (ERLN). Taking these averages as a cut-off point, there are 9 cities which are mentioned more than expected (i.e. for which the distance from the prediction is more than those averages) in all three newspapers. These are listed in Table 6.8 along with their residuals (the distance from the prediction – positive values are those describing more mentions than expected, negative values less mentions than expected).

Table 6.8 9 British cities which are mentioned in all 3 newspapers more often than expected given their population (residuals provided in per-million-words)

City	Residuals in ERLN	Residuals in PMGZ	Residuals in RDNP
bath	28.8	28.21	22.68
brighton	68.67	73.82	28.91
cambridge	44.31	64.74	28.26
chester	35.93	16.2	21.88
derby	63.14	39.23	44.15
liverpool	95.54	55.89	67.97
manchester	38.58	33.86	21.74
oxford	74.06	90.25	37.18
york	158.02	163.38	122.68

We already know from section 6.3.1.2 that Oxford and Derby are liable to over-estimation because many of their instances do not refer to the cities. It is easy to see how a similar effect may be affecting 'Bath' (which is homonymous with the washing facility) and ‘york’ (which can also be part of a reference to New York): perhaps these over-estimations are skewing the regression analysis? Figure 6.19 shows side-by-side graphs for RDNP only, with the non-adjusted figures on the left and the adjusted figures on the right. Although the values seem slightly closer together on the adjusted than on the non-adjusted graph, the overall pattern looks very similar. The best fitting regression model of the adjusted figures is a linear model instead of a power model, and the fit is much better (R^2 of 0.52 instead of 0.36 for the non-adjusted figures) but the two models are not actually very different one from another.

Figure 6.19 Relationship between population and mentions of British cities in RDNP (non-adjusted and adjusted figures side-by-side)

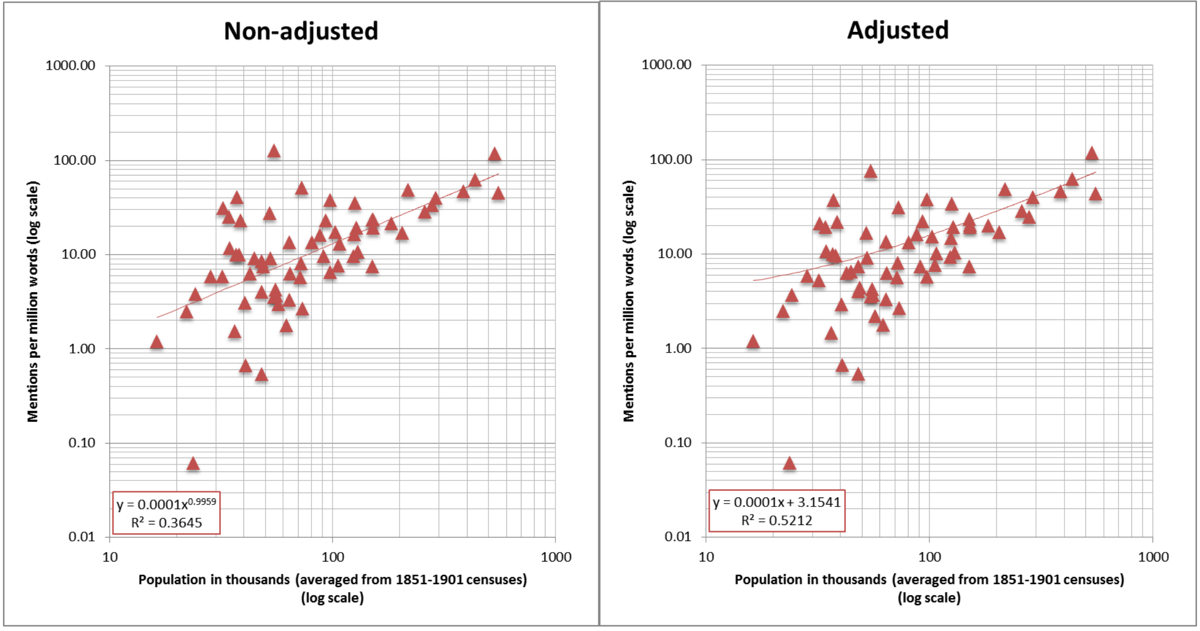
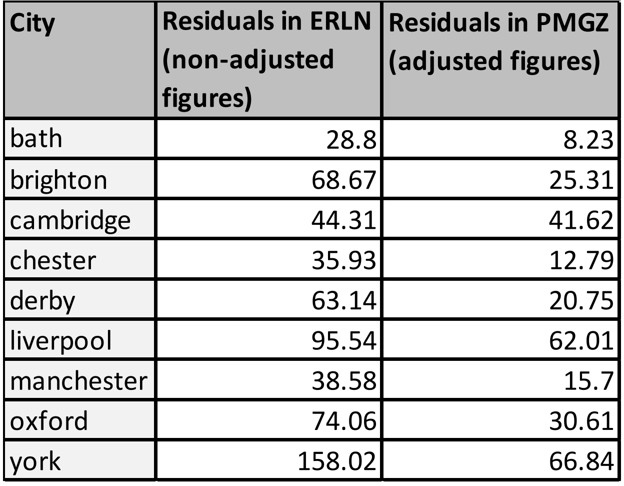


Table 6.9 shows the RDNP non-adjusted and adjusted figures side-by-side for those 9 cities mentioned more than expected in all three newspapers. Although all the figures have reduced dramatically, all figures except those for Bath are still above the new average distance of 8.24. Hence, although less dramatically so, all of the 9 cities still occur more than expected given their population even in the adjusted figures. This suggests that although the non-adjusted figures are an approximation which may render some of the values more extreme than they ‘really’ are, the approximation is accurate enough to identify overall trends. I hence continue the analysis in this section with non-adjusted figures.

Table 6.9 9 British cities which are mentioned in all 3 newspapers more often than expected given their population: non-adjusted and adjusted figures for RDNP



City	Residuals in ERLN (non-adjusted figures)	Residuals in PMGZ (adjusted figures)
bath	28.8	8.23
brighton	68.67	25.31
cambridge	44.31	41.62
chester	35.93	12.79
derby	63.14	20.75
liverpool	95.54	62.01
manchester	38.58	15.7
oxford	74.06	30.61
york	158.02	66.84

So far, only cities mentioned more often than expected in all three newspapers were considered. What of the cities mentioned more or less than expected in only one or two newspapers? Do overall geographical patterns emerge? Figure 6.20, Figure 6.21 and Figure 6.22 show the residuals (in per-million-words) for all 65 cities considered in this section. Although differences between the newspapers begin to be identifiable, these maps do not take into account the variation which is expected to occur.

Figure 6.20 Difference between observed and expected mentions of British cities given their population (1851-1901), ERLN

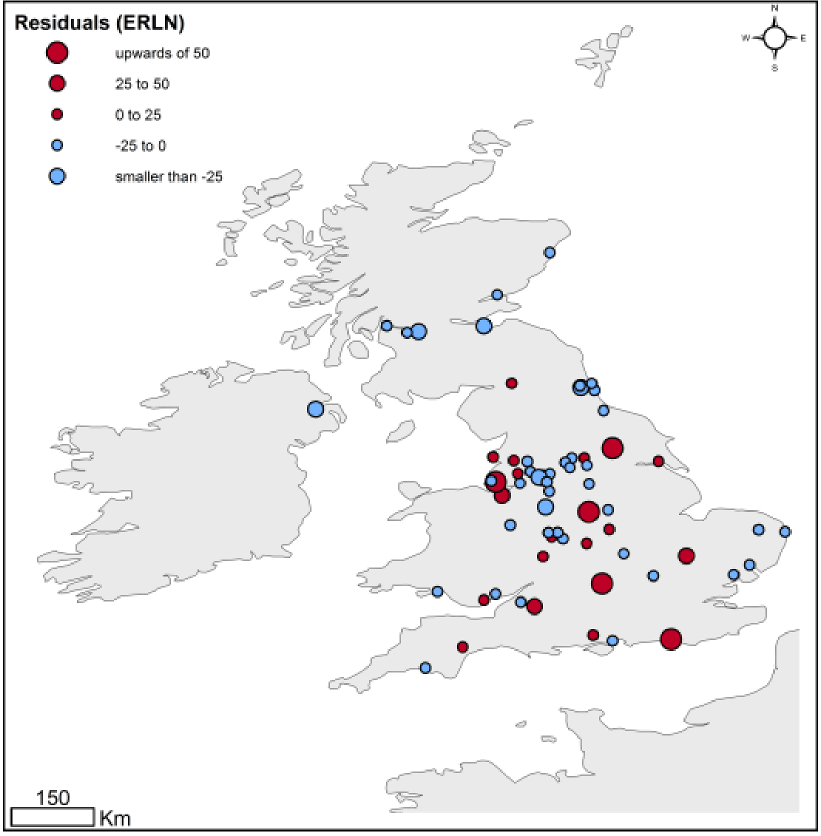


Figure 6.21 Difference between observed and expected mentions of British cities given their population (1851-1901), PMGZ

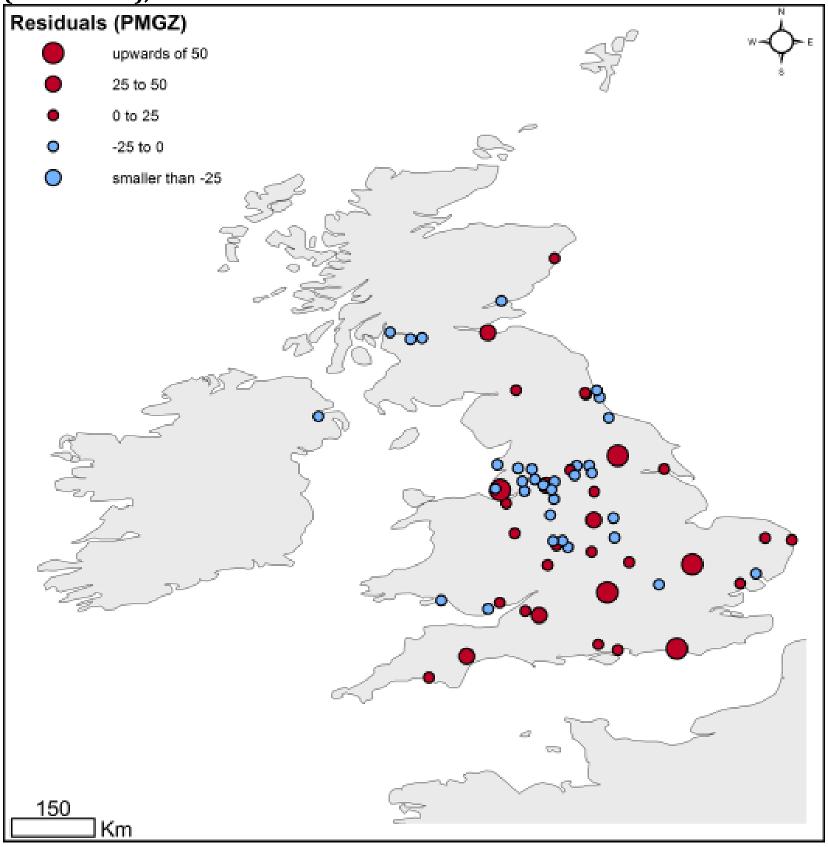


Figure 6.22 Difference between observed and expected mentions of British cities given their population (1851-1901), RDNP

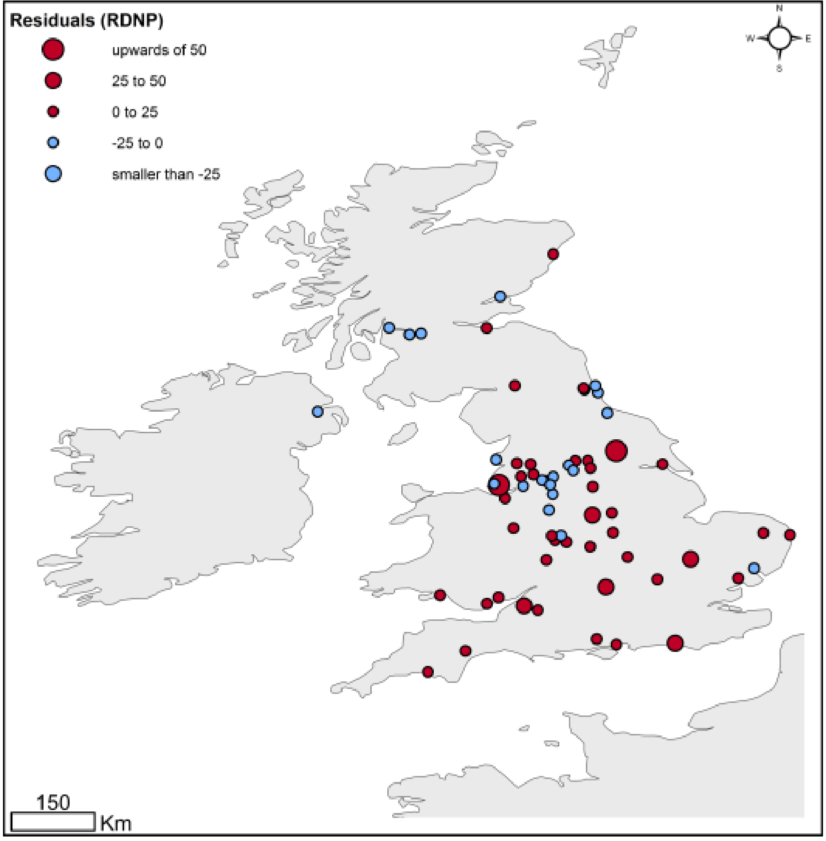


Figure 6.23, Figure 6.24 and Figure 6.25 show the same data as Figure 6.20, Figure 6.21 and Figure 6.22, with the same categories (the value boundaries are the same), with the exception that values within the average distance to the predicted values have been grouped into a central category (in black). The coloured points on these figure are hence those which are more extreme than the average distance from the predicted values – those which could be considered 'surprising'.

Figure 6.23 Above average differences between observed and expected mentions of British cities given their population (1951-1901), ERLN

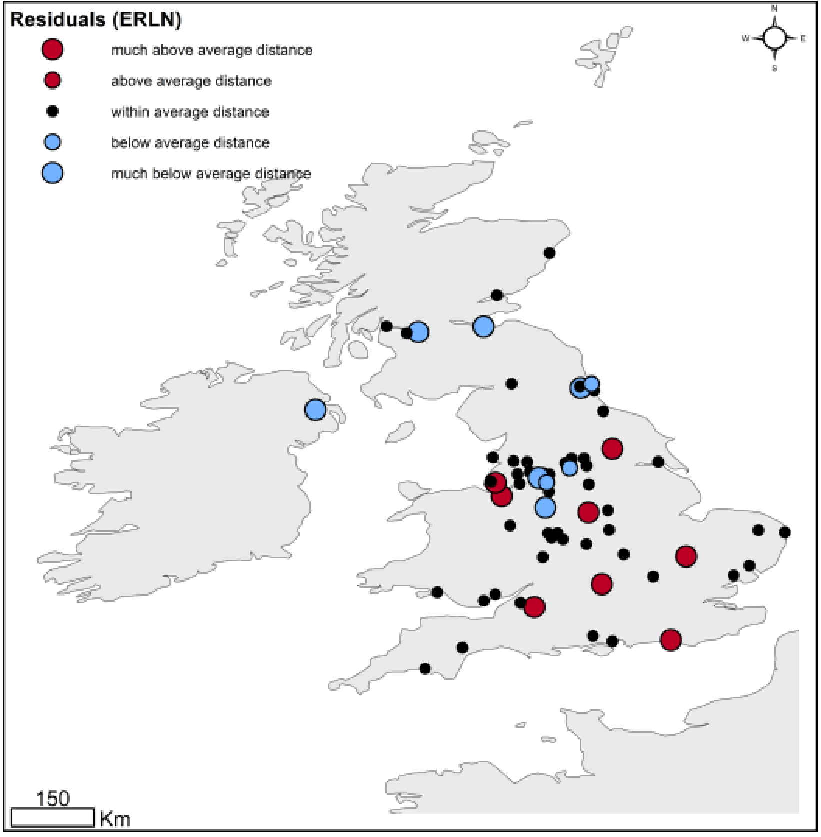


Figure 6.24 Above average differences between observed and expected mentions of British cities given their population (1951-1901), PMGZ

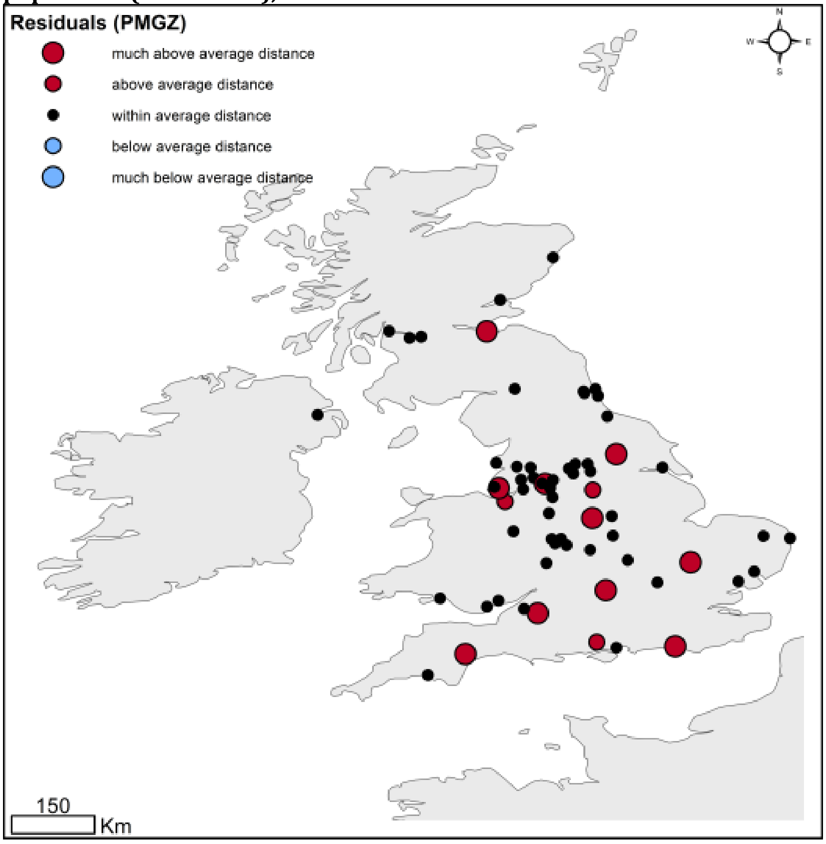
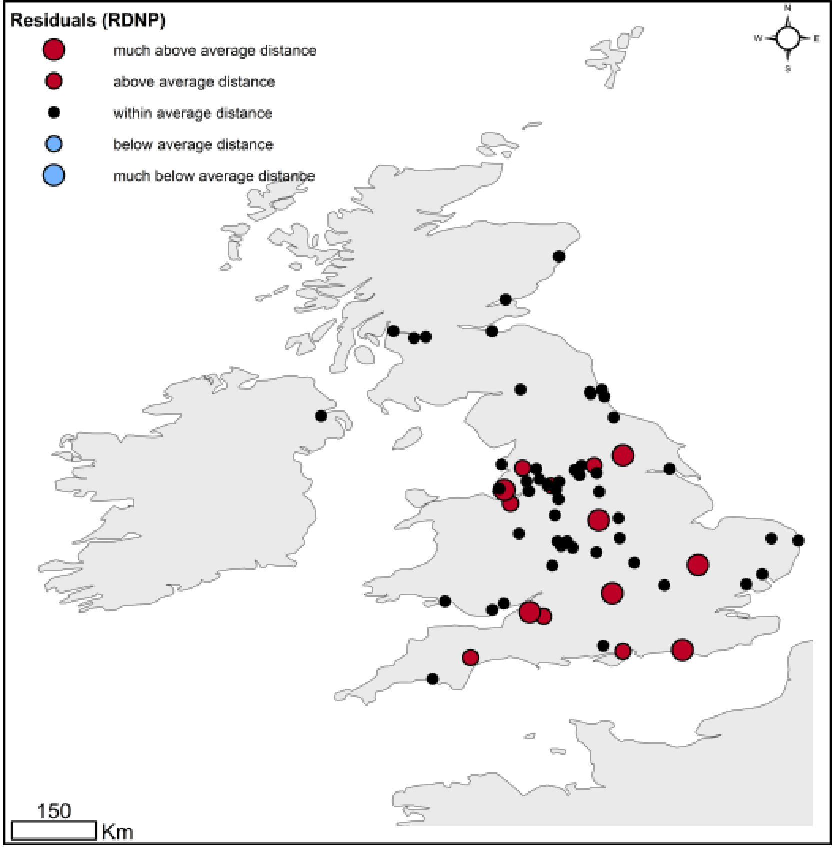


Figure 6.25 Above average differences between observed and expected mentions of British cities given their population (1951-1901), RDNP



Perhaps the clearest tendency discernible from Figure 6.23, Figure 6.24 and Figure 6.25 is for cities in England, and especially but not exclusively the South of England, to be mentioned more than expected given their population. This is a discernible trend in all three newspapers, and is in contrast to cities in the North of England, Scotland, and Ireland. In ERLN, there is even a relatively strong tendency for places in the North of England, Scotland, and Ireland, to be mentioned *less* than expected given their population. The exception to this pattern is Edinburgh, which is mentioned more than expected given its population in two out of three newspapers – considerably so in PMGZ, and slightly so in RDNP.

Cities in the area around Manchester (in the North-West) are mentioned less than would be expected given their population in all three papers, but most dramatically in ERLN where 4 cities (Stockport, Huddersfield, Stoke, Salford) are mentioned much less than would be expected given their population. The more extreme character of the pattern in ERLN may be simply due to the earlier time-period covered in ERLN: the places in the industrial pocket around

Manchester grow very fast in population in the second half of the c19th, and since ERLN begins publication earlier than the other two newspapers, the *average* population figures for these cities may be too high for the time-period covered by ERLN. However, since the trend is discernible in all three papers, it seems likely that the pattern overall is capturing a 'real' trend. Indeed, the pattern is especially unexpected in RDNP, a newspaper which supposedly circulates especially among the working-class population (see section 6.2.2). Also of interest is that Chester and Liverpool escape this trend and are in fact mentioned more than expected given their population in all three papers.

6.4.3 MENTIONS AND ARTICLE-GENRES

In section 6.4.1, it was suggested that different article-genres may be associated with different concentrations of mentions of place-names. Is this the case? It is difficult to answer this fully, since that would require being able to consider all place-names, which would require geo-parsing the whole corpus. Nevertheless, looking at frequent or populous cities is a good place to start. In this section, I will consider the 20 most populous British cities mentioned in the three newspapers (as determined using the averaged population figures from the previous subsection). Figure 6.26 and Figure 6.27 show how mentions of these 20 British cities are distributed across article genres in all three newspapers.

Figure 6.26 Distribution of mentions of 20 most populous British cities mentioned in ERLN, RDNP and PMGZ, per article-genre, with adverts

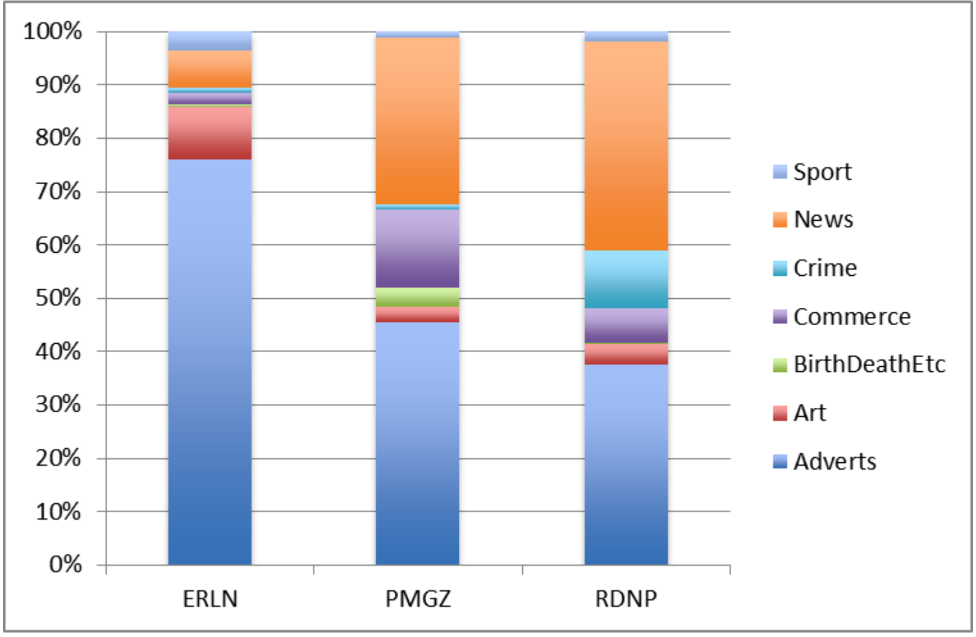
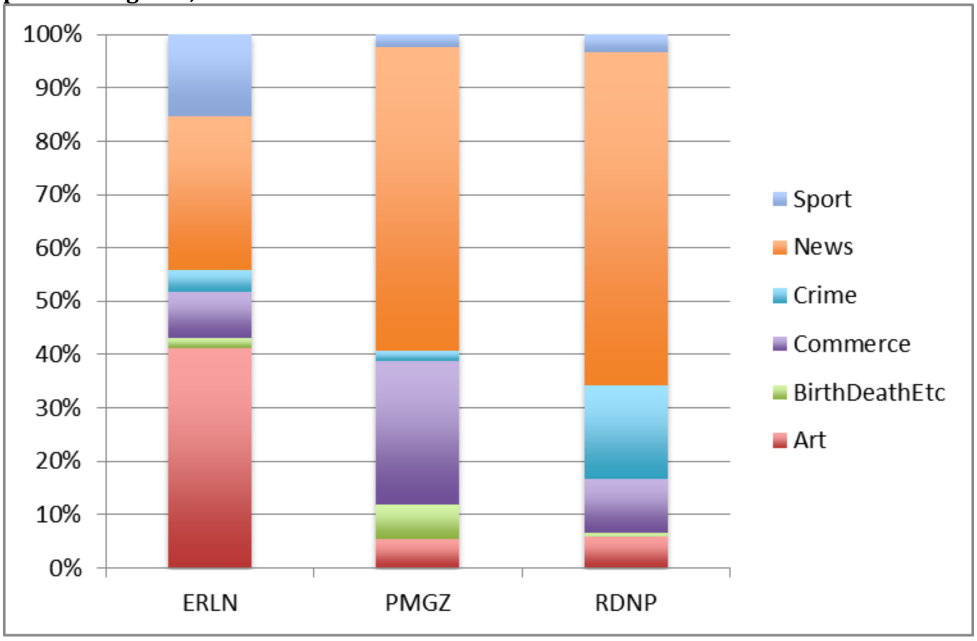


Figure 6.27 Distribution of mentions of 20 most populous British cities mentioned in ERLN, RDNP and PMGZ, per article-genre, without adverts



It is clear that there is a remarkable difference between the three newspapers. In ERLN, 76% of mentions of these British cities occur in advertisements, contrasted with 46% of mentions in advertisements in PMGZ and 38% in RDNP. The other category with more than 20% of mentions in the other two newspapers is News, which attracts 32% of mentions of the British cities in PMGZ, and 39% of mentions in RDNP. In contrast, Art is the second most important category in ERLN, far behind with only 9% of mentions. These observations lend support to the hypothesis formulated in section 6.4.1 that ERLN has relatively more mentions of place-names than the other two newspapers because it contains more adverts and arts content than the other two papers (as seen in section 6.2.3).

Is there much variation in the distribution among article-genres from one city to another? Figure 6.28, Figure 6.29, Figure 6.30, Figure 6.31, Figure 6.32 and Figure 6.33 show the distribution among article-genres for each of the 20 most populous cities mentioned in the three newspapers.

Figure 6.28 Distribution across article-types of raw mentions of 20 most populous British cities mentioned in ERLN, with adverts

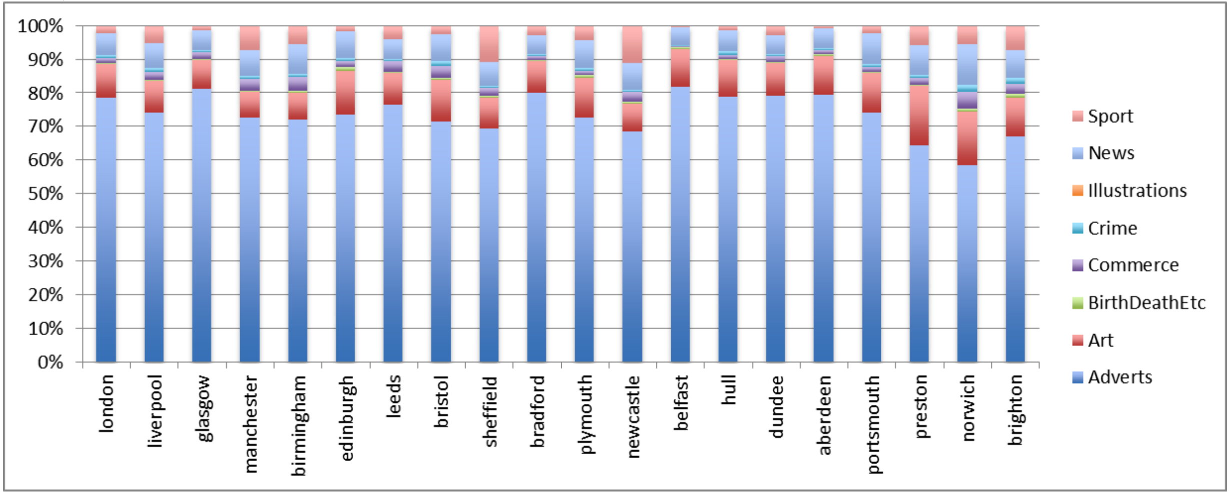


Figure 6.29 Distribution across article-types of raw mentions of 20 most populous British cities mentioned in ERLN, without adverts

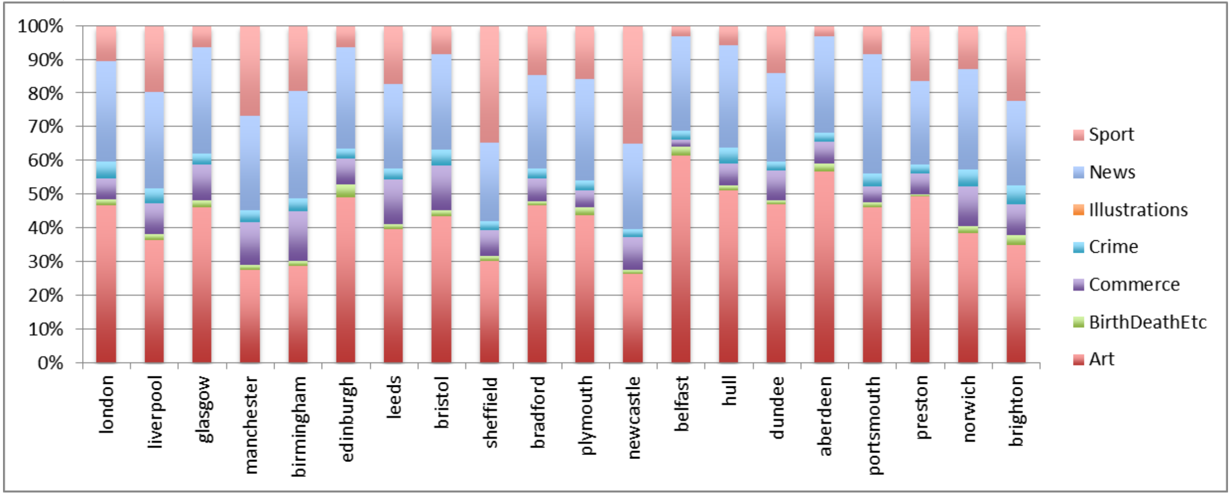


Figure 6.30 Distribution across article-types of raw mentions of 20 most populous British cities mentioned in PMGZ, with adverts

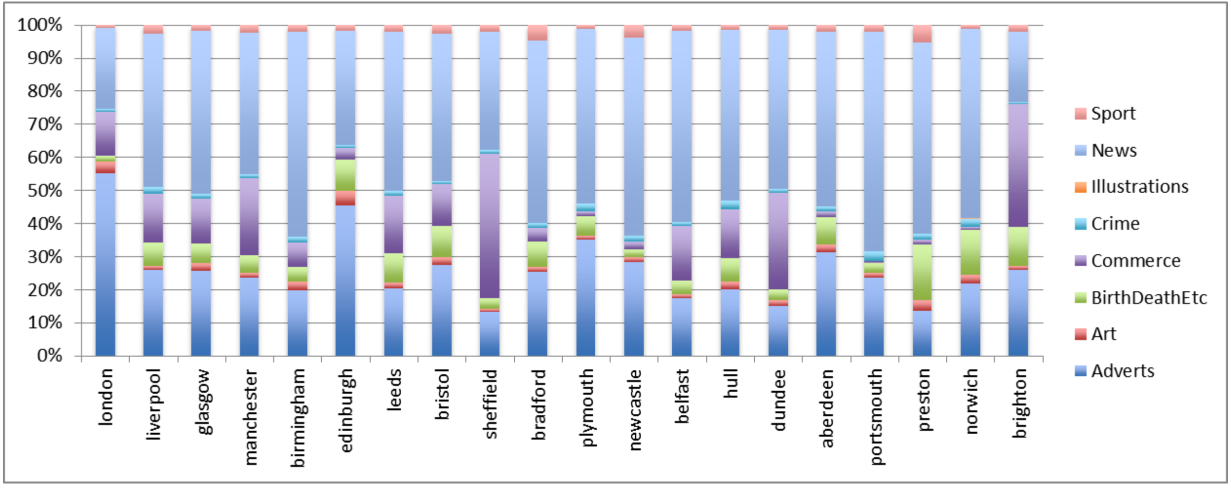


Figure 6.31 Distribution across article-types of raw mentions of 20 most populous British cities mentioned in PMGZ, without adverts

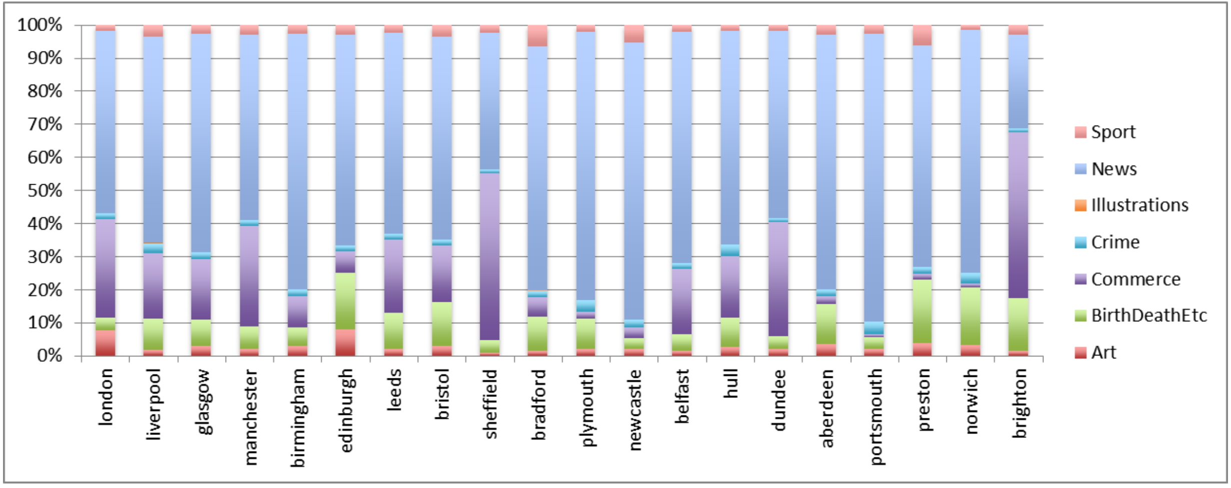


Figure 6.32 Distribution across article-types of raw mentions of 20 most populous British cities mentioned in RDNP, with adverts

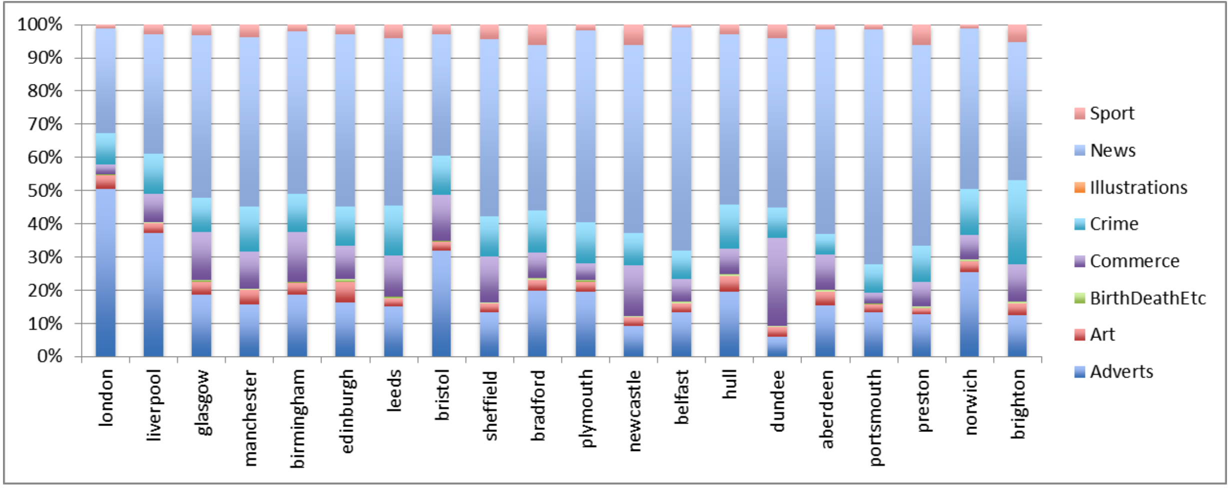
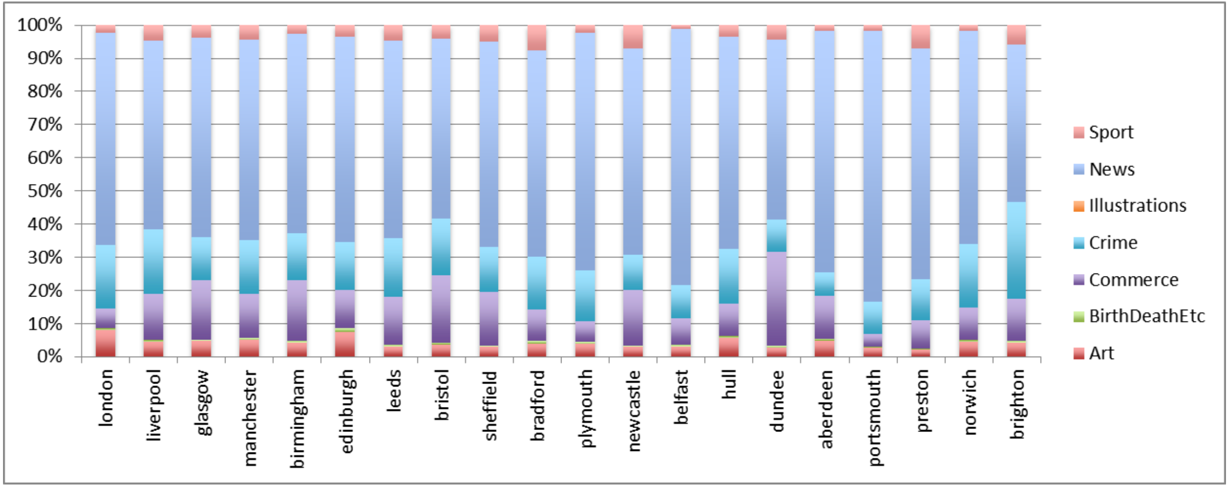


Figure 6.33 Distribution across article-types of raw mentions of 20 most populous British cities mentioned in RDNP, without adverts



The most striking pattern shown in Figure 6.28 to Figure 6.33 is that there is more similarity between the distribution across article-genres of different cities mentioned within the same newspaper, than there is for a single city across newspapers. This suggests that looking at the generic distribution of a place-name tells us more about the newspaper in which it is mentioned than about the specific discursive treatment of that place. Nevertheless, cities do have different profiles. In PMGZ, for example, London, Edinburgh, Dundee, Sheffield and Brighton appear to have unusual generic distributions. In RDNP, London, Bristol, Dundee and Brighton are the ones which stand out. In contrast, London, Dundee and Brighton do not particularly stand out in ERLN. In fact, the profiles of cities look more similar in ERLN than they do in the other two newspapers – for the most part simply because the cities are mostly mentioned in adverts, which dwarfs the differences between the distribution of the remaining mentions among the remaining genres in ERLN.

It is also interesting to look at the cities which were discussed in the previous stage. Glasgow and Edinburgh, for example – which are relatively neglected in ERLN, positively emphasized in PMGZ, and somewhat in the middle in RDNP – have very different generic distributions in the different newspapers. The two cities get over 40% of their mentions from news in RDNP, but not even 10% in ERLN; in PMGZ, the profile of both cities are more different from one another, with around 30% of mentions in news for Edinburgh and 50% for Glasgow.

6.4.4 *ATTEMPTING A MORE EXHAUSTIVE SURVEY OF BRITISH CITIES MENTIONED*

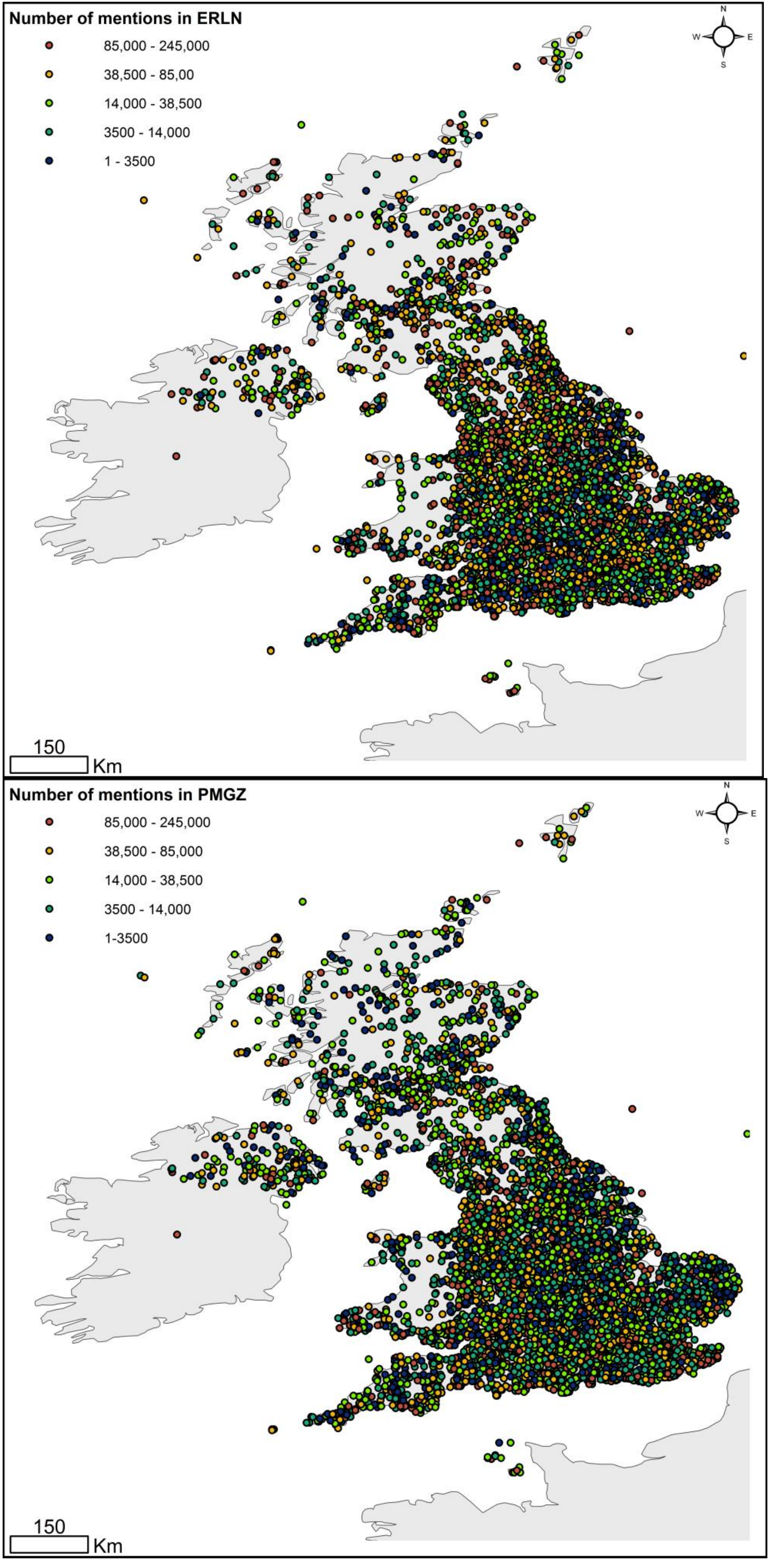
How much more information would we get about patterns of mentions of places in the three newspapers if we could consider all of the cities mentioned in the newspapers, and not just a small selection as we have done so far? Using the approach outlined in section 6.3.1.3.3, of matching a list of words tagged 'Z2' to a gazetteer, allows this question to be addressed. I will illustrate the procedure by reference to RDNP, but see Table 6.10 for the equivalent figures for the two other papers. There are over 22 million word-types in RDNP, 296 thousand of which have received 'Z2' tags. Searching for common place-names between the words tagged as 'Z2' and the *geonames*[6] list of over 30 thousand British place-names, I find 4,346 candidates. These candidates are shown, with colour indicating their frequency, on Figure 6.34.

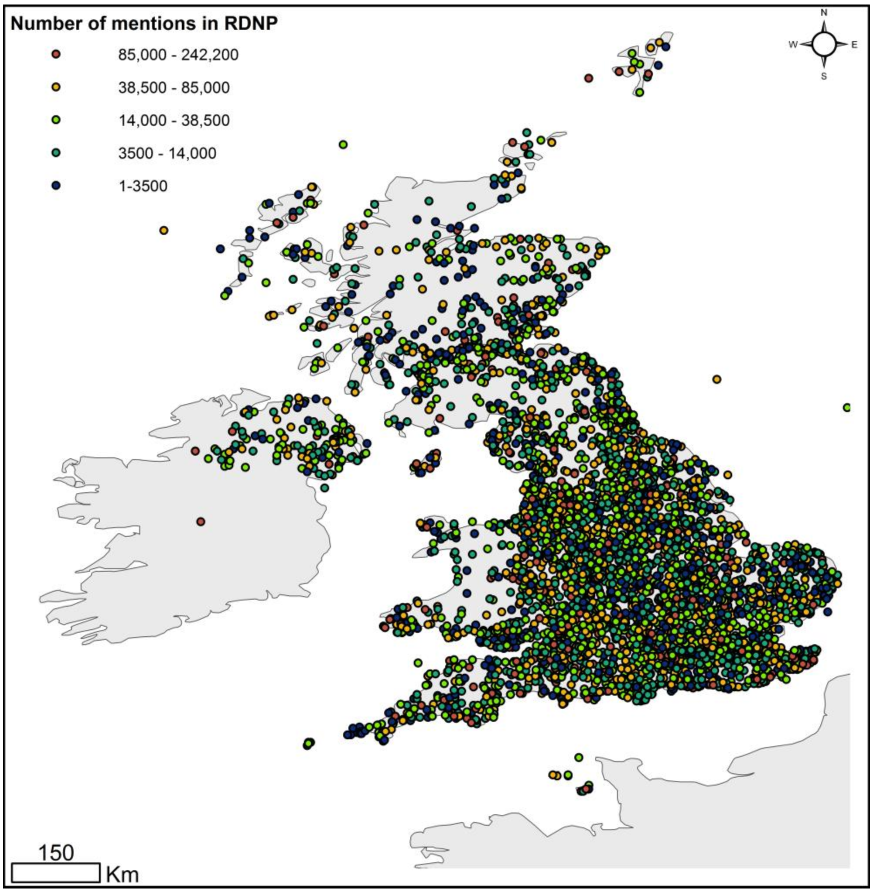
Table 6.10 Number of words tagged 'Z2' which occur in a geonames list of British place-names

	ERLN	**PMGZ**	**RDNP**
Total types	24,333,352	23,386,491	22,153,306
Types tagged 'Z2'	394,996	394,996	295,530
Types tagged 'Z2' which match entries in the geonames list of British place-names (incl. homonyms)	3749 (5,031)	4500 (5,871)	3,163 (4,346)

[6] A free online geographical database, see www.geonames.org.

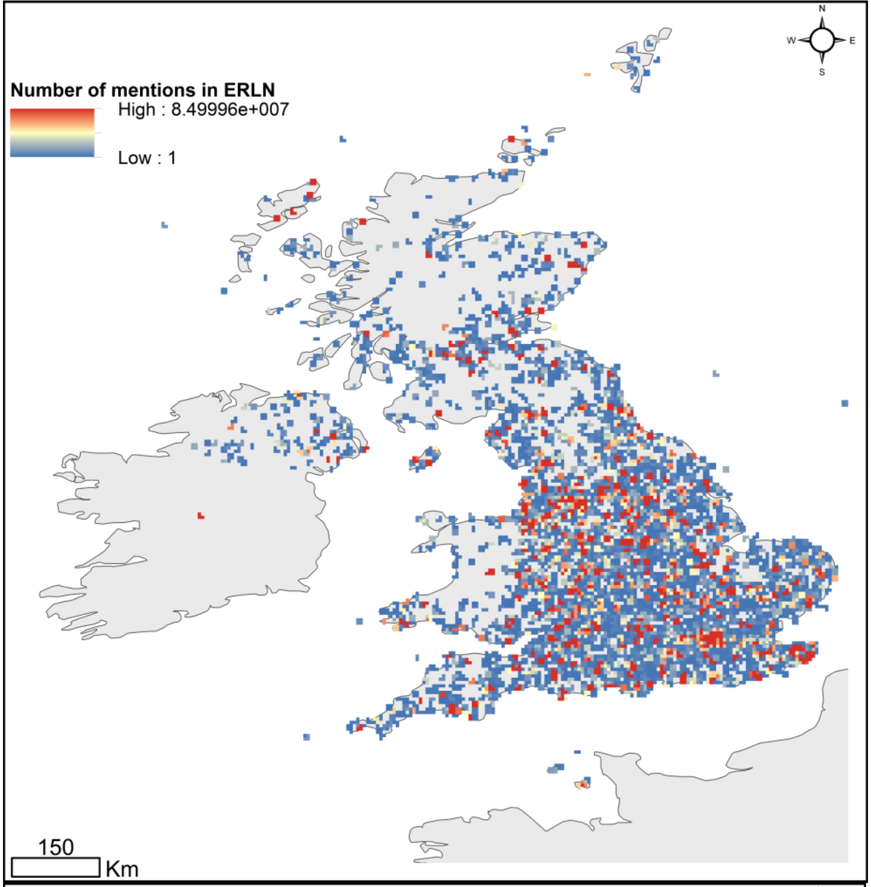
Figure 6.34 Words tagged as 'Z2' in ERLN, PMGZ and RDNP which occur in a geonames list of British place-names, with their overall number of mentions

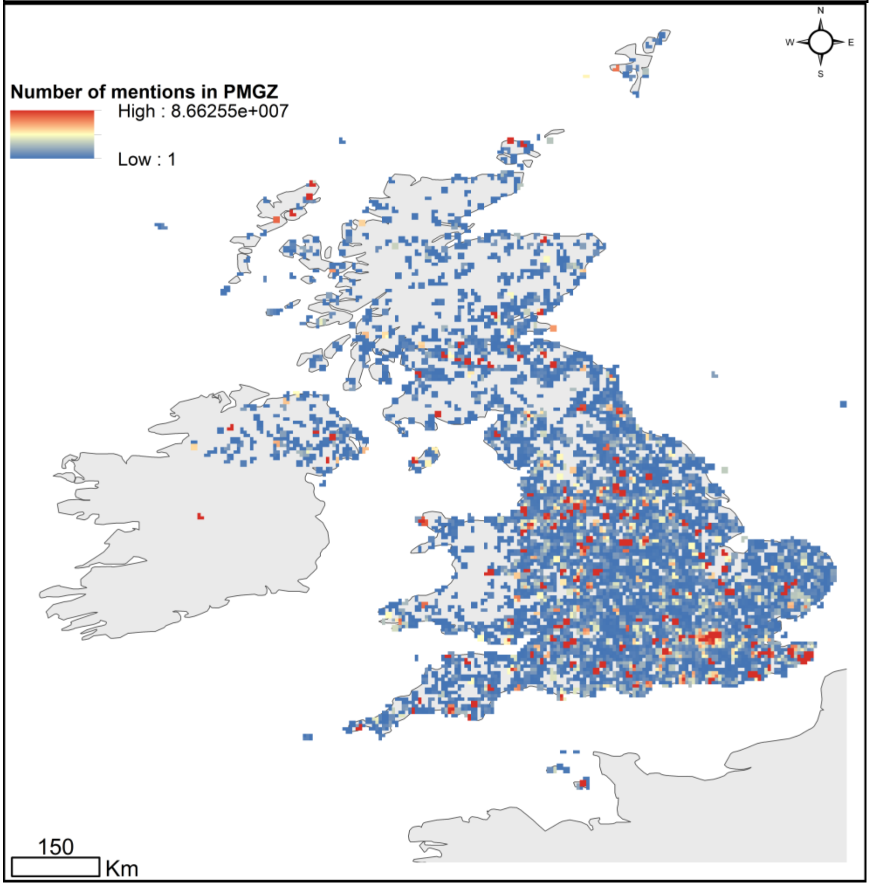


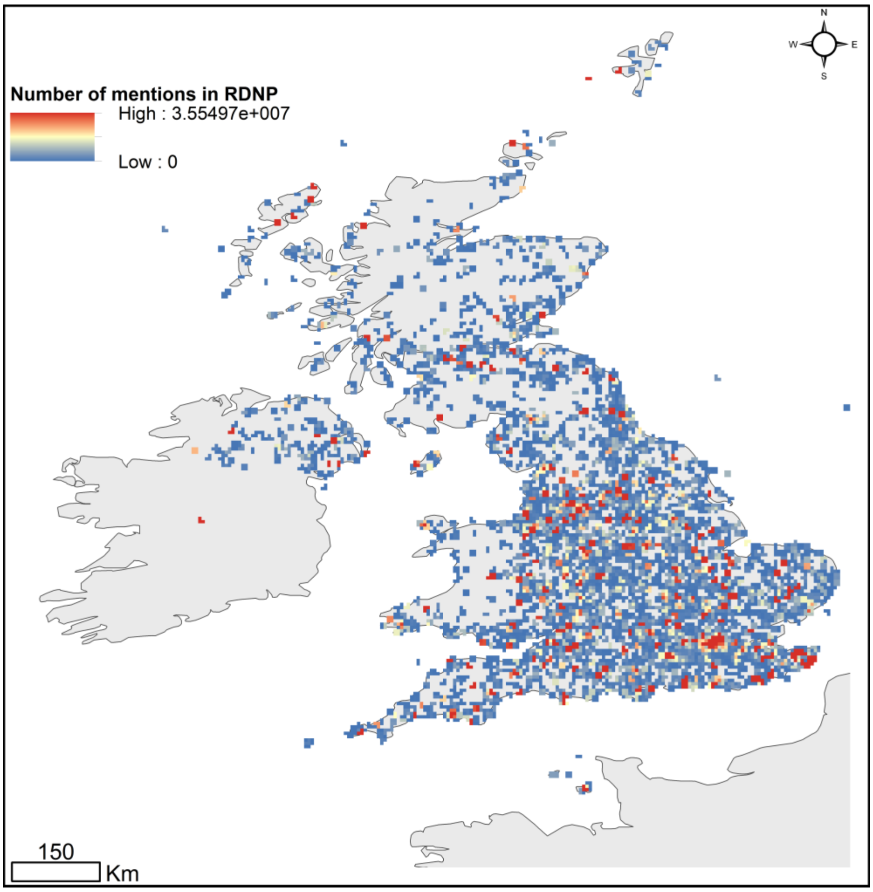


It is difficult to conclude anything from Figure 6.34 – the density of points is such that most of the UK is covered in dots. A solution to this is to represent the data as a density map instead. Figure 6.35 shows the same data, but each cell on the map has been coloured accorded to the number of mentions which any place-name within its coverage has attracted. The new map is easier to read, but no clear overall trend emerges.

Figure 6.35 Words tagged as 'Z2' in ERLN, PMGZ and RDNP which occur in a geonames list of British place-names, with their overall number of mentions (density)

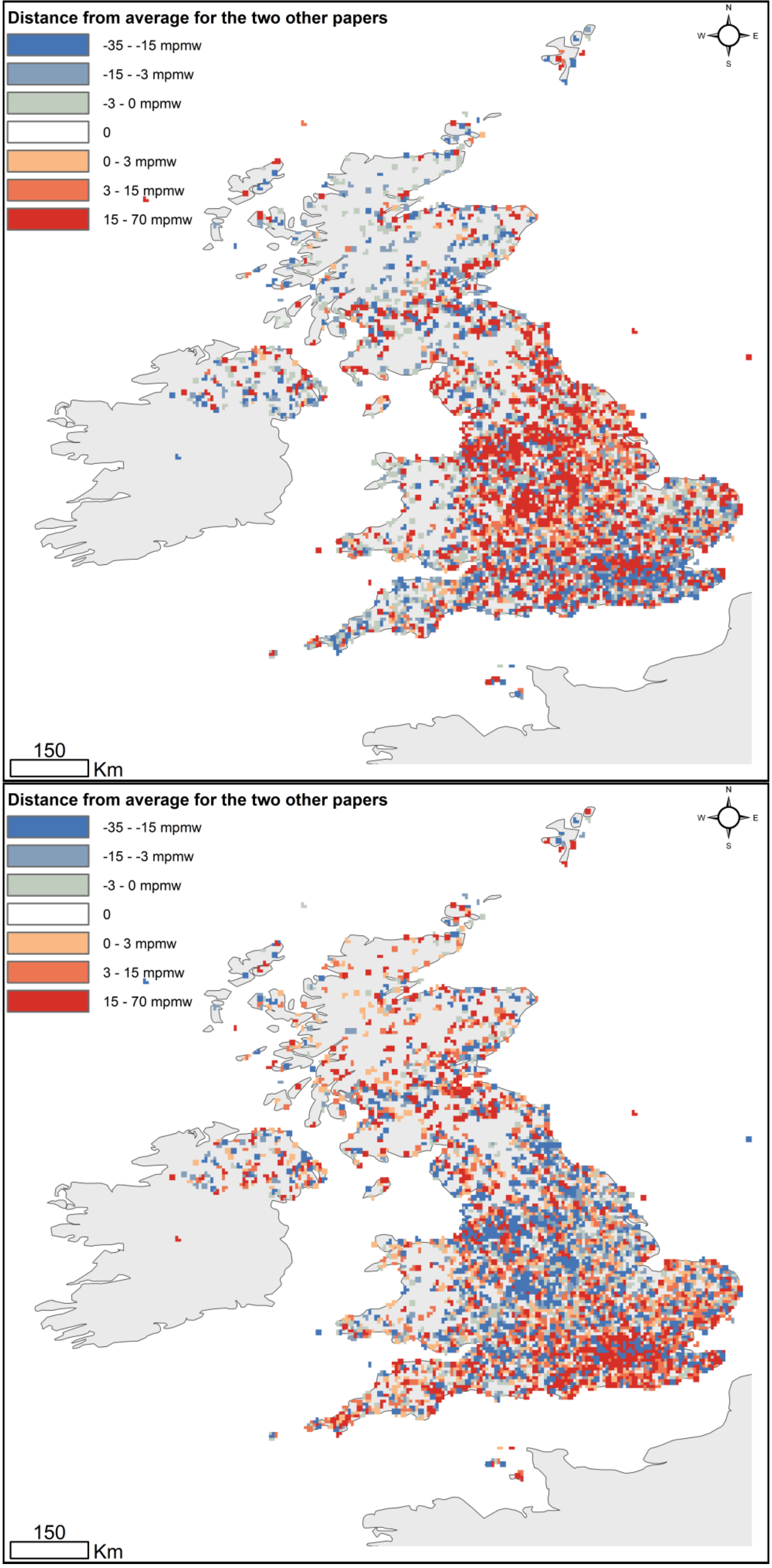


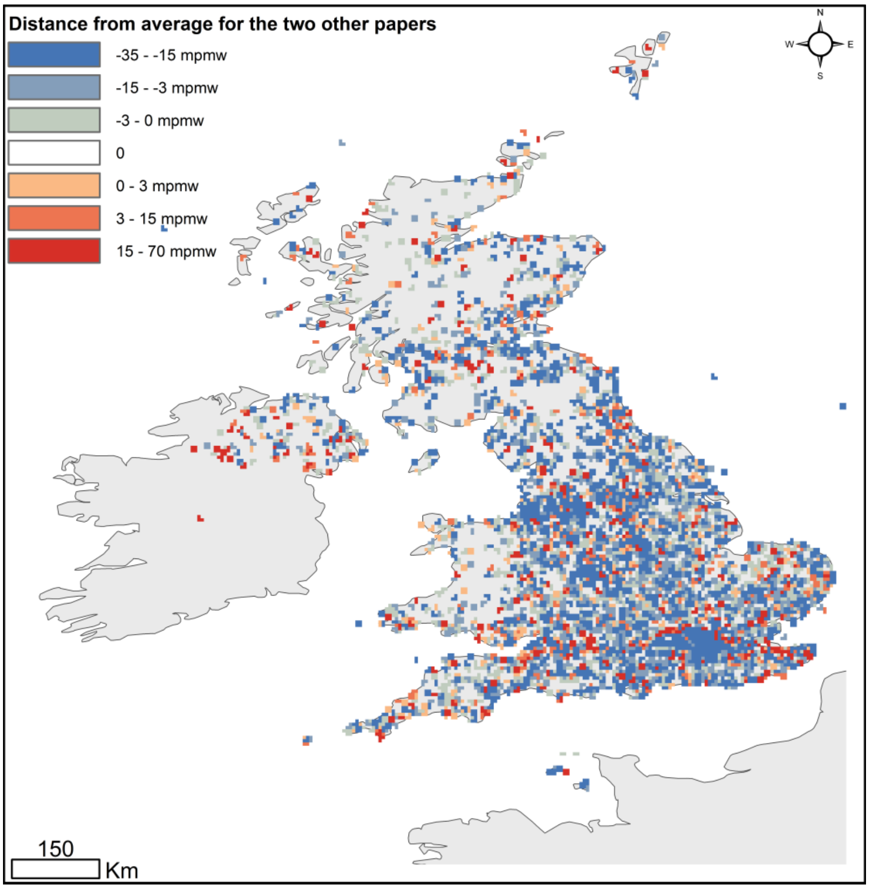




In part, the difficulty in interpreting the maps stems from the lack of any benchmark. What we want to know, after all, is whether there are particular geographical patterns of over- or under-emphasis. One question is, for example, 'which places (or regions) are mentioned more or less in a given newspaper, as compared to the other two newspapers?'. This can be answered by comparing the values for the different newspapers. In Figure 6.36, the number of mentions per million words of each place has been averaged for two newspapers, and the average subtracted from the value for the third, in order to obtain positive values where a place is mentioned more often in the third newspaper than in the other two, and negative values where a place is mentioned less often. The figures show positive values in warm colours and negative values in cold colours, and the data is represented in the form of a density map, to be more easily interpretable.

Figure 6.36 Places which occur more or less often in one newspaper compared to the other two (density) (ERLN, PMGZ, RDNP) [in mentions per million words (mpmw)]





These maps suggest that, generally speaking, British places are mentioned more often (relative to overall content) in ERLN than in the other two papers, and that places are mentioned least often (relative to overall content) in RDNP. This observation matches what was found in section 6.4.1. In terms of regions which consistently attract over- or under-emphasis in one paper as compared to the other two, there do not seem to be overwhelming patterns, although ERLN seems to show a tendency towards mentioning places in the Midlands more than the other two papers (i.e. the contrast seems greatest there), and PMGZ seems to mention places in the South of England more than the other two papers. Again, these observations match those found in section 6.4.1.

However, since no filtering other than excluding place-names with several entries has been applied to the results, the data includes a certain amount of noise. In particular, places which are homonyms of famous people's names (such as 'Gladstone' and 'Newton'), places which are homonyms of common words (such as 'more' and 'stone'), places which have more famous namesakes abroad (e.g. 'Kingstown', 'Melbourne') and places which are homonyms of common first or family names (such as 'Charles' and 'Paul') will have inflated, and in some cases vastly inflated, the reported frequencies. Another issue which needs considering: when trying to answer a question such as 'what places are mentioned in a corpus?' is how to deal with different scales within one analysis; for example, how should 'Maidstone', 'Kent' and 'England'[7] be represented (since each is contained inside the next)? Finally, place-names which refer to more than one place within the UK also need to be dealt with.

Other interesting lines of subsequent enquiry which are beyond the scope of this chapter include: what would an analysis of the relationship between population and mention of place-names reveal when dealing with a geo-parsed list of words tagged Z2? And to what extent could a geo-parsed list of words tagged Z2 be useful as benchmark against which to compare findings obtained by exploring questions of the type 'what places are mentioned in relation to theme X?'?

6.5 Conclusion

In this chapter, various approaches to answering the question 'which places are mentioned in this corpus?' have been explored. Since geo-parsing a whole corpus of the size of the datasets at hand is not technically feasible (yet), alternative approaches have been tested. Three promising approaches are identifying place-names from lists of words tagged with the semantic tag for geographic names 'Z2'; starting from a list of place-names obtained from an external source such as a population table; and geo-parsing a list of words tagged 'Z2'. Whilst the two first approaches only feasibly allow analysis of a limited number of place-names, the last approach allows a large number of place-names to be considered. All three approaches involve a degree of approximation. In particular, all the approaches surveyed require dealing with the polysemic nature of most place-names.

[7] Maidstone is a town located in Kent, a county (i.e. a region) of England.

In terms of feasibility, although all approaches involve some degree of manual work, the last approach is probably the most time-consuming: the amount of manual work required to produce 'clean' results is extensive. Overall, however, the rough results simply reinforce observations obtained using the external-source (population table) approach. Hence this external-source approach may prove more effective relative to the level of analyst time that must be invested.

Ultimately, it is also clear that although each method has allowed us to answer specific questions which cannot quite be answered by the other two, in all cases, there are still issues associated with combining both scale and granularity in a single analysis. In fact, in all cases, it remains necessary to select which dimension will be analysed in some depth, and which dimensions will be considered only in a fairly general way. A choice needs to be made between working at different geographical scales, comparing different genres, and achieving a certain amount of granularity in a diachronic analysis.

Let us now summarize this chapter's findings related to the place-names mentioned in three Victorian British newspapers, *The Era* (ERLN), *The Pall Mall Gazette* (PMGZ) and *Reynold's Newspaper* (RDNP). Using the Z2-list approach suggested that all three newspapers mention cities located across the UK, but that ERLN tended to mention all British cities more often than the other two newspapers. It was hypothesized that this difference may be related to differences in the generic make-up of the newspapers. Exploring the relationship between genre and mentions of the most populous cities provided support for this hypothesis, with most mentions of place-names in ERLN occurring in Adverts and Arts articles, two genres which are more represented in ERLN than in the other two newspapers. In fact, looking at the distribution of mentions of specific cities across genres within a newspaper suggested that the generic distribution of mentions of a city within a newspaper tells us more about that newspaper than it does about treatment of that city, because generic distributions of cities were more similar between cities within the same newspaper than between the same city across newspapers.

Another hypothesis formulated by looking at the most mentioned cities in the three newspapers was that there was a relationship between mentions of cities and population. This relationship was explored using Mitchell and Deane's (1971) population table. The findings suggested that there is indeed a relationship between mentions of cities and their population, and that this relationship is present in all three newspapers, although strongest in ERLN.

Finally, two more specific methodological findings were, first, that it is preferable to use full counts rather than Z2 counts because the 'Z2' tag does not reliably exclude all cases where the place-name does not refer to the place, and also erroneously excludes some cases where the place-name does refer to the place; and second, that although it is possible to geo-parse a list of words tagged Z2 as a potential substitute for geo-parsing the whole corpus, the method involves substantial manual work for little additional insight compared to investigating a limited but principled list of place-names.

7 INVESTIGATING DISCOURSES SURROUNDING PLACE-NAMES

7.1 INTRODUCTION

In the previous chapter, I considered methods for finding out which places were mentioned in large amounts of text. Once we know that a place is mentioned in a set of texts, how can we answer the question 'what is being said about this place in these texts?'. To illustrate approaches to answering that question, this chapter will explore discourses surrounding 'France' and 'Russia' in three newspapers, *The Era* (ERLN), *The Pall Mall Gazette* (PMGZ) and *Reynold's Newspaper* (RDNP). An introduction to these newspapers is provided in section 6.2.

The first part of this chapter provides some brief background about the relationship between Britain, France and Russia in the relevant period and sets out some hypotheses as to what may be found about France and Russia in British newspapers. The rest of the chapter explores approaches to investigating the discourses which surround these two country-names. Section 7.3 discusses overall trends revealed by looking at frequencies of mentions of the two countries. Section 7.4 introduces what I call the *global approach*, a collocation-via-significance approach (see section 2.3.2.3.1). Section 7.5 introduces what I dub the *sampling approach*, a collocation-via-concordance approach (see section 2.3.2.3.1).

7.2 BACKGROUND: THE RELATIONSHIP BETWEEN FRANCE, RUSSIA AND GREAT BRITAIN

Nineteenth-century Europe is considered to have been exceptionally peaceful. This peace was maintained by a complex system of interrelationships between nations which emerged as a consequence of the Congress of Vienna. This 'balance of power' was enacted within intricate diplomatic dances which took place both on a continental and global scale: the era of imperialism meant that countries vied for influence both within and outside of Europe. France and Russia were perhaps the two most serious global rivals of Great Britain. The fragile balance of power meant that although skirmishes and direct military confrontations did occur among Europe's Great Powers, they did not side with one another in fixed configurations – Britain could side with Russia on one occasion and be against her in another, sometimes even simultaneous, conflict (Gildea 1996: 55-59; Schroeder 2000: 159-60,64).

The relationship among the countries was thus an interesting, complex one, and it is pertinent to ask how British people might have conceived of France and Russia. Would the countries have been thought of mostly through the 'lens' of rivalry? Or would they have been perceived of as fellow 'Europeans', partners in the complex dance of the 'balance of power'? Would British people have held deeply engrained suspicions, perhaps outright hatred, of these 'others', or more nuanced and even positive views?

It is of course impossible to know what perceptions the British public actually held. This is impossible at any time (thoughts are invisible even to neurology), but particularly for publics of the past, since we no longer have direct informants to interview. Newspapers may serve as an alternative source of insight into discourses circulating at a point in the past, and these discourses in turn may be taken as indicative of perceptions of that time. This is in fact emphatically argued by Gleason (1950):

> Each of the many modes in which opinion finds expression, letters, speeches, books, pamphlets, periodicals, and newspapers, as well as government documents, has its special value. Newspapers are perhaps the single richest source. The frequency of their publication and their general dependence upon public favor render unlikely their total disregard of any important element in the formation of public sentiment. Since they are the primary medium through which opinion becomes articulate, it may be assumed that at least one organ reflected, if it did not actually generate, the opinion of each significant segment of the community. (Gleason 1950: 7)

Contra Gleason, I would argue that how representative such newspapers and discourses are of views held by the wider population is to some extent a moot point – we can only begin to answer this question by surveying ever more discourses. Sooner or later, we must fall back on the historical imagination in order to consider whether the discursive landscape we have uncovered is a plausible one, without significant omission.

What might we expect to find when we explore discourses surrounding Russia and France in British Victorian newspapers? In very general terms, since France and Russia are rival military powers to Britain, we might formulate the following hypotheses:

(*h1*) We might expect discussions surrounding the two countries to focus on reports of their actual or possible involvement in armed conflicts. This expectation would be met if an important proportion of these discussions related to military matters.

(*h2*) We might expect mentions of the countries to be driven by the onset or imminent onset of armed conflicts involving them. This expectation would be met if peaks in the mentions of countries corresponded in time with the onset or imminent onset of armed conflicts involving the countries.

(*h3*) We might expect most mentions of the countries to occur in news articles. This expectation would be met if these mentions occurred disproportionately in news articles relative to the proportion of the corpora made up be news articles.

(*h4*) We might expect a degree of negative bias towards the countries. This could be in the form of fear, suspicion or hatred. On the other hand, the rivalry between countries might instead be associated with sentiments of respect, admiration and jealousy. It is harder to define what evidence would support this expectation. Minimally, the presence of evaluative associations (positive or negative) alongside mentions of the two countries would support this hypothesis.

In addition, since France is geographically closer to Britain than Russia, we might also formulate the following hypotheses regarding *differences* in the representations of France and Russia:

(*h5*) We might expect more travel reports, or reports of events happening inside the country, for France than for Russia. These expectations rest on the assumption that greater geographical proximity would allow for more extensive travel to and from the country (and hence more commercial and cultural ties) as well as better journalistic networks (hence more information being gathered and returned to Britain). This expectation would be met if France was mentioned more than Russia, and if it was mentioned in a more diverse set of contexts than Russia.

(*h6*) We might also expect that more would be known about France than Russia, possibly thus leading to more prejudice towards Russia than France. As with *h4*, it is harder to define what evidence would support this expectation. We might consider it met if France was represented using more diverse *discursive strategies* – defined in a relatively open-ended way.

These hypotheses are fairly general, and not derived from prior literature. It would be possible to formulate further hypotheses related to the differences in representations of the two countries in the three newspapers under consideration. Formulating such hypotheses would, however, require making extensive reference to social and political research into the relevant period(s), which would unduly lengthen the present discussion.

Nevertheless, to help to situate and to reflect on the findings presented in the rest of the chapter, let us consider a restricted subset of the existing relevant scholarship. The body of work on Victorian British perceptions and discursive representations of France and Russia is not extensive, but neither is it entirely non-existent. France and Russia are both discussed as the objects of negative sentiments, although the extent of these feelings is debated. A seminal work on 'Russophobia' is Gleason (1950), which discusses the existence and manufacture of anti-

Russian sentiment in Britain in the early Victorian period. Other studies of British perceptions of Russia in the Victorian period include Hughes (2011) and Hughes (2015). Studies of British perceptions of France in the Victorian period include Lewis (2014) and Parry (2001); I am not aware of any work on France comparable in extent to Gleason (1950). These publications make a series of broad and specific arguments pertaining to the representations of these countries, overall and in relation to specific historic episodes. It is beyond the scope of this thesis to review them all in depth. Key points made by these authors include the following:

- In the British imagination, Russia had a two-sided, contradictory image. It was conceived as a distant, alien, other, both dangerous and yet also exciting curiosity – 'both unsettling yet intriguingly exotic', 'a dangerous imperial rival', the more so because it was 'profoundly backward' (Hughes 2015: 1-2).
- There was a recurrent 'Russophobic' undertone in British public discourse. Yet Russophobia was something of a paradox. It seems to have been based on the fear that 'Russia's control of the Straits would endanger Britain's Levantine trade, her naval power in the Mediterranean, and her position in India'; yet as a consequence of this fear, Britain 'pursued a policy designed to preserve the independence and the territorial integrity of Turkey', an aim precisely in line with Russia's policy (Gleason 1950: 2).
- France was considered a (more) immediate threat to Britain (than other European powers, presumably including Russia) during the Victorian period, in part because it was thought that France had ambitions and that to restore France's self-image, Napoleon III in particular would need to succeed where Napoleon I had failed: in invading Britain (Parry 2001).
- France was used in public discourse as a moral counterpart against which Britain could define itself in self-flattering ways (Colley 1992 cited in Lewis 2014: 209; Parry 2001).

7.3 Overview of mentions of France and Russia

Before looking at 'what is said' about France and Russia, it is useful to apply some of the methods used in the previous chapter to look at how often the two countries are mentioned, how that compares to mentions of other countries, and how this varies over time and across genres in the three newspapers.

7.3.1 Mentions overall (compared to other countries)

Table 7.1 shows the frequency of mentions of the 10 countries which occur highest on a Z2 list (see section 6.3.1.2); the frequency counts provided are overall frequency counts *whether tagged as Z2 or not*. In all three papers, France and Russia are among the 10 most mentioned countries (including UK country-names) and among the 5 most mentioned countries excluding UK constituent-country names (see Table 7.2). Nevertheless, the counts are very different from one paper to another; ERLN mentions both countries less often, and PMGZ mentions them more often, than the other papers in both absolute and relative terms.

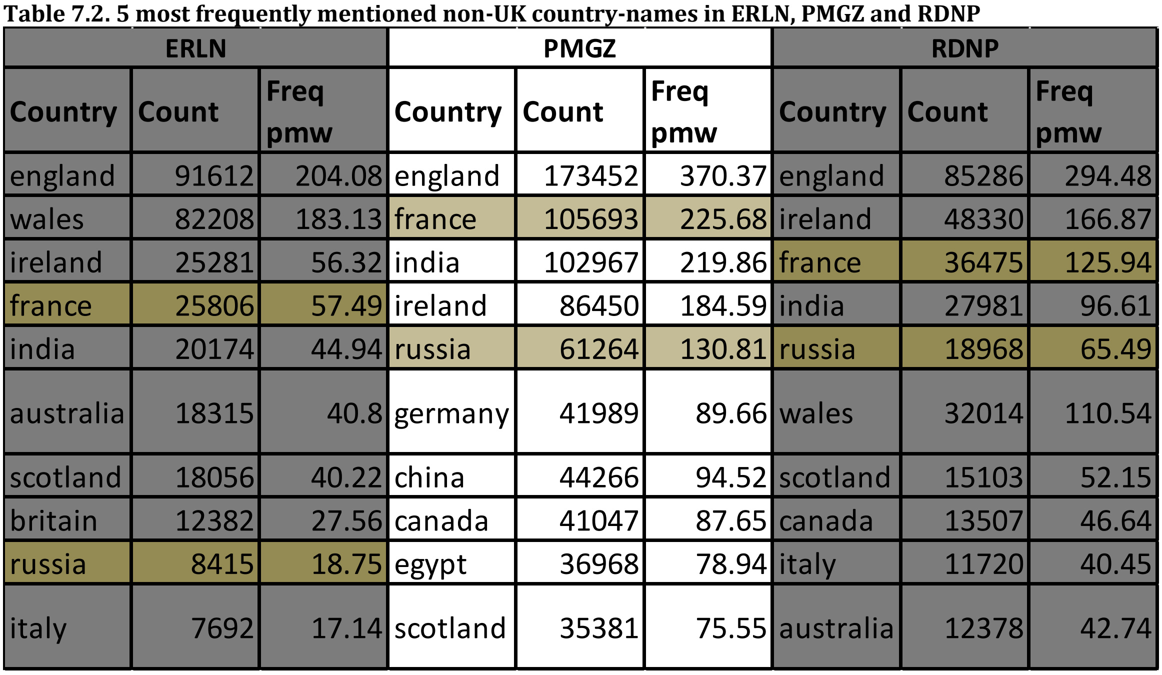
Table 7.1 10 most frequently mentioned country-names in ERLN, PMGZ and RDNP

ERLN			PMGZ			RDNP		
Country	Count	Freq pmw	Country	Count	Freq pmw	Country	Count	Freq pmw
england	91612	204.1	england	173452	370.4	england	85286	294.5
wales	82208	183.1	france	105693	225.7	ireland	48330	166.9
ireland	25281	56.32	india	102967	219.9	france	36475	125.9
france	25806	57.49	ireland	86450	184.6	india	27981	96.61
india	20174	44.94	russia	61264	130.8	russia	18968	65.49
australia	18315	40.8	germany	41989	89.66	wales	32014	110.5
scotland	18056	40.22	china	44266	94.52	scotland	15103	52.15
britain	12382	27.56	canada	41047	87.65	canada	13507	46.64
russia	8415	18.75	egypt	36968	78.94	italy	11720	40.45
italy	7692	17.14	scotland	35381	75.55	australia	12378	42.74

The newspapers do not, in fact, differ only in terms of how often they mention France and Russia, they also differ in terms of how much attention they devote to countries in general. ERLN seems to mention all countries less than the two other papers. This may reflect an overall difference in the orientation of the newspaper, with ERLN being more Britain-focused than the other two papers. It is easy to see how this might relate to the generic differences between the papers which have been discussed previously (see especially sections 6.2 and 6.4.3), with ERLN focusing on domestic sports and artistic events more than the other two papers. The difference between PMGZ and RDNP (with the former mentioning the countries in the top 10 more often than the latter) is also interesting, though harder to explain a priori.

Although the frequency counts are very different, all three newspapers mention France much more than Russia. This effect is strongest in ERLN, which mentions France three times more often than Russia, followed by RDNP then PMGZ (both mentioning France just under twice as often as Russia).

Table 7.2. 5 most frequently mentioned non-UK country-names in ERLN, PMGZ and RDNP



ERLN			PMGZ			RDNP		
Country	**Count**	**Freq pmw**	**Country**	**Count**	**Freq pmw**	**Country**	**Count**	**Freq pmw**
england	91612	204.08	england	173452	370.37	england	85286	294.48
wales	82208	183.13	france	105693	225.68	ireland	48330	166.87
ireland	25281	56.32	india	102967	219.86	france	36475	125.94
france	25806	57.49	ireland	86450	184.59	india	27981	96.61
india	20174	44.94	russia	61264	130.81	russia	18968	65.49
australia	18315	40.8	germany	41989	89.66	wales	32014	110.54
scotland	18056	40.22	china	44266	94.52	scotland	15103	52.15
britain	12382	27.56	canada	41047	87.65	canada	13507	46.64
russia	8415	18.75	egypt	36968	78.94	italy	11720	40.45
italy	7692	17.14	scotland	35381	75.55	australia	12378	42.74

There is hence a clear quantitative difference between the papers in terms of how often they mention countries in general, and France and Russia in particular. Is this quantitative difference associated with qualitative differences, or is it just 'more of the same' in the newspapers which mention both countries the most? This will be explored in section 7.4 and 7.5. Before that, the next subsection (7.3.2) looks at frequency of mentions over time.

7.3.2 Mentions over time

Figure 7.1 plots the mentions per million words of the two countries per year in each of the three newspapers. Figure 7.2 does likewise but aggregating the figures per decade. Looking at the distribution over time reveals a different picture from looking only at mentions overall (as in the previous section). Looking at mentions overall or the distribution over decades suggests that France is always mentioned more often than Russia in all three newspapers. But the distribution over time on a year-by-year basis shows that there are in fact three periods when the mentions of Russia overtake those of France: the period between 1852 and 1856 (most strongly for RDNP but also observable in ERLN – PMGZ was not yet published at the time), the period between 1876 and 1879 (most strongly for PMGZ but also very strong for RDNP and observable in ERLN), and 1885 (in PMGZ and RDNP only).

Most immediately apparent from looking at the graphs is also that there is a clear difference between the mentions of both countries over time (observable both per decade and, most dramatically, per year). Both countries have a pattern of 'peaks' and 'troughs' which is mostly present in all three newspapers, if not necessarily proportionally. It was hypothesized in section 7.2 that the distribution in mentions over time may be related to discussions of ongoing military conflict. It is interesting to consider the timing of the 'peaks' and whether they overlap (i.e. roughly coincide) in time with prominent military conflicts involving the countries. It would be beyond the scope of this chapter to summarize these military conflicts; the point is not so much to explore whether or not these peaks *are* in fact to related to these military conflicts, but to establish whether, at a cursory glance, they may *appear* to be. France presents the following highly prominent peaks:

- 1860: RDNP only, not observable in ERLN (PMGZ not yet published). Does not overlap with a major military conflict involving France.
- 1870: most prominent in RDNP, also very prominent in PMGZ, observable in ERLN. Overlaps with the Franco-Prussian War (see e.g. Howard 1988).

Russia presents the following highly prominent peaks:

- 1852-1858: RDNP and ERLN (PMGZ not yet published). Overlaps with the Crimean War (see e.g. Royle 1999).
- 1874-1880: most prominent in PMGZ, still very prominent in RDNP, observable in ERLN. Overlaps with the Russo-Turkish War (see e.g. Barry 2012).
- 1885: PMGZ and RDNP, not observable in ERLN. Overlaps with the Bulgarian crisis (see e.g. Lowe 1994: 63-66).

For Russia, in all cases, the peak is most prominent in PMGZ, followed by RDNP. The first two peaks are observable in ERLN, but not the last.

Figure 7.1. Mentions of France and Russia over time in ERLN, PMGZ and RDNP (per year)

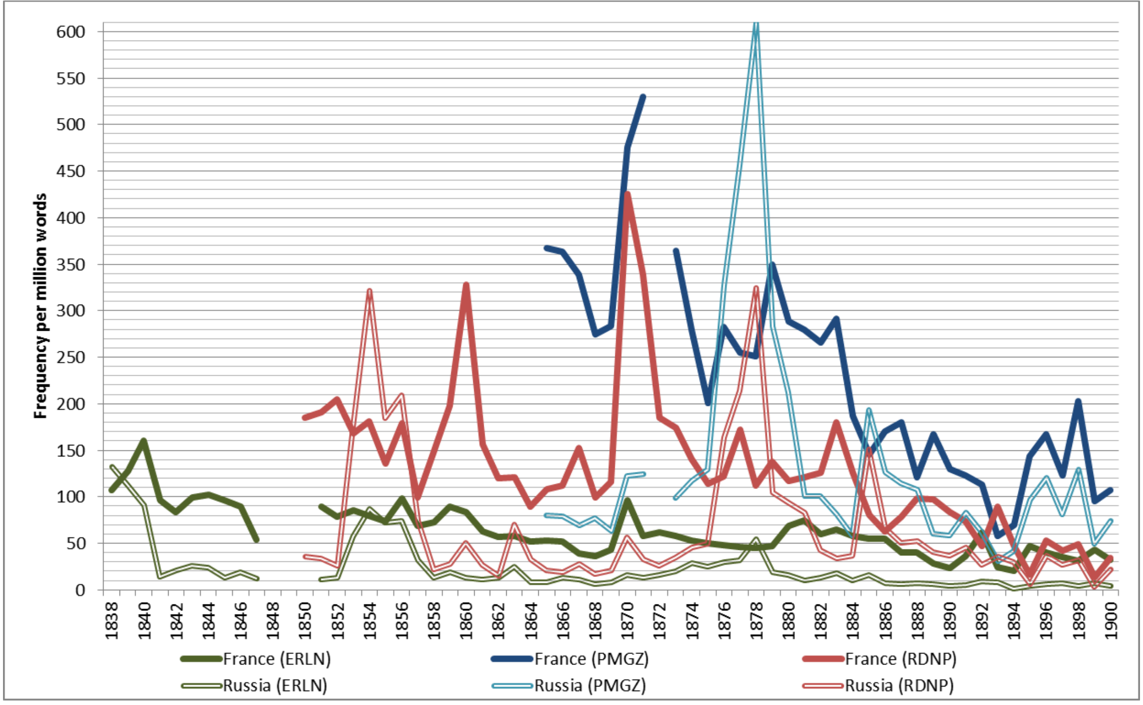
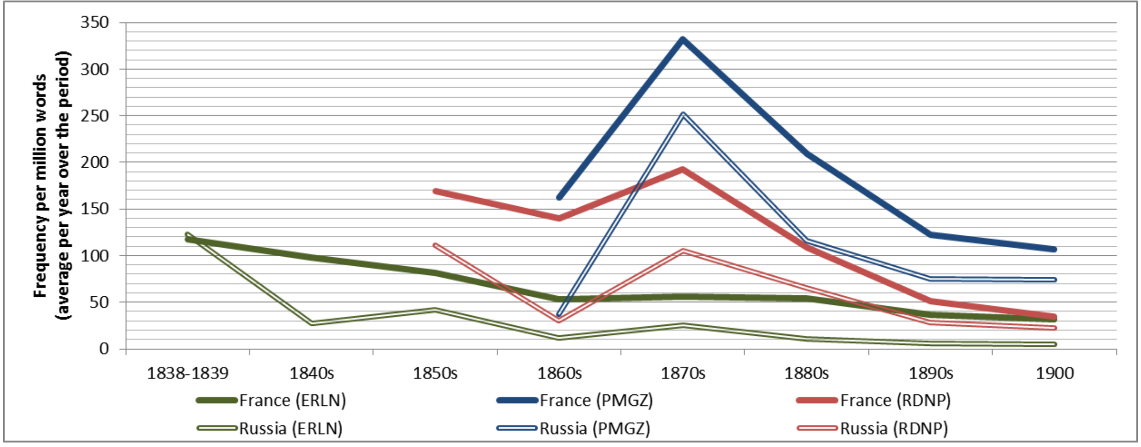


Figure 7.2. Mentions of France and Russia over time in ERLN, PMGZ and RDNP (per decade)



In all cases but one, the most prominent peaks did indeed overlap with an important military conflict involving that country. That the dates happen to coincide, however, does not provide evidence that this is indeed what is happening in the text. A quick verification of this involves looking at words in the semantic field of WAR. If the countries tended to co-occur with such words, and the proportion of mentions which co-occurred with such words were not smaller during peaks than during troughs, we might conclude that the discussion of the countries was indeed driven by discussions of military issues. As the data presented below suggests, however, this does not seem to be the case. Even if this *were* the case, fully confirming *h2* (as outlined in section 7.2) would also involve finding that the periods when the countries are *not* mentioned very often correspond to periods when the countries are *not* militarily engaged. This is harder (though possible in theory) to do, and I will not attempt it here.

Figures 7.3 to 7.5 show for each newspaper a comparison of occurrences per million words of all instances of 'France' or 'Russia', with the occurrences when mentioned within 10 words to the left or right of 'war' or words tagged with the USAS category 'G3', which stands for words in the semantic field of WAR (see also sections 3.2 and 6.3.1.2). Of course, these searches will not capture absolutely all discussions of the countries in relation to military issues. It is conceivable, though not highly likely, that in some cases discussions of military issues could occur without use of words in the semantic field of WAR. More likely are cases where the countries are being discussed in military terms, but without direct mention of ‘Russia’ or ‘France’. (The countries can be referred to in other ways, including demonyms, names of cities, or simply pronouns; some of these will be explored in section 7.3.4 below.) Nevertheless, this quantification should give us a reasonable idea of the extent to which mentions of the countries are associated with discussions of military issues.

Figure 7.3. Occurrences of France and Russia overall, with ‘war’ and with words tagged ‘G3’ in a span of 10LR, per decade, in ERLN (first France, then Russia)

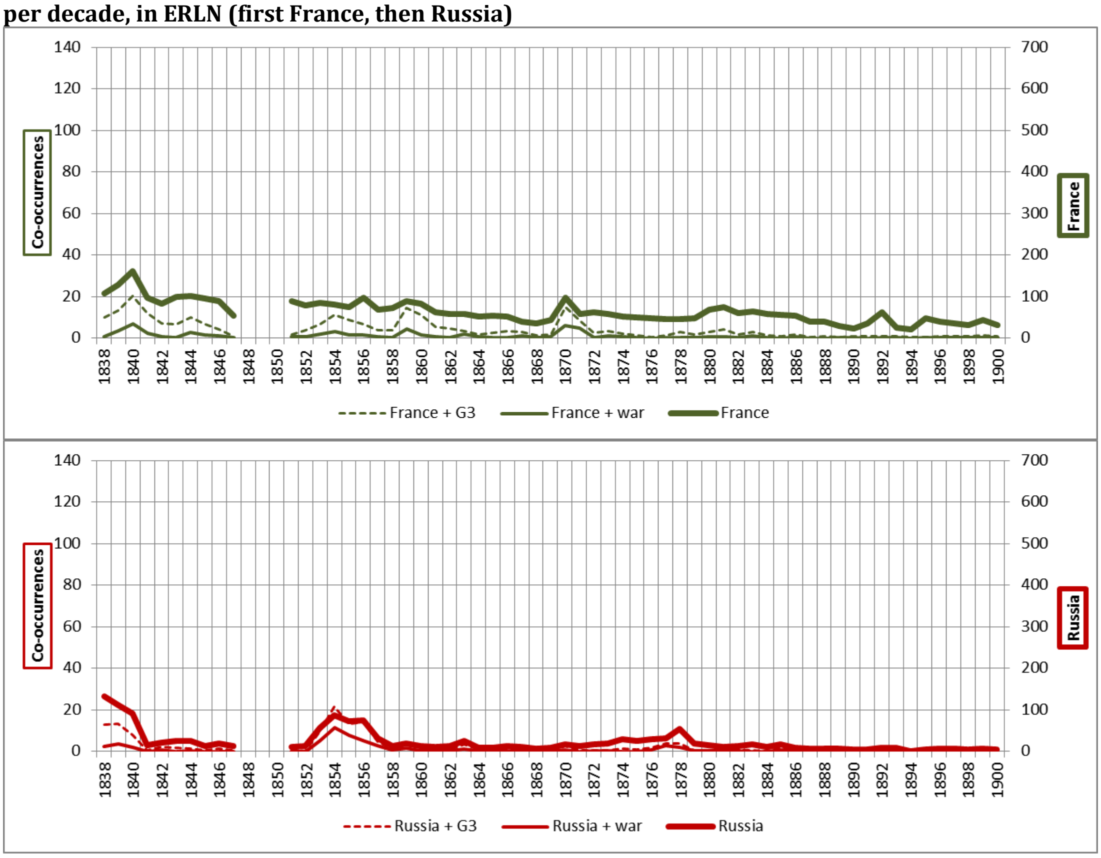


Figure 7.4. Occurrences of France and Russia overall, with 'war' and words tagged 'G3', in a span of 10LR, in PMGZ (first France, then Russia)

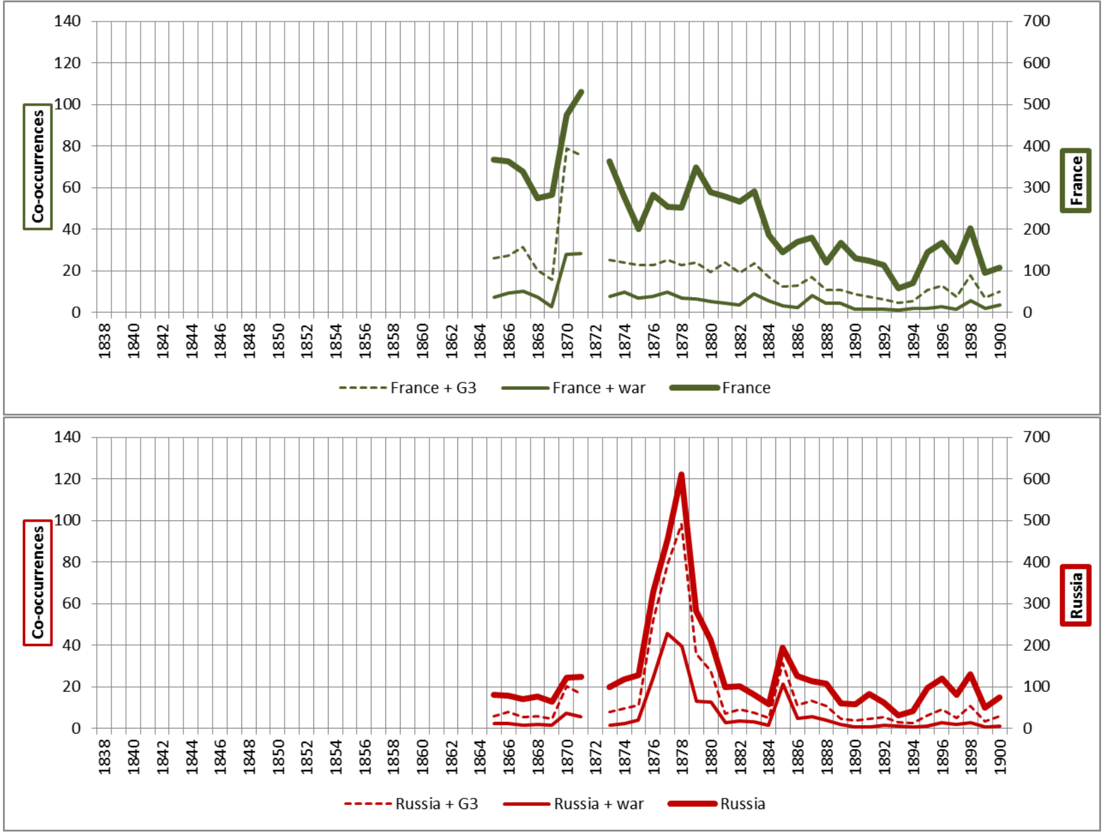
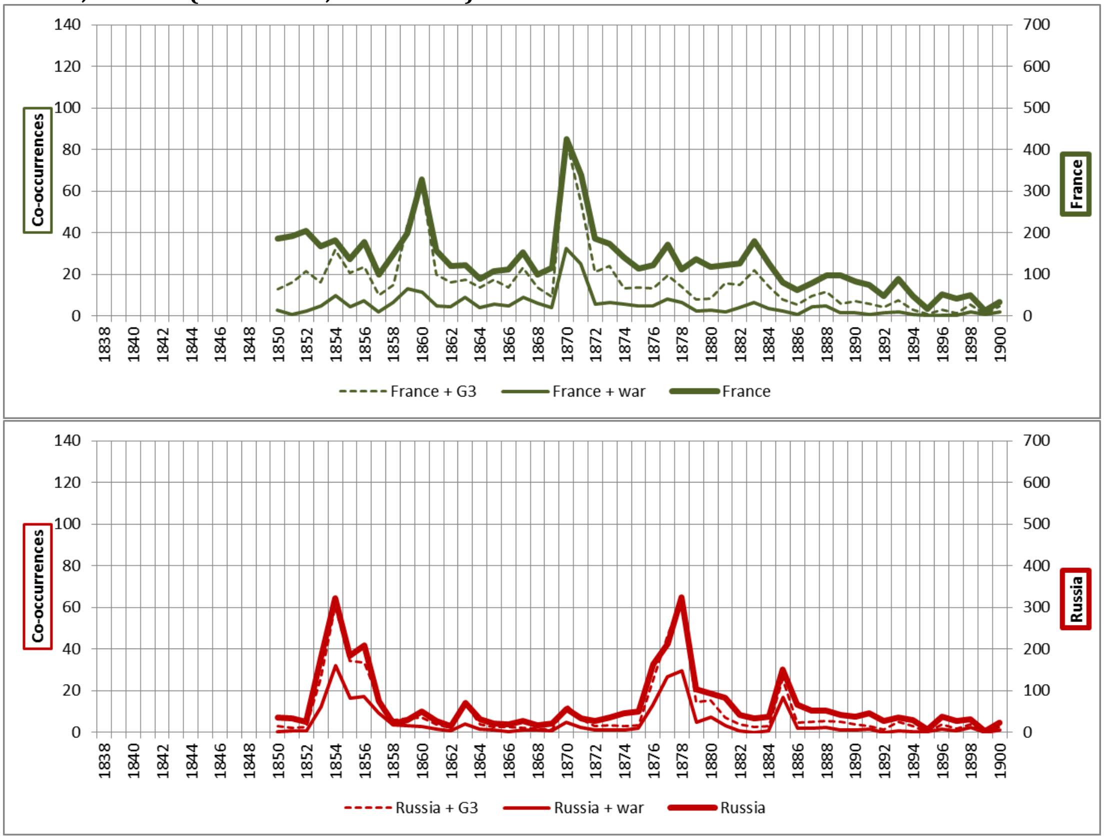


Figure 7.5. Occurrences of France and Russia overall, with 'war' and words tagged 'G3' in a span of 10LR, per decade, in RDNP (first France, then Russia)



What Figure 7.3, Figure 7.4 and Figure 7.5 show is that although the countries definitely occur as part of military discussions throughout the period, the proportion of mentions of the countries which these discussions represent vary over time but are never overwhelming. For France, co-occurrences with the word 'war' accounts for at most 8.16% of occurrences in ERLN, 5.84% in PMGZ and 7.67% in RDNP. For Russia, the maxima are respectively 13.16%, 10.93% and 16.95%. Extending this to co-occurrence with the semantic field of WAR (i.e. any word tagged 'G3'), these maxima become 16.17%, 16.55% and 21.87% for France, and 24.45%, 17.34% and 27.12% for Russia. This tells us than in no year in any newspaper is either country ever mentioned near to a word related to WAR in more than 27% of cases. The converse of this observation is that *in every year in which the countries are mentioned, in all three newspapers, there are always at the very least 72% of mentions of the country which occur without any word in the semantic field of WAR occurring within 10 words to the left or right of the country*.

Of course, OCR errors might mean that some WAR-related words are not included in these counts. But OCR errors would have to disproportionately affect WAR-related words to alter the finding that the countries overwhelmingly occur *away* from words in the semantic field of WAR. To help clarify this, we can think in terms of averages: the average (as opposed to maximum) proportion, per year, of occurrences of 'Russia' within 10 words to the left or right of any word semantically related to WAR is 6.87% in ERLN, 9.68% in PMGZ, and 12.40% in RDNP. For 'France', the averages are 5.71%, 8.30% and 11.30%. These are non-negligible proportions, but they are not overwhelming either. It is clear that something *else* is being discussed around mentions of the countries – more likely, several other things.

It is interesting to observe that in all three newspapers, the proportion of mentions nearby to a WAR-related word is higher for 'Russia' than for 'France'. This fits in with our expectation that since Russia is more geographically distant from France, the ways of representing Russia may be narrower (*h5* and *h6*). There are also observable differences between the newspapers. The highest average over the whole period for both countries occurs in PMGZ, followed by RDNP and lastly ERLN. This is interesting, since it matches our finding that PMGZ mentions countries most, followed by RDNP and then ERLN. It suggests that the increased interest in countries in PMGZ and RDNP over ERLN may be related to PMGZ and RDNP's greater interest in political/diplomatic issues (including warfare). So this difference may again relate to differences in the generic make-up of the newspapers. An obvious next step is to look at the distribution of the mentions of countries across genres; I do this in the next section.

7.3.3 MENTIONS ACROSS GENRES

Are France and Russia mentioned only in news articles or also in other article genres? Figure 7.6 and Figure 7.7 show the proportion of raw and relative counts for each country distributed among the article genres identified by Gale/Cengage (see also section 6.2) in each of the three newspapers. It is immediately obvious from Figure 7.6(top) that there is a difference between the newspapers in terms of the distribution across article genres, with ERLN contrasting with PMGZ and RDNP. PMGZ and RDNP have at least 80% of mentions of both countries occurring in news articles, whereas in ERLN, the proportion of mentions of France in news articles is only 34%, and of Russia only 52%. So although in all cases, mentions in news represent the greatest proportion of mentions across the categories, the proportion in news is very different in ERLN versus the two other newspapers. As this difference dwarfs all other differences, Figure 7.6(bottom) shows the same data, but omitting news articles. It is possible that news seems to attract most mentions because news may be less vulnerable to OCR errors than other newspaper genres. This is a possibility which I cannot test here, but the effect is considerable enough that the skew in OCR errors would have to be very great to nullify this finding.

With mentions in news set aside, there are still obvious differences between the newspapers. Although in all cases, the second most prominent article category is advertisements, and the third (other than for France in PMGZ) is arts, the fourth category is crime in RDNP but commerce in PMGZ, whereas neither crime nor commerce attract more than a few percent of mentions of the countries in ERLN. By and large, then, the greatest differences revealed in Figure 7.6 are those related to generic differences between the newspapers. This brings up the question – are these differences merely proportional to the distribution of content across article genres in these newspapers, or are there further differences beyond this?

Figure 7.7 provide an answer to this question. Figure 7.7(top) shows how many mentions per million words *in that article category of that newspaper* each country attracts; Figure 7.7(bottom) shows us the proportion of mentions, per million words in that article category of that newspaper, which is accounted for by mentions in a particular article category. Figure 7.7(top) is hence intuitively easier to understand, but the differences in raw mentions between the newspapers is such that comparisons across newspapers become difficult: Figure 7.7(bottom) facilitates such comparisons. What we see in these graphs is that *once the difference in generic make-up of the newspapers is taken into account* there is in fact not a very big difference in the distribution of mentions of countries across article genres within each newspaper. In other words, the differences observed in Figure 7.7 are mostly attributable to overall differences in the generic make-up of the newspaper and cannot be ascribed to differences in how the newspaper represent the countries.

Nevertheless, some differences among newspapers remain. Taking into account the relative amount of content per article category, we find that news remains, in all cases, the category in which the countries are mentioned most (relatively speaking). The ranking of other categories varies from newspaper to newspaper. In ERLN, the next categories are arts, then 'birth, death, etc.'. In PMGZ, the next categories are crime, commerce and arts in more-or-less equal proportions. In RDNP, the next categories are 'birth, death, etc.' followed (with some lag) by arts.

We thus observe that differences between the newspapers are greater than differences between the representations of the two countries within any given newspaper: when we look at distribution of mentions of the countries across article genres, France and Russia have more in common within a single newspaper than does either country across newspapers.

Figure 7.7 (bottom) also shows that there are systematic differences in the distribution of mentions across genres between the two countries. Although this difference is small compared to the differences between newspapers identified above, nevertheless, it is observable that, in all newspapers, Russia attracts a greater proportion of mentions in news articles than France does, and that France attracts a greater proportion of mentions in commerce articles than Russia does. This is in line with *h5* and *h6*.

Figure 7.6. Proportion of raw mentions of France and Russia in different article genres in ERLN, PMGZ and RDNP (first with, then without news articles)

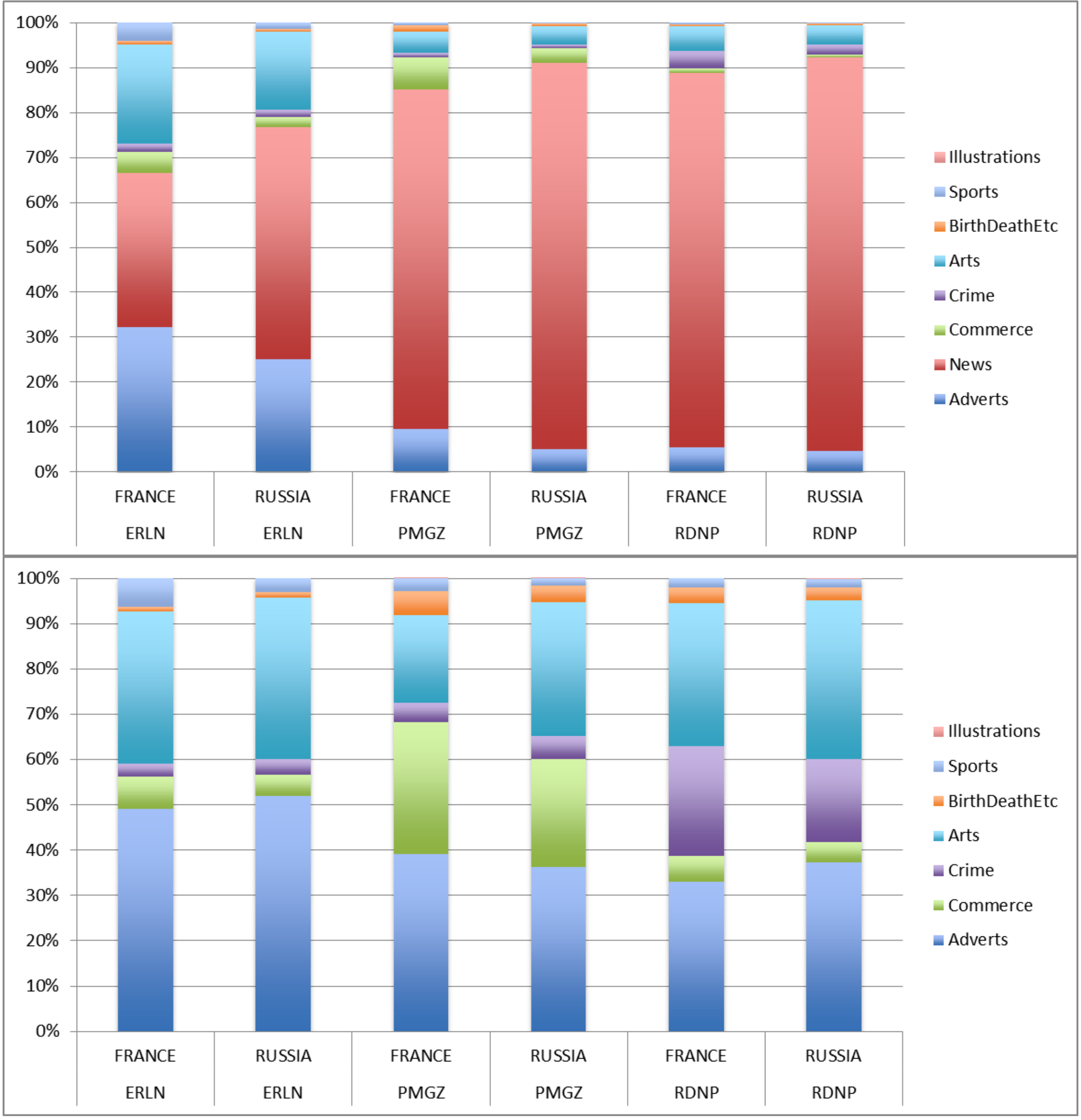
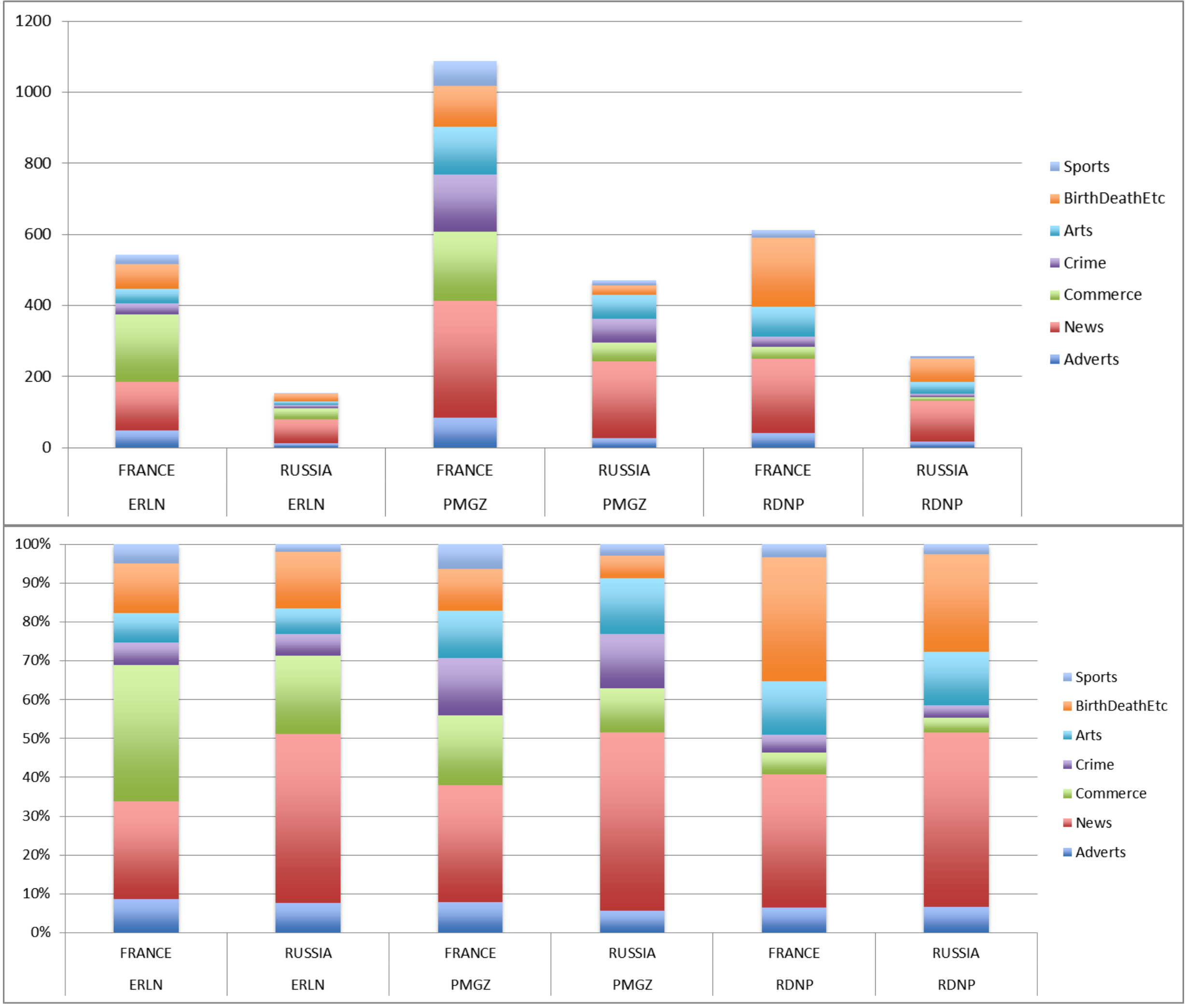


Figure 7.7. Proportion of mentions of France and Russia per million words in each article genre in ERLN, PMGZ and RDNP (first in counts per million words, then in percentages)



7.3.4 Ways of referring to France and Russia

The next two sections will explore approaches to investigating discourses surrounding 'France' and 'Russia' in more detail. First, however, it is important to address an important limitation of the discussion which follows, namely that I shall be exploring *only* the words 'France' and 'Russia'. I limit myself to these two words because the purpose of this chapter is to explore and illustrate approaches to answering the question 'what is said about this place?' in a large set of texts, and exploring other words would unduly lengthen the discussion.

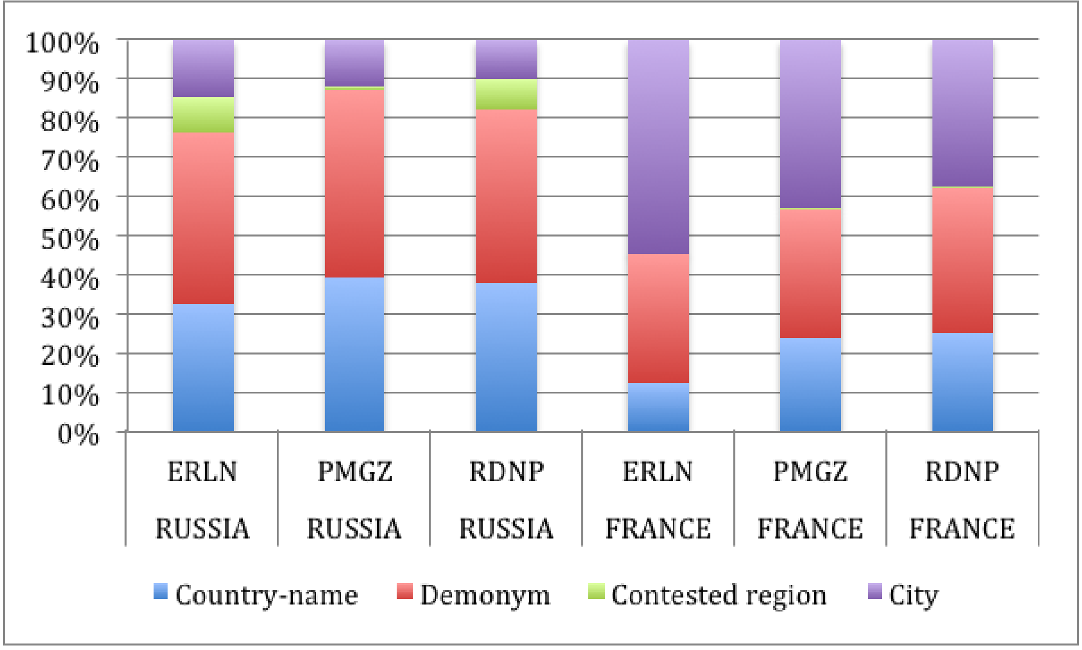
Nevertheless, it is obvious that France and Russia, the countries/nations, can be referred to in different ways, not just by those two words. Alternative referring expressions include pronouns ('she' or 'it'), geographical terms at different scales ('Crimea' or 'Paris'), and ways of referring to people from those places, whether neutral ('Russian', 'French') or derogatory ('Frogs'). A complete investigation of the discourses surrounding those countries would want to explore such terms, perhaps uncovering different discourses associated with different terms.

In lieu of such a broadly-focused analysis, I will look briefly at the distributions of mentions of such terms, in order to get an impression of the extent to which focusing narrowly on the terms 'Russia' and 'France' might miss relevant content. Figure 7.8 shows the distribution, per country and per newspaper, of mentions across the country-name, demonym(s), names of cities within the country, and names of contested regions. It is clear from this graph that there is a great difference between Russia and France, and that this difference cuts across the different newspapers. For Russia, mentions of the country-name 'Russia' account for between 30% and 40% of mentions of all the Russia-related terms considered. By contrast, 'France' attracts only between 12% and 25% of mentions of all the France-related terms considered. For Russia, the demonyms 'Russian' and 'Russians' are the most important category, attracting around 45% of mentions. But the most important category for France is cities ('Paris', 'Marseilles', etc.), which attracts between 39% and 53%, followed by demonyms, which attract between 30% and 35% of mentions.

What this tells us is that there are important differences in how the countries are represented, with more mentions of places within France than places within Russia, and more mentions of Russia in global ways (using the country-name or demonym); these observations are in line with *h4*.

Moreover, it is also clear that the discussion in the next two sections, which focuses on the country-name, may better capture how Russia is represented overall than France, since Russia attracts more mentions with its country-name than does France.

Figure 7.8. Ways of referring to France and Russia in ERLN, PMGZ and RDNP



I used two methods to discover which cities within the country were mentioned. First, I looked at the list of Z2 places (see section 6.3.1.2). This allowed me to find the most mentioned places, but the problem with this approach is that some of the French and Russian cities mentioned are not mentioned very frequently, so they only appear very low down the Z2 list. Using the Z2 list, I found Moscow and St Petersburg for Russia, and Paris and Calais for France. To supplement these, I looked in each newspaper for words tagged Z2 which occurred within 5 words to the left and right of the country-names. I then read through 200 random concordance lines in each newspaper to identify other cities mentioned. This method helped me identify the most mentioned cities; cities with very few mentions may still have been missed out. The following additional place-names were added to my list: for Russia, variant spellings of St Petersburg, Crimea and Circassia (two contested regions); and for France, Alsace (a contested region), Lyon, Versailles, StEtienne, Cannes, Bordeaux, Toulon, Rouen, Nantes, Nice, Le Havre, Dieppe, Perpignan, and St Valery. I then searched for the frequency of mentions of each of these place-names (except St Valery and St Etienne which, being multi-word place-names, do not occur in a frequency list).

For demonyms, I simply used ‘French’ and ‘Frenchs’ (which is rare) and ‘Russian’ and ‘Russians’ (which is common). Pejoratives could be added in theory (e.g. ‘Froggy’ for the French, or ‘Hun’ for the Russians[1]) but my preliminary searches identified few instances of these, which moreover did not systematically refer to the French or Russians, so I decided to leave them out.

7.3.5 Summary of findings so far

So far, I have found a number of differences in the coverage of the two countries, relative to one another and across newspapers, time, and genres. Here is a summary of my findings so far, and how they relate to the hypotheses set out in section 7.2.

- All three newspapers mention France and Russia; they are among the most mentioned countries in all three newspapers. This suggests that France and Russia are indeed relevant to the British public. (Section 7.3.1)
- In all three newspapers, France is mentioned much more than Russia overall. This meets *h5*. (Section 7.3.1)
- However, in all three newspapers, there exists specific periods in time when Russia is mentioned more often than France. This is in line with *h2*, though it is not sufficient to accept the hypothesis without reservation. (Section 7.3.2)
- The newspapers differ in terms of how often (in both relative and absolute terms) they mention France and Russia; this difference is in line with how often they mention other

[1] The term 'Hun' as a pejorative reference to the Russians in the nineteenth century was suggested to me by Michael Hughes.

countries. This suggests that differences in general frequency of mentions of France and Russia may be a reflection of the overall orientation of the paper and its generic make-up, rather than a difference in representation of the countries. (Section 7.3.1)

- There is a difference over time in terms of how often the countries are mentioned. Both countries have patterns of 'peaks' and 'troughs', which are roughly similar across newspapers but are different for each country. This is in line with *h2* though it is not sufficient to accept the hypothesis without reservation. (Section 7.3.2)
- The newspapers differ in terms of the extent to which particular 'peaks' are evidenced in the data. For Russia, the most prominent peaks are most prominent in PMGZ, followed by RDNP. For France, the most prominent peaks are most prominent in RDNP, followed by PMGZ. For both countries, the peaks are much less prominent in ERLN. This is surprising relative to the hypotheses set out in section 7.2. We might surmise that it is related to differences in the overall orientation of the newspapers (including the generic make-up), but also that it might tell us something about the level of interest in particular events in different sections of the population (since the newspapers circulated to different audiences, see section 6.2). (Section 7.3.2)
- Both countries occur nearby to the word 'war' and words semantically related to WAR (as identified using the USAS tag 'G3') throughout the period and in all three newspapers. This is in line with *h1* and *h2*. What is *not* in line with these expectations is that the proportions of instances of the countries which co-occur with WAR-related words are small: within a span of 10 words to the left and right, neither country occurs more than 27% of the time in proximity with one such word in any given year in any of the newspapers. The averages per year, per country, per newspaper are in fact in the range 5% to 13%. This suggests that, counter to *h1* and *h2*, the countries are not overwhelmingly mentioned in military discussions, or that if they are, these discussions do not make extensive use of military terms. (Section 7.3.2)

- A greater proportion of instances of Russia than France occur around military terms. This difference is small but observable. This is in line with *h5* and *h6*. (Section 7.3.2)
- Looking at distribution across genres, both countries are mentioned most in news articles, but are also mentioned in all the other categories considered. This is in line with *h3* (which predicts that most mentions will be in news articles), but it also suggests that the framework set out in *h3* is limiting, since the countries are in fact represented in a wide diversity of contexts, not just news articles. (Section 7.3.3)
- In terms of distribution across genres, both countries have more in common *within* a newspaper than *across* newspapers. Here, differences between newspapers appear to be more important than differences between the representations of the two countries. This might be taken to indicate that the countries have similar representations overall, and/or that differences between newspapers' representations of the countries may be important. In any case, it is clear that it is important to take into account the generic make-up of newspapers, and that comparisons of discourses surrounding places (or, more broadly, of *any* discourses) across newspapers which do not take these generic make-ups into account will miss an important explanatory factor. (Section 7.3.3)
- A greater proportion of mentions of Russia occur in news articles than is the case for France. In turn, a greater proportion of mentions of France occur in commerce articles than is the case for Russia. This is in line with *h5* and *h6*. (Section 7.3.3)
- Many French cities get mentioned in these three newspapers, but only two Russian cities (Moscow and St Petersburg) are mentioned often enough to have been picked up by my search methods. This is an interesting finding in itself which is in line with *h5* and *h6*. (Section 7.3.4)
- Within the range of terms considered, 'Russia' attracts between 30% and 40% of mentions per newspaper but 'France' attracts only between 12% and 25% of mentions per newspaper. This, like the previous finding, is in line with *h5* and *h6*. Moreover, it suggests that the additional analyses of 'France' and 'Russia' in the sections that follow

are likely to capture more of the discourses surrounding Russia than France. (Section 7.3.4)

7.4 The global approach

7.4.1 Introducing the approaches

This section and section 7.5 illustrate two further, more comprehensive, approaches to answering the question ‘what is said about this place in these texts?’. Since this question focuses on a pre-selected linguistic expression (the place-name), only the expression-intensive and the tracking-expressions approaches are relevant (out of the approaches outlined in section 2.3.3)[2]. Since I am interested in comparing corpora, the most relevant approach is the tracking-expression approach. And since I am working with broad rather than thematic corpora[3], the approach that I exemplified in section 2.3.3.5 via a review of Caldas-Coulthard and Moon (2010) – which involves comparing patterns of frequency and collocation of the expressions of interest in the various corpora – is most relevant. It would however also be interesting to construct thematic corpora in order to attempt the other possibility; this might be explored in future research.

Hence, both approaches explored here will involve comparing patterns of collocation surrounding the terms 'France' and 'Russia'. As discussed in section 2.3.2.3, there are many different ways of operationalizing collocation. It is beyond the scope of this thesis to investigate them all. The two approaches chosen here will nevertheless address the broad distinction made in section 2.3.2.3: the first, the 'global approach', will be a 'collocation-via-significance' approach to collocation, whereas the other, the 'sampling approach', will be a 'collocation-via-concordance' approach. The two methods are further compared in section 7.6. For now, it is enough to point out that whilst the first (the ‘global’ approach) will start from the whole corpus and then zoom in, the second (the ‘sampling’ approach) will start by looking at a small sample

[2] The selection-via-concordance approach would also be relevant, as a means of broadening the set of linguistic expressions to investigate, but pursing that approach is beyond the scope of this analysis.

[3] I.e. with corpora which are not assembled on the basis of the topics treated in the texts; see also section 2.3.3.5.

and then zoom out. Throughout, I will use the notation ‘COUNTRY’ in capitals to refer to whichever of France or Russia is under study at any given moment.

The first approach consists of the following steps:

- search for the word of interest (the ‘node’, i.e. COUNTRY) in the whole corpus (e.g. search for ‘Russia’ in all of PMGZ)
- bring up a list of collocates for the node in the whole corpus
- categorize the collocates according to the context evident in most of the concordance lines where COUNTRY and the collocate co-occur (taking a random subset if there are many co-occurrences). The question here is: ‘what is each collocate telling us about how COUNTRY is referred to?’.
- compare the results of the preceding step between countries, newspapers, etc.

Generating a list of collocates requires choosing three parameters: the span, frequency thresholds, and a statistic. Here, I use a span of 3 words left and right of the node. This decision is to some extent arbitrary (some span or other needs to be chosen); section 4.5 found that collocation statistics would be more reliable with smaller spans in OCR data, so I err on the side of smaller rather than larger.

For frequency thresholds, I chose a minimum frequency of 10 for the co-occurrences. Section 4.4.1 showed that using a threshold would be useful, as it excludes many OCR errors (since many erroneous word-forms occur only once).

For the statistic, I chose Log Ratio (LR), which is equivalent to using LL as a cut-off point and MI as the ranking statistic. This is recommended by Hardie (forthcoming); moreover, section 4.5 showed that using MI in combination with LL would be more reliable than using either LL or MI on its own with OCR data.

By contrast, the second approach, which will be used in section 7.5, consists of the following steps:

- search for the node (i.e. COUNTRY) in the whole corpus
- take a random sample of concordance lines
- create categories describing the phraseologies involving the place, as instantiated in that sample of concordance lines
- take a new random sample of concordance lines
- describe the new sample using the categories obtained by analysing the first sample; update the set of categories as required
- continue analysing more random samples until the system of categories has stabilized
- continue analysing more random samples until there are enough examples to support all findings
- search for 'interesting' categories in the whole corpus (where interesting categories are those associated with specific findings) to further explore/refine findings.

7.4.2 RESULTS

Having generated lists of collocates for France and Russia in each newspaper, the first step is to decide which collocates to investigate further. Table 7.3 shows the number of collocates which co-occur at least 10 times with each country in each newspaper (as well as being above the LL cut-off point). From these collocates, words containing OCR errors and tokenization errors are excluded (such as 'TURKEY.', 'Ger-' and 'wvar').[4]

Of the remaining collocates it makes sense to focus on those which have a 'strong enough' effect. What counts as 'strong enough' is to some extent arbitrary, but it makes sense to decide on a number because it gives a principled reason to investigate *all* the collocates which are above that point. A number which is often used is MI=3 (see section 4.5). Since MI and LR will in practice generate similar numbers, it makes sense to adopt LR=3 as a cut-off point. Table 7.3 shows how many of the remaining collocates are above that point. If the main aim of the

[4] I also excluded *all* word forms with hyphens as most word-types containing hyphens proved to result from OCR errors. This, however, also excluded genuine words, such as *jew-baiting*. The number of excluded but correct hyphenated words is provided in **Table** 7.3.

thesis was to investigate discourses surrounding France and Russia, at this point I would investigate *all* of these collocates. However, here, my aim is to test the approach, so it is appropriate to limit the number of collocates to investigate. I have hence used the relatively high threshold of Log Ratio=6. This leaves me with between 12 and 54 collocates to investigate per country and newspaper (see Table 7.3).

Table 7.3. Number of collocates of France and Russia meeting various conditions in ERLN, PMGZ and RDNP

	FRANCE			RUSSIA		
	ERLN	PMGZ	RDNP	ERLN	PMGZ	RDNP
Number of collocates co-occurring at least 10 times with COUNTRY	209	341	201	147	837	276
Number of collocates remaining after errors are excluded (hyphenated words)	189 (4)	324 (14)	192 (2)	145	812 (12)	267
Of the remaining collocates, number with a Log Ratio above 3	189	267	192	91	294	137
Of the remaining collocates, number with a Log Ratio above 6	54	20	12	33	20	25

In the list of collocates from the overall data using Log Ratio, what is prominent are either unusual words which are commonsensically associated with the country (people's names, foreign words, names of places), errors (including hyphenated words), or words from ads that were published repeatedly. This is unfortunate, because these things may not capture the patterns most of interest to a researcher. One solution to this is to exclude ads; see section 7.3.3. In the meantime, the approach still allows us to find out some things of interest about the representation of the countries. To illustrate, I discuss below the collocates most strongly associated with the two countries in these three newspapers overall.

Once the collocates to explore further have been identified, the next step is to look at the concordance lines in which the countries co-occur with these collocates, in order to figure out what we can learn about how the country is represented from the fact that that country co-occurs with that collocate. This amounts to grouping the collocates together in terms of what they tell us about how the country is represented in these newspapers.

Perhaps the pattern to stand out most strongly, even before exploring the collocates any further, is that both France and Russia have a strong tendency to co-occur with mentions of other countries and cities. Table 7.4 shows the countries and cities (from the top collocates considered here) which France and Russia co-occur with in each newspaper.

Table 7.4. Countries and cities co-occurring with France and Russia at least 10 times in ERLN, PMGZ and RDNP, and with a Log Ratio above 6

		ERLN	PMGZ	RDNP
FRANCE	Countries	Belgium, Germany, Portugal, Italy, Sardinia, Austria, Prussia, Switzerland, Spain, Algeria,	Germany, Belgium	Denmark, Germany, Tonquin, Siam, Belgium, Sardinia, Madagascar
	Cities	Roubaix, Grenoble, Nimes, Lille, Toulouse, Toulon, (Le) Havre, Rheims, Nantes, Cannes, Amiens, Bordeaux, Marseilles, (St) Etienne, Rouen	Nantes, Dinard	-
RUSSIA	Countries	Turkey, Prussia, Austria, Italy, Sweden, Germany, Poland	Bessarabia, Austria, Finland, Turkey	Portugal, Turkey, Austria, Bulgaria, Germany, Prussia
	Cities	Riga, (St) Petersburg, Moscow, Warsaw	-	(St) Petersburg

It is clear from looking at the table that there are important differences in this respect between newspapers:

- Cities are especially strong collocates in ERLN as opposed to the two other newspapers for both countries.

- Although countries are strong collocates of both countries in all three newspapers, there are more of them in ERLN and RDNP than in PMGZ.

The table also reveals that, curiously, it is not the same countries which are strong collocates of France and Russia in all three newspapers. For France, only Belgium and Germany are strong collocates in all three papers, with Sardinia being the only strong country collocate in two newspapers (ERLN and RDNP). For Russia, only Turkey and Austria are strong collocates in all three newspapers, with Prussia and Germany being strong collocates in two newspapers (ERLN and RDNP). In a moment, I will explore what these collocations tell us about the representations of France and Russia, but before I do so, I want to point out a problem with this kind of comparison between collocates across newspapers.

The strength of collocation quoted here tells us not only about the relationship of two given words but also about the rest of the corpus in which these words co-occur: that is, the statistics tell us that two words are unusually associated *relative to the baseline in that newspaper*. However, the newspapers are very different, so finding that one word collocates strongly with another in one newspaper but not another might tell us more about the newspaper than about the node. A hypothetical example might help. Let us say Russia is well-known for its ballerinas. Thus all three newspapers mention *Russia's ballerinas*. But ERLN is very interested in the performing arts and hence mentions ballerinas a lot in other contexts, whereas the other two newspapers do not. In the two other newspapers, 'ballerinas' will crop up as strongly associated with Russia, but in ERLN it will not. It would be wrong, however, to conclude that 'ballerinas' is only associated with Russia in the two other newspapers. What we find is that this technique is telling us not just about COUNTRY but also about the newspapers. And since we already know that there are systematic differences (in particular in generic make-up) between the newspapers, it is difficult to compare these results across newspapers.

One way to perform this comparison would be to regenerate the collocates, but using the three newspapers added together as the comparison baseline, rather than the individual newspapers. The collocates which emerged would then be those unusually associated with that COUNTRY given the discussions happening in all three newspapers. This would be a good solution but the current infrastructure of CQPweb does not allow for it[5]; it could be done outside of CQPweb but that exercise is beyond the scope of the present analysis. Another approach would simply be to use a different procedure, keyness analysis, which is also foundational in corpus linguistics. There is much to say about keyness analysis, but I have limited myself to collocation in this thesis, for reasons of space. A last approach would be to try and control for

[5] The analysis in this chapter and the previous one were done using CQPweb, see also section 6.2.

the differences between the newspapers by comparing only more comparable portions of text. I will adopt this final approach in order to explore the extent to which genre might impact on these overall results. In the next section, then, I look at the most strongly associated collocates of COUNTRY in each newspaper for the news section only.

Putting this issue aside, I will now discuss what the collocates tell us about how the countries are represented in the newspapers, providing illustrative examples from the newspapers as I go. Returning to the countries and cities which co-occur with COUNTRY, let us ask, in what kinds of contexts do they co-occur? First, let us consider the countries, looking at each newspaper in turn. For a summary of the patterns discussed below, see also Table 7.5.

Table 7.5. Predominant contexts in which strong country-collocates of France and Russia co-occur with France/Russia

	FRANCE			RUSSIA		
	ERLN	PMGZ	RDNP	ERLN	PMGZ	RDNP
Military/diplomatic discussions	Sardinia, Austria, Russia		Tonquin, Siam, Sardinia, Madagascar	Persia, France	Austria, Turkey	Turkey, Austria, Bulgaria, Prussia
Diplomatic discussions (no military overtones)	Algeria					
Adverts	Prussia, Austria, Italy, Sweden, Denmark		Denmark	Belgium, Portugal, Italy, Switzerland		Portugal
Various contexts	Germany, Prussia, Spain	Germany, Belgium	Germany,	Turkey, Germany, Poland	Finland	Germany

Both France and Russia strongly co-occur with 11 countries in ERLN. In both cases, the most frequent context of co-occurrence is an advertisement addressed to artists looking for employment, which lists countries in which the advertiser has correspondents:

> "Messrs Parravicini and Corbyn have Special Correspondents in France, Belgium, (…), Prussia, Russia, Italy (…) Artistes applying for Engagements must state (…)" (ERLN 28/06/1874)

For Russia, this accounts for 6 (Prussia, Austria, Italy, Sweden and Denmark) out of the 11 strong country-collocates in ERLN. For France, this accounts for 4 (Belgium, Portugal, Italy, Switzerland) out of the 11 strong country-collocates in ERLN. In ERLN, 2 further countries (Persia and France) co-occur with Russia predominantly in the context of diplomatic or military discussions:

> "One thing is clear; that Russia calculated upon Persia as the means by which she would march her armies towards Hindoostan." (ERLN 09/12/1838)
>
> "Our relations with the great empires of France and Russia, as well as on the Continent of Europe, unless we may except that of Spain, happily continue to be of the most friendly character " (ERLN 15/01/1860)

The last 3 remaining strong country-collocates of Russia in ERLN (Turkey, Germany and Poland) occur in various contexts, including diplomatic/military discussions and arts advertisements, such that it is not possible to summarize the patterns of co-occurrences with a single label.

For France in ERLN, 3 out of the strong country-collocates (Sardinia, Austria and Russia) co-occur predominantly in the context of diplomatic or military discussions:

> "Under these circumstances, he could not sympathise with France, and Sardinia in a war which he believed to be unnecessary" (ERLN 14/08/1859)
>
> "Even the alleged coldness between France and Austria has ceased to form a topic of conversation" (ERLN 27/06/1858)
>
> "This understanding between France and Russia extends also to a solution of the Schleswig question" (ERLN 13/07/1862)

One strong country-collocate of France in ERLN (Algeria) also occurs in diplomatic discussions, but which have no (potential or effective) military overtone:

> "The new post office convention between England and France came into operation on 1st of June by which the British rate on all letters to France and Algeria (…) " (ERLN 04/06/1843)

Finally, the three remaining strong country-collocates of France in ERLN (Germany, Prussia and Spain) occur in various contexts, including diplomatic/military discussions and arts advertisements, such that it is not possible to summarize the patterns of co-occurrences with a single label.

The examples above illustrate that diplomatic/military discussions where Russia and France co-occur with other countries may be discussions of ongoing, past, or potential military conflict; they may be discussions of military conflicts or simply diplomatic relationships more broadly (including commentary on the absence of ongoing conflicts); and these conflicts and relationships may be between COUNTRY and the co-occurring country, or may be between COUNTRY and Britain (with the co-occurring country being a third party on one side or the other). That such discussions occur could be considered in line with *h1* (see section 7.2), although their diversity shows that *h1* is reductive and potentially misleading. Further, that such discussions do not constitute an overwhelming proportion of the discussions mentioning these countries is not in line with *h1* and suggests that this expectation is, in the final analysis, misguided. Other contexts in which France and Russia are mentioned will be discussed further in this section.

Without going into as much detail for PMGZ and RDNP as for ERLN, the countries which co-occur with France and Russia in these newspapers can be placed into similar categories as in ERLN according to the predominant context in which they co-occur with France and Russia. In PMGZ, both countries (Germany and Belgium) which co-occur with France occur in various contexts such that they cannot be summarized under one label. This is also the case for one of the countries which co-occurs with Russia (Finland), but the two remaining countries (Austria and Turkey) which co-occur with Russia in PMGZ do so predominantly in diplomatic/military discussions:

> "There is but one way by which the position of Austria and Russia, as well as of Serbia and Montenegro, may be sharply defined, and that is by bold action on the part of Turkey." (PMGZ 21/09/1875)
>
> "They mean to watch the fight between Russia and Turkey with complete philosophy, caring not at all which of the two wins." (PMGZ 27/04/1877)

In RDNP, most country-collocates of France and Russia occur predominantly in the context of diplomatic/military discussions. This is case for Tonquin, Siam, Sardinia and Madagascar, which co-occur strongly with France, and Turkey, Austria, Bulgaria, and Prussia, which co-occur strongly with Russia.

> "FRANCE AND TONQUIN. Paris, April 4 . The Temps publishes the following telegram, dated Hanoi yesterday : - " The French positions at Chn remain unmolested, and the district continues quiet" (RDNP 05/04/1885)
>
> "The dispute between France and Siam is reported at length in another column" (RDNP 30/07/1893)
>
> "the allied armies of France and Sardinia may be said to have been acting on the defensive" (RDNP 12/06/1859)
>
> "ENGLAND, FRANCE AND MADAGASCAR. I (says the Paris correspondent of the Times) have received the following letter (…) - " My dear Sir, - (…) The peaceful relations of England and France are in danger in this part of the world; (…)"(RDNP 16/09/1883)
>
> "Mr. I. InorsiDE moved the first resolution : - That the unjustifiable aggressions of Russia upon Turkey, the meanness and the base duplicity which Russia has manifested in support of those aggressions, and itself long-continued forcible occupation of Turkish territory, without any celourable pretext whatever, imperatively call for every nation having any regard for the principles of justice, honour, and international law to take such prompt and decisive measures as shall cause the rights of Turkey to be respected" (RDNP 25/09/1853)
>
> "A Vienna dispatch of Tuesday says: - “In well- informed quarters it is believed that an alliance between Russia and Austria is on the point of being concluded” " (RDNP 22/10/1876)

> "The Sultan is the Suzerain or Sovereign of the Soudan as he was of Bulgaria, but we objected to Russia interfering in Bulgaria, although without any regard to the Sultan we interfere in the Soudan " (RDNP 09/03/1884)

> "Russia and Prussia have met with a view of conciliating the interests and necessities of France with those of the other States. " (RDNP 04/11/1860)

For both France and Russia, one collocate occurs predominantly in advertisements: Denmark for France and Portugal for Russia. They in fact occur predominantly in the same advertisement, for Reynold's Newspaper itself:

> "According to the new postal rates *Reynolds's Newspaper* can now be forwarded to the following foreign countries for 2s. 2d. per quarter, payable in advance- viz., Austria, Australia, (…), Denmark, France, Germany, (…), Portugal, Russia, Servia, (…)" (RDNP 21/05/1876)

Finally, Germany is a strong collocate of both France and Russia, but for both France and Russia, it co-occurs in contexts so varied that they cannot be summarized under one label.

Although the strong country-collocates of Russia and France in all three newspapers can be grouped into similar categories, it is nevertheless interesting that the strongest country-collocates of France and Russia are not the same across newspapers. This may, however, be telling us more about the newspapers as a whole than about France and Russia, as mentioned above.

What of the cities? All of the cities which co-occur with France and Russia in ERLN co-occur predominantly or exclusively in the context of advertisements or discussions surrounding artistic events:

> "Grand casino, Toulouse, France. Every evening." (ERLN 08/03/1884)

> "All Engagements must in future be Addressed to AMon CLEO, CIRCUS SALAMONSKY, Warsaw (Russia)." (ERLN 16/03/1873)

In PMGZ and RDNP, the situation is very different. The single city which strongly co-occurs with Russia in RDNP (St Petersburg) occurs in various contexts which cannot be summarized by a single label. The city is indeed usually mentioned in the article byline, as an indication of the source of the article. The article that follows is then most often either a report on events happening in Russia or an article about the doings of people of standing in Russia (a kind of article which I will refer to as 'political/cultural gossip'):

> "KORE ARRESTS IN RUSSIA. St. Petersburg, August 13. Three well known lawyers have been arrested here a charge of political disaffection." (RDNP 17/08/1879)
>
> "THE EMPEROR OF RUSSIA. St. Petersburg, April 24th. The Emperor and Empress of Russia left here to-day at ten a.m. for Livadia. They were accompanied by (...)" (RDNP 27/04/1879)

No cities were among the strong collocates of Russia in PMGZ, or France in RDNP. The two cities which strongly co-occur with France in PMGZ do so predominantly in a well-defined, repetitive context. Dinard co-occurs with France in all but 1 instance in death notices. Nantes co-occurs with France predominantly in an advertisement for the Union Bank of London published repeatedly in 1876-1877.

> "I. HARRISON , Henry A. , late of the Bombay Civil Service , at Dinard , France , aged 75 , Dec. 2o" (PMGZ 03/01/1878)
>
> "The Union Bank of London. BRANCHES AT- Lyons, Marseilles, Nantes (France), Brussels (Belgium), (...). The Bank grants DRAFTS and LETTERS of CREDIT on all their Branches and Correspondents on the Continent and the East, and transacts Banking business of every description." (PMGZ 16/02/1876)

Let's now look at the other collocates of France and Russia. Looking first at France, the top collocates of France overall (in all three newspapers) can be summarized as pointing to the following patterns:

- France is mentioned in the context of diplomatic/military discussions in all three newspapers: see country-collocates referred to above plus collocates 'Governments' in ERLN and 'dismemberment', 'dismembered' and 'rapprochement' in PMGZ. RDNP does not have further top collocates (other than countries) which occur predominantly in diplomatic/military discussions.

"it is hard to believe that he [Bismarck] could have (…) read recent history to so little purpose , as to believe that the dismemberment of France would make war less likely in the future" (PMGZ 10/03/1874)

- In ERLN, France is also mentioned in the context of diplomatic discussions without (potential or effective) military overtone: see 'Algeria' above plus collocate 'passports'.

"It is now necessary that travellers to France should be provided with passports" (ERLN 12/02/1854)

- There is also some interest in all three newspapers in political events happening in France beyond their explicit relationship with other countries. Collocates 'Orient' in ERLN, 'Agriculteurs' in PMGZ and 'ruler' in RDNP predominantly occur in such contexts.

"The triennial meeting of the Grand Orient of France for the election of a new Grand Master is about to take place " (ERLN 19/05/1861)

"The annual session of the Société des Agriculteurs de France will be opened to-morrow"(PMGZ 18/02/1879)

"it is a matter for congratulation that the ruler of France has practically proclaimed a republic" (RDNP 17/11/1872)

- A number of collocates in all three papers point to the existence of various types of commercial ties between Britain and France. These include some of the city-collocates mentioned above, as well as the collocates 'exported' in ERLN and 'Montauger' in PMGZ.

"Of the brandy exported from France this year England took 84,439 hectolitres" (ERLN 31/07/1864)

"PEAT COAL AND CHARCOAL (…) Monsieur Challeton de Brugisat has been engaged at Montauger , in France , in improving this manu-facture (…)" (PMGZ 29/08/1873)

- A number of collocates in all three papers also point to the existence of ties at cultural and personal levels. These include the sports-related collocate 'bred' in ERLN, collocates which point to the existence of British tourism in France such as ‘voyager’ in RDNP and ‘southwest’ in PMGZ (from a weather report), collocates which occur in political/cultural gossip articles such as ‘Marshals’ and ‘tricolour’ in ERLN, ‘Dinard’ in PMGZ which occurs primarily in death notices as mentioned above, and the collocate ‘inundations’ in ERLN (which often occurs in reports of charity fund-raisers in Britain).

"horses of 3 yre and upwards foaled and bred in France " (ERLN 06/06/1855)

"G. W. M. Reynolds fait voyager en France M. Pickwick , au grand amusement de ses lecteurs[6]" (RDNP 17/04/1853)

"Heavy rain prevails in the southwest of France " (PMGZ 20/07/1874)

"she sang for the benefit of the sufferers of the inundations in France" (ERLN 01/08/1875)

Together, all these collocates suggest a complex, multi-faceted relationship between Britain and France which cannot be reduced simply to military rivalry; this is hence not in line with *h1* (see section 7.2), but may contribute to meeting *h5*. No negative bias emerged from looking at the top collocates for France in the three newspapers. This is hence not in line with *h4*, but neither does it provide any conclusive evidence on this, since only a small selection of collocates (those with the strongest effect size) were considered in each newspaper.

Moving to other collocates (beyond cities and countries) of Russia, some patterns similar to those identified for France emerge, but there are also differences. The main patterns which stand out are:

- Like France, Russia is mentioned in the context of diplomatic/military discussions in all three newspapers, with, in addition to the countries mentioned above, ‘invasion’, ‘Baltic’ and ‘Porte’ in ERLN, ‘coquetting’, ‘embroil’ and ‘rapprochement’ in PMGZ, and ‘aggressions’, ‘preponderance’, ‘encroachments’, ‘Emperors’, ‘autocrat’, ‘aggressive’ and ‘Asia’ in RDNP all occurring predominantly in diplomatic/military discussions.

[6] In French in RDNP. My translation: ‘G. W. M. Reynolds has Mr. Pickwick travelling in France, for the amusement of his readers.’

"Letters from Constantinople of the Ist of March , hint that a serious misunderstanding has arisen between Russia and the Porte in reference to Servia " (ERLN 26/03/1843)

"Whether such an enterprise embroiled us with Russia or not, it appears to us that to enforce the letter of the Treaty of Berlin in this particular would be almost hopeless" (PMGZ 11/11/1878)

"The more we learn of the Czar's recent massacres at Warsaw, the deeper our indignation at the almost unparalleled perfidy of the anointed despot, and the firmer our conviction that the present autocrat of Russia is one of the most cold-blooded, cunning, cruel, and hypocritical monarchs who ever sat upon a throne. " (RDNP 21/04/1861)

- For France, we saw that two of the strong collocates in ERLN occurred predominantly in political/cultural gossip. For Russia, a number of strong collocates in all three newspapers occur predominantly in this context. This is the case for, in ERLN, ‘emperor’, ‘Vladimir’, ‘czar’, ‘Constantine’, and ‘Empress’, in PMGZ, ‘Michaelovitch’, ‘Michailovitch’, ‘Sergius’, ‘Alexandrovitch’, ‘Vladimir’, ‘tsars’, ‘Alexis’, ‘Serge’ and ‘Constantine’, and in RDNP, ‘empress’. It was also mentioned above that St Petersburg, a strong collocate in RDNP, also occurs part of the time in such a context in RDNP.

"By a preconcerted arrangement the Emperor of Russia, accompanied by the Crown Prince of Wurtemberg and the Prince of Hesse, his brother-in-law, with their suites, started exactly at eleven from the villa of the Grand Duchess Olga to Canstadt, to witness this ceremony, and to honour it with their presence" (ERLN 04/10/1857)

"They have, to begin with, photographed the Queen over a hundred times, and have taken, in addition, almost every crowned head in Europe, (...) they have taken three Tsars of Russia, three Sultans of Turkey, the King and Queen of Denmark, (…)" (PMGZ 01/06/1897)

"The Emperor and Empress of Russia arrived here at eleven o'clock this morning, and were received by the principal civil and military authorities and the foreign ministers." (RDNP 02/09/1883)

- As was observed for France, a number of the strong collocates of Russia in all three papers indicate that there is interest in political events happening in Russia, beyond their explicit relationship with other countries. The following collocates hence co-occur with Russia predominantly in the context of discussions of events internal to Russia: in ERLN, ‘aggressive’, ‘intrigues and ‘Jews’, in PMGZ, ‘nihilism’, and in RDNP, ‘nihilism’, ‘nihilists’, ‘nihilist’, ‘autocratic’ and ‘serfs’.

"When first the outrages against the Jews in Russia became known the horror and commiseration of the British nation were aroused" (ERLN 01/07/1882)

"The Daily Telegraph asks if there is Nihilism today in Russia, how much of its portentous growth is due to the cold tyrant whose heart was broken by the Crimean War!" (PMGZ 02/03/1880)

"We know how the minds of the wretched serfs of Russia are trained to tolerate and uphold the horrible system of government which exists in that country."(RDPN 27/08/1854)

- For France, many of the strong collocates indicated cultural and commercial ties between Britain and France. No such trend emerges for Russia, although a few strong collocates of Russia in ERLN and PMGZ come from ads: ‘Livadia’ and the cities mentioned above in ERLN and ‘tsarina’ in PMGZ.

"August and September, St. Petersburg, Theatre de Livadia, Russia." (ERLN 04/07/1885)

"Superb and richly arranged Drawing-room Tableau, the late Lord Randolph Churchill, M. Casimir-Perier, the late President of the French Republic, her Majesty the Queen holding a Drawing Room, the Tsar and Tsarina of Russia, our leading Ecclesiastical Dignitaries , &c., &c." (PMGZ 06/02/1895)

- Finally, one trend which emerges in the strong collocates of Russia in all three papers which did not emerge for France is one of negative bias. A number of the collocates of Russia occur in contexts in which Russia is clearly cast in an unfavourable light: in ERLN, ‘aggressive’ and ‘intrigues’, in PMGZ, ‘embroiled’, ‘coquetting’ and ‘embroil’, and in RDNP, ‘aggressions’, ‘preponderance’, ‘encroachment’, ‘autocratic’, ‘autocrat’, ‘aggressive’, and ‘serfs’. Some examples featuring these collocates were already presented above, but here are some more:

"The complaints of the enormous intrigues of Russia are becoming universal." (ERLN 23/10/1842)

"There has been a great deal of coquetting with Russia." (PMGZ 01/06/1871)

> "Now is the time for the Western Powers to adopt bold and decisive measures to arrest the encroachments of Russia." (RDNP 10/07/1853)

For France, we could say confidently that the collocates pointed to a complex, multi-faceted relationship between Britain and France. For Russia, it is less straightforwardly so. Many of the collocates appear in political and diplomatic/military contexts, and there was no emergent pattern pointing to the existence of a rich web of cultural and commercial ties as was found for France. Conversely, a number of the strong collocates of Russia appeared predominantly in contexts which featured Russia in a bad light, whereas this pattern did not emerge for France. I have of course only looked at the strongest collocates, so the patterns which were not found here may be evident when looking at other collocates.

7.4.3 SUMMARY OF FINDINGS IN THIS SECTION

The global approach turned out not to be well-suited for comparisons between newspapers. Other approaches, such as using keyness analysis, would probably be more helpful for this. Nevertheless, the approach *can* tell us about some of the main contexts in which the countries are mentioned. Looking at the strongest collocates for the countries, we find the following:

- Both countries have a strong tendency to co-occur with mentions of other countries. These co-occurrences arise in a variety of contexts, including, but not restricted to, military/diplomatic discussions. That these military/diplomatic discussions occur is in line with *h1*, but since these discussions do not constitute an overwhelming proportion of the discussion surrounding these countries, overall this finding undermines *h1*.
- Both countries have a strong tendency to co-occur with cities. In ERLN, this is almost exclusively in the context of advertisements or discussions surrounding artistic events. In PMGZ and RDNP, the contexts are more diverse. For Russia in RDNP, only one city (St Petersburg) strongly co-occurs with Russia, and it

usually occurs in the article byline, most often followed by a report of happenings in Russia, or by political/cultural gossip. France does not have strong city collocates in RDNP, nor Russia in PMGZ. Two cities co-occur strongly with France in PMGW. Dinard and France almost always co-occur in death notices; Nantes and France predominantly co-occur in an advertisement for the Union Bank of London published repeatedly over a two year period. These points show us that the countries occur in a variety of contexts, but they also point to limitations in the approach. The fact that the approach does not allow for comparisons between newspapers reduces the value of these findings. Another limitation is that very repetitive genres, such as advertisements, will yield very strong collocations; this may be misleading, and means that collocation patterns must always be considered with respect to their dispersion, as well as their strength. In our case, the dispersion which is relevant is across articles, and article genres, but these dispersions are difficult to generate automatically[7]. This is an issue which could benefit from further research.

- Looking at other strong collocates of France, we find that France is mentioned in the following contexts: diplomatic/military discussions in all three newspapers; diplomatic discussions without military overtones in ERLN; and political happenings in France without explicit relation to other countries in all three newspapers. In addition, collocates in all three papers point to the existence of commercial, cultural and personal ties between Britain and France. Together with the findings about France and countries and cities, this undermines *h1*, may contribute to meeting *h5*, and fails to provide evidence for *h4.*
- Looking at other strong collocates of Russia, we find that Russia is mentioned in the following contexts: diplomatic/military discussions in all three newspapers, political/cultural gossip in all three newspapers; and political happenings in

[7] This will be easier to do in future versions of CQPweb (Andrew Hardie, personal communication, 17th Jan. 2017).

Russia without explicit relation to other countries in all three newspapers. In addition, collocates in all three newspapers betray a negative bias towards Russia.

With regard to the hypotheses formulated in section 7.2, we can say in summary that:

- With regard to *h1*, both countries are mentioned in diplomatic/military contexts in all three newspapers, which is in line with *h1*, but also in other contexts, which is not in line with *h1*, and suggests that this hypothesis is reductive or simply incorrect.
- Nothing can be said with regard to *h2*, since this approach does not consider change over time.
- Nothing can be said with regard to *h3*, since this approach does not take article genre into account.
- With regard to *h4*, this approach may not allow us to uncover all relevant patterns. Nevertheless, we have clearly found evidence of some anti-Russian angles in all three newspapers, which may be in line with *h4*. We do not find a corresponding pattern for France, though the approach cannot be taken as providing evidence for its absence. The presence of political/cultural gossip, which was pointed to by collocates of Russia in all three papers, is perhaps also in line with the other side of *h4*. Traces of this for France were found in ERLN but not in the other two papers, although again the approach cannot be taken as providing evidence for their absence. Nevertheless, on the basis of what has been found so far, it seems to be the case that this hypothesis fits the case of Russia more closely than that of France.
- With regard to *h5* and *h6*, findings from looking at the strongest collocates of France and Russia can be considered in line with these hypotheses, since we found evidence of cultural and commercial ties with France, but little evidence of this with Russia (although, again, this approach cannot generate evidence of absence). As with *h4*, *h5* and *h6* appear more tailored to the case of Russia than to that of France.

7.5 THE SAMPLING APPROACH

As introduced in section 7.4.1, this second approach to answering the question 'what is said about these places in these texts?' starts by looking at a small sample and then zooms out. The approach involves two main phases: developing a framework for analysis based on as many samples as required (described in section 7.5.1), and exploiting the framework (section 7.5.2).

7.5.1 DEVELOPING THE FRAMEWORK

7.5.1.1 Method

Unlike the previous approach, which starts with identifying patterns of collocation in the whole dataset, then 'zooms in' to read selected concordance lines in more detail, this approach starts with reading concordance lines and describing them. Based on these descriptions, patterns are identified which can then be investigated in the whole dataset. By necessity, then, the approach requires down-sampling[8], since a researcher cannot, at the outset, read all relevant concordance lines in the kind of large datasets considered in this thesis.

Hence, the first two steps (as outlined in section 7.4.1) are simply to search for the node in the whole corpus, then take a random sample of concordance lines. I initially developed the framework based on data from ERLN. I started off with samples of 10 concordance lines: 2 samples from the last decade (1890-1899) of ERLN for each country, then 2 samples from the first decade (1840-1849) for each country. As the framework started to take shape, I expanded the samples to 60 concordance lines for each country: 10 from each of 6 decades in ERLN (1840-1849, 1850-1859, 1860-1869, 1870-1879, 1880-1889, 1890-1899). After considering 5 such samples (or 600 concordance lines: 300 per country), I then tested the framework on 5 further samples of 10 concordance lines (for each country and decade) from PMGZ and RDNP.

For each sample considered, I simply read through each concordance line and attempted to describe the phraseology within which the name of the country was being used. My definition of 'phraseology' was kept intentionally loose to begin with, in order to remain open to what

[8] i.e. producing a smaller sample which is of a size that can manageably be analysed, see section 2.3.3.3.

would emerge from my observation of the data. At first, it simply referred to the immediate linguistic context of the mention of the country. But as the framework started to take shape, the definition of 'phraseology' also became more precise. In the end, the framework developed into a slot-schema framework[9], where each phraseology is described using a sequence of 'slots' which are more or less specified depending on the similarity of the concordance lines which are described by that slot-schema.

Before presenting the system of categories, let us consider a few implications of the method of developing the system. First, since the system is developed based on random concordance lines, it can be expected to work well for common patterns. For uncommon patterns, however, the system may not be very effective, either because too few instances of that pattern were encountered during the development phase, so that the descriptions of the pattern are not very accurate, or because *no* instances were encountered during the development phase. Second, some categories are more general than others; similar phraseologies which are very common may end up described by several categories, whereas similar phraseologies which are uncommon will end up lumped together in a single category.

Third, the system of categories ultimately depends not only on the examples encountered, but also on the gaze of the researcher. I did, however, want to check whether the system I developed would be usable by someone else. In order to test this, I gave a fellow corpus linguist documentation (consisting of Figure 7.9 and Table 7.6 to Table 7.10) describing the framework I had developed, as well as 20 concordance lines (10 per country) taken at random from the whole period covered in each of three newspapers (i.e. amounting to a total of 60 concordance lines): ERLN, the *Hampshire/Portsmouth Telegraph* (HPTE), another c19th British Library newspaper covering a similar historical period but which had not been involved in the development of the framework; and the *New York Times* (NYT), a twentieth-century newspaper.

[9] By *slot schema*, I simply mean a (linear) sequence of slots, with each slot being filled by one or more words. Although some slot schemas may end up describing a typical grammatical construction, the slot schemas are not *intended* to capture specific constructions; they are simply a description of which words tend to follow each other.

Both of us independently categorized these concordance lines using the framework. A high level of agreement was achieved: our classification was the same for all NYT and HPTE concordance lines. For ERLN, all but 4 concordance lines were given the same rating. For 3 out of these 4 diverging ratings, the difference was simply human error – upon reconsidering, the other linguist decided they would have actually placed the line in a different category (the same as the one I used). The last case was a genuine case of disagreement, but it also concerned an unusual case which had not been encountered whilst developing the framework.

The results of this test hence suggest that the framework is reasonably objective – it can be used by different researchers with very similar results. They further suggest that the framework may lend itself to the analysis of newspaper texts from various time-periods. The next section describes the framework.

7.5.1.2 The final framework

7.5.1.2.1 Structure of the framework: slot-schemas

The process described in the previous section resulted in a hierarchical framework containing, at the lower level, *labels* (or categories), and at the higher level *over-arching categories*. Each *label* refers to a set of similar phraseologies which can be described using a *slot-schema*: a sequence of slots specifying the grammatical and/or semantic characteristics of the word(s) which can occupy that slot. The description of each slot starts out very specific, then becomes more abstract as similar phraseologies are grouped together. For example, the following three concordance lines contain very similar uses of the country-names France/Russia (in bold, the words covered by the slot-schema):

> "…negotiations were on foot for the conclusion of **a commercial treaty between England and France**…" (ERLN, 06/01/1861)
>
> "We may expect any day to hear of **a new combination between Austria and Russia**…" (ERLN, 15/08/1875)

> "The Excitement in Constantinople , caused by **the late War between Russia and Turkey**…" (ERLN, 08/12/1878)

Initially, these three examples are described using the slot-schema **NOUN_PHRASE between COUNTRY and COUNTRY**, which is given an alphanumeric *label* (e.g. **A1**) which is, at first, arbitrary. In the slot-schema, elements in normal letters represent fixed parts of the phraseology: the slot described by 'between' can *only* be filled by the word 'between'. By contrast elements in capital letters represent non-fixed parts: the slot described by 'COUNTRY' can be filled by any country-name (France, Russia, Austria, etc.). Underscores (_) are used to join together words which are part of the description of the same slot. Hence this first slot-schema contains 5 slots: 'NOUN_PHRASE', 'between', 'COUNTRY', 'and', 'COUNTRY'.

Later in the process, concordance lines such as the following are encountered:

> "he considered that a **war with Russia** would be a great advantage to this country…" (ERLN, 16/06/1839)

Upon reflection, and in comparison with other phraseologies encountered, I decide that this example represents a very similar usage of COUNTRY to those presented above: COUNTRY is being referred to as part of a noun phrase modifier, in a way which emphasizes the relationship between several countries. The slot schema presented above is then modified to incorporate this new set of examples: 'between' becomes the more abstract RELATIONAL_PREPOSITION (i.e. a preposition which emphasizes the relationship between countries), and the slot-schema as a whole becomes simply **NOUN_PHRASE RELATIONAL_PREPOSITION COUNTRY (and) (COUNTRY)**, with two optional elements indicated in brackets.

7.5.1.2.2 Over-arching categories

As more and more phraseologies are encountered, over-arching groupings emerged. At first, I identified two over-arching categories: *locational* phraseologies and *personifying* phraseologies, which describe, respectively, phraseologies where COUNTRY is used to indicate a geographical location or space, and phraseologies where COUNTRY is used as an agent, capable of action and emotion. Examples of locational phraseologies:

> "**the news from France**" (*The Era*, 27/02/1848)
>
> "a naval fight occurred on **the coast of France**" (*The Era*, 22/02/1896)
>
> "We are the only Agents who have **travelled over India , (…) Germany , Russia**" (*The Era*, 12/6/1886)

Examples of personifying phraseologies:

> "existing treaties between Great Britain and Russia" (*The Era*, 05/04/1840)
>
> "the wish and dictation of France" (*The Era*, 25/04/1858)
>
> "if any European power opposes Russia in her projects" (*The Era*, 10/6/1866)

Two further over-arching categories were later added. The first, *specialized* categories, grouped together 4 labels for phraseologies which had in common a limited diversity of expression, and occurrence in a limited set of contexts (see Table 7.9), including examples such as :

> "Engagement with **CIRCUS CINISELLI , ST . PETERSBURG , RUSSIA** , on their Three Horizontal Bars" (*The Era*, 17/2/1883)
>
> "**France , Mlay 25th , 1888** . My dear Goddard , Your letter of 2ed has reassured me…" (*The Era*, 23/6/1888)
>
> "Tire total quantity amounted to 2,689,000 bottles , which were thus distri-buted : **England and British India , 467,000** ; Russia and Poland , 502,000 ; " (*The Era*, 24/10/1852)

> "Penetrating Hair Brushes, with the durable unbleached **Russia Bristle**, which do not soften like common hair." (*The Era*, 05/5/1850)

The last over-arching category, *affiliative* categories, grouped together a set of 4 labels, all described using a similar **NOUN_PHRASE of COUNTRY** slot-schema, but encompassing cases where the noun phrase referred to a person, group of persons, or institution:

> "the famous Ciniselli family of Russia" (*The Era*, 15/04/1899)
>
> "the merchants and manufacturers of Great Britain, of France" (*The Era*, 22/03/1840)
>
> "the monthly return of the Bank of France has been received today" (*The Era*, 13/09/1857)

These labels were grouped together into a separate over-arching category because it was difficult to conceptually group them strictly either with the *locational*, or with the *personifying* phraseologies. However, since these phraseologies occur frequently in the data I considered, grouping them with either over-arching category would have had an important impact on the quantitative picture which emerged.

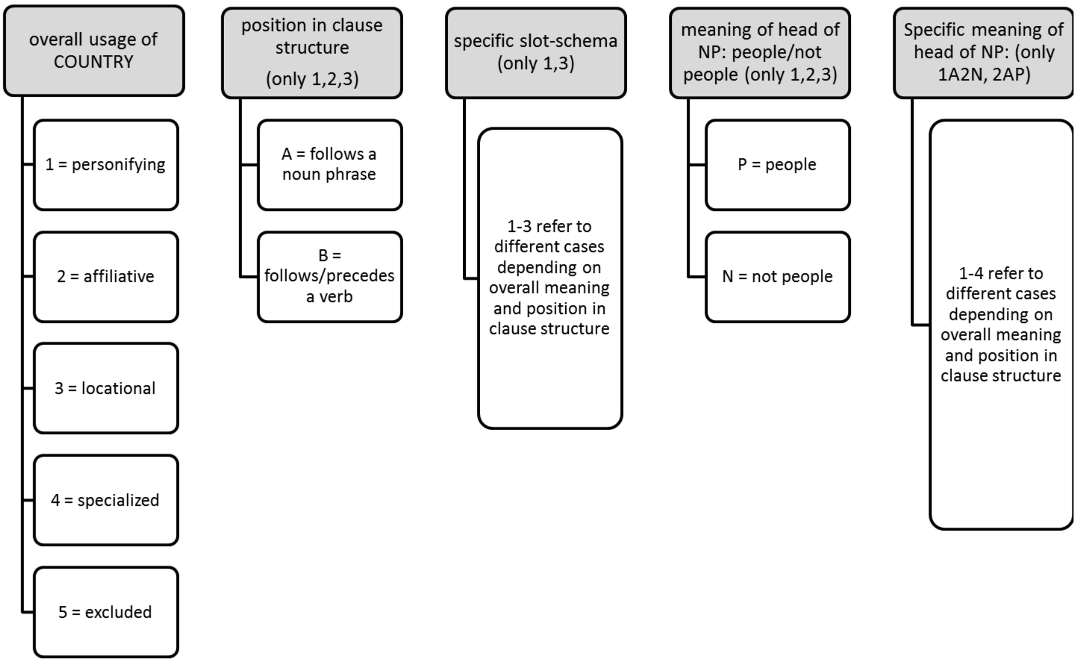
Finally, some concordance lines had to be excluded altogether, for one of three reasons. First, when the poor quality of the surrounding OCR precludes identification of the correct label. Second, when COUNTRY occurs as part of the title of an article, book, performance, etc. Third, when COUNTRY does not in fact refer to the country-name. Here is one example of each of these cases, respectively:

> "hnv depsrtisUe **thatI llgt for Russia**, whither business... " (*The Era*, 16/1/1876)
>
> "By:FRANCIS PARRIklIAN , Author cf **" Pioneers of France in the New World , "** Map . 8vo , l0ex 6d" (*Pall Mall Gazette*, 01/1/1870)
>
> "entertainments , in which Messrs . Charles Foster , Marden , France, Raymond , **Miss Kate France** , and Miss Connelly sustained..." (*The Era*, 16/4/1871)

7.5.1.3 Labels

Once the framework reached a fairly stable form, it became possible for me to organize the labels in a non-arbitrary manner; see Figure 7.9. Labels contain between 2 and 5 characters, the first of which is a number between 1 and 5 which refers to the over-arching category or exclusion. For 4 (*specialized* categories) and 5 (*excluded* cases), this first number is followed by a second number which refers to the specific case or slot-schema. For the other 3 over-arching categories, the first number is followed by a letter which refers to the position of COUNTRY in the clause structures – A for noun phrase modifier, and B for verb complementation (including subjects). This is then followed by up to 3 characters referring to the specific slot-schema.

Figure 7.9. Reading a label


overall usage of COUNTRY
1 = personifying
2 = affiliative
3 = locational
4 = specialized
5 = excluded
position in clause structure (only 1,2,3)
A = follows a noun phrase
B = follows/precedes a verb
specific slot-schema (only 1,3)
1-3 refer to different cases depending on overall meaning and position in clause structure
meaning of head of NP: people/not people (only 1,2,3)
P = people
N = not people
Specific meaning of head of NP: (only 1A2N, 2AP)
1-4 refer to different cases depending on overall meaning and position in clause structure

Tables 7.6 to 7.10 summarize the framework. The 1.B2 label is worth noting for its peculiar slot-schema. Phraseologies in this category are very diverse: each individual phraseology is rare, so that the resulting category exists at a high level of abstraction than the other labels. Hence, its slot-schema no longer captures the details of the sequences of words which may occur in this category, and the description is reduced simply to the abstract observation that COUNTRY will be some kind of complement of the verb.

Table 7.6. Labels in the personifying category

Label	Schema	Definition and examples
1.A1	**NOUN_PHRASE (relational_preposition) COUNTRY**	**the preposition (typically ‘with’, ‘between’ or ‘against’) emphasizes the relationship between COUNTRY and another entity**
		--> “through a declaration of war with Russia” (*The Era* , 07/11/1885)
		--> “existing treaties between Great Britain and Russia” (*The Era* , 05/04/1840)
1.A2.N1	**NOUN_PHRASE OF COUNTRY**	**The noun phrase is a material possession (military, financial)**
		--> “the combined fleets of France and Spain” (*The Era* , 03/02/1867)
1.A2.N2	**NOUN_PHRASE OF COUNTRY**	**The noun phrase is an immaterial possession (political, ideological)**
		--> “the emblem of France is the figure of Liberty” *(The Era* , 10/02/1878)
1.A2.N3	**NOUN_PHRASE OF COUNTRY**	**The noun phrase is a property or action (nominalization)**
		--> “the wish and dictation of France” (*The Era* , 25/04/1858)
		--> “take place without the consent of England and France” (*The Pall Mall Gazette* , 07/06/1882)
1.B1	**COUNTRY PREDICATE**	**COUNTRY is the subject of a verb (possibly separated by ‘to’)**
		--> “Russia is greedy for Batoum” (*The Era* , 07/7/1878)
		--> “What is it that France wants?”(*The Pall Mall Gazette* , 01/01/1870)
1.B2	**COUNTRY=VERB_COMPLEMENT**	**COUNTRY is a non-locative complement or modifier of a verb, e.g. (in)direct object, agent of passive**
		--> “if any European power opposes Russia in her projects” (*The Era* , 10/6/1866)
		--> “several extraditions have been obtained by and accorded to France” (*The Pall Mall Gazette* , 24/02/1880)

Table 7.7. Labels in the affiliative category

Label	Schema	Definition and examples
2.A.P1	NOUN_PHRASE OF COUNTRY	**The noun phrase is a named individual (e.g. a personality) or group (e.g. an artistic group)** --> “the famous Ciniselli family of Russia” (*The Era* , 15/04/1899)
2.A.P2	NOUN_PHRASE OF COUNTRY	**The noun phrase is an individual(s) referred to by their position in the political hierarchy** --> “honoured last week by the presence of the Emperor of Russia” (*The Era* , 16/06/1867)
2.A.P3	NOUN_PHRASE OF COUNTRY	**The noun phrase is a broad group of people or unnamed individual** --> “the merchants and manufacturers of Great Britain, of France” (*The Era* , 22/03/1840)
2.A.P4	NOUN_PHRASE OF COUNTRY	**The noun phrase is a legal person/entity (i.e. an institution or business)** --> “the monthly return of the Bank of France has been received today” (*The Era* , 13/09/1857)

Table 7.8. Labels in the locational category

Label	Schema	Definition and examples
3.A1	NOUN_PHRASE (LOCATIONAL_PREPOSITION) COUNTRY	**the preposition (typically ‘to’, ‘from’, ‘through’) emphasises movement or situation to, from, or within COUNTRY (possibly metaphorical)** --> “the news from France” (*The Era* , 27/02/1848) --> “an extended Tour through France” (*The Era* , 26/05/1878)
3.A2	NOUN_PHRASE OF COUNTRY	**The noun phrase is a geographical feature** --> “a naval fight occurred on the coast of France” (*The Era* , 22/02/1896)
3.A3.P	NOUN_PHRASE IN COUNTRY	**The noun phrase is a person or group of people** --> “I received my education as an artiste in France” (*The Era* , 08/04/1860)
3.A3.N	NOUN_PHRASE IN COUNTRY	**The noun phrase is not a person or group of people** --> “a great deal of public remark, both in this country and in France” (*The Era* , 06/08/1843)
3.B1	VERB (LOCATIONAL_PREPOSITION) COUNTRY	**COUNTRY is a locative complement/modifier of a verb** --> “We are the only Agents who have travelled over India , (...) Germany , Russia” (*The Era* , 12/6/1886)
3.B2	ADJECTIVE LOCATIONAL_PREPOSITION COUNTRY	**COUNTRY complements/modifies an adjective (incl. comparative or superlative)** --> “It was permissible in France to perform little vaudeville” (*The Era* , 14/05/1892)
3.B3	IN COUNTRY	**COUNTRY occurs in a prepositional phrase to mark location, but is separated from the main verb** --> “In France, an official of a more refined grade of intelligence is consulted when ...” (*The Era* , 17/06/1866)

Table 7.9. Labels in the specialized category

Label	Schema	Definition and examples
4.1	(TROUP/VENUE) (TOWN) COUNTRY	**COUNTRY occurs as part of an address, or to locate (or provide the affiliation) of a venue, person, group (e.g. of performing artists), settlement, etc.** --> "Engagement with CIRCUS CINISELLI, ST . PETERSBURG, RUSSIA, on their Three Horizontal Bars" (*The Era* , 17/2/1883) --> "A native of Grodno, Russia, shot himself dead" (*Reynold's Newspaper* , 08/06/1890)
4.2	COUNTRY DATE	**COUNTRY is immediately followed by a date, as part of date-stamping a letter or article** --> "France, May 25th, 1888. My dear Goddard, Your letter of 2nd has reassured me..." (*The Era* , 23/6/1888)
4.3	COUNTRY NUMBER	**COUNTRY occurs in a list of numbers and places** --> "The total quantity amounted to 2,689,000 bottles, which were thus distributed: England and British India, 467,000; Russia and Poland, 502,000;" (*The Era* , 24/10/1852)
4.4	COUNTRY NOUN	**COUNTRY is used as a noun adjunct** --> "Penetrating Hair Brushes, with the durable unbleached Russia Bristle, which do not soften like common hair." (*The Era* , 05/5/1850)

Table 7.10. Labels for the excluded cases

Label	Schema	Definition and examples
5.1	UNCLEAR: OCR	**(poor-quality OCR precludes identification of the phraseology)** --> "hnv depsrtisUe thatI llgt for Russia, whither business..." (*The Era* , 16/1/1876)
5.2	FRAGMENT: TITLE	**(COUNTRY is part of a title, e.g. of an article, book or show)** --> "By:FRANCIS PARRIklIAN , Author cf " Pioneers of France in the New World , " Map . 8vo , l0ex 6d" (*Pall Mall Gazette* , 01/1/1870)
5.3	EXCLUDED: not referring to country	**(COUNTRY does not refer to the country, e.g. it is a person's name)** --> "entertainments , in which Messrs . Charles Foster , Marden , France, Raymond , Miss Kate France , and Miss Connelly sustained..." (*The Era* , 16/4/1871)

7.5.2 Exploiting the framework

There are two main ways to use the framework once it has been developed. One is to take samples from the whole corpus and analyse them using the entire set of categories in the framework. This procedure is virtually identical to the analysis involved in the development of the framework. The only difference is that the framework is no longer amended as the analysis proceeds. This approach is illustrated in section 7.5.2.1.

The other way is to select one or several specific labels to focus on. Search queries can then be devised to locate most instances of these labels in the entire dataset. In principle, the entirety of the results thus extracted can then be analysed further. This approach is illustrated in section 7.5.2.2.

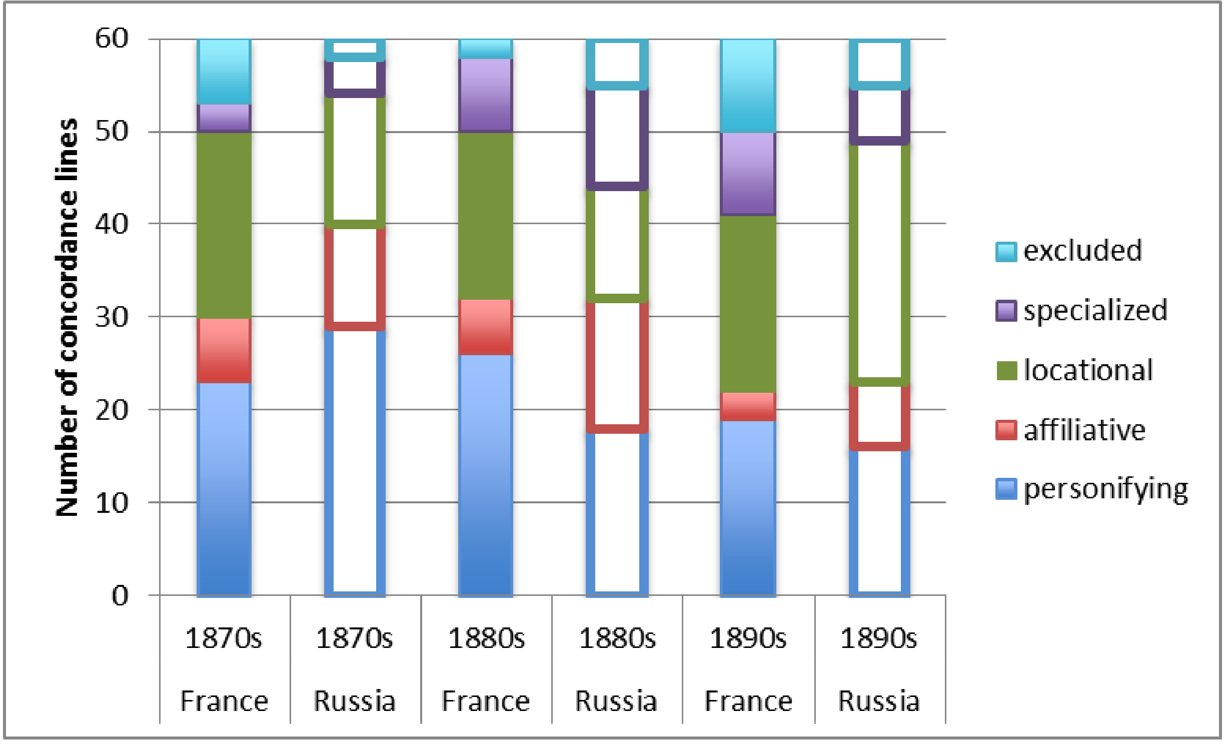
7.5.2.1 All the categories in the sample

Using the framework as a whole can be a broad-stroke approach to comparing various datasets – whether over time, from different publications, and so on. To illustrate this, I present results from comparing the phraseologies associated with France and Russia in three decades of ERLN, PMGZ and RDNP: 1870-1879, 1880-1889 and 1890-1899. The analysis is based on relatively small samples: 20 concordance lines taken at random for each country/decade/newspaper combination (totalling 360 concordance lines in total). It is presented here for purposes of illustration rather than as an attempt at a comprehensive account of the representations of France and Russia in these newspapers. For a more in-depth study, the analysis could easily be extended to more concordance lines, more newspapers, more decades, more countries, etc.

Figure 7.10 shows the distribution of concordance lines between the different labels for France and Russia in all three newspapers put together. Each bar represents a total of 60 concordance lines (20 from each newspaper); this is in fact true for all the figures in this section (except Figure 7.11 where the newspapers are shown in separate graphs, so that each bar represents a total of 20 concordance lines). On the whole, not very much can be deduced from

Figure 7.10. This is in part because it rests on relatively little data. But another reason is that, with respect to diachronic patterns, it is almost always possible to find a (so-called) pattern when looking at data plotted on a timeline with only three measuring points. Reliably establishing such a pattern, however, will require more evidence (such as data plotted over a longer time-period, or with more granularity in the periodization).

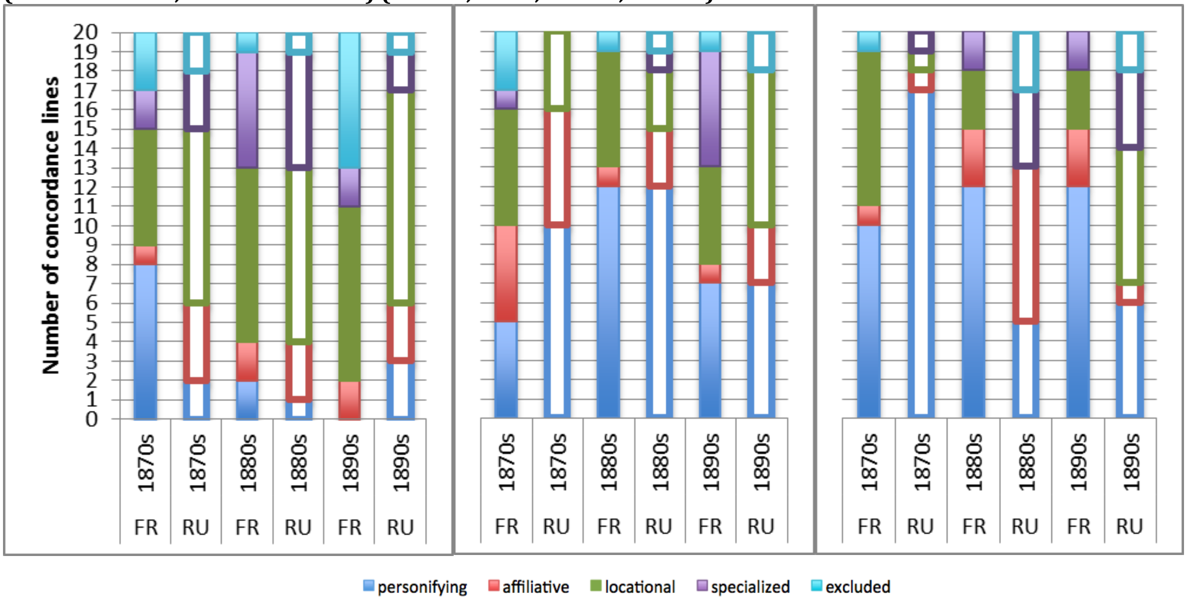
Figure 7.10. Proportion of concordance lines assigned to each category of labels, for France (filled) and Russia (hollow), for samples of 20 concordance lines per decade from each of ERLN, PMGZ and RDNP (put together)



Here are some observations which *can* be made from this figure. Both countries attract labels in all 5 over-arching categories (including exclusion cases) in all three decades, and for the most part the profile of both countries looks more similar than different. In the 1870s and 1880s, the personifying labels have the greatest proportion of concordance lines for both countries, closely followed by the locational ones; in the 1890s, this is reversed, with the locational labels attracting the greatest proportion, closely followed by the personifying ones. The sum of personifying and locational phraseologies encompasses, in all cases, at least 50% of the concordance lines, and in all but 2 cases (Russia in the 1880s and France and the 1890s), at least 70%. Specialized and affiliative phraseologies are in all cases the least numerically important categories, though in the case of affiliative phraseologies, little should be made of this observation, since, as shown in Table 7.6 to Table 7.8, these are similar in structure as well as meaning to phraseologies which are counted as personifying and locational phraseologies.

Both countries are thus referred to using varied language which, overall, is neither dominated by the personifying, nor by the locational, phraseologies. At this level, similarity between the presentations of the country hence stands out more than any differences. If any difference can be discerned, it is only that Russia, in each decade, attracts around double the number of affiliative phraseologies as France.

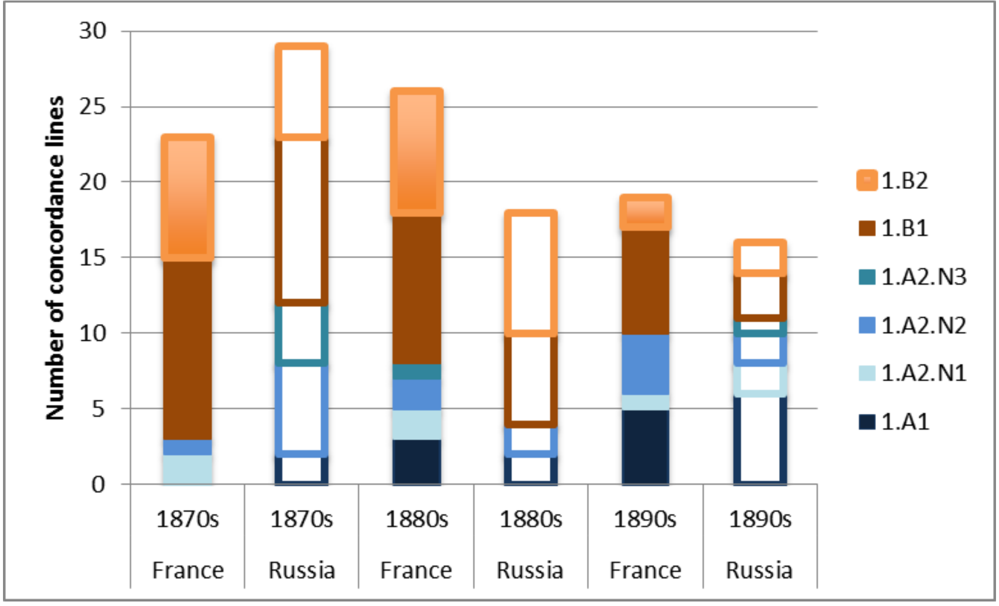
Figure 7.11. Proportion of concordance lines assigned to each category of labels, per decade and newspaper (in order: ERLN, PMGZ and RDNP) (France, filled; Russia, hollow)



This overview, however, hides differences among the newspapers. Figure 7.11 splits the numbers across newspapers. Most strikingly, the most important category for both countries in ERLN is the locational category, followed by the specialized category. In PMGZ and RDNP, by contrast, the most important category for both countries is the personifying category, followed by the locational category. The specialized category is less prominent in these two newspapers, and there are also fewer excluded cases, than in ERLN. There are also noticeably fewer personifying phraseologies in ERLN in all decades for both countries, except in the 1870s for France. The most obvious explanation for this is that we are seeing here again the impact of the differences in generic make-up between the newspapers, with ERLN having many more articles in the arts and entertainment and advertisements genres than the other two papers in the decades considered here. Although some differences between France and Russia are observable – most notably that France seems to have a distribution which is slightly more stable from one decade to the next than Russia in all three papers – no clear trend emerges.

What seems most clear is that looking at newspapers together actually hides a lot of variation. If we increased the chronological granularity, we might well find that what we see here also, in fact, hides a lot of variation over time. Nevertheless, the observation above that both countries attract phraseologies in all the over-arching categories in all decades seems to hold for ERLN; for PMGZ and RDNP, however, some decades do not present specialized phraseologies and/or excluded cases for one or both of the countries. However, we cannot know whether these are absent outside the random samples of concordance lines.

Figure 7.12. Labels within the personifying category, all newspapers together (France, filled; Russia, hollow)

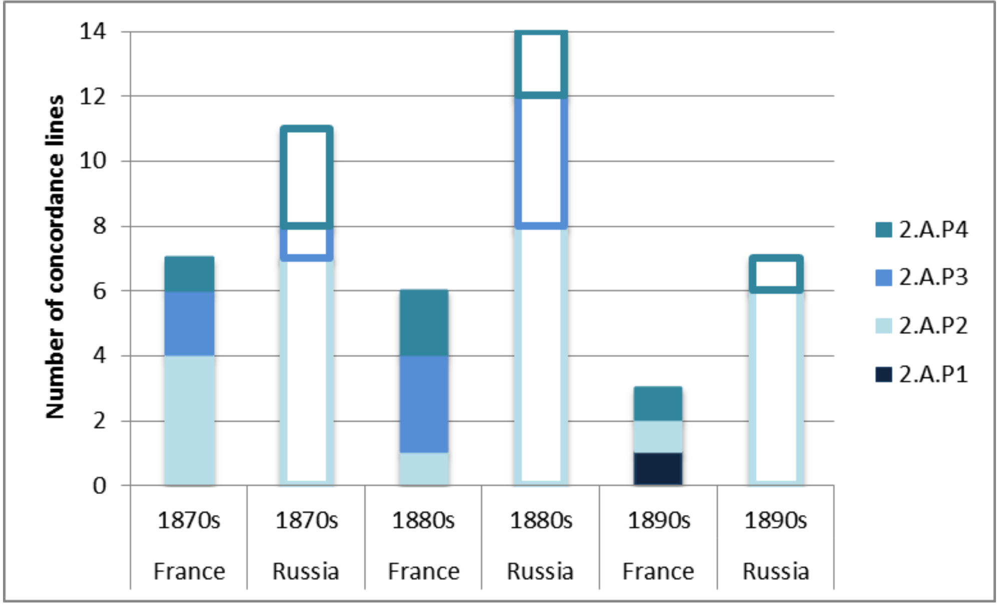


What about the different labels within categories? More differences are found in the distribution of labels within each category than between the three over-arching categories. Figure 7.12 shows the distribution of concordance lines from all three newspapers put together, across the various **personifying labels**. For France, the verb complement patterns (1.B1 and

1.B2) together encompass most of the concordance lines overall, although their proportion diminishes drastically, from 88% of the concordance lines in the 1870s to just under 50% of the concordance lines in the 1890s. These categories are slightly less important for Russia overall, and the trend of diminution over time is not observed (the proportion is greater in the 1880s than in the 1870s or the 1890s), although it is conceivable that looking at more data would reveal such a trend as well (or cast doubt on that trend for France). Perhaps the greatest difference seen here, however, lies in the distribution of concordance lines within the noun phrase modifier labels: 1.A2.N3 phraseologies (properties or actions) seem mostly reserved for Russia, whilst 1.A2.N1 phraseologies (material possessions) seem mostly reserved for France. 1.A1 phraseologies (which emphasize the relationship between countries) and 1.A2.N2 (immaterial possessions) are the main categories for both countries, and possibly increase in frequency over the period, mainly at the expense of 1.B2 phraseologies (where COUNTRY complements a verb other than as its subject).

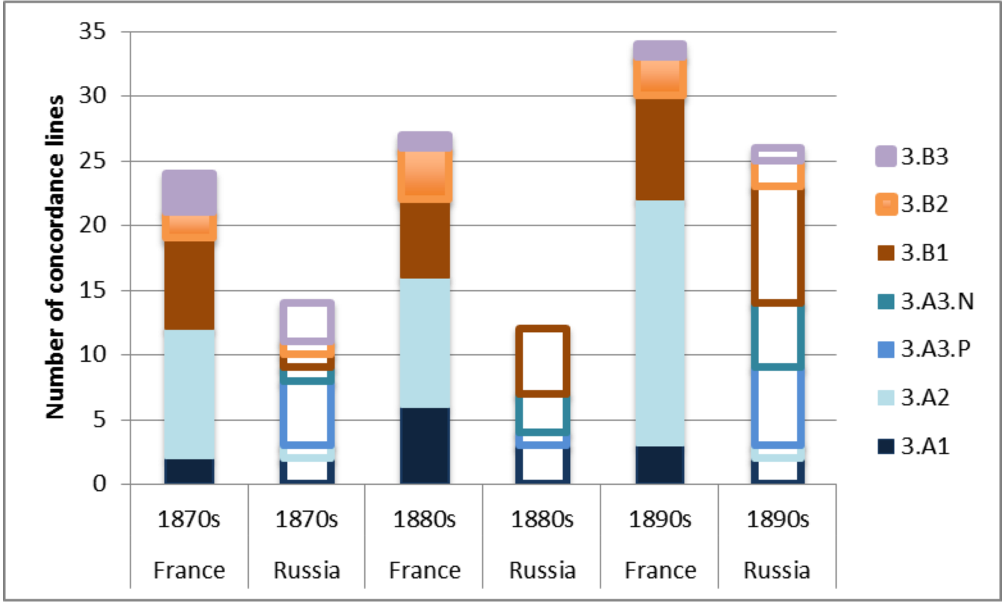
Figure 7.13 shows likewise the distribution of concordance lines across labels in the **affiliative category**. Again, more differences between the countries are obvious at this level than when looking at the level of over-arching categories. 2.A.P1 (named individual or group) is a rare label, which in the samples appears only for France in the 1890s. 2.A.P2 (individual(s) referred to by their position in the hierarchy) is the most important category for Russia in all decades, but for France only in the 1870s, after which it declines markedly. 2.A.P3 (broad group) is much more frequent for France than Russia, but does not appear in the samples for either country in the 1890s. Finally, 2.A.P4 (legal person/entity) occurs in all decades for both countries, but it is overall more important for France than for Russia.

Figure 7.13. Labels within the affiliative category, all newspapers together (France, filled; Russia, hollow)



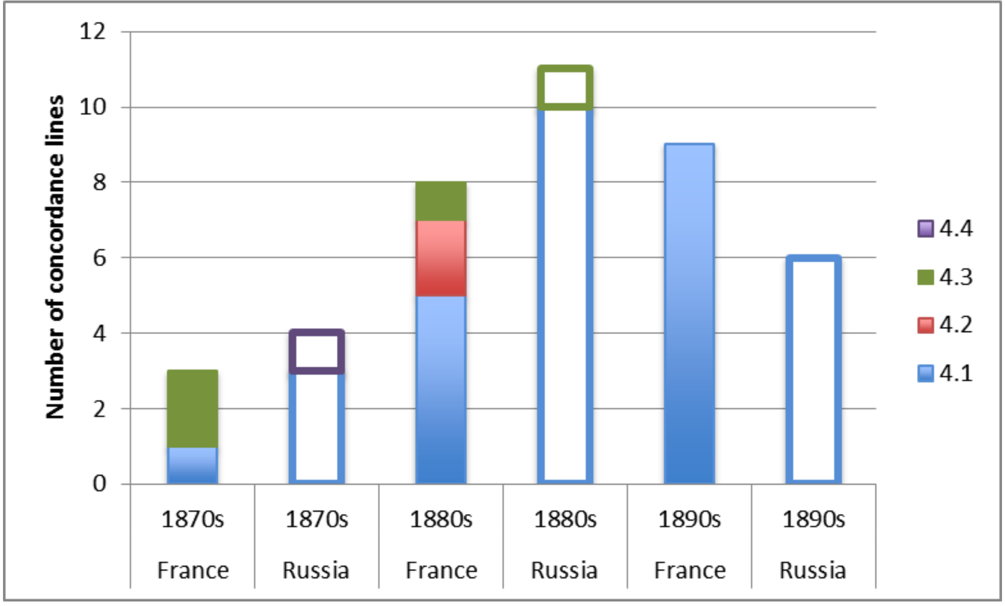
Within the **locational phraseologies** (see Figure 7.14), both countries have, in all decades, over half their phraseologies in the form of noun phrase modifiers, but the distribution across labels differs. For France, the most important phraseology is 3.A2 (geographical feature), and the only other noun phrase modifier phraseology to appear is 3.A1 (with a preposition emphasizing movement to or from the country). The phraseologies 3.A3.P and 3.A3.N (NOUN_PHRASE in FRANCE) only occur for France, although they account for at least 30% of the phraseologies of France in each decade considered. In the verb complementation phraseologies, there are differences between the countries too, chief of which is that the distribution of instances across labels is more stable for France over time than it is for Russia. All verb complementation phraseologies occur for both France and Russia (but not in every decade).

Figure 7.14. Labels within the locational category, all newspapers together (France, filled; Russia, hollow)



Finally, for the **specialized** phraseologies (see Figure 7.15), France seems to present more different cases than Russia does, though the figures are too small to be sure.

Figure 7.15. Labels within the specialized category, all newspapers together (France, filled; Russia, hollow)

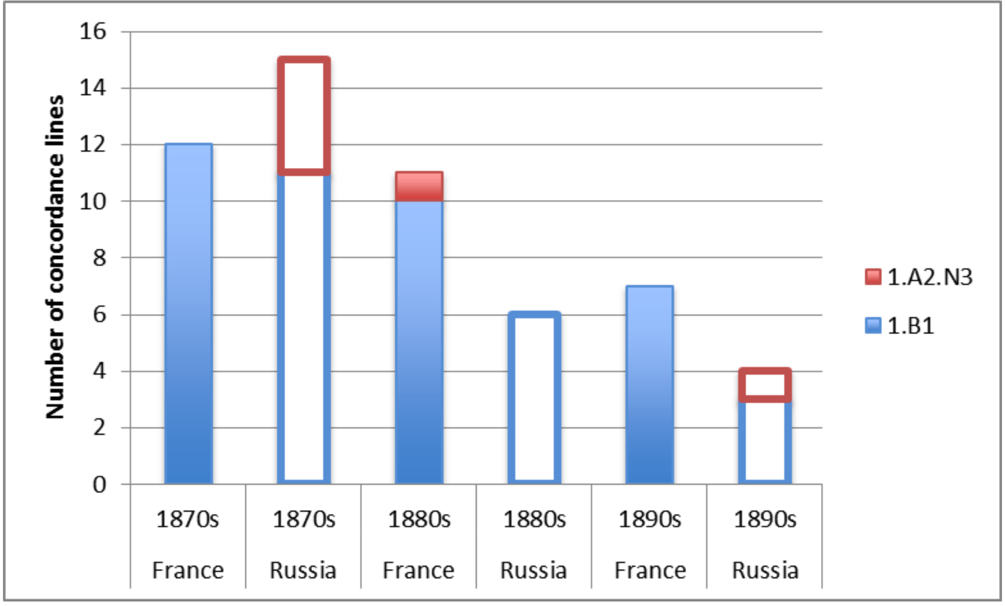


Beyond these general observations, more specific questions can be asked at this level. One interesting question is, for example, 'given that the countries are discussed with personifying phraseologies – which treat them as animates capable of doing or being done unto – to what extent are they attributed agency, and how is this agency realized in the texts?'. A starting-point for this investigation is simply to compare the frequency of cases where the country's agentivity within actions is expressed directly in the grammar (i.e. where the subject is the doer and the action is expressed in the main verb), which corresponds to label 1.B1 – to cases where the actions are expressed indirectly in the grammar through nominalization (i.e. the action is expressed by a noun, in these cases the doer is not always expressed) which corresponds to label 1.A2.N3[10]; in other words, comparing cases such as 'Russia demands' to cases such as 'the demands of Russia'. Arguably, the latter case, by its grammatically indirect expression of agency, subtly downplays the agency of the country.

Figure 7.16 shows the comparison of just those two phraseologies for Russia and France across the three newspapers. Although the numbers are very small, the figure suggests that the actions attributed to France and Russia through those phraseologies diminish for both countries over the time-period. Further, overall, in each decade, fewer actions are attributed through direct means, and more through indirect means, to Russia than to France. It would be fascinating to explore this in more depth, but space is lacking. In such an exploration, additional questions would rapidly arise, such as 'what kinds of actions are the countries portrayed as doing' and also, 'what is portrayed as being done to the countries', etc. which would involve both taking into account further labels, and constantly referring back to the source material.

[10] Though note that not all 1.A2.N3 represent actions *done by* the COUNTRY; some 1.A2.N3 will represent actions which the COUNTRY is *subjected to*. Distinguishing them is done by perusing the concordance lines.

Figure 7.16. Agency of France and Russia (across all three newspapers): comparing COUNTRY PREDICATE phraseologies to nominalization phraseologies



An analysis at this level helps draw out broad similarities and differences in the discursive strategies used to represent France and Russia in these newspapers. Most of all, however, they help identify interesting questions which then need to be explored further, to a great extent by looking back more extensively at the source material. One way to proceed with such an analysis is to focus on one or several labels which have been identified as interesting at this stage; for each such label, a greater number (ideally all) of the concordance lines in the source material can then be considered systematically. In the next section, I illustrate this approach.

7.5.2.2 A subset of categories in the whole data

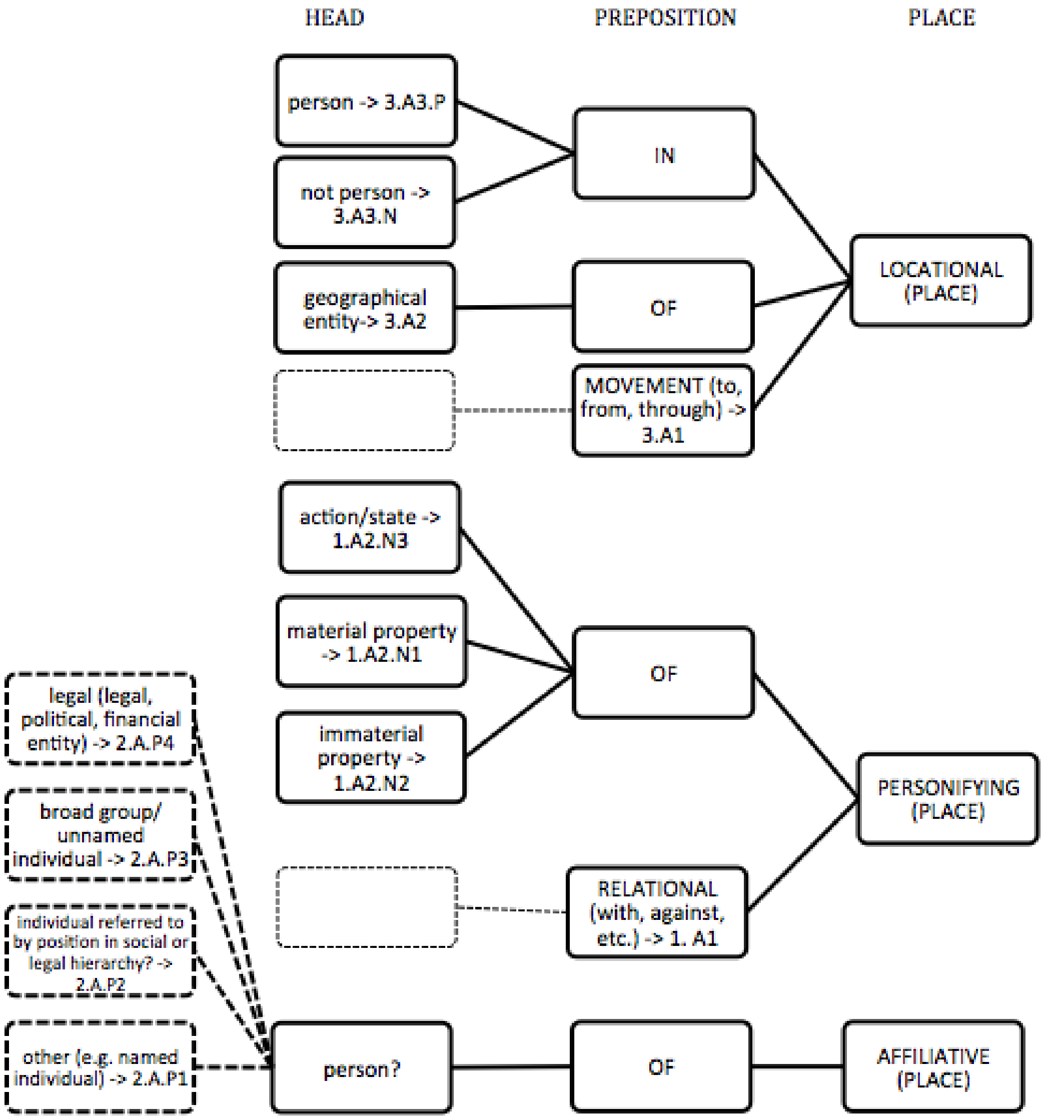
Not all labels lend themselves equally to analysis in the whole data. In particular, the verb complementation labels (type B) tend to encompass a range of phraseologies which are very diverse in their realizations. This makes it very difficult to devise a search query which can capture a significant number of such instances without simultaneously retrieving many irrelevant examples. In contrast, the noun phrase modifier phraseologies (type A) are easier to locate in the corpus as a whole since their linguistic expression is less diverse. In particular, in the samples, a large number of NOUN_PHRASE of COUNTRY phraseologies (hereafter referred to as ‘OF-phraseologies’) were encountered, which were, depending on the semantic nature of the NOUN_PHRASE, classified as labels 1.A2.N1, 1.A2.N2, 1.A2.N3, 2.A.P1, 2.A.P2, 2.A.P3, 2.A.P4 or 3.A2; see Figure 7.17 for a summary of the decision-making process for assigning labels to noun phrase modifier phraseologies.

Figure 7.17. Labelling phraseologies where COUNTRY is a noun phrase modifier

Three questions to determine the label for such phraseologies:

1. Which preposition is it?
2. (if needed) What is the semantic nature of the head?
3. (rare) In more detail, what is the semantic nature of the head?

Labels in relation to preposition and meaning of preposition and/or noun phrase:


HEAD
PREPOSITION
PLACE
person -> 3.A3.P
not person -> 3.A3.N
geographical entity-> 3.A2
IN
OF
MOVEMENT (to, from, through) -> 3.A1
LOCATIONAL (PLACE)
action/state -> 1.A2.N3
material property -> 1.A2.N1
immaterial property -> 1.A2.N2
OF
RELATIONAL (with, against, etc.) -> 1. A1
PERSONIFYING (PLACE)
legal (legal, political, financial entity) -> 2.A.P4
broad group/ unnamed individual -> 2.A.P3
individual referred to by position in social or legal hierarchy? -> 2.A.P2
other (e.g. named individual) -> 2.A.P1
person?
OF
AFFILIATIVE (PLACE)

In this section, I focus on the NOUN of COUNTRY phraseologies as an illustration. The following steps are involved in this approach:

1. Formulate a search query which achieves a satisfactory precision/recall balance for the phraseology/ies of interest.
2. Run the query to retrieve the relevant concordance lines
3. Proceed with analysing the concordance lines.

Step 3 will involve multiple steps, but their number and nature depend on the aims and interest of the researcher. Inevitably, however, it will involve a certain amount of categorization (which probably involves a preliminary exploratory phase to develop categories which are relevant and useful) for a broad overview of the data, following which more specific questions can be formulated, for which the answer will often involve focus on a subset of the data, which can be systematically isolated with the help of such a question. Below, my step 3 involves the following steps:

1. Focusing on the head nouns[11] in OF-phraseologies, classify them using the labels presented in the previous section (e.g. 1.A2.N3, 2.A.P3, etc.)
2. 'Zooming in' to consider each label in turn, classify the head nouns further into categories (e.g. in the 1.A2.N1 category, 'actual objects', 'financial resources', 'military resources' and 'other').
3. Focusing on head nouns in the 'moral property' category (identified in step 2) of the 1.A2.N2 label (identified in step 1), group concordance lines according to questions which emerge from reading them, such as whether the properties carry 'good' or 'bad' evaluations, and whether the properties are exercised 'by' or 'towards' the countries.

[11] A *head noun* is a noun which is part of a *phrase* (a sequence of grammatically related words), and which is qualified by other words in the phrase, e.g. in the phrase 'the king of Russia', king is the head noun, and *of Russia* qualifies (i.e. adds to the meaning of) the word *king*. In OF-phraseologies, the head noun always comes before the word *of*.

Each of these levels requires a back-and-forth between looking at examples in context (concordance lines and whatever additional context is required for the current level of analysis) and looking at more abstract lists of results (in this case usually lists of head nouns). The analysis at each of these levels may reinforce or challenge the analysis and categorization of the previous level[12].

It is a painful realization for the perfectionist in every researcher that no categorization is final or can account for every subtlety in the data. The purpose of a system of categories is to help identify broad trends in complex data. Accounting for more detail is often done at the expense of clarity, and sometimes of usefulness and practicality. A pragmatic compromise needs to be achieved between workability, usefulness and accuracy in categorization, whilst bearing in mind that the balance between these will shift constantly as the research proceeds. The compass, as always, remains the research question. The difficulty (stemming, paradoxically, from the advantage) of the approaches outlined in this chapter is that they lend themselves well to exploratory research which allows research questions to emerge from the data, in cases where the researcher is unable, or does not want, to formulate research questions before looking at the data. Hence, bereft of a north-point, the researcher can easily get bogged down in endless or unwieldy categorization.

Scholars from the Humanities newly encountering corpus linguistics sometimes show a measure of reluctance to adopt any form of computer-based text analysis; often, this reluctance stems from a misconception that *all* such methods aim to answer research questions at the press of a button. In part, this misconception originates from a confusion between corpus linguistics and text mining, the goals of which were distinguished in section 2.2.1. The discussion in this section should be sufficient to dispel any such misconceptions by illustrating that corpus linguistic approaches are entirely suited to, and often entail, research involving long

[12] This is in fact precisely how the 'moral properties' category of head noun within 1.A2.N2 came to my attention, as when I first encountered these examples, I hesitated between labelling them 1.A2.N2 (immaterial possessions) or 1.A2.N3 (properties or actions). I still think this is debatable.

explorations and perusal of the source material; this is not a methodology designed to produce press-of-a-button answers.

Having described the general approach of this section, let us now proceed to briefly illustrate this analysis. Retrieving NOUN_PHRASE of COUNTRY phraseologies can be done by using a CQPweb query such as (in CQP syntax):

[pos= "N.*"] [word = "of" %c] [word = "france" %c]

which simply allow for any word tagged as noun, followed by the word 'of' (case-insensitive[13]), followed by the word 'france' (case-insensitive). This will not retrieve cases where the noun phrase consists of more than a single noun, nor cases where COUNTRY is preceded by an adjective. As ever, errors will also preclude retrieval: OCR errors, and POS tagging errors (which may be more frequent due to OCR errors). Nevertheless, the query ensures a high degree of precision, and the recall should be reasonably high based on the examples encountered in the samples considered in the previous section.

These queries can also be restricted in various ways: per year, per newspaper genre (if this layer of annotation is present, as in our case), and so on. In this section, I will focus, simply for illustrative purposes, only on phraseologies present in news articles. The queries which I run as a first step, then, are the following two queries, which need to be run once per newspaper, and which include a restriction to news articles:

[pos="N.*"] [word="of" %c] [word="france" %c] :: match.article_category="News"

[pos="N.*"] [word="of" %c] [word="russia" %c] :: match.article_category="News"

How much of the total data does this constitute? Table 7.11 shows the number of times 'France' and 'Russia' occur in News articles and what proportion those occurrences constitute of the occurrences across all genres per newspaper. It also shows the number of instances of OF-phraseologies in News articles, and what proportion those instances constitute of occurrences

[13] I.e. whether in upper-case, lower-case, or a combination thereof.

of COUNTRY in News articles for that newspaper. Since the number of times the two countries are mentioned is very different across newspapers and also between the countries, the number of instances of OF-phraseologies also vary greatly. The proportion of total mentions of the countries in News articles in each newspaper, however, is remarkably constant – around 20% in all newspapers for both countries, though ERLN presents a slightly greater proportion, and PMGZ a slightly smaller proportion, for both countries, than RDNP.

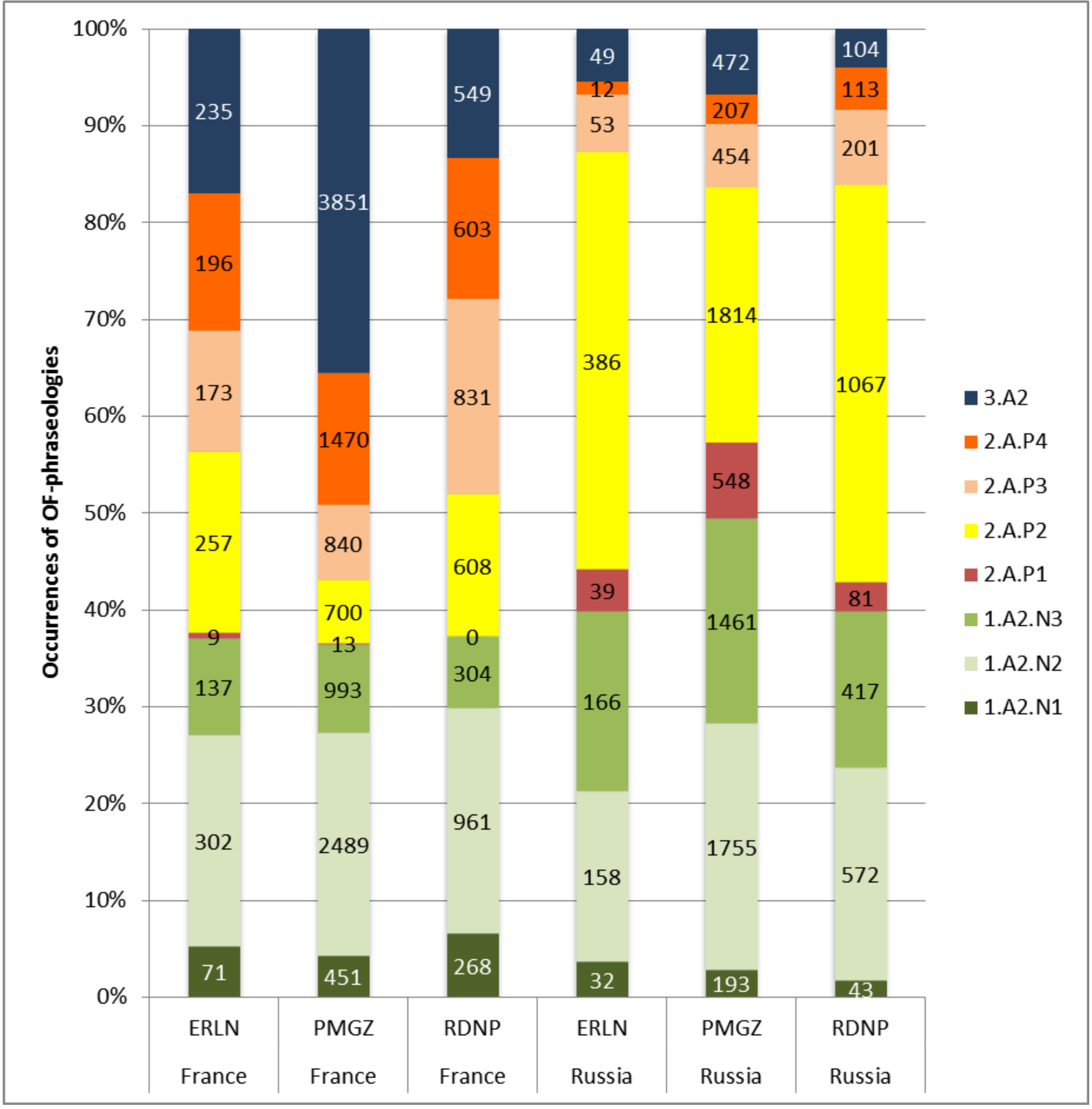
Table 7.11. Proportion of mentions of France and Russia in News articles

Label		FRANCE	RUSSIA
ERLN	Raw hits of COUNTRY in NEWS	8866	4340
	(% hits of COUNTRY in all genres)	(76%)	(86%)
	Raw hits of NP of COUNTRY in NEWS	1915	1266
	(% hits of COUNTRY in NEWS)	(22%)	(29%)
PMGZ	Raw hits of COUNTRY in NEWS	79923	52565
	(% hits of COUNTRY in all genres)	(74%)	(85%)
	Raw hits of NP of COUNTRY in NEWS	15384	9619
	(% hits of COUNTRY in NEWS)	(19%)	(18%)
RDNP	Raw hits of COUNTRY in NEWS	30370	16640
	(% hits of COUNTRY in all genres)	(83%)	(88%)
	Raw hits of NP of COUNTRY in NEWS	6202	3685
	(% hits of COUNTRY in NEWS)	(20%)	(22%)

How are OF-phraseologies distributed across the different labels that share this structure? Figure 7.18 shows figures for the 250 most common nouns per country and newspaper, excluding OCR errors, occurring in the NOUN_PHRASE slot; these account for between 66% and 72% of the total number of OF-phraseologies per country per newspaper. It is interesting that the OF-phraseologies cover the full range of possible labels for both countries, with the exception of the 2.A.P1 label (named individual), which occurs for France only very infrequently in ERLN and not at all in the two other newspapers (among the 250 most frequent single nouns considered here).

As was found above, although overall the OF-phraseologies account for a similar proportion of the instances of France and Russia in News articles in their respective newspapers, there are differences between the two countries in terms of the nature of these OF-phraseologies. For Russia, in all three newspapers, the most frequent category is 2.A.P2 (an individual referred to by their position in the hierarchy) – of which most instances are the phrase 'the emperor of Russia' – and this is followed by 1.A2.N2 (immaterial possessions), then 1.A2.N3 (properties/actions) in all three papers. In contrast, the phraseologies are somewhat more evenly spread out for France, although the two most prominent categories in all three papers are 3.A2 (geographical features) – which is relatively infrequent for Russia – and 1.A2.N2 (immaterial possessions).

Figure 7.18 OF-phraseologies across labels; distribution of raw instances of 250 most frequent single nouns



Again, at this level, interesting observations can be made, but they raise more questions than answers, and they constitute mostly exploratory steps which require further investigations to yield meaningful interpretations. Looking at the specific nouns occurring as head-nouns in these phraseologies, for example, is fascinating, and reveals both that there are broad similarities between the language used to discuss both countries, but also differences. For example, in the 1.A2.N1 category (material possessions), an infrequent category (among the top 250 single nouns considered) in all newspapers for both France and Russia, we find many references to France's financial resources, with terms such as 'resources', 'finances', 'debt', 'credit', 'commerce', share', trade', 'levies', 'industry', 'wealth', 'expenditure', 'expense', and 'revenue'. References to the financial resources of Russia are also made, but fewer of the 250 most frequent single nouns pertain to them – 'resources', 'trade', 'commerce', 'finances', 'debt', and 'pay'. It is interesting that some words are used in common for both countries – 'resources', 'finances', 'debt', 'commerce', trade', and moreover frequently so (since these words are among the 250 most frequently occurring nouns in the type of phraseologies considered here), although some of these words are more strongly associated with one country than with the other. An interesting avenue of analysis from this point would be to use contrastive collocation (see section 2.3.2.4 on concordance-corpora), to identify words used in common, and words used significantly more often with one country than with the other; there is insufficient space to pursue this here.

Table 7.12. Broad categories and examples of nouns occurring as head of OF-phraseologies

Label		FRANCE	RUSSIA
1.A2.N1 (material possession)	**Actual object**	'wines'	(none)
	Military resource	'fleets', 'armaments'	'armaments'
	Financial resource	'levies'	'commerce'
	Other	'notes'	'tools'
1.A2.N2 (immaterial possession)	**Moral property**	'honour', 'jealousy'	'honour', 'jealousy'
	Political object	'policy'	'policy'
	Planning/objective	'ambition', 'interests'	'ambition', 'interests'
	State/condition	'difficulties', 'destiny', 'peace'	'situation', 'welfare'
	Attribute of the nation	'flag', 'Eagle', 'spirit'	'history', 'flag'
	Political characteristic	'influence'	'influence'
	Art	'literature'	(none)
	Other	'sense'	'sense'
1.A2.N3 (property/action)	**Conduct**	'pretensions'	'pretensions'
	Demand/offer	'demand', 'proposal'	'assurance', 'threat'
	Planning	'designs', 'will'	'designs', 'schemes'
	Aid	'aid', 'protection'	'aid', 'protection'
	Action affecting territory	'invasion', 'dismemberment'	'conquest', 'encroachment'
	General reference to action	'action', 'efforts'	'activity', 'instigation'
	Change in political state	'ascent', 'defeat'	'progress', 'ascendancy'
	Relation with other countries	'alliance, 'exclusion'	'alliance', 'subjugation'
	Movement	(none)	'march', 'approach'
	Other	'taxation'	'appearance'
2.A.P1 (named individual)	**Ruler**	'Philip', 'Louis'	'Alexander', 'Catherine'
	Other	'Lily'	'Marie', 'Olga', 'Michael'
2.A.P2 (person referred to by their position in the hierarchy)	**Ruler**	'King', 'Sovereign', 'Dauphin'	'emperor', 'tzar', 'despot'
	Other political/diplomatic position	'Minister', 'Consul'	'ambassadors, 'Minister'
	Military position	'Marshal', 'Admiral'	(none)
	Other	'vassal', 'dynasty'	'family'
2.A.P3 (broad group or unnamed individual)	**Referred to by their economic situation/occupation**	'press', 'clergy', 'peasants'	'press', 'statesmen', 'serfs'
	Referred to by their social category	'children', 'women', 'Conservatives', 'nobles', 'Catholics'	'youth', 'autocrat', 'Jews', 'Nihilists'
	Military-related designation	'armies', 'squadrons', 'troops'	'forces', 'fleets', 'legions'
	Relational term	'enemies', 'allies'	'friends', 'opponent', 'vassal'
	General term	'population', 'natives', 'subjects'	'inhabitants', 'hordes'
2.A.P4 (legal entity)	**Commercial**	'Bank', 'Company'	'Bank'
	Political-legal	'government', 'institutions'	'Government', 'Court'
	Political-symbolic	'throne', 'sceptre'	'thrones'
	Religious	'church'	'Church'
	Educational	'University', 'schools'	(none)
	Other	'protectorate', 'republics'	'empire', 'protectorate'
3.A2 (geographical feature)	**Part/whole**	'parts', 'half', 'size'	'parts', 'portions'
	Region	'districts', 'department', 'interior'	'districts', 'provinces'
	Geographical feature	'shores', 'fields', 'frontier'	'ports', 'steppes', 'frontier'
	Urbanization	'towns', 'capital'	'capital'
	Orientation	'North', 'Western', 'north-west'	'north', 'south,
	Building	'houses', 'Theatres'	(none)
	Other	'map'	'map'

Table 7.12 presents some broad categories and examples of nouns occurring for each country within each label. This table also shows the extent to which this level of analysis is qualitative and subjective. Whereas the classification of labels presented in the previous section could be expected to yield similar results when used by different researchers, the classification

in Table 7.12 is to a greater extent in the eye of the beholder. Here, maybe more than elsewhere, the shape and form of the analysis, as well as its potential to yield insightful interpretations, will depend on the background of the researcher. Hence, for these methods to serve to further our understanding of history, they should be used ideally, either in collaboration with, or by historians themselves.

To illustrate a more in-depth analysis, I will consider in depth one sub-category, 'moral properties', within the 1.A2.N2 category. This category accounts for around 10% of the 250 most frequently occurring nouns in OF-phraseologies (36 out of 391[14] for France, 35 out of 347 for Russia) in the news articles of the three newspapers. I have chosen it because it intrigued me, and because it is common to both countries. Two questions which are relevant specifically to phraseologies that express such moral properties are: 'do these properties have a clear moral charge – i.e. are they considered good or bad in context?', and 'are these properties exercised by or towards COUNTRY?'.

Properties generally considered 'good' which are mentioned for both countries include 'hono(u)r', 'dignity', 'glory', 'pride', 'grandeur', 'confidence', 'greatness', 'prestige', 'strength', 'faith', 'authority', 'sincerity', 'might'. Properties generally considered 'bad' which are mentioned for both countries include 'jealousy', 'hostility', 'hatred', 'vanity', 'fear', 'distrust', 'tyranny', 'weakness', 'duplicity', 'dread', 'suspicion', 'despotism'. Providing examples of these used in context generally requires lengthy quotations, as short excerpts tend to be hard to interpret due to the complexity of the language used and the circumstances described. I present here two for each country, one 'good' and one 'bad'.

> "The Times observes that Count Bismarck's object and Germany's first necessity were that the Fatherland, or at least that part of it which lies north of the Main, should present itself to foreign Powers as a strongly organized and compact country. The **jealousy of France** succeeded where, perhaps, all the strong sense and stubborn will of Bismarck, of Von Roon, and of Moltke might have

[14] The total number of nouns considered amounts to less than 750 (the number expected if the 250 most frequently occurring nouns in the head position of the OF-phraseologies in each newspaper were entirely different) since there is a large amount of overlap between the lists from each newspaper.

failed, and the military budget, which extends to the whole Confederacy the burdens of the Prussian service up to the 3Ist of December, 187i, has been accepted." (PMGZ, 15/04/1867)

"France has compelled Portugal to succumb upon the question of the Charles et Georges, the French slaver seized by the Portuguese vessel. The lesser power concedes of the demands of the greater, but declares that she does so only on account of the superior force of the plaintiff. (...) it certainly appears that France is strong and wrong. And her ready appeal to menaces, her sending her ships-of-war into the Tagus to enforce her demands (...) do not look as if she thought she had a good case. But a bit of violence is always acceptable to the people under whose flag it is committed, and the Emperor of the French knows his subjects perfectly well. His papers tell them that the **honour of France** has been triumphantly sustained – and are believed." (PMGZ 31/10/1858)

"Both Prince Orloff and General Ignatieff continue to protest that Russia desires peace, the chief reason being an intense distrust of Germany, for the conduct of Prince Bismarck has awakened serious misgivings in the mind of the Cabinet of St. Petersburg. Every time that Russia has taken a step in advance she has found the Cabinet of Berlin paying court to that of Vienna, and every time she has taken a step to the rear she has been sarcastically reproached for being alarmed at imaginary dangers, and not taking advantage of an opportunity not likely to present itself again. This policy has alarmed Prince Gortschakoff and his diplomatists, who think that Russia might be a match for Austria and Turkey if Germany would preserve a friendly neutrality, but not otherwise. Negotiations between Russia and this country have been carried on for some time past in the most friendly tone, a fact which has given umbrage at Berlin. (...) If war be averted now in consequence of the mutual **distrust of Russia** and Germany, it is quite possible that it may be settled for next year or the year after; that Russia will attack Turkey at the same time that Germany attacks France. In that manner each of the two great Northern Powers would be protected against the treason of the other." (PMGZ 13/03/1877)

"Russia has asserted that a regard for her dignity precludes her from acceding to the terms proposed by the Allies on the third point. But the **dignity of Russia** can not require that she should keep up in time of peace, and on the immediate threshold of her weaker neighbour, a force wholly unnecessary for the purposes of self-defence, but enabling her at the shortest notice to subvert the independence of that neighbour, and to change the territorial distribution of Europe. Yet such is the position which Russia has maintained in the Black Sea, and which she has even now publicly avowed her determination not to renounce." (PMGZ 18/01/1871)

These examples show how these moral properties of the countries are used in very interesting ways in the context of discussions about the diplomatic and military relationships between countries, particularly in attempts to monitor shifts in the alliances constituting the European balance of power (see section 7.2), and to predict military developments. These discussions, in which the moral dimension of the nations is central, seem strongly relevant to the thesis which was expressed specifically about France – but appears equally valid about Russia – that the country 'was used in public discourse as a moral counterpart against which Britain could define itself in self-flattering ways' (section 7.2).

Asking whether these moral properties are 'good' or 'bad' in context may seem trivial, but the examples above show how difficult answering this question can turn out to be. Passages including such discussions are often very nuanced and report on various perspectives. In the case of the 'honour' of France, above, for example, it is clear that, for France, this 'honour' is an important thing which merits preservation, but it is also clear that the writer is casting doubt on its legitimacy (or, to be more accurate, on the morality of its usage in that situation). In my classification of 'good' and 'bad' properties, then, I have normally adopted the most obvious definition; in this case, the 'honour' of France is a 'good' property because it is desirable for (and by) France.

Keeping these caveats in mind, Table 7.13 provides a tentative quantification of the distribution between 'good' and 'bad' properties. 'Good' properties seem always to be exercised by the country which is modifying the noun (e.g. the 'honour' in 'the honour of France' is wholly France's) whereas 'bad' properties are sometimes exercised by the country and sometimes directed towards it; a quantification of this is also provided in Table 7.13. (These figures are obtained by perusing the concordances in which each property is mentioned.)

Table 7.13. Moral properties of France and Russia in news articles of ERLN, PMGZ and RDNP: good or bad? By or towards COUNTRY?

	FRANCE			RUSSIA		
	ERLN	PMGZ	RDNP	ERLN	PMGZ	RDNP
GOOD	56	288	142	12	174	59
BAD (by)	6	37	7	5	46	22
BAD (towards)	3	39	16	3	88	32
RATIO (good/bad)	6.22	3.78	6.17	1.5	1.29	1.09

Table 7.13 reveals that, again, there are, overall, more similarities than differences between the representations of the two countries. On the similarity side, both countries are discussed using a similar language of what is called here 'moral properties'. These properties, as was illustrated in examples above, are used as part of sophisticated discussions of complex diplomatic relationships between countries with the looming threat of military repercussions. The way in which they are used in discourse is particularly interesting, as these properties are discussed as if they possessed a degree of agency, or acted as forces pushing animates into various forms of actions and behaviours. As Table 7.13 shows, these moral properties are, on the whole, overwhelmingly often 'good' properties, although the balance between 'good' and 'bad' (as shown in the last line of Table 7.13) is more favourable to France than it is to Russia, and varies across the newspapers (with, strikingly, the balance being much worse for France in PMGZ than in the two other papers considered). Nevertheless, for both countries, 'bad' properties are also mentioned in each newspaper – though these are especially present in PMGZ, much less so in the two other papers, though in RDNP still more so than ERLN, and this for both countries. Still on the similarity side, it is also interesting that many of these properties are the same for both countries, including the list provided above.

On the difference side, the balance between 'good' and 'bad' properties is more favourable to France than it is to Russia. Overall, then, this particular strand of discourse suggests a less favourable position on Russia than on France. A difference is observable also in terms of agency in this strand of discourse, at least in PMGZ. (The mentions are too few in ERLN and RDNP to be worth elaborating on.) Of the 'bad' properties mentioned in PMGZ for France, reference is virtually equally often made to 'by' than to 'towards'. For Russia, in contrast, references to 'towards' are almost twice as often as to 'by'.

7.5.3 SUMMARY OF FINDINGS IN THIS SECTION

Bearing in mind that only data from news articles was discussed in this section, we can say that, in broad terms, the countries are discursively represented with more similarities than differences. The overall properties of the language used to refer to them is hence similar: the phraseologies which surround mentions of the countries can be described using the same set of labels and categories; personifying and locational phraseologies are the most frequent over-arching categories, and together account for over half the mentions of both the countries in each newspaper; and so on. Some similarities can also be observed at greater levels of detail. In the personifying categories, for example, 1.A1 phraseologies (which emphasize the relationship between countries) and 1.A2.N2 phraseologies (which involve immaterial possessions) are the most frequent phraseologies for both countries. Within the 1.A2.N2 phraseologies, it was also observed that discussions of the countries in moral terms are an important feature of some the discourse surrounding these countries, with both broad similarities between the countries, and differences between them at the level of further detail. Indeed, OF-phraseologies account for around 20% of mentions of the countries in all newspapers; in many cases, the exact same nouns are used in the NOUN_PHRASE slot for both countries. This includes many of the moral properties identified. For both countries, these moral properties can carry both 'good' and 'bad' evaluations, though the 'good' are mentioned more often than the 'bad' for both countries in all newspapers – although the balance is less favourable for both countries in PMGZ, which has a particularly high number of instances of this type of phraseology.

Beyond these similarities, however, differences between the countries can be observed, often at the more detailed level of analysis. Often, although a broad trend is observed for both countries, quantitative differences can be observed on closer inspection. One broad difference identified is that France has a tendency to be discussed in more diverse ways than Russia, both in terms of the greater number of categories which tend to be expressed around France than Russia, and in terms of the more even distribution of mentions of France across the different categories than Russia. This is in line with *h6* (which states that since France is geographically closer to Britain than Russia, less may be known about Russia than France, leading to more prejudice towards Russia than towards France and a more diversified representation of France than Russia, see section 7.2).

Some further differences include the following. In personifying phraseologies, Russia presents more 1.A2.N3 (properties/actions) phraseologies and fewer 1.A2.N1 (material possessions) phraseologies than France does, which suggests a more abstract discussion of Russia overall than France, which is, perhaps, in line with *h6.* Russia presents more affiliative phraseologies overall; in terms of distribution, Russia presents more 2.A.P2 (position in hierarchy) and fewer 2.A.P3 (broad group) and 2.A.P4 (legal entity) than France does. The observations regarding 2.A.P2 and 2.A.P4 could be taken to indicate that Russia is associated with more 'political gossip' than France, and that, again, these discussions tend to be less concrete. This is, perhaps, in line with Hughes's (2015) observation (see section 7.2) of the exoticism of Russia for the British in this period. That France presents more 2.A.P3 phraseologies is perhaps more surprising in light of *h6,* although it is also, in a way, congruent with it. It is surprising since it seems to indicate a higher level of stereotyping regarding France than Russia, but on the other hand it also reveals more discussion of the goings-on within France, which is what is expected under *h6.* In locational phraseologies, France presents more 3.A2 (geographical features), 3.A3.P and 3.A3.N (people/not people *in* France) phraseologies than does Russia, which is line with *h6*, as is the observation that France presents more specialized phraseologies overall than Russia.

My analysis also observed briefly that agency appeared to be expressed more indirectly for Russia than for France, which could well be in line with Parry's (2001) point that France seemed a more immediate threat to Britain than Russia did (see section 7.2). The differences identified within OF-phraseologies specifically were in line with those identified overall, with Russia presenting more 2.A.P2 and fewer 3.A2 phraseologies than France as well as, more specifically, fewer references, and less diverse references, to financial resources. Finally, the focus on moral properties revealed a balance of mentions of 'good' to 'bad' properties less favourable to Russia than France (although for both the overall balance was positive in all three newspapers). Here is perhaps some evidence for the thesis of 'Russophobia' (see section 7.2), though this would need to be explored in much more depth to be established with confidence; on the whole negative references to Russia have been much more evasive and subtle than might have been expected based on that thesis.

This approach, then, has proved useful for exploring representations of France and Russia in these newspapers. The conclusions derived therewith have been, on the whole, in line with expectations based on existing literature, although there have also been some surprises. Most of all, perhaps, the approach has highlighted questions which could prove interesting for further investigation.

I was unable, in this section, to do adequate justice to questions of differences across genres and over time. Nevertheless, it seems clear that this approach would allow such differences to be explored. Differences over time could not be explored satisfactorily here since, as was mentioned above, these require, for any degree of effectiveness, more than three points on a time-series.

Differences across newspapers were evident at various points in the analysis, and it was also clear that these are often larger than the differences in representations of the two countries within a single newspaper. This is interesting, since, given the difference in social locatedness of the newspapers, it suggests that there was, overall, little differentiation in how different countries were discussed in the c19th press. This means conclusions based on analysing the discourses surrounding a single country may be misleading, as they may emphasize as unique to that country discourses which were actually similar for most countries. It would also be very interesting, for further research, to investigate the extent to which similar frameworks to those used here to analyse the language surrounding countries in c19th newspapers would also be effective for analysing the language surrounding other types of place-names, such as cities, and also more varied texts, such as newspapers from other periods and perhaps other types of texts altogether.

7.6 Conclusion

In this chapter, in addition to an initial survey based on frequency counts, two approaches to answering the question 'what is said about these places?' have been explored. The two approaches have in common that they exploit the concept of 'collocation', allow for both quantitative and qualitative forms of analysis, focus on capturing the most frequent patterns, and are able to take parts-of-speech into account. The approaches differ in that one, the global approach, focuses on the re-occurrence of lexical patterns of co-occurrence, whereas the other, the sampling approach, focuses on syntactic and semantic patterns. Moreover, the global approach starts with the whole dataset and eventually looks at samples of text, whereas the sampling approach starts with samples of text and eventually (in some cases) works with the whole dataset. Both approaches are able to contribute to answering the question 'what is said about these places?', though both have limitations. Both approaches are particularly well-suited for exploratory research, as they are able to generate unexpected research questions which would ideally be followed up using a combination of corpus linguistic and more traditional historical methods.

The first approach, *the global approach*, consists in generating a list of collocates of the nodes in the whole dataset, then focusing on the strongest collocates, categorizing them by looking at all, if possible, or a random sample if not, of concordance lines in which the nodes co-occur with each collocate. The broad question answered is 'what do these patterns of co-occurrence tell us about the kinds of contexts in which the nodes are referred to?'. The approach proved feasible and scalable, though ill-adapted to robust and granular comparisons between newspapers, across genres and over time. It seems most helpful for initial, exploratory overviews of a whole dataset; its principal strength is its ability to accommodate large amounts of data, quickly identifying broad common patterns, and facilitating the analysis of these patterns both quantitatively and qualitatively.

The second approach, *the sampling approach*, consists of two phases. The first phase involves reading through successive random samples of concordance lines for the nodes, in order to elaborate a framework for analysis. This is done by creating descriptions of the phraseological context of the nodes. The second phase exploits this framework either by simply analysing more random samples, or by focusing on one or more of the categories constituting the framework, locating concordance lines fitting these categories from across the whole dataset. This approach is more time-consuming than the previous approach. It is scalable in theory, though in practice some categories will be easier to locate in the entire dataset than others, thus making some patterns more easily analysable at scale than others. It is also suited to exploratory research, but is moreover also able to accommodate more rigorous and in-depth, if time-consuming, research.

In terms of findings related to France and Russia, let us consider what was learned about the hypotheses formulated in section 7.2. The first two hypotheses stated that since France and Russia are rival military powers to Britain, we might expect discussions of the countries to pertain to the potential or ongoing military involvement of the countries. *h1* suggested that an important proportion of the discussion surrounding the countries would relate to military matters, whilst *h2* suggested that the peaks in mentions of countries would coincide with the onset or imminent onset of military conflicts involving those countries.

The global approach, in addition to looking at overall frequencies, was able to address *h1*. The findings were overwhelmingly *not* in line with this hypothesis. Looking at the frequency of co-occurrence of the countries around 'war' and WAR-related words revealed that the countries never occurred with WAR-related words more than 27% of the time in any given year of either newspaper, a proportion which can certainly *not* be considered overwhelming. The diversity of contexts of mentions of the countries uncovered using the global approach suggested that *h1* was at best reductive and potentially misleading, with the relationship between Britain and France in particular having been revealed as a complex, multi-faceted one which could not be reduced to a simple military rivalry. The sampling approach seemed least suited to exploring this hypothesis, although the findings betray a diversity of representations of the countries which undermines *h1*.

The approaches did not allow either full confirmation or rejection of *h2*. The presence of patterns of peaks and troughs in the frequency of mentions of the countries over time and of military/diplomatic discussions surrounding the countries affirm the pertinence of this hypothesis, but more research would be required to confirm or reject it.

The third hypothesis stated that since France and Russia are rival military powers to Britain, we might expect a disproportionate number of mentions in news articles. Looking at overall frequencies was the approach most suited to exploring this hypothesis. Although it was indeed found that the greatest proportion of mentions occurred in news articles, this proportion remained relatively small, and the countries were also mentioned in all other article genres considered. This suggests that although *h3* is not entirely incorrect, it is only weakly supported by the data.

The fourth hypothesis stated that since France and Russia are rival military powers to Britain, we might expect a degree of negative bias towards both countries. On the other hand, we might also expect expressions of respect, admiration or jealousy. Both approaches were able to contribute findings relevant to this hypothesis, though neither approach appeared ideally suited to assessing it in full. Both approaches suggested that the hypothesis is not incorrect, but perhaps simplistic. Although a degree of negative bias can be found with regards to Russia, the expression of negative bias towards France was relatively minimal. Since both countries were military rivals to Britain, it suggests that there are other factors at play in determining the cause of negative bias towards the country. The global approach showed that some of the strong collocates of Russia betrayed anti-Russian angles in all three newspapers, but not so for France. The sampling approach showed that although both countries are associated with moral properties which are both good and bad, more negatively-evaluated moral properties were associated with Russia than with France. These findings suggest that this is an area which would merit further investigation.

The last two hypotheses stated that since France is geographically closer to Britain than Russia, we might expect more ties of various nature with France than with Russia, leading to more mentions of France than Russia (*h5*), mentions of France in more diverse contexts than Russia (*h5*), and mentions of France less stereotypical than Russia (*h6*). Both approaches were able to address these hypotheses, and the findings were overwhelmingly in support of them. In all three newspapers, France is mentioned more often than Russia; a greater proportion of mentions of Russia occur near WAR-related words; a greater proportion of mentions of Russia occur in news articles and a smaller proportion in commerce article; more French cities are mentioned than Russian cities; France is referred to proportionally less often by its country-name than Russia; strong evidence of a multi-faceted relationship including cultural, commercial, and personal ties between Britain and France, but not between Russia and France, were uncovered using the global approach; and the sampling approach showed that France was referred to in more diverse ways than Russia within news articles, with further evidence of more detailed references to the geographical interior of France than Russia.

It is interesting that, in the end, although there was a degree of overlapping in the findings elicited with each approach, some of the findings were elicited only with one of the approaches. This suggests that both approaches can be used in isolation, but may also productively be used together to generate a richer understanding of the data, an argument reminiscent of the argument developed by Baker and Egbert (2016) in favour of methodological 'triangulation'.

PART 4: CONCLUSION

8 CONCLUSION

8.1 INTRODUCTION

In chapter 1, I noted that this thesis pursues two major research aims. The first pertains to broadening understanding of OCR errors and their impact on corpus linguistic methods, and the second to establishing and evaluating a methodology for investigating spatial patterns in large amounts of text. Sections 8.2.1 and 8.2.2 summarise and discuss my findings related to the first and second aims respectively. Section 8.3 evaluates them; section 8.4 sets out directions for future research; and section 8.5 provides some final remarks.

8.2 SUMMARY OF FINDINGS AND DISCUSSION

8.2.1 PERTAINING TO OCR ERRORS

This section summarises my findings on OCR errors and their impact on corpus linguistic methods, and specifically two collocation statistics: Mutual Information (MI), a commonly-used effect size statistic, and Log-Likelihood (LL), a commonly-used significance statistic.

8.2.1.1 What is, in theory, the impact of OCR errors on the statistics of collocation?

MI and LL are vulnerable to OCR errors in the first place because OCR errors impact the form (i.e. spelling), presence, and number of word-tokens, which in turn affects wordcounts, which form the basis for the calculation of MI and LL. If OCR errors are assumed to be distributed homogenously throughout a corpus, then they will not have any problematic effect on MI and LL (section 3.3.2). Indeed, if that assumption holds, the corpus which results once the OCR errors are excluded is simply a random sample of the original (error-free) corpus. In that hypothetical smaller corpus, MI will remain unchanged whilst LL will be smaller: MI is, by nature, not altered by such sample reduction, and LL is *meant* to be responsive to corpus size since it is a measure of the amount of evidence available for a given pattern. Unfortunately, as was shown in section 4.5.2, OCR errors are *not* distributed homogenously.

Two factors related to OCR errors were identified as theoretically affecting the variables from which MI and LL are calculated (corpus size, span, node frequency, collocate frequency, and number of co-occurrences) and hence the resulting statistics. These factors are the distribution of errors across instances of the same word-type, and the distribution of spurious characters and spaces (chapter 3). Since these factors influence several variables and may offset each other, it is difficult to predict the overall direction and magnitude of any impact without empirical analysis.

8.2.1.2 In practice, what are OCR errors like and how do they impact frequency figures?

The DICER analysis (section 5.3.2) which I applied during the VARD training phase provided evidence that OCR errors are very varied (DICER identified 2,152 letter replacement rules, of which three quarters applied only once). OCR errors are also more distant from their correct forms than is observed in natural spelling variation (50% of errors were at edit-distance 1 from their corrections, 25% at edit-distance 2, 12% at edit-distance 3, but still over 10% at edit-distance 4 or greater), suggesting that OCR errors will be harder to correct than natural spelling variation. But these observations are *underestimates* of the variation and distance of OCR errors, because of limitations in training VARD (which supports only the correction of the most straightforward errors affecting only single words mappable on a one-to-one basis to an original word). The DICER analysis also showed that most corrective operations (65%) were substitutions, which were very varied, with ~25% being deletions and ~10% being insertions. However, these figures may tell us more about how VARD works than about what OCR errors are like.

Comparing the uncorrected and gold (i.e. hand-corrected) portions of the CNNE matching corpus (see section 4.4) showed that there was only a small difference in corpus size (i.e. overall token count), with the OCR corpus having 1.25% more tokens. There was, however, a large difference in the type count, with roughly twice as many types in the OCR corpus as in the gold corpus, and likewise a roughly double OCR type/token ratio. Unsurprisingly then, a majority of OCR types did not occur in the gold corpus. However, not all these spurious types were infrequent: around 7% of types occurring at least 10 times in the uncorrected corpus did not occur in the gold corpus. This implies that using a frequency floor may be helpful but will not eliminate all OCR errors.

Real-word errors (erroneous tokens which happen to coincide with a correct type) are problematic in theory: they are harder to recognize (since the word looks correct) and affect wordcounts twice (one less count for the correct type, one additional count for the incorrect type). Looking at these comparative frequencies provided evidence of the presence of real-word errors: around 9% of types occurring at least 10 times in the OCR corpus occurred *more* in the OCR corpus, an unexpected result suggesting that at least one of the occurrences of each of these types is a real-word error.

Comparing MI and LL statistics generated from the uncorrected and gold CNNE samples suggested that OCR errors were *not* homogenously distributed. Indeed, MI should be identical in both samples if OCR errors were homogenously distributed; instead, some uncorrected statistics are smaller, and some greater, than their gold counterparts. Additionally, some uncorrected LL statistics are greater than their gold counterpart; since OCR LL statistics should be smaller in a corpus with homogenously distributed OCR errors, this provides further evidence that OCR errors are not homogenously distributed. Two factors were identified in chapter 3 as potentially affecting the distribution of OCR errors: the distribution of stray characters and spaces, and the distribution of errors across instances of a single word-type. The observation that the overall wordcount of the uncorrected sample is greater than the overall wordcount of the gold sample suggests that there are indeed stray characters and spaces which affect wordcount, though I did not attempt to assess their distribution. The observation that the quality of the OCR varied extensively between files in the uncorrected CNNE sample further underlines the point that errors are not distributed homogenously, and hence are unlikely to be distributed homogenously across instances of a single word-type. I did not attempt a more thorough investigation of this issue.

It might have been expected that a determining factor for OCR reliability would be image readability. However, the variation in OCR errors across the CNNE matching corpus files, which were hand-picked for their readability, suggests that there may be little relationship between human and computer image readability[1].

8.2.1.3 In practice, how do OCR errors impact on two common collocation statistics?

I found evidence that OCR errors *do* have an impact on MI and LL (section 4.5). Considering pairs of 140 nodes and all other words in the CNNE matching corpus, I found that although OCR-derived MI values are often close to their gold counterpart as would be expected, they are also often *over*-estimations – an undesirable result which could lead to drawing unwarranted conclusions from observations of patterns exhibited by OCR data. OCR-derived LL statistics were found to often be *smaller* than their gold counterpart, as would be expected, but there were also some *over*-estimations. OCR MI rankings were found to be broadly reliable for small spans, especially when used in combination with an LL threshold. Both MI and LL were found to have non-negligible rates of false positives. For all measures considered, large spans were found to be less reliable than small spans, and best avoided. The use of a frequency floor, though recommended for the reason outlined in the previous section, did not improve the reliability of the statistics; in fact, rates of false positives were higher when considering only node/collocate pairs above the frequency floor. The use of MI in combination with an LL threshold emerged as the most reliable set-up, preferable to using MI or LL individually.

[1] It is a truism that there is, similarly, little relationship between human and computer text readability, since humans can often guess the correct form of a word including spelling errors which current correction software may not be able to correct.

8.2.1.4 How effective are two existing automated correction techniques for correcting OCR errors?

I found that VARD corrected few errors (the maximum recall achieved was 23%), and introduced on average as many errors as it corrected (the maximum precision achieved was 52%) (section 5.3.3). VARD thus yields no net benefit with the level of training which was practicable within this project. The variation exhibited by OCR errors, and the long list of DICER letter replacement rules required to handle it, suggested that a correction approach centred around letter replacement rules would be ineffective: additional training would not be enough to improve the situation.

In contrast, Overproof proved promising, attempting a correction for a majority (56%) of errors, and getting most corrections (83%) right (section 5.4.2). The improvement was variable from file to a file, but every file improved, with each corrected file being on average more similar to its gold counterpart by 8 words than the uncorrected version. Overproof also yielded considerable improvement to the type count, as well as the type/token ratio, with a dramatic halving of the number of hapaxes (section 5.4.3). Nevertheless, the number of types with frequency 10 or more which occurred more often in the OCR than in the gold corpus almost doubles. This suggests that real-word errors may be *more* of an issue in an Overproof-corrected corpus than in an uncorrected corpus. Since real-word errors may be considered more problematic than other types of errors, this observation should be weighed carefully when considering the costs and benefits of using Overproof. However, as outlined below, the rate of false positives from MI and LL actually *improves* in Overproof-corrected data, suggesting that this may well be a useful solution for researchers interested in collocation patterns in OCR data.

8.2.1.5 How much of an impact does the most promising correction technique have on two common collocation statistics?

Overproof corrections improve the reliability of MI and LL (section 5.4.4). All measures used showed that the most extreme values for MI and LL in uncorrected OCR data were ironed out in Overproof-corrected data. Although rates of false negatives for both MI and LL remained comparable (and negligible), the rates of false positives fell, especially dramatically for MI and for the larger spans. The reliability of rankings also improved for both LL and MI, with the greatest improvements affecting the larger spans. This suggested that researchers interested in working with large spans in OCR data may find Overproof-corrections particularly attractive.

8.2.2 Pertaining to the investigation of spatial patterns in large amounts of text

In chapter 1, I noted that there were two main types of research questions for which I wished to investigate a methodology: 'what places are mentioned in this corpus?' and 'what is said about this place?'. Section 8.2.2.1 will summarize the findings pertaining to the first, whilst section 8.2.2.2 will summarize the findings pertaining to the second.

8.2.2.1 What places are mentioned in this corpus?

Out of various approaches to identifying the places mentioned in a large amount of text, I singled out three as promising, and implementable without recourse to a geo-parser. One method involves reading through a list of words tagged 'Z2' (USAS tag for geographic names). A Z2 list does not have high precision (many items on the list will not be the city-names or country-names), rather it relies on the researcher recognizing the place-names (i.e. it relies on the researcher's prior knowledge of place-names), and it precludes the identification of multi-word place-names. Nevertheless, it can help the researcher identify the most frequent place-names fairly rapidly, and is hence scalable. A problem with the approach is that it is not well-equipped to deal with polysemy: place-names may often refer to several places, or to non-places as well as places. I suggested a procedure for dealing with this problem: (a) reading a random sample of concordance lines to determine the proportion of concordance lines which refer to a given place of interest, and then (b) using this proportion as a corrective factor to produce an estimate of the actual number of mentions of a given place from the frequency of the place-name as found on the Z2 list. Using this approach, I found that all three newspapers considered mentioned British cities throughout the UK, but that one of them (ERLN) tended to mention all cities more than the other two. This may arise from generic differences between the newspapers. Using the same approach on subsets of the newspaper texts divided by genre, I did indeed find large differences between the newspapers in terms of the distribution of mention of British cities across genres, and that the distributions across article-genres of *different* cities mentioned in the *same* newspapers were often more similar than the distributions across article-genres of the *same* city in *different* newspapers.

A second approach involved starting with a principled text-external list of place-names, and simply locating these place-names in the text. The advantages of this approach are that it can potentially provide context for interpreting the mentions of place-names, can help identify places which are *not* mentioned in a set of texts, and can also accommodate multi-word place-names. A disadvantage is that this approach is not necessarily very scalable, depending on the method used to locate the predefined place-names in the text. Using this approach with census data, I found that there was a weak relationship between the number of mentions of a British city and its population, but that population could not explain all the variance in the data. Moreover, the strength of the relationship between the number of mentions of cities and their population varied across the newspapers considered. This may be explained by generic differences between the newspapers.

The third approach consisted of using a list of words tagged Z2 and comparing it (automatically) to a gazetteer (a detailed list of place-names). Like the previous approach, this approach cannot help to identify multi-word place-names; however, it allows more than just the most frequent place-names to be identified, and it removes the reliance on a source of prior knowledge. The problem of polysemy remains unchanged. Using this approach, I found that, as mentioned above, one newspaper (ERLN) tended to mention all cities more than the two other newspapers. I also found that there were no overwhelming patterns of over- or under-emphasis on particular regions of the UK in one newspaper compared to the other two, though one newspaper (ERLN) displayed some tendency to mention places in the Midlands more than the other two papers, and another (PMGZ) contrariwise tended to mention places in the South of England more than the other two.

8.2.2.2 What is said about this place?

Out of various possible approaches to investigating the discourses surrounding place-names, I singled out two scalable operationalisations of collocation analysis. The first, the global approach, involved starting from a list of statistical collocates of the place-name, and categorizing these collocates according to the context in which the place-name and the statistical collocate co-occurred in most of a random subset of concordance lines for that co-occurrence. The approach was found to be scalable, and suitable for comparing discourses surrounding different place-names, but not well-suited to reliable comparisons between newspapers. Moreover, an interpretative difficulty associated with the approach is that more repetitive genres yield stronger collocates, so that looking at a list of strong collocates may lead to more attention being given to repetitive genres than may be thought warranted. Using this approach, I found that France and Russia occurred in a variety of contexts, including military/diplomatic discussions, advertisements, discussions surrounding artistic events and political gossip. I also found that collocates of France pointed to the existence of commercial, cultural and personal ties between Britain and France. A similar pattern was not observed for Russia. Collocates of Russia in all three newspapers were also found to betray a negative bias towards Russia; a similar pattern was not observed for France.

The second, the sampling approach, involved starting from a random sample of concordance lines featuring the place-name. The syntactic and semantic patterns in which the place-name occurred were then described, and the process repeated for a new random sample of concordance lines. This generated a framework for describing the phraseologies in which place-names were described. The framework was considered final once patterns which could not be accounted for using the existing framework ceased to be observed. Once finalized, the framework could be used to analyse more samples of concordance lines, as well as to help formulate queries for specific phraseologies throughout the whole corpus. Using this approach, I found that the overall properties of the language used around France and Russia was similar and could be described using the same set of labels and categories, with personifying and locational phraseologies accounting for more than half the mentions of either country in each newspaper. Within the personifying phraseologies, phraseologies which emphasise the relationship between countries and phraseologies which attribute immaterial possessions to the countries are the most frequent, suggesting that discussions in moral terms, and particularly about relations between countries, are an important feature of the discourse surrounding the countries. Looking at OF-phraseologies (which account for 1/5 of mentions of the countries in all newspapers) showed that the very same phrases were often used to refer to both countries, including references to moral properties which encode evaluations. These evaluations were more often good than bad for both countries, though the balance was less favourable for Russia than for France. Although the more abstract levels of analysis often revealed similarities between the discourses surrounding both countries, more detailed analysis also showed some differences. France was hence found to have a tendency to be referred to in more diverse and less abstract ways than Russia, with Russia's agency being expressed using more indirect means than France's, and France's interior and financial resources being referred to more often than Russia's.

8.3 EVALUATION

8.3.1 CONTRIBUTIONS TO THE FIELD OF HISTORY

The increasing abundance of digitised historical sources is a boon for the field of History. As was discussed in chapter 2, there is a growing awareness of the need for new historiographical methodologies appropriate for analysing large amounts of digital text. This thesis hence makes a major contribution to the field of History by demonstrating some of the ways in which corpus linguistic methods can be harnessed for the benefit of historical research.

In particular, this thesis outlines and illustrates various approaches to answering questions such as 'what places are mentioned in these texts?' and 'what is said about these places?'. These are important contributions not just to the field of History, but also for scholars from any field interested in investigating spatial patterns in text. Indeed, there remain technical challenges limiting the methodological possibilities for investigating spatial patterns in large amounts of text, primarily that of reliably geo-parsing entire large corpora. By outlining approaches to identifying place-names in large amounts of text without recourse to a geo-parser, this thesis hence contributes to what may be termed the *Spatial Humanities* as a whole[2].

Additionally, there is at present no existing standard methodology for investigating the representation of place-names in large amounts of text. This thesis begins to bridge this gap by outlining various approaches to doing so, including the provision of a method for devising a framework to describe the language surrounding mentions of place-names which can be used by historians not trained in Linguistics.

Of course, using corpus linguistic methods on digitised sources is not without its challenges. One of the major challenges is that of understanding the impact of OCR errors on the process of drawing conclusions on the basis of OCR data. Remarkably little work has been done on the impact of OCR errors on computer-based text analysis techniques. This thesis sets the agenda by exploring the nature and distribution of OCR errors and outlining their impact on two common collocation statistics. The issue is far from resolved, and this thesis simultaneously highlights issues worth pursuing in future research and formulates recommendations for best practice to enable researchers to continue using corpus linguistic techniques on OCR data whilst further research is under way.

[2] Based on this work, I was able to offer advice on spatial analysis of text to research projects at Lancaster University which were unable to make use of a geo-parser.

Further, the thesis evaluates two OCR post-correction solutions, rejecting one as unpromising, and highlighting another as helpful[3], especially for researchers interested in investigating patterns of collocation across large spans. However, even a very effective OCR post-correction solution will not achieve 100% accuracy; the issue of how OCR errors impact collocation statistics hence remains relevant even when the OCR data has undergone OCR post-correction. Another contribution of this thesis is thus that it provides the first assessment of the impact of OCR post-correction on collocation statistics.

8.3.2 Limitations

Practical considerations limited the size of the corpus which I could use to test the impact of OCR errors on MI and LL. As a consequence, the size of the differences in MI and LL which I could uncover was also limited. I mitigated this factor as much as I could by choosing words with frequencies ranging from 1 to 14,559. Nevertheless, it would be useful to replicate this work on a larger corpus. However, the major limitation to the size of the corpus is the need to produce a gold standard version of the OCR data; this requires a major investment of hand-correction time which is likely to prove limiting for future studies as well. In addition, the sampling of the corpus, also dictated by practical considerations, is not ideal. This limits the amount which could be said about differences in the importance of various factors including publication title and year. Nevertheless, I was able to consider thousands of node/collocate pairings, exhibiting a wide range of MI and LL values, and wide ranges of differences between the statistics derived from the various CNNE samples. This proved a sufficient amount of data to be able to provide the first account of the impact of OCR errors on OCR-derived MI and LL statistics.

An important limitation in the discussion of approaches to investigating spatial patterns is the limited amount of historical contextualisation provided in the discussion of the patterns

[3] Based on my recommendation of this solution, the British Library recommended the Overproof software to Hannah-Rose Murray, runner-up in the British Library Labs competition (2016) (Mahendra Mahey, Hannah-Rose Murray, personal communications, 7th Nov. 2016).

observed in the newspapers. The task of relating observations generated using corpus linguistic approaches to the kind of rich in-depth small-scale analyses which historians master is not a trivial one. That being the case, my contention is that great advances in the field of History (and other fields in the Humanities) will be achieved by scholars collaborating across disciplinary boundaries. This thesis is an important step in this direction, as I apply my knowledge of Linguistic issues to questions of relevance to the field of History. I do this by outlining and illustrating methodological approaches to observing textual sources; the next step is for historians to combine these approaches with those they are already familiar with[4].

8.4 DIRECTIONS FOR FUTURE RESEARCH

8.4.1 PERTAINING TO OCR ERRORS AND THEIR IMPACT ON CORPUS LINGUISTIC METHODS

- Various results have suggested that real-word errors may be a real issue in OCR data. Investigating their prevalence, their impact on corpus linguistic techniques, and ways of identifying them could be useful. Contextual data, such as n-grams and/or grammatical parsing, may prove helpful, for example, since they could help identify unlikely sequences of words which probably include a real-word error.
- My results have suggested that MI and LL are insufficiently conservative when applied to OCR data. Future work might to explore the possibility of using higher cut-off points to compensate for this.
- I have tested the impact of OCR errors on two collocation statistics only, MI and LL. Future work could usefully consider other collocation statistics, such as t-score and z-score.
- I have only explored the impact of OCR errors on collocation statistics, but there are other central corpus techniques, such as keyness analysis, which I have not explored. Future work could usefully consider these.

[4] Work in this direction is in fact under way, with Ruth Byrne, a trained historian, investigating discourses surrounding immigration in c19th newspapers using corpus linguistic techniques.

- My evidence has suggested that OCR errors are *not* distributed homogenously throughout a corpus. I did not attempt a thorough investigation of the impact of various factors such as publication date, publication title, article genre, font-type, word-type, etc. on the distribution of OCR errors. This would be useful to investigate further since it could, in particular, help to predict OCR error rates in new OCR collections.
- Future work could also explore heuristics for predicting the rate of OCR errors under certain conditions, such as the rate likely to affect a particular word-type, or the rate likely to affect word-counts from a particular article genre.
- I have shown theoretically that the distribution of stray characters and spaces, which affect wordcount, would have an impact on MI and LL. Evidence suggested that there are indeed stray characters and spaces which affect wordcount, but it was not possible in this thesis to explore their distribution throughout the corpus. Future work could explore this.
- If OCR errors are non-homogenous across variables such as publication title or publication year, comparisons between subcorpora assembled on the basis of these factors may be unreliable. I have not tested this, and future work may usefully consider this.
- OCR errors have been found to affect MI and LL. Future work could explore heuristics for predicting the *likely value* of the statistic under certain conditions; ultimately, it would be helpful to be able to provide and/or adjust confidence intervals for OCR-derived MI and LL statistics; this would allow for more intuitive and reliable use of OCR-derived statistics. Such work could explore, for example, the use of known OCR error indicators, such as the ratio of hapaxes to corpus size, to predict impact on MI and LL.
- To help gauge the impact of OCR errors on MI and LL, I calculated rates of false positives and false negatives, using a single commonly-used cut-off point per

statistic. Future work could replicate this, but using (and, thus, comparing) several cut-off points, as users are likely to do in practice.

8.4.2 Pertaining to the investigation of spatial patterns in large amounts of text

- One of the approaches I outlined for identifying place-names in large amounts of text without using a geo-parser, that based on the Z2 semantic tag, employs USAS, which is extensively used and supported. Future work could consider ways of refining this method with the aim of producing an integrated method of geo-parsing which could easily fit in with existing infrastructures.
- I have devised a framework for describing the phraseologies surrounding country-names in c19th newspapers. It would be interesting for future work to test and extend the framework to other text-types and types of place-names.
- At several points in this thesis, I have pointed out that it would be interesting to test contrastive collocation (introduced in section 2.3.2.4). For example, future work could explore the use of contrastive collocation for exploring stability and change over time, or variation across publications and text-types, in the representation of places.

8.5 Final remarks

The work in this thesis was intended to contribute to the methodology of History. I hope to have demonstrated that it is possible to analyse large amounts of text – both quantitatively and qualitatively – in ways that are useful to the historian. Since no one person can be an expert in all fields, it seems obvious to me that the best results in this new digital age will be achieved by scholars from various fields collaborating. To fully elucidate the historical questions which emerge from the consideration of vast amounts of text will hence likely require extensive knowledge about the historical context of the texts considered, as well as expertise in the analysis of large amounts of text, including computational aspects. My hope is that this work will inspire others in the various relevant disciplines to reach out across disciplinary divides in pursuit of answers to a myriad of yet unanswered historical questions.

9 REFERENCES

Albakry, M. (2007). Usage prescriptive rules in newspaper language. *Southern Journal of Linguistics, 31*, 28-56.

Alex, B., & Burns, J. (2014). *Estimating and rating the quality of optically character recognised text.* Paper presented at the Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, Madrid, Spain.

Attar, K. (2010). S and Long S. In M. F. Suarez & H. R. Woudhuysen (Eds.), *The Oxford Companion to the Book* (Vol. 2, pp. 1116). Oxford: Oxford University Press.

Aull, L. L., & Brown, D. W. (2013). Fighting words: a corpus analysis of gender representations in sports reportage. *Corpora, 8*(1), 27-52.

Baird, H. S., & Tombre, K. (2014). The Evolution of Document Image Analysis *Handbook of Document Image Processing and Recognition* (pp. 63-71): Springer.

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

Baker, P. (2010a). Representations of Islam in British broadsheet and tabloid newspapers 1999-2005. *Journal of Language and Politics, 9*(2), 310-338.

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

Baker, P., & Egbert, J. (2016). *Triangulating methodological approaches in corpus linguistic research*. New York: Routledge.

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyzanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refuges and asylum seekers in the UK press. *Discourse and Society, 19*(3), 273-306.

Baker, P., Gabrielatos, C., & McEnery, T. (2013). Sketching Muslims: A Corpus Driven Analysis of Representations Around the Word 'Muslim' in the British Press 1998-2009'. *Applied Linguistics, 34*(3), 255-278.

Balk, H. (2009). *Poor access to digitised historical texts: the solutions of the IMPACT project*. Paper presented at the Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data, Barcelona, Spain.

Balk, H., & Conteh, A. (2011). *IMPACT: centre of competence in text digitisation*. Paper presented at the Proceedings of the 2011 Workshop on Historical Document Imaging and Processing, Beijing, China.

Baron, A. (2011). *Dealing with spelling variation in early modern English texts.* (Unpublished PhD thesis), Lancaster University.

Baron, A., & Rayson, P. (2009). *Automatic standardization of texts containing spelling variation, how much training data do you need?* Paper presented at the Corpus Linguistics Conference (CL 2009), 20-23 July 2009, University of Liverpool, UK.

Barry, Q. (2012). *War in the East: A military history of the Russo-Turkish war 1877-8*. Solihull: Helion & Company Ltd.

Bauer, L. (1994). *Watching English change: an introduction to the study of linguistic change in standard Englishes in the twentieth century*. London: Longman.

Bauer, L. (2002). Inferring Variation and Change from Public Corpora. In J. K. Chambers, P. Trudgill & N. Schilling-Estes (Eds.), *The Handbook of Language Variation and Change* (pp. 97-114). Oxford: Wiley-Blackwell.

Bennamoun, M., & Mamic, G. J. (2012). Optical Character Recognition *Object recognition: fundamentals and case studies* (pp. 199-220): Springer Science & Business Media.

Bingham, A. (2010). The Digitization of Newspaper Archives: Opportunities and Challenges for Historians. *Twentieth Century British History, 21*(2), 225-231.

Blanke, T., Bryant, M., & Hedges, M. (2012). Ocropodium: open source OCR for small-scale historical archives. *Journal of Information Science, 38*(1), 76-86.

Bloor, M., & Bloor, T. (2007). *The Practice of Critical Discourse Analysis: An Introduction*. London: Hodder Arnold.

Bos, B. (2012). From 1760 to 1960: Diversification and Popularization. In R. Facchinetti (Ed.), *News as Changing Texts: Corpora, Methodologies and Analysis* (pp. 91-144). Newcastle upon Tyne: Cambridge Scholars Publishing.

Brake, L. (2001). On Print Culture: the State We're In. *Journal of Victorian Culture, 6*(1), 125-136.

British Library. (n.d.). More about the newspaper titles in the 19th Century British Library Newspapers database. Retrieved from British Library, Help For Researchers, 19th Century Newspapers Database website: http://www.bl.uk/reshelp/pdfs/headnotesconsolidatedlist.pdf

Brownlees, N. (2012). The Beginning of Periodical News (1620-1665). In R. Facchinetti (Ed.), *News as Changing Texts: Corpora, Methodologies and Analysis* (pp. 5-48). Newcastle upon Tyne: Cambridge Scholars Publishing.

Bryant, J., & Oliver, M. B. (Eds.). (2009). *Media effects: Advances in theory and research* (3d ed.). New York: Routledge.

Bublitz, W., Jucker, A. H., & Schneider, K. P. (2010). Preface to the handbook series. In A. H. Jucker & I. Taavitsainen (Eds.), *Historical Pragmatics* (pp. v-vii). Berlin: De Gruyter.

Caldas-Coulthard, C. R., & Moon, R. (2010). 'Curvy, hunky, kinky': Using corpora as tools for critical analysis. *Discourse and Society, 21*(2), 99-133.

Cantos, P. (2012). The Use of Linguistic Corpora for the Study of Linguistic Variation and Change: Types and Computational Applications. In J. M. H. Hernández-Campoy (Ed.), *The Handbook of Historical Sociolinguistics* (pp. 99-122). Oxford: Wiley-Blackwell.

Carrasco, R. C. (2014). *An open-source OCR evaluation tool*. Paper presented at the Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, Madrid, Spain.

Chase, M. (2009). Digital Chartists: Online Resources for the Study of Chartism. *Journal of Victorian Culture, 14*(2), 294-301.

Cheng, W., & Lam, P. W. Y. (2013). Western perceptions of Hong Kong Ten Years On: A Corpus-driven Critical Discourse Study. *Applied Linguistics, 34*(2), 173-190.

Claridge, C. (2010). News discourse. In A. H. Jucker & I. Taavitsainen (Eds.), *Historical Pragmatics* (pp. 587-620). Berlin: De Gruyter.

Clark, C. (2010). Evidence of *evidentiality* in the quality press 1993 and 2005. *Corpora, 5*(2), 139-160.

Cohen, D. J., & Rosenzweig, R. (2006). Becoming digital - How to Make Text Digital: Scanning, ORC, and TypingDigital History: a guide to gathering, preserving, and presenting the past on the web (Vol. 28). Philadelphia: University of Pennsylvania Press.

Colella, S. (2013). "That Inscrutable Something": Business in the periodical press. *Victorian Periodicals Review, 46*(3), 317-342.

Coleman, R., McCombs, M., Shaw, D., & Weaver, D. (2009). Agenda Setting. In Wahl-Jorgensen & T. Hanitzsch (Eds.), *The Handbook of Journalism Studies* (pp. 147-160). New York: Routledge.

Cressie, N., & Read, T. R. C. (1984). Multinomial goodness-of-Fit Tests. *Journal of the Royal Statistical Society, Series B (Methodological), 46*(3), 440-464.

Culpeper, J., & Kytö, M. (2010). *Early Modern English dialogues: Spoken interaction as writing*: Cambridge University Press.

Daðason, J. F., Bjarnadóttir, K., & Rúnarsson, K. (2014). *The Journal Fjölnir for Everyone: The Post-Processing of Historical OCR Texts.* Paper presented at the Language Resources and Technologies for Processing and Linking Historical Documents and Archives-Deploying Linked Open Data in Cultural Heritage–LRT4HDA Workshop Programme.

Deswarte, R. (2010). Growing "The Faith in Numbers": Quantitative Digital Resources and Historical Research in the Twenty-First Century. *Journal of Victorian Culture, 15*(2), 281-286.

Dobrosklonskaya, T. (2013). Media linguistics: a new paradigm in the study of media language. In M. N. Volodina (Ed.), *Mediensprache und Medienkommunikation* (pp. 37-48). Mannheim: Institut Für Deutsche Sprache.

Duguid, A. (2010a). Investigating anti and some reflections on Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS). *Corpora, 5*(2), 191-220.

Duguid, A. (2010b). Newspaper discourse informalisation: a diachronic comparison from keywords. *Corpora, 5*(2), 109-138.

English Department Uppsala University. (undated). The Corpus of Nineteenth-Century Newspaper English (CNNE). Retrieved 26/08/2015, from http://www.engelska.uu.se/Forskning/engelsk_sprakvetenskap/Forskningsomraden/Electronic_Resource_Projects/Nineteenth-century_Newspaper/

Evershed, J., & Fitch, K. (2014). *Correcting noisy OCR: context beats confusion.* Paper presented at the Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage.

Evert, S. (2005). *The statistics of word cooccurrences.* (PhD Thesis), Stuttgart University.

Fairclough, N., & Graham, P. (2002). Marx as critical discourse analyst: the genesis of a critical method and its relevance to the critique of global capital. *Estudios de Sociolinguistica, 3*(1), 185-229.

Fitzmaurice, S. M. (2010). Mr Spectator, identity and social roles in an early eighteenth-century community of practice and the periodical discourse community. In P. Pahta, M. Nevala, A. Nurmi & M. Palander-Collin (Eds.), *Social Roles and Language Practices in Late Modern English* (pp. 29-53). Amsterdam: John Benjamins Publishing.

Fitzsimmons-Doolan, S. (2009). Is public discourse about language policy really public discourse about immigration? A corpus-based study. *Language Policy, 8*(4), 377-402.

Fleming, P., & King, E. (2009). The British Library newspaper collections and future strategy. *Interlending & Document Supply, 37*(4), p. 223-228.

Fowler, R. (1987). Notes on critical linguistics. In R. Steele & T. Threadgold (Eds.), *Language Topics: Essays in honour of Michael Halliday* (Vol. II, pp. 481-492). Amsterdam: John Benjamins Publishing.

Freake, R., Gentil, G., & Sheyholislami, J. (2011). A bilingual corpus-assisted discourse study of the construction of nationhood and belonging in Quebec. *Discourse & Society, 22*(1), 21-47.

Fusari, S. (2010). Sweetheart deals, wildcat strikes and other dangerous things: Metaphorical representations of Alitalia bailout and privatization in the British, US American and Italian press. *Facta universitatis-series: Linguistics and Literature, 8*(2), 91-104.

Fyfe, P. (2009). 2008 VanArsdel Prize Graduate Student Essay: The Random Selection of Victorian New Media. *Victorian Periodicals Review, 42*(1), 1-23.

Fyfe, P. (2016). An Archaeology of Victorian Newspapers. *Victorian Periodicals Review, 49*(4), 546-577.

Gabrielatos, C., & Baker, P. (2008). Fleeing, Sneaking, Flooding: A Corpus Analysis of Discursive Constructions of Refugess and Asylum Seekers in the UK Press, 1996-2005. *Journal of English Linguistics, 36*(1), 5-38.

García-Marrugo, A. (2013). What's in a name ? The representation of illegal actors in the internal conflict in the Colombian press. *Discourse and Society, 24*(4), 421-445.

Garside, R., Leech, G. N., & McEnery, T. (1997). *Corpus annotation: linguistic information from computer text corpora*. London: Longman.

Garside, R., & Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In R. Garside, G. Leech & A. McEnery (Eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora.* (pp. pp. 102-121). London: Longman.

Gee, J. P. (2011). *An introduction to discourse analysis : theory and method*. New York: Routledge.

Gibbs, F. W., & Cohen, D. J. (2011). A Conversation with Data: prospecting Victorian Words and Ideas. *Journal of Victorian Studies, 54*(1), 69-77.

Gildea, R. (1996). *Barricades and Borders, Europe 1800-1914* (second edition ed.). New York: Oxford University Press.

Gleason, J. H. (1950). *The genesis of Russophobia in Great Britain*. Cambridge, Mass.: Harvard University Press.

Gregory, I. N., & Hardie, A. (2011). Visual GISting: bringing together corpus linguistics and Geographical Information Systems. *Literary and Linguistic Computing, 26*(3), 297-313.

Hakam, J. (2009). The 'cartoons controversy': a Critical Discourse Analysis of English-language Arab newspaper discourse. *Discourse and Society, 20*(1), 33-57.

Hardie, A. (2012). CQPweb—combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics, 17*(3), 380-409.

Hardie, A. (forthcoming). A dual sort-and-filter strategy for statistical analysis of collocation, keywords, and lockwords.

Hardie, A., & McEnery, T. (2010). On two traditions in corpus linguistics, and what they have in common. *International Journal of Corpus Linguistics, 15*(3), 384-394.

Heuser, R., & Le-Khac, L. (2011). Learning to Read Data: Bringing out the Humanistic in the Digital Humanities. *Journal of Victorian Studies, 54*(1), 79-86.

Hirst, G. (2013). Computational linguistics. In K. Allan (Ed.), *The Oxford Handbook of the History of Linguistics* (pp. 707-726). Oxford: Oxford University Press.

Hitchcock, T. (2013). Confronting the Digital: Or How Academic History Writing Lost the Plot. *Cultural and Social History, 10*(1), 9-23.

Holley, R. (2009). How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. *D-Lib Magazine, 15*(3/4).

Howard, M. (1988). *The Franco-Prussian War: The German invasion of France, 1870-1871.* London: Routledge.

Hughes, M. J. (2011). British Opinion and Russia Terrorism in the 1880s. *European History Quarterly, 41*(2), 255-277.

Hughes, M. J. (2015). *Exotic Friend, Exotic Enemy: Nineteenth-Century Russia in the British Imagination*. Unpublished manuscript.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hunston, S. (2012). Corpus Linguistics: Historical Development. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 1366-1372). London: Wiley and Sons.

James, L. (2009). The Era. In L. Brake & M. Demoor (Eds.), *Dictionary of Nineteenth-Century Journalism* (pp. 206). Gent and London: Academia Press and the British Library.

Jaworska, S., & Krishnamurthy, R. (2012). On the F word: A corpus-based analysis of the media representation of feminism in British and German press discourse, 1990-2009. *Discourse and Society, 23*(4), 401-431.

Jeffries, L. (2009). *Critical Stylistics: the Power of English*. Basingstoke: Palgrave Macmillan.

Jeffries, L., & McIntyre, D. (2010). *Stylistics*. Cambridge: Cambridge University Press.

Jeffries, L., & Walker, B. (2012). Key words in the press: A critical corpus-driven analysis of ideology in the Blair years (1998- 2007). *English Text construction, 5*(2), 208-229.

Johnson, S., Culpeper, J., & Suhr, S. (2003). From Politically Correct Councillors to Blairite Nonsense Discourses of Political Correctness in Three British Newspapers. *Discourse & Society, 14*(1), 29-47.

Käding, F. W. (1897). *Häufigkeitswörterbuch der deutschen Sprache*. Steglitz: privately published.

Kent, C. A. (2009). The Pall Mall Gazette. In L. Brake & M. Demoor (Eds.), *Dictionary of Nineteenth-Century Journalism* (pp. 478). Gent and London: Academia Press and the British Library.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). the sketch engine. *Information Technology, 105*, 116.

Kim, K. H. (2014). Examining US news media discourses about North Korea: A corpus-based critical discourse analysis. *Discourse & Society, 25*(2), 221-244.

King, E. (2005). Digitisation of Newspapers at the British Library. *The Serials Librarian, 49*(1), 165-181.

King, E. (2007). Digitisation of British Newspapers 1800-1900 *19th Century British Newspapers*. Detroit: Gale Cengage Learning.

Kress, G. (1990). Critical Discourse Analysis. *Annual Review of Applied Linguistics, 11*, 84-99.

Kukich, K. (1992). Techniques for automatically correcting words in text. *ACM Computing Surveys (CSUR), 24*(4), 377-439.

Kytö, M. (2010). Data in historical pragmatics. In A. H. Jucker & I. Taavitsainen (Eds.), *Historical pragmatics* (pp. 33-68). Berlin: De Gruyter.

Leary, P. (2005). Googling the Victorians. *Journal of Victorian Culture, 10*(1), 72-86.

Lewis, M. D. (2014). The Edinburgh and Quarterly Reviews in 1848: Allies against French Revolution and British Democracy. *Victorian Periodicals Review, 47*(2), pp. 208-233.

Liddle, D. (2012). Reflections on 20,000 Victorian Newspapers: ‘Distant Reading’ The Times using The Times Digital Archive. *Journal of Victorian Culture, 17*(2), 230-237.

Liddle, D. (2015). Genre:" Distant Reading" and the Goals of Periodicals Research. *Victorian Periodicals Review, 48*(3), 383-402.

Lopresti, D. (2009). Optical character recognition errors and their effects on natural language processing. *International Journal on Document Analysis and Recognition (IJDAR), 12*(3), 141-151.

Lowe, J. (1994). *The Great Powers, Imperialism, and the German Problem, 1865-1925*. New York: Routledge.

Mahlberg, M. (2007). A corpus stylistic perspective on Dicken's *Great Expectations*. In M. Lambrou & P. Stockwell (Eds.), *Contemporary Stylistics* (pp. 19-31). London: Routledge.

Mahlberg, M. (2010). Corpus Linguistics and the Study of Nineteenth-Century Fiction. *Journal of Victorian Culture, 15*(2), 292-298.

Mahlberg, M. (2012). Corpus Stylistics-Dickens, Text-Drivenness and the Fictional World. In J. John (Ed.), *Dickens and Modernity* (pp. 94-114). Woodbridge, England: Brewer.

Mahlberg, M. (2013). *Corpus stylistics and Dickens's fiction*. New York: Routledge.

Marchi, A. (2010). "The moral *in* the story”: a diachronic investigation of lexicalised morality in the UK press. *Corpora, 5*(2), 161-189.

Mautner, G. (2007). Mining large corpora for social information: the case of *elderly*. *Language in Society, 36*, 21-72.

McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

McEnery, T., & Hardie, A. (2013). The history of corpus linguistics. In K. Allan (Ed.), *The Oxford Handbook of the History of Linguistics* (pp. 727-746). Oxford: Oxford University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies : an advanced resource book*. London: London : Routledge.

McGann, J. (1998). Textual scholarship, textual theory, and the uses of electronic tools: A brief report on current undertakings. *Victorian Studies, 41*(4), 609-619.

McGann, J. (2008). The Future is Digital. *Journal of Victorian Culture, 13*(1), 80-88.

McGann, J. (n.d.). The Complete Writings and Pictures of Dante Gabriel Rossetti: A Hypermedia Research Archive. Retrieved 23/05/2014, from http://www.rossettiarchive.org/

Mills, S. (2004). *Discourse*. London: Routledge.

Mitchell, B. R., & Deane, P. (1971). Population and Vital Statistics 8. Population of the Principal Towns of the United Kingdom - 1801-1951 *Abstract of British Historical Statistics* (pp. 24-27). Cambridge: Cambridge University Press.

Moretti, F. (2000). Conjectures on World Literature. *New Left Review, 1*(Jan-Feb 2000), 54-68.

Mussell, J. (2008). Ownership, Institutions, and Methodology. *Journal of Victorian Culture, 13*(1), 94-100.

Mussell, J. (2009). Cohering Knowledge in the Nineteenth Century: Form, Genre and Periodical Studies. *Victorian Periodicals Review, 42*(1), 93-103.

Mussell, J. (2010). Processing the Past. *Journal of Victorian Culture, 280*.

Mussell, J. (2012a). *The Nineteenth-Century Press in the Digital Age*. New York: Palgrave Macmillan.

Mussell, J. (2012b). The Passing of Print. *Media History, 18*(1), 77-92.

Mussell, J. (2012c). Teaching Nineteenth-Century Periodicals Using Digital Resources: Myths and Methods. *Victorian Periodicals Review, 2*(45), 201-209.

Nevalainen, T. (2010). Historical Sociolinguistics. In R. Wodak, B. Johnstone & P. Kerswill (Eds.), *The SAGE Handbook of Sociolinguistics* (pp. 279-295). Thousand Oaks, Cal.: Sage Publications.

Nicholson, B. (2012a). Counting Culture; or, How to Read Victorian Newspapers from a Distance. *Journal of Victorian Culture, 17*(2), 238-246.

Nicholson, B. (2012b). *Looming Large: America and the Late-Victorian Press, 1865-1902.* (PhD thesis), University of Manchester.

Nicholson, B. (2013). The Digital Turn: Exploring the Methodological possibilities of digital newspaper archives. *Media History, 19*(1), 59-73.

O'Halloran, K. (2010). How to use corpus linguistics in the study of media discourse. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 563-577). New York: Routledge.

Oakes, M. P. (1998). *Statistics for corpus linguistics*: Edinburgh : Edinburgh University Press.

Orpin, D. (2005). Corpus Linguistics and Critical Discourse Analysis: Examining the Ideology of Sleaze. *International Journal of Corpus Linguistics, 10*(1), 37-61.

Paltridge, B. (2012). *Discourse analysis : an introduction*. London: Bloomsbury Academic.

Parry, J. (2001). The Impact of Napoleon III on British Politics, 1851-1880. *Transactions of the Royal Historical Society, 11*, 147-175.

Partington, A. (2010). Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS) on UK newspapers: an overview of the project. *Corpora, 5*(2), 83-108.

Partington, A. (2012). The changing discourses on antisemitism in the UK press from 1993 to 2009: A modern-diachronic corpus-assisted discourse study. *Journal of Language and Politics, 11*(1), 51-76.

Partington, A., Duguid, A., & Taylor, C. (2013). *Patterns and meanings in discourse: theory and practice in corpus-assisted discourse studies (CADS)*. Amsterdam: John Benjamins Publishing.

Pearce, M. (2008). Investigating the collocational behaviour of man and woman in the BNC using Sketch Engine. *Corpora, 3*(1), 1-29.

Pearson, R. (2008). Etexts and Archives. *Journal of Victorian Culture, 13*(1), 88-93.

Percy, C. (2012). Early Advertising and Newspapers as Source of Sociolinguistic Investigation. In J. M. H. Hernández-Campoy (Ed.), *The Handbook of Historical Sociolinguistics* (pp. 191-210). Oxford: Wiley-Blackwell.

Perrin, D. (2013). *The linguistics of newswriting*. Amsterdam: John Benjamins Publishing.

Perse, E. M. (2000). *Media effects and society*. Mahwah, N.J.: L. Erlbaum Associates.

Pionke, A. D. (2014). Excavating Victorian Cuba in the British Periodicals Database. *Victorian Periodicals Review, 47*(3), 369-397.

Potter, P. A. (1998). Centripetal textuality. *Victorian Studies, 41*(4), 593-607.

Prensky, M. (2001). Digital natives, digital immigrants part 1. *On the horizon, 9*(5), 1-6.

Prentice, S., & Hardie, A. (2009). Empowerment and Disempowerment in the Glencairn Uprising: A Corpus-Based Critical Analysis of Early Modern English News Discourse'. *Journal of Historical Pragmatics, 10*(1), 23-55.

Prescott, A. (2014). I'd Rather be a Librarian: A Response to Tim Hitchcock, 'Confronting the Digital'. *Cultural and Social History, 11*(3), 335-341. doi: 10.2752/147800414X13983595303192

Pumfrey, S., Rayson, P., & Mariani, J. (2012). Experiments in 17th Century English: Manual Versus Automatic Conceptual History. *Literary and Linguistic Computing, 27*(4), 395-408.

Rasinger, S. M. (2010). "Lithuanian migrants send crime rocketing":representation of 'new' migrants in regional print media. *Media, Culture, Society, 32*(6), 1021-1030.

Rayson, P., Archer, D., & Smith, N. (2005). *VARD versus Word: A comparison of the UCREL variant detector and modern spell checkers on English historical corpora*. In Proceedings of Corpus Linguistics 2005 , University of Birmingham, Birmingham, UK.

Reynaert, M. (2008). Non-interactive OCR post-correction for giga-scale digitization projects *Computational Linguistics and Intelligent Text Processing* (pp. 617-630): Springer.

Robertson, A. M., & Willett, P. (1992). *Searching for historical word-forms in a database of 17th-century English text using spelling-correction methods.* Paper presented at the Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval.

Royle, T. (1999). *Crimea: The great Crimean War 1854-1856.* London: Little, Brown.

Rupp, C., Rayson, P., Gregory, I., Hardie, A., Joulain, A., & Hartmann, D. (2014). *Dealing with heterogeneous big data when geoparsing historical corpora.* Paper presented at the 2014 IEEE International Conference on Big Data.

Sanders, M. (2009). The Chartist Text in an Age of Digital Reproduction. *Journal of Victorian Culture, 14*(2), 301-307.

Saussure, F. d. ([1916] 1986). *Course in General Linguistics* (R. Harris, Trans.). Peru, Ill. (US): Open Court Publishing Company.

Schroeder, P. W. (2000). International politics, peace, and war, 1815-1914. In T. C. W. Blanning (Ed.), *The Nineteenth Century: Europe 1789-1914*. Oxford: Oxford University Press.

Shaw, J. (2007). Selection of Newspapers *19th Century British Newspapers*. Detroit: Gale Cengage.

Shirley, M. (2009). Reynold's Weekly Newspaper. In L. Brake & M. Demoor (Eds.), *Dictionary of Nineteenth-Century Journalism* (pp. 541). Gent and London: Academia Press and the British Library.

Smitterberg, E. (2014). Syntactic Stability and Change in Nineteenth-century Newspaper Language. In M. Hundt (Ed.), *Late Modern English Syntax* (pp. 311-330). Cambridge: Cambridge University Press.

Stauffer, A. (2011). Introduction: Searching Engines, Reading Machines. *Journal of Victorian Studies, 54*(1), 63-68.

Stubbs, M. (1997). Whorf's Children: Critical Comments on Critical Discourse Analysis (CDA). In A. Ryan & A. Wray (Eds.), *Evolving Models of Language* (pp. 100-116). Clevedon: Multilingual Matters.

Taavitsainen, I., & Jucker, A. H. (2010). Trends and developments in historical pragmatics. In A. H. Jucker & I. Taavitsainen (Eds.), *Historical Pragmatics* (pp. 3-32). Berlin: De Gruyter.

Tabbert, U. (2013). *Crime through a corpus: The linguistic construction of offenders, victims and crimes in the German and UK press.* (PhD thesis), University of Huddersfield.

Tanner, S., Muñoz, T., & Ros, P. H. (2009). Measuring mass text digitization quality and usefulness. *D-Lib Magazine, 15*(7/8), 1082-9873.

Taylor, C. (2008). What is *corpus linguistics*? What the data says. *ICAME Journal, 32*(179-200).

Taylor, C. (2010). Science in the news: a diachronic perspective. *Corpora, 5*(2), 221-250.

Taylor, C. (2013). Searching for similarity using corpus-assisted discourse studies. *Corpora, 8*(1), 81-113.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins Publishing Company.

Toolan, M. (1998). *Language in Literature: An Introduction to Stylistics*. London: Taylor & Francis.

Towheed, S. (2010). Reading in the Digital Archive. *Journal of Victorian Culture, 15*(1), 139-143.

Turner, M. W. (2006). Time, Periodicals, and Literary Studies. *Victorian Periodicals Review, 39*(4), 309-316.

Varieng. (2014). The Corpus of Nineteenth-Century Newspaper English (CNNE). Retrieved 28/10/2015, from http://www.helsinki.fi/varieng/CoRD/corpora/CNNE/index.html

Vessey, R. (2014). Borrowed words, mock language and nationalism in Canada. *Language and Intercultural Communication, 14*(2), 1-15.

Volk, M., Furrer, L., & Sennrich, R. (2011). Strategies for reducing and correcting OCR errors *Language Technology for Cultural Heritage* (pp. 3-22): Springer.

Vuohelainen, M. (2013). "Contributing to Most Things": Richard Marsh, Literary Production, and the Fin de Siecle Periodicals Market. *Victorian Periodicals Review, 46*(3), 401-422.

Welling, G. (2001). Can computers help us read history better? Computerized text-analysis of four editions of the outline of american history. *History and Computing, 13*(2), 151-160.

Westin, I., & Geisler, C. (2002). A multi-dimensional study of diachronic variation in British newspaper editorials. *ICAME Journal, 26*, 133-152.

Wick, M. L., Ross, M. G., & Learned-Miller, E. G. (2007). *Context-sensitive error correction: Using topic models to improve OCR.* Paper presented at the Ninth International Conference on Document Analysis and Recognition, 2007. ICDAR 2007.

Wilson, A., & Thomas, J. A. (1997). Semantic annotation. In R. Garside, G. Leech & A. McEnery (Eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora* (pp. 53-65). London: Longman.

Wodak, R., Johnstone, B., & Kerswill, P. (2010). Introduction. In R. Wodak, B. Johnstone & P. Kerswill (Eds.), *The SAGE Handbook of Sociolinguistics* (pp. 1-7). Thousand Oaks, Cal.: Sage Publications.

Wodak, R., & Meyer, M. (2009). Critical discourse analysis: History, agenda, theory and methodology. In R. Wodak & M. Meyer (Eds.), *Methods for Critical Discourse Analysis* (pp. 1-33). London: Sage.

Wynne, M. (Ed.). (2005). *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books.

10 APPENDICES

10.1 NEWSPAPER CODES

Code	Newspaper Title
ANJO	Aberdeen Journal
BNER	Baner
BNWL	Belfast News-Letter
BDPO	Birmingham Daily Post
BRPT	Brighton Patriot
BLMY	Bristol Mercury
CNMR	Caledonian Mercury
CHPN	Champion
CHTR	The Charter
CHTT	Chartist
CTCR	Chartist Circular
CWPR	Cobbet's Weekly Political Register
DNLN	Daily News
DYMR	Derby Mercury
ERLN	The Era
EXLN	Examiner
FRJO	Freeman's Journal
GNDL	Genedl
GWHD	Glasgow Herald
GLAD	Goleuad
GCLN	Graphic
HPTE	Hampshire/Portsmouth Telegraph
HLPA	Hull Packet
IPNW	Illustrated Police News
IPJO	Ipswich Journal
JOJL	Jackson's Oxford Journal
LEMR	Leeds Mercury
LVMR	Liverpool Mercury
LINP	Lloyd's Illustrated Newspaper
MRTM	Manchester Times
MCLN	Morning Chronicle
NECT	Newcastle Courant
NRWC	North Wales Chronicle
NREC	Northern Echo
NRSR	Northern Star
ODFW	Odd Fellow
OPTE	Operative
PMGZ	Pall Mall Gazette
PMGU	Poor Man's Guardian
PNCH	Preston Chronicle
RDNP	Reynold's Newspaper
SNSR	Southern Star
TEFP	Trewman's Exeter Flying Post
WMCF	Western Mail

10.2 Test nodes and their frequencies

Here are listed the nodes used for assessing the impact of OCR errors on MI and LL statistics, along with their frequency in the gold, uncorrected and Overproof-corrected sections of the CNNE matching corpus, see discussion in chapters 4 and 5.

N	node	gold	OCR	Ovprf
1	after	207	191	217
2	afternoon	47	36	45
3	agricultural	9	8	8
4	altercation	2	1	2
5	although	40	29	38
6	americans	7	4	7
7	animated	3	3	3
8	arthur	4	4	5
9	asylum	6	5	6
10	attacked	10	7	10
11	audacious	2	2	2
12	austria	3	2	3
13	banstead	7	7	7
14	battalion	10	10	10
15	battle	9	6	8
16	bearing	11	8	8
17	beautifully	3	3	3
18	belgian	7	3	7
19	belgium	7	1	4
20	betwixt	2	1	1
21	beyond	20	15	17
22	birmingham	16	9	12
23	blockade	5	5	5
24	bradford	25	11	18
25	bristol	24	22	22
26	brown	2	2	2
27	brussels	9	1	3
28	building	52	43	49
29	bulgarians	2	1	1
30	cambridge	7	7	7
31	cash	3	3	3
32	cheap	4	4	4
33	cholera	2	1	3
34	constructing	3	3	3
35	creature	5	3	3
36	creatures	9	6	8
37	daily	13	12	12
38	declaration	6	4	5
39	deed	4	4	6
40	despotism	3	2	3
41	disease	10	10	11
42	dublin	8	7	8
43	efficient	8	8	9
44	emigrate	3	3	3
45	empire	10	9	10
46	energetic	4	3	4
47	engines	19	18	19
48	england	60	42	48
49	english	58	45	53
50	englishman	4	2	3
51	englishmen	4	4	4
52	enjoy	7	8	8
53	europe	10	5	7
54	european	4	3	4
55	fall	25	31	32
56	females	3	2	2
57	flag	5	1	7
58	france	14	9	12
59	fresh	12	11	13
60	government	137	97	121
61	grave	6	6	6
62	great	222	204	221
63	hailes	3	1	3
64	handkerchief	4	3	3
65	heard	58	45	56
66	indian	9	8	8
67	inquired	3	3	3
68	kitchen	4	3	4
69	knots	5	5	5
70	lancashire	8	5	6
71	lancaster	4	3	4
72	law	105	90	92
73	liverpool	79	61	69
74	london	107	85	97
75	loud	9	10	9
76	lunatic	3	3	3
77	manchester	44	25	32
78	mansion	18	7	8
79	medical	25	20	23
80	merchant	10	7	7
81	meredith	8	5	6
82	ministry	17	16	17
83	murder	52	45	49
84	netherlands	7	2	6
85	noticeable	3	3	3
86	of	7298	6989	7352
87	offences	6	4	4
88	outer	4	3	4
89	owners	14	12	12
90	paris	14	14	14
91	perform	5	4	5
92	pinioned	3	1	3
93	place	153	127	137
94	plain	15	12	12
95	police	158	111	123
96	power	65	59	62
97	powers	15	13	13
98	radicals	3	1	1
99	railway	72	59	67
100	raised	16	15	19
101	reckon	4	3	4
102	religion	11	8	10
103	render	11	7	12
104	reports	11	10	12
105	rod	6	5	5
106	rothschild	5	3	3
107	rushing	3	3	3
108	russia	7	6	6
109	russian	7	4	5
110	science	6	5	6
111	serious	31	24	29
112	settlement	13	7	11
113	sewage	6	6	6
114	sheet	4	4	4
115	sheffield	6	2	4
116	sitting	8	6	6
117	somewhat	23	19	20
118	songs	3	2	2
119	spaces	3	3	3
120	spent	7	7	9
121	statesmanship	1	1	1
122	stop	10	10	9
123	tearing	3	3	3
124	that	1878	1727	1840
125	the	14559	12838	13935
126	thinking	9	8	9
127	tickets	4	3	3
128	time	270	244	285
129	to	4646	4500	4669
130	unconscious	5	5	5
131	used	28	26	28
132	using	3	3	5
133	vaccination	3	2	2
134	valentia	1	1	1
135	wanting	6	4	6
136	war	17	15	19
137	weather	35	27	35
138	were	1013	858	964
139	who	529	455	501
140	yarmouth	4	2	2

10.3 Issues in the VARD training sample

Here are listed all of the sources from which partitions for the VARD training sample were obtained, see discussion in section 5.3.2.

Aberdeen Weekly Journal

25/10/1881	p.8

Baner ac Amserau Cymru

09/01/1889	p.7
18/07/1866	p.5

Birmingham Daily Post

04/02/1895	p.7
24/02/1860	p.2
29/12/1870	p.2

Brighton Patriot and South of England Free Press

16/08/1836	p.1

Caledonian Mercury

20/08/1832	p.4
08/12/1800	p.2
13/08/1801	p.3
30/12/1813	p.4
23/08/1819	p.3
11/05/1820	p.2
27/12/1828	p.2

Cobbett's Weekly Political Register

21/06/1834	p.19
01/04/1815	p.3
23/03/1822	p.6
11/12/1830	p.2
23/06/1832	p.16

Daily News

06/11/1866	p.2

Glasgow Herald

26/04/1893	p.7
24/09/1821	p.1
12/05/1856	p.1

Hampshire Telegraph and Sussex Chronicle

22/02/1830	p.1
29/08/1803	p.2

Jackson's Oxford Journal

02/09/1865	p.2
23/04/1808	p.1

Liverpool Mercury

01/03/1900	p.6
26/03/1824	p.5

Lloyd's Weekly London Newspaper

11/04/1847	p.1
28/07/1844	p.9

Lloyd's Weekly Newspaper

28/12/1856	p.3
03/03/1850	p.3
16/04/1854	p.1

Manchester Times

27/04/1872	p.5

North Wales Chronicle

18/02/1860	p.6
12/04/1831	p.1
18/08/1835	p.2

Northern Echo

06/10/1886	p.2

Preston Chronicle

27/08/1831	p.3

Reynolds's Newspaper

15/01/1854	p.14

Reynolds's Weekly News

01/12/1850	p.2

The Aberdeen Journal

03/07/1805	p.1

The Belfast News Letter

30/08/1900	p.1
18/11/1831	p.4

The Bristol Mercury and Daily Post

16/05/1888	p.1
04/11/1880	p.5
04/09/1885	p.5

The Champion and Weekly Herald

29/01/1837	p.12

The Charter

17/03/1839	p.2

The Chartist

02/03/1839	p.3

The Chartist Circular

30/01/1841	p.1

The Derby Mercury

06/01/1875	p.7
30/04/1807	p.2

The Era

16/04/1887	p.1
20/11/1853	p.3
03/01/1858 (supplement)	p.1
27/02/1859	p.9
24/08/1862	p.11
17/05/1863	p.16
27/03/1864	p.1
13/12/1874	p.14

The Examiner

12/07/1879	p.10
01/04/1810	p.1
25/10/1812	p.15
06/11/1814	p.13
22/02/1818	p.2
16/08/1818	p.10
23/02/1823	p.6
27/09/1829	p.14
11/12/1836	p.1
10/09/1837	p.16
25/04/1841	p.8
11/11/1843	p.4
20/06/1846	p.12
08/12/1849	p.5
22/11/1851	p.8
20/03/1852	p.5
25/07/1857	p.12
11/01/1868	p.12

The Graphic

22/05/1897 (supplement)	p.9
21/10/1876	p.22
08/11/1879	p.16

The Hull Packet

10/07/1804	p.1

The Hull Packet and East Riding Times

03/04/1885	p.7

The Illustrated Police News

26/07/1884	p.4

The Ipswich Journal

10/11/1810	p.4

The Leeds Mercury

01/11/1886	p.6
25/10/1845 (supplement)	p.6
27/01/1855	p.6

The Morning Chronicle

04/04/1860	p.7
22/09/1809	p.3
12/04/1811	p.1
08/10/1816	p.2
04/08/1817	p.4
22/04/1833	p.2
20/11/1838	p.1
09/05/1839	p.1
30/05/1840	p.3
21/01/1848	p.4

The Newcastle Courant

21/04/1871	p.2
10/06/1826	p.3
08/09/1827	p.1

The Northern Star and Leeds General Advertiser

05/02/1842	p.33

The Northern Star and National Trades' Journal

29/01/1848	p.21

The Odd Fellow

27/04/1839	p.3

The Operative

20/01/1839	p.11

The Poor Man's Guardian

07/01/1832	p.4

The Pall Mall Gazette

25/05/1899	p.3
18/04/1865	p.11
18/06/1866	p.6
29/07/1867	p.6
28/08/1868	p.11
23/03/1869	p.2
30/06/1871	p.10
14/03/1872	p.4
27/01/1873	p.9
24/06/1874	p.4
04/10/1875	p.8
17/10/1877	p.1
09/07/1878	p.11
23/02/1880	p.2
02/07/1881	p.20
24/02/1882	p.10
05/07/1883	p.10
06/03/1884	p.3
03/05/1890	p.5
19/06/1891	p.2
29/04/1892	p.4
21/07/1894	p.2
18/11/1896	p.8
05/12/1898	p.2

The Southern Star and London and Brighton Patriot

01/03/1840	p.4

Trewman's Exeter Flying Post

02/10/1806	p.3

Trewman's Exeter Flying Post or Plymouth and Cornish Advertiser

19/11/1892	p.1
30/05/1822	p.3
14/07/1825	p.3
06/12/1855	p.2
23/10/1861	p.1

Western Mail

24/06/1875	p.8

Y Genedl Cymreig

02/05/1878	p.5

Y Goleuad

26/04/1884	p.3

10.4 Issues in the CNNE matching corpus

This table shows the full list of issues from which were selected the articles which make up the CNNE matching corpus, introduced in section 4.2. For a key to the abbreviations, see appendix 10.1.

N°	Issues
1	BDPO 17/01/1879 p.04
2	BDPO 18/01/1879 p.08
3	BDPO 27/01/1879 p.08
4	BDPO 03/02/1879 p.05
5	BDPO 31/01/1879 p.07
6	BDPO 04/02/1879 p.08
7	BDPO 24/02/1879 p.04
8	BDPO 11/02/1879 p.05
9	DNLN 26/01/1846 p.06
10	DNLN 26/01/1846 p.06
11	DNLN 16/02/1846 p.04
12	DNLN 07/02/1846 p.06
13	DNLN 21/02/1846 p.06
14	DNLN 01/04/1846 p.04
15	DNLN 26/02/1846 p.06
16	DNLN 18/03/1846 p.05
17	DNLN 03/04/1846 p.06
18	DNLN 31/03/1846 p.04
19	DNLN 13/02/1879 p.05
20	DNLN 15/02/1879 p.06
21	DNLN 12/02/1883 p.03
22	EXLN 02/01/1831 p.09
23	EXLN 23/01/1831 p.13
24	EXLN 30/01/1831 p.01
25	EXLN 20/03/1831 p.10-11
26	EXLN 05/01/1834 p.10
27	EXLN 28/08/1831 p.12-13
28	EXLN 06/11/1831 p.06-08
29	EXLN 08/04/1832 p.11
30	LVMR 09/07/1892 p.05
31	LVMR 14/07/1892 p.06
32	LVMR 20/07/1892 p.05
33	LVMR 28/07/1892 p.06
34	LVMR 30/07/1892 p.06
35	LVMR 02/08/1892 p.05
36	LVMR 06/08/1892 p.06
37	LVMR 24/08/1892 p.08
38	LINP 22/01/1843 p.03
39	LINP 13/08/1843 p.07
40	LINP 22/01/1843 p.04
41	LINP 05/03/1843 p.05
42	LINP 12/03/1843 p.04
43	LINP 02/04/1843 p.05
44	LINP 21/05/1843 p.08
45	LINP 28/05/1843 p.05
46	LINP 30/07/1843 p.05
47	LINP 29/05/1887 p.06
48	LINP 05/06/1887 p.01
49	LINP 05/06/1887 p.07
50	LINP 12/06/1887 p.12
51	LINP 07/08/1887 p.07
52	LINP 11/09/1887 p.01
53	LINP 27/11/1887 p.01
54	LINP 27/11/1887 p.07
55	LINP 08/01/1888 p.07
56	LEMR 15/01/1880 p.05
57	LEMR 24/08/1833 p.08
58	LEMR 13/11/1830 p.03
59	LEMR 20/08/1831 p.04
60	LEMR 17/03/1832 p.03
61	LEMR 22/01/1880 p.05
62	LEMR 13/05/1887 p.03
63	LEMR 26/01/1880 p.05
64	LEMR 06/02/1880 p.08
65	LEMR 11/02/1880 p.08
66	MCLN 19/05/1830 p.03
67	MCLN 22/05/1830 p.03
68	MCLN 24/05/1830 p.03
69	MCLN 09/05/1831 p.04
70	MCLN 17/06/1830 p.03
71	MCLN 15/01/1831 p.03
72	MCLN 17/01/1831 p.04
73	MCLN 17/01/1831 p.04
74	MCLN 14/02/1831 p.04
75	MCLN 28/04/1831 p.04
76	NREC 02/09/1876 p.03
77	NREC 13/09/1876 p.04
78	NREC 16/09/1876 p.03
79	NREC 18/09/1876 p.03
80	NREC 19/09/1876 p.04
81	NRSR 06/01/1838 p.06
82	NRSR 20/01/1838 p.01
83	NRSR 05/05/1838 p.04
84	NRSR 09/06/1838 p.03
85	NRSR 22/09/1838 p.08
86	NRSR 12/01/1839 p.06
87	NRSR 13/04/1839 p.03
88	NRSR 20/04/1839 p.03
89	PMGU 03/09/1831 p.06
90	PMGZ 23/07/1886 p.01
91	PMGZ 31/07/1886 p.05
92	PMGZ 02/08/1886 p.05
93	PMGZ 14/08/1886 p.05
94	PMGZ 16/08/1886 p.04
95	PMGZ 20/08/1886 p.10
96	PMGZ 03/09/1886 p.01
97	PMGZ 25/09/1886 p.01
98	RDNP 05/05/1850 p.03
99	RDNP 12/05/1850 p.04
100	RDNP 26/05/1850 p.03
101	RDNP 07/07/1850 p.02
102	RDNP 07/07/1850 p.01
103	RDNP 18/06/1882 p.01
104	RDNP 25/06/1882 p.06
105	RDNP 29/10/1882 p.01
106	RDNP 13/08/1882 p.08
107	RDNP 02/07/1882 p.06

10.5 List of rules (for VARD corrections)

Two lists of rules were used for the VARD corrections described in section 5.3.2. VARD rules specify whether they are allowed to apply to the beginning, middle or end of a word, or whether they can apply to all of these positions. For both lists, all rules can apply anywhere, except the ‘insert space’ rule which can only be applied to the middle of a word. The lists are shown below.

Long list					
N	Type of rule	Rule	N	Type of rule	Rule
1	**Deletion**	1	41	**Substitution**	i > s
2	**Deletion**	a	42	**Substitution**	i > t
3	**Deletion**	b	43	**Substitution**	ii > i
4	**Deletion**	c	44	**Substitution**	ii > in
5	**Deletion**	e	45	**Substitution**	in > m
6	**Deletion**	f	46	**Substitution**	l > d
7	**Deletion**	h	47	**Substitution**	l > h
8	**Deletion**	i	48	**Substitution**	l > i
9	**Deletion**	l	49	**Substitution**	l > ll
10	**Deletion**	n	50	**Substitution**	l > t
11	**Deletion**	o	51	**Substitution**	li > h
12	**Deletion**	r	52	**Substitution**	mn > n
13	**Deletion**	s	53	**Substitution**	n > m
14	**Deletion**	t	54	**Substitution**	n > u
15	**Deletion**	u	55	**Substitution**	ni > m
16	**Deletion**	v	56	**Substitution**	o > c
17	**Insertion**	space	57	**Substitution**	o > e
18	**Substitution**	i > n	58	**Substitution**	o > n
19	**Substitution**	a > n	59	**Substitution**	o > ou
20	**Substitution**	a > s	60	**Substitution**	o > s
21	**Substitution**	a > u	61	**Substitution**	oo > ou
22	**Substitution**	b > h	62	**Substitution**	ooo > oo
23	**Substitution**	be > h	63	**Substitution**	r > n
24	**Substitution**	c > ch	64	**Substitution**	s > a
25	**Substitution**	c > e	65	**Substitution**	s > n
26	**Substitution**	ci > ch	66	**Substitution**	si > sh
27	**Substitution**	e > a	67	**Substitution**	t > f
28	**Substitution**	e > c	68	**Substitution**	t > ht
29	**Substitution**	e > ee	69	**Substitution**	t > n
30	**Substitution**	e > o	70	**Substitution**	t > r
31	**Substitution**	e > s	71	**Substitution**	t > th
32	**Substitution**	f > s	72	**Substitution**	tb > th
33	**Substitution**	gi > gh	73	**Substitution**	ti > n
34	**Substitution**	gt > ght	74	**Substitution**	ti > th
35	**Substitution**	gt > gth	75	**Substitution**	tt > t
36	**Substitution**	i > a	76	**Substitution**	u > n
37	**Substitution**	i > e	77	**Substitution**	v > w
38	**Substitution**	i > l	78	**Substitution**	v > y
39	**Substitution**	i > n	79	**Substitution**	wv > v
40	**Substitution**	i > r			

Short list		
N	Type of rule	Rule
1	**Deletion**	a
2	**Deletion**	e
3	**Deletion**	i
4	**Deletion**	l
5	**Deletion**	o
6	**Deletion**	r
7	**Deletion**	s
8	**Deletion**	t
9	**Insertion**	space
10	**Substitution**	b > h
11	**Substitution**	c > e
12	**Substitution**	e > s
13	**Substitution**	f > s
14	**Substitution**	ii > i
15	**Substitution**	mn > n
16	**Substitution**	o > e
17	**Substitution**	u > n
18	**Substitution**	wv > v