The Written British National Corpus 2014:

Design, compilation and analysis

Abi Hawtin

ESRC Centre for Corpus Approaches to Social Science

Department of Linguistics and English Language

Lancaster University

A thesis submitted to Lancaster University for the degree of
Doctor of Philosophy in Linguistics

December 2018

Table of Contents

Declaration

This thesis has not been submitted in support of an application for another degree at this or any other university. It is the result of my own work and includes nothing that is the outcome of work done in collaboration except where specifically indicated. Many of the ideas in this thesis were the product of discussion with my supervisor Professor Tony McEnery.

Abi Hawtin (BA, MA, PhD)

Lancaster University, UK

Abstract

The ESRC-funded Centre for Corpus Approaches to Social Science at Lancaster University (CASS) and the English Language Teaching Group at Cambridge University Press (CUP) have collaborated to compile a new, publicly accessible corpus of contemporary Written British English, known as the Written British National Corpus 2014 (Written BNC2014). The Written BNC2014 is an updated version of the Written British National Corpus (Written BNC1994) which was created in the 1990s. The Written BNC1994 is often used as a proxy for present day British English, so the Written BNC2014 has been created in order to allow for both comparisons between the two corpora, and also to allow for research on British English to be carried out using a state-of-the-art *contemporary* data-set. The Written BNC2014 contains approximately 90 million words of written British English, published between 2010-2018, from a wide variety of genres. The corpus will be publicly released in 2019.

This thesis presents a detailed account of the design and compilation of the corpus, focusing on the very many challenges which needed to be overcome in order to create the corpus, along with the solutions to these challenges which were devised. It also demonstrates the utility of the corpus, by presenting a diachronic comparison of academic writing in the 1990s and 2010s, with a focus on the theory of colloquialisation.

This thesis, whilst not a Written BNC2014 user-guide, presents all of the decisions made in the design and creation of the corpus, and as such, will help to make the corpus as useful to as many people, for as many purposes, as possible.

List of Tables

List of Figures

List of Appendices

Acknowledgements

Completing this thesis would not have been possible without the help and encouragement of the many people around me over the last 4 years. Firstly, I must thank my supervisors, both past and present. Dr Andrew Hardie and Professor Tony McEnery have both helped me immensely throughout this process. They have provided guidance, technical expertise, and constructive criticism, but most importantly they gave me time and encouragement when I needed it. I must also thank Dr Vaclav Brezina who, in many ways, has been an unofficial supervisor during the final year of this process. He has helped me not only to finish this thesis, but also to continue moving towards finishing the construction of the corpus.

The Written BNC2014 project would also not have been possible without the many contributions we received from both the public and UK publishing companies. My sincere thanks goes to both Dunedin and John Benjamins publishing companies, and also to every person who contributed data to the corpus.

I must also mention the various members of the Written BNC2014 project team, both past and present. All of the interns, technical staff, and research assistants who have worked with me on this project have been fantastic colleagues, and wonderful friends. A special mention must go to Mathew Gillings and Isolde Van Dorst, for keeping my spirits high, and providing an understanding ear when I was feeling frustrated.

Finally, none of this would have been possible without the support of my family and friends. My fellow CASS PhD students, Gillian Smith, Jennifer Hughes, Robbie Love, and Rosie Knight, have provided advice, encouragement, understanding, and plenty of laughs along the way. My parents and sister always gave me a place where I could forget the PhD for a while, and where I knew they would be proud regardless of the outcome. Most importantly, a huge thank you to my fiancé, Seb. He has put up with endless conversations about corpus linguistics and the frustrations of writing a thesis, and has never once complained. His support has been utterly invaluable, and I could not have done this without him.

Chapter 1: Introduction

1.1 The British National Corpus 2014

The ESRC-funded Centre for Corpus Approaches to Social Science at Lancaster University (CASS; see appendix A for a consolidated list of acronyms used in this thesis) and the English Language Teaching Group at Cambridge University Press (CUP) have collaborated to compile a new, publicly accessible corpus of contemporary British English, known as the British National Corpus 2014 (BNC2014)[1]. British English refers here to the language produced by native speakers of the variety of English spoken in either England, Scotland, Wales, or Northern Ireland. Of course, native speaker is also a condition which needs defining: for the purposes of this thesis native speaker is used to mean a person whose first language is English, which they learnt in a British context (as previously defined). On some occasions throughout this thesis, native speaker status is also self-defined by the people who contributed data to the corpus (see chapter 7 for examples of this in the context of e-langauge; see section 4.3.1 for more information about population definition). The corpus contains both spoken and written data. The spoken section of the corpus (the Spoken BNC2014) has already been compiled and released (Love et al., 2017a). This thesis outlines the challenges and solutions in the design and compilation of the written component of the corpus (the Written BNC2014), and also presents an analysis of some of the data. The BNC2014 is an updated version of the British National Corpus which was created in the 1990s (henceforth known as the BNC1994; see section 1.3 for more information on the BNC1994). The BNC1994 is often used as a proxy for present day British English (see section 1.3), so the

[1] The project was supported by ESRC grants no. EP/P001559/1 and ES/K002155/1.

BNC2014 has been created in order to allow for both comparisons between the two corpora, and also to allow for research on British English to be carried out using a state-of-the-art *contemporary* data-set (I return to this issue in section 1.3). The Written BNC2014 contains approximately 90 million words of written British English, published between 2010-2018, from a wide variety of genres. The corpus will be publicly released in 2019.

In this thesis I give a thorough account of the very many important challenges and decisions made in the design and compilation of the Written BNC2014, along with an analysis of the data to illustrate the corpus' potential for the research community. A running theme throughout this thesis is the need for a balance between what is *ideal* in a project such as this, and what is *possible*. Thus, many of the decisions which I discuss throughout this thesis centre on reaching a compromise between the desires of a corpus linguist and the time and budget constraints of the project; all of these compromises are laid out transparently throughout the thesis in the hope that users of the corpus will assess for themselves the impact that these compromises may have on their research. Whilst this thesis is not a Written BNC2014 user guide[2], it is a detailed and thorough account of all of the careful decisions which I made during the design and creation of the corpus, and should be read by all users of the corpus. At the time of submitting this thesis, data collection is not yet fully complete. Thus, some of the numbers quoted in this thesis may differ slightly from the numbers in the corpus when released.

[2] See Love et al. (2017b) for the BNC2014 user manual and reference guide

1.2 Distinguishing between spoken and written language

As stated, the BNC2014 will be split into two sections: the spoken BNC2014 and the Written BNC2014. Whilst the distinction between ‘spoken’ and ‘written’ language may at first seem to be a straightforward one, i.e. that spoken language is delivered orally whereas written language is delivered graphically, this thesis will demonstrate that this is in fact not the case. In section 2.2 I note that the CRFC labels scripts, text messages, and online fora as spoken language, whilst these genres are considered written language in the Written BNC2014 (and in many other corpora). Furthermore, in section 8.4 I highlight the difficulty encountered when considering whether Hansard texts are written language (and included in the Written BNC2014) or spoken language. Thus, it seems clear from these examples that whilst there are prototypical, easily-defined members of the written and spoken mediums (e.g. a fiction book and a spontaneous telephone conversation respectively), there is a significant degree of overlap when it comes to the representation of one medium in the form of another (e.g. representing speech in written form, as in a script for a play). I will consider how to resolve this overlap here, in order to clearly define boundaries for the types of language to be included in the Written BNC2014.

Nencioni (1983 [1976]; cited in Zago, 2016) addresses this problem by drawing a distinction between ‘spoken speech’ and ‘written speech’, with the former being a spontaneous conversation and the latter being spoken dialogue in books, or play scripts. Nencioni seems to suggest that ‘written speech’ is its own medium of language, separate from ‘spoken speech’ and writing, which perhaps suggests that a separate corpus would be needed for this type of writing.

However, Gregory (1967: 189) neatly captures this blurred boundary in a diagram (see figure 1a) which could be used to clarify this issue further. In this model, language is split into speech and writing, with speech being further divided into spontaneous and non-spontaneous language. Non-spontaneous speech can be split into ‘reciting’ and ‘the speaking of what is written’. ‘The speaking of what is written’ is also connected, in this model, to written language, and this is where the blurred boundary between the two mediums occurs. However, using this model one may infer that speech and writing can be distinguished at the point of delivery. It is possible to arrive at any of the three types of writing (‘to be spoken as if not written’, ‘to be spoken’, and ‘not necessarily to be spoken’) from a starting point of either speech or writing. It is the first distinction between oral or graphical delivery which distinguishes the types of writing. Thus, a play script would be considered spoken language (‘the speaking of what is written to be spoken as if not written’) when receiving the language orally (as when watching the play performed, for example), and considered written language (‘writing to be spoken as if not written’) when reading the script.

Thus, in corpus construction, we could use this model to define any text which is recorded (i.e. delivered ‘orally’) and then transcribed as spoken language, and any text which is collected in written form (e.g. a play script, rather than recording and transcribing a performance) as written language. This is a definition which would seem to work well within the constraints of the present project. The Spoken BNC2014 (Love et al., 2017a) has already been created and only contains transcripts of spontaneous, recorded speech. The Written BNC2014 will include language which was collected in written form. Although this does not resolve the blurred boundary between these types of language completely, it does provide boundaries for what can and cannot be considered written language for the purposes of the present corpus. This discussion will be returned to on several occasions throughout this thesis as individual genres are discussed.

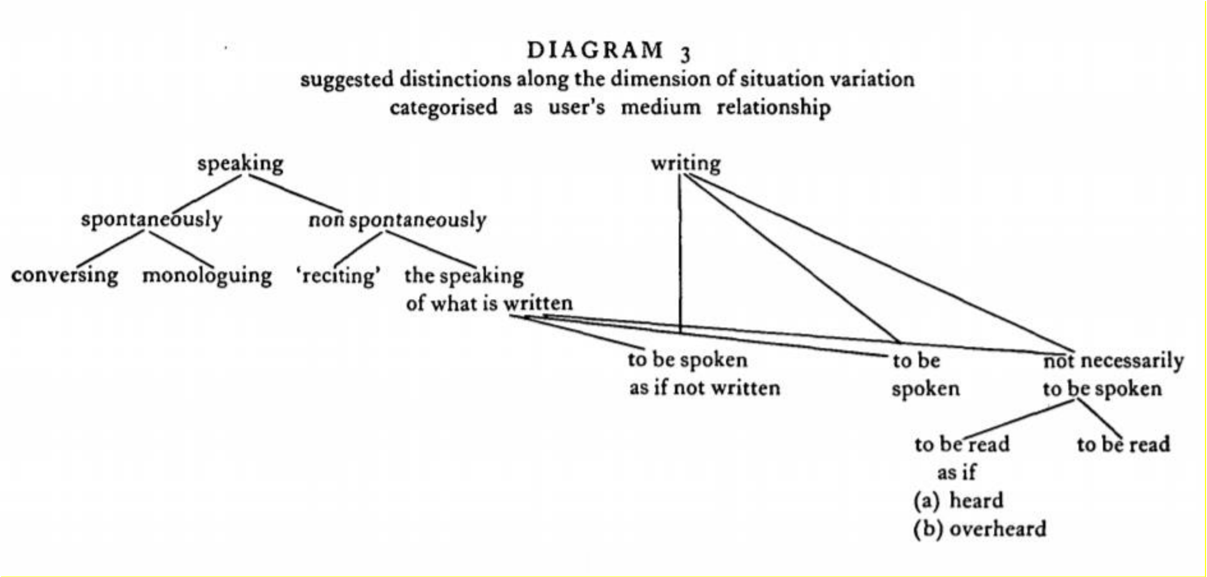


DIAGRAM 3
suggested distinctions along the dimension of situation variation
categorised as user's medium relationship
speaking
spontaneously
non spontaneously
conversing
monologuing
'reciting'
the speaking of what is written
writing
to be spoken as if not written
to be spoken
not necessarily to be spoken
to be read as if
(a) heard
(b) overheard
to be read

Figure 1a: "Suggested distinctions along the dimension of situation variation categorised as user's medium relationship." (Gregory, 1967:189).

1.3 Justifying the Written BNC2014 project

1.3.1 Introduction

In this section I will justify the *need* for the Written BNC2014 project, despite the very many contemporary corpora of British English which have been created in the years since the BNC1994 was created. I begin, in section 1.3.2, by giving some detail about the BNC1994 project, specifically the goals which the creators had in mind when designing and compiling the corpus. In section 1.3.3 I discuss the enduring popularity of the BNC1994 by highlighting the many areas of research in which the BNC1994 has been, and still is being, used. I then consider why this is, when so many other corpora of British English have been created since. In section 1.3.4, I argue that the BNC1994 continues to be used so frequently, despite its age, because none of the corpora which contain more contemporary data meet all of the same goals which the BNC1994 does. I conclude, in section 1.3.5, that a new version of the BNC1994 is needed in order to allow all of the research which was fostered by the BNC1994 to continue, but with data and results which are truly representative of *contemporary* British English.

1.3.2 The British National Corpus (BNC1994)

A corpus which aims to be representative of the language used in a particular national community is known as a national corpus (e.g. the Czech National Corpus, the Thai National Corpus, and the American National Corpus; see chapter 2). For example, the BNC1994 is a 100 million word corpus of written and spoken British English which has been described as the "first and best-known national corpus" (Xiao, 2008: 384). It was compiled in 1990-1994 (Burnard, 2002), although some texts within the corpus date back as far as the 1960s (Burnard, 2000). The project to create the BNC1994 brought together dictionary publishers, the British library, and Lancaster and Oxford Universities (Burnard, 2002). This consortium worked toward several goals which, if achieved, would make the BNC1994 unique (Burnard, 2002). These goals were:

- To create a corpus an order of magnitude larger than any currently freely available corpus.
- To create a synchronic corpus.
- To create a corpus of contemporary language.
- To include a range of samples from the full range of both spoken and written British English.
- To create the corpus using a non-opportunistic design.

- To include automatic word class annotation, and detailed contextual information.
- To make the corpus generally available.

  (Burnard, 2002: 53).

The creators of the BNC1994 did indeed achieve all of these goals, which is likely the reason why its popularity as a data set for research endures to the present day (an issue which I return to in sections 1.3.3 and 1.3.4).

The goal of creating the corpus using a non-opportunistic design, where target amounts or types of texts are set out prior to data collection commencing, was particularly important to the creators of the BNC1994. This is because, at the time that the BNC1994 was created, it was not common for texts to be digitised prior to their publication. This meant that corpus creators at that time tended to include in their corpora only those texts which had been digitised, without much consideration given to what those texts actually represented (Burnard, 2002: 57). In contrast to the norms at the time, the BNC1994 creators established a set of design criteria at the outset of the project, which proposed target text characteristics and proportions. Thus, the creators had to seek out texts which fitted their criteria, rather than taking just those texts which were readily available in a digitised format. Burnard (2002: 57) claims that this allowed the BNC1994 to be used to “say something about language in general”, which had not been possible using many previous corpora. The eventual composition of the BNC1994 is shown in tables 1a and 1b. The design and composition of the BNC1994 will be discussed in more detail throughout this thesis where relevant.

Table 1a: Composition of BNC World Edition (Burnard, 2000).

Text type	**Texts**	**W-units**	**S-units**	**Percent**
Spoken demographic	153	4.30	610563	10.08
Spoken context-governed	153	6.28	428558	7.07
All spoken	910	10.58	1039121	17.78
Written books and periodicals	2688	80.49	4403803	72.75
Written-to-be-spoken	35	1.35	120153	1.98
Written miscellaneous	421	7.55	490016	8.09
All written	3144	89.39	5013972	82.82

Note: The BNC World Edition was the second edition of the corpus, with some improvements made to the tagging over the first edition of the corpus

Table 1b: Domains in the Written BNC1994 (Burnard, 2000).

Domain	**Texts**	**W-units**	**Percent**	**S-units**	**Percent**
Applied Science	370	7104635	8.14	357067	7.12
Arts	261	6520634	7.47	321442	6.41
Belief and thought	146	3007244	3.44	151418	3.01
Commerce and finance	295	7257542	8.31	382717	7.63
Imaginative	477	16377726	18.76	1356458	27.05
Leisure	438	12187946	13.96	760722	15.17
Natural and pure science	146	3784273	4.33	183466	3.65
Social science	527	13906182	15.93	700122	13.96
World affairs	484	17132023	19.62	800560	15.96

The creators of the BNC1994 originally thought that the corpus would only be of interest to a few researchers – those working in Natural Language Processing (NLP) or lexicographers (Burnard, 2002: 67). However, it rapidly became clear that this was not the case; as we now know, the main users of the BNC1994 would be those working in applied linguistics, particularly researchers concerned with language learning and teaching (Burnard, 2002: 67; see section 1.3.3).

1.3.3 The enduring popularity of the BNC1994

In this section I will show that, despite being created in the 1990s and containing data from as far back as the 1960s, the BNC1994 is still extremely widely used in linguistic research to this day. This is perhaps surprising because the BNC1994 certainly no longer represents *contemporary* British English, and yet it is being used as though it does.

A simple search for the term *"British National Corpus BNC"* in Lancaster University's online library catalogue yields 1,845 results (although this figure does include some repeats). These are mostly articles, but there are also several books and conference proceedings. Almost 60% of these results were works which were published from 2010 onwards, and 11% of these results were published in 2017 or 2018. This shows that the BNC1994 continues to be a very productive data source for research right up to the present day.

A more specific example of just how productive the BNC1994 still is comes from the abstract book for the ICAME36 conference held in 2015 (ICAME36, 2015). 20 of the research papers, posters, and works in progress which were presented at the conference used the BNC1994 as a data source. Many of these works used the BNC1994 as a "present day" corpus despite its age. Several pieces of research used the BNC1994 as comparable to the Corpus of Contemporary American English (COCA; Davies, 2013b), despite the fact that COCA contains data collected as recently as 2017. For example, Smith (2015) compares the use of temporal phrasal adverbials, such as *the moment*, in British and American English using COCA and the BNC1994.

As highlighted by the two examples, above, the BNC1994 has been, and continues to be, an extraordinarily productive source of data for researchers from many disciplines within linguistics. There has been far too much research for a comprehensive discussion of it all within this thesis. Thus, I will briefly highlight three areas of linguistics where the BNC1994 has been widely used: language teaching research, discourse analysis, and grammar research.

Use of the BNC1994 has made language teaching a particularly productive area of research in linguistics because "[c]omputational linguistics and corpus linguistics enable people to look beyond the word level into the chunk of language, which is actually the key to develop writing competence" (Sha, 2010: 390). Many researchers use the BNC1994 to show how useful corpora can be for language pedagogy. Chujo and Utiyama (2006) apply various statistical measures to the 'commerce and finance' section of the BNC1994 in order to see if a level-specific list of technical vocabulary can be generated for learners. This research is based on the idea that having students use word lists can speed up vocabulary acquisition and expansion (Chujo and Utiyama, 2006: 256). However, previous research (Thorndike and Lorge, 1944; Harris and Jacobson, 1972; Engels et al., 1981) has used objective measures such as frequency, or subjective measures such as teacher's intuitions to generate vocabulary lists for learners (Chujo and Utiyama, 2006: 256). Chujo and Utiyama (2006) use a 2973 word list from the BNC1994, and apply statistical measures to compile vocabulary lists for various levels of language learner. They find that the cosine is an effective measure for identifying beginner-level basic business words, the log-likelihood and chi-square tests are effective at extracting intermediate-level business words, and the mutual information and McNemar's test are effective at generating lists of advanced-level business words (Chujo and Utiyama, 2006: 255). This study shows that corpora such as the BNC1994 are a valuable tool for "automatically extract[ing] various types of specialized lists that can be quickly and accurately targeted to learners' vocabulary or proficiency levels" (Chujo and Utiyama, 2006: 266).

Some researchers, such as Sha (2010), use the BNC1994 as a basis of comparison when trying to improve methods of language teaching. Sha (2010) compares the effectiveness of the use of traditional corpora, such as the BNC1994, to the use of the search engine Google in language learning. Sha finds that search-engine based corpora are more effective than traditional corpora. This is because Google returns many more results than the BNC1994. Providing learners with many examples is advantageous because it allows them to see the phrase they are learning about in many different contexts and varieties of English. Also, Google was found to perform searches faster than the BNC1994, and Google's built in spellchecker aids beginners whose spelling is not as advanced, whereas the BNC1994 cannot do this (Sha, 2010: 391). However, this is a category error; Sha is equating the software used to perform searches on the BNC1994 with the actual data contained within the corpus. The fact that searches are slower and the software does not incorporate a spellchecker has nothing to do with how useful the actual *data* contained within the BNC1994 is to learners. Despite this, this study shows that even in research where the BNC1994 is not used directly in language teaching, it is helping to improve teaching methods by being used as a comparison tool.

For more language teaching research using the BNC1994 see: Hsu (2013), Lin (2014), Zhao and Feng (2014), Perez-Paredes et al. (2011), Cheng et al. (2003), Bartley and Benitez-Castro (2013), and Siyanova and Schmitt (2008).

Another area where the BNC1994 has been used widely in research is discourse analysis. Corpora can be very useful tools in discourse analysis because they allow “researchers to objectively identify widespread patterns of naturally occurring language and rare but telling examples, both of which may be over-looked by a small-scale analysis” (McEnery and Baker, 2005:198). Some researchers, such as Norberg (2012), use the BNC1994 in order to identify discourses around particular topics or events. Norberg (2012) uses the BNC1994 to examine the discourses around male and female shame. Norberg searches the BNC1994 for singular and plural forms of the noun ‘shame’, all of the inflected forms of the verb ‘shame’, and the adjectives ‘ashamed’, ‘shameless’, and ‘unashamed’. Of the 435 results found, 159 references to ‘shame’ were able to be coded for gender (Norberg, 2012: 163). There were roughly the same amounts of instances relating to men and women, which may lead one to conclude that men and women are “equally shame-prone in the corpus” (Norberg, 2012: 164). However, upon further qualitative analysis Norberg (2012: 165) finds that men and women are ‘shame-prone’ in very different situations. Men feel shame as a result of ‘nonacheivements’ and ‘physical weakness’, whereas women do not. In contrast, women feel shame due to ‘personal shortcomings’ and ‘exposure’ whereas men do not. This study exemplifies how quantitative corpus methods can complement qualitative discourse analysis methods in order to give a more in depth analysis.

Other discourse analysis researchers, for example McEnery and Baker (2005), have used the BNC1994 as an example of British English to which their data can be compared. McEnery and Baker (2005) use a corpus-based approach to examine the discourses surrounding refugees and asylum seekers in a corpus of newspaper articles and a corpus of articles published by the UN. They choose to use the BNC1994 as a third corpus in their analysis because "it can reveal normative patterns of language use which can then be compared against the findings in the two more specific corpora" (McEnery and Baker, 2005: 200). In this way, the BNC1994 helps to reveal how movement descriptors construct discourses around refugees. Refugees are found to often be described as *streaming, overflowing* and *swelling* in the newspaper corpus. All of these words are found to occur in negative contexts in the BNC1994; *streaming* collocates with *tear, water,* and *rain*; *overflowing* collocates with *leaking* and *water*; and *swell* collocates with words connected to water. McEnery and Baker (2005: 204) conclude that these collocations lead to refugees being "constructed as a 'natural disaster' like a flood, which is difficult to control as it has no sense of its own agency" (McEnery and Baker, 2005: 204). This study highlights the utility of the BNC1994 as a point of reference for 'normal' language.

For more examples of discourse analysis research using the BNC1994 see: Wang (2005), Poole (2015), O'Halloran (2009), Steen et al. (2010), Dilts and Newman (2006), Yamasaki (2008), and Pearce (2008).

The BNC1994 has also been used in grammar research, as a way of examining the existence or usage of various grammatical constructions. Schonefeld (2013) uses the BNC1994 to research whether the English *go un-V-en* construction varies depending on the register it is used in. Schonefeld (2013: 17) first extracts all examples of this construction from the BNC1994 and correlates these results with the more general *go adjective* pattern. Next, Schonefeld reduces these instances to those occurring in the academic prose, newspaper, fiction, and conversation categories of the BNC1994, and finds that register has a big effect on the usage of the *go adjective* pattern:

> Academic prose has a remarkable number of attributive readings of the general go adjective pattern; in newspaper texts the different (sub-) constructions are more evenly distributed, with three readings represented by the top 4 ranking collexemes. Fiction noticeably favours the resultative pattern (17 collexemes, among them the first 12 ranks) and conversation is a mixed bag indeed, though it does not have depictives. (Schonefeld, 2013: 29)

Schonefeld (2013: 17) argues that these findings show a need for usage-based approaches to constructions to incorporate extra-linguistic factors, such as register variation.

For more examples of grammar research using the BNC1994 see: Liu (2011), Erman (2014), Breul (2014), Weichmann and Kerz (2013), Van Bogaert (2010), Berg (2011), and Tottie (1997).

This section has shown just how much research has been based on the BNC1994, continuing right up to the present day. However, the reason for its continued popularity has not yet been addressed. Many other general language corpora have been created since the BNC1994 (e.g. BE06, the Bank of English, ukWaC, ICE-GB; see section 1.3.4), and yet the BNC1994 still retains its popularity as a tool for investigating contemporary British English. I will consider why this is in section 1.3.4.

1.3.4 Other corpora of written British English

In this section I will consider some well-known general corpora of written British English which have been created since the BNC1994, and aim to assess why they have not been as widely used. To do this it will be important to return to the goals of the BNC1994 project, introduced in section 1.3.2. I argue that the reason that none of the following corpora have enjoyed the level of uptake of the BNC1994 is because none of them meet all of these goals.

1.3.4.1 The Brown Family

The Brown family consists of multiple corpora, which are all considered to be comparable in McEnery and Hardie's (2012: 20) sense of comparable corpora: "a corpus containing components that are collected using the same sampling method" (see section 3.3 for more information about comparable corpora; see section 3.3.3 for a more in depth discussion of the Brown Family). The first member of the Brown Family was the Standard Corpus of Present-Day American English (later renamed the Brown Corpus) which consists of approximately 1 million words of American English prose produced during 1961 (Francis and Kučera, 1979). The corpus contains 500 samples of 2,000 words each, with samples representing a wide range of styles and varieties. The sampling frame used to construct the corpus then became the model for all subsequent members of the Brown family which have been created (see table 1c for the sampling frame, and table 1d for all members of the Brown Family). As can be seen from table 1d, there are many members of the Brown Family, and all represent a particular language variety at a particular point in time.

For the purposes of this discussion I am interested in BE06, because it is a general corpus of written British English which was created after the BNC1994. BE06 contains 1 million words of written British English from the mid-2000s (Baker, 2009) and so represents a much more contemporary form of British English than the BNC1994. The motivation for building BE06 was similar to the motivation for building the BNC2014 – the then-current Brown family corpora did not adequately represent contemporary British English (Baker, 2009: 315). For instance, Baker (2009:315) notes that FLOB contains no instances of the words ‘internet’ or ‘www’, strongly suggesting that a new corpus was needed to properly reflect current British English.

BE06 does indeed represent contemporary British English (much more contemporary than the BNC1994), and has been used to investigate frequent linguistic phenomena (see Baker, 2009; Ramírez, 2015; Brezina and Gablasova, 2015), but why has it not been used more widely by researchers instead of the BNC1994? Whilst BE06 does meet the majority of the goals of the BNC1994 project, it falls short on one extremely important factor: size. BE06 contains just 1 million words, whereas the BNC1994 contains 100 million. Baker (2009: 314) acknowledges this issue: “It is likely to be the case that one million word samples are not large enough to reveal definitive conclusions about linguistic variation and change”, but suggests that 1 million words may be enough to examine the usage of very high frequency words. Thus, although BE06 represents a more contemporary form of British English than the BNC1994, it has been used far less because its size is too small to draw reliable conclusions from for anything other than very frequent phenomena (and, of course, because it is much newer). Furthermore, BE06 is not available to download freely. It can only be accessed via CQPweb because of concerns over copyright issues which may arise by making the corpus freely available in its entirety. Thus, BE06 does not fully meet the BNC1994’s goal of being generally available.

Table 1c: Sampling frame for the Brown family of corpora (McEnery and Hardie, 2012: 97).

Text categories	Broad Genre	No. of texts	% of corpus
A Press: reportage	Press	44	8.8
B Press: editorial	Press	27	5.4
C Press: reviews	Press	17	3.4
D Religion	General prose	17	3.4
E Skills, trades and hobbies	General prose	36	7.2
F Popular lore	General prose	48	9.6
G Belles lettres, biography, essays	General prose	75	15
H Miscellaneous (government & other official documents)	General prose	30	6
J Learned and scientific writings	Learned	80	16
K General fiction	Fiction	29	5.8
L Mystery and detective fiction	Fiction	24	4.8
M Science fiction	Fiction	6	1.2
N Adventure and western fiction	Fiction	29	5.8
P Romance and love story	Fiction	29	5.8
R Humour	Fiction	9	1.8

Table 1d: Corpora within the Brown Family.

Corpus	Language variety	Period
B-Brown	American English	1931 +/- 3 years
Brown	American English	1961
Frown	American English	1991-1992
AmE06	American English	2006 +/- 1 year
BLOB	British English	1931 +/- 3 years
LOB	British English	1961
FLOB	British English	1991-1992
BE06	British English	2006 +/- 1 year
Kolhapur	Indian English	1978
ACE	Australian English	1986
WWC	New Zealand English	1986-1990

1.3.4.2 The Bank of English

The Bank of English (BoE) corpus was initiated in 1991 as part of the Collins Birmingham University International Language Data-base (COBUILD) project (Xiao, 2008: 394). The BoE is a monitor corpus, which means that it is constantly updated in order to track rapid language change (Xiao, 2008: 394). The corpus contains both written and spoken data from British English (70%), American English (20%), and other English varieties (10%) (Xiao, 2008: 394). As of 2008, the BoE contains 524 million words (Xiao, 2008: 394).

So the BoE is larger and more up to date than the BNC1994, and yet is not as widely used. This may be because the BoE does not meet the BNC1994's goal of being a synchronic corpus. The BNC1994 aimed to provide a snapshot of British English in the 1990s, whereas the BoE is a monitor corpus which tracks language over time. The advantage of a synchronic corpus over a monitor corpus is that it can provide "a fixed point which can be referred to in years to come" (Jakubíček et al., 2013: 126). Furthermore, a synchronic corpus ensures that analyses done using the corpus are fully replicable; this replicability is lost once more data is added to a monitor corpus (unless, of course, the monitor corpus is structured so as to allow you to effectively recover earlier versions of the corpus). However, the advantage of a monitor corpus is that it remains contemporary.

The BoE also fails to meet the BNC1994's goal of being generally available. This is because the corpus was created by a publishing company (Collins) and so they understandably would not want to give away their commercial advantage. Only a 56 million word sampler of the corpus (half the size of BNC1994) is accessible for free online, whereas access to the whole corpus is only granted by special arrangement (Xiao, 2008: 395).

1.3.4.3 ukWaC

ukWaC is a 2 billion word corpus of English produced by web-crawling the .uk internet domain in 2007 (Ferraresi et al., 2008). The creators of ukWaC wanted to ensure that the corpus was representative of general English rather than just web-specific genres, and so the corpus contains texts that can be found in print as well as on the web (e.g. recipes, sermons, transcripts of spoken language etc.), and also web-specific texts (e.g. blogs and forums etc.) (Ferraresi et al., 2008).

ukWaC would seem to be an ideal candidate to replace the BNC1994 in popularity; it is extremely large, represents written and spoken British English, is modern, and is freely available. However, two of these points may be called into question: ukWaC's modernity, and its representativeness of *British* English. Firstly, although the corpus was collected in 2007, Ferraresi et al. (2008) do not give any indication of whether the age of the web pages included were taken into account. This means that, although the data in the corpus is surely more contemporary than that in the BNC1994, users cannot know exactly what time span of British English they are looking at. Furthermore, not all of the texts in the corpus may actually be representative of British English. Only crawling the .uk domain does indeed increase the likelihood of the authors writing these pages being British, however it does not ensure it. Thus, some of the texts included in ukWaC may come from speakers of different varieties of English, or non-native English speakers. This means that ukWaC does not meet one of the fundamental goals of the BNC1994: to represent spoken and written *British English.* Of course, it is also the case that any text-types which are not present online will not have been collected, further limiting the corpus' representativeness.

UkWaC also fails to meet the BNC1994's goal of being of a non-opportunistic design. A typical web crawl selects web pages entirely at random and so there is no way of knowing anything about the actual contents of the corpus. It cannot be known if the corpus is balanced for register, domain, or demographic factors, and so any conclusions drawn from the corpus would not be generalisable beyond the corpus itself.

The criticisms discussed in this section are also true for the many other web-crawled corpora of English which have been created since the BNC1994 (e.g. enTenTen; Jakubíček et al., 2013). Whilst web-crawled corpora are more contemporary, larger, and easier and quicker to create than a 'hand-made' corpus such as the BNC1994, they also do not represent a synchronic 'snapshot' of language, they cannot be guaranteed to contain the target variety of a language, and they also may not contain samples from the full variety of the language. This is not to say that web-crawled corpora are not themselves valuable resources for particular purposes, most notably purposes for which the combination of size and speed of collection is especially desirable. A corpus of the size that has taken the BNC2014 project team years to create could be created by a web-crawler in a matter of hours. However, smaller 'hand-made' corpora such as the BNC1994 and the BNC2014 have in their own sphere equally many advantages.

1.3.4.4 ICE-GB

The International Corpus of English (ICE) is a family of corpora which were designed specifically for the synchronous study of world Englishes (Xiao, 2008: 398).

It consists of 20 different corpora, each containing 1 million words of written and spoken data produced during 1990-1994 in countries where English is the first or official language (Xiao, 2008: 398).

ICE-GB also seems, in some respects, to be a good replacement for the BNC1994. Despite being produced at the same time as the BNC1994, it contains a larger proportion of contemporary data (BNC1994 contains data dating back as far as the 1960s). However, the reason why ICE-GB has not replaced the BNC1994 is that, similarly to BEO6 (see section 1.3.4.1), the corpus only contains 1 million words of data in comparison to the BNC1994's 100 million.

1.3.5 Summary and justification for the project

In this section I have described how and why the BNC1994 was created, and shown that, despite its age, the corpus is still used today as a proxy for present day British English. The BNC1994 was created in the 1990s with the aims of being a generally available, synchronic corpus, of contemporary written and spoken British English, on a scale larger than anything then available (Burnard, 2002: 53). Although many corpora of British English have been created since the BNC1994, none of them have met all of these goals, and this, I argue, is the reason for the enduring popularity of the BNC1994 despite its age: there are simply not any corpora available to researchers which are as large as the BNC1994, are synchronic, and contain the kind of carefully selected data which the BNC1994 does.

Thus, the enduring popularity of the BNC1994 despite its age implies a need in the research community for a new BNC which meets all of the same goals as the BNC1994 but contains truly contemporary British English. This is the gap which the BNC2014 will fill.

1.4 The project team and my ownership of the research

1.4.1 The Written BNC2014 project team

My PhD project, rather than being a traditional single-person project driven solely by that researcher's particular interests, is a major resource creation exercise which is externally funded and has a team of people working together on the project. As already stated, the BNC2014 project is a collaboration between CASS at Lancaster University and Cambridge University Press (CUP). However, CUP had much less involvement with the Written BNC2014 project than with the Spoken BNC2014 project, and simply acted as interested advisors to the project team. The Written BNC2014 project team consists of the following members:

Project Committee: Tony McEnery (TM), Vaclav Brezina (VB)

Technical Staff: Matt Timperley (MT), Andrew Hardie (AH)

Research Assistants: Mathew Gillings (MG), Carmen Dayrell (CD), Isolde Van Dorst (IVD), Andressa Gomide (AG)

Research Student: Abi Hawtin

Members of the team will henceforth be referred to using the initials given above. The project also benefitted from generous contributions from several publishers (Dunedin Academic Press, John Benjamins; see chapter 5), and from texts obtained through public participation in scientific research (PPSR; Shirk et al., 2012). All contributors are fully credited in the corpus documentation.

1.3.2 Ownership of research

It is important to establish firmly that, although I did not originate the idea for the project, and although I worked as part of a project team, the work contained in this thesis is my own original work. The project committee provided oversight and governance for the project, and then I was responsible for actually implementing the creation of the corpus. The role of the committee was to act as advisors, and to be involved in major decision making on the project. I was responsible for being the main decision maker and driving-force behind the project. All research (unless otherwise explicitly stated) has been done by me, with advice sought from the committee where necessary. In this sense, the committee has acted much like an extended supervisory panel, and has met regularly to discuss progress on the project.

In practice this means that the division of work on the Written BNC2014 has been as follows. The original goal of a "new" BNC corpus was a product of the committee, who identified funding and recruited me as a research student to implement the written corpus construction; I then took on full responsibility for investigating previous relevant corpus construction research to inform the new corpus, surveying relevant opinion, and developing a schema for the design of the corpus including the different genre categories. This was put to the committee for discussion, but with the ultimate decisions left to me. After this, I began the work of developing methods for collecting each section of the corpus (including, where necessary, fact-finding research on the text types in question, e.g. the research into magazines in print versus on the web; see section 6.6). The majority of manual text collection was done by me. Exceptions to this occurred when the time-constraints of the project simply would not allow for all data to be collected by one individual, and in such cases assistance was given by the research assistants (CD, MG, IVD and AG) on the project team, or from undergraduate or Masters students who interned on the BNC2014 project from time to time. These people were, however, working at my sole direction. Automated text collection (and text format management) was done according to designs which I developed but implemented by the project's technical staff (MT and AH) following the parameters I laid out (for this reason, parts of this thesis which discuss automated text collection focus on the conceptual aspects of text collection rather than the programming done to implement my decisions).

Thus, all work presented in this thesis is my own. The Written BNC2014 represents the result of decisions which I made (and which will be justified in this thesis) and text-collection efforts which I have either done entirely myself or directed others to do, with the input and advice of the project committee not going beyond the level of direction and support which any PhD student would receive from their PhD supervisor. Pronouns will be used systematically throughout this thesis to support this: first person singular pronouns are used to indicate work which was conducted solely by me, whereas first person plural pronouns, initials of project team members, and third person references to "the Written BNC2014 project team" are used when discussing decisions I made with the team, or data collection carried out by someone other than myself.

1.5 Copyright and Permissions

1.5.1 Introduction

The Written BNC2014 includes a wide variety of contemporary British English texts. This means that many texts which are protected by copyright will be included in the corpus. Thus, this section discusses the issue of copyright law, and gives some useful definitions which will be referred to throughout the thesis. Section 1.5.2 outlines the current copyright law in the UK, and discusses some exceptions which are relevant to the project. Section 1.5.3 presents some expert opinions on the issue of copyright in corpus creation. I then give some definitions of terms which will be relevant throughout this thesis, and conclude briefly by stating in broad terms how I will approach copyright law in the creation of the Written BNC2014.

1.5.2 Copyright law

The current act which covers issues of copyright in the UK is the Copyright, Designs and Patents Act 1988 (United Kingdom Intellectual Property Office, 2017). The current law gives the creators of literary works (and others not relevant to the present discussion) the right to have control over how their material is used (United Kingdom Intellectual Property Office, 2017). It is an offence to copy a work, or to "rent, lend or issue copies of the work to the public" without the consent of the copyright holder (UKCS, 2017). This will clearly be relevant to the creation of the Written BNC2014, as I will be copying authors' works to include them in the corpus and also issuing copies of these works to people who want to use the completed Written BNC2014 for research or teaching. This was an issue encountered in the creation of the American National Corpus (Ide, 2008; see section 2.5). The creators of the corpus only sought to include data which was not protected by copyright, which hugely limited the potential pool of data for the corpus, and ultimately proved to be a huge stumbling block for them in the project.

However, there are several exceptions to copyright protection, two of which are relevant to the current project: 'Non-commercial research and private study' and 'Text and data mining for non-commercial research' (Gov.uk, 2017). The Written BNC2014 is a non-commercial project, meaning that no money will be made through the licensing of the corpus, and also that those who use the corpus cannot do so for commercial purposes. This means that under the 'Non-commercial research' exception it will be acceptable for me to "copy limited extracts of works" (Gov.uk, 2017). This use must be within ‘fair dealing’ (which I will explain fully later), and there must be no financial impact on the copyright holder because of the use (Gov.uk, 2017). It is highly unlikely that there would be any financial impact on any of the copyright holders of works included in the Written BNC2014, because the eventual texts will be so heavily transformed with XML markup and word-level annotation that it is doubtful that anyone would try to read the text in the Written BNC2014 rather than the original. Furthermore, although I anticipate that the Written BNC2014 will be a very widely used resource within the fields of linguistic research and teaching, this actually represents a tiny proportion of the possible audience for most copyrightable works. Thus, most potential readers of the copyrighted texts will have no idea that they are present in the corpus, much less any idea of how to access them.

‘Fair dealing’ is “a legal term used to establish whether a use of copyright material is lawful or whether it infringes copyright” (Gov.uk, 2017). There is no formal definition of fair dealing; it is determined on a case-by-case basis (Gov.uk, 2017). It is suggested that you should ask yourself the question “how would a fair-minded and honest person have dealt with the work?” (Gov.uk, 2017). Factors which have been deemed relevant by courts in determining fair dealing include whether the use of the work affects the market for the original work (discussed above), and whether the amount of the work used was reasonable, appropriate, and necessary (Gov.uk, 2017). As discussed above, the use of works in the Written BNC2014 should not affect the market for the original, and only samples will be taken from copyrighted books (other procedures will be used for other types of text, this will be discussed later in this section), so this use should be considered to fall within the limits of fair dealing. In the creation of the Thai National Corpus (TNC), Aroonmanakun et al. (2009) were advised by Thai lawyers that their samples of 40,000 words would be too big to fall within the bounds of fair dealing. However, the samples included in the Written BNC2014 are much smaller than this (typically 5000 words; see section 4.3.2). Furthermore, it is likely the case that lawyers would err on the side of caution when advising on these matters as it is their job to protect their clients. The boundaries of fair dealing will be assessed separately for each type of text collected for the corpus (see chapters 5, 6, 7, and 8). Thus, the ‘Non-commercial research and private study’ exception has applied to the collection of texts for the Written BNC2014 project.

The other exception to copyright law which may be relevant to the project is the exception for ‘Text and data mining for non-commercial research’. Text and data mining is defined by Gov.uk (2017) as “the use of automated analytical techniques to analyse text and data for patterns, trends and other useful information”. This definition seems to describe the sort of work which will be done with the Written BNC2014, and thus permits the copying of texts to be included in the corpus. To make use of this exception, the researchers must already have lawful access to the work (e.g. own a copy of the book, have a subscription to access material etc.) (Gov.uk, 2017). The copied work must also be accompanied by a full acknowledgment of the original author where possible (Legislation.gov.uk, 2014). However, copyright of the work is infringed if the copy is transferred to another person (Legislation.gov.uk, 2014). The aim of the Written BNC2014 project is to make the corpus widely available, so this means that the text and data mining exception cannot be used for the creation of the Written BNC2014. Interestingly, this also means that any research which utilises this exception to collect data will not be replicable, as the data cannot be shared with others.

It will be possible to collect some texts for the corpus without reference to the ‘Non-commercial research’ exemption. Only academic journal articles which are available under an open access license are included in the corpus (see chapter 6). Two commonly used open access licenses are the CC BY 4.0 and CC BY-NC-ND 4.0 license. These allow a user to freely "redistribute or republish" a work and allow the reuse of "portions or extracts from the [text] in other works" (Elsevier). Including books and articles published under this type of license in the corpus is fully compliant with the terms of the license, and also allowed me to copy the entire work rather than trying to stay within the, unclear, boundaries of fair dealing.

1.5.3 Expert opinions

There is currently no case law which relates directly to the issue of copyright in creating a corpus, and so it is impossible to know how this would be viewed in a court of law. In order to get an idea of how these issues may be interpreted I contacted several experts in this area. I will briefly outline the findings here.

Firstly, Lancaster University's copyright officer was contacted for advice. She indicated that it would likely be necessary to gain permission to use all texts which will be included in the corpus, including e-language (electronic language; see chapter 7 and Knight et al., 2014). She also suggested that taking any more than just a few lines from a text may be seen as operating outside of fair dealing. Secondly, Professor Christopher May of Lancaster University, who specialises in intellectual property rights, was contacted. His response was quite the opposite; he suggested that because the corpus will be non-commercial and only used for research and educational purposes, it will not be necessary to gain permission to include *any* text in the corpus, providing that we stay within the bounds of fair dealing. Finally, the legal team at Cambridge University Press were contacted to see how an actual publisher would view these matters. They indicated that they may be able to create a license to use their works in a corpus, but that it would be very restrictive (no display or distribution of the texts would be permitted). They also indicated that for texts other than books (such as e-language), permission from the individual copyright holders must be gained before the texts can be included in the corpus.

Professor Christopher May's response should be given most weight here, not only because his interpretation provides the most freedom, but also because he is in the least biased position. It is the job of both Lancaster University's copyright officer and Cambridge University Press' legal team to protect their institutions and so the advice they give is understandably much more cautious. Therefore, I decided to take the advice of an impartial expert, Professor Christopher May, when dealing with issues of copyright throughout the project.

1.5.4 Definitions

The following definitions will be applied to these terms throughout this thesis:

The 'Non-commercial research' exception to UK copyright law: allows researchers to copy extracts of works for the purposes of non-commercial research, within the bounds of fair dealing.

Fair dealing: the amount of a work taken which does not affect the market for the original work, and which is reasonable, appropriate and necessary.

Open access license: a licence which allows a user to redistribute and republish the work.

1.5.5 Conclusion

Issues of copyright will be discussed in detail for each medium of text (see chapters 5, 6, 7, and 8), but briefly, I consider the 'Non-commercial research and private study’ exception to UK copyright law to cover the collection of any type of text which I have legal access to for the Written BNC2014 (providing that the collection is within the bounds of fair dealing). This means that, for most texts collected for the corpus, permission was not sought from copyright holders because it was deemed unnecessary to do so.

1.6 Research aims and the structure of the thesis

Section 1.3 has clearly articulated the *need* for a new version of the BNC1994, and thus, the need for the Written BNC2014 project. As such, the research aims of the Written BNC2014 project were simple: we would aim to create a widely-available corpus of contemporary written British English, which is of the same magnitude as the BNC1994. This would allow the same kind of research fostered by the BNC1994 to continue, but with results that are representative of truly contemporary British English. Furthermore, diachronic comparisons with the BNC1994 would become possible for the first time.

This thesis will detail the design and compilation of the corpus, along with presenting an analysis undertaken using the data. As such, the aims of the thesis are as follows:

(1) To survey relevant literature in the field of corpus creation, and to use this to design a sampling frame for the Written BNC2014

(2) To test and implement methods of collection for all of the data types to be included in the corpus

(3) To implement the findings of (1) and (2) in order to create the Written BNC2014

(4) To use the Written BNC2014 to carry out a novel analysis (the specific aims of which will be discussed in chapter 9)

As the focus of this thesis is expressly methodological, the standard approach to a thesis, wherein relevant literature is reviewed firstly, and then a methodological approach is outlined, before presenting an analysis in the rest of thesis, is not suitable. Instead, I take a chronological approach by firstly addressing each stage in the design process, and then discussing each medium of data collection separately. I finish by presenting an analysis using the data collected for the corpus in order to demonstrate the potential of the corpus. Rather than one, over-arching, literature review, I instead discuss relevant literature in each chapter of the thesis in order to contextualise all of the decisions detailed within. The thesis is divided into the following chapters:

- Chapter 2: Contemporary National Corpora

This chapter presents a discussion and comparison of six contemporary national corpus projects. I introduce these projects in this early chapter because a good understanding of other projects, similar to the Written BNC2014 project, will be essential in order to contextualise the decisions presented throughout this thesis.

- Chapter 3: Creating Representative and Comparable Corpora

In this chapter I present an in depth discussion of two issues which were key in the design of the Written BNC2014: representativeness and comparability. I first look in detail at the issue of creating a representative corpus, considering both how and whether this can be done. I then present a discussion of research into methods for creating and testing comparable corpora. I draw these issues together by showing how they can often be at odds with one another, and, importantly, propose a solution to this for the Written BNC2014 project.

- Chapter 4: Designing the Written BNC2014 Sampling Frame

In this chapter, I use all of the knowledge compiled in chapter 3 to design a sampling frame for the Written BNC2014. I discuss in detail the issue of how the texts in the corpus will be classified, and how this compares to other corpus projects. I also return to the issues of population definition, sample size, number of samples, corpus size, and sampling methods, which were introduced in chapter 3, and explain how these issues will be addressed in the design and creation of the corpus. Importantly, I consider how the decisions made in the design of the corpus will affect its representativeness and comparability with the BNC1994.

- Chapter 5: Collection of Books for the Written BNC2014

This chapter considers the collection of books for inclusion in the corpus. I present an account of the many methods trialled for the collection of books, and critically evaluate their success. I present a detailed account of the main method used for the collection of fiction books, and assess its success. I finish by comparing the books section of the sampling frame to the eventual composition of this section which was achieved.

- Chapter 6: Collection of Periodicals for the Written BNC2014

This chapter considers the collection of periodicals for inclusion in the corpus. I first discuss the details of how this section of the corpus sampling frame was designed, before moving on to discuss the collection of each type of periodical for the corpus. This chapter focuses, in particular, on the collection of magazines for the corpus. I present an investigation into the similarity of print and online magazines, and critically consider how the results of this study may impact the data collected for the

Written BNC2014. I finish by comparing the periodicals section of the sampling frame to the eventual composition of this section which was achieved.

- Chapter 7: Collection of e-language for the Written BNC2014

In this chapter I will discuss the rationale for, the design of, and the construction of the e-language section of the corpus. In this chapter I investigate what the composition of the web is, in order to help design the section. I also draw heavily on previous corpora of e-language. Furthermore, I will investigate the very important legal and ethical considerations which must be addressed in the creation of an e-language corpus. I then detail the design of the e-language section, and how the data was collected. I finish by comparing the e-language section of the sampling frame to the eventual composition of this section which was achieved.

- Chapter 8: Collection of Miscellaneous and Written-to-be-spoken Genres for the Written BNC2014

This chapter presents an account of the design and construction of two mediums within the corpus: miscellaneous, and written-to-be-spoken. I discuss the rationale for these sections, the design of these sections in the sampling frame, and give details of how these data types were collected. I finish by comparing the miscellaneous and written-to-be-spoken sections of the sampling frame to the eventual composition of these sections which was achieved.

- Chapter 9: Colloquialisation in Academic British English

This chapter presents a study which I have carried out using some early parts of the Written BNC2014. The analysis presented in this chapter focuses on the theory of colloquialisation, as applied to academic writing. As such, I analyse a sub-set of the academic writing data which has been included in the Written BNC2014 (academic books and academic journal articles). Several comparisons are carried out to assess whether linguistic features associated with colloquialisation have changed in frequency over time, using data from the BNC1994 and the Written BNC2014. I find that some features of academic writing have certainly become more colloquial over time, and that this change is much more pronounced in academic books than in academic journal articles.

- Chapter 10: Conclusion

Finally, I conclude by summarising the work presented in this thesis, and I discuss the main successes and limitations of my work. I also consider future research directions of the project, both using the Written BNC2014, and, more broadly, whether another BNC should be created in another twenty years time.

As stated, before moving on to discuss the design and creation of the Written BNC2014, I first need to contextualise how this project fits in with other contemporary national corpus projects. This is the focus of the next chapter.

Chapter 2: Contemporary National Corpora

2.1 Introduction

In this chapter I will discuss other national corpus projects which have taken place (or are currently taking place) within the last ten years. A good understanding of other projects, similar to the BNC2014 project, will help to contextualise the challenges I faced in the design and construction of the corpus. Such a review is also helpful as a way of identifying the various options for overcoming these challenges. The projects discussed in this chapter will be referred to throughout this thesis in order to help contextualise and provide a rationale for the decisions detailed within.

There have been a great many national corpora constructed around the world (for example, Korean (Kang & Kim, 2004), Albanian (Arkhangelskiy, 2012), Russian[3], and Croatian (Tadić, 2002) national corpora), but I focus here on *contemporary* projects as these are the most relevant and directly comparable to the BNC2014 project, and thus the most useful for seeing how others have undertaken similar projects. I will focus my discussion on aspects of design and compilation of these corpora, rather than details of annotation and tagging for example, as design and compilation are the primary focuses of this thesis. I also will not give details about the design and compilation of the spoken sections of these corpora, as they are not relevant for the *Written* BNC2014 project (readers interested in spoken corpus design and compilation should see Love et al., 2017a). Web-crawled corpora, such as the TenTen family (Jakubíček et al., 2013) and ukWaC (Ferraresi et al., 2008) have already been discussed in chapter 1, where I justified why the BNC2014 would be a 'hand-made' corpus rather than a web-crawled corpus (see section 1.3.4.3). Thus,

[3] http://www.ruscorpora.ru/en/index.html

web-crawled corpora will not be discussed in this section as they are not relevant for the decisions being made in this project. The corpora which will be discussed in this chapter are: the Corpus de référence du français contemporain (CRFC; section 2.2), the Czech National Corpus (SYN2015; section 2.3), the Thai National Corpus (TNC; section 2.4), the American National Corpus (ANC; section 2.5), the Corpus of Contemporary American English (COCA; section 2.6), and the Deutsches Referenzkorpus (DeReKo; section 2.7).

2.2 The Corpus de référence du français contemporain (CRFC)

Siepmann et al. (2015: 63) note that the analysis of French from a corpus linguistics perspective has been lagging behind that of other major languages, both in terms of "the diversity and availability of corpora as well as the sophistication of statistical analysis". This was the motivation for constructing the Corpus de référence du français contemporain (henceforth CRFC). The first version of the CRFC is a 310 million word, genre-balanced corpus of French from 1945 to 2014, with more than 90% of the data coming from the last two decades (Siepmann et al., 2015). The corpus is a monitor corpus, and, as such, it is planned for the corpus to be updated as new material becomes available. The corpus is the largest French corpus which is not based solely on internet sources, and includes spontaneous speech as well as writing (Siepmann et al., 2015). The creators anticipate that the corpus will be a source for the construction of dictionaries, grammars, and language teaching materials (Siepmann et al., 2015). The CRFC will be available online from 2018 (Siepmann et al., 2015).

The composition of the corpus was modelled on the BNC1994 and COCA, but with an even greater diversity of genres (Siepmann et al., 2015: 70; see table 2a for the composition of the corpus). The creators claim that the CRFC is the first corpus to include equal amounts of spoken and written data. However, this claim rests largely on how the blurred boundaries between speech and writing discussed in section 1.2 are defined. What they term ‘pseudo-spoken data’ would be termed ‘written-to-be-spoken’ data in many corpora. Written-to-be-spoken texts are texts which were originally written, but with the intention of them being spoken aloud at a later time (or “writing to be spoken as if not written” in Gregory’s (1967) model (see section 1.2). For example, the BNC1994 incudes play and television scripts as written-to-be-spoken texts, and, along with many other corpora, classifies these as written language. The decision to classify these texts as pseudo-spoken in the CRFC was taken because they tend towards ‘communicative proximity’ rather than ‘communicative distance’ (Siepmann et al., 2015: 71). However, referring back to section 1.2, this decision could also me made based on how the texts were delivered at the point of collection (i.e. whether they were recorded or collected in written form). This demonstrates that this distinction can be approached in many ways, and highlights the need for corpus users to assess how a corpus was constructed before deciding whether two corpora are comparable. The corpus creators’ claim that the corpus contains equal amounts of speech and writing, but the majority of the ‘spoken’ data (see the composition of the corpus in table 2a) would not be considered ‘spoken’ by most corpus linguists. Indeed, the amount of data in the Written BNC2014 would be reduced and the amount of data in the Spoken BNC2014 would be increased if drama and television scripts, IM messages, and discussion forums were reclassified as spoken data, as they are in the CRFC.

The collection of the pseudo-spoken texts was done in two ways (I am detailing the collection of these texts here because, although they are classed as spoken texts in the CRFC, they will be classified as written texts in the Written

BNC2014, and so their collection is relevant to the present discussion). Firstly, the stage plays, film scripts, and subtitles were all downloaded from various internet sites, and secondly, the text messages and discussion forums were sourced from other corpora (Siepmann et al., 2015). For the written category of the CRFC, care was taken to not allow any of the eight sections to overwhelm the others (see table 2a for details of these sections), as, the creators note, has often been the case with newspaper texts in previous corpora (Siepmann et al., 2015: 74). For the academic section of the corpus, journals were sampled from a wide range of arts and science journals available on the web. Non-academic language was sourced from samples and complete books available online, with preference being given to ‘general’ rather than ‘technical’ language (although the creators do not state how this distinction was made in practice) (Siepmann et al., 2015: 74). Complete novels and short stories were sourced online for the prose fiction section, along with some children’s books which were typed by the compilers. National and regional newspapers were obtained from the relevant websites. Sample issues of several magazines were downloaded online, with a balance between domains. No information is given about how the remaining three sections of the written corpus were compiled (Siepmann et al., 2015). The written sections of the corpus differ in size due to differing availability of material (Siepmann et al., 2015).

In addition to organising the corpus according to medium and genre (see section 4.2.3 for discussion and definitions of these terms), the creators also organise the corpus by broad and individual subject areas (Siepmann et al., 2015). The texts in the corpus are divided into 16 thematic subject areas: “arts; business; politics, government and law; computing; the environment; science and scholarship; health and medicine; belief and thought; psychology and social relationships; leisure, entertainment, sports; nutrition; clothing and fashion; travel, tourism and transport;

home and gardening; history, communication and mass media” (Siepmann et al., 2015: 75). The creators give the example that this could facilitate an investigation of the lexico-grammatical patterns used in the subject of football. The broad subject area ‘leisure, entertainment, sports’ will contain, for example, books and articles about football, discussion about football on forums, television commentaries on football etc. (Siepmann et al., 2015: 75).

Table 2a: Composition of the CRFC (adapted from Siepmann et al., 2015: 70).

Category	**Section**	**Proportion**	**Words (tokens)**
Spoken	Formal	9.7%	30,000,000
	Informal	9.7%	30,000,000
Pseudo-spoken	Stage plays and film scripts	9.7%	30,000,000
	Film and daily soap subtitles	0.8%	2,500,000
	Text messages/chat	0.8%	2,500,000
	Discussion forums	19.4%	60,000,000
		50%	**155,000,000**
Written	Academic	9.7%	30,000,000
	Non-academic books	9.7%	30,000,000
	Prose fiction	9.7%	30,000,000
	Newspapers	14.5%	45,000,000
	Magazines	3.2%	10,000,000
	Diaries and blogs	1.6%	5,000,000
	Letters and e-mails	0.3%	1,000,000
	Miscellaneous	1.3%	4,000,000
		50%	**155,000,000**

Note: I have added the proportions column to the authors’ table to aid comparability with the other composition tables presented in this chapter.

2.3 The Czech National Corpus (SYN2015)

SYN2015 is a 100 million word, representative corpus of contemporary (2010-2014) written Czech, published within the framework of the Czech National Corpus, and released in 2015 (Křen et al., 2016). The Czech National Corpus project aims to provide extensive and continuous representation of Czech in all varieties and forms, and contains many general and specialised corpora. For example SYN2013PUB contains newspaper and magazine texts from 2005-2009, and SYN2000 is a representative corpus of texts from 1990-1999 (en:cnk:uvod, 2017). I will focus this discussion on SYN2015 as it is a contemporary corpus which contains similar data to the Written BNC2014 (i.e. written texts), and thus will be most useful as a basis of comparison.

SYN2015 can be described as a 'hand-made' corpus, similar to the Written BNC2014, as opposed to a web-crawled corpus, and features "cleared copyright issues, well-defined composition, reliability of annotation and high-quality text processing" (Křen et al., 2016: 2522). The authors do not detail what copyright issues were cleared, although presumably they are suggesting that they had legal access to copy all of the texts within the corpus – how this access was obtained is not clear. SYN2015 is designed to be representative, that is, it contains "a large number of texts that cover all the varieties the corpus aims to represent" (Křen et al., 2016: 2523), but is not claimed to be balanced proportionally, because, amongst other arguments, the population of texts to be represented was unknown. Texts within SYN2015 are classified according to text type and genre, and information regarding medium, periodicity, and audience is also available for every text. A full list of text types and genres included in the corpus, along with their proportions can be seen in table 2b.

The three top-level categories are fiction, non-fiction, and newspapers and magazines. E-language is not included in the corpus as this is covered in a different series of corpora within the Czech National Corpus project. The three top-level categories each comprise one third of the corpus – proportions are set arbitrarily, but close to the figures seen in earlier corpora within the series (Křen et al., 2016: 2523). The texts included in the corpus are claimed to be representative of the period 2010-2014, but, due to the variation of the borders of synchronicity across texts types, the date of first publication of some texts is much earlier than this period. For example, fiction must have been published within the previous 25 years and *first published* within the previous 75 years.

Table 2b: Composition of SYN2015 in terms of the major classification categories (adapted from Křen et al., 2016: 2524).

Text type	Genre	Category	Proportion	Words (tokens)
Fiction (FIC)			**33.33%**	**33,330,000**
NOV		Novels	26%	26,000,000
COL		Short stories	5%	5,000,000
VER		Poetry	1%	1,000,000
SCR		Drama, screenplays	1%	1,000,000
X		Other	0.33%	330,000
Non-fiction (NFC)			**33.33%**	**33,330,000**
SCI – scientific POP – popular PRO - professional	HUM	Humanities [sub-classified into: ANT – anthropology, THE – theatre, PHI – philosophy and religion, HIS – history, MUS – music, LAN – philology, INF – library and information science, ART – arts and architecture]	7%	7,000,000
	SSC	Social sciences [sub-classified into: ECO – economics, POL – politics, LAW – law, PSY, psychology, SOC, sociology, REC – recreation, EDU – education]	7%	7,000,000
	NAT	Natural sciences [sub-classified into: BIO – biology, PHY – physics, GEO – geography and geology, CHE – chemistry, MED – medicine, AGR – agriculture]	7%	7,000,000
	FTS	Technical sciences [sub-classified into: MAT – mathematics, TEC – technology, ICT – information and communications technology]	7%	7,000,000
	ITD	Interdisciplinary	1%	1,000,000
MEM		Memoirs, autobiographies	4%	4,000,000
ADM		Administrative texts	0.33%	330,000
Newspapers and magazines (NMG)			**33.33%**	**33,330,000**
NEW	NTW	Nationwide newspapers – selected titles [equal shares of HN, LN, MFD, Právo]	10%	10,000,000
	NTW	Nationwide newspapers	5%	5,000,000
	REG	Regional newspapers	5%	5,000,000
LEI		Leisure magazines [sub-classified into: HOU – hobby, LIF – life style, SCT – society, SPO – sports, INT – curiosities]	13.33%	13,330,000

Note: I have added the words (tokens) column to the authors' table to aid comparability with the other composition tables presented in this chapter.

2.4 The Thai National Corpus (TNC)

The Thai National Corpus (TNC) is a general corpus of written Thai which is designed to be comparable to the Written BNC1994 (Aroonmanakun et al., 2009). The corpus project is ongoing, and the creators aim to collect 80 million words all together (ibid.). 90% of the texts in the corpus will have been published between 1998 and the present day, and 10% of texts will have been published earlier than this (ibid.). However, as the project is still ongoing at time of writing, it could certainly be argued that this date range does not represent only *contemporary* Thai.

Texts within the corpus are classified similarly to the Written BNC1994: texts are classified according to medium and domain, and also according to genre (see section 4.2.4.3 for more information on text classification in the BNC1994). The classification system for texts in the TNC can be seen in table 2c (as the project is ongoing, it is not possible to provide detailed information on proportions and word counts, as has been done for other corpora in this chapter). Since it was not possible for the creators to know the proportions of texts within their population, they decided to make the TNC comparable to the BNC in terms of the medium and domain proportions (Aroonmanakun, 2007). 75% of the texts in the corpus will come from the 'informative' domain, and 25% will come from the 'imaginative' domain. This decision was based on the belief that generally "people read or write informative texts, e.g. newspapers […] more often than imaginative texts, e.g. novels" (Aroonmanakun, 2007: 7). In the medium dimension, an 'internet' medium has been included to reflect the fact that many texts are now published on the web – this medium will replace the written-to-be-spoken medium found in the BNC1994 (Aroonmanakun, 2007).

Priority was given to collecting texts which were read by lots of people, produced by famous writers, or recognized as valuable works (although I have been unable to find information about how this was determined for a given text). The maximum sample size for texts in the corpus is 40,000 words or 80 pages of A4 paper. Text samples are randomly taken from either the beginning, middle, or end of the text. If a text is less than 40,000 words then only 90% of the text is used (Aroonmanakun, 2007).

For texts which were protected by copyright (e.g. books), the creators of the TNC contacted publishers to ask for their permission to include their copyrighted texts in the corpus (Aroonmanakun et al., 2009). However, initial response rates to this request were very low, with publishers not seeming to understand the purpose of a corpus project (Aroonmanakun et al., 2009). It also transpired that for many texts the copyright was in fact owned by the author of the text rather than the publisher. The creators of the TNC contacted 22 publishers in total, 7 of whom were able and willing to provide a list of contact details for the copyright holder of each text which they had published (Aroonmanakun et al., 2009). Each author then had to be contacted directly to ask for their permission to include their text in the corpus. For texts which were not protected by copyright (e.g. news articles), the creators of the TNC selected texts from internet sources.

Table 2c: Design of the Thai national Corpus (Aroonmanakun et al., 2009: 159).

Domain		**Medium**	
Imaginative	25%	Book	60%
Informative *Applied science* *Arts* *Belief and thought* *Commerce and finance* *Leisure* *Natural and pure science* *Social science* *World affairs*	75%	Periodical	20%
		Published miscellanea	5-10%
		Unpublished miscellanea	5-10%
		Internet	5%
		Time	
		1998-present	90-100%
		1988-1997	0-10%
		Before 1988	0-5%

Genres	**Sub-genres**
Academic	Humanities, e.g. Philosophy, history, literature, art, music
	Medicine
	Natural sciences, e.g. Physics, chemistry, biology
	Political science – Law – Education
	Social sciences, e.g. Psychology, sociology, linguistics
	Technology & Engineering, e.g. Computing, engineering
Non-academic	Humanities
	Medicine
	Natural sciences
	Political science – Law – Education
	Social sciences
	Technology & Engineering
Advertisement	
Biography – Experiences	
Commerce – Finance – Economics	
Religion	
Institutional documents	
Instructional – DIY	
Law & Regulation	
Essay	School
	University
Letter	Personal
	Professional
Blog	
Magazine	
News report	
Editorial – Opinion	
Interview – Questions & answer	
Prepared speech	
Fiction	Drama
	Poetry
	Prose
	Short stories
Miscellanea	

Note: As the project is ongoing, more detailed information about proportions and word counts is not available.

2.5 The American National Corpus (ANC)

The American National Corpus (ANC) is intended to be a carefully designed corpus containing 100 million words of written and spoken American English which follows the general framework of the BNC1994 (Reppen and Ide, 2004). The currently released version of the corpus contains 22 million words (Ide, 2008; see table 2d for the composition of the currently released version of the corpus). Additional genres are included in the corpus which did not exist when the BNC1994 was created, such as e-language (Ide, 2008).

The creators of the corpus initially hoped that the publishers within their project consortium would contribute data to the project, but very few did (Ide, 2008). As such, data acquisition has been the major issue faced in the development of the corpus (Ide, 2008). Ide (2008) points out that many linguists have turned to using the web as a source of data, and that hand-made corpora, such as those discussed in this chapter, have been seen as outdated. However, Ide (2008) notes that this approach was not desirable or possible for the creators of the ANC. Firstly, Ide (2008: 110) notes that web-crawled corpora are not representative of general language use. Secondly, and most significantly, studies using web-crawled corpora are not replicable because in the U.S. “all web data are copyrighted unless explicitly indicated to be in the public domain or licensed to be redistributable through a mechanism such as Creative Commons” and so the corpora created from the web cannot be released to others (Ide, 2008: 110). Although this argument provides justification for the creation of hand-made corpora rather than web-crawled corpora, it was also the greatest obstacle in the creation of the ANC (Ide, 2008: 110). As a result of web-data being protected by copyright, the creators have had to rely on government sites for public domain documents, and web archives which are under creative commons licenses (Ide, 2008).

The creators have also reached out to the public to invite them to submit texts to the corpus (Ide, 2008).

The current release of the corpus contains 3.8 million words of spoken data, from unscripted conversations and interviews (see table 2d). The corpus also contains 18.5 million words of written data, from sources such as government reports, travel guides, web forums, academic journals, fiction, magazines, newspapers, and non-fiction books (American National Corpus Project, 2015; see table 2d). These texts are not, however, reflective of the balance seen in the BNC1994 (which was a goal of the project at its outset). The difficulty presented by the creators' desire to only include texts which can be legally redistributed means that obtaining large amounts of data has had to be prioritised over balance.

Table 2d: Composition of the second release of the American National Corpus (adapted from American National Corpus Project, 2015).

Spoken				
Corpora	**Domain**	**No. files**	**Proportion**	**Words (tokens)**
callhome	telephone	24	0.2%	52,532
charlotte	face to face	93	0.9%	198,295
micase	academic discourse	50	2.6%	593,288
switchboard	telephone	2,307	13.5%	3,019,477
Spoken Totals		**2,474**	**17.2%**	**3,863,592**
Written				
Corpora	**Domain**	**No. files**	**Proportion**	**Words (tokens)**
911 report	government, technical	17	1.3%	281,093
berlitz	travel guides	179	4.5%	1,012,496
biomed	technical	837	15%	3,349,714
buffy	Blog	143	13.8%	3,093,075
hargreaves	Fiction	106	1.8%	405,195
eggan	Fiction	1	0.3%	61,746
icic	Letters	245	0.4%	91,318
nytimes	newspaper	4,148	16.2%	3,625,687
oup	non-fiction	45	1.5%	330,524
plos	technical	252	1.8%	409,280
slate	Journal	4,531	18.9%	4,238,808
verbatim	Journal	32	2.6%	582,384
web data	government	285	4.7%	1,048,792
Written Totals		**10,821**	**82.8%**	**18,530,112**
Corpus Totals		**13,295**	**100%**	**22,393,704**

Notes: 1) I have added the proportion column to the authors' table to aid comparability with the other composition tables presented in this chapter. 2) The Corpora column gives information about the corpora from which the data in these sections was sourced.

2.6 Corpus of Contemporary American English (COCA)

The Corpus of Contemporary American English (COCA) was created to address the limitations of the ANC discussed in section 2.5. As work on the ANC has been halted by issues of copyright, Davies (2009) decided that a new project was needed to represent American English. Davies (2009: 159) claims that COCA is the “first large and diverse corpus of American English”. COCA is a monitor corpus to which 20 million words of data have been added for every year between 1990 and 2017 – the corpus now contains more than 560 million words. The composition of the corpus can be seen in table 2e.

For each year, the proportions of texts in the corpus are evenly divided (roughly 20% each) between spoken, fiction, popular magazines, newspapers, and academic journals. The creators of COCA chose not to include e-language in the corpus for two reasons. Firstly, the corpus was designed to facilitate diachronic studies and, as such, the creators wanted each year contained within the corpus to have the same proportions of genres. It would have been practically impossible to collect enough e-language for the earlier years in the corpus as, for example, blogs did not exist until the early 2000s (Davies, 2009). Additionally, the creators wanted each genre of texts to be present in equal proportions in each year of the corpus, and it was felt that it would have been extremely difficult to collect 20 million words of e-language for any year. Secondly, as the corpus is designed for the study of *American* English, the texts contained within must be produced in the United States. It is very difficult to ensure the location of production when collecting e-language and so it would not have been desirable to include this data. Davies (2009: 160) claims that COCA is the “first large corpus of American English that contains data from a wide range of genres”. However, it is questionable whether five genre categories really represents a “wide range” of the genres found within American English. For example, the ANC contains three different types of spoken data and nine different types of written data (although these are referred to as ‘domains’ rather than genres in the ANC; see table 2d). It should be said though, that Davies (2009) does not actually make any claims that the corpus is *representative* of American English.

Texts for the corpus were mostly downloaded from text archives which contain, for example, TV transcripts, short stories, magazines, newspapers, and academic articles. Some texts were retrieved manually, and some were downloaded automatically using a script which detects the sources needed for the corpus. Automatic downloading is advantageous as it allows the corpus to be added to regularly with little manual effort. In order to circumvent the copyright issues faced by the creators of the ANC (see section 2.5), the creators of COCA, rather than giving users full text access, chose to limit KWIC displays to a limited number of words. This is compliant with US Fair Use law as there is “no competition with and no adverse economic impact on the copyright holder” (Davies, 2009: 164).

Table 2e: Composition of COCA.

Genre	**Proportion**	**Words (tokens)**
Spoken	20.5%	118,000,000
Fiction	19.7%	113,000,000
Magazines	20.5%	118,000,000
Newspapers	19.8%	114,000,000
Academic journals	19.5%	112,000,000
	100%	575,000,000

Note: This table is accurate at the time of writing. As COCA is a monitor corpus the word counts will change over time, but proportions should stay roughly the same.

2.7 Deutsches Referenzkorpus (DeReKo)

The Deutsches Referenzkorpus (DeReKo) is “one of the major resources for the study of the German language” (Kupietz et al., 2010: 1848). The project was started in 1964, and is a monitor corpus which is regularly added to, with the corpus now containing over 42 billion words (Instutut Für Deutsche Sprache, 2018). Kupietz et al. (2010) state that the corpus contains fiction, scientific, and newspaper texts, as well as other text types which Kupietz et al. (2010) do not name explicitly. There is no published account of the composition of DeReKo in either Kupietz et al. (2010), Kupietz and Lüngen (2014) or Kupietz et al. (2018), and information regarding the composition of the corpus is also not available from the Instutut Für Deutsche Sprache (2018). Thus, it was not possible to produce a composition table for this corpus similar to those given for the other corpora discussed in this chapter.

The texts contained within the corpus are complete and unaltered, and only licensed material is included in the corpus. As such, the corpus is not available to download due to the creators of the corpus not owning the rights to the texts. The rights to use the texts are heavily regulated, for example “(i) only academic use is allowed whereas direct or indirect commercial use is explicitly forbidden; (ii) access is only allowed through specialized software; (iii) only authenticated users may be granted access; (iv) full texts must not be reconstructable from the output of this software; (v) all traffic must be logged; and (vi) abuse must be, as far as possible, prevented by technical precautions” (Kupietz et al., 2010: 1849). However, in 2018 an alteration to German copyright law came into effect, which altered how copyright protected content can be used in “the spheres of education and research, and within so-called knowledge institutions” (Kupietz et al., 2018: 4353). This change now allows available content to be “automatically reproduced, structured, and categorised for building a corpus” without gaining permission from the copyright holder (Kupietz et al., 2018: 4353). This new law will allow the creators of DeReKo to legally collect any documents which are freely available on the web and publish them in DeReKo without explicit permission from copyright holders. This law seems similar to the ‘non-commercial research’ exception to UK copyright law discussed in section 1.5. However, the new German law explicitly mentions corpus creation, which the UK law does not. This may suggest that corpus creation would be viewed favourably under UK copyright law, if attitudes are changing in line with those of German law.

DeReKo is not designed to be either balanced or representative, as the creators felt that these issues should be decided by individual researchers using the corpus (Kupietz et al., 2010). As such, although the whole corpus may be used as a sample, the principle purpose of the corpus is to be used as a large sample from which smaller, specialized samples can be drawn. This means that the project focuses on the maximization of the size of the corpus, with the issue of sampling left to the users of the corpus. This is not uncommon in monitor corpora, where users of the corpus will often want to select what time period within the corpus they are interested in. This idea in the DeReKo corpus simply extends this to also encourage users of the corpus to select what genres of data, and in what amounts, they want to use. As will be seen in chapter 3, the issue of creating a representative corpus is fraught with difficulty, and an approach where users are aware of the unbalanced or unrepresentative nature of the corpus is often suggested as a solution to this problem.

2.8 Conclusion

This chapter has shown the various approaches to creating national corpora which have been taken in recent years. Table 2f summarises some of the basic features of the corpora. Of the six corpora discussed, three are synchronic corpora (SYN2015, TNC, and ANC) and three are monitor corpora (CRFC, COCA, and DeReKo). As one would expect, because they are regularly added to and so continuously grow in size, the three monitor corpora are the biggest of the six, with DeReKo containing the most words (42 billion).

The most common ways of classifying texts are according to medium (CRFC and TNC) and genre (CRFC, SYN2015, and TNC) (see section 4.2 for a more in depth discussion of this issue). All of the corpora claim, to greater or lesser extents, to include a wide range of genres. For the most part the corpora seem to contain similar genres, e.g. newspaper articles, magazine articles, fiction books etc. However, some corpora choose to include e-language (CRFC, TNC, ANC) whereas others do not (SYN2015, COCA, DeReKo). The corpus creators who chose to include e-language did so to ensure that the corpus was as representative of contemporary language as possible. The creators of SYN2015 did not include e-language because it was already represented in a different corpus within the SYN series, whereas the creators of COCA chose not to include e-language because the different years within the monitor corpus would not then be comparable.

An issue which four of the corpus creators discuss explicitly is that of balance and proportionality. The creators of SYN2015, TNC, COCA, and DeReKo all reached the conclusion that representing the genres within their corpora in the proportions in which they occur in the real world was not possible. This is an issue faced by all corpus creators (see section 3.2.3 for a more in depth discussion of this issue), and was, indeed, an issue I faced when designing the sampling frame for the Written BNC2014 (see chapter 4).

Another issue faced by many of the corpus projects discussed here was copyright and/or legal access to texts. The TNC, ANC, COCA, and DeReKo projects all faced these issues to varying extents. One of the biggest issues faced by the creators of the TNC and ANC was access to published books. Both corpus projects attempted to contact publishers for their permission to be given access to their texts to include them in the corpora, but the response rate in both cases was extremely low. This is an issue which has been particularly apparent in the creation of the Written BNC2014 (see chapter 5). These copyright issues have greatly stalled the creation of the TNC and ANC, whereas these issues did not halt the creation of COCA or DeReKo. They did, however, impact the final release of both (texts within COCA can only be viewed in KWIC view, and DeReKo is not downloadable). The struggles faced by the creators of the TNC and ANC further justify my decision to collect the majority of the texts for the Written BNC2014 using the 'non-commercial research' exception to UK copyright law. The creators of the TNC and ANC followed their countries' copyright laws very strictly and the projects stalled, whereas the creators of DeReKo and COCA have both made use of exceptions to copyright law in their jurisdictions which allow them greater access to texts for inclusion in their corpora. This is the approach which has been followed in the creation of the Written BNC2014.

Now that I have an understanding of previous national corpus projects which may be useful in the design of the Written BNC2014, I must now consider the various aspects of actually designing a representative and comparable corpus. This is the focus of the next chapter.

Table 2f: A comparison of contemporary national corpus projects.

	Size	**Dates represented**	**Classification of texts**	**Includes e-language?**	**Balance/proportionality**	**Difficulties encountered regarding legal access to texts/copyright?**
CRFC	310 million	1945-2014 (monitor)	Medium, genre, subject areas	Yes	Not proportional, spoken and written sections equally balanced	Not discussed
SYN2015	100 million	2010-2014	Text type, genre	No	Not proportional	Not discussed
TNC	80 million (aim, currently 33 million)	1998-present	Medium, domain, genre	Yes	Not proportional, based on BNC1994 proportions	Yes
ANC	100 million (aim, currently 22 million)	1990-present	Not discussed	Yes	Aimed to be balanced the same as BNC1994, not possible due to copyright issues	Yes
COCA	560 million	1990-2017 (monitor)	Not discussed	No	Not proportional, each genre equally balanced	Some (can only view texts in KWIC view)
DeReKo	42 billion	1964-present (monitor)	Not discussed	No	Not designed to be balanced or representative	Some (corpus cannot be downloaded)

Chapter 3: Creating Representative and Comparable Corpora

3.1 Introduction

In this chapter I will discuss two issues which have been key in the design of the Written BNC2014: representativeness and comparability. In section 3.2 I will discuss previous research into the issue of how to make a corpus representative, and the varying views regarding whether representativeness can be achieved. In section 3.3 I discuss the various ways of defining 'comparable' in terms of corpus creation, and consider research which has investigated methods for creating and testing comparable corpora. In section 3.4 I will draw these two issues together by showing how they can often be at odds with one another in the design and creation of a corpus. I link this to problems which were encountered in the design of the Written BNC2014 sampling frame, and then outline the solution to this problem which has been developed for this corpus.

3.2 Corpus Representativeness

In this section I will discuss previous research on the contentious issue of corpus representativeness. I will begin by giving some definitions of representativeness and reasons why the issue of representativeness is important in corpus design. I will then introduce the idea of sampling procedures, and discuss the related issues of population definition, random sampling, proportional sampling, balance, sample size, number of samples, and corpus size. I will then discuss the view that representativeness cannot be achieved in corpus construction and the various approaches which have been suggested to deal with this problem. I will finish by outlining how the BNC1994 dealt with the issue of representativeness.

3.2.1 Defining representativeness

Hunston (2008: 154) suggests that corpus planning and compilation are "prone to paradox, where even the apparently simplest decisions can have extensive ramifications". This is perhaps never truer than when dealing with the issue of corpus representativeness. Representativeness in corpus design is achieved when the texts selected for inclusion in the corpus represent the full range of variability in the language which the corpus aims to represent. In other words, representativeness "means that the study of a corpus (or combination of corpora) can stand proxy for the study of some entire language or variety of language" (Leech, 2007: 135).

Leech (2007) states that representativeness is always desired in corpus design, but that in practice this issue has not been treated as seriously as it should be. Similarly, Köhler (2013) argues that whilst representativeness is often claimed for large corpora, this claim is rarely justified. In Leech's (2007: 135) view this is unacceptable, as "unless the claim that a corpus is representative can be substantiated, we cannot accept such findings [of corpus research]. Without representativeness, whatever is found to be true of a corpus, is simply true of that corpus – and cannot be extended to anything else". However, once one begins to consider how to *prove* that a corpus is representative, what initially seems like a fairly simple concept quickly becomes very complex. The remainder of section 3.2 will discuss the various considerations which researchers must make in order to achieve a representative corpus.

3.2.2 Sampling

One of the most important elements of creating a representative corpus is sampling (Váradi, 2001; Biber, 1993). Sampling refers to the decisions made regarding what texts to include in a corpus, and includes considerations such as population definition, approaches to considering the importance of texts, sampling methods, and sample and corpus sizes, all of which will be considered in the following sections. The sampling techniques employed in corpus design are important because "a corpus is not simply an archive of texts but rather a principled collection of texts" (Váradi, 2001: 588). Each decision made regarding what to include in the corpus is a sampling decision, and so must be considered carefully. Indeed, Bauer and Aarts (2000:22) state simply that "The moral is clear: pay more attention to sampling".

3.2.2.1 Population definition

The first sampling issue which linguists must tackle is defining the population which they want their corpus to be representative of. Biber (1993: 243) states that there are two aspects involved in defining the target population: the population boundaries (what texts are included and excluded), and what text categories are included in the population, along with definitions of these text categories. Biber (1993: 243) argues that population definition is often not paid enough attention in corpus design, which means that for many corpora there is no way to assess their representativeness, because what the corpus was intended to represent was never explicitly defined. The representativeness of a sample will depend upon the extent to which it represents the full range of variation within the target population. There will be a range of linguistic distributions (i.e. the different ways in which linguistic features are distributed within texts, across texts, and across text types) in a population, and a representative corpus should allow full analysis of all of these distributions (Biber, 1993).

Both Biber (1993) and McEnery et al. (2006) agree that, when defining a population, register/genre (see chapter 4 for discussion and definition of these terms) distinctions are more important than text type distinctions (again, see chapter 4 for discussion and definition). Register/genre distinctions are based on factors which are external to the corpus, such as the purpose and function of a text, whereas text type distinctions are based on linguistic criteria which are internal to the corpus. McEnery et al. (2006: 14) explain that it would be circular to use internal criteria to select data from a population because a corpus is "typically designed to study linguistic distributions". Thus, if the linguistic distributions are already known when the corpus is designed, then there is nothing to be gained from analysing the corpus. However, some researchers, such as Otlogetswe (2004), have indicated that they believe internal criteria to be the best selection tools.

An important sampling decision, which is closely tied to population definition, is whether you will define the population in terms of i.) language production; ii.) language reception or iii.) texts as products. Representing the population in terms of language reception would mean giving great significance to the language of the very few people within a population who produce language which is heard or read by many (for example, published authors), whereas representing the population in terms of language production would give great significance to texts such as everyday conversations and emails, each of which is often only heard or read by a very few people. Both of these definitions of the population result in a 'demographically organised' corpus, where the data is collected and organised according to statistics about the producers or receivers of the language. Atkins et al. (1992) favour representing language production as much as possible because, although texts with a wide reception are easier to come by, for the corpus to be a true reflection of language in use, as much production material must be included as is possible. However, Biber (1993: 244) takes a different route. He suggests that the population should be defined in terms of "texts as products", because there are many types of texts (such as insurance documents) which are very rarely produced or received, and so these text types would not be properly represented in a demographically organised corpus (regardless of whether it was demographically organised according to production or reception). Thus, Biber (1993: 245) suggests that "A corpus organized around texts as products would be designed to represent the range of registers and text types rather than the typical patterns of use of various demographic groups". In order to define the population in this way, Biber (1993: 245) proposes a set of sampling strata which should be considered in turn when defining a population (see table 3a).

Table 3a: "Situational parameters listed as hierarchical sampling strata" (from Biber, 1993:245).

1.	*Primary channel.* Written/spoken/scripted speech
2.	*Format.* Published/not published (+ various formats within 'published')
3.	*Setting.* Institutional/other public/private-personal
4.	*Addressee.* (a) Plurality. Unenumerated/plural/individual/self (b) Presence (place and time). Present/absent (c) Intercativeness. None/little/extensive (d) Shared knowledge. General/specialized/personal
5.	*Addressor.* (a) *Demographic variation.* Sex, age, occupation, etc. (b) *Acknowledgement.* Acknowledged individual/institution
6.	*Factuality.* Factual-informational/intermediate or indeterminate/imaginative
7.	*Purposes.* Persuade, entertain, edify, inform, instruct, explain, narrate, describe, keep records, reveal self, express attitudes, opinions, or emotions, enhance interpersonal relationship,…
8.	*Topics…*

So it seems clear that in order to create a representative corpus the first task must be to define the population to be represented. However, this may not, in fact, be possible. Hunston (2008), Bauer and Aarts (2000), and Atkins et al. (1992) agree that delimiting the total population in any systematic way is often impossible because there are no exhaustive lists of, for example, genres or social groupings in a population. This is an issue also encountered by Křen et al. (2016:2523) when constructing the SYN2015 corpus (a corpus of contemporary written Czech), and by Aroonmanakun et al. (2009) when creating the Thai National Corpus (TNC) (see chapter 2). Furthermore, Atkins et al. (1992: 4) point out that "even if the population could be delimited […] it will always be possible to demonstrate that some feature of the population is not adequately represented in the sample." Siepmann et al. (2015:70) believe that because of these difficulties in defining a population, "standard approaches to statistical sampling are hardly applicable to building a language corpus".

3.2.2.2 Random sampling

Once the population has been defined, different sampling techniques can be used to select items from the population for inclusion in the corpus. The most basic of these techniques is simple random sampling. In simple random sampling, all members of a population are assigned a number, and then a table of random numbers is generated in order to facilitate random selection of members of the population (McEnery et al., 2006; Bauer and Aarts, 2000; Biber, 1993). This gives every item an equal chance of being selected, which would seem to be a good sampling method. However, simple random sampling works against selecting items which are rare in the population, and favours those which are common (McEnery et al., 2006; Váradi, 2001). Researchers are often interested in the rare items within a population, and a means of accomplishing this is to use stratified random sampling. Stratified random sampling first divides the population into strata, and then samples randomly from within these strata. Biber (1993: 244) suggests that "This approach has the advantage of guaranteeing that all strata are adequately represented while at the same time selecting a non-biased sample within each stratum […] a sample that forces representation across identifiable groups will be more representative overall." Of course, deliberately seeking to include rare items in a population has the effect that the sample is no longer quantitatively representative of the population, however, one must always think about what the end-user of the corpus wants to research, and if the users are interested in rare occurrences in a population then stratified random sampling is entirely appropriate. Váradi (2001: 590) points out that the granularity of the strata will have a direct bearing on the quantitative results drawn from the corpus. He uses the example of reviews to illustrate this: if you have a stratum for reviews, chance will dictate whether any reviews of travel books are selected; however, if a specific stratum for reviews of travel books is set up then the sample is sure to include reviews of travel books. A further difficulty of random sampling relates back to the discussion of population definition in section 3.2.2.1. It is often impossible to know the exact members of a population, because this information often does not exist, and yet without an itemised list of the population random sampling is not possible. This again reinforces Siepmann et al.'s (2015) belief that traditional sampling procedures are not suitable for investigating language.

In the creation of the ARCHER corpus, a multi-genre corpus of British and American English covering the period 1600-1999, Biber et al. (1994) used random sampling within their population of the research libraries of the University of Southern California, the University of California at Los Angeles, and the Huntington Library in

San Marino. They used random sampling within bibliographies to identify double the number of samples they would need to fill each of their chosen strata. These were then checked for availability and suitability, until the target amounts had been met. This illustrates how the random sampling method can actually be used in practice.

3.2.2.3 Proportional sampling

An important aspect of stratified random sampling is proportionality. In order for a corpus to be considered representative, it is commonly claimed that the amount of text in each stratum should be proportional to its frequency in the population as a whole (McEnery et al., 2006; Biber, 1993). However, Biber (1993) argues that proportional sampling is not suitable for language corpora. This is because, Biber argues, a proportional language corpus would have to be organised demographically based on people's language production (as there is no way to determine the proportions of all registers within a language). This would result in a corpus of "roughly 90% conversation and 3% letters and notes, with the remaining 7% divided among registers such as press reportage, popular magazines, academic prose, fiction, lectures, news broadcasts, and unpublished writing" (Biber, 1993: 247). This would be a proportional representation of the language, but would only allow for generalizations about language which are not particularly interesting to researchers (Biber, 1993), because "Linguists consider the rare event, while representative sampling would suggest ignoring it" (Bauer and Aarts, 2000: 29). Biber (1993) instead suggests that researchers actually require corpora which are representative in the sense that the full range of linguistic variation is adequately represented. Biber (1993: 247-248) concludes that there are two main factors which make proportional sampling unsuitable for language corpora. The first is that proportional samples only reflect the numerical frequencies of registers in a language, rather than being representative of a register’s importance within a language. Biber (1993) argues that registers such as books and newspapers are much more important than their numerical frequencies would indicate (see discussion of language production and reception in section 3.2.2.1), and so perhaps proportional representation based on frequency is not the best solution for representing language. Secondly, Biber (1993) argues that we already know that 90% of texts in a language (i.e. the conversations) are linguistically similar, so we do not need a corpus to find this out. Rather, we should be creating corpora which are representative of the other 10% of language, since this is where the majority of variation lies.

However, Váradi (2001) strongly rejects Biber’s arguments. Váradi (2001) argues that using a notion of importance derived from culture is far too subjective for corpus linguistics. He states that there is no way to establish this notion of importance in language, and that using this method would result in “subjective judgement in the compilation of the body of data that is expected to provide empirical evidence for language use” (Váradi, 2001: 592). He also claims that it is misleading to criticise proportional sampling for failing to do something which it was never intended to do (Váradi, 2001). Váradi (2001) also rejects Biber’s (1993) attempt to reframe the definition of representativeness for corpus design. Biber (1993) argues that linguists want a corpus which is representative in the sense that it represents the full range of variation in a language. But Váradi (2001) sees this re-definition (from the well-understood definition of representativeness equalling proportionality) as detracting from the field of corpus linguistics. He argues it is somewhat like cheating in that it allows researchers to claim that their corpora are representative by simply redefining the notion of representativeness.

Leech (2007) proposes a method which he feels is a solution to both Biber's (1993) criticisms of proportionality, and also to Váradi's (2001) criticisms of Biber. Biber (1993) stated that proportionality is not appropriate for linguistic corpora because it would be based on language users' production. However, Leech (2007) shows that this does not have to be the case; he proposes that representation should be proportional to both language production and language reception. Leech (2007) labels this measure of significance an "Atomic Communicative Event (ACE)". Thus, "a radio programme that is listened to by a million people should be given a much greater chance of being included in a representative corpus than a conversation between two people, with only one listener at any one time" because the radio programme has a million ACEs (Leech, 2007: 138). This also solves Váradi's (2001) issue with the subjectivity of measuring cultural importance – in this method a text's importance can be measured in terms of its ACEs. Leech (2007) does concede that for most texts in a population it would be impossible to actually obtain the information which would allow a researcher to calculate its ACE value, and this is a common issue which is raised when considering the value of proportional sampling. Váradi (2001: 590) argues that this kind of information about a population is simply not available. This is also suggested to be the case by the creators of the SYN2015 corpus (Křen et al., 2016: 2523) and the TNC (Aroonmanakun et al., 2009). As such, both of these corpora are not sampled proportionally. However, Leech (2007) responds that this does not mean that ACE-proportionality is not worth pursuing. He suggests that even when these figures are not known that they can be estimated, and that proportional representativeness should be viewed as a scalar phenomenon and something which should be aimed for, rather than something which can be proven to have been achieved (see section 3.2.3).

3.2.2.4 Balance

An issue which is intertwined with proportionality is that of balance. Hunston (2008) views balance as at odds with representativeness, because, in her view, balance requires all text types in a corpus to be equally represented, whereas representativeness requires all text types to be represented proportionally. However, Leech (2007: 136) argues that “for a corpus to be balanced is an important aspect of what it means for a corpus to be representative.” Many other linguists agree with Leech (2007) that balance and proportionality are essentially the same things (Atkins et al., 1992; McEnery et al., 2006). Indeed, in the creation of the SYN2015 corpus, Křen et al. (2016: 2523) consider a balanced corpus to contain “varieties in proportions that correspond to the reality of a (sub)language in question”. In other words, a corpus can be said to be properly balanced when all of the text types within it are represented proportionally to their occurrence in the total population. This is the definition of balance which I will use in this thesis.

3.2.2.5 Sample size

Once your target population has been defined, and you have decided on your sampling method, there are three final factors to consider, one of which is sample size. However, before you can consider your sample size, you must consider where to select your sample from within a text. Should samples be taken from the beginning, middle, or end of a text? Or should they be made up from a combination of locations within a text? Or should we include whole texts as our sample, rather than sampling sections of texts? Sinclair (1991: 19) suggests that sampling texts creates a risk of differences between parts of texts being overlooked, and therefore advocates the use of whole texts rather than sampling. However, many linguists disagree with this opinion. McEnery et al. (2006: 20) point out the difficulties which would be encountered in terms of copyright if whole texts were used – copyright holders are unlikely to agree to their entire texts being reproduced in a corpus (see chapter 5 for a discussion of this issue in relation to the collection of books). Furthermore, use of whole texts would require the eventual corpus to be extremely large to avoid the problem of one or two large texts skewing the results (McEnery et al., 2006: 20; Hunston, 2008: 166). Although, when sampling rather than using whole texts, it is important to ensure that you are balancing samples from text initial, middle, and end position so that features which are particular to certain locations within a text are not over- or under-represented compared to others (McEnery et al., 2006: 20).

So it seems that most linguists favour sampling texts, which leads us to the consideration of sample size. Sample size refers to the decision which must be made regarding how long the text chunks included in the corpus should be in order to reliably represent the linguistic distributions in the population. Biber (1990) conducted a study in order to attempt to identify what length of sample was necessary for a corpus to be representative of the population. He compared a variety features (e.g. first person pronouns, contractions, prepositions etc.) in 1000 word extracts of texts from the LOB (a member of the Brown Family; see section 1.3.4) and London-Lund (a 500,000 word corpus of spoken British English) corpora in order to see if internal variation was stable. He found that most linguistic features had fairly stable variations across 1000 word samples, which indicates that 1000 word samples within corpora would reliably represent the variation within common features. However, the stability of rarer features, such as conditional subordination, was weaker, leading Biber to suggest that the larger 2000 and 5000 word samples which are common in many corpora would be satisfactory for this type of analysis. Biber (1993: 252) continues this research and finds that for linearly distributed features (i.e. those features which occur the same amount of times in each equally sized sample), the required sample length will depend on the overall stability of the feature, whilst for curvilinear features (i.e. those features where each new sample contributes fewer new instances) a cut-off point must be decided where ‘adequate’ representation has been reached (Biber suggests when additional material is adding less than 10% new types). Overall, Biber (1993: 252) acknowledges that much more research is needed in order to propose specific recommendations for sample length, particularly focusing on less stable linguistic features, and other types of features such as discourse features. This is particularly important when making recommendations for general corpora, as these will be used to study a wide variety of linguistic features rather than the fairly common ones which Biber’s research focuses on.

In dealing with these issues in the creation of the ARCHER corpus, Biber et al. (1994) used 2000 word text samples. For short texts, they grouped together individual texts to achieve their target, and for longer texts they sampled the first and last 500 words, and the middle 1000 words. Similar approaches were also used in the creation of the Brown Family corpora (see section 3.3.3).

3.2.2.6 Number of samples

The next sampling decision to consider when trying to create a representative corpus is how many samples you will need to reliably represent the registers within your corpus. Biber (1990) examines mean frequency counts across 10 text samples and five text samples to assess the reliability of each set in representing the extent of internal variation within registers. Biber’s results show a very high level of stability for the linguistic features analysed in the 10 text samples, which leads Biber (1990:

263) to conclude that “the coverage of most categories in the standard corpora, which typically include anywhere between twenty and eighty texts per category, is adequate for these types of analyses.” However, Biber (1993: 253) notes that the linguistic features considered in this study were all very common, and that the study did not address “the representation of linguistic diversity in registers.” Biber (1993) proposes that, in order to calculate the number of samples required to represent registers, a measure of variance within each register must be calculated. Registers with more variation are then allotted proportionally larger samples: a minimum number of samples should be allocated to all registers and then the remaining samples should be distributed proportionally based on the relative variance of each register. Biber (1993: 254) does however stress that this is not the same as proportional representation (as discussed in section 3.2.2.3).

3.2.2.7 Corpus size

The final issue related to sampling, and one which is intertwined with the issues of sample length and number of samples, is corpus size. Many researchers view representativeness and size as connected (Hunston, 2008; Leech, 2007; Biber, 1990; McEnery et al. 2006). Hunston (2008: 165) claims that “some of the difficulties posed by seeking to make a corpus balanced and representative can be lessened by having a corpus large enough for each of its constituent components to be of a substantial size”. Similarly, Leech (2007: 138) suggests that “There is one rule of thumb that few are likely to dissent from. It is that in general, the larger a corpus is, and the more diverse it is in terms of genres and other language varieties, the more balanced and representative it will be.” This view is echoed by the creators of The Corpus de référence du Français contemporain (CRFC) who claim to have included a greater diversity of genres than in any previous corpora, in order to “ensure a reasonable degree of balance and representativeness" (Siepmann et al., 2015: 70). So it seems that some researchers firmly believe that representativeness lies in making a corpus as large as is possible. However, Baker (2009) counters that there is still value in using small corpora, such as those of the Brown Family which are 1 million words in size, which have been carefully balanced and sampled, in order to research fairly common linguistic features. Similarly, McEnery et al. (2006) and Hunston (2008) agree that corpus size cannot be set at a 'one-size-fits-all' level, but rather the appropriate size for a corpus depends on the research aims at hand. Furthermore, Köhler (2013) suggests that no corpus can be large enough to represent all linguistic variation, even in a limited population, because enlarging a corpus causes an increase in the diversity of the data.

Biber (1990: 269) aims to identify how big a corpus needs to be to be representative. He finds that "the underlying parameters of text-based linguistic variation […] can be replicated in a relatively small corpus, *if* that corpus represents the full range of variation". He concludes that "the total number of texts included in existing computer-based corpora are adequate for multivariate statistical analyses" but that more research needs to be done to examine the extent to which existing corpora actually represent the full range of variation in the populations which they claim to represent (Biber, 1990: 269). As with Biber's other (1990, 1993) research discussed above, this conclusion is problematic because Biber's (1990) study focused on common grammatical features. When researching features which are less common, a much bigger corpus will be needed to ensure that there are enough instances of the feature under examination to make the study possible. It is important to note that Biber was writing in 1990, and so the "total number of texts included in existing computer-based corpora" (Biber, 1990: 269) which Biber was discussing would have been significantly less than corpora which are used today. Thus, Biber's conclusion may have strengthened over time, as corpora have increased in size.

3.2.3 Representativeness is not possible

Section 3.2 has so far shown that there has been extensive research into what makes corpora representative; however, I have also shown that there are problems with achieving these ideals for a representative corpus. This leads many researchers to conclude that representativeness, or at least *proving* representativeness, is simply not possible. Hunston (2008: 156) claims that "All corpora are a compromise between what is desirable, that is, what the corpus designer has planned, and what is possible". There are many issues, such as copyright and text availability, which may stop a corpus from being representative even if all of the points above are given thorough consideration (Hunston, 2008). Hunston (2008: 162) goes on to outline three responses to the problem of achieving representativeness. One response would be to forgo the notion of representativeness altogether and simply view a corpus as a collection of different registers which are frequent in the target population, but without any claim of representativeness. This was the approach taken by the creators of the DeReKo corpus (see section 2.7), who do not claim that their corpus is representative, but rather focus on maximising the size of the corpus (Kupietz et al., 2010). A second approach would be to allow the corpus user to assess the degree of representativeness of a corpus by making all of the design decisions taken in the creation of the corpus public. Again, this approach was taken in the creation of the DeReKo corpus, where the goal was to create a large corpus from which users could create their own sub-corpora based on their research needs (Kupietz et al., 2010). Hunston's final response is very similar to her first: to treat corpora as collections of sub-corpora rather than as single entities, however this would only be possible if each sub-corpus was of a “reasonable size” (Hunston gives no indication of what this “reasonable size” is, or how to calculate it).

Köhler (2013: 81) argues that “It is not possible to assess representativeness of a corpus because we lack the theoretical previous knowledge about the hypothetical population that would be needed.” He also suggests that obtaining this knowledge would be impossible because the number of parameters that could be considered is infinite. This leads Köhler (2013: 81) to conclude that “no corpus can be representative in a scientifically meaningful sense, in particular not with respect to statistical methods”.

Leech (2007) agrees that representativeness is something which is unattainable in corpus creation, but he maintains that representativeness is still a goal which should be aimed for. Leech (2007) favours Bungarten’s (1979) idea of an ‘exemplary corpus’ which is a term used when a corpus has been created to be as representative as is possible, but when this representativeness cannot be proven. Leech (2007: 143-144) believes that “We should aim at a gradual approximation of these goals [representativeness], as crucial desiderata of corpus design. It is best to recognize that these goals are not all-or-nothing: there is a scale of representativity […] We should seek to define realistically attainable positions on these scales, rather than to abandon them altogether.”

Many researchers view ‘cyclical procedures’ as the best solution to the problems faced in creating a representative corpus (Atkins et al. 1992; Biber, 1993; McEnery et al., 2006; Bauer and Aarts, 2000). These ‘cyclical procedures’ all involve theoretical research to begin with, which Biber (1993: 243) believes should always be “prior in corpus design”, followed by creation of the corpus, and then testing of the corpus by users to investigate where the corpus is lacking. This procedure is neatly illustrated by Biber (1993: 256) in figure 3a. Biber (1993: 256) claims that “the design of a representative corpus is not truly finalized until the corpus is completed, and analyses of the parameters of variation are required throughout the process of corpus development in order to fine-tune the representativeness of the resulting collection of texts.” Bauer and Aarts (2000) also suggest that all of the decisions made during these cyclical procedures should be well documented, so that corpus users can assess the reliability of their results for themselves. However, these cyclical procedures risk falling prey to the circularity problem mentioned in section 3.2.2.1. If enough research and testing has been done that the linguistic distributions of the population are known, then there is nothing to be gained from analysing the corpus.

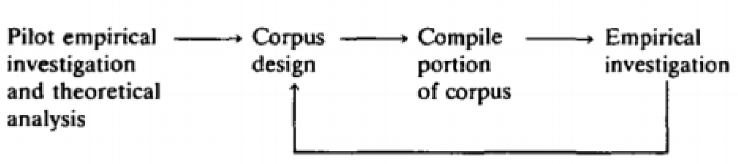

Pilot empirical investigation and theoretical analysis
Corpus design
Compile portion of corpus
Empirical investigation

Figure 3a: Schematic representation of cyclical corpus creation (Biber, 1993: 256).

Some researchers have also highlighted the fact that just because a corpus’s representativeness cannot be proven, that this does not make the corpus useless. Leech (2002: 71) argues that the difficulties presented by representativeness do not justify “a response of extreme scepticism”. Rather, results should be treated as provisional and further research should be done to corroborate findings (Leech, 2002). Atkins et al. (1992: 6) agree with this notion; when discussing balance they claim that any corpus, regardless of how well balanced it is, is a source of information, and that “Knowing that your corpus is unbalanced is what counts.”

3.2.4 The BNC1994

In this section I will outline the approach taken by the creators of the BNC1994 to some of the issues discussed above. In terms of population definition, the creators of the BNC1994 wanted to take account of both language production and language reception (Burnard, 2000). Books form the greatest part of the Written BNC1994 because, although they are written by very few people, they are read by a large proportion of the population (Burnard, 2000). Bestseller lists, prize winner lists, library lending statistics, and periodical circulation figures were used to ensure that where a particular type of text was needed, one with a greater reception was prioritised for collection (Burnard, 2000).

Despite some statistics being available for books, for the majority of the population there were not enough objective measures of the target population for the creators to implement a proportional sampling method (BNC Document Register, 1991). Thus, the creators utilised a stratified random sampling method. Texts were chosen based upon three features (domain, time, and medium), and these selection features were further subdivided into strata with target amounts of text set. These target percentages were decided by the corpus creators, sometimes based on similar factors to Biber’s (1993) suggestion of selecting texts based on cultural importance. For example, it was found that imaginative works accounted for far less than 25% of published and unpublished writing. However, the target percentage was set at 25% because of the “influential cultural role of literature and creative writing” (Burnard, 2000: 7). For books, roughly half of the texts were selected randomly from Whitaker’s ‘Books in Print’ (1992), and the remaining half were chosen systematically, based on the reception criteria outlined above, to fill the remaining target percentages.

The sample size used for books in the corpus was 40,000 words. According to Biber's (1993) work this would appear to be more than ample for reliably representing a text. Texts which were shorter than 40,000 words were reduced by a further 10% to avoid copyright issues. Samples are continuous stretches, and were selected randomly from the beginning, middle, or end of the whole text. Convenient points, such as chapter or section ends, were chosen as end points for samples in order to preserve high-level discourse units (Burnard, 2000). For some text types, such as newspapers, multiple articles were included in one sample, but in these instances articles were always grouped together with other articles from the same domain.

So, the BNC1994 is an example of how all of these problems and recommendations can be dealt with in practice. Of course, the BNC1994 was created before much of the literature discussed in this chapter was written, but it remains a good example of a compromise between what is desirable, and what is possible.

3.2.5 Conclusion

This section has discussed the issue of representativeness in corpus creation, and has investigated some of the problems associated with this issue. The first issue which must be considered is population definition, but, as section 3.2.2.1 showed, even this is not always as straightforward as one may think. Defining a population in any systematic way can often be very difficult, or even impossible. Decisions must then be made about what sampling procedures to use. Proportional sampling seems to be the method which is considered most representative, but again this is often not possible in practice. Sample size, corpus size, and number of text samples are further sampling decisions which must be made. Whilst there has been valuable research into all three of these issues (Biber, 1990; 1993) there is still no definitive consensus on what will achieve representativeness in these areas. Despite all of this research, many researchers still feel that true representativeness is unattainable. However, researchers have stressed that this doesn't mean that representativeness should not be strived for (Leech, 2007), and have proposed cyclical procedures as a way of getting closer to representative corpora (Atkins et al. 1992; Biber, 1993; McEnery et al., 2006; Bauer and Aarts, 2000). Finally, I outlined the decisions of the creators of the BNC1994 in relation to some of these areas, and showed that when it comes to putting these ideas into practice compromises must sometimes be made.

3.3. Comparability in corpus design

3.3.1 Introduction

In this section I will introduce the concept of corpus comparability and discuss some of the different ways in which a corpus can be considered comparable. Of course, this is a very important consideration in the creation of the Written BNC2014 as the corpus will inevitably be used in research which compares the BNC1994 and the BNC2014. It is therefore important to gain a full understanding of the different ways of considering, realising, and using comparable corpora. Firstly, I will discuss the varying definitions of 'comparable' in corpus design. I will then discuss some research into how to create and test comparable corpora. I will also give an overview of two well-known sets of comparable corpora – The Brown Family and ARCHER.

For some linguists the term 'comparable corpus' is synonymous with 'parallel corpus'. A parallel corpus is a corpus which contains (usually) an "authentic translation" (Sharoff et al., 2013: 1) of a corresponding corpus in another language. For example, Sharoff et al. (2013: 3) suggest that there are four different levels of comparability within parallel corpora: parallel, strongly comparable, weakly comparable, and unrelated. Parallel texts are direct translations of the same text, but in another language. Strongly comparable texts are "heavily edited translations" (Sharoff et al., 2013: 3) or strongly related texts which report on the same event or subject in different languages. Weakly comparable texts are from "the same narrow subject domain and genre, but describing different events" (Sharoff et al., 2013: 3) or "texts within the same broader domain and genre, but varying in subdomains and specific genres" (Sharoff et al., 2013: 3). An example of unrelated texts are the majority of texts on the internet, which can still be used for comparative research. For example, two comparable corpora could be created, one representing a random snapshot of the Chinese web and one representing a random snapshot of the French web.

However, McEnery and Hardie (2012) view parallel and comparable corpora as very different things, rather than as varying levels on a single scale. They define a comparable corpus as "a corpus containing components that are collected using the same sampling method", and define a parallel corpus as "a corpus that contains native language (L1) source texts and their (L2) translations" (McEnery and Hardie, 2012: 20). Typically these two types of corpora are used for different types of studies; parallel corpora are used for translation research and comparable corpora are used for contrastive studies (McEnery and Hardie, 2012). Furthermore, McEnery and Hardie (2012: 20) also point out that they are designed with very different focuses:

> "For a comparable corpus, the sampling frame is essential. All the components must match with each other in terms of what types of texts they sample, in what proportions, from what periods. For the translated texts in a parallel corpus, the sampling frame is irrelevant, because all of the corpus components are exact translations of each other. Once the source texts have been selected

> in the first place, there is no need to worry about the sampling frame in the other language.” (McEnery and Hardie, 2012:20).

However, McEnery and Hardie do stress that this does not mean that creating a parallel corpus is easier than creating a comparable corpus.

So far, all of these definitions, regardless of the variation in how the terminology is used, have been based on the premise that a parallel or comparable corpus will vary in the dimension of the language of the texts in the corpora. However, Leech (2007: 141-142) views comparable corpora as “a set of two or more corpora whose design differs, as far as is possible, in terms of only one parameter: the temporal or regional provenance of the textual universe from which the corpus is sampled.” This suggests that comparable corpora cannot only be used to investigate translation or differences between languages, but that they can also be used to investigate diachronic changes or dialectal differences within the same language. In these types of comparable corpora “the language dimension is fixed and it is one of the other dimensions which varies” (Sharoff et al., 2013: 5), i.e. the time period or dialect. It seems that McEnery and Hardie’s (2012) definition of a comparable corpus can be neatly expanded to include this type of comparability: two corpora created using the same sampling frame, but from different time periods or dialects of the same language. An example of these types of comparable corpora are those of the Brown Family (which will be discussed in more detail in section 3.3.3).

As it seems that there are many different definitions for what constitutes a comparable corpus, it is important to consider in what way the Written BNC1994 and the Written BNC2014 will be comparable. The two corpora both represent British English, so are not parallel corpora by any of the definitions given above. Rather, they represent comparability of the kind discussed by Leech (2007) and McEnery and Hardie (2012), that is, they have (as far as is possible, see chapter 4) been created using the same sampling frame and vary only in the dimension of time. Thus, for the remainder of this thesis, when I discuss the comparability of the Written BNC1994 and the Written BNC2014 I will be using this definition of comparability.

3.3.2. Methods for creating and testing comparable corpora

In this section I will discuss some of the methods which linguists have proposed for creating comparable corpora and for testing corpus comparability. The methods for corpus creation which I will discuss are all variants on automatic web crawling – a procedure which allows large corpora to be created very quickly.

3.3.2.1 Creating comparable corpora using web crawling

All of the methods for creating comparable corpora discussed in this section create the type of comparable corpora which Sharoff et al. (2013) would class as ‘unrelated’ and which McEnery and Hardie (2012) would class as ‘comparable’ rather than ‘parallel’. In other words, these methods will create 2 (or more) corpora in different languages using the same sampling method.

The first method for creating comparable corpora which I will discuss uses the BootCaT toolkit (Baroni and Bernadini, 2004) to ‘bootstrap’ corpora from the web. Very briefly the process works as follows:

1. An initial list of words are defined which are expected to be relevant to the domain being researched.
2. The words are randomly combined and used as search terms in a web search engine.

3. The top *n* pages returned by the search engine are selected and converted to text files which are included in the corpus
4. New search terms are then generated from these pages and the process runs iteratively until stopped.

Baroni and Bernadini (2004) use this process to create an English and an Italian corpus, with limited success. Of 30 pages randomly selected from the English corpus, ten were found to be unacceptable, and in a random selection of 30 from the Italian corpus nine were found to be unacceptable. They also found that the newly generated search terms (step 4) were limited in their acceptability (Baroni and Bernadini, 2004). Thus, BootCaT is certainly a useful tool for generating corpora from the web, but is limited in its ability to create comparable corpora which closely match each other for topic.

A similar method for creating comparable corpora from the web is discussed by Talvensaari et al. (2008). This method uses focused web-crawling, and works as follows:

1. A set of URLs which are known to be relevant to the topic to be collected are specified.
2. These URLs are placed in a queue, and are loaded one by one.
3. Out-links of the page are extracted and added to the queue.
4. The queue can be prioritised using a *driver query* (words from the desired domain) which is specified at the start of the process. Pages are compared to the driver query to see which are most relevant to the topic.
5. Process continues until the queue is empty or the process is stopped manually.

Talvensaari et al. (2008) use this method to create two comparable corpora in order to test domain-specific translation.

Ghani et al. (2005) propose another method, called ‘CorpusBuilder’, for generating comparable corpora from the web. This process works in a similar way to those outlined above:

1. Decide on two sets of initial documents: one set which has been judged to be relevant to the given query, and one set which has been judged as irrelevant.
2. These documents are used for query generation based on the odds-ratio of each word (probability of the word occurring in the relevant and non-relevant documents).
3. The three words (both positive and negative) with the highest odds-ratios are used for query generation.
4. After each retrieval operation, the first document is automatically analysed to check that it is in the target language.
5. If the document is in the target language then the set of documents are updated and query generation is performed again. If the new document does not change the query then the next document is used.
6. This process is performed iteratively.

Ghani et al. (2005) illustrate that this method can be used to create corpora of under-resourced languages, such as Slovenian.

A problem with all of these methods is that by using a set of initial queries or pages, you end up only finding what you set out to find, at least in terms of topic. For the creation of comparable corpora where it is the language dimension which varies this poses no problems, rather, it is the whole point of the process. However, for the creation of corpora which will be comparable across time periods this is detrimental. By limiting what you collect to what is identified using your initial search terms you risk missing out on the collection of new topics which have become relevant since the earlier corpus was created (e.g. 'Brexit' is a new topic within politics which did not exist just a few years ago). Furthermore, it is unlikely that these processes would be suitable for the collection of corpora where the time dimension varies. These web-crawling methods do not take account of the date that a text was published, and so cannot limit the crawl to texts published in a particular time period. Jakubíček et al. (2013: 126) state that, in the creation of the enTenTen corpus, most web pages do not reliably state when a text was written; the only information available is the date that the crawl was carried out. Le and Quasthoff (2016), in their construction of the Vietnamese Corpus, search the web-crawled corpus for the frequencies of the years 1980-2030, and propose that "the distribution of these numbers is strongly correlated with the origin of the texts" (Le and Quasthoff, 2016: 412). However, this is purely an assumption, and relies totally on the date of publication being listed within the text of the web page. Of course, an archive such as the WayBack machine could be used to access web pages from a particular time period (Arora et al., 2015). However, this presents a similar problem in that the search will return anything present on a website within the analysts selected time period, and does not guarantee that the text was written on that particular date. Also, if the corpus being created aims to represent texts from any further back than the early 2000s, then there would probably not be enough texts from the target time period which have been digitised and put online anyway. Similarly, these methods are all designed to aid collection of texts about very specific and narrow topic fields, which would be unsuitable for the creation of comparable general corpora. The fact that all of these methods use web-crawling techniques is also problematic in some cases. For example, one of the main aims of the Written BNC1994 was to be non-opportunistic (see Chapter 1), and as such the Written BNC2014 will also be created non-opportunistically. A web-crawl is an opportunistic method of data collection, and so these techniques are unsuitable for anyone wishing to create comparable corpora non-opportunistically. All of these factors make these methods unsuitable for collecting data for the Written BNC2014.

3.3.2.2 Testing corpus comparability

Another problem with creating corpora from the web in the ways described in section 3.3.2.1 is that once they have been created, unless they are very small, we cannot know their composition (without undertaking an extremely time consuming manual analysis). Sharoff (2013) notes that this problem is exaggerated when creating and using comparable corpora because we cannot know if "we get comparable pages by sending comparable queries" (Sharoff, 2013:114). Sharoff (2013) proposes a method of analysing the contents of corpora generated from the web. Very briefly, this method involves statistically identifying 'clusters' and 'topic models' within the corpora under evaluation which can then be compared to those identified in the other corpora to which they are claimed to be comparable, in order to assess the level to which these corpora are truly comparable with one another (Sharoff, 2013). This could be a potentially useful way of analysing how comparable the Written BNC2014 is to the Written BNC1994; however, it does of course require both corpora to be created before such an analysis can be performed. Thus, this method cannot help us to *create* a corpus which is comparable to the Written BNC1994, although it may be useful in designing the comparable sub-corpus of the Written BNC2014 (discussed in section 3.4).

In addition to Sharoff's (2013) method for testing the comparability of corpora, other methods have been designed by linguists. Kilgarriff (2001) presents a method for use with monolingual corpora, but Sharoff (2013) notes that this method could also work for testing the comparability of corpora of different languages. Kilgarriff (2001) proposes a method for testing corpus similarity using 'Known Similarity Corpora' (corpora composed of documents judged to be similar within categories, but different across categories). The distance between the corpora under question can then be measured by the overlap in their keywords.

Köhler (2013) also outlines a method for assessing the comparability of corpora, but first notes that "If a systematic test for comparability is intended, a number of predicates come into play which are logically connected to comparability and must be discussed before." (Köhler, 2013:80). These predicates are representativeness, homogeneity, homoscedasticity and skewness, and corpus balancing (see section 3.2 for a discussion of some of these issues). Once these issues have been considered, Köhler (2013) outlines a method for testing the comparability of a corpus whilst creating it. The method (greatly simplified) works as follows (for creating comparable corpora in different languages):

1. Firstly, you must have one corpus already created, and wish to create another, comparable corpus in another language.
2. Create documents which are direct translations of some of the texts in the already created corpus.
3. Use statistical tests to determine how the translations behave in relation to the original corpus for a parameter which you are interested in (Köhler, 2013, uses the example of sentence length). This becomes your hypothesis upon which your text collection will be based.

4. Then test the documents which are being considered for inclusion against this hypothesis. If they do not fulfil the hypothesis then they cannot be included in the corpus; if they do then they are included.

This method must be repeated for every parameter which is expected to be relevant to the comparable corpora, so would be extremely time consuming. It must also be noted that by collecting a corpus based on parameters which you expect to be relevant, you will greatly limit the diversity of the corpus, and, similarly to the above, may only find what you set out to find.

3.3.3 The Brown Family

In this section I will introduce the Brown Family of corpora, and outline some of the research which has been done using it. The Brown family consists of multiple corpora, which are all considered to be comparable in McEnery and Hardie's (2012) sense of comparable corpora.

The first member of the Brown Family was the Standard Corpus of Present-Day American English (later renamed the Brown Corpus) which consists of approximately 1 million words of American English prose produced during 1961 (Francis and Kučera, 1979). The corpus contains 500 samples of 2,000 words each, with samples representing a wide range of styles and varieties. The corpus was built in two phases: an initial classification of samples and decisions regarding how many samples of each category would be included, and then a random selection of the samples for each category (Francis and Kučera, 1979). This sampling frame then became the model for all subsequent members of the Brown family which have been created (see table 3b for the sampling frame, and table 3c for all members of the Brown Family). As can be seen from table 3c, there are many members of the Brown

Family, and all represent a particular language variety at a particular point in time. The fact that they are all created according to the same sampling frame means that they can be used to make diachronic comparisons within language varieties and comparisons between language varieties at various time periods.

Table 3b: Sampling frame for the Brown family of corpora (McEnery and Hardie, 2012: 97).

Text categories	**Broad Genre**	**No. of texts**	**% of corpus**
A Press: reportage	Press	44	8.8
B Press: editorial	Press	27	5.4
C Press: reviews	Press	17	3.4
D Religion	General prose	17	3.4
E Skills, trades and hobbies	General prose	36	7.2
F Popular lore	General prose	48	9.6
G Belles lettres, biography, essays	General prose	75	15
H Miscellaneous (government & other official documents)	General prose	30	6
J Learned and scientific writings	Learned	80	16
K General fiction	Fiction	29	5.8
L Mystery and detective fiction	Fiction	24	4.8
M Science fiction	Fiction	6	1.2
N Adventure and western fiction	Fiction	29	5.8
P Romance and love story	Fiction	29	5.8
R Humour	Fiction	9	1.8

Table 3c: Corpora within the Brown Family.

Corpus	**Language variety**	**Period**
B-Brown	American English	1931 +/- 3 years
Brown	American English	1961
Frown	American English	1991-1992
AmE06	American English	2006 +/- 1 year
BLOB	British English	1931 +/- 3 years
LOB	British English	1961
FLOB	British English	1991-1992
BE06	British English	2006 +/- 1 year
Kolhapur	Indian English	1978
ACE	Australian English	1986
WWC	New Zealand English	1986-1990

Much of the research done using the Brown Family has, unsurprisingly, focused on the investigation of diachronic change in the languages within the family which have multiple corpora from different time periods. Much of this diachronic style of research has focused on researching the specific social change of ‘colloquialisation’, which Leech (2002: 72) defines as “a tendency for the written language gradually to acquire norms and characteristics associated with the spoken conversational language” (see chapter 9 for an exploration of this issue using data from the Written BNC2014). Mair (1997: 206) notes that in studies which have compared the LOB and FLOB corpora, “very few genuine instances of grammatical change were noted” and instead suggests that most changes are simply “a result of the colloquialisation of the norms of written English which has taken place over the last thirty years”. For example, Leech (2002) compares LOB and FLOB and finds that there is a trend of colloquialisation (features typical of spoken language spreading in written language; see chapter 9). The findings indicating colloquialisation include the use of the present progressive construction increasing, contractions increasing, a decline in the use of the passive, and an increase in questions (Leech, 2002: 74).

Furthermore, Mair et al. (2003) compare tag frequencies in LOB and FLOB and their findings echo those of Mair (1997) and Leech (2002). They conclude that the change in tag frequencies which they observe, for example a 7.3% rise in verbs in the reportage samples, is not a direct indicator of grammatical change but is rather a style change indicative of colloquialisation. Baker (2009) builds on this research by comparing pronoun usage in BLOB, LOB, FLOB, and BE06. He concludes that "the higher frequencies of first and second person pronouns in the BE06 are indicative that colloquialisation or 'involved' discourse appears to be higher now in written British English than in previous sampling periods" (Baker, 2009: 327), but does note that more linguistic features would need to be investigated before stronger claims could be made.

The corpora of the Brown family have also often been used to investigate cultural differences and cultural change. Leech and Fallon (1992) compare word frequency lists in Brown and LOB in order to attempt to identify cultural differences between America and Britain. They take the linguistic items with the greatest significance from each corpus and categorise them into groups such as 'sport', 'business', 'military' and 'education' in order to see the differences between the two cultures. Leech and Fallon (1992: 44-45) sum up their findings as follows (although they do acknowledge that this is a "wild generalisation"):

> [W]e may propose a picture of US culture in 1961 - masculine to the point of machismo, militaristic, dynamic and actuated by high ideals, driven by technology, activity and enterprise - contrasting with one of British culture as more given to temporizing and talking, to benefitting from wealth rather than creating it, and to family and emotional life, less actuated by matters of

> substance than by considerations of outward status. (Leech and Fallon, 1992: 44-45).

Much research since has built on these findings using more corpora. Oakes (2003) conducts a very similar study using FLOB and FROWN in order to see if the cultural differences identified by Leech and Fallon (1992) still held true 30 years later. He found that some cultural differences had changed, for example, America no longer had a greater interest in sport or transport than Britain, and America had lost the masculine bias found in Leech and Fallon's (1992) study. However, on the whole Oakes (2003) found that most of the differences found by Leech and Fallon (1992) "still held true for a comparison of UK and US English using texts written in the 1990s" (Oakes, 2003: 221). Baker (2011) uses BLOB, LOB, FLOB, and BE06 to investigate language change within British English. Amongst his findings about linguistic features such as grammatical change, Baker (2011) also hypothesises that some of his findings could be indicative of cultural change. For example, the word 'children' was found to increase in frequency over time which could represent a cultural shift towards greater anxiety about dangers posed to and by children. Potts and Baker (2012) draw together these cross-cultural comparisons and diachronic investigations in their study which investigates whether semantic tags can show cultural change using Brown, Frown, AmE06, LOB, FLOB, and BE06. Potts and Baker's (2012) findings largely echo those of Leech and Fallon (1992) and Oakes (2003) in that they note "the continued focus of British English on words to do with time, if's, but's, and modality, and the continued American English focus on the military and weaponry, and IT and computing" (Potts and Baker, 2012: 321). However, Potts and Baker (2012) do admit that they are hesitant to conclude that their findings are firm proof of actual cultural differences, as many of the observed differences may simply be 'topic' differences.

Other researchers have used the Brown Family to compare aspects of British and American English. Hundt (1997) uses Brown, Frown, LOB, and FLOB to investigate whether British English has been catching up with American English in terms of morphological, syntactic and lexico-grammatical change. Hundt (1997: 146) finds that “AmE, with the occasional exception, is usually more advanced in ongoing morphological and syntactic changes”. Baker (2017) uses the Brown Family of corpora to compare a wide variety of linguistic phenomena in American and British English. Baker (2017) presents far too many findings to detail them all here, but this expansive study makes full use of the Brown Family of corpora to investigate differences and similarities between the two varieties. Baker (2017: 237) finds that American English is “at the forefront of change at the grammatical level”, or, in other words, British English is lagging behind American English in terms of grammatical changes. However, this same trend of ‘Americanisation’ did not hold true for spelling differences or semantic tag use (Baker, 2017: 237). Baker (2017: 236) links his findings to six major trends: “Americanisation, densification, democratisation, informalisation/colloquialisation, grammaticalisation and technologisation”.

Another use of the Brown family has been to investigate the differences between text types. Johansson (1985) compares the various LOB text categories and finds the most striking and consistent differences are between ‘fiction’ and ‘learned and scientific English’. They find that, amongst other things, verbs predominate in fiction texts whilst nouns predominate in learned and scientific texts; fiction favours adjectives describing personal qualities whilst learned and scientific texts favour adjectives describing non-personal qualities; and fiction texts favour the past tense whilst learned and scientific texts favour the present tense.

3.3.4 ARCHER

Another collection of corpora which allows for similar comparisons to those facilitated by the Brown Family is the ARCHER collection. ARCHER was designed to "investigate the diachronic relations among oral and literate registers of English between 1650 and the present" (Biber et al., 1994:1). ARCHER represents both written and spoken British and American English, and the breakdown of the corpus can be seen in table 3d. In total ARCHER contains approximately 1.7 million words, with around 2,000 words per register (see table 3e) in each corpus (Biber et al., 1994: 4).

Table 3d: Chronological and geographical coverage of ARCHER (Biber et al., 1994: 3).

British	**American**
1. 1650-1699	
2. 1700-1749	
3. 1750-1799	4. 1750-1799
5. 1800-1849	
6. 1850-1899	7. 1850-1899
8. 1900-1949	
9. 1950-1990	10. 1950-1990

Table 3e: The registers of ARCHER (Biber et al., 1994: 4).

Written	**Speech-Based**
Journals-Diaries	
Letters	
Fiction	Fictional conversation
News	Drama
Legal opinion (1750-; USA only)	Sermons-Homilies
Medicine (excluding 18th-cent. USA)	
Science (British only)	

Similarly to the Brown Family of corpora, and perhaps unsurprisingly, a predominant use of ARCHER has been for diachronic investigations. For example, Broccias and Smith (2010) use the British component of ARCHER to investigate the diachronic change of the simultaneity subordinator ‘as’. They find that there is “a dramatic increase in the frequency of simultaneity as-clauses from the first half of the nineteenth century onwards” (Broccias and Smith, 2010: 348). They hypothesise that “the spread of ‘as’ may be symptomatic of an evolution in narrative techniques, particularly in respect of the means by which complex events are typically represented” (Broccias and Smith, 2010: 348). Biber and Gray (2011) use ARCHER to investigate grammatical change in the noun phrase, and to consider whether linguistic innovation always occurs in spoken language before written. They research historical patterns in the use of, amongst other features, nouns as nominal premodifiers and prepositional phrases as nominal postmodifiers, and find that whilst “It is not possible to prove that these constructions were first used in writing rather than in speech” (Biber and Gray, 2011: 247), it is clear that they have become characteristic of written rather than spoken discourse over the past two centuries.

Another similarity in the use of the Brown Family and ARCHER is that ARCHER has also been used to investigate differences between text types. Pérez-Guerra and Martínez-Insua (2010) compare the lexical and syntactic complexity of the news and letters text types in the British component of ARCHER. They find that the proportion of pronominal subjects is greater in the letters, the proportion of non-pronominal subjects and objects is greater in the news texts, and the average length of syntactic units in the news texts is greater than in the letters (Pérez-Guerra and Martínez-Insua, 2010). They interpret these findings as showing that news texts have a greater level of complexity than letters. Additionally, the diachronic nature of ARCHER allowed the researchers to show that these differences have not varied greatly over the last three centuries (Pérez-Guerra and Martínez-Insua, 2010).

3.3.5 Conclusion

This section has introduced the concept of corpus comparability, and has considered what will be meant by 'comparability' in reference to the Written BNC2014 throughout this thesis. Comparable here has nothing to do with comparable corpora which represent translations of texts in multiple languages, but rather refers to diachronic comparability, where there are two corpora, both created using the same sampling frame, which vary only in the dimension of time.

I discussed web-crawling methods which have been tested for creating comparable corpora, but found that, for various reasons, none of these would be suitable for collection of the Written BNC2014. I also discussed some methods for testing the comparability of two corpora, and argued that although Sharoff's (2013) method of identifying and comparing clusters and topic models would not be suitable for assessing the whole of the two corpora, it may be useful for guiding our creation of the comparable sub-corpus of the Written BNC2014 (see section 3.4.2).

I have also introduced two collections of comparable corpora (the Brown Family and ARCHER) and given a brief overview of the kinds of research done with these corpora in order to give an idea of the kinds of things that the Written BNC2014 may be used for. In section 3.4 (and chapter 4) I will discuss the decisions made about how I will ensure that the Written BNC2014 is comparable to the Written BNC1994. I will also discuss the important interaction between comparability and representativeness, and the extent to which this will impact on how comparable the corpora will be.

3.4. Representativeness vs. Comparability

3.4.1 The Problem

As sections 3.2 and 3.3 have shown, both representativeness and comparability are complex and important issues which must be considered when creating a corpus. However, they can often be at odds with one another. Leech (2007:142) points out that “an attempt to achieve greater comparability may actually impede representativity and vice versa”. This is because of ‘genre evolution’, where over time new genres emerge and old genres decay. Thus, corpora which are created to be comparable to corpora from a previous time period may lose their representativeness because they must include old genres which have disappeared, because they were included in the older corpus, and cannot include new genres which have emerged, because they were not included in the older corpus. Of course, as well as decaying or emerging, genres can also ‘shift’. For example, there is no guarantee that a genre labelled ‘X’ in a corpus 20 years ago will contain the same type of data as a genre labelled ‘X’ nowadays. A good example of this is newspapers: in the past there was a clear distinction between ‘broadsheet’ and ‘tabloid’ new articles, however this distinction has lessened over time and what was once a broadsheet article may now be classified as a tabloid article. This is certainly the case in the BNC1994 and 2014, and has resulted in these genres been given different labels in the 2014 corpus (see section 6.5 for a full discussion of this). Baker (2009: 335) discusses genre evolution in relation to the Brown Family of corpora. He considers “whether a model that was developed in the early 1960s will always be appropriate”. For example, there was a great amount of science fiction being published in the 1960s, when the Brown Family sampling frame was first created. This has resulted in all members of the Brown Family having to include a greater amount of Science Fiction than is representative of the time period of the corpus being created because it is included in the original sampling frame. In creating the BE06 corpus, Baker (2009) only included texts which were originally published in paper form in order to stick more closely to the original sampling frame. However, he concedes that "if we limit corpus building projects to just texts that were originally published in paper form (as I did with the BE06), we risk building a rather anachronistic and idiosyncratic corpus that does not reveal much about the true pattern of language use in the twenty first century" (Baker, 2009: 335).

This problem was one which was encountered in the early stages of creating the sampling frame for the Written BNC2014 (see chapter 4). Once an initial version of the sampling frame had been created I sent it to various experts in corpus creation in the hope of getting their feedback on how the sampling frame could be improved. I contacted 27 experts, some because they had worked extensively with the BNC1994 in the past (either on the construction of it, or using it as a data source), some because they are established experts in the field of corpus linguistics, and some because they represented the end-users of the corpus. All of the experts were sent an email which introduced them to the Written BNC2014 project, and contained an attachment detailing (both in brief and in full) the decisions made in the creation of the initial sampling frame. All experts were asked to "take a look at either the executive summary or, if you prefer, the full document and provide us with any comments, suggestions, or opinions you may have".

Of the 27 emails sent, I received 11 detailed responses. One of the most obvious, and most often repeated, pieces of advice which I received was that the corpus was not comparable enough to the Written BNC1994 to be useful for diachronic studies, but was also not representative enough of current British English to allow research on contemporary language. A piece of feedback which neatly highlights this issue is "It's impossible to maximize both representativeness and comparability at the same time." In other words, in trying to make the corpus both comparable to the Written BNC1994 and representative of current British English, I had actually achieved neither to a sufficient degree.

Five of the respondents[4] felt that I should prioritise the comparability of the corpus:

> "For me personally, comparability with [the BNC1994] is probably more important than the representativeness issue".
>
> "For me, the effort to be representative is a bit of a wild-goose chase… For me, the top criterion would be EASY COMPARABILITY across the OLD AND NEW BNC."

However, there was one expert who felt that representativeness of contemporary language was the more important criterion, in order to make the project valuable in the future. Despite being the only person to explicitly state this opinion, this view has been given a relatively high weight in my decision making as this comment was from an expert who had not worked with the BNC1994 extensively and represented an end-user who was not invested in the past of the BNC project as many of the other respondents were.

[4] The names of the respondents are not given here as these comments were made as part of a confidential, early-stage consultation.

> “I think one thing to bear in mind is that the model doesn't just need to look backwards to be a match to the old BNC, but it should also try to be forward thinking - what will the next BNC look like (say in 30 years’ time?) And as language use changes, will sticking to an old model start to increasingly make the BNC project feel outdated and unworkable?”

3.4.2 The Solution

Clearly the sampling frame which I initially created was not suitable for either diachronic purposes, or for investigating contemporary British English. Thus, a resolution had to be found for this issue. Initially, it seemed that I would have to choose either comparability or representativeness as our top criterion and accept that whichever one I chose would limit the usefulness of the corpus in respect to the other. Personally, I felt that representativeness should be prioritised as this would ensure the longevity of the project and would avoid the problems encountered by the Brown Family, discussed above. On the other hand, I could absolutely see that diachronic studies would be an important and very interesting function of the Written BNC2014.

Despite it seeming initially impossible, I managed to arrive at a solution to the problem of representativeness and comparability without having to choose one or the other. In designing the corpus I have prioritised representativeness of contemporary British English. This takes the form of, for example, including new genres such as ‘e-language’ and by altering the proportions of genres compared to the Written BNC1994 (see chapter 4 for a full discussion of the sampling frame). The updating of the BNC1994 genres has a precedent in the American National Corpus (Ide, 2008; see section 2.5), which aimed to follow the framework of the BNC1994 but included ‘new’ genres which had emerged since the BNC1994, such as e-language.

However, once the corpus has been created I will create a sub-corpus which will be fully comparable to the Written BNC1994 (this will be done after the completion of this thesis, and, thus, will not be discussed further here). Thus, the corpus will be representative of contemporary British English, but it will also be possible for diachronic studies to be carried out using the comparable sub-corpus.

It is of course important to remember all that was discussed in section 3.2, which indicated that achieving full representativeness of a language is often impossible, or at least impossible to prove. I will not have time on this project to put any of the cyclical procedures mentioned into practice, and many of our sampling decisions will be influenced by availability of data. However, as recommended in section 3.2, I will strive for representativeness as much as I can whilst acknowledging that this will not be achieved perfectly. I will also provide users of the corpus with clear descriptions of how the corpus was created so that they can make their own assessments of the representativeness of the corpus. These issues will be reflected upon and discussed further in chapter 4, where I detail the design of the Written BNC2014 sampling frame.

Chapter 4: Designing the Written BNC2014 Sampling Frame

4.1 Introduction

In this chapter I will introduce the Written BNC2014 sampling frame. Section 4.2 discusses how the texts included in the corpus were classified in the sampling frame. Section 4.3 considers the design of the sampling frame, and returns to many of the concepts discussed in chapter 3 when considering creating representative corpora. I will discuss the decisions made relating to population definition, sample size, number of samples, corpus size, and sampling methods when designing the Written BNC2014 sampling frame. Section 4.4 considers the sampling frame in relation to its comparability with the Written BNC1994, both in terms of the genres included in the corpus, and the proportions in which these genres are represented. I finish, in section 4.5, by summarising how the design of the Written BNC2014 sampling frame will affect the representativeness and comparability of the corpus.

4.2 Classifying texts in corpora

4.2.1 Introduction

In this section I will explain how the texts within the Written BNC2014 sampling frame were classified and labelled. I first briefly introduce the concept of *genre theory*, and consider what this approach can bring to the discussion of the use of the term *genre* in corpus creation. I will then consider some of the most common ways of classifying texts in linguistics, namely *genre*, *register*, *style* and *text type*, before settling on clear definitions for these terms which will be used consistently throughout this thesis. I will then look at how the texts in three previous national corpora (the Brown Family, the CRFC, and the BNC1994) were classified. Finally, I will bring all of this information together to come to a conclusion about how the texts in the Written BNC2014 will be classified.

4.2.2 Genre Theory

4.2.2.1 Introduction

In this section I will discuss briefly the study of *genre theory*. Genre theorists are concerned with defining what a *genre* is, finding systematic ways in which to classify genres, and examining the way different structures of meaning are created through the various genres of writing which exist (Frow, 2006). I will only discuss genre theory briefly here as a full account would not be relevant to the aim of this section (to discover how different terms, including genre, have been used in the design of previous corpora). Moreover, as will be seen in section 4.2.2.2, there seems to be a general consensus that it is not possible to come up with a full list of genres, but it would be difficult to discuss genre without a mention of this extensive area of research. I will not explore the history of genre theory here, but, for those who are interested, Frow (2006) and Duff (1999) both provide interesting accounts of the history of genre theory.

4.2.2.2 Approaches to genre

As mentioned in section 4.2.2.1, many genre theorists are concerned with how to classify genre into *genre-systems*. Duff (1999: xiii) defines a genre-system as “a set of genres that is understood to form a coherent system of some kind; or a theoretical model that offers a comprehensive list of genres and an explanation of the relations between them”. In this section I will discuss some of the ways of classifying genres that genre theorists have proposed.

Fowler (1982) discusses the idea of classifying genres according to a logic of family resemblance. In this theory texts could belong to a genre if they had some common features, without necessarily all having any single feature in common. Frow (2006) extends this theory by discussing genres in terms of prototypes. So, it would be possible to think of a text which is a prototypical member of a genre and then classify other texts according to how similar to the prototype they are. However, Frow (2006) also points out that this still leaves the problem of how to know when a text is too dissimilar to the prototype to be included in the genre.

Another way in which genre theorists have framed genre is in terms of situation and behaviour. Frow (2006) discusses the idea that a text cannot actually *be* a particular genre, but rather it *participates in* one or more genres. Frow (2006) views situation as a very important part of this, defining genre as a relationship between a text and the situation that it occurs in. He then goes on to demonstrate how this can be seen in everyday life: genre tells us how to behave in certain situations by, for example, showing us, through a combination of text and situation, whether a story should be taken seriously or whether it is a joke. Dubrow (1982: 2) points out that genre, similarly, "functions much like a code of behaviour established between the author and his reader".

Frow (2006) and Rosen (2013) (amongst others) express the opinion that there is not, and cannot be, a complete list of all genres and how they relate to each other. In fact, Frow (2006: 2) begins his book by stating that he is not concerned with classifying genres or comprehensively covering the full range of genres because he believes that there is no "master list". Rosen (2013) echoes this, pointing out that the genre structures which have been invented are not fixed structures which have been deduced from empirical investigation, and further highlights that there is not even a consensus on what the word genre can be used to mean. Frow (2006: 52) believes that ways of thinking about genre using metaphor (such as the 'family' and 'social behaviour' metaphors discussed above) are ways "of thinking systematically about a form of ordering that is in many ways resistant to system".

Thus, it seems clear from this very brief overview of genre theory that this approach to the study of genre will not be relevant to the Written BNC2014 project. Whilst it is interesting to theorise about how one could classify genres within this perspective, it seems that there is no agreement on how one might do this. Thus, it will be more useful to look into how genre has been approached by corpus linguists previously, to see what approaches to classification have worked in the past. This will be addressed in section 4.2.4 of this chapter.

4.2.3 Genre, register, style, text type – some definitions

4.2.3.1 Introduction

As well as the term *genre*, the terms *register*, *style*, and *text type* are often used by linguists to describe and categorise the texts which they are working with or studying. However, the definitions of these terms are often unclear, overlap, and are used differently by different linguists. Biber and Conrad (2009: 21) note that "the terms register, genre, and style have been central to previous investigations of discourse, but they have been used in many different ways". They emphasise the importance of being aware that "there is no general consensus concerning the use" of these terms. Many other linguists also point out that these terms are used differently, and sometimes interchangeably, in the literature (Lee, 2001; Nunan, 2008; Taavitsainen, 2001). Lee (2001) believes that the terms genre and register are the most confusing precisely because they are often used interchangeably. Biber and Conrad

(2009) support this by pointing out many studies where one of these terms is adopted and the other simply disregarded. However, some linguists do make a distinction between the two terms, but define them very differently. For example, Taavitsainen (2001: 141) defines register as a broad term such that one register "may contain several genres", whereas Nunan (2008: 59) believes that register "offers a more fine grained analysis than genre", where the analyst begins by analysing the genre of a text and then goes deeper into the text to perform a "more fine grained register analysis" (Nunan, 2008: 60). So it seems that there are different definitions of these terms being used which are directly contradictory and irreconcilable.

This section aims to disentangle these definitions and develop clear definitions for these terms which will be used throughout this thesis. The following sections will consider some of the most widely used definitions of the different terms; there are many studies which use these terms in subtly (and sometimes less subtly) different ways, but due to space constraints I have limited my discussion to those definitions which are most commonly used.

4.2.3.2 Genre

Genre is a term which can be defined in terms of culture, and analysed in terms of linguistic factors. Some literature takes account of both aspects. For example, Hyland (2009: 15) defines a genre as a text which "has a specific purpose, an overall structure, specific linguistic features, and is shared by members of the culture." Much of the literature which focuses on defining genre focuses on the cultural context. Taavitsainen (2001: 139-140) states that genres "are inherently dynamic cultural schemata used to organise knowledge and experience through language". Trosberg (1997: 6) also views genre in cultural terms, stating that genres "are the text categories readily distinguished by mature speakers of a language". Lee (2001: 38) states that genres "have the property of being recognised as having a certain legitimacy as groupings of texts within a speech community".

However, Biber and Conrad (2009) emphasise the linguistic aspects of analysing genre. They state that in order to analyse a text from the genre perspective, account must be taken of purposes, situational context, and conventional structures within a complete text. The genre perspective often focuses on linguistic features which only occur once within a text, and for this reason Biber and Conrad (2009) state that genre analysis should only be performed on complete texts.

There are criticisms of these ways of defining and analysing genre. Trosberg (1997: 12) points out that "Texts within particular genres can differ greatly in their linguistic characteristics… On the other hand, different genres can be quite similar linguistically". Nunan (2008) also points out the 'fuzziness' of defining genre in these ways. He notes that there is great difficulty in knowing when two texts are different enough from each other to represent two different genres.

4.2.3.3 Register

Whilst the definitions of genre, discussed above, were concerned with culture, definitions of register are concerned with situation. Nunan (2008: 59) states that register analysis takes account of three situational variables: the subject matter of the text (field), the relationships between the producers and receivers of a text (tenor), and the channel of the communication (mode). Crystal (2008a: 295) defines register as "a variety of language defined according to its use in social situations". Lee (2001: 46) refers to this as "variety according to use".

Biber and Conrad (2009) also favour this situational definition of register, and contrast register analysis with genre analysis. Whilst genre analysis requires complete texts in order to find linguistic features which may only occur once within a text, a register analysis can be performed on any excerpt of a text because a register analysis "focuses on the pervasive patterns of linguistic variation" (Biber and Conrad, 2009: 23). Biber and Conrad (2009: 6) explain that "linguistic features are always functional when considered from the register perspective", or in other words, particular linguistic features occur in texts because they are particularly well-suited to the situational characteristics of that register.

Trosberg (1997: 6) argues that register analysis reveals relatively little about genres, and so registers are sub-divided into genres in order to reflect "the way social purposes are accomplished in and through them in settings in which they are used". Lee (2001: 46) provides a good example of this: "we talk about the existence of a *legal register* (focus: language), but of the instantiation of this in the *genres* of 'courtroom debates,' 'wills' and 'testaments,' 'affidavits,' and so forth (focus: category membership)".

4.2.3.4 Style

Biber and Conrad (2009: 23) state that, commonly, style "has been treated as a characteristic way of using language." Lee (2001: 45) similarly defines style as "to do with an individual's use of language." This perspective is often applied to literary language and is termed *stylistics* (Biber and Conrad, 2009: 23). This notion can also be applied to the study of conversational interactions, "where cultures can be described as having distinctive conversational styles" (Biber and Conrad, 2009: 23). The analysis of style can be seen to be similar to the analysis of register, because it focuses on linguistic features which are distributed throughout text samples in a variety. However, it is different to a register analysis because in the style perspective these linguistic features are due to the aesthetic preferences of particular authors or particular time periods, rather than being situationally motivated (Biber and Conrad, 2009: 2).

4.2.3.5 Text type

Biber (1989: 39) identifies text types as being defined strictly by linguistic criteria (as opposed to genres which are defined according to non-linguistic, cultural aspects). Thus, text types often cut across genre distinctions because "Linguistically distinct texts within a genre represent different text types; linguistically similar texts from different genres represent a single text type" (Biber, 1989: 6). Paltridge (1996) (referenced in Lee, 2001) proposes some examples of text types: 'procedure', 'anecdote', 'description' etc. However, Lee (2001: 40) points out that these would be better termed "discourse/rhetorical structure types" because the determinants are rhetorical features rather than Biber's (1988, 1989) "internal linguistic features".

Lee (2001: 41) is of the opinion that text type is still an "elusive concept which cannot be established explicitly in terms of linguistic features".

4.2.3.6 Terminology in this thesis

Moving forward in this thesis I will use these terms according to the definitions outlined below. I have chosen these definitions both because they are those most commonly found in the literature studied, and because they are the definitions which most convincingly differentiate the terms clearly.

Genre: A category of texts which is easily recognised by a member of the culture. Genres can be identified using external, non-linguistic criteria.

Register: A category of texts which are recognised according to their situation of use.

Style: A particular characteristic way of using language (e.g. a particular author's or time period's style; this is a term most often used in the literary analysis of language and was the least exemplified in the literature reviewed).

Text type: A category of texts which have similar internal, linguistic features.

4.2.4 Text classification in previous national corpora

When considering how to classify the texts included in the Written BNC2014, it has been important to consider how other national corpus projects have approached this issue. Thus, in this section I will briefly outline how previous national corpus projects have approached the classification and labelling of texts. This will allow me to see if there is a common standard for the classification of texts in national corpora, and also allow me to assess the success of the various decisions which have been made in corpora previously, in order to make an informed decision about how texts will be classified in the Written BNC2014. I will focus on the Brown Family of corpora, the Corpus de référence du Français contemporain (CRFC), and the British National Corpus 1994.

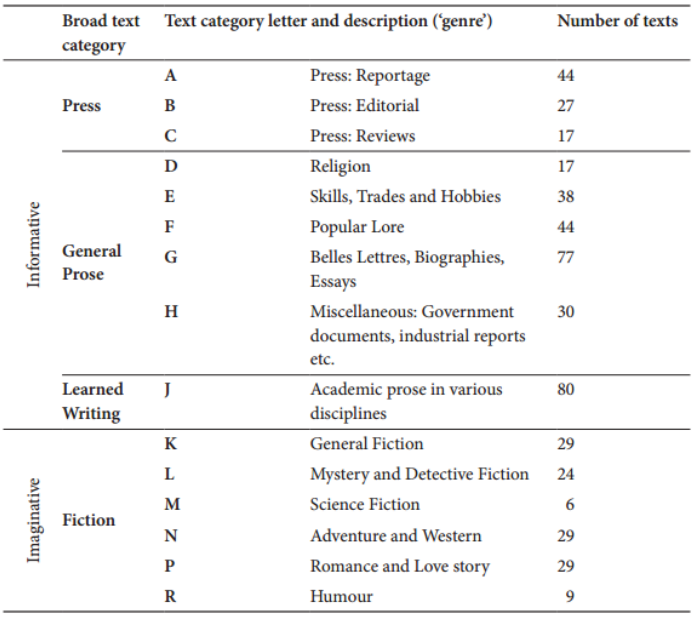
4.2.4.1 The Brown Family

As discussed in section 3.3.3 the Brown Family is a collection of corpora which all contain approximately 1 million words of some national language variety from a particular time period. All of the corpora within the family have been created using the same sampling frame (see section 3.3.3 for more details). In the Brown Corpus Manual, Francis and Kučera (1979) simply refer to the texts in the corpus being split into *categories*, rather than *genres* or *registers* etc. They do make one reference to the term *style* – "The samples represent a wide range of styles and varieties of prose" (Francis and Kučera, 1979), which may be referencing the fact that the text samples, when taken together, incorporate a wide range of characteristic ways of using language. However, in Baker's (2009: 313) discussion of the Brown Family of corpora he states that the corpus "consists of four main genres of writing…which were further divided into fifteen sub-genres". Baker (2009) is consistent in his use of the term *genre* to describe the texts in the Brown Family.

Table 4a shows the categorisation of the texts in the BE06 corpus (a corpus within the Brown Family). There are 3 'levels' of classification, and these do seem to fit best with the definition of *genre* discussed in section 4.2.3.6, i.e. they are mostly categories of texts which could be easily recognised by a member of the culture, without reference to internal criteria.

The genres which were included in the corpus were decided on during a conference at Brown University by a group of experts (Francis and Kučera, 1979), however Francis and Kučera (1979) give no more details regarding how these genres were selected.

Table 4a: Genres in the BE06 (Baker, 2009: 317).



	Broad text category	**Text category letter and description ('genre')**		**Number of texts**
Informative	**Press**	**A**	Press: Reportage	44
		B	Press: Editorial	27
		C	Press: Reviews	17
	General Prose	**D**	Religion	17
		E	Skills, Trades and Hobbies	38
		F	Popular Lore	44
		G	Belles Lettres, Biographies, Essays	77
		H	Miscellaneous: Government documents, industrial reports etc.	30
	Learned Writing	**J**	Academic prose in various disciplines	80
Imaginative	**Fiction**	**K**	General Fiction	29
		L	Mystery and Detective Fiction	24
		M	Science Fiction	6
		N	Adventure and Western	29
		P	Romance and Love story	29
		R	Humour	9

4.2.4.2 The Corpus de référence du Français contemporain (CRFC)

As discussed in section 2.2, the CRFC is a new “genre diverse” corpus of modern French (Siepmann et al., 2017: 63). The composition of the corpus has been inspired by the BNC1994 and COCA, but with an even greater diversity of genres. The composition of the CRFC can be seen in table 4b. The corpus is divided at the highest level according to *medium* (spoken, pseudo-spoken, and written), and then divided into genres. Medium is not a term which was discussed in the previous section as it is a much broader type of classification than genre or register, and refers to the channel through which language is broadcast (for example, speech or writing). The written medium contains 8 genres: academic, non-academic books, prose fiction, newspapers, magazines, diaries and blogs, letters and emails, and miscellaneous (Siepmann et al, 2017). These genres conform neatly to the definition of genre in section 4.2.3.6 as they are all easily identified by members of the culture based on external features (such as the format and structure of the text, the location in which it is found, the broad topic of the text etc.).

Table 4b: Categorisation of texts in the CRFC (adapted from Siepmann et al., 2017: 70).

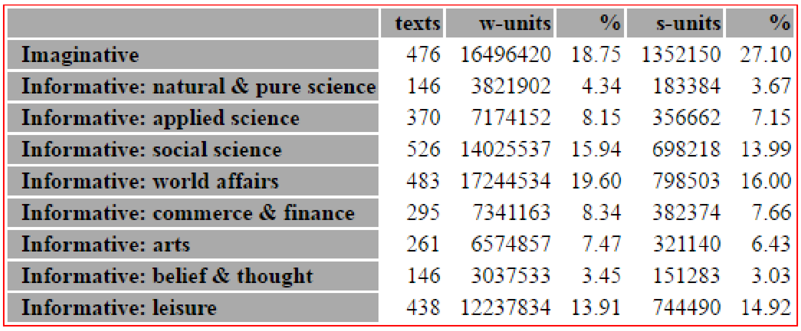
Medium	**Genre**	**Size**
Spoken	Formal	30m
	Informal	30m
Pseudo-spoken	Stage plays and film scripts	30m
	Film and daily soap subtitles	2,5m
	Text messages/chat	2,5m
	Discussion forums	60m
		155m
Written	Academic	30m
	Non-academic books	30m
	Prose fiction	30m
	Newspapers	45m
	Magazines	10m
	Diaries and blogs	5m
	Letters and e-mails	1m
	Miscellaneous	4m
		155m

4.2.4.3 The British National Corpus 1994

At the highest level, the BNC User Reference Guide (Burnard, 2007) describes the corpus as being divided into 5 text types: spoken demographic, spoken context-governed, written books and periodicals, written-to-be-spoken, and written miscellaneous. This use of text type is inconsistent with the definition discussed in this chapter, and the categories are actually more like genres, or mediums. It is unlikely that the creators actually carried out any research to determine whether these different text types were similar internally. Indeed, this would be problematic as the corpus creators would have had to already have created the corpus in order to do this, by which point a sampling frame is not necessary (see section 3.2.2.1).

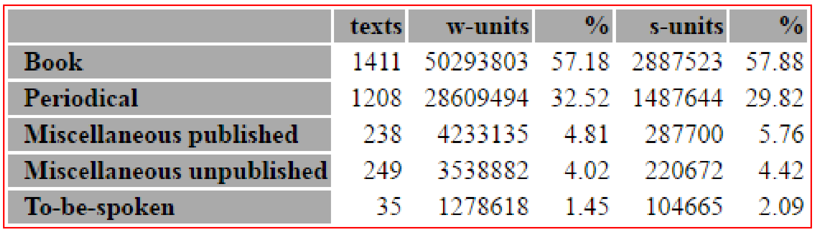
Within the written portion of the corpus, texts were categorised according to *domain*, and *medium* (see tables 4c and 4d). It is difficult to apply any of the labels discussed above to these categorisations. Indeed, Lee (2001:51) points out that "genres cannot easily be found at all under the current domain scheme". Texts in the BNC1994 are also classified according to time, author type, author sex, author age, author domicile, target audience, audience sex, publication place, and sampling type (Burnard, 2007).

Table 4c: Written domains in the BNC1994 (Burnard, 2007).



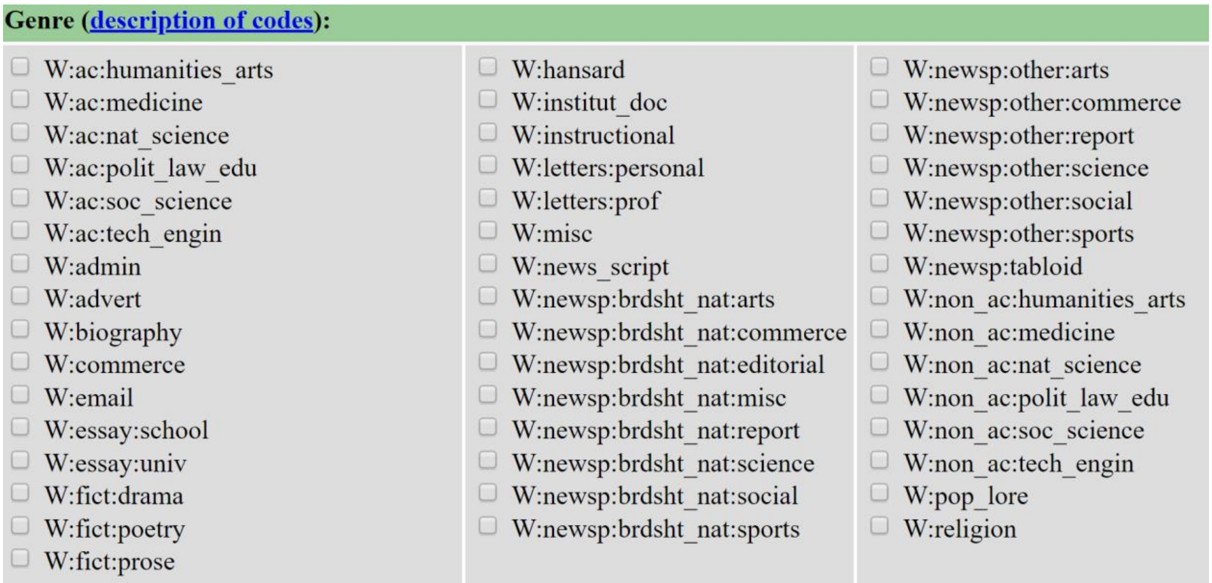
	texts	w-units	%	s-units	%
Imaginative	476	16496420	18.75	1352150	27.10
Informative: natural & pure science	146	3821902	4.34	183384	3.67
Informative: applied science	370	7174152	8.15	356662	7.15
Informative: social science	526	14025537	15.94	698218	13.99
Informative: world affairs	483	17244534	19.60	798503	16.00
Informative: commerce & finance	295	7341163	8.34	382374	7.66
Informative: arts	261	6574857	7.47	321140	6.43
Informative: belief & thought	146	3037533	3.45	151283	3.03
Informative: leisure	438	12237834	13.91	744490	14.92

Table 4d: Written mediums in the BNC1994 (Burnard, 2007).



	texts	w-units	%	s-units	%
Book	1411	50293803	57.18	2887523	57.88
Periodical	1208	28609494	32.52	1487644	29.82
Miscellaneous published	238	4233135	4.81	287700	5.76
Miscellaneous unpublished	249	3538882	4.02	220672	4.42
To-be-spoken	35	1278618	1.45	104665	2.09

In 2001, Lee created a new classification scheme for the texts in the BNC1994. The scheme gave each text a genre label, and some of these genres were grouped into super genres, with the aim of allowing “linguists, language teachers, and other users to easily navigate through or scan the huge BNC jungle more easily, to quickly ascertain what is there (and how much) and to make informed selections from the mass of texts available” (Lee, 2001: 37). Lee (2001) felt that the original BNC1994 classification system had many problems, which he aimed to solve by creating this new genre scheme. The first problem is that the categories are “overly broad” (Lee, 2001: 53). Lee (2001) points out that in the original domain classification there is no distinction made between academic and non-academic prose, despite the fact that the distinction between these genres was made in the Brown Family corpora, and has proved to be of great interest to researchers (Lee, 2001: 53). Additionally, whilst it is a very positive step that the BNC1994 contains a wide variety of imaginative texts such as novels, poetry, and drama (whereas the Brown Family only contains novels), there is no way to distinguish between these genres when searching the corpus using the original domain and medium classifications (Lee, 2001). Another problem with the original classification of texts in the BNC1994 is that there were many classification errors and misleading titles in the corpus (Lee, 2001). Some texts were classified as the wrong category because they had a misleading title; Lee (2001: 53) gives the example that “many texts with ‘lecture’ in their title are actually classroom discussions or tutorial seminars involving a very small group of people”. Another problem, but one which Lee (2001) emphasises has no real solution, is that some BNC files are too big and contain multiple different genres or sub-genres. For example, single newspaper files labelled as containing ‘editorial material’ can include letters-to-the-editor, institutional editorials, and personal editorials. A final problem which Lee (2001) points out is that a lack of genre classification means that the BNC1994 Sampler (a subset of the BNC1994 containing two collections of written and spoken material of about one million words each, originally compiled to mirror the composition of the full BNC as far as possible) cannot claim to be representative in terms of genre. Lee (2001: 54) believes that “it is because ‘domain’ is such a broad classification in the BNC that the Sampler turned out to be rather unrepresentative of the BNC and of the English language.” Lee’s (2001) genre scheme for the Written BNC1994 can be seen in figure 4a.



Genre (description of codes):

□ W:ac:humanities_arts
□ W:ac:medicine
□ W:ac:nat_science
□ W:ac:polit_law_edu
□ W:ac:soc_science
□ W:ac:tech_engin
□ W:admin
□ W:advert
□ W:biography
□ W:commerce
□ W:email
□ W:essay:school
□ W:essay:univ
□ W:fict:drama
□ W:fict:poetry
□ W:fict:prose

□ W:hansard
□ W:institut_doc
□ W:instructional
□ W:letters:personal
□ W:letters:prof
□ W:misc
□ W:news_script
□ W:newsp:brdsht_nat:arts
□ W:newsp:brdsht_nat:commerce
□ W:newsp:brdsht_nat:editorial
□ W:newsp:brdsht_nat:misc
□ W:newsp:brdsht_nat:report
□ W:newsp:brdsht_nat:science
□ W:newsp:brdsht_nat:social
□ W:newsp:brdsht_nat:sports

□ W:newsp:other:arts
□ W:newsp:other:commerce
□ W:newsp:other:report
□ W:newsp:other:science
□ W:newsp:other:social
□ W:newsp:other:sports
□ W:newsp:tabloid
□ W:non_ac:humanities_arts
□ W:non_ac:medicine
□ W:non_ac:nat_science
□ W:non_ac:polit_law_edu
□ W:non_ac:soc_science
□ W:non_ac:tech_engin
□ W:pop_lore
□ W:religion

Figure 4a: Lee’s (2001) genre classifications for the Written BNC1994 (as seen in BNCWeb).

4.2.4.4 Summary

It seems that, from this brief review of three previous national corpus projects, genre is the classification most commonly used in these kinds of corpora. The texts within the Brown Family of corpora, whilst not referred to as such at the outset of the project, have been referred to as genres, and neatly conform to the definition of genre given in section 4.2.3.6. The creators of the CRFC classify their texts into mediums at the highest level, and then into genres. The creators of the BNC1994 did not classify the texts included in the corpus into genres, but rather into text types, which were further divided by domain and medium. However, Lee (2001) highlighted the problems associated with this method of text classification, and designed a new classification system in which each text is classified according to super genre and genre. This new classification scheme was welcomed by users of the corpus, and has proved very useful. The common use of the term genre to classify texts in national corpora will be taken into account when making decisions regarding classification of texts in the Written BNC2014 in section 4.2.5.

4.2.5 Text classification in the Written BNC2014

This section has outlined the different ways in which texts can be considered and classified by linguists. It is important to consider all of these in order to come to a decision about how the texts included in the Written BNC2014 will be categorised. After considering the different options I have decided that the texts in the Written BNC2014 will be labelled as *genres* at the most detailed level, which will be grouped into *super-genres*, and which will be split into 5 different *mediums* at the highest level.

This decision has been taken for 2 principal reasons. The first reason is that this type of labelling fits well with previous corpus projects, and the definition of genre fits the texts which will be collected well. As was shown in section 4.2.4, previous corpora have used genre to classify their texts (e.g. the CRFC), and where corpora have not been classified according to genre (the Written BNC1994) this type of classification has been added in later with great success. This shows that genre is a way of labelling texts which researchers find useful, and so it seems natural that they will desire this kind of labelling in the Written BNC2014. Furthermore, I want the Written BNC2014 to be as comparable as possible with the Written BNC1994, so it must employ as similar a system of classification as possible. The top level split into mediums preserves some of the work done by the original creators, and then the subsequent split into super-genres and genres closely mirrors Lee's (2001) classification system. As shown in section 4.2.2, there is no established system for classifying genres, and so I will attempt to keep the labelling of the genres as close to Lee's (2001) labels as possible, for more details on this see section 4.4.1. Additionally, for the texts which I will include in the Written BNC2014, the definition of genre as 'a category of texts which is easily recognised by a member of the culture; genres can be identified using external, non-linguistic criteria' works extremely well.

A second reason for the use of the term genre to label the texts is that other linguists have argued in favour of doing so. Atkins et al. (1992) emphasise that selection of texts for a corpus must be based on external criteria because a corpus where the texts were selected based on internal criteria would give no information regarding the relationship between language and context (section 3.2.2.1). Indeed, none of the texts included in the Written BNC2014 will be selected based on internal criteria (see section 4.3); this rules out categorising texts into text types (see definition in section 4.2.3.6). Lee (2001: 37) supports the use of genre to classify texts in a corpus because it "is the level of categorisation which is theoretically and pedagogically most useful and most practical to work with”. However, some linguists may disagree on theoretical grounds with the use of genre to label the texts within the Written BNC2014. As we saw, Biber and Conrad (2009) believe that genre analysis can only be performed on whole texts, rather than samples. Many texts in the Written BNC2014 will be samples of texts rather than whole texts (see section 4.3.2), and so Biber and Conrad (2009) may not agree with labelling the texts according to genres. However, Biber and Conrad (2009) in framing their criticism are talking about performing a genre analysis, rather than simply labelling texts which can then be subsequently investigated from multiple perspectives, so this criticism may not be entirely relevant to the discussion here. Furthermore, I believe that the arguments in favour of a taxonomy based on genre for the Written BNC2014 far outweigh this potential criticism.

4.3 Design of the sampling frame

Once the classification system for the texts included in the sampling frame had been decided upon, I could then begin to design the sampling frame itself. This section outlines the major decisions made in the design of the sampling frame, many of which will be returned to in chapters 5, 6, 7, and 8 when discussing text collection in detail. The sampling frame can be found in appendix B. Additionally, the eventual composition of the corpus can be found in appendix C (although it should be remembered, as has been noted on other occasions in this thesis, that all numbers given in appendix C are provisional, as the corpus has not yet been finalised).

4.3.1 Population definition

As seen in section 3.2.2.1, defining the population is one of the first issues which must be tackled when designing a corpus. The population for the Written

BNC2014 can be defined quite simply as ‘all written texts which were produced by native speakers of British English in 2014’. This definition at first glance seems to be a useful one as it addresses Biber’s (1993: 243) first feature of population definition: what texts are included and excluded from the population. This definition makes it easy to see what texts would be acceptable as members of the population as the definition is very broad but also has strict boundaries in terms of date of production, and the language of the producer. However, this broadness also makes the definition less useful when we consider Biber’s (1993:243) second feature of population definition: what text categories are included in the population. With such a broad population definition it would be impossible to come up with a list of all of the possible text categories which could be included in the population (see section 4.4.1 for a discussion of the genres which will be included). This brings us back to arguments made by Hunston (2008), Bauer and Aarts (2000) and Atkins et al. (1992) (see section 3.2.2.1) that delimiting the population to be represented by a corpus is often impossible because there are no lists of all genres within a population. This is certainly the case for the Written BNC2014 – there are no listings of all of the genres which would be eligible for inclusion in the corpus according to the above population definition. This means that I will not be able to compare the Written BNC2014 to the total population, and thus will not be able to assess the eventual representativeness of the corpus.

This definition of the population will also prove to be problematic for other reasons. Whilst 2014 is the maximally desirable year in which texts included in the corpus have been published, it will not always be possible to collect as much data as is needed from this one year. In these instances the population definition will be expanded according to the date range policy set out below.

> *The Written BNC2014 date range policy:* Where it is not possible to collect all of the required data for a genre from 2014, the date range will be expanded forward one year at a time, until enough data is collected (i.e. firstly including 2015, then 2016 etc.). If enough data has not been collected after expanding the date range to include the years 2014–2018, then the date range will be expanded backwards by one year at a time, to no earlier than 2010 (i.e. firstly including 2013-2018, then 2012-2018 etc.).

I selected 2014 as the maximally desirable year for texts in the corpus to have been published as this was the year in which the project began. The creation of the Spoken BNC2014 (Love et al., 2017a) was already underway at this stage of designing the corpus, and it was known that the median collection point of the data included in the Spoken BNC2014 would be the year 2014. The Spoken and Written BNC2014 will eventually be combined into one corpus, and so keeping the date ranges similar is desirable. The date range policy, set out above, allows me to balance this desire for comparability with a desire to represent *contemporary* British English. By firstly stretching the date range forwards as far as possible, I ensure that the collection of more contemporary language is prioritised over less contemporary language, and, by limiting data collection to only texts published in the 2010s, language which is certainly not contemporary is excluded. Furthermore, even at its most stretched (i.e. 2010-2018), the date range policy designates a far smaller range of dates of publication than were included in the Written BNC1994, in which some texts were published more than 30 years before the release of the corpus (Burnard, 2000: 5).

Another important aspect of population definition is deciding whether you will define your population in terms of text production or reception. The creators of the BNC1994 decided to take account of both perspectives. This approach will also be followed in the Written BNC2014. As with the BNC1994, books are the largest genre (see appendix B for the sampling frame, and appendix C for the eventual composition of the corpus) as, whilst they are written by relatively few people, they are read by a far greater number of people. On the other hand, the genre of 'e-language' will also be present in the corpus, as although individual emails and instant messages (IMs), for example, are only read by a handful of people, many people produce this kind of text extremely often in their daily lives (Deloitte, 2014, estimate that 50 billion mobile IM messages were sent everyday worldwide in 2014).

4.3.2 Sample size

The typical sample size for the texts included in the Written BNC2014 is 5000 words. The reason that I am referring to this as the *typical* sample size is that, for some genres of text which were particularly difficult to collect, this sample size was increased to allow more text to be collected from fewer sources. This was particularly relevant for the collection of books (see chapter 5) where it was extremely difficult to collect data. In this case the sample size was increased to allow for roughly one third of a book to be collected as one sample. These samples are evenly balanced (as far as is possible) between samples from the beginning, middle, and end of texts to ensure that structural features of different texts are fully represented. I decided to use samples of texts rather than whole texts largely because of the difficulties which would be encountered regarding copyright if whole texts were to be used (see McEnery et al., 2006, and section 3.2.2.5 of this thesis). It is extremely unlikely that any publishers would allow me permission to include whole texts of their published books in the corpus, where those books are not already open access, due to worries about copyright and commercial rights. Publishers are typically financially invested in the texts which they publish, and would undoubtedly worry that releasing them for free in a corpus would affect the market for the original work. Of course, it is extremely unlikely that any member of the public would try to read an entire book in corpus format as it will be heavily altered with xml tags etc., but convincing a publisher (who may have no knowledge at all about corpora) of this is likely to be difficult. On the other hand, it is possible to argue that a 5000 word sample is rather similar to the kind of extract given away by publishers for free online. Publishers often make samples of their books available on Amazon.co.uk or Google Books to entice potential customers to purchase the book (see section 5.3.4). Thus, asking for a sample which is roughly the same size as these free samples would not worry publishers in the same way as asking for whole texts would. In addition, one can also suggest to publishers that being included in the corpus will act as a form of advertising for them, in the same way as the free previews which they release do. These arguments actually ended up being largely irrelevant once the collection of books began, due to the problematic nature of contacting publishers (see section 5.2). Nevertheless, at the planning stage of the corpus this was the rationale for selecting 5000 words as the typical sample size.

Furthermore, using text samples avoids the problem of some very large texts potentially skewing the results derived from the corpus (McEnery et al., 2006: 20; Hunston, 2008: 166). Whilst there are arguments for the use of whole texts in corpora (see section 3.2.2.5), it is likely that this decision will only be relevant for books, and some of the miscellaneous and periodical genres. For the majority of the genres included in the corpus (e.g. newspaper articles, blogs etc.) texts are typically less than 5000 words in length, and so the whole text *will* be included in the corpus. This presents a further aspect of sampling which a decision needs to be reached upon: whether or not to include these shorter individual texts as single texts within the corpus, or whether to group several texts together to create a 5000 word sample instead. There are pros and cons to both decisions. Including the texts on their own is easier and faster, as the corpus builder will not have to spend time fitting texts into groups and checking word counts. However, having all of the texts in the corpus of varying lengths makes comparing individual texts with each other more difficult. Although, this would not be commonly done in corpus linguistics, and if it were, normalised frequencies would account for differing text lengths so is perhaps not an important limitation. Additionally, as has already been discussed, some texts will be longer than 5000 words so there will already be variance in text length, regardless of this decision. Grouping texts together is more time consuming but means that all of the texts within the corpus will be directly comparable with one another. As there is no clear 'best' decision in this case, I will follow the decisions taken by the creators of the Written BNC1994 (and the creators of the ARCHER corpus, and the Brown Family of corpora) and group texts of the same genre to create samples which are 5000 words in length. However, to mitigate against the time consuming nature of grouping texts, and because the arguments for grouping texts seem to have limited relevance here, only texts shorter than 2000 words will be grouped. If a text is between 2000 and 5000 words in length, then it will be included as a single text.

The BNC1994 used a sample size of 40,000 words for books, so most sample sizes in the Written BNC2014 will be much smaller. As already mentioned, one reason for selecting 5000 words as the typical sample size is to reduce issues with copyright for published books – gaining permission to include whole books, or even 40,000 word samples in the present day would be almost impossible. Additionally, in the TNC project it was indicated that 40,000 word samples would be too big to be considered to fall within the bounds of 'fair dealing' (see section 2.4). As such, sample sizes needed to be smaller than this to ensure that the 'non-commercial research' exception to UK copyright law, discussed in section 1.5.2, could be utilised. A further reason is that 5000 words fits with Biber's (1990, 1993) recommendation that 2000 and 5000 word samples will be satisfactory for investigating both common and rarer features in a corpus. However, Biber (1993) does acknowledge that more research is needed to propose specific recommendations for sample length, particularly for less stable features, and other features such as discourse features (see section 3.2.2.5 for a full discussion of this research). However, in the absence of more specific recommendations, 5000 word samples seem to be a good balance between what is recommended and what is practical. Furthermore, having a smaller sample size means that more samples can be included in each genre (see section 3.2.2.6). This ensures that samples are taken from a wider range of sources within each genre, which should hopefully increase representativeness.

4.3.3 Proportions of each genre

Similarly to the Written BNC1994, the proportions of each genre, and thus the number of samples included in each genre, will vary greatly. This is largely because decisions regarding the proportions of the genres have often been made based on the practicalities of data collection – that is, for genres where data is easier to collect (e.g. newspapers) the number of samples is greater than for genres which will be more difficult to collect (e.g. emails). These considerations of practicality have also been balanced with a desire for the corpus to remain broadly comparable to the Written BNC1994, and also a desire for each genre within the corpus to be useful as an object of study in its own right as a representative sample of a particular kind of written British English. The desire for each genre to be large enough to be useful as an object of study in its own right means that few genres in the sampling frame contain less than 900,000 words of data (1% of the total corpus; see section 4.3.4). Baker (2009) suggests that a corpus of 1 million words in size is useful for investigating common linguistic features, so a size of 900,000 words for a single genre within the corpus seems a good balance between including enough genres in the corpus, and having each be big enough to be useful.

As a consequence of these decisions, the genres which were planned to make up the smallest proportions of the corpus were the individual blog genres (see appendix B; see appendix C for details of how this changed in reality). These were all allocated 180,000 words, which equates to thirty-six 5000 word samples per genre. According to Biber (1990) this should be plenty to investigate common linguistic features. Furthermore, if we consider all of the six 'blogs' genres together then we have a total of 1,080,000 words comprised of 216 samples, which is plenty to be considered useful as an object of study. Furthermore, this is plenty of samples considering that Biber's (1990) recommendations were based on an investigation of 10 text samples. This relatively large number of samples should also address the concerns raised by Biber (1993) regarding his 1990 recommendations (see section 3.2.2.6).

4.3.4 Corpus size

The Written BNC2014 was designed to be 90 million words in size, as this directly matches the size of the Written BNC1994. Much of the research discussed in section 3.2.2.7 showed that many linguists feel that bigger corpora are more representative and balanced (Hunston, 2008; Leech, 2007; Biber, 1990; McEnery et al. 2006). Whilst 90 million words may not seem like such a large number of words nowadays due to the rise of extremely large web-crawled corpora, such as enTenTen which currently contains 15 billion words and continues to grow (Jakubíček et al., 2013), it is still a relatively large size for a ‘hand-made’ corpus seeking to explicitly represent a range of genres. Baker (2009) suggests that a 1 million word corpus, such as the BE06, is large enough for investigating the use of high frequency words, and thus 90 million words should be sufficient for rarer items. 90 million words also seems to fit with Biber’s (1990:269) recommendations that “the total number of texts included in existing computer-based corpora are adequate for multivariate statistical analyses” (see section 3.2.2.7). Biber’s conclusions were drawn based on studies of relatively common grammatical features (e.g. first person pronouns, contractions, present tense verbs etc.) so can’t be extended to the study of rarer features, but 90 million words is likely larger than the “existing” corpora that Biber was discussing in 1990 and so this should go some way to addressing these limitations.

4.3.5 Sampling methods

As chapter 2 showed, another important sampling decision is the sampling method which will be used to select texts for inclusion in the corpus. This will not be straightforward in the creation of the Written BNC2014 and will change depending on what genre is being worked on. As far as is possible, sampling has been done randomly in order to prevent any bias in the data selection. However, as there is no list of members of the population (both for the population as a whole and for individual genres) it was not possible to use simple random sampling, where all members of a population have an equal chance of being selected (McEnery et al., 2006; Bauer and Aarts, 2000; Biber, 1993). Where possible, I gave all of the members of a population *which are known about* an equal chance of being selected. So, for example, using the LexisNexis method discussed in section 6.3.2, articles from a particular newspaper on any day between 2014-2016 which were available on LexisNexis had an equal chance of being included. Sometimes, random sampling was not appropriate, as I wanted to prioritise texts with a wide readership. So, for example, I was sampling texts from ‘popular’ blogs, rather than randomly sampling from *all* blogs (which would be impossible anyway without an exhaustive list of all blogs, which does not exist; see chapter 7 for more detail on the collection of blogs). The fact that the sampling frame is divided into genres means the sampling of texts for inclusion in the corpus was stratified. The genre distinctions laid out in the sampling frame (see appendix B) divide each of the super-genres, and determine what specific genres will definitely be represented. As discussed in section 3.2.2.2, this presents the problem that what is included in the corpus has already been predetermined by me; for instance, if there was one genre covering all blogs then chance would dictate whether any travel blogs were included in the corpus; but because a separate genre has been identified for travel blogs, at least some such texts will definitely be included. However, this issue is balanced by the fact that stratified sampling ensures that the full range of linguistic variation in each genre is represented in the corpus, including rarer items (Biber, 1993). The sampling frame for the corpus has, for the most part, not been designed to be proportional. This is largely because, as has already been mentioned, the proportions in the population are not known. When designing the SYN2015 corpus Křen et al. (2016: 2523) decided that “a general language corpus should primarily attempt to cover the variety of existing texts and their well-designed and documented classification rather than trying to estimate their […] proportions in a language”. This decision was reached because of the many factors which had to be taken into account when designing the corpus, for example: the population of texts is unknown, it is impossible to measure the real proportions of language in use, and corpus-interface software makes it increasingly easy for users to examine the composition of a corpus and adapt it to their needs, resulting in less need for exact proportional balance (Křen et al., 2016: 2523; see section 2.3 for more information on the SYN2015 corpus). As a result, the SYN2015 corpus was designed to be representative of written Czech, but not balanced. Likewise, in the Written BNC2014, as proportionality is not possible, the proportions in the sampling frame are not proportional to real language production or reception but do represent the full variety of texts in large enough amounts that they will be useful as objects of study in their own right (see section 4.3.3). The decisions regarding proportions have also been based on practicalities of data collection, and considerations of comparability (see section 4.3.3). One notable exception is the genres within the 'fiction' super-genre where the proportions in the sampling frame were based on the proportions found on a popular bookseller's website. More detail about the sampling methods used to collect each genre of text will be given in chapters 5, 6, 7, and 8.

4.4 Comparability with the Written BNC1994

4.4.1 Genres in the corpus

Appendix B shows the full sampling frame for the Written BNC2014, where all of the genres and their ideal proportions in the corpus can be found. Tables 4a, 4b, and appendix D show how these genres and proportions compare to the texts contained in the Written BNC1994. It is important to note here that this is a sampling frame, and as such is *not what the final corpus actually looks like*. The sampling frame was designed prior to data collection and shows the *ideal* make-up of the corpus. As can be seen in appendix C, and as will be seen in subsequent chapters of this thesis, some genres presented problems in their collection, and as such the eventual make-up of the Written BNC2014 is somewhat different to the sampling frame in appendix B.

The eventual make-up of the corpus can be seen in appendix C (although, as noted, the numbers quoted in this thesis are not finalised and are subject to change).

The genres included in the corpus sampling frame are largely similar to those in the Written BNC1994, but with the addition of some new genres which have emerged since (e.g. e-language). The genres have been kept largely comparable due to the advice of the experts who were consulted in the design of the sampling frame (see section 3.4). Most experts felt that all of the genres included in the Written BNC1994 should be preserved in the new corpus. However, e-language has been added in order to make the corpus more representative of present-day British English (again, see section 3.2 for a fuller explanation of these decisions).

I have maintained the use of Lee's (2001) genre labels (see section 4.2.4.3), but adapted the labelling slightly to fit the new corpus (this can be seen in appendix D). New labels have been added (e.g. the e-language labels), and some of the labels have been given more levels of distinction (e.g. splitting tabloid news into 7 different genres, rather than 1).

4.4.2 Proportions of genres in the corpus

As discussed in chapter 3, comparability with the Written BNC1994 is a secondary focus of the new corpus. My primary focus is to make a corpus which is as representative of present-day British English as is possible, with a comparable sub-corpus being created once the entire corpus is finished. As such, the proportions in the Written BNC2014 are not directly comparable with the proportions in the Written BNC1994, both due to the inclusion of new genres and also due to practical considerations regarding data collection which are relevant now but weren't as relevant in the 1990s. The inclusion of the e-language super genre in particular has meant that the proportions of other genres have had to be reduced compared to the Written BNC1994. Furthermore, due to practicalities, the proportion of books is smaller in the new corpus due to this type of data being much more difficult to collect nowadays (see chapter 5).

In terms of the mediums to be included in the corpus (see table 4e), books are the only medium which have decreased as a proportion (from 58.58% in 1994 to 41% in 2014). All other mediums have increased slightly. This is partly due to the desire, discussed above, for all of the sub-sections of the corpus to be useful in their own right, but also because the proportions in the 2014 sampling frame are ideals to aim for rather than the realistic results represented by the proportions in the actual Written BNC1994 corpus. As such, as noted, the percentages in the 2014 sampling frame are goals, rather than the reality represented by the 1994 proportions (the reality for the Written BNC2014 can be seen in appendix C).

Table 4e: Comparison of the proportions of the mediums included in the Written BNC1994 and the Written BNC2014 sampling frame.

Medium	**Proportion (BNC1994)**	**Proportion (BNC2014)**
Books	58.58%	41%
Periodicals	31.08%	35%
Miscellaneous	8.78%	10%
To-be-spoken	1.52	4%
E-language	0	10%

In terms of the super genres within the Written BNC1994 and the Written BNC2014 sampling frame, some have increased, some have decreased, and some have stayed much the same. Table 4f shows this comparison. Note that for this comparison I have re-categorised some of the BNC1994 genres into super genres in order to make the data comparable, e.g. I have combined drama texts and news scripts to make a ‘written-to-be-spoken’ super genre even though this is not identified as a super genre in Lee’s (2001) genre scheme. I have also slightly altered the super genres from the 2014 sampling frame in order to aid comparability, e.g. in the 1994 scheme there is no differentiation between academic and non-academic books and journals, so these have been combined in the 2014 comparison in table 4f. The actual super genres for the Written BNC1994 can be found in Lee (2001). The actual super genres for the Written BNC2014 can be found in appendices B and C.

The proportion of fiction texts has increased slightly in the 2014 sampling frame to represent their “influential cultural role.” (Burnard, 2000: 7). Academic and non-academic prose and periodicals have both decreased slightly for similar reasons. The proportion of newspapers (including broadsheet, regional & local, and tabloid) has doubled in the new sampling frame. This is because more newspaper texts are included in the new sampling frame to address the imbalance of newspaper types in the 1994 corpus. In the 1994 corpus much more data was included from broadsheet newspapers and regional and local newspapers than tabloid newspapers. The amount of texts from these three types of newspapers are equal in the 2014 sampling frame, to avoid the implication that one type is more important than another. Consequently newspapers overall were planned to be present in a much higher proportion, with the most significant increase planned to be in the proportion of tabloid news texts. The proportion of magazines in the sampling frame has stayed roughly the same (although magazines were labelled as ‘W:pop_lore’ in the 1994 corpus). E-language has, of course, increased in the 2014 sampling frame because there were only a handful of texts in the 1994 corpus which could be categorised as e-language. Essays, letters, and written-to-be-spoken texts have increased in the Written BNC2014 sampling frame. These super genres were present in very small amounts in the Written BNC1994, but I decided it was important to include them in the new corpus as the experts who were consulted in the design of the sampling frame stated that they felt all of the genres in the BNC1994 should be included in the 2014 corpus (see section 3.4). Thus, the size of these super genres needed to increase in the new corpus in order for them to be useful as objects of study in their own right.

Table 4f: Comparison of the proportions of the super genres included in the Written BNC1994 and the Written BNC2014 sampling frame.

Supergenre	**Proportion (BNC1994)**	**Proportion (BNC2014)**
Fiction	18.49%	21%
Academic (prose+periodical)	18.3%	12%
Non-academic (prose+periodical)	18.46%	14%
Broadsheet national newspapers	3.45%	7%
Regional & local newspapers	6.41%	7%
Tabloid newspapers	0.83%	7%
Magazines	8.42%	8%
E-language	0.24%	10%
Essays	0.23%	2%
Letters	0.14%	2%
Written-to-be-spoken	1.47%	4%

Note: The columns do not total 100% because in both corpora some texts are not categorised into super genres.

There are too many individual genres in both the Written BNC1994 and the Written BNC2014 sampling frame to present a full comparison here, thus, I will simply highlight some of the main differences. For a detailed comparison of the proportions of individual genres in the two corpora see the table in appendix D.

The main difference between the genres in the Written BNC1994 and the Written BNC2014 sampling frame is that the proportions of the genres in the Written BNC2014 sampling frame are all much more similar to each other than they are in the Written BNC1994. For example, only 14 out of the 80 genres included in the 2014 sampling frame do not comprise either 1% or 2% of the total sampling frame. The proportions of genres vary much more widely in the 1994 corpus. Of course, this is in part due to the fact that the 1994 proportions are actual collection figures whereas the 2014 proportions are ideal targets which may or may not be reached (see appendix C for the totals which were eventually reached). Furthermore, it is important to remember that Lee's (2001) genre labels were applied to the Written BNC1994 years after it had been collected. Thus, the creators were not attempting to balance the proportions of these genres when they were collecting the data.

As already mentioned, another notable difference is that tabloid news is now split into the same seven genres as the other two types of newspapers in the 2014 sampling frame, whereas this was not the case in the 1994 corpus. As a consequence, tabloid news is present in a much higher proportion in the 2014 sampling frame than it was in the 1994 corpus (7% as opposed to 0.83%).

As a consequence of the desire for each genre within the corpus to be useful as an object of study in its own right, some genres have increased greatly in the 2014 sampling frame. For example, drama scripts only comprise 0.05% of the 1994 corpus, but comprise 2% of the 2014 sampling frame. Similarly, university essays only comprise 0.06% of the 1994 corpus, but comprise 1% of the 2014 sampling frame.

4.5 Conclusion

This chapter has introduced the design of the Written BNC2014 sampling frame and has discussed the impacts that this design will have on the representativeness and comparability of the corpus. The corpus aims to be as representative of present-day written British English as practically possible (see chapter 3), and this is reflected in the design of the sampling frame. The population has been clearly defined as 'all written texts which were produced by native speakers of British English in 2014’. This definition provides clear boundaries for what is and is not included in the population, but has limited use in reality because, as many other linguists have pointed out (see section 3.2.2.1), it is impossible to create an exhaustive list of members of the population. This impacts on the sampling methods which will be employed in the creation of the corpus. Where possible, I endeavoured to use a stratified random sampling method in order to increase the representativeness of the corpus, but of course random sampling is not possible where all members of a population are not known. For this reason, in most genres, the sampling also could not be done proportionally. However, decisions regarding the typical sample size to be used in the corpus (5000 words), the proportions of each genre, and the overall corpus size were made according to recommendations in the literature discussed in chapter 3, as well as considerations of practicality, with the goal of increasing the representativeness of the corpus. These design decisions are a perfect example of what was discussed in section 3.2.3 – the idea that representativeness is not possible in corpora. It is important to acknowledge that, for the practical reasons discussed, it will not be possible for the Written BNC2014 to be fully representative of the population; however, as Leech (2007) suggests, representativeness is still something which I will aim for as far as possible. As Atkins et al. (1992:6) recommend, “knowing that your corpus is unbalanced is what counts”; thus, when the corpus is released users will have access to a reference guide (Love et al., 2017b) which will detail all of the design decisions taken so that they can assess the representativeness of the corpus for themselves.

The design decisions taken in the creation of the sampling frame also ensure that the Written BNC2014 is broadly comparable to the Written BNC1994. The Written BNC1994 and the Written BNC2014 sampling frame contain mostly the same genres, labelled in mostly the same way, with the notable addition of the ‘e-language’ section in the 2014 sampling frame. The proportions of the mediums, super genres, and genres in the 1994 corpus and the 2014 sampling frame vary somewhat due to the addition of these new genres, but are broadly similar. This is due to the fact that, as already mentioned, the corpus could not be proportionally representative of the population as a whole. Thus, I decided to, where possible, keep the proportions similar to the 1994 corpus. The decision to create a comparable sub-corpus once the whole corpus is completed means that the comparability of the Written BNC2014 to the Written BNC1994 is not too much of a concern, because comparative research will be able to be carried out using the comparable sub-corpus.

Chapter 5: Collection of books for the Written BNC2014

5.1 Introduction

Now that the corpus has been designed, I can begin to consider the collection of data for the corpus. In this chapter I will discuss the collection of published books to include in the Written BNC2014. The Written BNC2014 sampling frame planned for the corpus to contain 36.9 million words from published books (see sampling frame in Appendix B), taken from academic books (5.4 million words), fiction books (18.9 million words), and non-academic non-fiction books (12.6 million words). The majority of this chapter will discuss the collection of published books in relation to the figures laid out in the sampling frame. The actual composition of this medium which was achieved can be seen in appendix C, and will be discussed in section 5.5. Published books make up a smaller percentage of the Written BNC2014 sampling frame than they did in the Written BNC1994 (41% as opposed to 58%). This is a consequence of adding the e-language medium to the Written BNC2014, which led to the reduced percentage of books. More 'space' was taken from the books medium than from any other medium because I also knew from the outset of this project that published books would be amongst the hardest types of data to collect for the corpus (see section 5.2). Thus, percentages were set lower in order to reflect a balance between what I would like to include in the corpus and what would be possible.

The academic and non-academic prose samples in the sampling frame are split into the same genre categories which were used in the Written BNC1994 (Lee's 2001 classification system; see sampling frame in appendix B), thus making these texts directly comparable in the two corpora (however, see section 5.5 for a discussion of how and why this ultimately changed in the final composition of the corpus). The proportions of academic books and non-fiction books in the population could not be known beforehand, so were split equally between genres in the sampling frame. I attempted to infer the proportions of each genre in the population of academic books by using Lancaster University's online library system, but it was not possible to search books according to the genres being used in the corpus. No other websites could be identified which contained records of the vast majority of published academic books. The website of the popular UK book retailer Waterstones was consulted to attempt to infer the proportions of genres within the population of non-academic non-fiction books (similarly to the method used for fiction books, see below). However, there were so many books published under each genre of non-fiction writing that the website could not return exact numbers (simply '10,000+ items'), and so inferring proportions was not possible. The website of another popular book retailer, Amazon.co.uk, was also consulted, but this website did not give numbers of results within each of its top-level genre distinctions. Furthermore, the websites consulted did not categorise books according to the genre categories being used in the Written BNC2014. These categories were preserved in the sampling frame in order to attempt to increase comparability with the Written BNC1994, but this made it very difficult know exactly which categories on a retailer's website lined up with the categories in the sampling frame. As such, even if proportions were able to be calculated, these may not have accurately reflected the proportions of the genres being used in the corpus. For example, the Waterstones category 'Science, Technology & Medicine' would contain books which would fall under the medicine, natural science, and technology & engineering genres in the corpus sampling frame.

Thus, each genre of academic prose in the sampling frame contains 900,000 words, and each genre of non-academic (non-fiction) prose in the sampling frame contains 1.8 million words. These proportions will achieve the primary goals, set out in chapter 4, of the corpus being broadly comparable to the Written BNC1994 and also each individual genre being large enough to be useful as an object of research in its own right. Academic books made up 12% of the Written BNC1994 and make up 6% of the Written BNC2014 sampling frame. This is obviously a smaller proportion, but this reduction was necessary in order to accommodate the new e-language medium (as already discussed, above). Very few people ever write or read academic books, and thus I felt that lowering the proportion by half here seemed more defensible that taking texts away from the fiction section, which represents a type of text which is read by very many people. Furthermore, while less data is included from academic books in the Written BNC2014, each genre is more equally represented in the sampling frame than in the Written BNC1994. For example, in the Written BNC1994, 3.97% of the corpus consists of academic politics, law & education books, coming from a total of 108 texts. On the other hand, only 0.12% of the corpus consists of academic medicine books, comprising just 4 texts. This imbalance is resolved in the Written BNC2014 sampling frame.

Non-academic prose (non-fiction) made up 22% of the Written BNC1994, and makes up 14% of the Written BNC2014 sampling frame. Once again, this reduction is mainly due to the inclusion of the e-language medium in the corpus. Similarly to the academic books, the non-academic non-fiction genres are much more equally represented in the Written BNC2014 sampling frame than in the Written BNC1994. For example, non-academic non-fiction politics, law & education books make up 5.14% of the Written BNC1994, whereas non-academic non-fiction medicine books only make up 0.57% of the Written BNC1994. Both of these genres make up 2% of the Written BNC2014 sampling frame. Of course, this imbalance in these genres in the Written BNC1994 may actually reflect the population. In other words, these genres may be proportionally represented. As discussed in chapter 4, proportional representation was not used in the design of the Written BNC2014 sampling frame because the populations for the vast majority of genres could not be known in advance. However, it became apparent that the genre categories used and the proportions set out in the Written BNC2014 sampling frame were far from representative of the actual population of non-academic non-fiction books once collection began. As such, the eventual composition of this super-genre looks somewhat different to what was originally planned (see section 5.5, and appendix C).

There is no single widely accepted way of dividing fiction books into genres, and so the fiction samples in the sampling frame were split according to commonly used bookseller categories; the website of the popular UK book retailer Waterstones was investigated to see what genres they classify books into, and these genres were replicated in the sampling frame. However, in the time between creating the sampling frame and the present, Waterstones have changed the genre categories used on their website, which perhaps emphasises that there really is no commonly accepted classification system for these texts. Nevertheless, the genre categories used in the sampling frame for fiction books are: poetry, general fiction, children's fiction, teen fiction, science fiction & fantasy, crime, and romance. Fiction books make up 21% of the Written BNC2014 sampling frame as opposed to 18.49% of the Written BNC1994. As already mentioned, this increase, whilst the two other books super-genres have decreased in proportion compared to the Written BNC1994, is because many people read fiction books. As discussed in chapter 4, the design of the Written BNC2014 sampling frame took account of both language production and reception. Relatively few people will ever write any kind of book, but many people read fiction books (certainly more than read academic books, as can be evidenced informally by noting the lack of academic books in Amazon.co.uk's bestseller list), and thus allocating fiction books the largest proportion within this medium satisfies this criteria. Additionally, the creators of the BNC1994 sought to include more imaginative writing than was proportional to the population of British writing because of the "influential cultural role of literature and creative writing" (Burnard, 2000: 7). The fiction books genres were the sole genres for which the proportions within the population could be inferred prior to the design of the sampling frame. I, once again, used Waterstones' website to see how many books were listed for sale under each genre category. I then calculated from this the proportions of each genre, and divided the fiction books super-genre accordingly. While only one bookseller, this method at least allowed some approximation to the proportions of different volumes of texts by title in each of the genres used in the corpus. It should still be noted, however, that the length of the works was difficult to assess and, if this had been known, a quite different decision may have been made about the proportions of texts included in the sampling frame.

This chapter outlines in turn each method trialled for the collection of books to include in the Written BNC2014. Section 5.2 details my attempts to contact publishers for their permission to access and include their texts in the corpus. Section 5.3 then details the various other collection methods which were trialled, including using personal contacts, collecting open-access data, collecting free samples, and scanning books and converting them to text using OCR. In section 5.4 I summarise the most successful collection methods discussed in this chapter, before presenting a comparison of the sampling frame and the eventual composition of this medium in section 5.5.

5.2 Contacting publishers

5.2.1 Introduction

As discussed in section 1.5, the vast majority of the texts collected for inclusion in the corpus are exempt from copyright restrictions under the ‘Non-commercial research’ exemption to UK copyright law. This exemption could also be applied to the collection of books for the corpus, providing collection stays within the limits of fair dealing (see section 1.5.2). However, this is not as straightforward for books as it is for other mediums. The vast majority of published books are not freely available online, as is the case for almost every other type of text collected for the corpus. Thus, I was unable to access these texts in order to take extracts from them to include in the corpus.

This is a problem encountered by other corpus creators. It is unclear *precisely* how books were collected for inclusion in the BNC1994. However, due to the rarity of digitised texts in the early nineteen nineties, it is likely that the creators of the corpus either typed up print copies of books, or scanned them and converted them to digital text. Roughly half of the books included in the BNC1994 were selected randomly from Whitaker’s ‘Books in Print’ (1992), with each text being examined to ensure that it fitted all of the relevant criteria (published by a British publisher, fall within the designated time limits etc.) (Burnard, 2000: 10). The other half were selected strategically from bestseller lists, literary prize lists, and library lending statistics in order to make up the target percentages for each category. Before a selected text was included in the corpus, the creators sought to gain permission from the copyright owner (Burnard, 2000: 11). The creators drafted a standard Permissions Request but “some requests were refused, or simply not answered even after prompting, so that the texts concerned had to be excluded or replaced" (Burnard, 2000: 11). This is similar to the procedure used in the creation of the Thai National Corpus (see section 2.4), where publishers were contacted to attempt to gain permission to include their copyrighted texts in the corpus (Aroonmanakun et al., 2009). Aroonmanakun et al. (2009) had great difficulty in securing positive responses from publishers (only 7 out of 22 were able and willing to provide the details needed), and this greatly stalled the progress of the project.

In the creation of the BE06, Baker (2009) only included samples of books which were freely available online. He collected much of the fiction and non-fiction texts from publisher's websites where short samples are made available for free. He also collected free samples from author's own websites. However, this method did present some problems. Samples were sometimes very short, and so were not long enough to fit the sampling criteria (between 1,950 and 2,050 words). Furthermore, in the majority of cases authors only made extracts from the beginnings of their work available, which again did not fit with Baker's corpus sampling criteria.

Thus, it seems that I have several options for collecting published books to include in the Written BNC2014 corpus: i.) I could contact publishers and ask for access to their texts and permission to include samples of them in the corpus; ii.) I could take free samples of books which are available online or iii.) I could find print copies of books which I have legal access to and then convert these to digital text. After some consideration, I decided that the method to try first was to contact publishers to ask for access and permission. This method avoids the problems encountered by Baker (2009) when collecting free samples, and also promised to be much less time consuming than manually converting the very large number of book extracts which I needed for the corpus. Additionally, gaining permission from publishers meant that I would be able to negotiate sample lengths individually, rather than needing to stay within the bounds of fair dealing. The remainder of this section will detail my initial investigation into this method of text collection, and report on its success and the consequences of it for the collection of books for the corpus.

5.2.2 Method

In order to begin this process I first needed to identify a list of British book publishers. It was important that the publishers be British to increase the likelihood of them publishing books which had been written by British authors. Any books collected through this method would then be researched to ascertain, to the best of my ability, the author's native language. Initially, only large publishing houses which provided an email contact were identified in order to increase the likelihood that they would be able to offer us a large amount of data; these publishers were: Wiley, Bloomsbury, Hachette, Harper Collins, Oxford University Press, Palgrave Macmillan, Pan Macmillan, Routledge, Sage Publications, and Penguin Random House UK. The rights and permissions departments at all of these publishers were contacted via email. The publishers were sent an email containing a brief description of the project and an outline of what we needed from them, along with a document which gave more details about the project and how we would prioritise protecting the commercial value of their copyrights (the email text can be found in appendix E and the document can be found in appendix F). The suggested sample size to be taken from their books was 5,000 words (see section 4.3.2 for a justification of this decision).

Alongside this, I also contacted our project partners (see section 1.3.1) at Cambridge University Press (CUP). Our project partners already had a good understanding of what we needed for the corpus, and were happy to engage in discussion with their legal team in order to see if giving us texts for the corpus would be possible. I sent the same document which was sent to other publishers (see appendix F) along with a simple contract which both parties could sign in order to allow the texts to be included in the corpus.

After approximately 3.5 months, any publishers who had not responded to the initial email were contacted again. At this point, due to a very low response rate to the initial email (see section 5.2.3), a further group of 8 smaller British publishers were contacted with the same email as detailed above. These publishers were: Anthem Press, Dunedin Academic Press, Egmont, Hodder & Stoughton, Little Brown Book Group, Orion Publishing Group, Octopus Publishing Group, and Hodder Education. This low response rate was expected based on the findings of Aroonmanakun et al. (2009) when creating the TNC. They found that of the 22 publishers whom they contacted, only 7 were willing to provide them with any information (and the majority simply declined to reply).

5.2.3 Outcome

Of the 21 publishers who were contacted, only eight responded to my email(s) (a summary of the outcomes of this method can be seen in table 5a). Hachette responded stating that I would need to contact the individual permissions departments at each of their imprints. These were publishers whom I had already contacted so no further action was taken. Both Anthem Press and Egmont replied stating that they were discussing my request in house and would get back to me soon. Despite sending follow-up emails, I received no further contact from these publishers. Several of the publishers who responded were unable to process a request for extracts of *any* books which they published in 2014, but rather needed a specific list of books which I would like extracts from. Most of the publishers who I contacted were very large companies who publish very many books in a given year, and thus going through their online catalogues to make a list of books published in 2014 was extremely time consuming. For the publishers who needed this information, I settled on listing a sample of books (around 200 titles) from their online catalogue to give the publishers an idea of what I was looking for. This was the case for Palgrave Macmillan, but after the request list was sent to them I received no further contact. Lengthy discussions were had with both Oxford University Press and Harper Collins to clarify exactly how we would protect their copyrights in the corpus. Both publishers were concerned that the Written BNC2014 user license was a form of creative commons license, which would allow free use and distribution of their copyrighted works. I assured them that the user license was not a creative commons license, and was substantially more restrictive. Redistribution of texts in the corpus is not allowed, and neither is any commercial use of the texts. Any user of the corpus must register their agreement to the licence in order to get access, and we keep a record of who has signed up, and don't put the data online for people who haven't submitted their details. Both publishers were also sent a full copy of the user license, and Oxford University Press were also sent a request list which they had asked for. Despite this, contact from both publishers ceased, despite me sending further emails to both.

Contact with CUP was easier and more productive, although ultimately also did not result in gaining permission to use any of their texts. The legal team at CUP rejected the initial contract which I had sent them as being too simple, and sent back a much more detailed contract. Lancaster University agreed to sign this contract after a few adjustments, however CUP still needed to clarify whether they could legally agree to the contract without contacting individual authors for their permission. After some investigation, they found that it would be necessary to gain permission from individual authors for each book which we wanted to include in the corpus. This finding mirrors the findings of the TNC project - Aroonmanakun et al. (2009) also found that publishers themselves could not give permission to include texts in a corpus, but rather the author needed to be contacted for permission. Clearly, this would be far too time consuming a process for CUP to undertake. Furthermore, I cannot offer any financial incentives to publishers, meaning that they would be giving up a large amount of their time for free. Understandably, this meant that CUP could not grant us access to any of their texts for inclusion in the corpus.

The only publisher who did give us access to and permission to include their texts in the corpus was Dunedin Academic Press. A short discussion was had, in which I sent them the same simple contract which was initially sent to Cambridge University Press. Dunedin Academic Press quickly signed this contract and sent samples of eight of their books to be included in the corpus. The surprising ease of this process, when compared to my interactions with other publishers, is potentially due to Dunedin Academic Press being a smaller scale publisher, and thus not having the large legal departments which are present in bigger publishing houses.

Table 5a: Summary of contact with book publishers.

Publisher	**Contact level**	**Outcome**
Wiley	No response	No data collected
Bloomsbury	No response	No data collected
Hachette	Responded	No data collected
Harper Collins	Responded	No data collected
Oxford University Press	Responded	No data collected
Palgrave Macmillan	Responded	No data collected
Pan Macmillan	No response	No data collected
Routledge	No response	No data collected
Sage Publications	No response	No data collected
Penguin Random House UK	No response	No data collected
Anthem Press	Responded	No data collected
Dunedin Academic Press	Responded	Data collected
Egmont	Responded	No data collected
Hodder & Stoughton	No response	No data collected
Little, Brown Book Group	No response	No data collected
Orion Publishing Group	No response	No data collected
Octopus Publishing Group	No response	No data collected
Hodder Education	No response	No data collected
Cambridge University Press	Responded	No data collected

5.2.4 Conclusion

In summary, after months of input from me in trying to contact and negotiate with publishers, only eight text samples from one academic publisher were collected. This actually reduced to six samples once the authors of each book were researched and non-British authors were excluded. The amount of manual input and time taken to gain this extremely small amount of data clearly proved that this would not be a viable collection method for books in the Written BNC2014. The fact that even Cambridge University Press, who are partners on this project, could not grant us permission to include any of their texts in the corpus emphasises just how cautious publishers are regarding the commercial value of their copyrights. Publishers are bound by legal restrictions which protect their copyrighted material, and, as such, giving permission to use their texts in the way I wanted to is often simply not possible. Furthermore, as no financial incentive could be offered, the publishers would have to give up their time for free in order to work on my request, which they understandably did not want or were unable to do. This was a thorough investigation of this method of text collection for books in the UK, and as such the results of this investigation seem to indicate that, although this method was usable when the BNC1994 was compiled in the 1990s, compiling corpora of books in this way is simply not possible any more, at least in the UK.

5.3 Other book collection methods

The outcomes of the data collection method discussed in section 5.2 make it clear that, if published books are to be included in the corpus, collection will have to be done via a different method. Of course, one may ask why it is necessary to include books in the corpus at all. It would be much easier and quicker to simply fill the corpus with data scraped from the web, and not include published books at all. However, the inclusion of books in the corpus will be extremely important as they will be one of the distinctive contributions of the corpus, and will set the Written BNC2014 apart from the many other corpora of written British English which are available. For example, the enTenTen corpus (Jakubíček et al., 2013) contains around 15 billion words of data which has been crawled from the English web. As well as not representing *British* English specifically, the enTenTen corpus also does not contain any published books (other than any free samples which may have happened to be picked up by the web crawl). Published books are a very important part of British English as they are read by many people, and have an influential cultural role in British English (Burnard, 2000: 7). Thus, despite containing billions of words of data, the enTenTen corpus neglects this key type of data. Another example is the BE06 corpus (Baker, 2009). This corpus *does* contain some published fiction, however this is comprised entirely of free samples found on the web, and the entire corpus is only one million words in size. The books medium of the Written BNC2014 will be many times the size of the books samples included in the BE06. Baker suggests that a corpus the size of the BE06 can only be used to examine high frequency words, and that only very cautious conclusions could be drawn about any other lexis. This highlights the need for books to be included, in large quantities, in the Written BNC2014, in order to allow researchers to investigate less frequent phenomena in this medium.

As it is necessary to include books in the corpus, but getting access to copies of published books proved extremely difficult (see section 5.2), it may seem like a natural next step to collect unpublished and self-published books. These types of books are often made freely available online by the authors, and so would be easy to collect. However, there are several reasons why these types of data will not be suitable for inclusion in the corpus. Firstly, there is not always accurate information available about the authors of these books. This means that determining the ‘Britishness’ of the language being collected is extremely hard to do. The vast majority of published authors have either a Wikipedia page or a biographical page on their publisher’s website which, more often than not, gives information about where the author was born and grew up (see section 5.3.5). This is often not the case for self-published and unpublished authors. Furthermore, much of this type of data is ‘fan-fiction’, which is very often posted under a username which does not reflect the author’s real name. Thus, finding out their nationality would be impossible, even if this information was available online. Secondly, self-published and unpublished books are, for the most part, not professionally edited. This means that these books may contain typos and grammatical errors. Whilst these mistakes are of course a natural part of this particular type of data, they are not representative of the majority of books which people read (as most of these books will have been published and professionally edited). The question of whether these mistakes would then need to be corrected before including samples in the corpus arises. Correcting the mistakes would mean that searching this medium in the corpus would be easier and would give more accurate results. However, doing so would be extremely time consuming, and would also result in my own preconceptions about ‘correct’ British English being imposed on this medium of data. Finally, the end users of the corpus’ expectations must be taken into account. I believe that the vast majority of people using the corpus will assume that the books medium is comprised of published books. Of course, all decisions regarding the corpus and details of the data contained within it will be completely transparent, both in this thesis and in later documentation, so users will be able to find out exactly what is contained within the books medium. However, I believe that the majority of users will not look into this, and will assume that they are working with published books, as was the case in the Written BNC1994. Thus, including self-published and unpublished books in the corpus would be to potentially lead researchers and other users to draw unfounded conclusions from the data.

So it seems that published books must be included in the Written BNC2014, and consequently ways of accessing these texts, other than the method trialled in section 5.2, must be investigated. The remainder of this section reports on the other methods which I utilised in order to collect published books, and discusses the outcomes of these methods.

5.3.1 Professional contacts

5.3.1.1 Method

The first alternative collection method trialled was to contact several publishers, but this time using personal contacts which senior members of the project team had. The publishers who we were able to contact in this way were John Benjamins, Elsevier, Routledge, and Bloomsbury. Unsurprisingly, as the project team is comprised of academics, the majority of these publishers were academic publishers (the exception being Bloomsbury who publish both academic and fiction books), and so this method had limited viability for the collection of fiction or non-academic non-fiction books. Personal contacts at each of these organisations were contacted by the relevant members of the project team, and were given some brief information about the project and details of what we wanted from them.

5.3.1.2 Outcome

No response was received from either Routledge or Bloomsbury through this method, although this was perhaps to be expected based on the findings of section 5.2.

Elsevier seemed initially keen to be involved with the project, and I sent them further details and a request list (as discussed in section 5.2). However, contact ceased after the request list was sent. John Benjamins, on the other hand, were very interested in being involved with the project and quickly sent samples of 45 of their books which were published in 2014. However, unfortunately only seven of these samples were written by British authors and could be included in the corpus.

Thus, although this method did yield some data, it will not be a viable method for collecting large amounts of data for the Written BNC2014. It is likely that the success of this method was limited by similar factors to those found in section 5.2. Publishers are bound by strict legal procedures regarding the copyright of their texts, and, as commercial companies, are understandably very concerned with protecting the value of their commercial properties. Although the fact that including their texts within the corpus would not impact their commercial products in any way was explained fully, section 5.2 showed that publishers were, understandably, still cautious about this. Furthermore, the lack of financial incentives also has a large impact. Although in this method we were contacting people who the project team already had personal or professional relationships with and so would presumably be more willing to help us, it is still the case that we are asking these people to give up a significant amount of their time to liaise with their companies legal departments for free, which they may be unable or unwilling to do.

5.3.2 Contact authors directly

5.3.2.1 Method

The next collection method to be attempted also relied on utilising the contacts of the project team – this time published fiction authors rather than employees at publishing companies. This method is similar to the method utilised by the creators of the TNC once their attempts to contact publishers had proved unhelpful. Aroonmanakun et al. (2009) accessed contact details for many authors (via publishers, internet searches etc.) and wrote to them individually to seek permission to access and include their texts in the corpus.

Two authors were contacted. The first, despite wanting to help, was unable to do so without the permission of her publisher. The second author was enthusiastic, and agreed to help me develop and implement a way in which authors could easily submit extracts of their own published writing for inclusion in the corpus. We decided that the most effective way of doing this would be via the creation of an online form (using Google Forms, see Appendix G for a copy of the form). I created a form which explained to writers what the project was and how they could contribute. The form then asked for the following information: title of book, date of publication, publisher, name of author, author gender, genre of the book, and the author's native speaker status. The authors were then invited to submit an extract (or multiple extracts) of their published books in any widely used file format. For this method I extended the date range from just books published in 2014 to books published between 2013 and 2018, in accordance with the date range policy set out in section 4.3.1. This is because individual authors often do not publish very frequently, and so to limit our collection to only those authors who had published in 2014 would result in a lot of data potentially being lost. The form was publicised via Twitter by the author who had agreed to work with us.

5.3.2.2 Outcome

Unfortunately, this method was almost entirely unsuccessful. In the several months since the form has been available, only one published book extract has been submitted. This seems to suggest that either our promotion of the form via social media channels did not reach our target audience, that authors are not interested in submitting their work to the project, or that they are unable to do so. Based on the response from the first author we contacted (see section 5.3.2.1), it seems that the most likely explanation is that authors are simply not contractually allowed by their publishers to redistribute their published works.

This method was the last possible way of collecting data via the owners or creators of the works. The remainder of the methods discussed in section 5.3 focus on the collection of texts which I have legal access to under either an open-access license or through the ‘Non-commercial research’ exception to UK copyright law (see section 1.5).

5.3.3 Collect open-access data

5.3.3.1 Method

After the lack of success of the methods discussed above, other ways of legally accessing data needed to be sought. For the collection of academic books, one way that this could be done was by collecting books which had been published under an open-access license (see section 1.5.4). An open-access license permits the reuse and redistribution of texts, and so I could collect any texts published under this type of license and include them in the corpus. Academic books are increasingly being published open-access, and so this presented a rich source of data for this super-genre. Of course, it may be suggested that open-access books do not represent the whole population of academic books. This is of course true, and it may be the case that books which are published open access are in some ways linguistically different to academic books which are not published under an open-access license (although this seems unlikely). However, as the methods outlined in previous sections were unsuccessful, and the method used for fiction and non-academic non-fiction collection (see section 5.3.5) would have been too time consuming to extend to academic books, open-access books represented the only viable way of quickly collecting lots of this type of data.

A list of all books listed on a web repository[5] of open access academic books was generated by MT (see section 1.3.1 for a full description of the project team). I then manually narrowed this list to only include books published between 2013-2018 (in accordance with the date range policy set out in section 4.3.1), written in English, and published by a British publisher. The criteria of being published by a British publisher was included to increase the likelihood of the books being written by a British author, but of course by no means guarantees this. Of course, researching each author individually, as was done for the samples sent by Dunedin Academic Press and John Benjamins, would have been the most effective way of guaranteeing ‘Britishness’. However, researching this information for each author would have been far too time consuming, and so for this data collection method, publisher location was the only criteria for indicating ‘Britishness’. It is the case that all books published by a British publisher will have gone through a British editorial process, so even if the author is not British their language will have been standardised to some degree. Furthermore, academia is extremely international, and it is certainly the case that British academics are very frequently reading academic books which were not written by native speakers of British English. Thus, whilst this method does not ensure that

[5] http://oapen.org/

the language contained within this super-genre represents what is being *produced* by British academics, it is at least representative of the language which they are *receiving.*

The identified 463 (out of a possible 4588 books listed on the website) books were then automatically downloaded by MT, as manual collection would have been far too time consuming. A script was written to collect the text from the open-access sources, and then another script was written to transform the text into a format suitable for inclusion in the corpus. Samples of around 10,000 words were then taken from each book, ensuring that a balance was kept between samples from the beginning, middle, and end of the books. In theory, I could have included the books in their entirety in the corpus, because they are published under an open-access license. However, samples were taken in order to avoid any very long books skewing results, and also to maximise the amount of books which could be included in the corpus.

The text still needed cleaning manually, in order to remove any text (such as page numbers, reference lists etc.) which was present but which should not be included in the corpus. I manually removed all page numbers, any chapter or book titles which were present as headers on the pages of the books, and all reference lists, glossaries, and indexes. Reference lists, glossaries and indexes were removed because they added a lot of words to a sample, and I did not want to populate the corpus with excessive amounts of very predictable and linguistically uninteresting language. For the majority of books, reference lists, glossaries, and indexes were presented at the end of books, and were often many thousands of words long. This resulted in the removal of the majority of the data from some of the end samples (this will be returned to in section 5.3.3.2).

5.3.3.2 Outcome

The amount of data collected via the first trial of this method can be seen in table 5b. The target of 900,000 words was only reached for one genre, but for three out of the remaining five genres over 80% of the required data was collected. The population of British academic books was not known prior to collection and so the sampling target for each genre was set at 900,000 words (see discussion in section 5.1). However, after this initial round of data collection it seemed that medicine and natural science books may comprise a smaller percentage of the population than the other genres of academic books. However, I did not deliberately seek to replicate the proportions seen here in the full sample, as it is important to remember that this only indicates the population of *open-access* academic books. It may be the case that, rather than less medicine or natural science books being published, these genres of books are simply not published under open-access licenses as often as books within the other academic genres. Thus, I do not want to give too much weight to the proportions found in this sample.

I increased the sample size for each book in order to make up the desired word counts. This is particularly important for the end samples, as some of these ended up being less than 1000 words long after cleaning (discussed in section 5.3.3.1). Overall then, this collection method was highly successful. Via a combination of automatic and manual procedures all of the data needed for this super-genre was collected in a relatively short period of time.

Table 5b: Number of words of academic books collected via the initial trial of the open-access method.

Genre	**Words collected**	**Number of books sampled**	**Target words**
Humanities & Arts	741,762	99	900,000
Medicine	270,688	12	900,000
Natural Science	482,340	15	900,000
Politics, Law, & Education	845,165	25	900,000
Social Science	900,990	25	900,000
Technology & Engineering	891,875	25	900,000

5.3.4 Collect free samples

5.3.4.1 Method

The success of the method discussed in section 5.3.3 still left fiction books and non-academic non-fiction books to be collected. As getting samples directly from publishers or authors was not feasible, and as fiction and non-academic non-fiction books are not typically published under open-access licenses, collection had to be done within the bounds of the 'Non-commercial research' exemption to UK copyright law (see section 1.5). The easiest way to do this would have been to collect free samples from online, in a similar way to Baker (2009) in the creation of the BE06. Collecting free samples would fall under this exemption as I have legal access to the work, am only taking a small amount so am staying within the bounds of fair dealing, and the inclusion of the samples within the corpus will not harm their commercial value in any way.

Two main sources of data were targeted in this collection method: free samples on publishers' websites, and free samples available on Amazon.co.uk. Amazon.co.uk represents a large potential data source as they make free samples available for almost all of the many thousands of fiction and non-fiction books which they sell. Some British publishers also release short extracts of their books for free on their websites. As Amazon.co.uk represented the largest potential data source, this was where mine and the team's efforts were focused for this method.

However, we discovered that data could not be collected manually from Amazon.co.uk, as the majority of the free samples are not able to be copied through a web browser. Thus automatic collection by MT was trialled. However, it quickly became clear that Amazon.co.uk heavily protect their free samples against being collected, as it proved impossible to collect these samples via automatic methods either.

5.3.4.2 Outcome

This method, although seeming initially promising, yielded very little data. No data was able to be collected from Amazon.co.uk, and only a very few samples were collected from publishers' websites. As was found by Baker (2009) these samples were often very short, and almost exclusively from the beginnings of books. Thus, this method was not viable for the vast majority of the collection of fiction and non-academic non-fiction for the corpus.

5.3.5 Manual scanning and OCR

5.3.5.1 Methodology

After trialling numerous other methods (detailed above), the only remaining feasible method for the collection of fiction and non-academic non-fiction books was to scan print copies of books and convert these to text using Optical Character Recognition (OCR) software. The collection of data in this way will fall under the 'Non-commercial research' exemption to UK copyright law (section 1.5) as I was scanning works which I have legal access to, I was only taking samples of each book so was staying within the bounds of fair dealing, and the inclusion of the samples within the corpus will not harm the text's commercial value in any way.

The procedure for this method was as follows: first identify a book, either from a library or from mine or friends and colleagues' personal collections, which was written by a British author between 2010 and 2018. Books were initially selected at random, in order to speed up data collection. Later, bestseller lists were used to ensure that any gaps in genres were being filled by books which had been read by a large number of people. This is the same method used by the creators of the BNC1994 (Burnard, 2000). The Britishness of each author could be quickly identified through a Google search, as most published authors have a Wikipedia page which contains biographical information, or have a biography available on their publisher's website. This was the widest date range used for any type of data collection in the corpus, and was expanded to this extent in accordance with the date range policy set out in section 4.3.1. As every previous method of data collection had failed, I felt strongly that I did not want to limit the possible sources of collection for this method too much. Stretching the date range back to 2010 still ensures that the data collected is representative of the fiction and non-academic non-fiction published in this decade, whilst not unnecessarily excluding sources of data. This is still a much smaller date range than was used for some data in the Written BNC1994, and is also a much smaller date range than was used in the collection of books for the SYN2015 corpus (Křen et al. 2016). Křen et al. (2016: 2523) designed the SYN2015 to be representative of contemporary written Czech, but collected fiction books which were published within the previous 25 years (and first published within the previous 75 years). So it seems that increasing the date range for the collection of books to be included in contemporary corpora has a precedent.

Once a suitable book had been identified, around 50 double pages from either the beginning, middle, or end of the book were scanned. 50 double pages was set as the average length for a sample, but this was flexible and was adjusted to ensure that no more than 50% of a book was scanned. Fair dealing is typically assumed to be a smaller proportion of a text than this, however there is no formal definition of fair dealing and it is assessed on a case by case basis. Gov.uk (2017) suggest that relevant factors are whether the use of the work affects the market for the original work, and whether the amount of work used was reasonable, appropriate and necessary. As discussed in section 1.5, the inclusion of samples of texts in the corpus will certainly not affect the market for the original work. For the purposes of this project, as other data collection methods had failed, the copying of around half of a book was certainly *necessary* in order to collect the amount of data needed. I would also argue that this was *reasonable* and *appropriate,* although these are of course very subjective criteria.

To carry out this procedure for the full amount of books needed for the corpus would be extremely time consuming. Thus, myself and the project team decided to take a ‘public participation in scientific research’ (PPSR; see Shirk et al., 2012) approach to this problem. We ran a data collection training session, in which participants had the opportunity to learn more about the Written BNC2014 project, learn about corpus creation methods, and, critically, help us collect data for the corpus. The event was advertised to students both at Lancaster University and nationwide, and 50 people signed up to attend the event. Participants were given instructions on how to select and scan books (see appendix H), and were given access to the university library and scanners. After they had spent several hours scanning a large number of book samples, they were taken to a computer lab where they could submit their scans via a Google form (see appendix I). Feedback from participants indicated that they had enjoyed and valued the opportunity to be involved in the project, and everyone was encouraged to keep collecting book samples and submitting them via the Google form. In order to incentivise people to do so, the team has given out small prizes to the ‘contributor of the month’ since the event. Everyone who submits a book scan via the Google form will be fully credited in the corpus documentation.

5.3.5.2 OCR comparison

Finally, the scanned texts needed to be converted from image files into text files. This was done using OCR. Several different OCR programmes are available, so I carried out an experiment on some initial data samples to identify which programme would work best for this project.

In this small study I compared three different OCR programmes: Adobe Pro OCR, Tesseract OCR and Google OCR. I selected 10 scanned books at random using a random number generator, and then carried out the same tests on them using each OCR tool. Before this could be done, each book sample had to be stitched back together from the individual scans (see appendix J for a full list of instructions for the OCR conversion), which was very time consuming. Each book sample was converted to text using each of the different tools, and the word counts of each sample were compared. The results of this comparison can be seen in table 5c.

Table 5c: A comparison of word counts for each document when converted to text using three different OCR tools.

Document	Google OCR word count	Tesseract OCR word count	Adobe Pro OCR word count
1	9,149	9147	9,079
2	15,676	15,953	15,790
3	30,775	31,031	31,139
4	12,290	12,296	12,431
5	24,763	5,222	25,500
6	33,770	34,427	34,133
7	28,218	28,556	28,681
8	17,957	18,074	18,081
9	24,812	25,362	25,093
10	27,482	27,612	27,736
Total	**224,892**	**207,680**	**227,663**

Google OCR and Adobe OCR produce fairly similar results, although different enough to clearly indicate that the two tools work differently. Tesseract OCR produces a significantly lower overall word count than the other two tools. However, this difference is mostly contributed by text 5. Tesseract OCR completely failed to convert text 5 to any kind of recognisable text, perhaps suggesting that Tesseract OCR is more error prone than the other tools.

The next comparison which I carried out was a detailed analysis of the types and amounts of errors which each tool produces when converting an image to text. The first five pages of five of the texts were compared in each tool. The results of this comparison can be seen in table 5d.

Table 5d: A comparison of the amounts and types of errors found when using three different types of OCR tool.

Text	Wrong character	Extra character	Missing character	Extra space	Missing space	Total
1 (Google)	1	1	2	0	0	4
1 (Adobe)	2	0	0	2	0	4
1 (Tesseract)	10	0	0	0	0	10
2 (Google)	0	0	0	0	0	0
2 (Adobe)	1	1	0	0	0	2
2 (Tesseract)	6	0	0	0	2	8
3 (Google)	0	0	1	0	0	1
3 (Adobe)	8	8	11	8	17	52
3 (Tesseract)	38	18	31	15	3	105
4 (Google)	0	1	0	0	0	1
4 (Adobe)	0	0	0	0	2	2
4 (Tesseract)	0	2	0	0	1	3
5 (Google)	3	0	0	7	2	12
5 (Adobe)	27	10	22	9	18	86
5 (Tesseract)	-	-	-	-	-	-

Note: The row labelled ‘5 (Tesseract)’ is left blank because the text was converted completely incorrectly, and thus the entire text was comprised of errors.

The types of errors encountered in each tool were wrong characters (e.g. an exclamation point being converted to a colon), extra characters (i.e. characters being introduced where none are present in the original), missing characters, extra spaces, and missing spaces. In this small comparison Google OCR far outperformed the other tools, with only 18 errors across all 5 texts, compared to 146 in Adobe OCR and over 126 in Tesseract OCR (an exact figure is not given as text 5 was so badly converted in Tesseract OCR). It seems that for relatively straightforward scans all three tools perform fairly similarly (e.g. texts 1 and 4). However, when a ‘messier’ scan is encountered (i.e. a scan where the book was not placed at 90 degrees on the scanner, or where the pages were not flattened properly, or where the book is printed in an unusual font, or includes pictures), Adobe OCR and Tesseract OCR seem to perform much worse than Google OCR (e.g. texts 3 and 5). As we hoped to gain most of the scans for this data collection method via submission by members of the public, it was fair to assume that many of the scans may be imperfect, as we will not be observing them to ensure they are done perfectly. Thus, Google OCR was the obvious choice for the conversion of scans of fiction and non-fiction books.

However, even when using Google OCR many mistakes are still introduced to the text which need to be cleaned manually. This is extremely time consuming, and so an intern was hired to the project to help with this aspect of data collection. She continued to scan books, and also converted the scans using OCR, and cleaned them. She found that this was a very time consuming process, sometimes needing to spend several hours cleaning just one book sample thoroughly. Based on this, I decided that, when cleaning books, the individual should spend 15 minutes working on a book and if after this time the individual felt that this book would take longer than one hour to clean, then the sample should be discarded. A record of any discarded samples was kept, in the hope that they could be scanned again in a way which would introduce fewer errors.

5.3.5.3 Outcome

This method of data collection has proved successful for the collection of fiction and non-academic non-fiction books. However, as detailed above, this method of data collection is extremely time consuming and requires by far the most manual input of any method of data collection used for any genre in the corpus. Nevertheless, this is the data collection method which was used for the collection of fiction and non-academic non-fiction texts in the corpus because it is the only remaining feasible collection method.

5.4 Conclusion

At the beginning of the project, it was assumed by myself and other members of the project team that contacting publishers would be the easiest and quickest way of collecting published books to include in the corpus. However, as this chapter has shown, this was far from the case. The three methods trialled which involved contacting publishers or authors yielded very few results, and were also time consuming. This made them unfeasible for data collection for the Written BNC2014. This is a finding echoed in other national corpus projects. Aroonmanakun et al. (2009) cite gaining access to and permission to include copyrighted texts in the corpus as the biggest obstruction causing a delay to the TNC project.

One clear best option for the collection of academic books emerged – the collection of books published under open-access licenses. This data collection method allowed academic books to be collected quickly and easily, with no consideration of UK copyright law, and is the method which has been used for the collection of all academic books within the Written BNC2014. The method which will be used to collect fiction and non-academic non-fiction books for the corpus did not necessarily emerge as a *best* option, but was the best *remaining* option after all other collection methods had been trialled. The scanning and OCR conversion procedure allowed very targeted collection of books, as any gaps in the sampling frame could be filled by choosing a book from the library which exactly matched the required criteria. The time consuming nature of the method was sped up by involving members of the public in the project. However, this method, particularly the OCR conversion, was still extremely time consuming.

5.5 Composition of the books medium of the Written BNC2014

The majority of this chapter has discussed the collection of books in relation to the sampling frame set out in appendix B. However, it has become clear throughout this project that the ideal design of a corpus is very difficult to achieve in practice. Thus, this section will compare the sampling frame in appendix B to the reality achieved for the collection of books for the corpus (the full corpus proportions can be seen in appendix C), and discuss any differences and why these occurred. Table 5e shows the books medium of the sampling frame, and table 5f shows the eventual composition of the books medium of the corpus (although at the time of writing, the numbers given are still provisional).

As can be seen from tables 5e and 5f, the amount of data collected from books is in line with what was hoped for in the sampling frame, i.e. 41% of the corpus is indeed comprised of books. For the academic prose super genre, an exact match with the sampling frame was achieved, using the method discussed in section 5.3.3. For the fiction super genre, the overall amount of data collected is in line with what was hoped for in the sampling frame, but the genre labelling and the distribution of this data between the genre categories has changed somewhat. Firstly, it became apparent that it would be difficult to collect enough data from children's books to fill the 'W_fict_prose_childrens' genre. This was because, by their nature, children's books are short, and so only taking a sample results in very few words being collected per book. Secondly, it became clear that distinguishing between children's fiction and teen fiction was difficult, and very subjective. For these reasons, the children's fiction and the teen fiction genres were merged to create one, larger genre.

The second change made to the fiction super genre was the redistribution of data from the poetry genre to the general fiction genre. It was originally planned for 2% of the corpus (or 1.8 million words) to contain poetry. However, similar problems were encountered to the children's books in terms of text length. Books of poetry tend to be short, and the poems themselves often contain few words. When only taking a sample of a book, this results in very little data being collected per book. Therefore, the poetry genre had to be reduced in the eventual make-up of the corpus, shrinking from a planned 1.8 million words to just 100,000. The remaining data was redistributed to the general fiction genre, as this genre classification was the most broad and so would be the easiest to collect more data for.

A very obvious change has occurred in the collection of the non-academic non-fiction books between the sampling frame and reality. Whilst the amount of data collected is consistent with the sampling frame, this super-genre has been reduced to just one genre, rather than the planned seven. It became clear when it came to classifying the non-fiction books that the genre labels devised by Lee (2001), and used in the corpus sampling frame, were wholly inadequate for the data which had been collected. Rather than falling into the academic classifications used in the BNC1994 corpus, the texts collected for the BNC2014 were much more often to do with hobbies (e.g. gardening, sports) and topics of interest (e.g. celebrities, pop culture). However, there were not enough similarities amongst the texts to devise a new classification scheme. Therefore, the genres within this super genre were condensed into one, general genre into which all non-fiction books will be categorised. Of course, if it seems useful, these texts can always be further classified by end users of the corpus, as was done by Lee (2001) for the BNC1994.

Overall then, the books medium of the Written BNC2014 looks, for the most part, very close to what was planned in the corpus sampling frame. The amount of data planned for all of the super genres has been achieved in collection, with minor changes to how this is distributed within the super genres. The success of the collection of this medium is largely due to the high level of manual input in all forms of collection. This allowed the team to target the exact genres of books which were needed to achieve the sampling frame proportions.

Table 5e: The books medium of the Written BNC2014 sampling frame.

Medium	**Super genre**	**Genre**	**Target**	**Words**
Books (41%)	Academic Prose (textbooks, academic books etc.)	W_ac_book_humanities_arts	1%	900,000
		W_ac_book_medicine	1%	900,000
		W_ac_book_nat_science	1%	900,000
		W_ac_book_polit_law_edu	1%	900,000
		W_ac_book_soc_science	1%	900,000
		W_ac_book_tech_engin	1%	900,000
	Fiction	W_fict_poetry	2%	1,800,000
		W_fict_prose_general	9%	8,100,000
		W_fict_prose_childrens	2%	1,800,000
		W_fict_prose_teen	2%	1,800,000
		W_fict_prose_sf_fantasy	2%	1,800,000
		W_fict_prose_crime	2%	1,800,000
		W_fict_prose_romance	2%	1,800,000
	Non-academic prose (non-fiction)	W_non_ac_humanities_arts	2%	1,800,000
		W_non_ac_medicine	2%	1,800,000
		W_non_ac_nat_science	2%	1,800,000
		W_non_ac_polit_law_edu	2%	1,800,000
		W_non_ac_soc_science	2%	1,800,000
		W_non_ac_tech_engin	2%	1,800,000
		W_non_ac_biography	2%	1,800,000

Table 5f: The eventual composition of the books medium of the Written BNC2014.

Medium	**Super genre**	**Genre**	**Target**	**Words**
Books (41%)	Academic Prose (textbooks, academic books etc.)	W_ac_book_humanities_arts	1%	900,000
		W_ac_book_medicine	1%	900,000
		W_ac_book_nat_science	1%	900,000
		W_ac_book_polit_law_edu	1%	900,000
		W_ac_book_soc_science	1%	900,000
		W_ac_book_tech_engin	1%	900,000
	Fiction	W_fict_poetry	0.11%	100,000
		W_fict_prose_general	10.89%	9,800,000
		W_fict_prose_childrens_teen	4%	3,600,000
		W_fict_prose_sf_fantasy	2%	1,800,000
		W_fict_prose_crime	2%	1,800,000
		W_fict_prose_romance	2%	1,800,000
	Non-academic prose (non-fiction)	W_non_ac_book_general	14%	12,600,000

Chapter 6: Collection of periodicals for the Written BNC2014

6.1 Introduction

In this chapter I will discuss the various processes used to collect the data for the periodicals medium in the Written BNC2014. The majority of the discussion in this chapter will focus on the collection of periodicals in relation to the Written BNC2014 sampling frame (see appendix B). Section 6.5 will compare the periodicals medium in the sampling frame to the actual collection figures achieved for the corpus (see appendix C for the eventual composition of the full Written BNC2014). As has been mentioned before in this thesis, data collection is still in its final stages, and as such, any figures given are provisional and subject to change.

The periodicals medium in the Written BNC2014 sampling frame is comprised of 5 super-genres: academic prose (journal articles), broadsheet national newspapers, regional & local newspapers, tabloid newspapers, and magazines (see chapter 4 and the sampling frame in appendix B for more details). I aimed to collect 5.4 million words of academic prose, split equally across 6 different genres (humanities & arts, medicine, natural science, politics, law & education, social science, and technology & engineering). These disciplines were chosen because they are directly comparable to the genre distinctions made for this type of text in Lee's (2001) BNC1994 genre scheme. The goal of 900,000 words per genre was selected because it allows each genre to be of a useful size for analysis in its own right (see section 4.3.3). This is a larger amount of data from journal articles than was included in the BNC1994, which included just under 2.7 million words across 153 academic periodical texts.

Each super-genre of newspaper in the sampling frame is comprised of the same 7 genres: arts & entertainment, commerce, editorial, reportage, science, social, and sports. It was aimed for each of these genres contains 900,000 words of data, for a targeted 18.9 million words of newspaper data overall. As already discussed in chapter 4, the proportion of newspapers in the Written BNC2014 sampling frame has doubled compared to the BNC1994. More newspaper texts were included in the new sampling frame to address the imbalance of newspaper types in the 1994 corpus. In the 1994 corpus much more data was included from broadsheet, and regional and local newspapers than tabloid newspapers. The amount of texts from these three types of newspapers are equal in the 2014 sampling frame, avoiding any implication that one type is more 'important' than another. Consequently, I aimed for newspapers overall to be present in a much higher proportion in the Written BNC2014 relative to the Written BNC1994, with the most significant increase being in the proportion of tabloid news texts. Additionally, I predicted that newspaper texts would be relatively straightforward to collect for the new corpus (see section 6.3). The genres within each type of newspaper in the sampling frame are the same as those into which broadsheet newspapers were split in Lee's (2001) genre scheme. Again, these genres have been extended to the other types of newspaper in order to avoid any implication that one type of newspaper deserves more attention than another. As discussed in chapter 4, proportional representation within the sampling frame was not possible for the vast majority of genres because these populations simply could not be known before collection was underway. Thus, each newspaper genre was allocated the same proportion in the sampling frame. However, it became clear once collection began that collecting each newspaper genre in equal amounts would not be possible. Thus, the newspaper genres actually ended up being proportionally represented for each type of newspaper (see section 6.5, and appendix C for more details).

My design aimed to include 7.2 million words of magazine data in the Written BNC2014 (see sampling frame in appendix B). This decision was made because magazines made up roughly 8% of the BNC1994 (although labelled ‘popular lore’ rather than ‘magazines’), and I wanted to keep this proportion similar in the Written BNC2014. The magazines super-genre is divided into 8 genres in the sampling frame (lifestyle, men’s lifestyle, TV & film, motoring, food, music, science & technology, and sports), each of which was allocated 900,000 words. There are no widely accepted genre classifications for magazines, so I developed these labels after becoming familiar with a wide variety of magazines and seeing what categories naturally emerged (see section 6.4).

In this chapter I will discuss the collection process for each of the above types of data within the periodicals medium. Section 6.2 discusses the collection of academic prose (journal articles); section 6.3 discusses the collection of the 3 super-genres of newspaper; and section 6.4 discusses the collection of the magazine texts. Finally, in section 6.5 I discuss the eventual composition of the periodicals medium, and compare this to the BNC1994 and to the original sampling frame for the Written BNC2014.

6.2 Collection of academic prose (journal articles)

The broad, initial parameters for the collection of academic prose, in line with the goals of the corpus overall, were that the data should be from journal articles written by native speakers of British English, published in 2014. However, both of these seemingly simple parameters presented problems. To ensure that all authors were native speakers of British English would have required manual research into each author for every one of the journal articles considered. Even with substantial manual effort, moreover, it might prove impossible to obtain this information. Whilst it would be largely possible to obtain information about the academic institution where each author currently works, many people do not make public online the country in which they were born or what their first language is. Clearly, this method would be far too time consuming to implement, and may not have yielded results even if undertaken.

Another approach to ensuring the collection of British English would be to screen articles automatically for the presence of British spelling variants, excluding texts where American spelling variants were used. However, it is quite clear that when creating a resource for the study of contemporary British English, the creators should not predetermine what is seen as 'British' English orthography. In other words, if American spelling variants have become a part of Written British English, then this method would lead to this development being entirely missed.

The final option considered, and the option which was ultimately used, was to only collect articles from journals published by British publishers in which at least one author was affiliated to a British institution. This was relatively easy to do as the websites which were used to compile potential journal sources have a filter for the publishing location of a journal. Whilst collecting articles from British publishers does not, of course, ensure that the authors of those articles are speakers of British English, it does ensure that those articles have been through a process of review and editing in order to ensure that they conform to 'British' standards. As mentioned, the British publication criteria was supplemented by manually searching each article for the author affiliations. If at least one author was affiliated to a British institution then the article was included, otherwise the article was excluded from the corpus. Of course, being affiliated to a British institution is no guarantee of native British status. However, I felt that relying on British publications alone was not enough, as anyone can submit work to a British journal. Thus, combining these two methods increases the likelihood of at least one author being British, or at least one author being familiar with British academic writing standards. Furthermore, the international nature of academic communities (for example, many of the journal articles cited in this thesis have been written by multiple academics who are affiliated to institutions in different countries, and a great many have been written by academics affiliated to institutions outside of the UK) may mean that the 'Britishness' of an author is not particularly important when considering language reception rather than production. Many of the academic journal articles read by British people will not have been written by British authors, and so, from a language reception perspective, including authors of other nationalities is representative of the academic language which British people are reading.

It became apparent during collection that limiting collection to only articles published in 2014 would make it difficult to collect enough data. In light of this, and referring to the date range policy set out in section 4.3.1, I decided to collect data from 2014 and 2015, as this ensured that I would be able to collect enough data, whilst still representing contemporary British English.

Another parameter to consider was what journals to collect data from. As discussed above, the journals would need to be published by a British publisher, and would need to fit into one of the categories in the sampling frame. However, it was also important that the text could legally be made publicly accessible. One potential way of doing this would be to contact the journal publisher who could potentially provide data and ask for their permission to include excerpts of their articles in the corpus. However, based on the lack of success when this approach was applied to books (see chapter 5), this would probably be time consuming and give little in the way of results. It is likely that many publishers would simply not respond, and that others would require lengthy legal procedures to be followed where contracts would need to be drawn up and agreed upon. A much easier, and less time consuming, solution was to only collect data from journal articles which are published under an open-access license (section 1.5). This process had already been used with great success for the collection of academic books (see Chapter 5). An open-access license means that the articles are freely available online, and that redistribution or republishing of the articles is permitted (see section 1.5). This also meant that entire articles could be collected for inclusion in the corpus because I did not need to consider the limitations of working within fair dealing. This was helpful because fewer samples overall would be required, which went some way to balancing the high level of manual input to check the author affiliations for every article.

Following these decisions, a list of potential journals for each genre from which to collect articles was drawn up. Journals in which all articles are published under an open-access license were considered first, in order to maximise the amount of data being collected. However, later on in the process individual articles which were published under an open-access license in a journal which was not otherwise open-access were also considered. This was not as straightforward as it may sound. Many journal article topics are interdisciplinary, for example, an agriculture journal could be considered natural science, social science, or technology & engineering depending on the exact focus of any article in it. In these cases journals were categorised according to the publisher's classification, or where this information was not available I made a decision about which category they fitted best. I felt that manual collection of this text type would be far too time consuming, so the collection process was automated by MT (see section 1.4.1 for a full discussion of the project team). Very briefly, this involves writing a script which collects the text from a specific issue of a specific journal, and then running another script on this text to ‘clean’ it and ensure that it is suitable for inclusion in the corpus. As already mentioned, I then manually searched each article for affiliations, and excluded those articles which did not have at least one author affiliation to a British institution. At this stage I also carried out some basic cleaning of the data, including the removal of reference lists and the correction of any other errors which had been introduced in the collection process (such as hyphens at the ends of lines). Reference lists were removed because they added a lot of words to a sample, and I did not want to populate the corpus with excessive amounts of very predictable and linguistically uninteresting language. This first round of automatic collection was then supplemented by manual collection to achieve the target amounts of data in each genre. Manual collection was used at this stage as the cleaning process for the automatically downloaded texts was extremely time consuming, and collecting texts manually meant that a lot of this cleaning could be done at the same time as collection.

6.3 Collection of newspapers

The collection of newspaper texts began with the same broad parameters as journal articles: written by a native speaker of British English, and published in 2014. Once again, it would have been impossibly time consuming to manually research every journalist whose work we may want to include in the corpus. Part of the solution used for the collection of journal articles is also applicable here: looking only at British newspapers ensures that the texts have been through a British editorial process, and are representative of the newspaper texts which British people are reading. Once again, collecting articles only from 2014 may have resulted in too small an amount of data being collected, so articles were collected from 2014, 2015, and 2016, in accordance with the date range policy set out in section 4.3.1.

Two approaches were taken to the collection of newspaper articles, both of which will be discussed in the following sections. Neither of these approaches were carried out by me – all factors were discussed by myself and the project team, but MT and CD (see section 1.3.1 for a full discussion of the project team) were ultimately responsible for the collection of these texts. For this reason, I keep my explanations of these processes brief.

6.3.1 Automatic Scraping

The first method used for collecting newspaper texts was the automatic scraping of British newspaper websites. This process was carried out by MT and was used initially because it can be automated so that a very large amount of text can be acquired with relatively little manual input. As in the collection of journal articles, a list of potential sources were identified, and a script was written to download all of the web pages from the relevant website; and then another script was run on this text to extract and 'clean' the articles to ensure that they are suitable for inclusion in the corpus.

As with journal articles, it was important to ensure that I could legally take extracts from newspapers and redistribute them to users of the corpus. As we will not be seeking permission from the copyright holders of these articles, they will be collected under the 'Non-commercial research' exception to UK copyright law (see section 1.5). The inclusion of the texts in the corpus will have no impact on their commercial value; most users of the corpus will only see small extracts of the texts, and even for users who download the corpus, the texts will be heavily marked-up with xml which will make them extremely difficult to read. Furthermore, the texts are all already freely available online, and so anyone can read these texts for free already – their inclusion in the corpus will not affect this. We must also satisfy the requirements of fair-dealing when using the texts for non-commercial research. As shown in section 1.5, fair dealing has no formal definition, but it has been suggested by courts that the use of work should be reasonable, appropriate and necessary. The use of these texts is certainly appropriate and necessary for the project. Additionally, when considered in relation to the many millions of words present on these newspaper websites, the amount of articles which we have taken from just a few years of reporting is likely to be considered entirely reasonable.

In the case of newspapers, not all of the articles present in a print copy of a newspaper may be replicated online, and not all of the articles present online may have appeared in a print newspaper. However, I felt that for newspapers this discrepancy would likely not be enough to impede the accurate representation of British newspaper texts. In this context it is relevant that, according to Ofcom's (2017) study of news consumption in the UK, circulation of national daily titles has decreased from 9.2 million in 2010 to just 6 million in 2016, but with online readership adding considerably to overall consumption figures. In fact, Ofcom (2017) find that the only 2 titles which have more print readers than online are The Times/Sunday Times and the Metro. This indicates that most people consume their news online nowadays, and so collecting newspaper articles from online sources may actually be more representative of what a contemporary speaker of British English reads than collecting print articles would be.

Although it was originally hoped that this collection method would be quick, easy, and generate lots of data, this turned out only to be partially true. This method resulted in the collection of huge amounts of data, in fact, much more data than we could ever need for the corpus. However, this method was not as time efficient as had been originally hoped. Due to the fact that each newspaper website is different, MT had to write new scripts for each website. Additionally, each script takes a long time to run for each paper – sometimes meaning that it would take months to extract all texts from a given newspaper. Furthermore, the large amount of data which was generated, whilst excellent in terms of sheer amounts of words, created problems. Such a large dataset needed to be down-sampled, and lots of time was dedicated to deciding how this would be done. Next, the texts needed to be categorised into genres, but information which could help with this (such as the section of the website which the article came from) was not always preserved, and so this process could take a great deal of manual input and time.

6.3.2 LexisNexis

It became clear, due to the factors discussed in section 6.3.1, that another method would be needed for the collection of newspapers, which was more time-efficient. LexisNexis is an online database which contains, amongst other things, copies of many British newspapers. In brief, the process carried out by CD was as follows:

1. Identify a target newspaper which is available on LexisNexis
2. Generate a random set of dates between 2014 -2016. Enough days were generated to guarantee that the amount of data collected would exceed the target amount for that newspaper, based on an estimate of how much data

would be available for one day in the newspaper. Typically, this was 40 days for tabloids and broadsheets and 20 days for regional & local newspapers. Where newspapers published Sunday editions, care was taken to ensure that some Sundays were present in the sample.

3. Every article published (and available through LexisNexis) on the randomly generated days were collected.
4. Steps 1-3 repeated for every newspaper.
5. All data is checked for duplicates, to ensure that the same news article is not included in the corpus more than once.
6. Articles are categorised into genres.

Whilst this method of collection does require much more manual input throughout the whole collection process, its chief benefit over the automatic scraping method is that one newspaper can be collected in a matter of hours. The data which is collected required very little cleaning, and, because not as much data was collected as with the automatic scraping method, nowhere near as much down-sampling was required. Furthermore, LexisNexis allows users to collect data from the print copies of newspapers. Where possible, print copies of newspapers were collected in this method, thus ensuring that the data directly matches print copies and avoids the problems discussed in section 6.3.1 and 6.4. LexisNexis also provides information about what section of the newspaper an article came from, which has greatly assisted in the categorisation of the articles into genres.

6.4 Collection of magazines

6.4.1 Introduction

In this section I will investigate to what extent print and online magazine content are the same, and explore how this will impact on the collection of magazine articles for inclusion in the Written BNC2014. It is important here to make clear what I mean by an 'online magazine'. I am not referring to e-copies of print magazines, but rather to the freely accessible websites which are run by many popular print magazines. It was clear from the collection of books (see chapter 5) that magazines would have to be collected online, rather than asking publishers for print copies of their magazines to include in the corpus.

I decided to begin data collection with this super-genre as it would be easy to determine objectively what magazines should be targeted (by looking up readership statistics to determine popularity, see section 6.4.2), and I would not need to seek permission to collect the data (under the copyright exceptions discussed in section 1.5; see section 6.3.1 for a full discussion of how this exception is utilised when collecting data from web pages). However, an initial look at several magazine websites immediately revealed that the content on them seemed to be rather different to that which appears in the corresponding print magazines (much more different than online and print newspapers). For example, articles were frequently very short, contained lots of pictures or videos, or were written in list form. Thus, I decided that more investigation needed to be done before I could be confident in online magazines as an adequate substitute for print magazines in the Written BNC2014.

6.4.2 Methodology

6.4.2.1 Data selection

To begin my data collection, I looked at statistics for the top 100 print and digital magazines by circulation in 2014 (Durrani, 2015). I narrowed this list according to which magazines had freely accessible websites, which left me with 71 magazines which could be considered. I categorised these magazines according to topic (men's lifestyle, TV and film, motoring, food, music, lifestyle, technology, sports, and miscellaneous), and also according to publisher. I then selected ten of these titles, representing a spread of topics and publishers, and purchased a copy of each in print. The magazines considered were: *Good Housekeeping*, *Stuff, Empire*, *Cosmopolitan*, *Q, Good Food, Top Gear, Mountain Biking UK*, *Tatler*, and *British GQ*.

6.4.2.2 Analysis

Using these single issues of each of the ten magazines under analysis, I systematically went through each print magazine and checked whether each print article was present on the magazine's website (excluding editor's letters, promotional articles, competitions, and other material other than standard articles). I used the Google search engine to query the website, using words from the article's headline, or key phrases from within the article. This method does not ensure 100% success because some articles from the print magazines were present online but with changes (which I will discuss shortly). However it seems likely that for the most part I was able to identify which articles were replicated online. There were also problems with determining what constituted an 'article' in a magazine. Are features which consist of only pictures articles? Are multi-page features which contain multiple pieces under separate headlines but all within the same topic one article or many? I decided to take a non-exclusionary approach, and left out nothing from my search other than those exceptions already mentioned above.

I coded each article as either ‘replicated online’, which meant that the article was present online with no, or very minor (such as formatting) changes; ‘replicated online with changes’, which meant that the article was present online but with omissions, additions etc.; or ‘not replicated online’, which simply meant that the article was not present online at all.

6.4.3 Findings

Of the 710 articles investigated, 10% were replicated online, 8% were replicated online with changes, and 82% were not replicated online. This shows an overwhelming trend for print articles in magazines not to be replicated online.

Figure 6a and table 6a show the percentages of replication in each magazine, and figure 6b and table 6b show the percentages of replication by publisher.

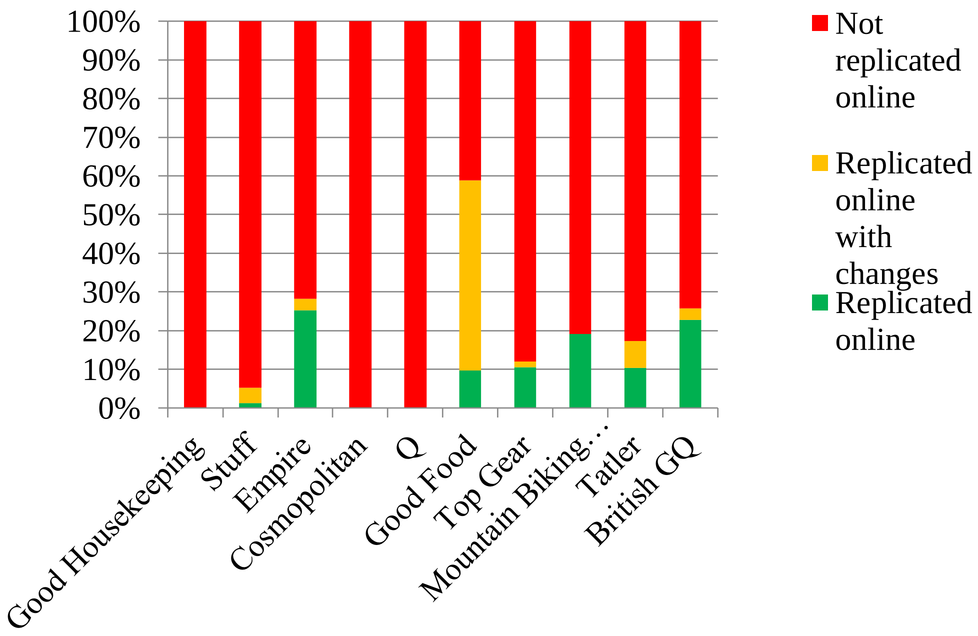


Figure 6a: Replication of print articles online in each magazine (%).

Table 6a: Replication of print articles online in each magazine (%).

	Replicated online (%)	Replicated online with changes (%)	Not replicated online (%)
Good Housekeeping	0	0	100
Stuff	1	4	95
Empire	25	3	72
Cosmopolitan	0	0	100
Q	0	0	100
Good Food	10	49	41
Top Gear	10	2	88
Mountain Biking UK	19	0	81
Tatler	10	7	83
British GQ	23	3	74
Total replication	**10**	**8**	**82**

Note: The 'Total replication' row does not display the totals of the columns in the table, but rather the total % of replication (or lack of) for *all* magazines.

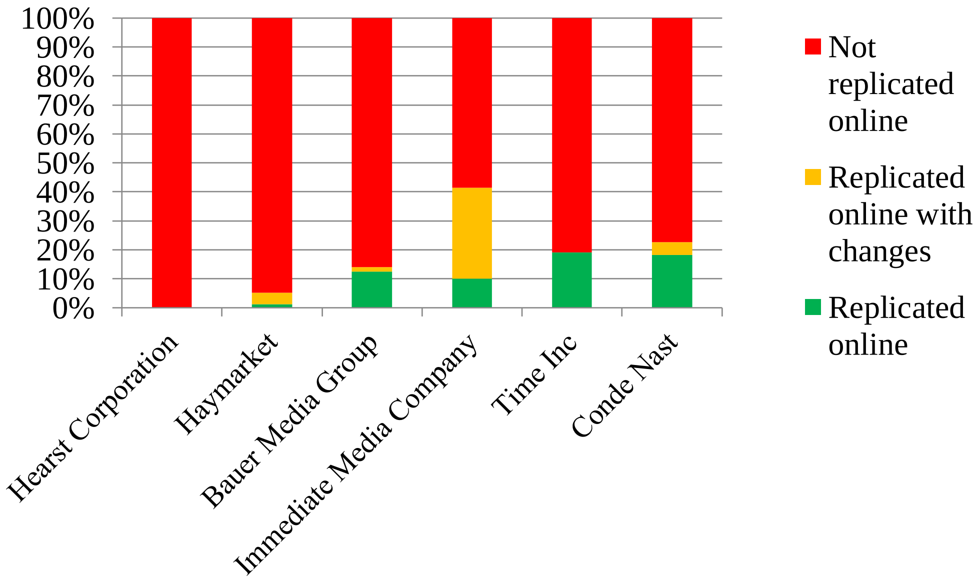


Figure 6b: Replication of print articles online for each publisher (%).

Table 6b: Replication of print articles online for each publisher.

	Replicated online (%)	Replicated online with changes (%)	Not replicated online (%)
Hearst Corporation	0	0	100
Haymarket	1	4	95
Bauer Media Group	13	1	86
Immediate Media Company	10	31	58
Time Inc	19	0	81
Conde Nast	18	5	77
Total replication	**10**	**8**	**82**

Note: The 'Total replication' row does not display the totals of the columns in the table, but rather the total % of replication (or lack of) for *all* publishers.

6.4.4 Discussion

The findings in section 6.4.3 clearly suggest that the majority of print magazine articles are not replicated online. Figure and table 6a show that the amount of replication does not seem to be related to the type of magazine; among the four magazines from the 'Lifestyle' and 'Men's lifestyle' categories (Good Housekeeping, Cosmopolitan, British GQ, and Tatler) the percentage of replication in the sample examined (including replication with changes) varies from 0% (Good Housekeeping and Cosmopolitan) to 26% of articles within each magazine (British GQ). It might be suspected that different publishers have different rules about whether or not they replicate articles online. I attempted to contact all of the publishers represented in figure 6b to ascertain whether they had regulations regarding what they published online, but received no response. Figure 6b suggests that the Hearst Corporation may have a blanket rule of not reproducing any of their print content online. However, these can only be speculations, as my sample size was not big enough to draw firm conclusions about this issue of online replication. With that said, though, the ten magazines investigated were chosen at random and are all popular, widely-circulating magazines. The general pattern that I found is thus likely to be generalisable to popular British magazines as a genre, and so I assume that, in a majority of cases, articles published in print will not be replicated online.

It might be queried whether it is the case that, despite not being the same articles, online magazines are written in the same style as print magazines. If a similar style of writing is used both in print and online, that could alleviate the need to be concerned about using online magazines despite their dissimilarity in terms of exact content to print magazines. However, it is beyond the scope of this study to investigate this. My general impression after becoming very familiar with these websites and magazines during this study, is that this varies between each magazine, and would thus need to be addressed separately for each publication.

6.4.5 Decisions regarding the collection of magazines

Despite the findings of this study, I continued with the plan to collect magazine texts for the Written BNC2014 from online magazines. This is because it would simply be too time consuming and expensive to purchase print copies of all of the required magazines and then convert them to digital text. Additionally, contacting publishers for access to and their permission to include magazines in the corpus did not seem like a feasible option as I received no responses to my query discussed in section 6.4.4. However, the findings of this investigation have helped to guide my collection to ensure that I am collecting the most ‘print-like’ online articles.

To do this I could consider eliminating magazines which are published by companies which this study indicated do not replicate their articles online. This would mean removing all magazines published by the Hearst Corporation, thus losing 14 magazines as potential sources of data. As I cannot be certain about specific publishing companies' regulations, and as total removal of certain companies' titles would result in losing multiple potential data sources, this solution was not desirable.

Another option would be to conduct a preliminary investigation of each magazine before collecting data from their website. This would involve purchasing a print copy and performing a similar study to this one to assess how similar the online and print content is. This would almost certainly be too time consuming, and so was not a practical solution.

It would seem, then, the best solution will be to devise a set of criteria which will allow selection of those online articles which most resemble print articles. This is an admittedly heuristic, but practical, approach. During this study I became familiar with the style of articles which appear both in print and online, and so was able to devise criteria to select articles in a systematic manner. The criteria I devised are:

- Articles must be over 400 words long
- Articles must not consist of mostly pictures
- Articles must not be in a 'slideshow' format
- Articles must not be in the form of a list
- Where possible, I will prioritise collection of 'feature' articles

Another point to consider is whether representing print magazines is important at all. Durrani (2015) notes that 'vogue.co.uk' has 2,217,678 unique users, with 2.3 million followers on Twitter and 2.5 million Facebook fans, whereas Vogue's combined print and digital circulation (digital circulation here means a digital copy of the print magazine, which is purchased by readers, not simply articles on their website) is just over 200,000. Thus, although some top-circulating magazines have seen an increase in circulation (Durrani, 2015), it seems that many more people consume magazine content through websites rather than traditional print magazines. This may suggest that I do not need to worry at all about including print magazine articles, or even ‘print-like’ online magazine articles, because texts on magazine websites will be more representative of what is read by people in Britain than are printed articles. This then suggests another consideration – should we be basing our collection on the most visited magazine websites, rather than top circulating print and digital magazines? And furthermore, should we consider including texts from websites which are ‘magazine-like’ but have no print counterpart, such as Buzzfeed.com? For the purposes of this project, the answer to both of these questions is ‘no’. I have been unable to find any comprehensive list of UK magazine websites along with visitor numbers, so it was not possible to pursue collection of data from the most visited magazine websites. Collecting data from websites such as Buzzfeed.com is impractical because it will be much harder to determine whether these types of content represent British English; online magazine websites usually have a ‘.co.uk’ site (or have some other explicit marker of the site being ‘British’, such as the website’s description) which contains the content which they view as British, but this is not the case for many other websites (including Buzzfeed.com).

6.4.6 Magazine collection in practice

In section 6.4.5 I outlined what I initially believed was the best method for collecting those online magazine articles which were most like those in print magazines. However, once collection began, it became clear that this method was not going to be workable. The method as outlined required all of the magazine texts to be collected by myself individually, so that I could look at them and assess their suitability against the listed criteria. This proved extremely time consuming. It quickly became clear that this would not be a workable method for the number of texts which I needed to collect.

Thus, the decision was made to automate collection of magazine texts. This means that the only criterion from section 6.4.5 which has remained usable for data collection is “Articles must be over 400 words long”. A script was written by MT which generated a list of all articles from the relevant websites, and also calculated the word count of the articles. This produced a list of possible articles which I then manually filtered to extract only those articles published in 2014 with word counts of over 400 words, and sent this list back to MT. MT then used a script to scrape the selected articles from the web pages, and clean them to remove adverts etc. After this process was completed it became clear that the initial word counts which were calculated were not accurate, due to the presence of adverts and html in the original texts. Thus, many of the articles collected were less than 400 words in length. This method, although problematic in some aspects, has proven to be the only method which would allow me to collect the data needed in the time frame which I had.

6.5 Composition of the periodicals medium of the corpus

As was the case with the books medium of the corpus, some changes had to be made to the periodicals medium of the corpus when compared to the sampling frame (see table 6c and appendix B for the periodicals medium of the sampling frame, and table 6d and appendix C for the eventual composition of the periodicals medium). The collection of journal articles, as shown in this chapter, was relatively straightforward, and as such, the amounts of these genres collected exactly matches the targets in the sampling frame. However, this was not the case for the newspaper genres or magazine genres.

The three newspaper super genres have undergone the greatest number of changes to their sampling frame of any super genre in the corpus. The total amount of data included from newspapers has increased, the genre labels have been altered, and the proportions of each genre in the corpus have changed greatly. However, it should be stressed that none of these changes impacts the representativeness of the super genres, and in some ways even increases their representativeness. Firstly, the amount of newspaper data included in the corpus has increased by 3%, from 21% in the sampling frame to 24% in final corpus. This was due to the redistribution of some words from the miscellaneous medium (see sections 8.6 and 8.8). The decision to redistribute the words in this way was largely taken because more newspaper data had been collected than was needed, and so adding words in these super genres would not require any extra data collection. Furthermore, letters have been excluded as a genre from the final corpus (see section 8.6), but are present in the form of 'letters to the editor' in some newspapers. Thus, redistributing some data here increases the representation of this genre, which was removed from the miscellaneous medium.

Secondly, many of the genre labels have been changed in the final corpus, compared to the sampling frame. At the super genre level, broadsheet national newspapers have been renamed as 'serious' newspapers, and tabloid newspapers have been renamed as 'mass market' newspapers. These changes are reflected in the genre labels. These changes were made because it is becoming increasingly difficult to determine whether a British newspaper is definitively a tabloid or a broadsheet. Strictly speaking, the terms 'tabloid' and 'broadsheet' referred historically to the size of a newspaper, with broadsheets being printed on bigger pages than tabloids. This is no longer the case; for example, in 2018 The Guardian newspaper (which has traditionally been a broadsheet) started being printed in tabloid format. Additionally, many people consume news online nowadays, in which case the size of a printed page is irrelevant. For this reason, I decided to rename these super genres to reflect the style of writing contained within them, rather than attempt to classify each newspaper as broadsheet or tabloid. Broadsheet newspapers are traditionally seen as providing quality journalism covering a wide range of serious topics in depth. Tabloid newspapers are traditionally viewed as being more widely accessible than broadsheets due to the more informal style of writing, and the less serious topics typically covered (celebrity gossip, for example). Thus, these newspaper categories will be referred to in the Written BNC2014 as ‘serious’ and ‘mass market’ respectively.

At the genre level, some of the labels have undergone further changes. The ‘commerce’ genre is now ‘commerce and business’, and the ‘social’ genre is now labelled ‘lifestyle’. These changes were made after I categorised the data which had been collected, and found that the original descriptors did not cover the full range of what was included in each category. Many of the articles which I categorised into the ‘commerce’ category were about businesses and business practices, rather than being specifically about commerce. Many of the articles which I categorised into the ‘social’ genre, did not seem to really be adequately described by the label ‘social’. For example, articles about travel, food, or fashion. Thus, I renamed this category as ‘lifestyle’, in order to better represent what is contained within the genre.

The final change to these genres is that they are now represented proportionally to their occurrence in the real world. When designing the corpus it was not possible to ascertain what the proportions within the populations of newspapers were, but collecting and then categorising texts made this possible. Texts were collected equally from each type of newspaper, and equally from each section of each newspaper, so the proportions found should be representative of the population of British newspapers. After categorising the texts it became clear that there were big imbalances between the different types of newspaper, and between the individual genres within each type of newspaper. I calculated word totals for each genre, and then included data in the corpus in amounts which reflected these proportions. This means that serious newspapers make up 9.84% of the corpus, broadsheet newspapers make up 7.88% of the corpus, and mass market newspapers make up 6.28% of the corpus. In all types of newspaper 'reportage' was the most common genre, but other genres varied. For example, in regional and mass market newspapers the second largest genre is 'sports', whilst for serious newspapers it is 'commerce and business'. Details of all differences can be seen in table 6d and appendix C.

Similarly, the collection of magazine articles actually ended up tending more towards proportional representation than equal representation. Once collection of all available data had been completed, it was clear that representing each genre equally would not be possible. Most significantly, it ended up being the case that *no* sports magazines were collected because none of the sports magazine websites could be scraped by MT. For the other genres, data was collected but the amount varied greatly. Thus, the proportions of these genres are now distributed according to the data which was available. The smallest genre is the 'food' genre which comprises just 0.06% of the corpus, and the largest is the 'science & technology' genre which comprises 1.56% of the corpus (see table 6d). However, despite these changes from the sampling frame, the magazine section of the corpus still consists of 8% of the corpus data, as originally planned.

Overall, the periodicals medium of the corpus has increased by 3%, from 35% in the sampling frame, to 38% in the eventual corpus. This was due to the redistribution of some data from the miscellaneous medium to the newspaper medium, as discussed. This slightly increased amount of data in this medium, alongside the shift towards proportional representation of the newspapers and magazines means that, despite some changes to the proportions of the genres in this medium, the periodicals medium is still highly representative of this type of written language.

Table 6c: The periodicals medium of the Written BNC2014 sampling frame.

Medium	**Super genre**	**Genre**	**Target**	**Words**
Periodicals (35%)	Academic Prose (journal articles)	W_ac_journal_humanities_arts	1%	900,000
		W_ac_journal_medicine	1%	900,000
		W_ac_journal_nat_science	1%	900,000
		W_ac_journal_polit_law_edu	1%	900,000
		W_ac_journal_soc_science	1%	900,000
		W_ac_journal_tech_engin	1%	900,000
	Broadsheet national newspapers	W_newsp_brdsht_nat_arts_ent	1%	900,000
		W_newsp_brdsht_nat_commerce	1%	900,000
		W_newsp_brdsht_nat_editorial	1%	900,000
		W_newsp_brdsht_nat_reportage	1%	900,000
		W_newsp_brdsht_nat_science	1%	900,000
		W_newsp_brdsht_nat_social	1%	900,000
		W_newsp_brdsht_nat_sports	1%	900,000
	Regional & local newspapers	W_newsp_other_arts_ent	1%	900,000
		W_newsp_other_commerce	1%	900,000
		W_newsp_other_editorial	1%	900,000
		W_newsp_other_reportage	1%	900,000
		W_newsp_other_science	1%	900,000
		W_newsp_other_social	1%	900,000
		W_newsp_other_sports	1%	900,000
	Tabloid newspapers	W_newsp_tabloid_arts_ent	1%	900,000
		W_newsp_tabloid_commerce	1%	900,000
		W_newsp_tabloid_editorial	1%	900,000
		W_newsp_tabloid_reportage	1%	900,000
		W_newsp_tabloid_science	1%	900,000
		W_newsp_tabloid_social	1%	900,000
		W_newsp_tabloid_sports	1%	900,000
	Magazines	W_magazines_lifestyle	1%	900,000
		W_magazines_mens_lifestyle	1%	900,000
		W_magazines_TV_film	1%	900,000
		W_magazines_motoring	1%	900,000
		W_magazines_food	1%	900,000
		W_magazines_music	1%	900,000
		W_magazines_science_tech	1%	900,000
		W_magazine_sports	1%	900,000

Table 6d: The eventual composition of the periodicals medium of the Written BNC2014.

Medium	Super Genre	Genre	Target	Words
Periodicals (38%)	Academic Prose (journal articles)	W_ac_journal_humanities_arts	1%	900,000
		W_ac_journal_medicine	1%	900,000
		W_ac_journal_nat_science	1%	900,000
		W_ac_journal_polit_law_edu	1%	900,000
		W_ac_journal_soc_science	1%	900,000
		W_ac_journal_tech_engin	1%	900,000
	Serious newspapers	W_newsp_serious_arts_ent	0.98%	885,600
		W_newsp_serious_commerce_business	1.97%	1,771,200
		W_newsp_serious_editorial	0.39%	354,240
		W_newsp_serious_reportage	3.74%	3,365,280
		W_newsp_serious_science	0.12%	106,272
		W_newsp_serious_lifestyle	1.14%	1,027,296
		W_newsp_serious_sports	1.50%	1,346,112
	Regional & local newspapers	W_newsp_regional_arts_ent	0.15%	142,560
		W_newsp_regional_commerce_business	0.45%	413,424
		W_newsp_regional_editorial	0.29%	263,736
		W_newsp_regional_reportage	4.68%	4,212,648
		W_newsp_regional_science	0.02%	21,384
		W_newsp_regional_lifestyle	0.24%	220,968
		W_newsp_regional_sports	2.05%	1,853,283
	Mass market newspapers	W_newsp_mass_market_arts_ent	0.18%	168,480
		W_newsp_mass_market_commerce_business	0.16%	146,016
		W_newsp_mass_market_editorial	0.25%	224,640
		W_newsp_mass_market_reportage	3.51%	3,161,808
		W_newsp_mass_market_science	0.01%	5,616
		W_newsp_mass_market_lifestyle	0.06%	56,160
		W_newsp_mass_market_sports	2.11%	1,853,280
	Magazines	W_magazines_lifestyle	1.55%	1,400,000
		W_magazines_mens_lifestyle	1.04%	940,000
		W_magazines_TV_film	0.67%	600,000
		W_magazines_motoring	1.55%	1,400,000
		W_magazines_food	0.06%	55,000
		W_magazines_music	1.55%	1,400,000
		W_magazines_science_tech	1.56%	1,405,000

Chapter 7: Collection of e-language for the Written BNC2014

7.1 Introduction

In this chapter I will discuss the rationale for, the design of, and the construction of the e-language (electronic language) section of the Written BNC2014. *E-language* (as used by Knight et al., 2014) or *computer mediated communication (CMC)* is a way of referring to the language used in online spaces; some examples include email, SMS, blogs, tweets, and discussion forums. Knight et al. (2014: 30) simply define e-language as “language communicated through any digital device”, however, it is important to make clear exactly what the definition of e-language, as used in this thesis, is, because the boundaries of this type of language can vary. Whilst almost any genre of writing can be found online, many of these genres are also present in offline spaces (e.g. news articles, recipes, short stories etc.). These types of texts do *not* fall within the definition of e-language used in this thesis. E-language, for the purposes of this project, encompasses texts which are *unique* to an online environment. To make this distinction clear throughout this chapter, I will refer to e-language which is unique to online spaces as type-A e-language, and the broad definition of e-language as type-B e-language. The choice to only include type-A e-language in the e-language medium of the Written BNC2014 was motivated by the fact that type-B e-language (e.g. news articles) will be present in all mediums of the Written BNC2014 but not classified as e-language. As chapters 5 and 6 have shown, some of the data for both the books and periodicals mediums of the corpus has been collected from online sources, but clearly not categorised as e-language. Thus, the e-language medium is reserved for language which can *only* be found online (i.e. type-A e-language).

The Written BNC1994 contains very little type-A e-language[6] (only emails from a Leeds United email list) for the simple reason that e-language was a very marginal part of language when the texts included in the Written BNC1994 were collected. However, this has changed completely since the creation of the Written BNC1994 and to not include a diverse range of e-language in the Written BNC2014 would be to ignore a very important part of contemporary British English. In 2018 86% of British adults accessed the internet every day (Office for National Statistics, 2018), which suggests that an extremely large amount of the British public's reading and writing is being done online. It has been found that e-language has its own unique set of features, such as vocal spellings and emoticons (Riordan and Kreuz, 2010), which set it apart from other written language. For these reasons the Written BNC2014 must seek to fully represent this type of written British English. This decision is similar to that of the creators of the ANC (Reppen and Ide, 2004; see section 2.5) who also included e-language in their corpus as an update to the BNC1994 sampling frame. On the other hand, this decision is in contrast to the creators of COCA's decision to not include e-language in their corpus (Davies, 2009; see section 2.6). E-language was not included in COCA for 2 reasons. Firstly, the corpus was designed to be a monitor corpus containing balanced data from each year since 1990. The creators strongly felt that it would not be possible to collect enough e-language for the earlier years in the corpus to keep the corpus balanced for genres, and balanced for each year. Secondly, the creators noted the difficulty of ascertaining precisely who is producing e-language. As the corpus aims to represent American

[6] The Written BNC1994 also contains very little type-B e-language because, at the time the corpus was compiled, very few texts were digitised. However, the corpus certainly does contain many texts which *nowadays* could be classified as type-B e-language.

English, this was a problem. This was also a problem in creating the e-language section of the Written BNC2014, and will be discussed in section 7.4.2-7.4.8.

In order to design the e-language section of the Written BNC2014 I first need to consider exactly what I am aiming to represent. In order to do this I first consider the composition of the World Wide Web, and decide which elements of this composition I will need to reflect in the e-language section. It has also been informative to look at previous corpora of e-language in order to understand what has been found to work and what has been found to limit the utility of these corpora. Once the genres of e-language to be included in the corpus have been determined, I decide in what proportions these different genres of e-language will be present.

I begin in section 7.2 by providing an overview of literature on web registers, specifically focusing on the work of Biber et al. (2015). I will then discuss previous corpora of e-language, moving from specific to general corpora. In section 7.3 I consider the very important legal and ethical considerations which must be addressed in the creation of an e-language corpus. Finally, in section 7.4 I discuss the design, collection, and composition of the e-language medium of the Written BNC2014.

7.2 Literature review

7.2.1 Introduction

In this section I will first discuss literature on the identification of web registers (section 7.2.2). It was important, in the creation of the e-language section of the Written BNC2014, to represent the full range of type-A e-language genres on the World Wide Web, and to consider what proportions the different e-language genres should be present in. Biber et al. (2015) provide a recent and comprehensive ‘taxonomy’ of web registers which has proved useful in the design of this section of the corpus. In section 7.2.3 I discuss previous corpora of e-language; firstly I discuss specific corpora which contain only one genre of e-language, and then some general corpora of e-language, with a focus on CANELC (Knight et al. 2014), the most recent corpus of British English e-language. I finish in section 7.2.4 by outlining how this literature has influenced the design of the e-language section of the Written BNC2014.

7.2.2 Web registers

There have been many studies of register in various contexts (Biber and Conrad, 2009: Appendix A). However, until recently little of this research has focused on identifying registers on the web. Biber et al. (2015) note that, although over 3 billion people worldwide use the internet (Internet World Stats, 2014), surprisingly little is known about the actual composition of the World Wide Web. This can be problematic for researchers using a web-as-corpus approach (Jakubíček et al., 2013; Ferraresi et al., 2008; Davies, 2013a) because they do not know what registers of texts their corpora contain, nor whether the texts contained within their corpora reflect the actual composition of the Web (Biber et al., 2015: 12). Furthermore, it means that for anything observed using such data, researchers cannot be sure whether what they are observing is actually tied to a single genre or sub-set of genres within the whole data set. Similarly, using the whole corpus may give rise to an averaged view of language which, when considered at the genre level, may be wholly misleading, i.e. the behaviour observed at the genre level generates a set of frequencies which are nowhere near the mean, for example. Biber et al. attempt to address these problems by categorising web pages according to register in order to find out the composition of the web. This type of research has previously been attempted by researchers using Automatic Genre Identification (AGI) techniques - computational methods which aim to automatically classify web texts by genre (see Santini, 2007; Rehm, 2002;

Stamatatos et al., 2000) - but with limited success (Biber et al., 2015). Santini and Sharoff (2009: 131-3) point out that much AGI research has relied on small corpora whose representativeness of the web as a whole is unknown, and thus the accuracy rates reported in different studies using different corpora are not comparable. Biber et al. (2015) also question the reliability of the standard technique employed by AGI researchers of having a single ‘expert’ code the documents under question as this assumes that this ‘expert’ will correctly identify the genre of all documents to a high degree of accuracy. This approach also assumes that the ontology to be applied is meaningful and agreed upon. As Biber et al. (2015) point out, this kind of consensus about genres on the web does not exist, hence the whole exercise can become relatively subjective as there is no expert view of what the web consists of in practice.

In order to address the problems of AGI research, Biber et al. (2015) use a large and representative corpus of web documents in their study, and have them coded by a large group of lay-users of the internet (each document being coded by four different users). The corpus was deemed to be representative because it was “obtained through random sampling from across the full range of documents that are publically available on the web” (Biber et al., 2015: 16). To create the corpus Biber et al. (2015) used 48,571 documents from the ‘General’ component of the Corpus of Global Web-based English (GloWbE) (Davies, 2013a). GloWbE contains approximately 1.8 million web documents, compiled by performing Google searches for frequent English 3-grams (Davies, 2013a). Lay-users of the web (who trained by watching a short video) were then asked to code the documents according to a decision tree which was designed based on the situational framework developed in Biber and Conrad (2009: Chapter 2). The findings of this are shown in table 7a. These general register categories were then broken down into many more sub-categories.

This study provides a reliable indication of the composition of the web, as it uses a large and representative corpus and the coding of documents was done manually by at least four coders for each document. However, despite the rigorous methods used for coding, inter-rater reliability was still relatively low, with all four coders agreeing on the general register category of documents in only 36.9% of instances, and all four coders agreeing on the specific sub-register of documents in only 24.2% of instances (Biber et al., 2015). Biber et al. (2015) suggest that this lack of agreement between coders is because of the presence of ‘hybrid’ registers, where documents on the web are actually combinations of several registers; some examples of frequent register combinations are ‘Narrative’ and ‘Informational Description/Explanation’, and ‘Narrative’ and ‘Opinion’. As well as these ‘hybrid’ registers Biber et al. (2015) also found that many of the documents in the corpus contained user comments; ranging from 7.2% of the documents in the Informational Description/Explanation documents to 37.3% in the ‘Opinion’ documents. This does seem to suggest that the ontology used by Biber et al. (2015) may need revising in order to account for these ‘hybrid’ registers.

Furthermore, the general register categories found by Biber et al. (2015) do not seem to conform to the definition of *register* given in section 4.2.3.5 (“a category of texts which are recognised according to their situation of use”). The registers ‘Narrative’, ‘Informational Description/Explanation’, and ‘Opinion’ are far too general to be considered associated with a particular situation of use. However, once the lowest levels of classification are reached it is clear that these truly are registers in the sense defined in chapter 4. For example, opinion blogs, encyclopaedia articles, recipes, and short stories all occur in particular situational contexts and have particular linguistic features associated with them.

A factor which limits the usefulness of these findings specifically in relation to the e-language section of the Written BNC2014, is that many of the registers found actually represent type-B e-language, i.e. many of the documents which were coded in Biber et al. (2015) were documents taken from the web but which may also be found offline. For example, news reports, short stories, novels, encyclopaedia articles, research articles, recipes, and instructions are all types of text which are not specific to an online environment. As discussed in section 7.1, the e-language section of the Written BNC2014 will seek to represent type-A e-language (i.e. those genres which are unique to an online environment), thus, the *proportions* of the registers found by Biber et al. (2015) (see table 7a) are of limited relevance to the design of the e-language section of the Written BNC2014.

Despite the problems with this research outlined above, it was still important to utilise the findings of this study in the creation of the e-language section of the Written BNC2014. The work done by Biber et al. (2015) was certainly pioneering, and undoubtedly is the best guide available as to the composition of the web. The fact that the corpus used was large and representative means that most, if not all, possible web registers were analysed in the study. Thus, I can ensure that all important registers which are unique to the web are included in the e-language section of the Written BNC2014.

Table 7a: Frequency information for general register categories found on the web (Biber et al., 2015: 23).

General register	No. of documents	Percent
Narrative	15,171	31.2
Informational Description/Explanation	7,042	14.5
Opinion	5,452	11.2
Interactive Discussion	3,104	6.4
How-to/Instructional	1,126	2.3
Informational Persuasion	794	1.6
Lyrical	605	1.2
Spoken	325	0.7
Hybrid (see below)	14,197	29.2
No agreement	755	1.6
Total	48,571	100

7.2.3 E-Language corpora

People are increasingly doing a large amount of their daily reading and writing online, for example, Statista (2018b) estimate that there are over 425 million blogs on the platform Tumblr alone. Thus it has been suggested that corpus researchers should take account of 'e-language' when researching a language or creating a new corpus (Beißwenger et al., 2013). Many researchers (Beißwenger et al., 2013; Knight et al., 2014; Tagg, 2011) have acknowledged the need to include this relatively new genre of language in their corpus research and many corpora of e-language have been created. However, most of these corpora tend to be very specific (in that they only represent one genre of e-language), very small, or not available to all researchers (examples of corpora with each of these problems will be discussed in sections 7.2.3.1-7.2.3.5). Of course, it should be noted that whilst many of these corpora only contain one genre of e-language, many genres of e-language actually overlap with each other in terms of their features. For example, Twitter is a form of blogging known as ‘microblogging’, email lists can be a lot like forums in that they are discussions between multiple people about a specific topic, and forums can be considered similar to a blog with comments because in a forum users are often responding to an opening post, much like comments responding to a blog post. I will outline some of these specific corpora below, before moving on to describe some, potentially more useful for present concerns, general corpora of e-language. A thorough understanding of previous corpora of e-language was essential for designing and collecting the e-language section of the Written BNC2014.

7.2.3.1 Corpora of emails

The most common corpora of e-language are those which contain emails. For example, the ENRON corpus contains approximately 0.5 million emails from 150 individuals, and is freely available online (Klimt and Yang, 2004). However, the vast majority of the contributors are members of senior management at ENRON (Klimt and Yang, 2004) and as such the corpus is only representative of a very specific genre of e-language (emails between senior management in a work setting), and results drawn from such a corpus cannot be generalised to the language of emails as a whole. The majority of the 1.3 million word Mini-McCALL corpus consists of emails (alongside forum discussions and assignments) (Deutschman et al., 2009). The nearly 6000 email messages were all produced by university students, which again, although the corpus is of quite a large size, means that the corpus is only representative of a very specific type of email discussion (Deutschman et al., 2009). The Junk Email Corpus was constructed in 2002 and contains 673 junk emails (Orasan and Krishnamurthy, 2002). This corpus is very small and likewise only represents a very specific type of email. In 2016, Krieg-Holz et al. (2016: 2543) created, what they claim to be, “the largest email corpus ever built” – CODE ALLTAG. The corpus consists of two sections – a set of 1.5 million German language emails downloaded from an online archive, covering a range of topics, and a smaller set of less than 1000 emails which were donated by participants and have rich metadata associated with them. Some researchers, such as Riordan and Kreuz (2010), have used listservs to create small corpora of emails to be used perhaps just once for a specific piece of research. To investigate the use of cues in e-language Riordan and Kreuz (2010) created the AIR-L corpus which contains 5770 emails between the Association of Internet Researchers, the Chalkhills corpus which contains 391 emails discussing movies and music, and the Luckytown corpus which includes 1562 emails between Bruce Springsteen fans. The small amount of e-language included in the BNC1994 was also taken from a listserv for Leeds United fans.

Something which almost all of these email corpora have in common is that they contain emails on some very specific topic (for example, the Luckytown corpus), or which were produced in very specific settings (for example, the ENRON corpus). This greatly reduces their representativeness, and research using them cannot claim strongly to be generalisable to emails as a whole. In the construction of the Written BNC2014 (see section 7.4.5) it will be important to collect emails from as wide a range of contributors and settings as possible, and, in order to monitor this, detailed metadata will need to be collected.

7.2.3.2 Corpora of SMS messages

Another form of e-language which has been compiled into corpora is Short Message Service (SMS) messages, also commonly known as ‘text messages’. Researchers may be interested in SMS messages because they represent a distinct genre of language, or because of popular concerns that text messaging could be damaging to language (Crystal, 2008b: 77). Alternatively, Knight et al. (2014: 41) simply suggest that texting has "become a very central part of communication in modern life". Thus, corpora of SMS messages can be very useful to language researchers. CorTxt is a corpus of 11,067 SMS messages collected by Tagg (2009). This is a large collection of data, but its representativeness is limited because almost all of the SMS messages were contributed by Tagg's friends and family (Tagg, 2009). The HKU SMS Corpus contains 853 SMS messages, and was created as part of a project to research the features of mobile phone communication (Baron et al., 2012). Choudhury et al. (2007) use a corpus of 854 SMS messages downloaded from Treasuremytext (an online SMS archive service) in order to test a technique for converting SMS messages into normalised Standard English. Some researchers, such as How and Kan (2005), have created corpora of SMS messages with the aim of improving the predictive text functions on mobile phones. How and Kan (2005) created the NUS SMS Corpus, which contains 10,117 SMS messages from Singaporean university students. An issue with the NUS SMS Corpus and the HKU SMS Corpus, when considering them in relation to the Written BNC2014 project, is that they contain messages written largely by speakers of English dialects other than British English; thus, no firm conclusions can be drawn about the language of SMS messages as a whole or SMS messages written by speakers of *British* English. If SMS messages are included in the Written BNC2014 (see section 7.4.6) it will be necessary to collect demographic information from all contributors to ensure that the data is coming only from native speakers of British English.

7.2.3.3 Corpora of forum discussions

Forums are online spaces where users can have discussions, usually centring on a specific topic, in an asynchronous fashion (meaning that messages are not posted and read at the same time, and there may be long time gaps between messages being sent; Baron et al., 2012). Research into this type of asynchronous e-language is longstanding (see Parks and Floyd, 1996; Baym, 1995), but in this section I will focus on corpora created in the last 15 years, as the Written BNC2014 project is focusing on *contemporary* language (i.e. language produced in since 2010; see section 4.3.1). Thus, contemporary projects will be most useful in exploring how this section of the corpus could be designed and created. The Mini-McCALL corpus mentioned in section 7.2.3.1 contains 462,890 words of forum data, accompanied by rich metadata for all of the contributors (Deutschman et al., 2009). However, the posts are all written by non-native speakers of English so its usefulness in examining British English is limited (Baron et al., 2012). Yahoo! Answers is a forum where users can ask questions about any topic and other users post answers. Yahoo! have created a database of 4,483,032 questions and their answers, and this data is available for use by academics upon request (Baron et al., 2012). Usenet, whilst not strictly a forum (Usenet predates forums; Baron et al., 2012), can be a valuable source of data for corpora of online discussions, such as those found in forums. A large corpus of over 30 billion words of English Usenet postings has been created and is freely available online; however, there is no metadata available for the individual contributors so the usefulness of the corpus is reduced (Baron et al., 2012). Hoffmann (2007) created a very large (over 150 million words) corpus of Usenet postings from 12 different newsgroups in his research. Furthermore, Hardaker (2012) created 2 corpora of Usenet postings with a combined size of over 86 million words in order to investigate negatively marked online behaviours such as trolling.

7.2.3.4 Other Specific Corpora

Other genres of e-language have been used to create specific corpora. One such genre is blogs; the Blog Authorship Corpus contains 681,288 blog entries contributed by 19,320 different bloggers across three different age groups, and is freely available online (Schler et al., 2006). Corpora of online chats or instant messaging have also been constructed; for instance, the NPS Chat corpus contains 10,567 messages in English from online chat rooms (Forsyth and Martell, 2007). Finally, microblogging data from Twitter has been compiled into corpora; Twitter_Smallcorp is a 2 million word corpus of tweets (Puschman, 2009) and Horn et al. (2011) created a 16 million word corpus of tweets.

7.2.3.5 General e-language corpora

Although the specific corpora discussed in sections 7.2.3.1-7.2.3.4 may have proven useful in researching specific genres of e-language, in order to research e-language in general, a corpus containing a variety of genres of e-language is needed. I will discuss some such corpora in this section.

The World Wide Web Consortium Corpus (W3C) (Craswell, 2005) is a corpus of over 200,000 files gathered from a ‘crawl’ of the World Wide Web Consortium’s sites (Riordan and Kreuz, 2010). The W3C contains emails, web pages, and texts in a variety of formats (.pdf, .ppt and.doc) (Riordan and Kreuz, 2010). EnTenTen (Jakubíček et al., 2013) is a corpus created from a crawl of the web, and contains over 10 billion words of internet texts. Another corpus created from crawling the web is ukWaC (Ferraresi et al., 2008; see section 1.3.4.3), which contains over 2 billion words. GloWbE (Davies, 2013a) is another large corpus of internet texts (1.9 billion words), which was created by Google searching high frequency n-grams to generate a random assortment of web pages. Despite the fact that all of these corpora are very large and contain a variety of internet texts, they will not actually be very useful for the design of the e-language section of the Written BNC2014. This is because they contain many texts which will not be considered e-language in the Written BNC2014, i.e. type-B e-language. This is one of the reasons why, despite the fact that it can yield huge amounts of data, web crawling will not be an option for data collection in the Written BNC2014 (see section 1.3.4.3 for a discussion of the other reasons).

A general corpus of e-language with similar goals to the one intended to form part of the Written BNC2014 is the *German Reference Corpus of Internet-based Communication* (DeRiK) (Beißwenger et al., 2013). DeRiK is a corpus of German e-language which will form part of a general reference corpus of the German language, much like the e-language section of the Written BNC2014. Unlike the Written BNC2014, DeRiK is a monitor corpus, meaning that the creators aim to add more data to the corpus over time (Beißwenger et al., 2013). The composition of the data added to DeRiK may change each time new data is added, as the creators plan to use the results of an annual survey of German internet usage to determine what types of e-language have been most popularly written and read each year. They will then base the composition of the corpus on this survey (Beißwenger et al., 2013). The results of the survey will provide an ideal framework for the creators of the corpus to strive for, but this will inevitably be changed due to the availability of data and any difficulties gaining permission to use certain types of data (Beißwenger et al., 2013). The first set of data in DeRiK contains “mostly discourse from Wikipedia talk pages, a selection of forum and weblog discussions, chat conversations, and postings of selected Twitter users who have published their tweets under a Creative Commons license" (Beißwenger et al., 2013: 533).

The most recent general corpus of British English e-language is the 1 million word Cambridge and Nottingham e-language Corpus (CANELC) (Knight et al., 2014). The composition of CANELC is shown in table 7b; it is implied that the inclusion of these five genres of e-language in the corpus is justified by their importance, both in terms of their use by people and in terms of their use as data in previous research (see Knight et al., 2014: 38-41 for extended discussion of how frequently people write and read these genres, and in what ways they have previously been researched). The data for the Twitter, blogs, and discussion boards sections of the corpus were selected from 'popular' Twitter accounts or websites, to ensure that the corpus would reflect both what is commonly written and also what is commonly read (Knight et al., 2014: 35-36). For example, only tweets from Twitter accounts with over 1000 followers were included in the corpus. However, with the rise of 'bots' online (an automated account online which can post, follow people, 'like' posts etc.) this measure of popularity may not be reliable. Another factor which affected what data was used in CANELC was ease of gaining permission to use the data. The CANELC creators sought permission from all authors whose writing was included, which meant that only data from sources where it would be easy to contact the author were considered (Knight et al., 2014: 36). For example, only blogs managed by one individual whose email address was easy to find were considered for inclusion in the corpus (Knight et al., 2014: 36). Thus, CANELC is arguably as balanced as was possible at the time of collection for the demographics of contributors and the topics covered in the texts. However, a perfect balance could not be achieved due to issues of availability and gaining permissions. Unfortunately, CANELC is not freely available for other researchers to use. This being the case, the creation of a well-balanced and representative e-language section in the Written BNC2014 will be invaluable, not only to researchers wishing to study British English as a whole, but also to those wishing to study specifically British English e-language.

Table 7b: The composition of CANELC (Knight et al., 2014: 35).

Data type	Number of contributors	Number of messages/ entries	Word count	
			Raw	Percent
Twitter	30	18,972	259,101	26
Blogs	36	1,101	267,983	27
Discussion boards	12	2,715	242,727	24
E-mails	Various	1,920	128,951	13
SMS	11	5,215	101,913	10
		29,923	1,000,675	100

7.2.4 Conclusion

In this section I have reviewed previous corpora of e-language, and have also discussed research which has aimed to help researchers using a web-as-corpus approach by trying to identify exactly what the composition of the World Wide Web is. The corpus of e-language which has been most influential for the design of the e-language section of the Written BNC2014 is CANELC (Knight et al., 2014) as it is large (one million words), it contains a variety of genres of e-language, and the researchers discuss at length their justifications for the decisions they made when creating the corpus. I have combined elements of the CANELC design with the findings of Biber et al. (2015) in order to ensure that the e-language section of the

Written BNC2014 contains samples of those genres of e-language which are most commonly found on the World Wide Web. I will discuss this further in section 7.4.

7.3 Copyright and ethical considerations

E-language presents a unique challenge in negotiating copyright law, because this type of language is often much more private and personal than a book or periodical. This means that legal concerns regarding gaining permission to use the work are combined with ethical considerations. In this section I will discuss the various views on the matter of whether permission needs to be gained from authors to use their e-language in the Written BNC2014.

7.3.1 Does permission need to be gained?

Legally, any original work which is posted online is protected by UK Copyright law (see section 1.5) and permission needs to be sought to reproduce this material. However, Herring (1996) suggests that it is unreasonable to assume that everything written on the internet is copyrightable, for example, the one word message “Hi” in a discussion forum could be considered trivial and unoriginal. It has also been indicated by Professor Christopher May (see section 1.5.3) that anything which is posted publicly online and which is not behind a paywall can be considered to be in the public domain (meaning that the text is no longer protected by copyright), and thus can be used by anyone. Also, as discussed in section 1.5, provided that I stay within the bounds of fair dealing, it will not be necessary to ask for permission in any case by utilising the ‘Non-commercial research’ exception to UK copyright law. In the creation of the ANC, Ide (2008) reports that web-data was not used to ensure that the creators were compliant with U.S. copyright law. However, this has been the biggest obstacle in the creation of the ANC, and has ultimately resulted in the project being stalled, and much smaller than was originally planned (see section 2.5). It is also important to consider the General Data Protection Regulation (GDPR) which has come into force across the EU since May 2018, with the aim of allowing EU citizens to better control their personal data. Very broadly, this means that any organisation which is using members of the public's personal data must do so "lawfully, fairly and in a transparent manner" (EUR-Lex, 2016: article 5). GDPR would seem to apply to the data which we would be collecting for the Written BNC2014, and as such means that we will have to be very careful to store data securely. Article 89 of the GDPR legislation seems particularly relevant to present purposes: "Processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, shall be subject to appropriate safeguards, in accordance with this Regulation, for the rights and freedoms of the data subject […] Those measures may include pseudonymisation". It seems then that anonymising all data will be very important, in order to comply with GDPR (discussed further in section 7.3.1). However, other than any identifying details present in the text which I collect (which will be anonymised), I will not be keeping records of any other personal information about authors of online texts. I am not interested in recording their location, gender, contact details etc., and so it seems that it should be fairly easy to comply with GDPR.

Gaining permission from all potential e-language contributors to use their texts would be extremely time consuming and potentially very difficult. Knight et al. (2014) sought permission from all potential contributors to CANELC (see section 7.2.3.5 for a full discussion of this corpus). During a pilot phase thirty prospective individuals were contacted with a request for permission to use their data; of these thirty only twelve responded, and only seven gave full permission to use their data (Knight et al., 2014). This response rate of less than 50% suggests that gaining permission from all potential contributors will simply not be practical given the time constraints of the project, and the fact that the e-language section of the Written BNC2014 will need hundreds of contributors. Knight et al. (2014) limited their potential contributors to only those with easily identifiable contact details online. However, comments have been collected for the Written BNC2014 (see section 7.4.8), which are not usually accompanied by user contact details. Thus, it will be practically impossible to collect permission from all contributors. It has been suggested by Koene et al. (2015) that in cases where all participants in online linguistic research cannot be contacted, it is good practice to contact a sample of the population to gain an idea of how the population would feel about their data being used – this suggestion will be addressed in section 7.3.3.

In addition, it may not actually be *desirable* to only use posts which I have permission to include as this will result in bias. Certain types of people, those who do not mind being associated with the texts which they have posted, are more likely to respond positively to a request to use their posts in the corpus. This means that a large proportion of, for example, aggressive, racist, or otherwise negative posts may be missed out of the corpus, as the people who post them, so-called ‘trolls’, are unlikely to be willing to be associated with them.

Ultimately, it seems that for legal and practical reasons permission will not have to be sought to include things which are posted publicly online in the Written BNC2014, providing that the terms of fair dealing are met.

7.3.2 Ethical considerations

Although there will be no legal need to gain permission from authors, there are ethical issues which must also be considered. Research which collects data from people (rather than, for example, scientific research which collects its data from inanimate objects) typically needs to be presented to, and approved by, a research ethics board before actual collection begins, as the research involves human beings. How collection of e-language for a corpus would be treated in such a context remains unclear. Harriman and Patel (2014) point out that there are currently no universal guidelines for how to handle the ethics of internet-based research. They contacted NHS ethics committees and found that these committees had no specific guidelines for dealing with internet based research, and had very little experience dealing with it (Harriman and Patel, 2014). The AOIR (2012: 7) recommendations suggest that it may be impossible to come up with a universal set of guidelines because "[t]he uniqueness and almost endless range of specific situations defy attempts to universalize experience or define in advance what might constitute harmful research practice". They instead propose that ethical questions should be taken into consideration at all stages of a research project and addressed as and when they arise. Following this approach I will outline some of the main ethical concerns involved in the collection of data for the e-language section of the Written BNC2014 here.

The first ethical consideration is the extent to which people consider their online posts to be private. Despite the public nature of online posts, most recommendations for ethical online research seem to assume that individuals view the things which they post online as fairly private, and only expect them to be read by certain people (AOIR, 2012 (Association of Internet Researchers); Herring; 1996; UoB Guidelines (University of Brighton ethics committee guidelines); BAAL (British

Association for Applied Linguistics)). AOIR (2012) and BAAL state that consideration must be given to an individual's expectations of what may be done with the text they post online. However, a study by Moreno et al. (2012) found that many older adolescents either felt 'fine' or 'neutral' about the idea of researchers using online information about them in their research. Where adolescents reported feeling negative, it seemed usually to be because they did not realise that this information was already publicly accessible online (Moreno et al., 2012). So not only may some people view their online posts as relatively private, despite knowing that they are public in nature, but also many people may simply be unaware that their online posts are public. However, it has been ruled by courts in the USA that "a person should have no reasonable expectation of privacy in writings that are posted on a social networking Web site and made available to the public" (Moreno et al., 2012: 440).

A further ethical issue is ensuring that vulnerable groups are protected. There are places on the internet which vulnerable people view as safe and private places (even if it is not the case that these spaces are truly private); research into these spaces could be viewed as an unwelcome intrusion (UoB Guidelines). Such an intrusion could harm the individuals involved; a group may be abandoned altogether if it is no longer viewed as a safe place by participants (UoB Guidelines). Sharkey et al. (2011) conducted research involving an online discussion group where young people who self-harmed could talk to each other. Their research was initially denied ethical approval on the basis of inadequate consideration of working with vulnerable groups (Sharkey et al., 2011). The UoB Guidelines also point out the possibility of a vulnerable person becoming *unwittingly* involved in research; for example, a child using a discussion forum intended for adults on which research is being carried out. This suggests that consent may need to be gained from all participants individually in order to ensure that they are not vulnerable. However, these considerations are mostly based on research with a high level of intrusiveness, which is in contrast to the very low level of intrusiveness involved in collecting publicly available online data to be used in a corpus. Thus, whilst protecting vulnerable groups is still important, it is highly unlikely that any harm would come to participants in the collection of data for the Written BNC2014, and so gaining permission may not be as important as guidelines for other types of research suggest (UoB Guidelines).

The outcomes of Sharkey et al.'s (2011) study mainly highlight the importance of the anonymisation of vulnerable groups; this is the third ethical consideration for internet based research. It is desirable to anonymise participants in research, especially vulnerable ones, so that they cannot be identified and are protected from potential harm. However, this can be very time consuming when using e-language data as a participant may be identifiable from references to places, other online spaces, or other people within the text. Furthermore, depending on how the data will eventually be presented, it may be possible for people to use a URL included in the metadata, or web-search engines to find the original post (Herring, 1996; AOIR, 2012). To deal with the first point in the creation of CANELC, Knight et al. (2014) anonymised all names, usernames, references to locations, and contact details. A similar approach was also taken in the creation of the Spoken BNC2014 (Love et al., 2017a). When transcribing the data for this corpus, names of people, locations, institutions, telephone numbers, and addresses were anonymised, but names of famous people or general locations were not. However, doing this can be "potentially detrimental" (Knight et al., 2014: 43) to the data, as some references are essential for understanding the meaning of a text. Also, anonymisation of data has much less practical impact in a spoken corpus because the anonymisation can take place at the same time as the transcription. For a written corpus the time needed to read and anonymise all of the texts would be a huge investment. It is possible that some of the anonymisation of texts could be automated, however, as the anonymisation needs to be *perfect*, manual checking of the anonymisation would be required which would, again, be time consuming. It is easy to assume that automatic methods are the way forward if they have accuracy rates of, for example, over 90%. However, where 100% accuracy is needed, and where the location of the errors is uncertain, an accuracy rate of over 90% is useless.

Another difficulty in anonymising e-language data is that it conflicts with the legal requirement to fully acknowledge authors (see section 1.5). Some interpretations of 'fair dealing' view the acknowledgment of authors as necessary (BAAL). This is obviously not compatible with full anonymisation. Herring (1996) suggests that in order to overcome this issue, it must be considered whether e-language is more like speech or writing (see table 7c; see section 1.2 for some of the difficulties of distinguishing between speech and writing).

Table 7c: Speech-like and writing-like properties of e-language.

Speech-like properties of e-language	**Writing-like properties of e-language**
Can 'overhear' something online which was not intended for you (Herring, 1996).	Can use anything you read as long as you cite it fully (Herring, 1996).
Informal (SMS, emails between friends, comments, discussion forums, Tweets from personal accounts)	Formal (blog posts, work emails, Tweets from institutions/businesses)
Multiple participants (SMS, email, comments, discussion forum, Tweets)	Single participant (blogs, Tweets)

Herring (1996) suggests that e-language is intermediate between speech and writing, and thus no firm decision can be reached about a universal best practice for anonymisation. However, this may not be the case anymore (note that Herring wrote over 20 years ago). From Gregory's (1967; see section 1.2) perspective, e-language would be considered written language as it is delivered graphically, and there is no intention of it being read aloud. As I have shown in section 7.2 there are many different genres of e-language, all with unique properties which may make them more speech-like or more writing-like (see table 7c). Nevertheless, the point remains that it is impossible to decide whether e-language is more like speech or writing, and so no universal guidelines for anonymisation can be agreed upon. However, if consent were to be gained from all participants then it would be possible to ask each individual whether they would prefer to be acknowledged or anonymised.

7.3.3 Public perceptions of research using e-language data investigation

7.3.3.1 Rationale and methodology

Although the majority of ethical guidelines are based on the assumption that people may view their online writing as very private and place a high value on anonymity (as outlined in section 7.3.2), there has been relatively little research to find out whether people's actual opinions match these assumptions. What research of this kind there has been may also suggest that people find using online data more acceptable than many ethical guidelines assume (see discussion of Moreno et al., 2012 in section 7.3.2). For this reason, I decided to conduct a small-scale questionnaire study to get an impression of: a) how aware people are of the legal situation regarding researchers using their public online writing and how they feel about this; b) how willing people are to allow their online writing to be used for research; and c) how people would feel if their public online writing were used in the creation of a corpus without their knowledge.

A copy of the questionnaire can be found in appendix K. The questionnaire was distributed both online and on paper, resulting in 71 responses overall. The demographic balance of the respondents was not ideal, as the majority of responses came from females aged 21-30. Furthermore, no under 18s are represented in the survey results, as this would have required getting parental consent to participate, which would have been too time consuming for the scope of this small survey. Many of the responses may also have come from linguists (so some repsones may have been more favourable than otherwise) as the questionnaire was distributed on my own and the BNC2014's Twitter accounts, which are largelt followed by linguists. However, for the purpose of getting an impression of people's attitudes towards their online writing being used in research, these responses are sufficient.

7.3.3.2 Results

Before beginning analysis of the results, I calculated an 'internet usage score' for all respondents. This was done by awarding points for answers to questions 5-9 (which enquired about the frequency with which participants posted online): an answer of 'never' scored 0 points, whilst an answer of 'daily' scored 5 points for questions 5, 7, 8 and 9 (with intermediate answers scored accordingly), and an answer of 'no' scored 0 points whilst 'yes' scored 2 points for question 6. Figure 7a shows the internet usage scores for all participants.

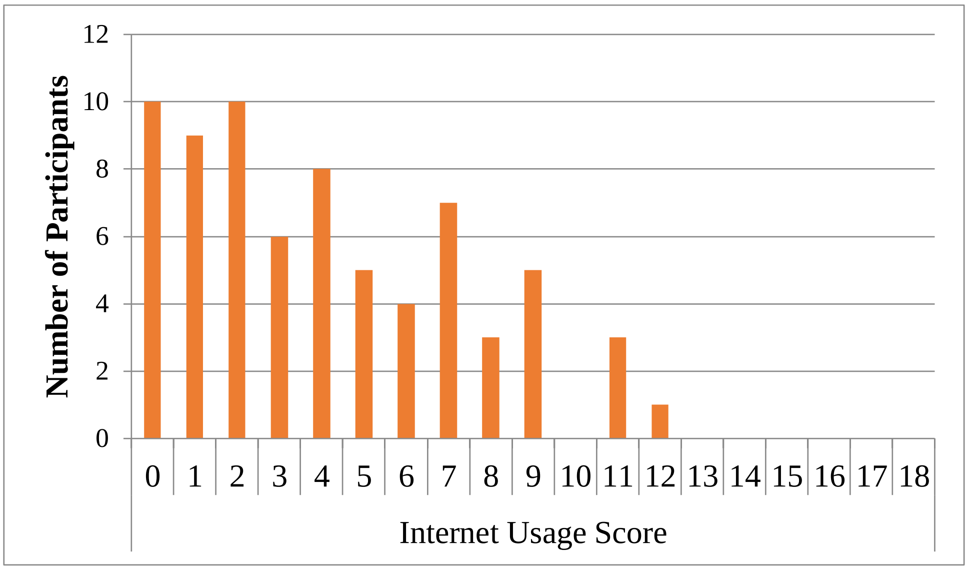


Figure 7a: A graph showing the internet usage scores for all participants.

In the first section of the questionnaire (after demographic information) I asked participants to read the statement: "*Currently, anything which you post publicly online (i.e. those things mentioned in questions 5-9) can be used by researchers without seeking your permission and without any obligation to anonymise you*". I then asked participants whether they were aware that this was the case, and what their response to this was. 60% of the respondents indicated that they were aware that this was the case. This suggests that a small majority of people are aware of how their public online posts can be used. Awareness appears to correlate with level of education; those who had currently achieved less than an undergraduate degree were the only group where more people were unaware than aware.

Participants were then asked to give their opinion on the situation in the form of an open-answer question. Following the procedure of Moreno et al. (2012) I then coded participant's answers as indicating that they felt 'fine', 'neutral' or 'concerned' about the situation. Results can be seen in table 7d.

Table 7d: Classification of responses to the legal situation outlined in the questionnaire.

	Awareness: YES	**Awareness: NO**	**Total**	**%**
Fine	21	3	**24**	**36%**
Neutral	9	6	**15**	**23%**
Concerned	11	16	**27**	**41%**

Table 7d shows that a similar number of respondents felt concerned about the situation as felt fine about it. Looking at the cross-tabulation with awareness, table 7d shows a clear division: those who were aware of the situation are more likely to be ‘fine’ with it, whereas those who were unaware are more likely to be ‘concerned’. A possible explanation for this is that the surprise of learning the situation may have made people feel more concerned; this is similar to Moreno et al.’s (2012) finding that those who were most concerned that they had been identified for research online were concerned precisely because they weren’t aware that they were identifiable. Perhaps if people had had more time to think about this new information then they may not have been as concerned as they were upon learning it.

The second section of the questionnaire aimed to find out how open participants would be to allowing their online writing to be used in research if they were asked for permission. Participants could choose as many as they liked from five options: ‘Yes’, ‘Yes, if I was anonymised’, ‘Some posts’, ‘It would depend on the research’ and ‘No’. Results are shown in table 7e.

Table 7e: Responses when asked whether participants would allow their posts to be used in research.

Response	**Total**	**% (of total respondents)**
Yes	20	28
Yes, if I was anonymised	38	54
Some posts	15	21
It would depend on the research	34	48
No	5	7

Note: the percentage in column 3 is calculated based on how many people out of the total amount of respondents selected that answer, not based on the proportion of all responses. This is because respondents could select multiple answers.

Table 7e shows that the majority of participants placed some conditions on whether they would allow their data to be used for research; the most important conditions being anonymisation and the purpose of the research. Encouragingly, only 5 participants said ‘No’, and one of these ‘no’ responses was qualified with ‘It would depend on the research’. All 4 unqualified ‘no’ responses were from people who scored 0 for internet usage, that is, people for whom the question is irrelevant anyway. Of the 20 ‘Yes’ responses, only 12 were unqualified. However, it is encouraging that half of these unqualified positive responses were from participants with some of the highest internet usage scores, suggesting that some people who may be targeted for data collection would be willing to contribute with no conditions.

The final part of the questionnaire aimed to assess how participants would feel if their online writing was used without their knowledge in the creation of a corpus. 59% of people indicated that they would be happy with this, 25% indicated that they would not be happy with this, and 15% indicated that they were unsure. These results are particularly reassuring in light of the finding that people feel that they would need to consider the purpose of the research being carried out before agreeing to participate, as it suggests that corpus creation is something which they would be happy to contribute to. This is backed up by comments from participants on this section of the questionnaire such as “Because I'm aware that creating a corpus can be a very useful tool to research language and I'd be happy to contribute to this information” and “Used for a beneficial purpose”. However, due to the fact that the questionnaire was distributed via my own and the BNC2014 project’s social media accounts, a large proportion of the responses may have come from linguists, which could have caused answers to be more favourable than if respondents had been better balanced.

7.3.3.3 Summary of questionnaire results

The results of this questionnaire suggest that, whilst people perhaps do not view their online posts as being as private as ethical guidelines would suggest, many would still rather their online writing was not used without their permission. Anonymisation is very important to most people, as is the purpose of the research that their data is being used for. However, it was encouraging to find that the majority of respondents felt that creating a corpus was worthwhile research and something which they would be happy to contribute to.

7.3.4 Copyright and ethics summary

This section has shown that there is no legal need to gain permission from authors to use their e-language in the Written BNC2014; however, gaining permission may be beneficial for ethical reasons. Ethically, gaining permission is desirable because it allows authors who view their e-language as private to deny permission; it allows for the protection of vulnerable groups as they can deny permission and it becomes possible to identify any vulnerable authors who are unexpected in the context; and it allows each author to decide whether being acknowledged or anonymised is more important to them. However, the findings of Knight et al. (2014)

suggest that gaining permission may be extremely difficult and can be very time consuming. Also, for some of the types of data included in the e-language section of the corpus, it will be impossible to contact the authors (for example, people leaving comments online very rarely have easily identifiable contact details associated with their account). This fact in turn suggests that much of the data collected for this section of the corpus is, in effect, pre-anonymised, and represents minimal ethical concerns from the perspective of anonymisation (discussed in section 7.3.2). As an intermediate solution, some potential participants have been contacted via a questionnaire to see how they would feel about their texts being included in a corpus (as suggested by Koene et al., 2015). The results of this, admittedly small-scale, survey suggested that most people feel that corpus creation is worthwhile research which they would be happy to contribute to. Furthermore, it is doubtful whether all of the ethical considerations discussed will truly apply in the creation of the Written BNC2014 because the gathering of texts will be extremely unintrusive and is unlikely to affect the authors. For these reasons, authors will not be contacted for their permission to include their data in the Written BNC2014 and will not be anonymised, with the caveat that I will only take posts which are openly available online – i.e. are not behind paywalls or login screens.

7.4 Composition and collection of the e-language medium of the Written BNC2014

In light of the research discussed in section 7.2, I designed the e-language section of the Written BNC2014 sampling frame to contain data from the following genres of e-language: Twitter, blogs, discussion boards, emails, SMS messages, reviews, and comments. The justification for the inclusion of these genres in the sampling frame will be discussed in sections 7.4.2 – 7.4.8, along with a discussion of how these genres were collected. Firstly, though, it is important to make clear in what proportions these genres are present in the sampling frame, and the reasoning behind this. This section will focus on the design of the e-language section of the Written BNC2014 sampling frame, the eventual composition of the e-language medium of the corpus can be seen in appendix C, and will be discussed in section 7.5.

7.4.1 Proportions

The e-language medium of the Written BNC2014 sampling frame contains 9 million words (10% of the total corpus; the eventual composition of the e-language medium of the corpus can be seen in appendix C and is discussed in section 7.5). The decision to include this amount of e-language in the corpus was reached after consideration of both the desire to fully represent this very important type of contemporary British English, and also the desire to not allow this type of language to overwhelm the corpus or to result in too great a reduction of the other mediums in the corpus. 10% of the corpus being e-language meant that for most genres no more than a few percent needed to be taken from any of the genres included in the Written BNC1994 (see chapter 4), but a corpus of 9 million words is certainly enough to, as far as is possible (see chapter 3), represent British e-language (and is 9 times the size of CANELC; Knight et al., 2014).

All of the genres included in the e-language medium of the Written BNC2014 sampling frame can be described as being either ‘broadcast’ (written with the intention of any person reading them) or ‘directed’ (written with the intention of specific individuals reading them), and as being restricted or unrestricted in their length (see table 7f).

Table 7f: Genres of e-language categorised according to length and broadcast/directed distinctions.

	Broadcast	**Directed**
Restricted Length	Twitter	SMS/IM
Unrestricted Length	Blogs, Discussion forums, Reviews, Comments	Emails

As illustrated in table 7f, most of the genres of e-language are broadcast texts of unrestricted length (accounting for 4 of the 7 different genres). For this reason, broadcast texts of an unrestricted length account for approximately 50% of the e-language medium in the sampling frame, with the remaining 50% being split between the other 3 genres of e-language. This composition was chosen because, as discussed in sections 7.4.2 – 7.4.8, it is necessary to represent all sections of table 7f, but also to represent all of the genres of e-language in proportions great enough to make the data useful in their own right (see section 4.2.3). Thus, the broadcast texts of unrestricted length have been allocated a greater proportion so that each of the four genres within this section will be present in a large enough proportion to make a useful contribution to the e-language section. As such, the smallest genres in the sampling frame are the individual blog genres with 180,000 words each. However, when combined, blogs comprise a total of 1.08 million words – i.e. plenty to be useful as an object of study in their own right (see section 4.2.3). The composition of the e-language medium of the Written BNC2014 sampling frame is shown in table 7g.

Table 7g: The composition of the e-language medium of the Written BNC2014 sampling frame.

Genre	Words (tokens)	% of the Written BNC2014
W_e_tweet	1,620,000	1.8
W_e_blog_news	180,000	0.2
W_e_blog_sport	180,000	0.2
W_e_blog_opinion	180,000	0.2
W_e_blog_personal	180,000	0.2
W_e_blog_informational	180,000	0.2
W_e_blog_travel	180,000	0.2
W_e_discussion_forum	1,170,000	1.3
W_e_email_prof	720,000	0.8
W_e_email_personal	720,000	0.8
W_e_SMS	1,530,000	1.7
W_e_review	1,080,000	1.2
W_e_comment	1,080,000	1.2
Total	9,000,000	10

7.4.2 Twitter

At the time of writing, eMarketer (2014) predicts that 17.1 million people in the UK will use Twitter in 2018 to update their friends and followers on their lives, thoughts and feelings. Researchers are fast realising what an important area of language ‘tweets’ have become, and as such much research is being carried out on the unique properties of the language of Twitter. Tweets are also being utilised as a resource for researching already established linguistic phenomena (see for example, Zappavigna, 2012; Cunha et al., 2014; Reyes et al., 2013; Herdadelen, 2013). It is important that the Written BNC2014 reflects this extremely prevalent and prominent form of language and provides data for researchers wishing to continue this recent trend of using Twitter data for research.

Tweets have been collected in two ways: through ‘public participation in scientific research’ (PPSR; see Shirk et al., 2012), and via other corpus projects. Firstly, the project team publicised, mainly via Twitter but also at conferences etc., the chance for the public to get involved with the project by submitting their Twitter archives to us (this is similar to the approach taken by the creators of the ANC, see section 2.5). This could be done very simply via an online form. The only restrictions were that the contributor had to be a native speaker of British English. It was made clear to contributors that their usernames would be fully anonymised before the tweets were included in the corpus, but that they were also welcome to anonymise any other identifying information that they wished before submission. All contributors, if they wished to be, are credited in the corpus documentation. In accordance with the date range policy set out in section 4.2.1, tweets written between 2014-2018 will be included in the corpus.

Secondly, the corpus has benefited from a generous collaboration with Simaki et al. (2017) who donated their corpus of tweets and Facebook statuses which were written by UK celebrities on their public social media profiles. This data was collected by Simaki et al. (2017) as part of a project which aimed to automatically detect what variety (US, UK, AUS, CAN, or NNS) of English each author uses. Thus, all of the data was carefully annotated for the native variety of the author, and as such we could be sure that the tweets and Facebook statuses donated to the Written BNC2014 were written by native speakers of British English. The data represents the online writing of 117 British celebrities in 2015, and thus fits within the date range for data in the Written BNC2014. The fact that this data represents British celebrities also satisfies the need to represent both production and reception of language by British English speakers. By including data written by celebrities, the corpus is representing a type of language which many British people write but also tweets which many people have likely read. It should be noted that this dataset does not only include tweets, but also Facebook statuses, which were not included in the sampling frame for the Written

BNC2014 (see appendix B). However, these are a very popular type of online language (Statista, 2018a, estimates that Facebook has 39.2 million UK users in 2018). The only reason Facebook statuses were not included in the sampling frame was because they are a private form of data which we would have to seek permission to access and include in the corpus (as was the case for emails and SMS and IM messages, sees sections 7.4.5 and 7.4.6). I knew that this would be very time consuming and possibly not very productive, and as Twitter was a publicly accessible form of microblog (a type of blog where users typically only post a sentence or two, such as Twitter and Facebook), Facebook data was not included in the sampling frame. However, rather than turn down data which had been kindly donated to us, I decided to expand this genre to include the Facebook statuses which were donated as part of this corpus. As a result of this, the name of this genre has been changed from W_e_tweet to W_e_microblog (this can be seen in the comparison in section 7.5, and in appendix C).

7.4.3 Blogs

Blogging rose to prominence in 1999 (Myers, 2010: 10), and in 2013 it was estimated that there were well over 152,000,000 blogs on the internet (WPVirtuoso, 2013). This number continues to grow, and, as of April 2018, there are over 425 million blogs on the popular blogging platform Tumblr.com alone (Statista, 2018b). Not only do millions of people write blog posts, but it is also estimated that most people read a blog at least once a day (WPVirtuoso, 2013). Thus, blogs represent an important area of both writing and reading in modern British English.

Biber et al. (2015) confirm that blogs are an important genre of language on the web, with blogs accounting for 30% of the web as a whole. However, Biber et al.’s

(2015) classification includes news and sports reports as the same as news and sports blogs. News and sports reports would not be considered e-language for the purposes of the Written BNC2014 (because they are Type-B e-language, as discussed in section 7.1) so the actual proportions of blogs on the web may be slightly lower when considered from this perspective. However, this does not diminish the fact that blogs are clearly a prevalent genre of e-language. Biber et al. (2015) found that across the web there were six different types of blogs (news, sport, opinion, personal narrative, informational, and travel), although these types are not particularly detailed and it seems likely that they could be broken down even further if desired. All of the types of blogs identified by Biber et al. (2015) are included in the Written BNC2014. Knight et al. (2014) choose to include blogs in CANELC (see section 7.2.3.5) in order to contribute to the continuation of research into the language of blogs. In CANELC blogs were chosen based on whether they occurred in a directory of popular blogs; sourced by Google searching “top ten blogs”, “popular blogs” etc. (Knight et al., 2014). However, although CANELC does include blog data covering a range of topics, no indication is given as to how balanced the corpus is across these topic areas (Knight et al., 2014).

It might be suggested that blogs should not be considered as a separate genre in linguistic research because you could read similar content in a magazine article or in books. However, Myers (2010) highlights some factors of blogs which make them different to magazines or books, and thus an important genre of language to study. Myers (2010) points out that blogs increasingly have an influence over people’s political decisions and their social and economic lives. If this is indeed the case, then the study of the language of blogs is important because “[t]he persuaded have to know what the persuaders are doing” (Myers, 2010: 3). Myers (2010) also suggests that looking at a medium as it emerges can help us to think better about other already established forms of media. Finally, Myers (2010) highlights the fact that blogs use the full scope of the web; they incorporate links, pictures, and videos, and have a potentially international reach, making them very different from the traditional forms of written language.

The blogs which are included in the Written BNC2014 were all collected manually using a similar procedure to Knight et al. (2014). To ensure that reception criteria were considered, the collection of 'popular' blogs was prioritised by Google searching, for example, 'popular travel blogs' or 'top UK travel blogs'. Blog posts which were published in 2014 or 2015 were then copied from the selected blogs and saved as .txt files, along with metadata (URL, title of blog, publication date, genre). Usually, no more than 10,000 words were collected from any one blog in order to avoid one author skewing the data, and also to ensure that I was staying within the bounds of fair dealing (see section 1.5.2). In practice, it was often difficult to distinguish between the different genres of blogs. For example, some blog posts could be classed as both personal and travel. In such cases, classification decisions were made based on the main content of the post and also the 'theme' of the blog itself. Furthermore, as predicted by Biber et al.'s (2015) findings, distinguishing between news reports and news blogs was problematic. Ultimately the distinction was made based on whether a post was published on a professional news outlets website (in which case the post was not considered) or on a personal website (in which case the post was considered for inclusion).

Another problem with the collection of blog data, and with many other types of data in the corpus (see discussions of this issue in chapters 5, 6, and 8), is ensuring that *British* English is being collected. Bloggers occasionally provide biographical details about themselves on their blogs, and in such cases these were utilised to ascertain whether the writer was a native speaker of British English. Of course, these biographies may not be accurate, but in the absence of contacting authors directly, it seems sensible to assume that they are correct. Additionally, where biographies were not given, only blog posts from websites with .co.uk, .ac.uk, or .org.uk URL endings were collected. Of course, this does not ensure that an author writing on these sites is a native speaker of British English, but it is the most practical way of having *some* degree of control over this factor.

7.4.4 Discussion forums

Unlike Twitter and blogs, where discussion is a possibility which may or may not be used, discussion forums are designed for the specific purpose of interaction between participants; they are spaces which are usually centred around a specific topic where users can respond to a message by voicing their own opinions and commenting on those of others. Discussion forums are a prominent genre of language on the web; Biber et al. (2015) find ‘Interactive Discussion’ to be a distinct register of language on the World Wide Web, with forums accounting for almost 90% of that register (Biber et al., 2015: 30). There has been much research into the language of discussion forums (for example, Moreno, 2011; Argan et al., 2011; Buil et al., 2012). However, until the creation of CANELC a corpus containing data from a wide variety of discussion forums covering a wide range of topics did not exist (Knight et al., 2014). I have continued the work of CANELC by creating a wide ranging and representative corpus of discussion forum data for inclusion in the e-language section of the Written BNC2014.

Discussion forum data was collected in a similar way to the blog posts. The process was done manually, by searching for ‘popular UK discussion forums’ and collecting from those websites which appeared at the top of results. No information was available about the *genres* of forums found on the web (as was the case for blogs) and so collection could not be stratified. However, I tried to ensure balance between various genres of discussion forum was achieved in collection. As with the blog collection, only discussion forums with a .uk domain name or which were otherwise explicitly marked as British were included in collection.

I collected discussion threads where the original post was published in 2014 or 2015, and prioritised the collection of ‘popular’ threads by collecting those threads which had the most responses. Where number of responses was not possible to determine, a random number generator was used to randomly select a thread from the lists available on the website. Threads were collected by copying the contents of a thread into a .txt file and preserving metadata (website name, URL of thread, date of first posting, title of thread). The cleaning process for discussion forums was lengthy – the metadata for each reply had to be removed (time of posting, name of poster etc.), and, where possible, repetitions were removed (it is common in forums to repeat the post which you are replying to at the beginning of your message). Removing these repetitions was not always possible as they were not always explicitly signalled via layout or a generic statement. In these cases the repetitions were: a) very difficult to spot when scrolling through, and so some were certainly missed, and b) even when I did spot them, the time needed to scroll back through the forum text and confirm that the text was indeed a repetition was too great to make this feasible. For this reason automatic procedures were also utilised to identify areas of repetition within the forum data, and remove these.

7.4.5 Emails

Unlike tweets, blogs, or discussion forums, emails are directed at a specific recipient or recipients rather than to the general public (with the exception of spam emails). Emails are a popular form of communication, perhaps because they are unrestricted in length, can be sent from any laptop, tablet, or phone with an internet connection, and can range in formality from messages between close friends to messages in a formal business setting. As outlined in section 7.2.3.1, research on the language of emails is longstanding; however most previous corpora of emails have only focused on particular genres of email. For example, the ENRON corpus (Klimt and Yang, 2004) and CANELC (Knight et al., 2014: 41) both contain mostly emails from a business setting. In the creation of the Written BNC2014, I have collected emails from a range of settings and with varying levels of formality.

As with the collection of tweets, and similarly to Krieg-Holz et al.'s (2016) approach to creating part of CODE-ALLTAG (see section 7.2.3.1), and to the ANC creators approach to accessing data (see section 2.5) a public participation in scientific research (see Shirk et al., 2012) approach was taken to the collection of emails. Alongside the call for the public to donate their tweets, the project team also asked people to donate their *sent* emails to the project. It was important that the emails were *sent* by the contributor and not *received*, as we would then have needed the permission of the original sender of the email to include the text in the corpus. As mentioned, emails are a private form of communication and, as such, are not covered by the exceptions discussed in section 1.5.2. Contributors could submit as many emails as they liked, provided they were sent in 2014-2018, and that they were a native speaker of British English. Participants were invited to anonymise their emails before sending them to remove any personal information that they wished.

The Written BNC2014 sampling frame (see appendix B) called for the emails to be categorised into *personal* and *professional* emails. However, it became clear once this categorisation was attempted that this would not be possible. There is simply too much overlap between personal and professional emails to reliably categorise them as one or another. For this reason, the two email genres were condensed into one (see section 7.5 for a full discussion of this). Spam and advertising emails were also collected, and included as a separate ‘advert’ genre of email. These emails were considered British as long as they were sent by a British company, and, as they were advertisements which were intended to reach as many people as possible, I considered them to be collectible under the copyright exceptions discussed in section 1.5 without any requirement to contact authors (see section 7.5 for a full discussion of this alteration to the sampling frame).

7.4.6 SMS and IM

Like emails, IM and SMS messages are directed at a specific recipient or recipients, but they are typically more informal than emails, and are often of a restricted length. An estimated 21 billion SMS messages were sent every day worldwide in 2014 (Deloitte, 2014), making SMS an important way in which we communicate with one another. Numerous researchers have investigated the language of text messages (Grinter and Eldridge, 2003; Faulkner and Culwin, 2005; Tagg, 2011) and as such this popular form of language was included in the Written BNC2014 sampling frame.

However, research suggests that SMS messages may be falling in popularity as IM (instant messaging) becomes more popular. Deloitte estimated that in 2014 mobile IM services (such as Snapchat and Whatsapp) were used more than twice as much as

SMS messaging, with an estimated 50 billion mobile IM messages sent everyday worldwide in 2014 (Deloitte, 2014). These figures were considered in the compilation of this section of the Written BNC2014 and, as such, IM messages are included in the sampling frame as well as traditional text messages.

SMS and IMs were collected in the same way as emails: a call was put out to the public to submit their data. This data could be in the form of text messages, Facebook messenger chats, or WhatsApp chats sent between 2014 and 2018. As before, participants were invited to anonymise their data before submission. Before the submission of these texts, contributors were required to indicate that they had gained permission from *all* authors within the conversation to submit their data.

7.4.7 Reviews

Reviews are a form of e-language not widely researched by linguists or often included in e-language corpora, although often used by NLP researchers for sentiment analysis (Liu, 2010; Burns et al., 2011). Biber et al. (2015) found that reviews accounted for 2.4% of the composition of the World Wide Web. Whilst this may seem like a very small proportion, it is actually a much larger proportion when only the registers which would be considered 'e-language' for the Written BNC2014 are considered (i.e. type-A e-language; see section 7.1 for further discussion of this issue). From this perspective, reviews actually account for a significant proportion of the composition of the web. Thus, they are included in the e-language section of the Written BNC2014.

Reviews were collected from Amazon.co.uk - a prominent and popular UK review site. This site was selected both for its popularity, but also because it has a .co.uk domain name. As it will not be possible to ascertain the native language of any of the authors of these reviews, taking reviews from UK sites seems to be the only feasible way of having *some* degree of control over the 'Britishness' of the reviews. Reviews were collected manually by IVD.

7.4.8 Comments

Another feature of the web found by Biber et al. (2015) which has been largely ignored in previous research is comments. Biber et al. (2015) find that a large proportion of the web pages they used in their analysis contain reader comments, ranging from 7.2% of the 'Informational Explanation/Description' pages to 37.3% of the 'Opinion' pages. This shows that a common way in which people use language on the internet is to voice their opinions or to respond to those of others via commenting on things which they have read or watched. This is distinct from the discussion forum register as comments are a response to something which has been read or watched, whereas messages in discussion forums are comments on a particular topic which users go to the forum specifically to discuss. Biber et al.'s (2015) research shows that comments are an important part of e-language, and so they are included in the Written BNC2014.

It is important to note here that I will be considering comments as separate entities to the posts on which they occur. This is different to the way in which many web-as-corpus researchers consider internet texts. For example, in the creation of ukWaC (Ferraresi et al., 2008) and EnTenTen (Jakubíček et al., 2013) the boilerplate is removed from the selected web page and then all of the text which is present on the page is considered together. However, I feel that a better approach is to consider each web page as logically designed and to follow the logical structure and sections of the page. In this way comments are very separate, and represent a different genre of e-language, to the posts which they accompany.

Comments were collected automatically by MT. A script was written to collect comments from various newspaper and magazine articles. These were often articles which had been collected without comments for inclusion in the periodicals medium of the corpus.

7.5 Composition of the e-language medium of the Written BNC2014

As was seen in the collection of the books and periodicals mediums of the corpus (see sections 5.5 and 6.5) achieving an exact match with the corpus sampling frame (see appendix B) when data collection is complete is highly unlikely. This is a pattern which is also true of the e-language medium of the Written BNC2014. This medium was planned to make-up 10% of the corpus, but in reality, due to changes made to the genres within the medium, this medium actually contains 11% of the data in the corpus. Various changes were made to this medium when compared to the sampling frame, and I will discuss each of these here. Table 7h shows the e-language medium in the corpus sampling frame, and table 7i shows the eventual composition of the e-language medium in the Written BNC2014.

The first change in this medium is the renaming of the ‘W_e_tweet’ genre to ‘W_e_microblog’. This is due to the inclusion of Facebook data in this genre, as discussed in section 7.4.2. The inclusion of this data necessitated a new name for this genre, so that it was not misleading as to the contents of the genre. Both tweets and Facebook statuses are types of microblogging, hence the new name for the genre. Another genre which has been renamed is the ‘W_e_SMS_IM’ genre, which is now labelled ‘W_e_IM’. This was due to the lack of SMS data which was submitted to the project, resulting in this genre being comprised entirely of IM data, and no SMS messages. The final genre renaming occurred with the 'W_e_email_prof' and 'W_e_email_personal' genres. It became apparent when trying to classify emails into these categories that this was not a workable distinction in practice. In reality, the majority of emails fell somewhere in between these two categories, for example, an email about a work topic but to a person the writer is friendly with outside of work, and thus containing personal questions or anecdotes. Thus, these categories were combined to create the 'W_e_email_prof_personal' genre. The proportions of these genres were combined also, so no data was lost.

Another change to this medium has already been discussed in section 7.4.5 – the addition of the 'W_e_email_advert' genre to this medium. This genre was created to include spam and other advertising emails sent by British companies. This genre contains 900,000 words of data (1% of the corpus); this amount was settled on because the genre was created from the movement of the 'W_advert' genre from the miscellaneous medium to the e-language medium (in the form of advertising emails). This genre was planned to comprise 1% of the corpus in the miscellaneous section, and so this data was simply moved to the e-language medium as this new genre. This is where the extra 1% of data comes from when comparing the e-language medium in the sampling frame and reality.

The final change to this medium occurs in the proportions of the various blog genres. Originally, each genre of blog was planned to contain 180,000 words (0.2% of the corpus), however this was not possible in reality. As predicted by Biber et al. (2015), news blogs were difficult to distinguish from news reports, and thus it was difficult to identify news blogs for the corpus. This meant that slightly less data was collected for this genre – only 90,000 words (or 0.1% of the corpus). This deficit was redistributed amongst the other blog genres, so that they each contain 198,000 words (or 0.22% of the corpus). This ensured that, despite the difficulty of collecting news blogs, the genre of blogs overall remained the same size in reality as in the sampling frame.

Overall then, the e-language medium of the Written BNC2014 has changed somewhat in reality when compared to the plans laid out in the sampling frame. In most cases, this was simply through renaming genres to better suit the data which was collected. However, some genres changed in size, and certainly the biggest change is the addition of a whole new genre in the form of advertising emails. The impact of this addition on the miscellaneous medium of the corpus will be discussed in section 8.8.

Table 7h: The e-language medium of the Written BNC2014 sampling frame.

Medium	**Super Genre**	**Genre**	**Target**	**Words**
E-language (10%)	E-language	W_e_tweet	1.8%	1,620,000
		W_e_blog_news	0.2%	180,000
		W_e_blog_sport	0.2%	180,000
		W_e_blog_opinion	0.2%	180,000
		W_e_blog_personal	0.2%	180,000
		W_e_blog_informational	0.2%	180,000
		W_e_blog_travel	0.2%	180,000
		W_e_discussion_forum	1.3%	1,170,000
		W_e_email_prof	0.8%	720,000
		W_e_email_personal	0.8%	720,000
		W_e_SMS_IM	1.7%	1,530,000
		W_e_review	1.2%	1,080,000
		W_e_comment	1.2%	1,080,000

Table 7i: The eventual composition of the e-language medium of the Written BNC2014.

Medium	**Super Genre**	**Genre**	**Target**	**Words**
E-language (11%)	E-language	W_e_microblog	1.8%	1,620,000
		W_e_blog_news	0.1%	90,000
		W_e_blog_sport	0.22%	198,000
		W_e_blog_opinion	0.22%	198,000
		W_e_blog_personal	0.22%	198,000
		W_e_blog_informational	0.22%	198,000
		W_e_blog_travel	0.22%	198,000
		W_e_discussion_forum	1.3%	1,170,000
		W_e_email_prof_personal	1.6%	1,440,000
		W_e_email_advert	1%	900,000
		W_e_IM	1.7%	1,530,000
		W_e_review	1.2%	1,080,000
		W_e_comment	1.2%	1,080,000

Chapter 8: Collection of miscellaneous and written-to-be-spoken genres for the Written BNC2014

8.1 Design of the miscellaneous and written-to-be-spoken mediums

This chapter will discuss the rationale for, design of, and collection of the miscellaneous and written-to-be-spoken mediums of the Written BNC2014. The majority of the discussion will focus on the miscellaneous and written-to-be-spoken mediums of the Written BNC2014 *sampling frame* (see appendix B), rather than the actual composition of the finished corpus (see appendix C). A comparison of the sampling frame will be presented in section 8.8, and can be seen by comparing appendices B and C. As already mentioned, the corpus is not complete at the time of submitting this thesis, and as such, any numbers quoted in this chapter are subject to change. The miscellaneous medium of the sampling frame contains ten genres (school essays, university essays, personal letters, professional letters, admin, advert, commerce, institutional, instructional, and religion), all of which would be difficult to categorise into another medium, hence they are all classified as *miscellaneous* genres (these genres will be defined in sections 8.2-8.7). Furthermore, the majority of these genres were not classified under a ‘super-genre’ in Lee’s genre classification scheme for the BNC1994 (see section 3.4.3) further implying their classification here as miscellaneous. The written-to-be-spoken medium is a small one, comprised of just two genres (news scripts and drama scripts), and so is considered in this chapter, alongside the miscellaneous medium, rather than given its own chapter.

The various miscellaneous and written-to-be-spoken genres were included in the sampling frame for two main reasons, despite there being good arguments for some of the genres *not* being included in the corpus. For example, when the sampling frame was originally conceived neither letters nor written-to-be-spoken texts were to be included in the corpus. For letters this was because I felt that the genre of emails had largely replaced letter writing, and that it would be extremely difficult to obtain people's private letters (even harder than obtaining people's private emails, see section 7.4.5). Written-to-be-spoken texts were not originally planned to be included in the corpus because of the blurred boundaries between speech and writing which these text types represent (see section 1.2 for a discussion of this). Of course, scripts are written, but are very often intended to appear to be like speech and are ultimately spoken in much the same way as a conversation. On the other hand, they are created in written form, and would be collected as pieces of writing for the corpus, making them fall within the 'written' medium according to the discussion in section 1.2. The first reason that these genres *were* ultimately included in the corpus, was that almost all of the experts who were consulted in the design of the sampling frame (see section 3.4.1) felt very strongly that all of the genres which were included in the Written BNC1994 should also be included in the Written BNC2014, in order to aid the comparability of the two corpora. The second reason was that, although these types of language are more marginal than some genres included in the corpus, they are of course types of written British English which do occur and are read and written by many. Thus, including these types of language in the corpus both aids the primary aim of making the corpus as representative of written British English as is possible, by increasing the genres of language which are represented, and also aids the secondary aim of making the corpus as comparable to the Written BNC1994 as is possible (see section 3.4.2).

The proportion of miscellaneous texts overall in the Written BNC1994 and the Written BNC2014 was planned to be very similar (see section 8.8 for discussion of how this changed in reality), with the Written BNC1994 corpus and the Written

BNC2014 sampling frame being comprised of 9.33% and 10% miscellaneous texts respectively (for the Written BNC1994 this figure is calculated based on only those genres which are directly comparable to those included in the Written BNC2014 sampling frame). The proportion of the written-to-be-spoken medium has increased in the Written BNC2014 sampling frame to 4% of the corpus compared to 1.47% of the Written BNC1994. This increase was due to the desire to make this medium useful as an object of study in its own right (see section 4.2.3).

The proportions of the individual genres within these two mediums in the sampling frame is split evenly, with the corpus containing 900,000 words of each miscellaneous genre and 1.8 million words of each written-to-be-spoken genre. This is in contrast to the proportions in the Written BNC1994 where university essays account for just 56,273 words of the corpus whereas commerce texts account for 3,807,342 words of the corpus. Of course, it is important to remember that this is not a criticism of the Written BNC1994; these genre categories were applied to the corpus *after* the corpus had been created, and so there could not have been any conscious effort to represent these genres at all, let alone in equal proportions or otherwise. The equal size of each individual genre within the mediums was chosen because, as with other genres in the sampling frame (see sections 5.1 and 6.1), there was no way of knowing the *real* proportions of these genres in the language, and so they could not be represented proportionally. The amounts were set at these levels in order to make each individual genre useful as an object of study in its own right (see section 4.2.3).

The remaining sections of this chapter will introduce each genre, define the texts which are included in it, and explain how (or indeed *if*) these texts were collected.

8.2 Essays

The Written BNC2014 contains both school level and university level essays (as did the Written BNC1994). These essays were collected in two ways. The first was through a PPSR approach (see Shirk et al., 2012); the British public were invited to submit any essay which they had written between 2014-2018 (in accordance with the date range policy set out in section 4.3.1) on any topic. Secondly, the corpus benefitted from a generous collaboration with Cambridge University Press and Cambridge Assessment English, who contributed data from the Cambridge Corpus of Academic English (CAMCAE[7]) to the project.

8.3 Admin

Lee (2001:65) defines the 'admin' genre as "administrative and regulatory texts, in-house use". The 'admin' genre in the Written BNC1994 contains texts such as county court practice handbooks, company manuals, and company system descriptions.

This definition has been maintained in the Written BNC2014. Most texts were collected manually from UK University and business websites. Text include, for example, Lancaster University's staff handbook and the New Look Group's code of business ethics.

8.4 Institutional

Lee (2001:65) defines the 'institutional' genre as "official/governmental documents/leaflets, company annual reports, etc.; excludes Hansard". The

[7] http://languageresearch.cambridge.org/camcae

‘institutional’ genre in the Written BNC1994 contains texts such as survey reports, official leaflets, annual council reports, and company reports and accounts.

In the Written BNC2014 this genre *does* contain Hansard (a daily edited record of what was said in Parliament, including votes and written statements). This decision was taken because having an entire genre of 900,000 words dedicated solely to Hansard seemed like it would be over-representing this, fairly unique, type of British English. Furthermore, Hansard is actually a written record of spoken language, so certainly not *Written* British English in the same way that most of the other types of data in the corpus are. In section 1.2 I concluded that speech and writing would be distinguished for the corpus dependent on the medium in which the texts were collected. Hansard texts are originally collected as speech, and so according to this definition should not be included in the Written BNC2014. However, Mollin (2007) undertook a study to see exactly how speech-like Hansard transcripts are. They found that Hansard transcripts are actually not transcripts at all, but heavily edited and decontextualized versions of what was said in parliament, with all of the edits making the text more writing-like. For example, repetitions and false starts are omitted, and words are changed to fit with the formal Hansard style. So, it seems that Hansard may actually not be as speech-like as some think. So with this in mind, the texts can be considered separately from the original spoken data, as written texts. Collected in their written form, Hansard has been included in the corpus in order to increase comparability with the Written BNC1994 (over 1 million words of Hansard was included in the Written BNC1994, and classified as a separate genre by Lee, 2001). Thus, the definition of ‘institutional’ texts in the Written BNC2014 has remained the same as in the Written BNC1994, but with the exception that Hansard is included.

Collection of this data type was done manually. Annual financial report texts were collected from publicly-available digital PDF annual reports published by a cross-section of companies listed on the London Stock Exchange. The method used to retrieve and parse annual report text is described in El Haj et al. (2017). Hansard statements were collected manually online, along with other publicly accessible government documents.

8.5 Instructional

Lee (2001:65) defines the ‘instructional’ genre as “instructional texts/DIY”. The ‘instructional’ genre in the Written BNC1994 contains texts such as recipes, and samples from instructional books and periodicals. This definition will be maintained in the Written BNC2014.

As with many texts in the corpus, the texts in this genre were sourced online. Instructional posts were downloaded from various UK websites, for example, ‘how to’ posts were downloaded from the Homebase website (a popular UK DIY supply store).

8.6 Letters, advert, commerce, and religion

For various reasons, although letters, adverts, commerce, and religion texts were included in the miscellaneous section of the sampling frame, they were not ultimately collected and included in this section of the corpus.

It became clear after discussing how to collect letters with the project team, that this would be practically impossible. None of the team could remember the last time they had sent a letter (apart from professional letters which they would not want to include in the corpus, e.g. returning a form to their bank). It became clear that, for all of the team members, email had replaced letter writing almost entirely. Indeed, a survey by the US Post Office found that the average US home received a personal letter only every seven weeks (cited in Schmid, 2011). This trend is likely to also be taking place in the UK, and the amount of letters sent may have reduced even further since the survey was carried out. Thus, the decision was made to focus on collecting emails rather than attempting to collect, what seems, anecdotally at least, to be a very marginal part of Written British English.

As more effort was put into the collection of emails, it also became clear that a very common way in which people are consuming adverts nowadays is via email. Lee (2001:65) defines the ‘advert’ genre in the Written BNC1994 as “print advertisements”. The ‘advert’ genre in the Written BNC1994 contains texts such as holiday brochures, tourist information leaflets, print adverts from magazines, and leaflets from various businesses. Many of these text types, e.g. brochures and leaflets, are now distributed via email. As we were already asking people to contribute their emails to the corpus (see section 7.4.5), it seemed that an easy way to collect this type of data would be to ask people to also send us any advertising emails they received (provided these were from a UK company or organisation). As these were advertising materials, I did not need to worry about gaining the senders permission because the purpose of the text was to be viewed by as many people as possible. As the advert genre was entirely populated by advertising emails, this genre has been moved to the e-language medium (see section 7.5 for a discussion of this).

Lee (2001:66) defines the ‘religion’ genre as “religious texts, excluding philosophy” and defines the ‘commerce’ genre as “commerce & finance, economics” but gives no indication of what *types* of texts this includes. The ‘religion’ genre in the Written BNC1994 contains samples from books about religion and the ‘commerce’ genre in the Written BNC1994 contains samples of books and periodicals on the topic of commerce. After some consideration, I felt that having two genres which consisted entirely of books and periodicals in the miscellaneous section of the corpus made no logical sense, when the books and periodicals mediums existed. Thus, these two genres were removed from the miscellaneous medium, but texts of these types will be included in the non-academic non-fiction books and in the newspaper articles.

8.7 News and drama scripts

Drama scripts were collected manually, from an online library of drama scripts. Manual collection was necessary in this case in order to confirm that the author of a play was British, which was determined by searching online for biographical information about the author. Samples were collected from plays written by British authors between 2010-2018 (in accordance with the date range policy set out in section 4.2.1). Up to one third (roughly) of each drama script was copied, in order to stay within the limits of fair dealing (see section 1.5).

News scripts proved very difficult to access. Transcripts were available online for news interviews, but these do not represent the scripted speech which was needed for a *written* corpus. Thus, this collection was expanded to include any UK television scripts published between 2014-2018 (in accordance with the date range policy set out in section 4.3.1). These were collected manually from various websites which provide scripts of television shows. Unfortunately, scripts could not be error checked due to being unable to access video of all of the television shows for which scripts were collected.

8.8 Composition of the miscellaneous and written-to-be-spoken mediums

As has already been discussed in section 8.6, the miscellaneous medium in particular changed quite drastically in reality compared to the corpus sampling frame. Table 8a shows the miscellaneous and written-to-be-spoken mediums in the Written

BNC2014 sampling frame, and table 8b shows the eventual composition of these mediums in the corpus (all numbers are provisional and subject to change at the time of writing).

The most obvious change to the miscellaneous medium of the corpus when compared to the sampling frame, is the removal of the two letters genres, the advert genre, the commerce genre, and the religion genre. These decisions have already been discussed in full in section 8.6 and so will not be discussed further here. The only other change to the miscellaneous genre is the doubling in size of the ‘W_institutional’ genre compared to the sampling frame (this genre has increased in size from 1% to 2 % of the corpus). This decision was made because a large amount of data had been lost from this medium due to the removal of the genres previously mentioned, and several readily available sources of data had been identified for the institutional genre (see section 8.4 for a discussion of these sources). Therefore, redistributing some of the data from the removed genres to the institutional genre seemed to be a wise choice. The overall impact of these changes on the miscellaneous medium of the corpus is that it now only comprises 6% of the corpus, rather than the planned 10%.

The written-to-be-spoken medium of the corpus is the least changed medium in the corpus when compared to the sampling frame. The only change which has occurred here is the renaming of the ‘W_news_script’ genre to ‘W_television_script’ in order to reflect the fact that all kinds of television scripts were included in this genre, rather than just news scripts. This decision is discussed in full in section 8.7.

Now that I have discussed the collection of all of the data for the corpus, it is important to demonstrate the utility of this data for research. This is the focus of the next chapter.

Table 8a: The miscellaneous and written-to-be-spoken mediums of the Written BNC2014 sampling frame.

Medium	**Super Genre**	**Genre**	**Target**	**Words**
Miscellaneous (10%)	Essays	W_essay_sch	1%	900,000
		W_essay_univ	1%	900,000
	Letters	W_letters_personal	1%	900,000
		W_letters_prof	1%	900,000
		W_admin	1%	900,000
		W_advert	1%	900,000
		W_commerce	1%	900,000
		W_institutional	1%	900,000
		W_instructional	1%	900,000
		W_religion	1%	900,000
Written-to-be-spoken (4%)	Written-to-be-spoken	W_news_script	2%	1,800,000
		W_fict_drama	2%	1,800,000

Table 8b: The eventual composition of the miscellaneous and written-to-be-spoken mediums of the Written BNC2014.

Medium	**Super Genre**	**Genre**	**% of corpus**	**Words**
Miscellaneous (6%)	Essays	W_essay_sch	1%	900,000
		W_essay_univ	1%	900,000
		W_admin	1%	900,000
		W_institutional	2%	1,800,000
		W_instructional	1%	900,000
Written-to-be-spoken (4%)	Written-to-be-spoken	W_television_script	2%	1,800,000
		W_fict_drama	2%	1,800,000

Chapter 9: Colloquialisation in academic British English

9.1 Introduction

This chapter presents a study which I have carried out using some early parts of the Written BNC2014. The analysis presented in this chapter focuses on the theory of colloquialisation, as applied to academic writing. As such, I will analyse a sub-set of the academic writing data which has been included in the Written BNC2014 (academic books and academic journal articles). Several comparisons will be carried out to assess whether linguistic features associated with colloquialisation have changed in frequency over time, using data from the Written BNC1994 and the Written BNC2014.

Section 9.2 presents a detailed overview of previous studies relating to colloquialisation, highlighting the linguistic features which have been found to be relevant to this phenomenon. Section 9.3 lays out the research questions which I will aim to answer in this chapter. This section also details the data used for this analysis, including my rationale for its selection. The section subsequently presents the methodology used, including the linguistic features which were studied, how these were searched for, and what statistics were used in the analysis. Section 9.4 presents my analysis, considering each research question in turn. I conclude in section 9.5 with a summary of the findings of this study, along with a consideration of some limitations as well as some directions for future research in this area.

9.2 Literature Review

Colloquialisation is "a tendency for features of the conversational spoken language to infiltrate and spread in the written language" (Leech, 2002: 72). Miller (2009: 210) characterises this phenomenon as a form of "stylistic drift" wherein the style of written language moves toward that of spoken language. Baker (2017: 243) suggests that colloquialisation of written language can make messages “more accessible to wider audiences”. This is because, whilst everyone is familiar with spoken language, many people are not familiar with the specifics of, for example, academic writing or business writing. Thus, a shift towards a more speech-like style in these written genres can make them more easily understood by the general public. Mair (2015: 6) suggests that colloquialisation of written language is a correlate of a societal trend towards “an informalisation of manner and codes of conduct”. However, Leech (2002: 76) warns that “Terms like colloquialization do represent some rather general attempt to explain change, but they do not amount to well-developed theories.” Note that for the purposes of this discussion, the prototypical cases of spoken and written language are considered (i.e. a spontaneous conversation and a book respectively), rather than the less easily defined cases (e.g. a play script; see section 1.2).

Leech (2002: 72) observes that there are two ways in which colloquialisation can be demonstrated quantitatively: “(a) by an increasing frequency of phenomena associated with spoken language, and (b) by a decreasing frequency of phenomena associated with the written language”. Thus, in order to research the phenomenon of colloquialisation it is first necessary to have an understanding of the typical features of spoken and written language as currently conceived. Biber and Finegan (1989) and Biber et al. (1999) are two key studies in the development of research into colloquialisation. They identify many evidence-based linguistic features which are more characteristic of written language or of conversation, and which can thus be used by researchers to explore how styles of English have developed over time in terms of the usage of these features.

Biber and Finegan (1989: 493), whilst not using the term ‘colloquialisation’, do refer to a distinction between *literate* and *oral* genres. A literate genre is one produced in a situation typical of writing, and an oral genre is one produced in a situation typical of speaking. Biber and Finegan (1989: 493) state that conversation and academic prose are stereotypical oral and literate genres, respectively. In their study of the evolution of linguistic style in three written genres of English over four centuries, they rely on three of the ‘dimensions’ of genre variation developed in Biber’s (1988) work on variation between speech and writing. The opposing poles of these dimensions represent literate and oral styles. Some of the features characteristic of literate styles are: a higher frequency of nouns, relative clauses, and passives; some of the features characteristic of oral styles are: a higher frequency of contractions, present-tense verbs, WH-questions, and first and second person pronouns (Biber, 1988). Biber and Finegan (1989) use the linguistic features associated with the literate and oral poles of the dimensions in order to track changes in style over time. They find that all of the genres which they analysed have drifted towards a more oral style over the four centuries studied. That is, the frequencies of features associated with literate styles have decreased, and the frequencies of features associated with oral styles have increased. Biber and Finegan (1989: 493) are careful to point out that there is not always a correspondence between literate genres and the physical mode of writing, and between oral genres and the physical mode of speech. For example, they find that personal letters are one of the most oral genres even though they are a form of written language.

Developing Biber and Finegan’s (1989) insights into oral and literate language, Biber et al. (1999) create a grammar of spoken and written English based on 40 million words of British and American English texts. This corpus includes data from various genres, and thus Biber et al. (1999) are able to give many examples of linguistic features which are more associated with speech or with writing. As a full-scale reference grammar, Biber et al. (1999) is impossible to summarise here, however, I will list a few of the most relevant observations for this thesis. Nouns, definite and indefinite articles, passive constructions, and-coordinated adjectives, prepositional phrases as post-modifiers, and relative clauses are all found by Biber et al. (1999) to be less common in conversation than in writing. Conversely, pronouns, lexical verbs, present tense verbs, semi-modal verbs, verb contractions, and negative contractions are all more common in conversation than in writing. All of the literature discussed in the remainder of this section relies, to varying degrees, on the findings of Biber and Finegan (1989) and Biber et al. (1999).

Leech (2002) compares the LOB (British English from 1961) and FLOB (British English from 1991-1992) corpora by extracting a range of grammatical features which have been associated with a trend of colloquialisation, and comparing their relative frequencies in the two corpora. Leech finds that many of these features have increased or decreased in frequency in a direction apparently indicative of colloquialisation[8]. The use of the present progressive and the progressive passive increase by 28.9% and 31.3% respectively in FLOB compared to LOB. Since Biber et al. (1999:461-463) had shown that the progressive is more common in speech than in writing, Leech (2002: 73) suggests that it is justifiable to consider colloquialisation to be a possible explanation for this change. As already discussed, Biber et al. (1999) find the passive to be associated more with the written medium, and so the 12.4% decline in the use of the passive observed by Leech (2002: 74) is also indicative of colloquialisation. Also within the verb phrase, Leech (2002) finds that the use of both

[8] See Leech, 2002: 73, Table 8 for a full list of the features and their frequencies

negative and verb contractions increases in FLOB, further reinforcing a pattern of apparent colloquialisation. However, Leech (2002: 74) warns that this increase can in fact be attributed, in part, to an increase in the proportion of reported speech in the corpus. Moving beyond the verb phrase, Leech finds an increase of 9.5% in the use of questions. However, once again, Leech (2002: 74) warns that the increase in quoted speech in FLOB may provide a readier explanation for the additional questions than colloquialisation. The choice between using an *'s* genitive and an *of*-construction has historically been seen "as a competition between more and less oral styles of expression" (Leech, 2002: 74). Leech (2002) finds that the use of genitives has increased by 24.1% in FLOB, whilst the use of *of*-phrases has decreased by 4.7%. However, if only those *of*-phrases which could be replaced by a genitive construction are considered, the decrease goes up to 23.6%. Leech (2002: 74) points out that this is an intriguing result because this decline almost exactly balances the increase in the use of the genitive. Turning to relative clauses, Leech (2002: 74) finds that the use of *wh*-relative pronouns has decreased, and thus so has the use of *wh*-relative clauses. On the other hand, he observes a large increase of 310% in the use of zero-relatives with a stranded final preposition (e.g. "someone I spoke to"; Leech, 2002: 74). Leech is careful to note that some of these findings are based on small samples and that the research thus represents a work in progress; as such, Leech's (2002) results can only be considered provisional.

Mair et al. (2003) also compare the LOB and FLOB corpora, with a focus on comparing part-of-speech tag frequencies in the two comparable corpora. The theory of colloquialisation would predict a decrease in the use of nouns in FLOB, but an increase in the use of verbs (Mair et al., 2003: 251). This prediction is based on Biber et al.'s (1999) findings that verbs are more frequent in conversation than in informative writing, whereas nouns are more frequent in writing than in conversation. However, Mair et al. (2003) find that, over the thirty-year period between LOB and FLOB, the frequency of verbs has remained virtually stable, whereas the use of nouns has increased by nearly 5%. This increase may not seem very big over a thirty-year period, but Mair states that the result is in fact statistically significant (Mair et al., 2003: 251). These results conform to neither prediction of the colloquialisation theory. Mair et al. (2003: 256) discuss these results in the context of corpus-internal variation. They state that any results from a comparison of the corpora must be interpreted with extreme caution because, in corpora such as LOB and FLOB which contain multiple genres, a small shift in tag frequencies over time may not mean much at all when there is "much greater scope for variation based on genre" (ibid.). They suggest that further work is required in order to decide how to interpret these diachronic comparison results in the context of the much greater contrasts in tag frequency that may be observed in a "corpus-internal synchronic analysis of genres" (Mair et al., 2003: 257).

Leech (2003) compares the frequencies of modal and semi-modal verbs in LOB, FLOB, Brown, and Frown. He finds that there is an overall decrease in the use of modal verbs in both American and British English between 1961 and 1991. This trend was found to be even more pronounced in the additional spoken data which he analysed. This leads Leech (2003: 96) to conclude that the decline of modal verbs is part of a "more general and long lasting trend". Leech (2003) investigates the theory that the increasing use of semi-modals may be the cause of the decline in the use of modal verbs. But the patterns for semi-modals is more mixed than for modal verbs; most of the semi-modals increase in frequency, but some (e.g. 'BE to') show a marked decline in frequency. Whilst some of the percentage increases are also quite pronounced, the absolute frequencies of the semi-modals are, when compared to the modals, relatively low. Thus, Leech (2003) concludes that the results do not appear to support the theory that modal verbs are being replaced with semi-modals. Leech (2003) finally suggests three hypotheses to explain the frequency changes that he found: *Americanisation, colloquialisation,* and *democratisation.* The American corpora show lower frequencies of modals in both time periods, perhaps suggesting that American English is leading a change in British English. The colloquialisation theory is supported by Leech's finding that the decline in modal verb usage was even more pronounced in spoken data. The democratisation theory suggests that the decline in the modals which convey stronger obligation (e.g. 'must') is part of a wider trend in society whereby etiquette increasingly demands that speakers "suppress or avoid overt claims to power and authority" (Leech, 2003: 237).

Millar (2009) seeks to replicate Leech's (2003) study of frequency changes in modal verbs using the TIME Magazine corpus. This corpus contains over 100 million words of data published in TIME magazine from 1923 to the present. Thus, any patterns of change found in this corpus can only represent this particular magazine, rather than English language as a whole (Millar, 2009: 206). Millar (2009: 206) does, however, suggest that the patterns observed may, by extension, "be expected to hold true for the genre of press reporting/magazines as a whole". Millar (2009: 199) finds that while some modal verbs have decreased in frequency across the corpus, the general trend is a 22.9% increase in overall frequency of modals (in contrast to Leech's 2003 finding). However, the patterns for individual verbs in Millar's results vary quite a lot. *Shall, ought,* and *may* show declines in frequency, whereas the frequencies of *can,* and *could* show large increases. All of the semi-modal verbs which were analysed were found to have increased considerably in frequency. Biber et al. (1999:486-490) find that semi-modal verbs are more closely associated with speech than with writing; thus, the increase in semi-modal verb constructions in the TIME corpus can be considered evidence of colloquialisation. Furthermore, Millar (2009) finds that the contracted form (*n't*) has become the preferred method of negating modal verbs; negative contraction is also closely associated with speech (Biber et al., 1999). This result thus lends more support to a theory of colloquialisation. Unlike Leech (2003), Millar (2009) observes an increase in the frequencies of *may*, *can* and *could*, whereas these modals are found to become less frequent by Leech (2003). Millar (2009: 208) suggests that this discrepancy may be due to the Brown family of corpora not representing enough data points. Millar (2009: 208) shows that in the TIME corpus *can* does decline (by 3.1%) between 1961 and 1991. However, over the full span of the data, from 1923 to 2009, *can* increases by 113.4%. Thus, comparing two intermediate data points can give a result which "radically contradicts the reality of the overall pattern" (Millar, 2009: 208) – and basing comparisons on the Brown family may produce such contradictions. That is to say, if, as it seems, changes in frequency can fluctuate over time, then "data from multiple chronological points appear to be essential to obtain a clear overview of any trend" (Millar, 2009: 208).

Baker (2009) also seeks to investigate language change using the Brown family of corpora. He compares lexical frequencies, pronoun usage, and keywords in BLOB, LOB, FLOB and BE06. When analysing pronouns, Baker (2009) encounters the fluctuations over time of which Millar (2009) warns. Baker (2009: 325) finds that the first person singular pronouns *I*, *me* and *my* increase slightly in frequency between 1931 and 1961. However, all first person pronouns are found to have decreased in frequency by 1991. This finding contradicts the theory of colloquialisation, which would suggest that first person pronouns should be becoming more frequent (Biber and Finegan, 1989). However, the 2006 data shows a higher number of first person pronouns than any other year, which suggests that the data for the year 1991 does not conform to the overall pattern (Baker, 2009: 325). The overall higher frequencies of first and second person pronouns found in the BE06 corpus are, therefore, indicative of a trend of colloquialisation, Baker (2009: 327) suggests. However, more linguistic features would need to be examined before any firm claims could be made (Baker, 2009: 327). One type of writing which bucks the overall trend is academic writing, in which the frequency of first person pronouns is found to decrease gradually over time (Baker, 2009: 329).

Like Leech (2003), Mair (2015) compares frequencies in Brown, Frown, LOB, and FLOB. Mair (2015: 1) finds "numerous statistically significant diachronic developments", including an increase in the use of the progressive and the *going-to* future, as well as an increase in the use of contracted forms. This section has already shown that these features have been linked to colloquialisation by multiple studies. Indeed, Mair (2015: 6) suggests that these changes are not due to a change in the grammar of the language, but rather they show that "informal options which have been available for a long time are chosen more frequently today than would have been the case thirty years ago". It is suggested that this trend of colloquialisation is driven by a "shift of public taste towards greater informality" (Mair, 2015: 1).

Finally, Baker (2017) compares a wide range of features in eight members of the Brown family of corpora, and finds numerous trends which can be linked to colloquialisation. Indeed, Baker (2017: 243) summarises his findings overall as "Writing's getting more like speech". He finds that both verb and negative contractions have become more frequent in both British and American English. Both varieties also show an increase in the use of the relative pronoun *who* and a corresponding decrease in the use of *whom*. Whilst, for the first two time periods represented by the corpora, British English had a higher frequency of the first person pronoun *I* than American English, American English shows a linear pattern of increase across all time periods. Baker (2017: 158) points out that this is a strong indication of colloquialisation as *I* occurs 6357 per million words in the written section of the BNC1994 and 16,560 times per million words in the spoken section. Contrary to what colloquialisation would predict, Baker (2017: 243) finds that the use of noun sequences (constructions which use multiple nouns in a row) has increased in written English over time. Baker attributes this to a trend of densification of language (i.e. a trend for information to be more densely packed in language, resulting in less function words being used between content words, and shorter sentences; Baker, 2017: 24). Further changes which can be associated with colloquialisation are: use of more affective language, less frequent use of personal titles, an increase in the use of discourse markers associated with speech (e.g. ‘sorry’ and ‘please’), and an increase in use of swearing and religious profanity. Baker (2017: 234) also addresses the suggestion made by Leech (2002: 74, discussed above) that some of the changes observed in the Brown family of corpora are due to an increase in the proportion of quoted speech in the newer corpora, rather than due to colloquialisation. Baker (2017: 234) is careful to point out that the discourse markers associated with speech and the use of swearing and religious profanities are increasing in frequency in writing outside of reported speech contexts. The shift in the generic balance in the Brown family is, therefore, shown not to be a confounding factor, and colloquialisation does seem to genuinely play a role.

9.3 Methodology

In this chapter, I will answer the following research questions:

RQ1: Have features of language associated with colloquialisation become more or less frequent in academic writing since 1994?

RQ2: Do the results of RQ1 differ between academic books and academic journal articles?

RQ3: Do the results of RQ1 differ across different genres of academic writing?

Academic writing was selected as the object of investigation in this chapter, rather than, for example, fiction or newspaper articles, based on the literature discussed in section 9.2. The literature shows that academic writing seems to be the least 'speech-like', and thus least colloquial, type of writing. Biber and Finegan (1989: 493) identify academic prose as the stereotypical example of a literate genre (as opposed to oral). Furthermore, Biber et al. (1999) find that, on many occasions, features which are most common in speech are least common in academic writing and vice versa. This implies that academic writing is the type of writing least like speech. For example, passives are by far most common in academic prose, and least common in conversation (Biber et al., 1999: 476). Similarly, Biber et al. (1999) find that "and-coordinated adjectives" are very common in academic prose but extremely rare in conversation (Biber et al., 1999: 601). The frequency of prepositional phrases used as postmodifiers varies across a scale, being extremely common in academic prose and relatively rare in speech (Biber et al., 1999: 606). Contractions are also found to be strongly associated with spoken language, and least common in academic writing (Biber et al., 1999: 1129). Another way in which academic language can be seen to be the least speech-like genre, is in its resistance to colloquialisation. For example, Baker

(2009: 329) finds that academic writing resists the trend when it comes to first person pronouns. While these pronouns increase in frequency over time in the corpora overall, they decrease over time in academic writing.

This evidence of academic writing being the least ‘speech-like’, and thus least colloquial, type of writing makes it an ideal choice for a diachronic study of colloquialisation. If academic prose is found to have become more colloquial over time, then it may be possible to tentatively infer that other types of written language will have become more colloquial to at least the same extent if not a greater extent. A swift reading of any piece of academic prose is sufficient to show that academic writing is constrained by very strong conventions and traditions regarding the expected level of formality of the language. Thus, evidence of colloquialisation in academic writing may point to an even greater degree of colloquialisation in other types of written language which are not constrained by similar conventions. To put it another way, we would expect academic prose to be the most stable type of writing and least affected by colloquialisation. Thus, if colloquialisation is observed in academic writing, it may be assumed that it has occurred in other, less formal, types of written language earlier, and more extensively.

Tables 9a and 9b show the word counts for each sub-genre of academic writing studied. All of the academic books and journals available in the BNC1994 were included in the study. I did not use *all* of the academic writing in the BNC1994, only the books and journals. This was an important distinction to make as the BNC1994 contains some texts categorised as academic which are not books or journals, but which are unpublished writing (such as theses), and which are thus not comparable to the data in the BNC2014. The academic books and journal articles from the BNC2014 used in this study represent a sub-set of the final corpus subsections.

The academic books amount to ~69% of the target amount for collection, and the journal articles amount to ~38% of the target amount. The sub-sets thus do contain plenty of data from each sub-genre. I would not expect the relative frequencies of any feature in the final data-sets to vary much from the results of the present analysis. Furthermore, the data yet to be incorporated into these parts of the Written BNC2014 will come from the same or similar sources to that which has already been collected. It is therefore a safe working assumption, albeit of course not a certainty, that this sub-set is representative of the final corpus subsections.

Table 9a: Details of the data from the BNC1994 used for analysis.

Sub-genre	**Word count (tokens)**	**Total academic books and total academic journals word count (tokens)**	**Total word count (tokens)**
W_ac_humanities_arts (books)	3,506,992	14,153,936	17,233,631
W_ac_medicine (books)	139,933		
W_ac_nat_science (books)	1,036,307		
W_ac_polit_law_edu (books)	4,425,905		
W_ac_soc_science (books)	4,513,798		
W_ac_tech_engin (books)	531,001		
W_ac_humanities_arts (journals)	190,532	3,079,695	
W_ac_medicine (journals)	1,497,792		
W_ac_nat_science (journals)	242,960		
W_ac_polit_law_edu (journals)	833,327		
W_ac_soc_science (journals)	282,622		
W_ac_tech_engin (journals)	32,462		

Table 9b: Details of the data from the BNC2014 used for analysis.

Sub-genre	**Word count (tokens)**	**Total academic books and total academic journals word count (tokens)**	**Total word count (tokens)**
W_ac_book_humanities_arts	741,762	4,132,820	6,395,903
W_ac_book_medicine	270,688		
W_ac_book_nat_science	482,340		
W_ac_book_polit_law_edu	845,165		
W_ac_book_soc_science	900,990		
W_ac_book_tech_engin	891,875		
W_ac_journal_humanities_arts	631,738	2,263,083	
W_ac_journal_medicine	231,428		
W_ac_journal_nat_science	646,114		
W_ac_journal_polit_law_edu	304,139		
W_ac_journal_soc_science	189,874		
W_ac_journal_tech_engin	259,790		

As discussed in section 9.2, the two ways in which colloquialisation can be quantified are: "(a) by an increasing frequency of phenomena associated with spoken language, and (b) by a decreasing frequency of phenomena associated with the written language" (Leech, 2002: 72). Thus, in order to answer the research questions defined at the beginning of this section, I first identified linguistic features strongly associated with either spoken or written language in the literature discussed in section 9.2. The features which I selected for analysis were:

- first and second person pronouns (Biber, 1988; Biber and Finegan, 1989; Biber et al., 1999; Baker, 2017)
- relative pronouns (Biber and Finegan, 1989; Biber et al., 1999; Leech, 2002; Baker, 2017)
- verb frequency (Biber et al., 1999; Mair et al., 2003)

- present tense verbs (Biber and Finegan, 1989; Biber et al., 1999)
- verb contractions (Biber and Finegan, 1989; Biber et al., 1999; Leech, 2002; Baker, 2017)
- negative contractions (Biber et al., 1999; Leech, 2002; Baker, 2017)
- questions (Biber and Finegan, 1989; Leech, 2002)
- *'s* genitives (Leech, 2002)
- semi-modals (Biber et al., 1999; Leech, 2003; Millar, 2009)
- passive forms (Biber and Finegan, 1989; Biber et al., 1999; Leech, 2002)
- noun frequency (Biber, 1988; Biber and Finegan, 1989; Biber et al., 1999; Mair et al., 2003).

Many more features are identified in the literature as being potentially associated with colloquialisation. However, these other features were not selected because the search terms needed to extract them from the corpora would have been too complex (e.g. prepositional phrases as post-modifiers; Biber et al., 1999), even using the state of the art corpus search tools available to me (as introduced below).

In order to answer RQ1, I searched for each feature across my defined sub-sets of the BNC1994 and the BNC2014. In order to answer RQ2, I searched for each feature in the academic books section of the BNC1994 and compared this with the same searches of the academic books in the BNC2014. I then did the same thing for the academic journals in both corpora. In order to answer RQ3, six different comparisons were carried out. Corpus sub-sets for each academic genre (i.e. discipline) were compared.

Sketch Engine (Kilgarriff et al., 2014) was used for all of the searches and analyses presented in this chapter. Sketch Engine allows users to upload their own corpora, as well as search the many corpora already available in the tool. This was perfect for this comparison, as I would be comparing the BNC1994 (available in Sketch Engine) with several different corpora from the BNC2014. The tool also allows for the searching of sub-corpora, which was necessary in order to pick out just those sections of the BNC1994 relevant to the analysis. The corpora used for this analysis were POS tagged using the tool TreeTagger with the 'English TreeTagger PoS tagset with Sketch Engine modifications'[9]. The list of features, their positive or negative associations with colloquialisation, and the search terms used to identify these features in the corpora are given in table 9c. Most of the search terms are straightforward, but I will give a brief explanation of each to make my analysis transparent and reproducible in terms of what exactly is being counted for each feature.

First and second person pronouns: This search returns any instances of the words *I*, *me*, *we*, *us*, and *you*. Note that this search misses the word *US* (in upper case) as this returns many instaces of *US* referring to the United States. Due to tagging errors in the corpus this could not be resolved using tag searches. Thus, any rare instances of upper case *US* are not found in this search.

Present tense verbs: This search returns any word tagged as a verb with present tense marking.

Verb contractions: This search returns any instances of the contracted forms of *is*, *have*, *are*, *will*, *would* and *am*. The search for *'s* is further restricted to only those instances tagged as verbs to prevent the *'s* genitive from being returned.

Negative contractions: This search returns any instance of the contracted form of *not*.

[9] The full tagset is available at: https://www.sketchengine.eu/english-treetagger-pipeline-2/

Questions (all): This search returns all question marks. Of course, questions are not *always* marked by question marks, but this search is expected to retrieve most, if not all, instances. In a genre such as academic writing, which is professionally copy-edited before publication, it is likely that question marks would be imposed for consistency of style on any occasion where an author had chosen to leave them out originally.

Verb frequency: This search returns any word tagged as a verb.

's **Genitives:** This search returns any possessive noun. However, this does mean that any rare instances of *'s* cliticised to anything other than a noun are missed (e.g. *the man I saw's hat*).

Semi-modals: This search is the most complex used for this analysis. It searches for any instances of the semi-modal verbs which were looked at in Leech (2003). There is not widespread agreement among linguists regarding a full set of semi-modal verbs, thus I based this search on a set which had already been utilised by a researcher carrying out a similar study. The semi-modals searched for are: BE *going to, gonna,* BE *to,* HAD *better, got to, gotta,* HAVE *to,* NEED *to,* WANT *to, wanna,* and *used to*. Capital letters here indicate a lemma. The word *to* in these constructions was always searched for using the tag for infinitive *to*, in order to increase the certainty that the results returned would be semi-modal. Most of the searches in the string are straightforward, but a few warrant further discussion. For BE *to,* HAD *better,* and HAVE *to* the interrogative structure can involve an auxiliary verb inverting with a subject, and thus 'interrupting' the semi-modal construction. Therefore, this must be taken into account in the searches. The searches for BE *to* and HAD *better* both allow for an optional pronoun, noun, or determiner in between the two core elements. Of course, in theory any kind of noun phrase could occur here, but I am only allowing for the most simple here which I would expect to be most common. This will of course miss some instances, but does avoid making the search string so loose that lots of things which don’t belong also get found. The search string also allows for zero, one, or two adverbs to occur before the infinitive. The number of adverbs is theoretically unlimited, but in practice, two would be the greatest number I would expect. The search for HAD *better* also contains some additions after the infinitive. Simply searching for [lemma = "have"] [tag="PP" | tag="N.*" | tag="DT"]? [tag = "RB.*"]{0,2} [word="better"] returned instances such as “she has a much better brain”, where *better* is modifying a noun, rather than acting as part of a semi-modal construction. Thus, [tag = "RB.*"]{0,2} [tag = "V.*"] was added to the end of this search string to ensure that the following word was a verb, optionally modified by up to two adverbs. In principle, HAVE *to* can also be interrupted by an inverted subject in an interrogative structure, but this would be what appears to me to be a very archaic structure (e.g. *Has he to do it?* compared to *Does he have to do it?*). However, trying to address this with the same search patterns already discussed does not work here, as they return instances such as “the court will always have jurisdiction to intervene”, where the noun phrase is an object rather than an inverted subject. There was no easy way to fix this, and as my native speaker intuition tells me that this construction is very archaic, I decided to leave this search as a simple one and not worry about the very few potential instances which I may miss. Optional adverbs were allowed for in this search in order to ensure instances such as “has still to” were found.

Passive forms (all): This search returns a string composed of any form of the verb *be*, followed by either zero, one, or two adverbs or particles, followed by a past participle verb. It is possible that more than two optional elements could occur within the passive construction, but this is unlikely.

Relative pronouns: This search returns any instance of the words *who*, *which*, *whose*, *whom,* and *what.* 'That' can also be considered to function as a relative pronoun, but its status as such is controversial, so it was not included in this analysis (Börjars and Burridge, 2010:57).

Noun frequency: This search returns any item tagged as a noun.

As can be seen from these explanations, some of these search terms have problematic elements. In some cases, there are alternate approaches to searching these features, which would likely return slightly different results. However, the exact same search term was used for each feature in each of the two corpora, so the results are directly comparable with one another, despite their potential imperfections. These search terms are written in the CQL search language required for use with Sketch Engine. These searches would have to be changed if they were to be carried out on, for example, BNCweb which uses CQP syntax as its search language. Sketch Engine CQL is a derivative of CQP syntax with minor differences. Thus, if these searches were rewritten in a different query language for use in a different tool, they may return slightly different results.

Table 9c: Searches used in Sketch Engine for each linguistic feature analysed.

Linguistic feature	**Positive/negative association with colloquialisation**	**Search**
First and second person pronouns	+	[word = "I|me|we|us|you|Me|We |Us|You|ME|WE|YOU"]
Present tense verbs	+	[tag="VBP|VBZ|VHP|VHZ|VVP|VVZ"]
Verb contractions	+	[word = "'s|'S" & tag = "V.*"|word = "'ve|'VE"|word = "'re|'RE"|word = "'ll|'LL"|word = "'d|'D"|word = "'m|'M"]
Negative contractions	+	n't
Questions (all)	+	\?
Verb frequency	+	[tag = "V.*"]
‘s Genitives	+	[tag = "NNSZ"|tag = "NNZ"|tag = "NPSZ"|tag = "NPZ"]
Semi-modals	+	([word = "going|Going|GOING"] [tag="TO"]|[word = "gonna|Gonna|GONNA"]|[lemma = "be|Be|BE"] [tag="PP" | tag="N.*" | tag="DT"]? [tag = "RB.*"]{0,2} [tag="TO"]|[lemma = "have|Have|HAVE"] [tag="PP" | tag="N.*" | tag="DT"]? [tag = "RB.*"]{0,2} [word="better|Better|BETTER"] [tag = "RB.*"]{0,2} [tag = "V.*"]| [word="got|Got|GOT"] [tag = "TO"]|[word = "gotta|Gotta|GOTTA"]|[lemma = "have|Have|HAVE"] [tag = "RB.*"]{0,2} [tag = "TO"]|[lemma = "need|Need|NEED"] [tag = "TO"]|[lemma = "want|Want|WANT"] [tag = "TO"]|[word = "wanna|Wanna|WANNA"]|[word = "used|Used|USED"] [tag = "TO"])
Passive forms (all)	-	[lemma = "be" & tag = "VB.*"] [tag = "R.*"] {0,2} [tag = "V.N"]
Relative pronouns	-	[word = "who|which|whose|whom| what|Who|Which|Whose|Whom|What|WHO|WHICH|WHOSE|WHOM|WHAT"]
Noun frequency	-	[tag = "N.*"]

Note: + indicates a predicted increase by colloquialisation theory, - indicates a predicted decrease by colloquialisation theory.

For each linguistic feature under analysis, a Bootstrap test was carried out which calculated a percentage change between the two corpora, along with a measure of statistical significance. A common significance test used in linguistic analyses is log-likelihood, and is indeed what was used by Leech (2003). However, Brezina (2018) points out that, when carrying out diachronic comparisons, differences in observed frequencies may not be due to changes in the language over time, but due to variation within the individual texts in the corpora, which the log-likelihood test does not take account of. Thus, a Bootstrap test (Lijffijt et al., 2014; Brezina, 2018) was carried out for all comparisons. A Bootstrap test takes account of variation within a corpus by resampling the corpus multiple times, and looking for a consistent difference between the resampled corpora. For my analysis, I considered any Bootstrap test results of $p<0.05$ to be significant enough to reliably indicate a change in language over time. The results of every comparison carried out for this analysis can be found in appendix L.

9.4 Analysis

In this section, I will discuss each of the research questions listed in section 9.3 in turn. I will present relevant data from my comparisons to illustrate my findings (the results of all comparisons can be found in appendix L), and discuss to what extent I can answer the research questions using the data analysed in this study.

9.4.1 Research question 1

RQ1: Have features of language associated with colloquialisation become more or less frequent in academic writing since 1994?

In order to answer this research question I compared the relative frequencies of the previously defined linguistic features (see section 9.3) in all 17,233,631 words of academic books and journals from the BNC1994 to their relative frequencies in all 6,395,903 words of academic books and journals from the BNC2014 academic sub-set (see section 9.3 for details about the data) using the Bootstrap test. Overall, four of the linguistic features have increased or decreased statistically significantly in the direction predicted by the theory of colloquialisation. I will now discuss each feature in turn (see table 9d for a summary of these results).

Table 9d: Results of the comparison of linguistic features in academic writing in the BNC1994 and the BNC2014.

	BNC1994 (freq per mill)	**BNC2014 (freq per mill)**	**Difference (+/- %)**	**Statistically Significant? (p<0.05)**
First and second person pronouns	4,526.78	5,848.72	+19.226	NO
Present tense verbs	38,018.11	37,957.90	-8.244	NO
Verb contractions	347.22	713.9	+90.094	YES
Negative contractions	189.75	384.03	+86.775	YES
Questions (all)	780.8	805.65	-4.808	NO
Verb frequency	134,352.71	122,456.80	-16.154	YES
's Genitives	3,664.52	3,675.00	-7.531	NO
Semi-modals	2,033.46	1,660.99	-24.624	YES (p<0.001)
Passive forms (all)	15,441.50	11,181.40	-33.176	YES (p<0.001)
Relative pronouns	8,420.04	6,010.71	-34.126	YES (p<0.001)
Noun frequency	254,176.85	273,095.00	-1.131	NO

The use of passive forms (-33.176%) and relative pronouns (-34.126%), both show statistically significant declines in usage between 1994 and 2014, as predicted by the theory of colloquialisation. The change in the use of passive forms and relative pronouns are found to be highly statistically significant by the bootstrap test at the p<0.001 level (95% CI [-37.21, -29.142]; 95% CI [-39.718, -28.535] respectively). The decreasing use of complex sentence structures such as passives and relative clauses introduced by relative pronouns seems to indicate that academic writing may be becoming increasingly simplified. As well as being evidence of colloquialisation, this change may also point to a trend of 'densification' (Baker, 2017: 24), whereby information becomes more densely packed in language, resulting in shorter sentences.

Table 9e: Examples of passive constructions and relative pronouns in the academic writing in the Written BNC1994 (where they are statistically significantly more frequent than in the Written BNC2014).

although, as	**is shown**	by the inscription
This will	**be described**	in Chapter 14
the most difficult questions	**which**	teachers ask themselves
The ex-patients	**who**	showed the heaviest

A further two of the linguistic features studied show large, statistically significant, increases in frequency. The use of verb contractions increases by 90.094%, and the use of negative contractions increases by 86.775%, in line with what would be expected by a theory of colloquialisation. Both changes are significant at the p<0.05 level in the bootstrap test (95% CI [59.155, 121.032]; 95% CI [53.281, 120.268] respectively). In Leech's (2002) study, he warns that the increase seen in verb and negative contractions may be due to an increase in reported speech, however, this seems an unlikely explanation here. Academic books and journals typically contain very little, or no, reported speech. Looking more closely at the BNC2014 data, it becomes apparent that there is *some* reported speech in the academic journal articles, in the form of short interviews with academics. This may account for some of the increase in the use of verb and negative contractions. However, the amount of reported speech is certainly not enough to account for such a large increase in the presence of these features. Indeed, when looking at books and journals separately (see section 9.4.2) it is clear that academic books show the same very large increase in these features, despite not being affected by the increase in reported speech present in the journal articles.

Table 9f: Examples of verb and negative contractions in the academic writing in the Written BNC2014 (where thay are statistically significantly more frequent than in the Written BNC1994).

So to reconstruct, if that	**'s**	what you want to do
an interesting sample of what	**'s**	possible when it comes to
Regular sugar is	**n't**	any more or less natural
which does	**n't**	signify that the

Two linguistic features show percentage changes in the direction predicted by colloquialisation, but not at a statistically significant level. The use of first and second person pronouns increase by 19.226%, and the use of nouns decreases by 1.131%. A further three features change in the opposite direction to that predicted by colloquialisation, but not to a statistically significant level. Present tense verbs decrease by 8.244%, use of questions decreases by 4.808% and use of *’s* genitives decreases by 7.531%. None of these changes are statistically significant, so there can be no certainty that they actually indicate a change over time. Rather, they likely indicate that the features fluctuate up and down over time, and are in fact relatively stable, rather than being evidence of a change in academic writing.

There are two features which show a statistically significant change in a direction *not* predicted by the theory of colloquialisation: semi-modal verbs decrease by a statistically significant 24.624% (p<0.001, 95% CI [-30.995, -18.254]) and overall verb frequency decreases by a statistically significant 16.154% (p<0.05, 95% CI [-22.717, -9.591]). Findings regarding semi-modal verbs in previous literature have been mixed, with results often contradicting each other (Leech, 2003; Millar, 2009), which may indicate that semi-modal verbs fluctuate across different types of language, and are therefore not stable enough to be used as a predictor of colloquialisation. However, this does not seem to be a robust enough explanation to account for the decrease seen in semi-modal verbs between the 1994 and 2014 academic writing. This change is statistically significant enough to conclude that semi-modal verbs are certainly being used less in the 2014 sample, rather than differences being due to small fluctuations over time. One possible explanation for this change may be that there is a decreased use of all modal verbs in academic writing since 1994. A quick analysis shows that this does indeed seem to be the case – the use of modal verbs has decreased by 29.5% in the BNC2014 data. The theory of colloquialisation would predict that this decrease in modal verbs would be matched by an increase in the use of semi-modals, but this does not seem to be the case here. In terms of overall verb frequency, there are other studies where noun and verb frequencies have not conformed to the changes that would be predicted by colloquialisation. Mair et al. (2003) find that over a thirty year period the use of verbs remains relatively stable. So, it seems that rather than providing evidence against the colloquialisation of academic writing, verbs may simply be an unreliable indicator of colloquialisation.

Table 9g: Examples of semi-modal verbs in the academic writing in the Written BNC1994 (where they are statistically significantly more frequent than in the Written BNC2014).

with group work,	**need to**	be admitted and addressed
these standards are	**going to**	have an impact

Overall then, it seems that the answer to RQ1 is not straightforward. Four features associated with colloquialisation have shown statistically significant changes in frequency (in a direction predicted by colloquialisation) between 1994 and 2014, and only two features have shown statistically significant changes contrary to what colloquialisation theory would predict. However, a further five features have shown only very small, not statistically significant fluctuations, indicative of a stability in frequency. When looked at together these results seem to indicate that academic writing is certainly not becoming *less* colloquial, and in some aspects is becoming markedly more colloquial than in the 1990s. However, before drawing any conclusions which go beyond the data used in this analysis, many more data points would be needed. Both Baker (2009) and Millar (2009) find that data from some years in their studies radically contradict the overall pattern observed. Thus, it may be the case that in the present study, some of these results are anomalous, and when taken together with more data points may present themselves as simple fluctuations in a larger overall pattern.

9.4.2 Research question 2

RQ2: Do the results of RQ1 differ between academic books and academic journal articles?

In order to answer this question I ran the same analysis as in section 9.4.1, but this time on four different corpora. I firstly compared the frequency changes between all academic books in the BNC1994 with all academic books in the BNC2014 sub-set. I then compared the frequency changes between all academic journal articles in the BNC1994 and all academic journal articles in the BNC2014 sub-set. The results of these comparisons can be seen in table 9e. There are some notable differences between the academic books and the academic journal articles, and also some notable differences between the comparisons in research question 1 and 2, indicated by the bootstrap test results.

The most immediately noticeable difference between the two comparisons is that none of the features tested show any similarities in their changes between the two comparisons. That is to say, none of the features which showed statistically significant changes in the direction of colloquialisation in one comparison showed the same change in the other comparison. This can be seen more clearly in table 9h. The corpora which showed the greatest amount of colloquialisation were the books corpora, with eight out of eleven features showing statistically significant changes in the direction of colloquialisation. The journals corpora only showed statistically significant changes in the direction of colloquialisation for three features, and actually found statistically significant changes in the opposite direction to colloquialisation for six features. I will discuss each of these comparisons in turn.

Table 9h: Results of the comparison of linguistic features in academic books and academic journals in the BNC1994 and the BNC2014.

	Frequency change in academic books (+/- %)	**Statistically Significant? (p<0.05)**	**Frequency change in academic journals (+/- %)**	**Statistically Significant? (p<0.05)**
First and second person pronouns	+189.395	YES (p<0.001)	-52.827	YES (p<0.001)
Present tense verbs	+104.141	YES (p<0.001)	-51.595	YES (p<0.001)
Verb contractions	+334.906	YES (p<0.001)	+75.067	NO
Negative contractions	+353.998	YES (p<0.001)	+24.73	NO
Questions (all)	+125.878	YES (p<0.001)	-49.277	YES (p<0.001)
Verb frequency	+85.753	YES (p<0.001)	-60.944	YES (p<0.001)
‘s Genitives	+98.243	YES (p<0.001)	-52.644	YES (p<0.001)
Semi-modals	+71.788	YES (p<0.001)	-59.343	YES (p<0.001)
Passive forms (all)	+40.504	YES (p<0.001)	-70.048	YES (p<0.001)
Relative pronouns	+52.596	YES (p<0.001)	-67.832	YES (p<0.001)
Noun frequency	+109.371	YES (p<0.001)	-54.546	YES (p<0.001)

It is important to note at this point that the figures in appendix L may initially seem confusing. For many of the comparisons, the relative frequencies may, for example, decrease over time, but the percentage change found shows an increase. This shows that the Bootstrap test was very much necessary for verifying these comparisons. In these instances, whilst the overall use of a feature may have decreased, when the corpora are resampled and individual texts are taken into account we see that the feature is actually increasing in use, and that actually just a few texts were skewing the results. A good example of this is the comparison of semi-modals in the academic books corpora: simple descriptive statistics would suggest that the use of this feature has decreased by 13.73%; however, when the Bootstrap test is applied we see that the use of semi-modals has actually increased by a statistically significant 71.788%.

In the comparison of the books corpora, all of the features tested showed statistically significant changes at the p<0.001 level (see table 9h). First and second person pronouns, present tense verbs, verb contractions, negative contractions, questions, overall verb frequency, *'s* genitives, and semi-modal verbs all changed frequency in the direction predicted by colloquialisation (see table 9i for examples). The use of passive forms, relative pronouns, and nouns all changed statistically significantly in the opposite direction to that predicted by colloquialisation. For some features, this is in direct contrast to the findings presented in section 9.4.1. For example, the comparison of all academic writing found that the use of all verbs and semi-modal verbs was changing significantly in the opposite direction to that predicted by colloquialisation. This is in contrast to the findings from the academic books, which show that the use of verbs and semi-modal verbs has increased by a statistically significant 85.753% (p<0.001, 95% CI [66.047, 105.459]) and 71.788% (p<0.001, 95% CI [50.978, 92.598] respectively. This finding is particularly interesting for semi-modal verbs, as this was a problematic change to explain in section 9.4.1, but here shows the expected direction of change for colloquialisation. On the other hand, the comparison of all academic writing found that the use of passive constructions and relative pronouns was changing statistically significantly in the direction predicted by colloquialisation. However, in the comparison of academic books these features are found to change statistically significantly in the opposite direction (+40.504%, p<0.001, 95% CI [24.839, 56.169]; +52.596%, p<0.001, 95% CI [34.74, 70.453] respectively).

Table 9i: Examples from the academic books section of the Written BNC2014 of features which increased in frequency, when compared to the academic books section of the Written BNC1994.

First and second person pronouns	In fact, as	**I**	hope my discussion
Verb contractions	can explain what you	**'re**	studying in terms of
Negative contractions	One grain of sand does	**n't**	constitute a heap
Questions	What is a just interpretation	**?**	What is justice
’s genitives	Of course,	**Darwin's**	'daring and momentous
Semi-modal verbs	USU experience	**needs to**	be nuanced accordingly

The comparison of the journals corpora, as mentioned, is very different to the comparison of the books corpora, but has more in common with the comparison of all academic writing than the books comparison does. All but two of the features compared between the journals corpora show statistically significant changes, three in the direction predicted by colloquialisation, and six in the opposite direction. Interestingly, the three features which showed statistically significant changes in the direction of colloquialisation in the journals comparison are the three features which showed statistically significant changes in the *opposite* direction to colloquialisation in the books comparison: passive forms, relative pronouns, and nouns. This is similar to what was found in the overall comparison, in which passive forms and relative pronouns also showed a statistically significant change in the direction of colloquialisation. Contrary to the books comparison, the journals comparison finds that the use of first and second person pronouns, present tense verbs, questions, overall verb frequency, genitives, and semi-modal verbs have changed statistically significantly (at the p<0.001 level), in the opposite direction to that predicted by colloquialisation (i.e. decreasing use). This contrasts sharply with the books comparison, and also with the overall comparison. In the overall comparison, verb frequency and semi-modal verb use was also found to decrease statistically significantly, however, the other features showed no statistically significant changes.

So it seems clear from this analysis that academic books are showing much greater levels of colloquialisation than academic journal articles. The books corpora showed many more changes which were statistically significant and in line with the predictions of colloquialisation theory than the journals did. It may be hypothesised that this greater level of colloquialisation seen in the books corpora is a result of the books corpus being markedly *less* colloquial than the journals in the BNC1994. This would mean that, rather than the books becoming more colloquial than the journals, the books simply had more room to change than the journals did. However, looking at the relative frequencies of the linguistic features for each corpus (see table 9h) this does not seem to be the case. This analysis has highlighted the importance of using a form of statistical significance testing which takes into account the distribution of the features under analysis across all of the texts within a corpus. Whilst the differences between the relative frequencies (see appendix L) seem to indicate that the books and journals corpora had many more changes in common with each other than were different, the statistical significance testing showed a rather different picture.

9.4.3 Research question 3

RQ3: Do the results of RQ1 differ between different genres of academic writing?

In order to answer this research question I compared each genre of academic writing (both books and journals) in the BNC1994 to each genre of academic writing (both books and journals) in the BNC2014. See table 9j for details of the genres analysed, and see chapters 5 and 6 for a discussion of the genres. Given the large amount of variables, I do not have space to report every finding here. However, details of all comparisons can be found in appendix L. In the majority of genres, many frequency changes (or lack of changes) are consistent with the results of one or another of the comparisons in RQ2. Results differ substantially between some genres, and vary in their consistency with the results of RQ1. These differences and similarities will be discussed in detail in this section. See table 9k for a summary of results.

Table 9j: Data analysed in the comparison of difference genres of academic writing.

Genre	**Word count in the BNC1994 (tokens)**	**Word count in the BNC2014 (tokens)**
Humanities and arts	3,697,524	1,373,500
Medicine	1,637,725	502,116
Natural science	1,279,267	1,128,454
Politics, law and education	5,259,232	1,149,304
Social science	4,796,420	1,090,864
Technology and engineering	563,463	1,151,665

Perhaps the most notable comparison to make when looking at the genres of academic writing is that the changes seen in the politics, law and education genre exactly mirror the changes seen in the academic books corpora in section 9.4.2. That is to say, first and second person pronouns, present tense verbs, verb contractions, negative contractions, questions, overall verb frequency, *'s* genitives, and semi-modals have all shown a statistically significant change in the direction of colloquialisation, whereas passive forms, relative pronouns, and overall noun use have shown statistically significant changes in the opposite direction to that predicted by colloquialisation. This was exactly what was found in the comparison of the academic books corpora. This is also the case to a slightly lesser extent in the social science genre, where all the same changes are seen with the exception that verb contractions and relative pronouns do not show statistically significant changes in frequency. Although, it should be noted that these two genres contain more data from books than journals, and so results similar to the books comparison may not be surprising. Nevertheless, it seems clear, from looking at these results that the politics, law and education genre is changing the most in the direction of colloquialisation out of all of the genres studied, closely followed by the social science genre. It is possible that this points to a divide in the types of genres studied; perhaps genres with a social aspect, such as the two mentioned here, are becoming more colloquial, or showing greater changes in line with colloquialisation, than the other genres studied.

Table 9k: Changes found for the comparison of linguistic features in different genres of academic writing in the BNC1994 and the BNC2014.

	Humanities and arts (+/- %)	**Medicine (+/- %)**	**Natural science (+/- %)**	**Politics, law and education (+/- %)**	**Social science (+/- %)**	**Technology and engineering (+/- %)**
First and second person pronouns	-42.433*	+98.399	-14.726	+215.846*	+139.531*	-79.174*
Present tense verbs	-49.359*	-42.042	-53.631*	+130.014*	+85.463*	-76.446*
Verb contractions	+32.131	+956.599	+391.451	+221.94*	+161.907	+448.786
Negative contractions	-15.761	+602.846	+16.604	+267.677*	+231.571*	+107.216
Questions (all)	-39.752*	+32.385	-12.479	+88.937*	+81.246*	-68.214
Verb frequency	-59.904*	-55.523*	-53.631*	+111.584*	+87.733*	-73.53*
‘s Genitives	-46.539*	-50.08	-23.626	+148.509*	+83.48*	-32.745
Semi-modals	-70.737*	-34.171	-33.825	+86.384*	+78.593*	-77.17*
Passive forms (all)	-67.047*	-68.851*	-66.421*	+51.743*	+49.484*	-81.956*
Relative pronouns	-67.653*	+0.304	-54.945*	+71.875*	+34.793	-45.379*
Noun frequency	-51.567*	-54.495*	-47.378*	+142.484*	+111.191*	-66.948*

Note: For reasons of space, positive statistical significance is marked with a *. Full results of the Bootstrap test can be seen in appendix L.

Conversely, the results of the humanities and arts genre comparison precisely mirror those of the academic journals comparison seen in section 9.4.2. That is, the use of passive forms, relative pronouns, and overall noun frequency have shown statistically significant changes in line with colloquialisation, whereas the use of first and second person pronouns, present tense verbs, questions, overall verb frequency, *'s* genitives, and semi-modal verbs have shown statistically significant changes in the opposite direction to that predicted by colloquialisation (verb contractions and negative contractions showed no statistically significant changes). This is exactly the same as the results seen in the comparison of the academic journals. The natural science and technology & engineering genres also closely mirror the results of the academic journals comparison, but with fewer statistically significant changes observed (see table 9k for the full results).

The medicine genre shows the least statistically significant changes overall, with only three features found to have changed to an extent that was statistically significant. Overall verb frequency decreased by a statistically significant 55.523% (p<0.05, 95% CI [-65.265, -45.782]); the use of passive forms decreased by a statistically significant 68.851% (p<0.001, 95% CI [-77.033, -60.668]); and overall noun frequency fell by a statistically significant 54.495% (p<0.05, 95% CI [-65.073, -43.916]). All other features showed no statistically significant changes between 1994 and 2014. Combined with the findings discussed above, it seems that it may be the case that the 'hard' science genres, i.e. medicine, natural science, and technology and engineering, are the most stable genres in terms of colloquialisation. Whilst all of the 'hard' science genres show statistically significant changes both in the direction of and in the opposite direction to colloquialisation, they also certainly show the fewest statistically significant changes overall. As discussed, the medicine genre only shows three statistically significant changes, the natural science genre only shows six, and the technology and engineering genre only seven. It may be hypothesised that the 'hard' science genres already showed greater levels of features associated with colloquialisation in 1994 than the other genres did, and therefore, are simply changing less because they were already more colloquial to begin with. However, a quick comparison of the relative frequencies of the features studied in 1994 between the 'hard' science genres and the other genres shows that this is not the case. Very often the 'hard' science genres show lower relative frequencies in 1994 than the other genres, suggesting that this hypothesis is false. It seems then that the 'hard' science genres are simply less susceptible to change when it comes to colloquialisation, both in the direction of colloquialisation and against it.

All of the genres compared in this analysis show some changes in common with the changes seen in RQ1. In particular, the use of passive forms and relative pronouns was found to decrease statistically significantly in the overall comparison, and this is also the case for the humanities and arts, natural science, and technology & engineering genres. All of the genres studied also show differences to the findings of RQ1, but these differences vary from genre to genre (all results can be seen in appendix L).

Overall then, the answer to RQ3 seems to be a resounding *yes* – the results of RQ1 certainly do differ between different genres of academic writing. Genres with a 'social' aspect (i.e. the social science and politics, law & education genres) show many more changes in line with colloquialisation theory than other genres, and show changes which are extremely similar, or identical to, the results found for the comparison of academic books in section 9.4.2. Furthermore, the 'hard' science genres seem to be changing the least out of all of the genres, with relatively few statistically significant changes found compared to other genres. This change is also shown not to be an effect of the 'hard' science genres being more colloquial to begin with. The results found in this comparison vary greatly in their level of similarity to the results of RQ1, but all genres have some similarities and some differences to the overall comparison.

9.4.4 Discussion

For academic writing, the analyses presented in this chapter show that some features associated with colloquialisation have certainly become statistically significantly more frequent since the 1990s, and only two features (semi-modal verbs and overall verb frequency) have changed in a way which statistically significantly contradicts colloquialisation theory. The frequency changes for these features seem to be variable across different genres, and when different mediums are compared. Leech (2002: 72) defines colloquialisation as "a tendency for features of the conversational spoken language to infiltrate and spread in the written language", and so it seems that I can conclude that, in several respects, academic writing has become more colloquial since the 1990s. Baker (2017:243) notes that colloquialisation "generally makes messages more accessible to wider audiences". It is possible that, given the growing trend for academic writing to be published in an open-access format, authors and editors have been deliberately moving towards a more colloquial style in order to make their content more easily understood by the wider audience to whom their work is now available. There is also a possibility that authors and editors may be keen to present their work in a more easily understood way in order to allow news outlets to easily pick up on key findings and report these more widely. However, only two of the features which showed relatively consistent changes – passive forms and relative pronouns – point to the use of simpler language and constructions. It is possible then that academics also want their writing to appear less formal (and thus, more like informal speech) in order to appeal to this wider potential readership. However, the various limitations of the data must be taken into account before assessing the validity of these conclusions – these will be discussed in section 9.5.2.

Turning to the colloquialisation of language in general between 1994 and 2014, it is much harder to draw any firm conclusions. As discussed in section 9.3, academic writing has been found to be the least ‘speech-like’ (and thus least colloquial) type of written language. We know then that academic writing appears to be the least susceptible type of writing to colloquialisation and so the finding that academic writing has become more colloquial may point to language in general becoming more colloquial. If this change has spread to academic writing, we can infer that it has already happened to the less formal and more colloquial types of language. Of course, this is purely a prediction based on the data and literature available. In order to draw firm conclusions, these analyses would need to be carried out on many other types of writing. One alternative theory is that as academic writing was the least colloquial type of language to begin with, it has simply become more colloquial in order to ‘catch up’ with the rest of written British English, whilst other types of language have remained stable. Thus, colloquialisation of academic writing, does not necessarily suggest colloquialisation of language in general.

9.5 Conclusion

9.5.1 Summary

This chapter has presented a study which aimed to answer the following research questions:

RQ1: Have features of language associated with colloquialisation become more or less frequent in academic writing since 1994?

RQ2: Do the results of RQ1 differ between academic books and academic journal articles?

RQ3: Do the results of RQ1 differ between different genres of academic writing?

In order to answer these research questions, a comparison of academic writing (both books and journals articles) was carried out using data from the BNC1994 and the Written BNC2014. The relative frequencies of eleven linguistic features which have been found to be associated with colloquialisation were compared in the two corpora. The linguistic features studied were: first and second person pronouns, present tense verbs, verb contractions, negative contractions, questions, verb frequency, *'s* genitives, semi-modal verbs, passive forms, relative pronouns, and noun frequency. This comparison was also carried out on the data separately for books and journals, and separately for each academic genre.

Overall, findings show that, at least in some aspects, academic writing is becoming more colloquial. In the comparison of all academic writing, four of the features studied were found to have changed statistically significantly in a direction predicted by colloquialisation, five were found to have remained relatively stable, and only two changed significantly in a direction contrary to the predictions of colloquialisation theory. When comparing books and journal articles, findings show that these mediums vary greatly in terms of colloquialisation. The books showed eight statistically significant changes in line with colloquialisation theory, whereas the journals only showed three. This difference may suggest that academic books are more susceptible to the colloquialisation of language than academic journal articles. When comparing the different genres of academic writing, a contrast was found between the 'hard' science genres (medicine, natural science, and technology and engineering) and the ‘social’ genres (politics, law and education, and social science). The ‘hard’ science genres showed the least changes, either in line with or contrary to colloquialisation theory. The ‘social’ genres, on the other hand, showed the most statistically significant changes in line with colloquialisation theory.

Whilst it can certainly be said that the academic writing included in the BNC2014 is, in many ways, more colloquial than that included in the BNC1994, it is much harder to draw any firm conclusions about the colloquialisation of written British English in general. If colloquialisation has occurred in academic writing, we may tentatively infer that it has already happened to the less formal and more colloquial types of language, such as newspaper articles, or fiction books. Of course, this is purely a prediction based on the data and literature available. In order to draw firm conclusions, these analyses would need to be carried out on many other types of writing (see section 9.5.3).

9.5.2 Limitations

Whilst this study has produced some interesting findings, the validity of these must be considered in relation to the limitations of the methodology and the data used in the analysis. Firstly, the search terms used, whilst capturing the features as accurately as possible, may have failed to capture some instances or may have captured instances which they should not have. Furthermore, whilst the accuracy of POS tagging is generally high, mistakes are always made which could have further limited the accuracy of the results returned by the search terms. Another methodological issue encountered was the lack of a clear way to *quantify* the amount of colloquialisation observed for each variable. This meant that conclusions were limited to comparisons of large numbers of variables, and made it difficult to assess to what degree academic writing had become more colloquial.

There were also several imperfections in the data which may limit the validity of the results of this study. Firstly, as discussed in section 9.3, the data used from the BNC2014 represents a substantial sub-set of the academic writing that is in the finished corpus. This means that results may be different if this study were to be replicated using the full data set. However, the remaining data to be collected will be gathered from the same or very similar sources to the data which has already been collected, which should mean that the academic sub-set used here is representative of the full data-set. Additionally, the data used represents both mediums (books and journal articles) and all six genres. The fact that only academic writing was used in this study limits my ability to draw any conclusions about British English in general. Whilst I can suggest that the colloquialisation of academic writing may point to the colloquialisation of other types of language, this cannot be a certainty until other types of data are included in the study.

There were also several less prominent issues with the data, which were only discovered once analysis had begun. Parts of the colloquialisation effects observed could be due to the fact that some of the data in the BNC2014 sub-set takes the form of interviews, which include a large volume of questions and reported speech. Of course, they did occur in the academic writing, so do represent the current nature of academic publishing. However, an informal interview is probably not what the general public would think of as an academic journal article. It may be possible to leave these samples in the final corpus, so as to accurately represent what is being published in these mediums, but categorise them in their own separate genre, so that people using the corpus can choose whether to include them in their analyses. A further difficulty with the data is the lack of certainty over the authors native languages (see Chapters 5 and 6 for a full discussion of this issue). Whilst all possible and practicable measures have been used to ensure the ‘Britishness’ of the academic writing included in the corpus, this remains an uncertainty for many of the authors included. Thus, it is possible to say that the findings are representative of the academic writing published in Britain, but drawing conclusions about the language used by native British academics is much harder to do based on this data.

Probably the biggest limiting factor on the validity of the conclusions presented here is the few data points used for analysis. Millar (2009: 208) points out that comparing only two data points can “radically contradict[s] the reality of the overall pattern”. For example, Millar (2009) finds that in the TIME corpus *can* decreases in frequency between 1961 and 1991, however when the full span of data is considered *can* actually increases by a large amount. In the present study, it is possible that using only two data points may have resulted in anomalous results obscuring the overall pattern. For example, if more data points were used from the time between the BNC1994 and the BNC2014 then we may see a gradual increase in semi-modal verbs, with the BNC2014 representing an anomalously low frequency.

9.5.3 Future research

The next steps for the expansion of this study are all natural progressions from the limitations discussed in section 9.5.2. The study needs to be replicated using the full academic data-set in order to confirm that the results from this study are representative of the academic writing included in the Written BNC2014. Next, the study would need to be replicated using other super-genres of data from the BNC2014, for example fiction books or newspapers, in order to draw firmer conclusions about the colloquialisation of written British English in general. Most importantly though, this study needs to be replicated with many more data points included in the analysis. This could be achieved by using the various British English members of the Brown family, although they are of course not perfect matches with the data included in the BNC1994 or the BNC2014. Another possibility would be the British Academic Written English corpus (BAWE), which contains 6.5 million words of academic writing, although the academic writing contained in the corpus is student writing, rather than professional, edited academic texts. Using these corpora would allow more certainty over whether results are showing a pattern, or are simply anomalous as part of a bigger picture.

Chapter 10: Conclusion

10.1 Overview of the thesis

This thesis has presented a detailed account of the design, construction, and initial analysis of a brand new, contemporary corpus of written British English – the Written BNC2014. My aim, throughout this thesis, has been to highlight the major challenges faced in the design and construction of the corpus, and to detail the decisions which I made to overcome these. I have also demonstrated the research potential of the corpus by conducting an analysis of some early data from the corpus. Based on this, I have no doubt that the Written BNC2014 will be widely used by the corpus linguistics community and beyond.

In chapter 1 I demonstrated the necessity of the creation of a new, contemporary corpus of written British English. It became clear that there had been no written corpora since the Written BNC1994 which had met all of the same goals, and thus, the linguistics research community was often forced to choose between using the BNC1994 as a proxy for present-day English, or using a more modern corpus which was less than ideal in other ways. Given this fact, I set out to create a corpus which *did* meet the crucial goals of the BNC1994 project, specifically:

- To create a synchronic corpus of contemporary language
- To include a range of samples from the full range of Written British English
- To use a non-opportunistic design
- To make the corpus generally available

Chapters 4-8 detailed precisely how all of these goals were achieved, and chapter 9 demonstrated the utility of a corpus which meets all of these goals. Chapter 1 also gave details of the current state of copyright law in the UK, and investigated the relevant exceptions to this law for the project at hand. I found that the ‘Non-commercial research’ exception to the law could be utilised to cover almost all types of data collection for the corpus, although would place some restrictions on the data being collected, chiefly the restriction of remaining within the bounds of fair dealing. Mine and the team’s adherence to UK copyright law was discussed in chapters 5-8, and the impact that this had on the data collected was considered.

Chapters 2, 3, and 4 satisfy research aim 1 of this thesis:

(1) To survey relevant literature in the field of corpus creation, and to use this to design a sampling frame for the Written BNC2014.

Chapter 2 gave a detailed account of other contemporary national corpus projects. This was vital in order to satisfy research aim 1, because it highlighted the issues which other corpus creators had faced in similar projects, and allowed me to adjust the design of the sampling frame accordingly. In particular, the issue of copyright restrictions emerged as something which had greatly damaged some of the projects considered, and so this issue was given careful consideration at all stages of the project (and thus, is considered throughout this thesis). The national corpus projects discussed in this chapter were returned to multiple times throughout the thesis in order to contextualise the many decisions which were made within.

Chapter 3 considered, in detail, what were, without a doubt, the two most important issues in the design of the Written BNC2014: representativeness of contemporary language, and comparability with the Written BNC1994. I found that there seemed to be a consensus amongst experts that creating, or at the very least proving that you have created, a truly representative corpus is simply not possible. However, it was also made clear in the literature that many feel that representativeness should still be aimed for, even if it cannot be fully achieved. I also considered previous sets of comparable corpora, and research on how comparable corpora can be created. It became clear through this that representativeness and comparability can often be at odds, because prioritising the following of an old sampling frame, for the sake of comparability, can limit a researcher's ability to fully represent their target population. This conflict led me to the first major decision which I made on the project: I would prioritise the representation of contemporary British English, with comparability with the Written BNC1994 being a secondary concern. I will, however, create a fully comparable sub-corpus of the Written BNC2014 once the corpus is completed, in order to allow for direct comparisons between the two corpora.

Chapter 4 detailed the very many decisions which needed to be made in the design of the Written BNC2014 sampling frame. I first considered the issue of text classification, and sought to find out what type of classification would be most useful in the corpus. I ultimately decided, based on extensive research, that the texts in the corpus would be classified into *genres*, *super genres*, and *mediums*. This decision both aids the comparability of the Written BNC1994 to the new corpus, and also takes account of other linguists' preferences. Once this had been decided, I needed to return to the many issues discussed in chapter 3 and decide how these would be dealt with in the design of the Written BNC2014. The population which I was aiming to represent was defined as 'all written texts which were produced by native speakers of British English in 2014'. However, this definition was not as straightforward as it at first seems, as there is no exhaustive list of members of this population. The decision was taken to represent both language production and language reception within the defined population. This was achieved by, for example, including books (which are produced by few, but read by many), but also e-language (which is produced by many, but often only read by a handful of people). Next, I had to make decisions regarding the sample size to be used for texts in the corpus. It was quickly decided that samples would be used rather than whole texts, chiefly due to copyright restrictions. The typical sample size was set at 5000 words, but with full acknowledgement that this would need to be made both bigger and smaller for certain types of texts. In this chapter I also decided on the overall corpus size (90 million words), what genres would be included in the corpus, the proportions of these genres, and the sampling methods that would be used to collect these genres. There were far too many decisions made in this chapter to recount them all here, but they were all made with great care and consideration of the previous literature discussed in chapters 2 and 3. This chapter culminated with the design of the Written BNC2014 sampling frame (see appendix B).

Chapters 5-8 then satisfied research aims 2 and 3:

(2) To test and implement methods of collection for all of the data types to be included in the corpus

(3) To implement the findings of (1) and (2) in order to create the Written BNC2014

Chapter 5 detailed the long series of trials which were conducted to ascertain the best ways to collect the various types of books needed for the corpus. Each method was tested, evaluated, and a conclusion was reached about whether it was a viable collection method for this data type. The two most successful methods which emerged were collecting open-access books, and scanning print books and converting them using OCR technology. This chapter included an investigation of OCR technology, with Google OCR emerging as the most reliable software in this case.

Chapter 6 detailed the design of the periodicals medium of the corpus and the methods used to collect periodicals. The chief focus of this chapter was on the collection of magazine articles. I carried out a detailed study in order to find out whether online magazine articles are comparable to their print counterparts. I found that the majority of print magazine articles are not replicated online. This meant that a decision had to be made between the much more time consuming collection of print magazine articles, or the faster and easier, but less representative of traditional magazines, collection of online magazine articles. I ultimately decided, for largely practical reasons, that online magazine articles would be collected, despite their differences from print versions of magazines. This piece of research may be very important to any other researchers investigating magazine data, who may assume that websites are an easy way to access copies of print magazine articles.

Chapter 7 discussed in detail one of the biggest innovations which I made on this project – the design and construction of the e-language medium of the corpus. It was necessary to include e-language in the corpus in order to satisfy the target of creating a maximally representative corpus, however, this was an entirely new medium of language compared to the Written BNC1994 and so I went to great efforts to ensure that this section of the corpus was well designed. I considered research into the composition of the web, and also previous corpora of e-language in order to provide context and guide the design of the medium. I also considered, in detail, the legal and ethical issues encountered when collecting this type of personal language. I carried out a study to investigate how people feel about the privacy of their public online posts, and found that many people do not view these posts as being as private as much previous research suggests. However, it did become clear that most people would rather be asked for permission to use their online posts, and that anonymisation was very important to most people. This chapter represents a great contribution to the corpus creation community. As the literature showed, this will be the biggest and most diverse corpus of type-A e-language which has been made publicly available, and so a detailed account of its design and construction is vital, both for users of the corpus, and for others who would also like to create a corpus of this kind.

Chapter 8 gave an account of the design and construction of the miscellaneous and written-to-be-spoken mediums of the corpus. I demonstrated that these mediums were straightforward to design, but that the miscellaneous medium underwent many alterations when compared to the sampling frame. These changes were justified, and the impact on the corpus explained.

Chapter 9 demonstrated the utility of the corpus to the research community by detailing an in depth study using some of the data from the corpus. I investigated the potential colloquialisation of academic British English, and found that, in some aspects at least, academic British English is certainly becoming more colloquial. These findings were backed up by robust statistical significance testing. The study gave a glimpse into the huge array of things which the corpus will certainly be used to research.

10.2 Successes, limitations, and future directions

In pursuing the aims of this thesis, I have achieved some notable successes. Clearly the biggest of these is the creation of a brand new corpus of contemporary written British English which I have no doubt will be of great use to the corpus linguistics community and beyond. The corpus meets all of the aims which I set out to meet: it is a synchronic corpus, which includes a range of samples from the full range of written British English, created using a non-opportunistic design, and will be made available for general use. This means that the corpus is highly likely to be extremely widely used by the research community, as it fills a gap left by other contemporary corpora (discussed in chapter 1). Additionally, this detailed record of the decisions made in the design and creation of the corpus will be essential for researchers using the corpus, but is not something typically available for many corpora. The ability to assess the suitability of the corpus for any given study using the details within this thesis will be invaluable for the research which will be carried out.

In addition to the creation of the corpus, I was also successful in updating and improving several aspects of the Written BNC1994 in the Written BNC2014. Notably, the addition of a substantial e-language medium in the corpus means that the data is as representative of contemporary written British English as is possible, while still being comparable to the BNC1994. Furthermore, I have improved upon Lee's (2001) genre scheme for the BNC1994. After critically reviewing the genre scheme I was able to make changes, such as categorising academic books and journals separately rather than together, and removing some miscellaneous genres which weren't really miscellaneous at all in reality (see discussion in section 8.6). Additionally, many of the names given to the genre categories in Lee's (2001) scheme have been updated to better reflect the data contained within that genre.

Despite returning throughout the thesis to the idea of compromising between what is ideal and what is possible, another major success which has been achieved is that the eventual corpus did in fact end up being very similar to the sampling frame. Of course, some changes were made, but overall the shape of the corpus is almost exactly as planned. The amount of books collected perfectly matches the sampling frame, and the amounts of periodicals and e-language have only increased slightly due to some compromises being made in the miscellaneous medium. The quality of the data collected was good, with compromises rarely being made. Where compromises were made, such as collecting only open-access academic books, this was due to a successful navigation of UK copyright law. Furthermore, this compromise was not one which compromised the quality of the corpus at all. The books had been through the same production processes as printed books, but had been released as open access rather than for sale in print form. Throughout the project, I dealt with the challenges presented by UK copyright law and successfully managed to find ways to collect all of the data types needed for the corpus. Overall, this success means that the goal of maximising the representativeness of the corpus as much as possible has been met, and thus, the corpus can certainly be used to investigate the written British English of the 2010s.

A further major success which I achieved in this project is the design of a bespoke sampling frame for the corpus, with careful thought given to all aspects of design, sampling, and data collection. The design of this sampling frame was based on extensive research into both other corpus projects, and empirical research regarding specific sampling issues. Without the sampling frame the corpus could not have achieved the goal of being non-opportunistic, and its representativeness of the population would have been greatly reduced.

Whilst I am confident that the Written BNC2014 will be of use to many, there are of course aspects of the project where compromises had to be made, resulting in some limitations. Firstly, although the eventual corpus is very similar to the sampling frame, it is not identical. Some genres could not be collected (e.g. letters), some genres could not be categorised as hoped (e.g. non-academic non-fiction books), and some genres were collected in different proportions than planned (e.g. newspaper articles). However, very often these changes actually reflected the reality of the population I was seeking to represent, and so the limitation caused is not too great, and in some cases the representativeness of the population is actually increased.

One of the major limitations encountered when collecting a lot of genres of data was the difficulty of identifying the native language of the author (academic books, journal articles, newspaper articles, blogs, discussion forums etc.). For many data types this information simply was not available, and so methods were employed to increase the likelihood of an author being British, but could not guarantee this. Clearly, this limits the representativeness of *British* English in the corpus. However, it is also the case that, whilst not all of the language in the corpus can be guaranteed to be representative of the language *produced* by British writers, it is certainly representative of the language *received* by British English readers.

As well as reflecting critically on the work which I did complete, it is also important to mention future work which needs to be, or which I hope will be, carried out on the project. At the time of writing, some final data (mostly from the e-langauge and miscellaneous genres) needs to be collected for the corpus, and this data then needs to be converted to a format which is usable in a corpus. The texts in the corpus will be xml tagged and POS tagged using the CLAWS7[10] tag-set. The corpus will then be made generally available for use. This is expected to be done by the end of 2019.

In terms of future research using the corpus, the first thing which I would like to do is replicate the study detailed in chapter 9 with the full academic data-set in the final corpus. This will allow me to confirm that the results found in chapter 9 hold true when all of the data is studied. I would also like to carry out similar investigations on other genres within the corpus, in order to compare how colloquialisation is affecting

[10] http://ucrel.lancs.ac.uk/claws7tags.html

different types of texts. It would also be interesting to revisit Biber et al. (1999), which most of the features related to colloquialisation were based on. Biber et al. (1999) identify whether certain linguistic features are more common in speech or various types of writing. It would be extremely interesting to search for some of these linguistic features in the Spoken and Written BNC2014 in order to find out whether the results found by Biber et al. (1999) are still true.

It will also become possible to use the corpus to answer some of the questions which many corpus linguists have hypothesised about, and which have been raised in this thesis. For example, chapter 7 highlighted the debate around whether e-language is more like speech or more like writing. Using the corpus, alongside the spoken BNC2014, it will now be possible to compare e-language to both contemporary spoken and written language in order to draw conclusions about this. The ability to use a diverse and contemporary data-set to answer these theoretical questions qill be invaluable.

The impact that the Written (and Spoken) BNC2014 is likely to have on the linguistics community is huge and widespread. As shown in chapter 1 of this thesis, the BNC1994 is many linguists first choice when a data-set of British English is required. However, the outdated nature of the corpus means that any conclusions drawn from such research cannot be generalised to contemporary language. I fully expect that the corpus will be quickly embraced by the linguistics community, and used for a variety of both diachronic comparisons, and also novel research on contemporary British English.

In the long-term, it is interesting to consider whether it will be possible, or indeed productive, to create another BNC in twenty years time – the BNC2034. Based on the success of this project, I certainly think that it would be possible, although perhaps even more time consuming as copyright laws become stricter. It was the case on this project that the only data collection which was significantly slowed by copyright restrictions was the collection of books (see chapter 5). However, with copyright holders becoming increasingly concerned about protecting the commercial value of their copyrights, it is likely that these laws may become harder to navigate for multiple genres of text. Indeed, during the project it became harder to collect some types of data – early in the project book extracts were available on line but, by the time we were ready to collect them, the downloading of these extracts had been made impossible. On the other hand, open-access publishing was an invaluable source of data in the present corpus, and, of course, runs against the trend of more constraints being put on text usage. This trend towards open-access publishing may mean that, in twenty years time, some sections of the corpus, particularly academic writing, will be fairly easy to collect. It may be asked, if it were possible to create another BNC corpus, what would this need to look like in order to be a productive linguistic resource? I strongly believe that the maximisation of the representativeness of the corpus should be prioritised in its design, and that backward comparability with the older corpora should be a secondary concern. As language inevitably evolves, I think that this project has shown that representing this, rather than sticking to an old and potentially out-dated model is the best way to create a maximally useful resource.

10.3 Summary

The design, compilation, and release of the Written BNC2014 is a much anticipated moment in the study of British English and beyond. The main contribution of this thesis, and the project itself, is evident: a brand new, generally available corpus of contemporary Written British English, which is comparable to the Written BNC1994.

Looking back over this project, the number of issues which I had to take into account, and the complexity of these, when designing and compiling the corpus was huge. Even something as seemingly simple as the initial decision of how to define the population to be represented was an issue which required much thought and research (see sections 3.2.2.1 and 4.3.1). The task of data collection was also complex and extremely varied; rarely could any single method of data collection be applied to more than one genre of text. This work has given me a new appreciation of previously created corpora, particularly the Written BNC1994, whose creators had to tackle all of these issues but often without the background of literature which was available to me, or the level of technology now available. However, it has also made me more cautious about using other corpora – now that I am aware of just how many issues must be resolved in the creation of a corpus, it highlights that for many corpora I have used in the past I have been unaware of how most of these were dealt with. This means that I do not know what affect this has had on the suitability of the corpus for the purposes I am employing it. In turn, this emphasises the major contribution to the field made by this thesis itself; detailed documentation of all of the decisions made in the design and compilation of the corpus will only serve to make the corpus even more useful for the research community, for as many purposes as possible. Overall, I feel extremely fortunate to have played a role in the creation of the Written BNC2014.

References

American National Corpus Project. (2015). *ANC Second Release.* Retrieved from: http://www.anc.org/data/anc-second-release/anc-second-release-contents/. Accessed on: 23/08/2018.

AOIR. (2012). *Ethical decision-making and internet research: Recommendations from the AoIR ethics working committee.* Retrieved from: http://aoir.org/reports/ethics2.pdf. Accessed on: 18/03/2015.

Argan, M.T., Argan, M. & Suher, I.K. (2011). Emergence of virtual communities as a means of communication: A case study on virtual health care communities. *Turkish Online Journal of Distance Education, 12*(3), 277-294.

Arkhangelskiy, T.A. (2012). Electronic corpora of the Albanian, Kalmyk, Lezgian, and Ossetic languages. *Automatic Documentations and Mathematical Linguistics, 46*(2), 118-123.

Aroonmanakun, W. (2007). Creating the Thai National Corpus. *Manusaya, 13,* 4-17.

Aroonmanakun, W., Tansiri, K. & Nittayanuparp, P. (2009). Thai National Corpus: A progress report. In: H. Riza & V. Sornlertlamvanich (Eds.), Proceedings of the 7th Workshop on Asian Language Resources (ALR7), (pp. 153-160). Suntec, Singapore: Association for Computational Linguistics.

Arora, S., Li, Y., Youtie, J. & Shapira, P. (2015). Using the wayback machine to mine websites in the social sciences: A methodological resource. *Journal of the Association for Information Science and Technology, 67*(8), 1904-1915.

Atkins, S., Clear, J. & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing, 7*(1), 1-16.

BAAL – The British Association for Applied Linguistics. *Recommendations on Good Practice in Applied Linguistics.* Retrieved from: http://www.baal.org.uk/dox/goodpractice_full.pdf. Accessed on: 28/07/2015.

Baker, P. (2009). The BE06 corpus of British English and recent language change. *International Journal of Corpus Linguistics, 14*(3), 312-337.

Baker, P. (2011). Times may change, but we will always have money: Diachronic variation in recent British English. *Journal of English Linguistics, 39*(1), 65-88.

Baker, P. (2017). *American and British English: Divided by a common language?* Cambridge: Cambridge University Press.

Baron, A., Rayson, P., Greenwood, P., Walkerdine, J. & Rashid, A. (2012). Children online: A survey of child language and CMC corpora. *International Journal of Corpus Linguistics, 17*(4), 443-481.

Baroni, M. & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In: M. T. Lino, M.F. Xavier, F. Ferreira, R. Costa & R. Silva (Eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation,* (1313-1316). Lisbon: ELRA.

Bartley, L. & Benitez-Castro, M.A. (2013). Accuracies and inaccuracies in EFL learners' written vocabulary use. *Spanish Journal of Applied Linguistics, 26,* 45-65.

Bauer, M. & Aarts, B. (2000). Corpus construction: A principle for qualitative data collection. In: M. Bauer & G. Gaskell (Eds.), *Qualitative researching with text, image and sound*, (19-37). London: Sage Publications.

Baym, N. (1995). The performance of humour on computer-mediated communication. *Journal of Computer-Mediated Communication, 1*(2).

Beißwenger, M. Ermakova, M., Geyken, A., Lemnitzer, L. & Storrer, A. (2013). DeRiK: A German reference corpus of computer-mediated communication. *Literary and Linguistic Computing, 28*(4), 531-537.

Berg, T. (2011). A diachronic frequency account of the allomorphy of some grammatical markers. *Journal of Linguistics, 47*(1), 31-64.

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

Biber, D. (1989). A typology of English texts. *Linguistics, 27*(1), 3-43.

Biber, D. (1990). Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing, 5*(4), 257-269.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*(4), 243-257.

Biber, D. & Conrad, S. (2009). *Register, genre, and style.* Cambridge: Cambridge University Press.

Biber, D & Finegan, E. (1989). Drift and the evolution of English style: A history of three genres. *Language: Journal of the Linguistic Society of America, 65*(3), 487-517.

Biber, D. & Gray, B. (2011). Grammatical change in the noun phrase: The influence of written language use. *English Language and Linguistics, 15*(2), 223-250.

Biber, D., Finegan, E. & Atkinson, D. (1994). ARCHER and its challenges: Compiling and exploring A Representative Corpus of Historical English Registers. In: U. Fries, P. Schneider & G. Tottie (Eds.), *Creating and using English language corpora. Papers from the 14th International Conference on English Language Research on Computerized Corpora, Zurich 1993,* (1-13). Amsterdam: Rodopi.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman grammar of spoken and written English.* London: Longman.

Biber, D., Egbert, J. & Davies, M. (2015). Exploring the composition of the searchable web: a corpus-based taxonomy of web registers. *Corpora, 10*(1), 11-45.

BNC Document Register (1991). *Corpus Design Specification*. Retrieved from: http://www.natcorp.ox.ac.uk/archive/vault/tgaw04.pdf. Accessed on: 17/12/2018.

Breul, C. (2014). The perfect participle paradox: some implications for the architecture of grammar. *English Language and Linguistics, 18*(3), 449-470.

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide.* Cambridge: Cambridge University Press.

Brezina, V. & Gablasova, D. (2015). Is there a Core General Vocabulary? Introducing the "New General Service List". *Applied Linguistics, 36*(1), 1-22.

Broccias, C. & Smith, N. (2010). Same Time, across Time: Simultaneity clauses from Late Modern to Present-day English. *English Language and Linguistics, 14*(3), 347-371.

Buil, I., Hernandez, B., Sese Fj. & Urquizu, P. (2012). Discussion forums and their benefits for e-learning: Implications for effective use. *Innovar-Revista De Ciencias Administrativas Y Sociales, 22*(43), 131-143.

Bungarten, T. (1979). Das Korpus als empirische Grunlage in der Linguistik und Literaturwissenschaft. In: H. Bergenholtz & B. Schaeder (Eds.). *Empirische*

Textwissenschaft: Ausbau und Auswertung von Text-Corpora, (28-51). Königstein: Scriptor.

Burnard, L. (2000). *Reference guide for the British National Corpus (World Edition)*. Retrieved from: http://www.natcorp.ox.ac.uk/archive/worldURG/urg.pdf. Accessed on: 02/07/2015.

Burnard, L. (2002). A retrospective look at the British National Corpus. In: B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus analysis*, (pp.51-72). Amsterdam – New York: Rodopi.

Burnard, L. (2007). BNC User Reference Guide. Retrieved from: http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html. Accessed on: 31/10/17.

Burns, N., Bi, Y., Wang, H. & Anderson, T. (2011). Sentiment analysis of customer reviews: Balanced versus unbalanced datasets. *Lecture Notes in Computer Science, 688*(1), 161-170.

Cheng, W., Warren, M. & Xu, X. (2003). The language learner as language researcher: Putting corpus linguistics on the timetable. *System: An international journal of educational technology and applied linguistics, 31*(2), 173-186.

Choudhury, M., Saraf, R., Jain, V., Mukherjee, A., Sarkar, S. & Basu, A. (2007). Investigation and modeling of the structure of texting language. *International Journal on Document Analysis and Recognition, 10* (3), 157–174.

Chujo, K. & Utiyama, M. (2006). Selecting level-specific specialized vocabulary using statistical measures. *System: An international educational journal of educational technology and applied linguistics, 34*(2), 255-269.

Craswell, N. (2005). *W3C test collection.* Retrieved from: http://research.microsoft.com/en-us/um/people/nickcr/w3c-summary.html. Accessed on: 10/03/2015.

Crystal, D. (2008a). *A Dictionary of linguistics and phonetics.* Oxford: Blackwell.

Crystal, D. (2008b). Texting. *ELT Journal, 62*(1), 77-83.

Cunha, E., Magno, G., Gonçalves, M.A., Cambraia, C. & Almeida, V. (2014). He votes or she votes? Female and male discursive strategies in Twitter political hashtags. *PloS One, 9*(1), 1-8.

Davies, M. (2009). The 365+ million word Corpus of Contemporary American English (1990-2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics, 14*(2), 159-190.

Davies, M. (2013a). *Corpus of Global Web-based English.* Retrieved from: http://corpus.byu.edu/glowbe/. Accessed on: 11/03/2015.

Davies, M. (2013b). Google Scholar and COCA-Academic: Two very different approaches to examining academic English. *Journal of English for Academic Purposes, 12*(3), 155-165.

Deloitte. (2014). *Short messaging services versus instant messaging: Value versus volume*. Retrieved from: https://www2.deloitte.com/content/dam/Deloitte/au/Documents/technology-media-telecommunications/deloitte-au-tmt-short-messaging-services-versus-instant-messaging-011014.pdf. Accessed on: 16/03/2015.

Deutschmann, M., Ädel, A., Garretson, G. & Walker, T. (2009). Introducing Mini-McCALL: A pilot version of the Mid-Sweden corpus of computer-assisted language learning. *ICAME Journal, 33*, 21–44.

Dilts, P. & Newman, J. (2006). A note on quantifying ‘good’ and ‘bad’ prosodies. *Corpus Linguistics and Linguistic Theory, 2*(2), 233-242.

Dubrow, H. (1982). *Genre.* London: Methuen.

Duff, D. (1999). *Modern Genre Theory.* London: Longman.

Durrani, A. (2015). Magazines ABCs: Top 100 at a glance. *MediaWeek.* Retrieved from: http://www.mediaweek.co.uk/article/1333599/magazines-abcs-top-100-glance. Accessed on: 17/02/2016.

El Haj, M., Alves, P., Rayson, P., Walker, M. & Young, S. (2017). Retrieving, classifying and analysing narrative commentary in unstructured (Glossy) annual reports published as PDF files. Retrieved from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2803275. Accessed on: 18/10/2018.

Elsevier. *Open access licenses.* Retrieved from: https://www.elsevier.com/about/policies/open-access-licenses. Accessed on: 18/07/2018.

eMarketer. (2014). *More than one-fifth of UK consumers use Twitter.* Retrieved from: http://www.emarketer.com/Article/More-than-One-Fifth-of-UK-Consumers-Use-Twitter/1010623. Accessed on: 12/03/2015.

en:cnk:uvod. (2017). In: Příručka ČNK. Retrieved from: https://wiki.korpus.cz/doku.php?id=en:cnk:uvod&rev=1513073509. Accessed on: 21/08/2018.

Engels, L.K., Beckhoven, B.V., Leenders, T. & Brasseur, I. (1981). *L.E.T. Vocabulary-List*. Leuven: Acco.

Erman, B. (2014). There is no such thing as a free combination: a usage-based study of specific construals in adverb-adjective combinations. *English Language and Linguistics, 18*(1), 109-132.

EUR-Lex. (2016). Regulation (EU) 2016/679 of the European Parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Retrieved from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv%3AOJ.L_.2016.119.01.0001.01.ENG&toc=OJ%3AL%3A2016%3A119%3ATOC. Accessed on: 18/10/2018.

Faulkner, X. & Culwin, F. (2005). When fingers do the talking: a study of text messaging. *Interacting with Computers*, *17*(2), 167–85 .

Ferraresi, A., Zanchetta, E., Baroni, M. & Bernardini, S. (2008, June 01). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In: S. Evert, A. Kilgarriff and S. Sharoff (Eds.), *Proceedings of the 4th Web as Corpus Workshop (WAC-4) – Can we beat Google?* pp. 47-54.

Forsyth, E.N. & Martell, C.H. (2007, September 19-26). Lexical and discourse analysis of online chat dialog. In: *International Conference on Semantic Computing (ICSC 2007)*. Irvine, California, pp. 19-26.

Fowler, A. (1982). *Kinds of Literature: An Introduction to the Theory of Genres and Modes.* Oxford: Clarendon.

Francis, W. M. & Kučera, H. (1979). Brown Corpus Manual, revised version. Providence, Rhode Island: Brown University. Retrieved from: http://www.hit.uib.no/icame/brown/bcm.html. Accessed on: 31/10/17.

Frow, J. (2006). *Genre.* Abingdon: Routledge.

Ghani, R., Jones, R., Mladenic, D. (2005). Building minority language corpora by learning to generate web search queries. *Knowledge and Information Systems, 7*(1), 56-83.

Gov.uk (2017). *Intellectual property – guidance. Exceptions to copyright*. Retrieved from: https://www.gov.uk/exceptions-to-copyright. Accessed on: 18/07/2018.

Gregory, M. (1967). Aspects of varieties differentiation. *Journal of Linguistics, 3*(2), 177-198.

Grinter, R.E. & Eldridge, M. (2003, April). Wan2tlk? Everyday text messaging. In: *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)* (441–8).

Hardaker, C. (2012). *Trolling in computer-mediated communication: Impoliteness, desception, and manipulation online.* (Unpublished thesis). Lancaster University, UK.

Harriman, S. & Patel, J. (2014). The ethics and editorial challenges of internet-based research. *Bmc Medicine, 12*, 59-66.

Harris, A.J. & Jacobson, M.D. (1972). *Basic Elementary Reading Vocabularies*. New York: Macmillan.

Herdadelen, A. (2013). Twitter n-gram corpus with demographic metadata. *Language Resources and Evaluation, 47*(4), 1127-1147.

Herring, S. (1996). Linguistic and critical analysis of computer-mediated communication: Some ethical and scholarly considerations. *The Information Society: An International Journal, 12*(2), 153-168.

Hoffmann, S. (2007). Processing internet-derived text: Creating a corpus of Usenet messages. *Literary and Linguistic Computing, 22*(2), 151-165.

Horn, C., Pimas, O., Granitzer, M., Lex, E. & Graz, K.C. (2011, November 15-18). Realtime ad hoc search in Twitter: know-center at TREC Microblog Track 2011. In: *Proceedings of the twentieth Text ReEtrieval Conference (TREC 2011)*. Gaithersburg, Maryland.

How, Y. & Kan, M.Y. (2005, July). Optimizing predictive text entry for short message service on mobile phones. In: M. J. Smith & G. Salvendy (Eds.), *Proceedings of Human Computer Interfaces International 2005 (HCII 05)*. Las Vegas: Lawrence Erlbaum Associates.

Hsu, W. (2013). Bridging the vocabulary gap for EFL medical undergraduates: The establishment of a medical word list. *Language Teaching Research, 17*(4), 454-484.

Hundt, M. (1997). Has BrE been catching up with AmE over the past 30 years? In: M. Ljung (Ed.), *Corpus-based Studies in English: Papers from the 17th International Conference on English Language Research on Computerized Corpora (ICAME 17),* (pp. 135-151). Amsterdam: Rodopi.

Hunston, S. (2008). Collection strategies and design decisions. In: A. Ludeling & M. Kyto (Eds.), *Corpus linguistics: An international handbook*, (154-167). Berlin, New York, NY: Walter De Gruyter.

Hyland, K. (2009). *Teaching and researching writing.* Harlow: Longman.

ICAME36 (2015). *Words, words, words – Corpora and lexis, ICAME36 Abstract book*. University of Trier.

Ide, N. (2008). The American National Corpus: Then, Now, and Tomorrow. In: M. Haugh, K. Burridge, J. Mulder & P. Peters (Eds.), *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages,* (pp.108-113). Sommerville, MA: Cascadilla Proceedings Project.

Institut Für Deutsche Sprache. (2018). Development and Maintenance of Contemporary Written Corpora. Retrieved from: http://www1.ids-mannheim.de/direktion/kl/projekte/korpora.html?L=1. Accessed on: 24/08/2018.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. & Suchomel, V. (2013). The TenTen Corpus Family. In: A. Hardie & R. Love (Eds.), *Corpus Linguistics 2013 Abstract Book* (pp. 125-127). Lancaster: UCREL.

Johansson, S. (1985). Word frequency and text type: Some observations based on the LOB corpus of British English texts. *Computers and the Humanities, 19*(1), 23-36.

Kang, B. & Kim, H. (2004). Sejong Korean corpora in the making. In: M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, R. Silva, C. Pereira, F. Carvalho, M. Lopes, M. Catarino, & S. Barros (Eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004),* (pp. 1747-1750). Lisbon, Portugal: ELRA.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics, 6*(1), 1–37.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography,* 1, 7-36.

Klimt, B. & Yang, Y. (2004). Introducing the Enron Corpus. In: *Proceedings of CEAS 2004 – First Conference on Email and Anti-Spam*. Mountain View, California, USA (30-31).

Knight, D., Adolphs, S. & Carter, R. (2014). CANELC: constructing an e-language corpus. *Corpora, 9*(1), 29-56.

Koene, A., Adolphs, S., Perez, E., Carter, C.J., Statche, R., O'Malley, C., Rodden, T. & McAuley, D. (2015). Ethics considerations for corpus linguistic studies using internet resources. In: F. Formato & A. Hardie (Eds.), *Corpus Linguistics 2015 Abstract Book*, (pp. 204-206). Lancaster, UK, 2015.

Köhler, R. (2013). Statistical comparability: Methodological caveats. In: S. Sharoff, R. Rapp, P. Zweigenbaum & P. Fung (Eds.), *Building and using comparable corpora*, (pp.77-91). Berlin: Springer.

Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kovářiková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P., Vondřička, P. & Zasina, A. (2016). SYN2015: Representative Corpus of Contemporary Written Czech. In: N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, (pp. 2522-2528). Portorož, Slovenia: ELRA.

Krieg-Holz, U., Schuschnig, C., Matthies, F., Redling, B. & Hahn, U. (2016). CODE-ALLTAG – A German-language e-mail corpus. In: N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, (pp. 2543-2550). Portorož, Slovenia: ELRA.

Kupietz, M. & Lüngen, H. (2014). Recent developments in DeReKo. In: N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J.

Odijk & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, (pp. 2378-2385). Reykjavik, Iceland: ELRA.

Kupietz, M., Belica, C., Keibel, H. & Witt, A. (2010). The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010),* (pp. 1848-1854). Valetta, Malta: ELRA.

Kupietz, M., Lüngen, H., Kamocki, P. & Witt, A. (2018). The German Reference Corpus DeReKo: New developments – new opportunities. In: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis & T. Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018),* (pp. 4353-4360). Miyazaki, Japan: ELRA.

Lee, D. (2001). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology, 5*(3), 37-72.

Le, D. & Quasthoff, U. (2016). Construction and Analysis of a Large Vietnamese Text Corpus. In: N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 412-416). Portorož, Slovenia: ELRA.

Leech, G. (2002). Recent grammatical change in English: Data, description, theory. In: K. Aijmer & B. Altenberg (Eds.), *Advances in Corpus Linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23),* (pp. 61-81). Amsterdam: Rodopi.

Leech, G. (2003). Modals on the move: The English modal auxiliaries 1961-1992. In R. Facchinetti, F.R. Palmer & M.G. Krug (Eds.), *Modality in contemporary English* (pp. 223-240). Berlin/New York: Mouton do Gruyter.

Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. In: M. Hundt, N. Nesselhauf & C. Biewer (Eds.), *Corpus linguistics and the web,* (pp. 133-150). Amsterdam: Rodopi.

Leech, G. & Fallon, R. (1992). Computer corpora — what do they tell us about culture? *ICAME Journal: Computers in English Linguistics, 16,* 29–50.

Legislation.gov.uk (2014). *The Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014*. Retrieved from: http://www.legislation.gov.uk/uksi/2014/1372/regulation/3/made. Accessed on: 18/07/2018.

Lin, P.M.S. (2014). Investigating the validity of internet television as a source for acquiring L2 formulaic sequences. *System, 42*, 164-176.

Liu, B. (2010). Sentiment analysis and subjectivity. In: N. Indurkhya & F.J. Damerau (Eds.), *Handbook of natural language processing. Second edition,* (pp. 627-666). London: Taylor and Francis Group.

Liu, H. (2011). Quantitative properties of English verb valency. *Journal of Quantitative Linguistics, 18*(3), 207-233.

Love, R., Dembry, C., Hardie, A., Brezina, V. & McEnery, T. (2017a). The Spoken BNC2014: designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics, 22*(3), 319-344.

Love, R., Hawtin, A., & Hardie, A. (2017b). *The British National Corpus 2014: User Manual and Reference Guide (version 1.0)*. Lancaster: ESRC Centre for Corpus Approaches to Social Science. Retrieved from: http://corpora.lancs.ac.uk/bnc2014/doc/BNC2014manual.pdf. Accessed on: 14/08/2018.

Mair, C. (1997). Parallel corpora: A real-time approach to the study of language change in progress. In: M. Ljung (Ed.), *Corpus-Based Studies in English: Papers from the Seventeenth International Conference on English Language Research on Computerized Copora (ICAME 17)* (pp. 195-209). Amsterdam: Rodopi.

Mair, C. (2015). Parallel corpora. A real-time approach to the study of language change in progress. *Diacronia, 1,* 1-9.

Mair, C., Hundt, M., Leech, G. & Smith, N. (2003). Short term diachronic shifts in part-of-speech frequencies: A comparison of the tagged LOB and F-LOB corpora. *International Journal of Corpus Linguistics, 7*(2), 245–264.

McEnery, T. & Baker, P. (2005). A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts. *Journal of Language and Politics, 4*(2), 197-226.

McEnery, T. & Hardie, A. (2012). *Corpus Linguistics: Method, theory and practice.* Cambridge: Cambridge University Press.

McEnery, T., Xiao, R. & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London: Routledge.

Millar, N. (2009). Modal verbs in TIME: Frequency changes 1923–2006. *International Journal of Corpus Linguistics, 14* (2), 191–220.

Mollin, S. (2017). The Hansard hazard: gauging the accuracy of British parliamentary transcripts. *Corpora, 2*(2), 187-210.

Moreno, M.A., Grant, A., Kacvinsky, L., Moreno, P. & Fleming, M. (2012). Older adolescents' views regarding participation in Facebook research. *Journal of Adolescent Health, 51*(5), 439-444.

Moreno, R.I.G. (2011). The role of discussion boards in a university blended learning program. *Profile Issues in Teachers' Professional Development, 13*(1), 157-174.

Myers, G. (2010). *Discourse of blogs and wikis.* London: Continuum.

Nencioni, G. (1983 [1976]). Parlato-parlato, parlato-scrit-to, parlato-recitato. In: G. Nencioni, *Di scritto e di parlato. Discorsi Linguistici,* (pp. 126-179). Bologna: Zanichelli.

Norberg, C. (2012). Male and female shame: a corpus-based study of emotion. *Corpora*, *7*(2), 159-185.

Nunan, D. (2008). Exploring genre and register in contemporary English. *English Today: The International Review of the English Language, 24*(2), 56-61.

O'Halloran, K. (2009). Inferencing and cultural reproduction: A corpus-based critical discourse analysis. *Text & Talk: An Interdisciplinary Journal of Language, Discourse & Communication Studies*, *29*(1), 21-51.

Oakes, M. (2003). Contrasts between US and British English of the 1990s. In: E. Oleksy & B. Lewandowska-Tomaszczyk (Eds.), *Research and Scholarship in Integration Processes Poland – USA – EU,* (pp. 213-222). Łódź: Łódź University Press.

Ofcom. (2017). *News consumption in the UK: 2016.* Retrieved from: https://www.ofcom.org.uk/__data/assets/pdf_file/0016/103570/news-consumption-uk-2016.pdf. Accessed on 12/02/2018.

Office for National Statistics. (2018). *Internet access – Households and individuals, Great Britain: 2018.* Retrieved from: https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/bulletins/internetaccesshouseholdsandindividuals/2018#daily-internet-use-has-more-than-doubled-since-2006. Accessed on: 11/09/2018.

Orasan, C. & Krishnamurthy, R. (2002). A corpus-based investigation of junk emails. In: *Proceedings of LREC-2002,* (pp. 1773-1779). Las Palmas, Spain: ELRA.

Otlogetswe, T. (2004). The BNC design as a model for a Setswana language corpus. *Paper presented at the 7th Annual CLUK Research Colloquim. University of Birmingham*, 6-7 January 2004. Retrieved from: https://www.academia.edu/3451597/The_BNC_design_as_a_model_for_a_Setswana_language_corpus. Accessed on: 24/06/2016.

Paltridge, B. (1996). Genre, text type, and the language learning classroom. *ELT Journal, 50*(3), 237-243.

Parks, M. & Floyd, K. (1996). Making friends in cyberspace. *Journal of Computer-Mediated Communication, 1*(4).

Pearce, M. (2008). Investigating the collocational behaviour of MAN and WOMAN in the BNC using Sketch Engine. *Corpora, 3*(1), 1-29.

Pérez-Guerra, J. & Martínez-Insua, A. (2010). Do some genres or text types become more complex than others? In: H. Dorgeloch & A. Wanner (Eds.), *Syntactic variation and genre,* (pp. 111-140). Berlin: Mouton de Gruyter.

Perez-Paredes, P., Sanchez-Tornel, M., Calero, J. & Jiminez, P.A. (2011). Tracking learners' actual uses of corpora: guided vs non-guided corpus consultation. *Computer Assisted Language Learning, 24*(3), 233-253.

Poole, B. (2015). Lord Lucan: 'missing' or 'on the run'? *English Today, 31*(2), 32-37.

Potts, A. & Baker, P. (2012). Does semantic tagging identify cultural change in British and American English? *International Journal of Corpus Linguistics, 17*(3), 295-324.

Puschmann, C. (2009, September 3-6). *Diary or Megaphone? The pragmatic mode of weblogs*. Paper presented at Language in the (New) Media: Technologies and Ideologies: Seattle, Washington, USA.

Ramírez, S.Q. (2015). Syntactic functions of infinitives in English. *International Journal of Applied Linguistics and English Literature, 4*(1), 182-190.

Rehm, G. (2002). Towards automatic Web genre identification: a corpus-based approach in the domain of academia by example of the Academic's Personal Homepage. In: *Proceedings of the 35th Annual Hawaii International Conference on System Sciences 2002,* (pp. 1143-1152).

Reppen, R. & Ide, N. (2004). The American National Corpus: Overall goals and the first release. *Journal of English Linguistics, 32*(2), 105-113.

Reyes, A., Rosso, P. & Veale, T. (2013). A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation, 47*(1), 239-268.

Riordan, M.A. & Kreuz, R.J. (2010). Cues in computer-mediated communication: A corpus analysis. *Computers in Human Behaviour, 26,* 1806-1817.

Rosen, M. (2013). How Genre Theory Saved the World. *Changing English, 20*(1), 3-10.

Santini, M. (2007, August). Automatic genre identification: towards a flexible classification scheme. In: A. MacFarlane, L. Azzopardi & I. Ounis (Eds.), *Proceedings of FDIA 2007 BCS IRSG Symposium: Future Directions in Information Access,* (pp.5-10).

Santini, M. & Sharoff, S. (2009). Web genre benchmark under construction. *Journal for Language Technology and Computational Linguistics, 25* (1), 129–45.

Schler, J., Koppel, M., Argamon, S. & Pennebaker, J. (2006, March). Effects of age and gender on blogging. In: *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, Vol.6, (pp. 199-205).

Schmid, R. (2011, October 4). You never write anymore; well, hardly anyone does. *The Washington Times.* Retrieved from: https://www.washingtontimes.com/news/2011/oct/4/you-never-write-anymore-well-hardly-anyone-does/. Accessed on: 14/12/2018.

Schonefeld, D. (2013). It is… quite common for theoretical predictions to go untested (BNC_CMH). A register-specific analysis of the English go un-V-en construction. *Journal of Pragmatics, 52,* 17-33.

Sha, G. (2010). Using Google as a super corpus to drive written language learning: a comparison with the British National Corpus. *Computer Assisted Language Learning, 23*(5), 377-393.

Sharkey, S., Jones, R., Smithson, J., Hewis, E., Emmens, T., Ford, T. & Owens, C. (2011). Ethical practice in internet research involving vulnerable people: lessons from a self-harm discussion forum study (sharptalk). *Journal of Medical Ethics, 37*(12), 752-758.

Sharoff, S. (2013). Measuring he distance between comparable corpora between languages. In: S. Sharoff, R. Rapp, P. Zweigenbaum & P. Fung (Eds.), *Building and using comparable corpora* (pp. 113-130). Berlin: Springer.

Sharoff, S., Rapp, R. & Zweigenbaum, P. (2013). Overviewing Important Aspects of the Last Twenty Years of Research in Comparable Corpora. In: S. Sharoff, R. Rapp, P. Zweigenbaum & P. Fung (Eds.), *Building and using comparable corpora* (pp. 1-17). Berlin: Springer.

Shirk, J. L., Ballard, H. L., Wilderman, C. C., Phillips, T., Wiggins, A., Jordan, R., McCallie, E., Minarchek, M., Lewenstein, B. V., Krasny, M. E., & Bonney, R. (2012). Public participation in scientific research: A framework for deliberate design. *Ecology and Society, 17*(2), 29.

Siepmann, D., Bürgel, C. & Diwersy, S. (2017). The Corpus de référence du Français contemporain (CRFC) as the first genre-diverse mega-corpus of French. *International Journal of Lexicography, 30*(1), 63-84.

Simaki, V., Simakis, P., Paradis, C. & Kerren, A. (2017). Identifying the authors' national variety of English in social media texts. In: R. Mitkov & G. Angelova (Eds.), *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017,* (pp. 671-678). Varna, Bulgaria: INCOMA Ltd.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Siyanova, A. & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review – Revue Canadienne des langues vivantes, 64*(3), 429-458.

Smith, A. (2015). The emergence of temporal subordinators across inner- and outer-circle varieties of English. In: *Words, words, words – Corpora and lexis, ICAME36 Abstract book,* (pp. 161-162). University of Trier.

Stamatatos, E., Kokkinakis, G. & Fakotakis, N. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics, 26*(4), 471-495.

Statista. (2018a). *Forecast of Facebook user numbers in the United Kingdom (UK) from 2015 to 2022 (in million users).* Retrieved from: https://www.statista.com/statistics/553538/predicted-number-of-facebook-users-in-the-united-kingdom-uk/. Accessed on: 17/10/2018.

Statista. (2018b). *Cumulative total of Tumblr blogs from May 2011 to July 2018 (in millions).* Retrieved from: https://www.statista.com/statistics/256235/total-cumulative-number-of-tumblr-blogs/. Accessed on: 13/09/2018.

Steen, G., Dorst, A.G., Herrmann, J., Kaal, A.A. & Krennmayr, T. (2010). Metaphor in usage. *Cognitive Linguistics, 21*(4), 765-796.

Taavitsainen, I. (2001). Changing conventions of writing: The dynamics of genres, text types, and text traditions. *European Journal of English Studies, 5*(2), 139-150.

Tadić, M. (2002). Building the Croatian National Corpus. In: *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002),* (pp. 441-446). Las Palmas, Canary Islands, Spain: ELRA.

Tagg, C. (2009). *A Corpus Linguistics Study of SMS Text Messaging*. (Unpublished Thesis). University of Birmingham, UK.

Tagg, C. (2011). *The discourse of text messaging: Analysis of text message communication.* New York, NY: Continuum.

Talvensaari, T., Pirkola, A., Järvelin, K., Juhola, M. & Laurikkala, J. (2008). Focused web crawling in the acquisition of comparable corpora. *Information Retrieval 11*(5), 427–445.

Thorndike, E.L. & Lorge, I. (1944). *The Teacher's Word Book of 30,000 Words*. New York: Bureau of Publications Teachers College, Columbia University.

Tottie, G. (1997) Relatively speaking: Relative marker usage in the British National Corpus. In: T. Nevalainen & L. Kahlas-Tarkka (Eds.), *To explain the present: Studies in the changing English language in honour of Matti Rissanen,* (pp.465-481). Helsinki: Société néophilologique.

Trosberg, A. (1997). Text typology: Register, genre and text type. In: Trosberg, A (Ed.), *Text Typology and Translation* (3-23). UK: John Benjamins.

UKCS (2017). *UK Copyright Law*. Retrieved from: http://www.copyrightservice.co.uk/ukcs/docs/edupack.pdf. Accessed on: 18/07/2018.

United Kingdom Intellectual Property Office. (2017). *Copyright, Designs and Patents Act 1988*. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/648444/copyright-designs-and-patents-act-1988.pdf. Accessed on: 18/07/2018.

UoB Guidelines - University of Brighton Health and Social Science, Science and Engineering Research Ethics and Governance Committee Guidelines. *Ethical issues for consideration when conducting internet research.* Retrieved from: http://about.brighton.ac.uk/hss/fregc/Ethical-internet.pdf. Accessed on: 18/03/2015.

Van Bogaert, J. (2010). A constructional taxonomy of I think and related expressions: accounting for the variability of complement-taking mental predicates. *English Language and Linguistics*, *14*(3), 399-427.

Váradi, T. (2001). The linguistic relevance of corpus linguistics. In: P. Rayson, A. Wilson, T. McEnery, A. Hardie & S. Khoja (Eds.), *Proceedings of the Corpus Linguistics 2001 Conference*, (587-593). Lancaster University: UCREL Technical Papers 13.

Wang, S. (2005). Corpus-based approaches and discourse analysis in relation to reduplication and repetition. *Journal of Pragmatics, 37*(4), 505-540.

Weichmann, D. & Kerz, E. (2013). The positioning of concessive adverbial clauses in English: Assessing the importance of discourse-pragmatic and processing-based constraints. *English Language and Linguistics, 17*(1), 1-23.

WPVirtuoso. (2013). *How many blogs are on the internet?* Retrieved from: http://www.wpvirtuoso.com/how-many-blogs-are-on-the-internet/. Accessed on: 13/03/2015.

Xiao, R. (2008). Well-known and influential corpora. In: A. Lüdelig & M. Kytö (Eds.), *Corpus linguistics: An international handbook Volume 1*, (pp. 383-456). Berlin – New York, NY: Walter de Gruyter.

Yamasaki, N. (2008). Collocations and colligations associated with discourse functions of unspecific anaphoric nouns. *International Journal of Corpus Linguistics, 13*(1), 75-98.

Zago, R. (2016). *From Originals to Remakes: Colloquiality in English Film Dialogue over Time.* Roma: Bonanno Editore.

Zappavigna, M. (2012). *Discourse of Twitter and social media.* London: Continuum.

Zhao, X. & Feng, Z. (2014). A dynamic study of English intertextual lexical repetition rates. *Journal of Quantitative Linguistics, 21*(1), 65-84.

Appendices

Appendix A: Acronyms used in this thesis

ACE - Atomic Communicative Event (Leech, 2007)

AH – Dr Andrew Hardie

AmE06 – American English 2006

ANC – American National Corpus

AOIR – Association of Internet Researchers

ARCHER - A Representative Corpus of Historical English Registers

BAAL – British Association for Applied Linguistics

BAWE – British Academic Written English

BE06 – British English 2006

BLOB – the British English 1931 member of the Brown family

BNC1994 – British National Corpus 1994

BNC2014 – British National Corpus 2014

CANELC – Cambridge and Nottingham e-language Corpus

CASS – Centre for Corpus Approaches to Social Sciences

CD – Dr Carmen Dayrell

CMC – Computer mediated communication

COCA – Corpus of Contemporary American English

CRFC - Corpus de référence du français contemporain

CUP – Cambridge University Press

DeReKo - Deutsches Referenzkorpus

DeRiK - German Reference Corpus of Internet-based Communication

E-langauge – electronic language

FLOB – Freiburg-LOB corpus. The British English 1991-1992 member of the Brown family.

FROWN – Freiburg-Brown corpus. The American English 1991-1992 member of the Brown family.

GDPR – General Data Protection Regulation

IM – Instant message

L1 – First language

L2 – Second language

LOB – Lancaster-Oslo-Bergen corpus. The British English 1961 member of the Brown family.

MG – Mathew Gillings

MT – Dr Matt Timperley

OCR – Optical Character Recognition

PPSR – Public Participation in Scientific Research

SMS – Short Message Service

SYN2015 – A corpus if contemporary written Czech

TM – Professor Tony McEnery

TNC – Thai National Corpus

Type-A e-language – e-language which can only be found online

Type-B e-language – e-language which may also be found offline

ukWaC – UK web-as-corpus

UoB – University of Brighton

VB – Dr Vaclav Brezina

XML – eXtensible Markup Language

Appendix B: The Written BNC2014 sampling frame

Medium	Super Genre	Genre	Target %	Words
Books (41%)	Academic prose (textbooks, academic books etc.)	W_ac_book_humanities_arts	1%	900,000
		W_ac_book_medicine	1%	900,000
		W_ac_book_nat_science	1%	900,000
		W_ac_book_polit_law_edu	1%	900,000
		W_ac_book_soc_science	1%	900,000
		W_ac_book_tech_engin	1%	900,000
	Fiction	W_fict_poetry	2%	1,800,000
		W_fict_prose_general	9%	8,100,000
		W_fict_prose_childrens	2%	1,800,000
		W_fict_prose_teen	2%	1,800,000
		W_fict_prose_sf_fantasy	2%	1,800,000
		W_fict_prose_crime	2%	1,800,000
		W_fict_prose_romance	2%	1,800,000
	Non-academic prose (non-fiction)	W_non_ac_humanities_arts	2%	1,800,000
		W_non_ac_medicine	2%	1,800,000
		W_non_ac_nat_science	2%	1,800,000
		W_non_ac_polit_law_edu	2%	1,800,000
		W_non_ac_soc_science	2%	1,800,000
		W_non_ac_tech_engin	2%	1,800,000
		W_non_ac_biography	2%	1,800,000
Periodicals (35%)	Academic Prose (journal articles)	W_ac_journal_humanities_arts	1%	900,000
		W_ac_journal_medicine	1%	900,000
		W_ac_journal_nat_science	1%	900,000
		W_ac_journal_polit_law_edu	1%	900,000
		W_ac_journal_soc_science	1%	900,000
		W_ac_journal_tech_engin	1%	900,000
	Broadsheet national newspapers	W_newsp_brdsht_nat_arts_ent	1%	900,000
		W_newsp_brdsht_nat_commerce	1%	900,000
		W_newsp_brdsht_nat_editorial	1%	900,000
		W_newsp_brdsht_nat_reportage	1%	900,000
		W_newsp_brdsht_nat_science	1%	900,000
		W_newsp_brdsht_nat_social	1%	900,000
		W_newsp_brdsht_nat_sports	1%	900,000
	Regional & local newspapers	W_newsp_other_arts_ent	1%	900,000
		W_newsp_other_commerce	1%	900,000
		W_newsp_other_editorial	1%	900,000
		W_newsp_other_reportage	1%	900,000
		W_newsp_other_science	1%	900,000

		W_newsp_other_social	1%	900,000
		W_newsp_other_sports	1%	900,000
	Tabloid newspapers	W_newsp_tabloid_arts_ent	1%	900,000
		W_newsp_tabloid_commerce	1%	900,000
		W_newsp_tabloid_editorial	1%	900,000
		W_newsp_tabloid_reportage	1%	900,000
		W_newsp_tabloid_science	1%	900,000
		W_newsp_tabloid_social	1%	900,000
		W_newsp_tabloid_sports	1%	900,000
	Magazines	W_magazines_lifestyle	1%	900,000
		W_magazines_mens_lifestyle	1%	900,000
		W_magazines_TV_film	1%	900,000
		W_magazines_motoring	1%	900,000
		W_magazines_food	1%	900,000
		W_magazines_music	1%	900,000
		W_magazines_science_tech	1%	900,000
		W_magazine_sports	1%	900,000
E-language (10%)	E-language	W_e_tweet	1.8%	1,620,000
		W_e_blog_news	0.2%	180,000
		W_e_blog_sport	0.2%	180,000
		W_e_blog_opinion	0.2%	180,000
		W_e_blog_personal	0.2%	180,000
		W_e_blog_informational	0.2%	180,000
		W_e_blog_travel	0.2%	180,000
		W_e_discussion_forum	1.3%	1,170,000
		W_e_email_prof	0.8%	720,000
		W_e_email_personal	0.8%	720,000
		W_e_SMS_IM	1.7%	1,530,000
		W_e_review	1.2%	1,080,000
		W_e_comment	1.2%	1,080,000
Miscellaneous (10%)	Essays	W_essay_sch	1%	900,000
		W_essay_univ	1%	900,000
	Letters	W_letters_personal	1%	900,000
		W_letters_prof	1%	900,000
		W_admin	1%	900,000
		W_advert	1%	900,000
		W_commerce	1%	900,000
		W_institutional	1%	900,000
		W_instructional	1%	900,000
		W_religion	1%	900,000
Written-to-be-spoken (4%)	Written-to-be-spoken	W_news_script	2%	1,800,000
		W_fict_drama	2%	1,800,000

Appendix C: The composition of the Written BNC2014

Medium	Super Genre	Genre	Target %	Words
Books (41%)	Academic prose (textbooks, academic books etc.)	W_ac_book_humanities_arts	1%	900,000
		W_ac_book_medicine	1%	900,000
		W_ac_book_nat_science	1%	900,000
		W_ac_book_polit_law_edu	1%	900,000
		W_ac_book_soc_science	1%	900,000
		W_ac_book_tech_engin	1%	900,000
	Fiction	W_fict_poetry	0.11%	100,000
		W_fict_prose_general	10.89%	9,800,000
		W_fict_prose_childrens_teen	4%	3,600,000
		W_fict_prose_sf_fantasy	2%	1,800,000
		W_fict_prose_crime	2%	1,800,000
		W_fict_prose_romance	2%	1,800,000
	Non-academic prose (non-fiction)	W_non_ac_book_general	14%	12,600,000
Periodicals (38%)	Academic Prose (journal articles)	W_ac_journal_humanities_arts	1%	900,000
		W_ac_journal_medicine	1%	900,000
		W_ac_journal_nat_science	1%	900,000
		W_ac_journal_polit_law_edu	1%	900,000
		W_ac_journal_soc_science	1%	900,000
		W_ac_journal_tech_engin	1%	900,000
	Serious newspapers	W_newsp_serious_arts_ent	0.98%	885,600
		W_newsp_serious_commerce_business	1.97%	1,771,200
		W_newsp_serious_editorial	0.39%	354,240
		W_newsp_serious_reportage	3.74%	3,365,280
		W_newsp_serious_science	0.12%	106,272
		W_newsp_serious_lifestyle	1.14%	1,027,296
		W_newsp_serious_sports	1.50%	1,346,112
	Regional & local newspapers	W_newsp_regional_arts_ent	0.15%	142,560
		W_newsp_regional_commerce_business	0.45%	413,424
		W_newsp_regional_editorial	0.29%	263,736
		W_newsp_regional_reportage	4.68%	4,212,648
		W_newsp_regional_science	0.02%	21,384
		W_newsp_regional_lifestyle	0.24%	220,968
		W_newsp_regional_sports	2.05%	1,853,283

	Mass market newspapers	W_newsp_mass_market_arts_ent	0.18%	168,480
		W_newsp_mass_market_commerce_business	0.16%	146,016
		W_newsp_mass_market_editorial	0.25%	224,640
		W_newsp_mass_market_reportage	3.51%	3,161,808
		W_newsp_mass_market_science	0.01%	5,616
		W_newsp_mass_market_lifestyle	0.06%	56,160
		W_newsp_mass_market_sports	2.11%	1,853,280
	Magazines	W_magazines_lifestyle	1.55%	1,400,000
		W_magazines_mens_lifestyle	1.04%	940,000
		W_magazines_TV_film	0.67%	600,000
		W_magazines_motoring	1.55%	1,400,000
		W_magazines_food	0.06%	55,000
		W_magazines_music	1.55%	1,400,000
		W_magazines_science_tech	1.56%	1,405,000
E-language (11%)	E-language	W_e_microblog	1.8%	1,620,000
		W_e_blog_news	0.1%	90,000
		W_e_blog_sport	0.22%	198,000
		W_e_blog_opinion	0.22%	198,000
		W_e_blog_personal	0.22%	198,000
		W_e_blog_informational	0.22%	198,000
		W_e_blog_travel	0.22%	198,000
		W_e_discussion_forum	1.3%	1,170,000
		W_e_email_prof_personal	1.6%	1,440,000
		W_e_email_advert	1%	900,000
		W_e_IM	1.7%	1,530,000
		W_e_review	1.2%	1,080,000
		W_e_comment	1.2%	1,080,000
Miscellaneous (6%)	Essays	W_essay_sch	1%	900,000
		W_essay_univ	1%	900,000
		W_admin	1%	900,000
		W_institutional	2%	1,800,000
		W_instructional	1%	900,000
Written-to-be-spoken (4%)	Written-to-be-spoken	W_television_script	2%	1,800,000
		W_fict_drama	2%	1,800,000

Appendix D: A comparison of the Written BNC1994 and the Written BNC2014 sampling frame

Genre (Written BNC1994)	Proportion (%)	Genre (Written BNC2014)	Proportion (%)
W: fict: poetry	0.25	W_fict_poetry	2
W: fict: prose	18.24	W_fict_prose_general	9
		W_fict_prose_childrens	2
		W_fict_prose_teen	2
		W_fict_prose_sf_fantasy	2
		W_fict_prose_crime	2
		W_fict_prose_romance	2
W:newsp:brodsht_nat: arts	0.40	W_newsp_brdsht_nat_arts_ent	1
W:newsp:brodsht_nat: commerce	0.49	W_newsp_brdsht_nat_commerce	1
W:newsp:brodsht_nat: editorial	0.12	W_newsp_brdsht_nat_editorial	1
W:newsp:brodsht_nat: misc	1.18		
W:newsp:brodsht_nat: report	0.76	W_newsp_brdsht_nat_reportage	1
W:newsp:brodsht_nat: science	0.07	W_newsp_brdsht_nat_science	1
W:newsp:brodsht_nat: social	0.09	W_newsp_brdsht_nat_social	1
W:newsp:brodsht_nat: sports	0.34	W_newsp_brdsht_nat_sports	1
W: newsp: other: arts	0.27	W_newsp_other_arts_ent	1
W: newsp: other: commerce	0.48	W_newsp_other_commerce	1
		W_newsp_other_editorial	1
W: newsp: other: report	3.11	W_newsp_other_reportage	1
W: newsp: other: science	0.06	W_newsp_other_science	1
W: newsp: other: social	1.31	W_newsp_other_social	1
W: newsp: other: sports	1.18	W_newsp_other_sports	1
W: newsp: tabloid	0.83	W_newsp_tabloid_arts_ent	1
		W_newsp_tabloid_commerce	1
		W_newsp_tabloid_editorial	1
		W_newsp_tabloid_reportage	1
		W_newsp_tabloid_science	1
		W_newsp_tabloid_social	1
		W_newsp_tabloid_sports	1
W:pop_lore	8.42	W_magazines_lifestyle	1

		W_magazines_mens_lifestyle	1
		W_magazines_TV_film	1
		W_magazines_motoring	1
		W_magazines_food	1
		W_magazines_music	1
		W_magazines_science_tech	1
		W_magazine_sports	1
W: essay: school	0.17	W_essay_sch	1
W: essay: univ	0.06	W_essay_univ	1
W: letters: personal	0.06	W_letters_personal	1
W: letters: prof	0.08	W_letters_prof	1
W: admin	0.25	W_admin	1
W: advert	0.63	W_advert	1
W: commerce	4.33	W_commerce	1
W: hansard	1.33	W_institutional	1
W: institut_doc	0.63		
W: instructional	0.5	W_instructional	1
W: religion	1.29	W_religion	1
W: news_script	1.42	W_news_script	2
W: fict: drama	0.05	W_fict_drama	2
		W_e_tweet	1.8
		W_e_blog_news	0.2
		W_e_blog_sport	0.2
		W_e_blog_opinion	0.2
		W_e_blog_personal	0.2
		W_e_blog_informational	0.2
		W_e_blog_travel	0.2
		W_e_discussion_forum	1.3
W: email	0.24	W_e_email_prof	0.8
		W_e_email_personal	0.8
		W_e_SMS	1.7
		W_e_review	1.2
		W_e_comment	1.2
W: non_ac: humanities_arts	4.26	W_non_ac_humanities_arts	2
W: non_ac: medicine	0.57	W_non_ac_medicine	2
W: non_ac: nat_science	2.88	W_non_ac_nat_science	2
W: non_ac: polit_law_edu	5.14	W_non_ac_polit_law_edu	2
W: non_ac: soc_science	4.22	W_non_ac_soc_science	2
W: non_ac: tech_engin	1.39	W_non_ac_tech_engin	2
W: biography	4.05	W_non_ac_biography	2
W: ac: humanities_arts	3.82%		
W: ac: medicine	1.63%		
W: ac: nat_science	1.28%		

W: ac: polit_law_edu	5.35%		
W: ac: soc_science	5.44%		
W: ac: tech_engin	0.78%		
		W_ac_book_humanities_arts	1
		W_ac_book_medicine	1
		W_ac_book_nat_science	1
		W_ac_book_polit_law_edu	1
		W_ac_book_soc_science	1
		W_ac_book_tech_engin	1
		W_ac_journal_humanities_arts	1
		W_ac_journal_medicine	1
		W_ac_journal_nat_science	1
		W_ac_journal_polit_law_edu	1
		W_ac_journal_soc_science	1
		W_ac_journal_tech_engin	1
W: misc	10.51		

Appendix E: Text from the email which was sent to UK publishers

Dear [publisher/copyright officer]

I write on behalf of the British National Corpus 2014 (BNC2014) project team at Lancaster University. The British National Corpus 2014 is a major resource creation exercise to build a 100 million word corpus (a large collection of 'real life' language) of modern-day British English. The BNC2014 will be a world-leading, next-generation resource for the study of the English language. The corpus consists of a Spoken part (recorded and transcribed conversations) and a Written part (books, website, magazines, newspapers, and many other kinds of text).

To collect samples of books for the Written BNC2014, we rely on the generosity of publishers and authors to give us access to texts and the necessary permission to incorporate them into the corpus. We therefore write in the hope that you will be able to allow us the necessary permissions to utilise, and access to the digital text of, 5000 word extracts of any fiction and non-fiction books which you published in 2014, from as wide a spread of genres as possible. If you are able to assist us in this matter, we would ask you to send us these extracts in any widely-used electronic format as is convenient to you. The corpus documentation will fully credit your generous contribution to the project.

I have attached to this email a document which provides more details about the project, and further information about how we prioritise protecting the commercial value of your copyrights.

I am aware that you usually require a permissions application form to be filled in for these types of requests. However, since we would like to include as many extracts as you can give us, rather than having a specific text in mind, I have not filled out the form as yet. However, if you would like me to fill out an adapted version of this form then I will of course be happy to do so.

Thank you very much for taking the time to read this email, and please do let me know if you would like any further information about the project.

Yours sincerely/faithfully,

Abi Hawtin

Appendix F: The document which was sent to UK publishers

The British National Corpus 2014

The *British National Corpus 2014* is a major project to create a 100 million word corpus (a large collection of 'real life' language) of modern-day British English.

The aim of the project is to create a successor to the extremely successful *British National Corpus (BNC)*, created in the early 1990s. The BNC consortium of universities and publishing houses collected samples of all kinds of English, including extracts from books, periodicals, unpublished texts, and spoken conversations. The result is an extremely widely-used standard resource – not only for research into English language and linguistics, but also for English language teaching. However, this crucial resource is now more than twenty years old. Now, a collaboration between Lancaster University and Cambridge University Press has begun to build a present-day companion corpus, the *BNC2014*, allowing this vital work to continue.

The BNC2014 will be a world-leading, next-generation resource for the study of the English language. Users of the corpus will be able to study the contemporary English language via data which truly represents modern British English, rather than data collected in the 1990s. The corpus consists of a Spoken part (recorded and transcribed conversations) and a Written part (books, website, magazines, newspapers, and many other kinds of text).

An appeal for publisher participation in the Written BNC2014 project

To collect **samples of books** for the Written BNC2014 we rely on the generosity of publishers and authors to give us access to texts and the necessary permission to incorporate them into the corpus.

What we need from you

We are asking publishers to provide us with 5,000 word extracts of any fiction and non-fiction books published in 2014, from as wide a spread of genres as possible. You can send us these extracts in any widely-used electronic format.

We also ask you to sign a letter confirming that we have permission to include them in the corpus, and to allow researchers to use these text extracts for non-commercial research (a copy of this letter can be seen in annex 1). If you would prefer to use a different form of agreement that is not a problem – we are happy to utilise any form of agreement that covers the necessary permissions.

The corpus documentation will fully credit your generous contribution to the project.

Intellectual property

We are as anxious as you are to protect the commercial value of your copyrights. We have designed our corpus collection procedure with this in mind:

- We are asking for access to excerpts **no longer than about 5,000 words**. This is one chapter, roughly speaking – that is, about the same length as the "samples" which publishers often release online. Donating a text to our corpus does not endanger your commercial interests any more than such "free samples" do. On the contrary, like such samples, a donated corpus text may even serve to spread awareness of – and thus demand for – your product.

- The corpus will **only** be made available to researchers under a licence (see annex 4) which **forbids any commercial exploitation**. This is *not* a "Creative Commons" licence or similar – it is a far more restrictive, non-commercial, non-transferable licence. The permission letter that we ask you to sign (see annex 1) makes clear that you are only giving us permission to reproduce your book excerpts under the specified licence.

- We will keep a record of all signatories of the licence (that is: everyone who has signed up to use the corpus), and we will make depersonalised, aggregate data on these users available to you on request, in perpetuity. You will always have the option of knowing **how many** and **what kind of** users **have signed up to access the corpus**.

- The corpus **will not include the original text in a "readable" format**. Instead, it will include a linguistically-annotated reformatting of the excerpt in XML. An example of how the data will appear is shown in annex 2.

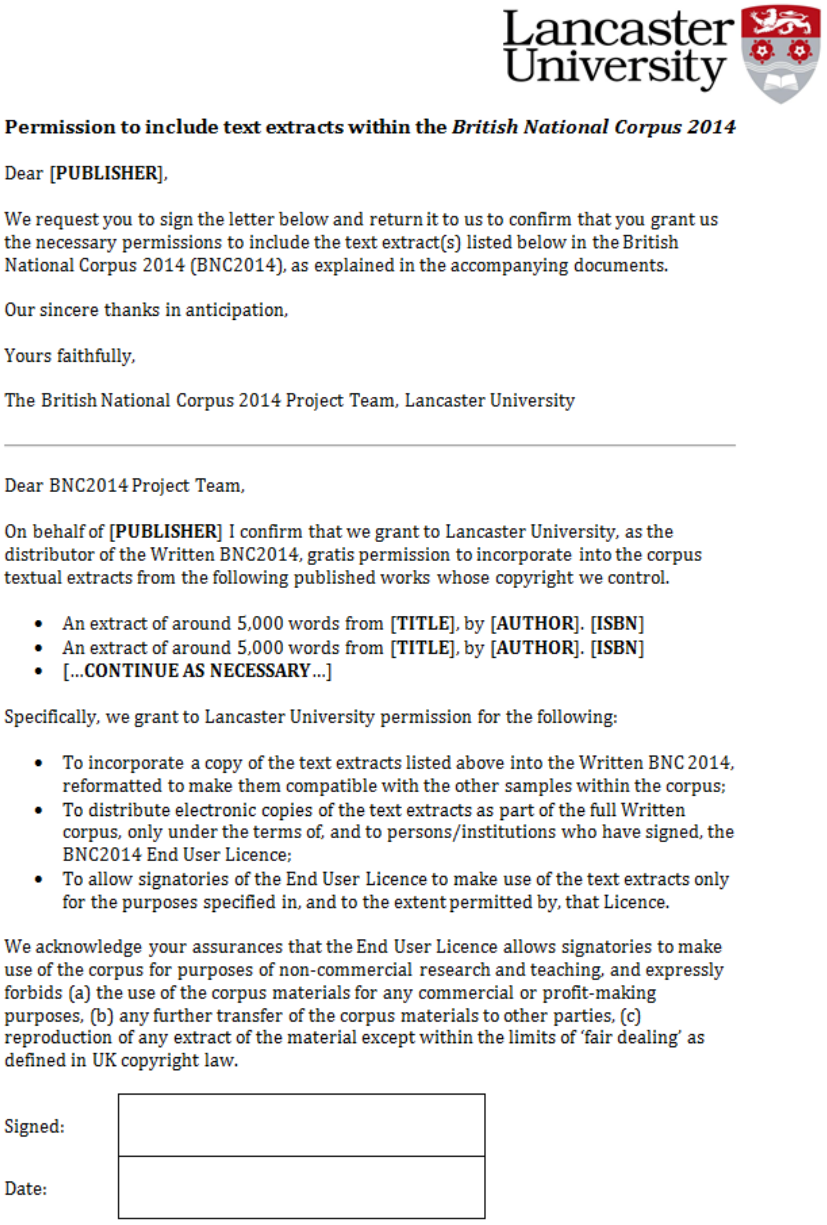
- Many users, rather than getting a full copy of the corpus, will access it via one or more specialised software tools. These software systems limit the amount of access any user has to the underlying "original" text to one or two sentences around the search result. An example of how the data will appear is shown in annex 3.

In short we are confident that there is no possibility that granting us permission to use an excerpt of one of your books in the corpus can endanger sales of the book.

Thank you very much for taking the time to read this document, and we hope that we will be able to work with you to create this world-leading language resource.

Annex 1

Below is a copy of the letter which we ask you to sign, agreeing to allow us to use extracts of your published material in the corpus. (If you would prefer to use a different form of agreement that is not a problem – we are happy to utilise any form of agreement that covers the necessary permissions.)



Lancaster University

Permission to include text extracts within the *British National Corpus 2014*

Dear **[PUBLISHER]**,

We request you to sign the letter below and return it to us to confirm that you grant us the necessary permissions to include the text extract(s) listed below in the British National Corpus 2014 (BNC2014), as explained in the accompanying documents.

Our sincere thanks in anticipation,

Yours faithfully,

The British National Corpus 2014 Project Team, Lancaster University

---

Dear BNC2014 Project Team,

On behalf of **[PUBLISHER]** I confirm that we grant to Lancaster University, as the distributor of the Written BNC2014, gratis permission to incorporate into the corpus textual extracts from the following published works whose copyright we control.

- An extract of around 5,000 words from **[TITLE]**, by **[AUTHOR]**. **[ISBN]**
- An extract of around 5,000 words from **[TITLE]**, by **[AUTHOR]**. **[ISBN]**
- [...**CONTINUE AS NECESSARY**...]

Specifically, we grant to Lancaster University permission for the following:

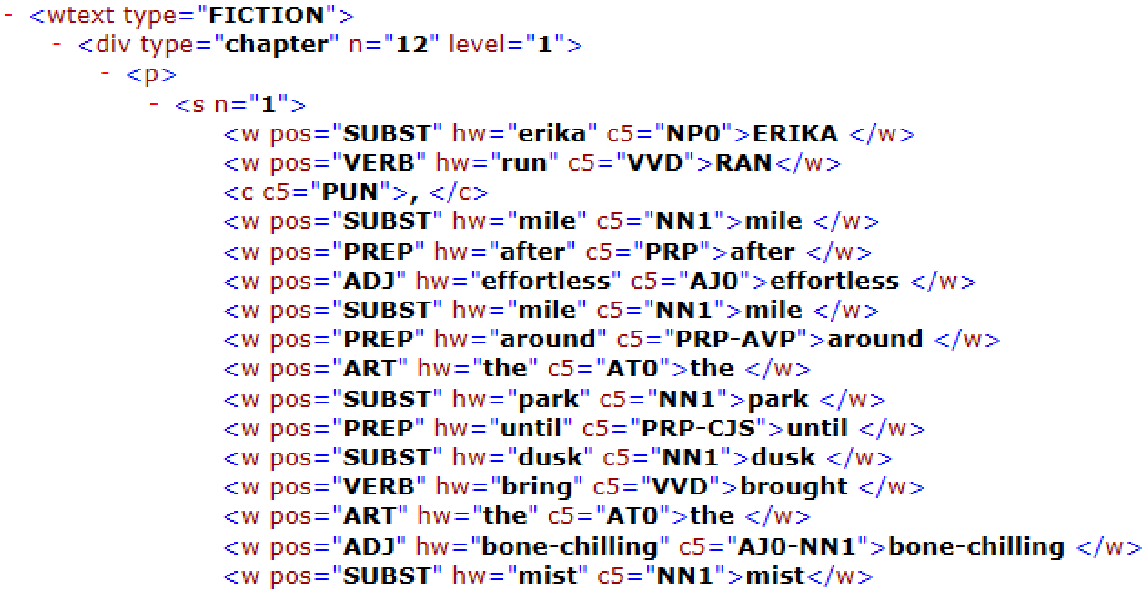
- To incorporate a copy of the text extracts listed above into the Written BNC 2014, reformatted to make them compatible with the other samples within the corpus;
- To distribute electronic copies of the text extracts as part of the full Written corpus, only under the terms of, and to persons/institutions who have signed, the BNC2014 End User Licence;
- To allow signatories of the End User Licence to make use of the text extracts only for the purposes specified in, and to the extent permitted by, that Licence.

We acknowledge your assurances that the End User Licence allows signatories to make use of the corpus for purposes of non-commercial research and teaching, and expressly forbids (a) the use of the corpus materials for any commercial or profit-making purposes, (b) any further transfer of the corpus materials to other parties, (c) reproduction of any extract of the material except within the limits of 'fair dealing' as defined in UK copyright law.

Signed:	
Date:	

Annex 2

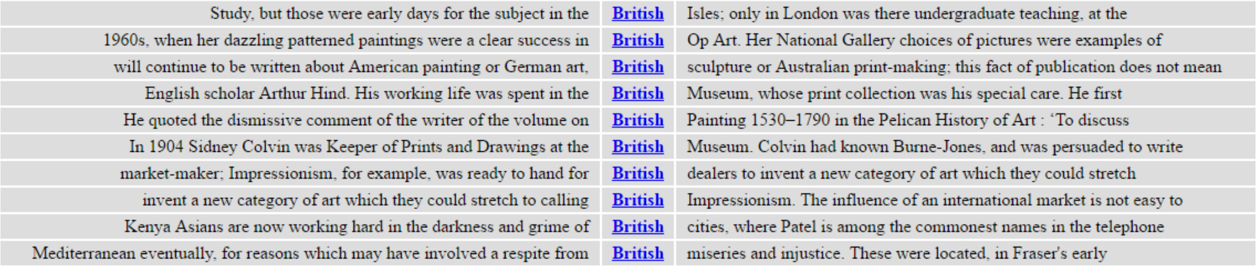
Below is an example of how the text would look to a user accessing a full copy of the corpus. This example shows the sentence "ERIKA RAN, mile after effortless mile around the park until dusk brought the bone-chilling mist", which comes from a chapter of a novel that was included in the original BNC. As you can see, the text is heavily annotated with XML tags, and is not a straightforward readable copy of the original sentence.



```
- <wtext type="FICTION">
  - <div type="chapter" n="12" level="1">
    - <p>
      - <s n="1">
          <w pos="SUBST" hw="erika" c5="NP0">ERIKA </w>
          <w pos="VERB" hw="run" c5="VVD">RAN</w>
          <c c5="PUN">, </c>
          <w pos="SUBST" hw="mile" c5="NN1">mile </w>
          <w pos="PREP" hw="after" c5="PRP">after </w>
          <w pos="ADJ" hw="effortless" c5="AJ0">effortless </w>
          <w pos="SUBST" hw="mile" c5="NN1">mile </w>
          <w pos="PREP" hw="around" c5="PRP-AVP">around </w>
          <w pos="ART" hw="the" c5="AT0">the </w>
          <w pos="SUBST" hw="park" c5="NN1">park </w>
          <w pos="PREP" hw="until" c5="PRP-CJS">until </w>
          <w pos="SUBST" hw="dusk" c5="NN1">dusk </w>
          <w pos="VERB" hw="bring" c5="VVD">brought </w>
          <w pos="ART" hw="the" c5="AT0">the </w>
          <w pos="ADJ" hw="bone-chilling" c5="AJ0-NN1">bone-chilling </w>
          <w pos="SUBST" hw="mist" c5="NN1">mist</w>
```

Annex 3

Below is an example of how the text would look to a user accessing the corpus via a specialised online software tool. This example shows the first ten results of a search for the word 'British' in the original BNC. As you can see, users only see the search term surrounded by a few words on either side (only users who have registered individually as licensees can see wider context for a search result). This way of accessing the corpus is much more commonly used than the method shown in annex 2.



Study, but those were early days for the subject in the	British	Isles; only in London was there undergraduate teaching, at the
1960s, when her dazzling patterned paintings were a clear success in	British	Op Art. Her National Gallery choices of pictures were examples of
will continue to be written about American painting or German art,	British	sculpture or Australian print-making; this fact of publication does not mean
English scholar Arthur Hind. His working life was spent in the	British	Museum, whose print collection was his special care. He first
He quoted the dismissive comment of the writer of the volume on	British	Painting 1530–1790 in the Pelican History of Art : 'To discuss
In 1904 Sidney Colvin was Keeper of Prints and Drawings at the	British	Museum. Colvin had known Burne-Jones, and was persuaded to write
market-maker; Impressionism, for example, was ready to hand for	British	dealers to invent a new category of art which they could stretch
invent a new category of art which they could stretch to calling	British	Impressionism. The influence of an international market is not easy to
Kenya Asians are now working hard in the darkness and grime of	British	cities, where Patel is among the commonest names in the telephone
Mediterranean eventually, for reasons which may have involved a respite from	British	miseries and injustice. These were located, in Fraser's early

<u>Annex 4</u>

Below is a copy of the end-user license which users of the corpus will have to agree to.

The Written BNC2014 User Licence

Preamble

The Written BNC2014 is a publically-accessible resource. This means that anyone may obtain a copy and use it for non-commercial research. However, the materials contained within the corpus are not in the public domain: they remain subject to copyright.

The copyright in the corpus texts **is owned by the respective originators of those texts**, who have granted permission for their material to be used under the terms of this licence ***only***. By adhering to the conditions of this licence, users respect the intellectual property of the copyright holders.

Use of the Written BNC2014 without registering and submitting a request via this form is forbidden. By registering and submitting this form you are entering into a licence with Lancaster University. When using the Written BNC2014 under this licence, you are bound by the following terms and conditions.

Terms used in this licence

- **The Corpus**: the Written British National Corpus 2014, including (a) the texts of the Corpus, (b) any modified versions of this corpus supplied alongside those texts, and (b) all supplementary documentation and other material supplied alongside those texts.
- **We/Us**: Lancaster University, distributor of the Corpus, acting on our own behalf and on behalf of the copyright holders in the material contained within the Corpus.
- **You**: the signatory of this licence, to whom permission to access and use the Corpus is granted.
- You may sign this licence either as an **individual**, or as a representative of an **institution**. In cases where different conditions apply to individual and institutional signatories, this is stated explicitly below: "If you are an **individual signatory**…" / "If you are an **institutional signatory**…"

General use of the Corpus

- **You may** make use of the Corpus only: (a) for purposes of non-commercial research, or (b) for purposes of teaching.

- If you are an **individual signatory**, **you must** ensure that your use of the Corpus, or use of the Corpus by other persons through any interface created or maintained by you, adheres to the terms of this licence.
- If you are an **institutional signatory**, **you must** ensure that use of the Corpus by any person affiliated to your institution, or who gains access to the Corpus through an interface created or maintained by your institution, adheres to the terms of this licence.

Publication of research

- **You may** publish the results of research that uses the Corpus.
- In any such publication, **you may** reproduce excerpts of the texts of the Corpus only within the limits of "fair dealing" as defined in UK copyright law (for example, by quoting individual sentences without extended context). **You must** clearly identify any such excerpt as originating from the Corpus. **You must not** include longer excerpts in any publication.
- In any such publication, **you must** acknowledge the use of the Corpus in your research, by citing the following standard reference (in your field's usual referencing style):

  [To be specified at completion of the project]
- **We ask you to** (*but you do not have to*) inform us of such publications through our website, so that we can list them in appropriate public bibliographies of research that uses the Corpus.

Reproduction and modification of the Corpus

- If you are an **individual signatory**, **you may** make an unlimited number of copies of the Corpus for your personal use only.
- If you are an **institutional signatory**, **you may** make an unlimited number of copies of the Corpus for use by people affiliated to your institution (that is: employees and, if you are an educational institution, students). Likewise, **you may** copy the Corpus to a shared drive or network location within your institution, so long as this location can only be accessed by people affiliated to your institution.
- **You must not** redistribute the Corpus. This means that **you must not** transfer, or allow to be transferred, any copy (in part or full) part of the Corpus or a modified version of the Corpus, to any other person or institution.
- You **must not** allow any other person or institution to access or use the Corpus, except under the conditions outlined below for online interfaces.
- **You must** store all your copies of the Corpus on computer equipment that you own and is under your direct control. In particular, **you must not** store any copy of the Corpus on any external "Cloud" Internet service.
- **You may** re-encode, reformat, annotate, and/or modify the Corpus in any way, and make use of, and/or make copies of, such a modified version of the Corpus in any of the ways that this licence permits.

- **You must not** pass copies of any such modified version of the Corpus, or any part of such a modified version, to any other person or institution (if you are an institutional signatory: to any person not affiliated to your institution as defined above).

- If you wish to allow others to obtain a copy of such a modified version of the Corpus, **you may** submit the modified version to us for distribution alongside the original Corpus.

- **We** reserve the right to accept or reject any such submission. If we accept such a submission, we reserve the right to distribute the modified version under a licence with more restrictive terms than those outlined here.

- **You agree** that any intellectual property that you hold in any modifications that you make to a version of the Corpus, that is submitted to us and that we accept, shall be assigned to Lancaster University.

Use in online interfaces

- **You may** allow others to make use of your copy of the Corpus, or any modified version, via an online interface.

- **You must** ensure that any such online interface only allows the Corpus to be used in accordance with the conditions of this licence. In particular:

    - If your online interface allows users to access only (a) minimal extracts of the Corpus within the limits of "fair dealing" as defined in UK copyright law and/or (b) statistical, graphical or other summaries of the Corpus, **you m**ay allow anyone to use the interface.

    - If your online interface allows users any more extensive access to the Corpus than that outlined above (e.g. if it allows partial or full texts of the Corpus to be read or downloaded), then **you must** ensure that only people who are signatories to this licence, or who are affiliated to an institutional signatory, are able to use the interface.

- **You must** provide us, on request, with access to any such online interface.

- **You must** provide us, on request, with full details in writing of how access to the corpus data in any such online interface is monitored and controlled in accordance with the conditions above.

Commercial use of the Corpus

- **You must not** make use of the Corpus for any commercial or profit-making purpose under this licence.

Other conditions of use

- **You** hereby **acknowledge** that the Corpus data is provided to you "as is", without any warranty, without even the implied warranty of fitness for a particular purpose.

- **You agree** that neither **we**, nor any copyright holder in the Corpus, will be liable to you for loss of profits, goodwill or any kind of consequential losses of any nature arising from your use of the Corpus, even if such loss was foreseeable.

Termination of the licence

- **We may** terminate this licence at any point, by giving you notice in writing. **You must** erase all copies of the Corpus in your possession upon receipt of such notice.
- **You may** terminate this licence at any point, by erasing all copies of the Corpus in your possession.

Data Protection

We will use your data in accordance with the Data Protection Act.

(At the bottom of the licence there is then either (a) in paper form, boxes for the licensee to insert their personal details, sign and date or (b) in online form, fields for contact details and a "typed" electronic signature – labelled as such, so that typing and pressing "I agree" indicates consent to the conditions of the licence.

Institutional signatories must also specify their official position within the institution.)

Appendix G: A copy of the Google form where authors could submit their texts

The British National Corpus 2014 - Book Collection

Dear writer,

Would you like to contribute to the British National Corpus - a very large research repository of modern British English?

The British National Corpus 2014 is a major project led by Lancaster University to create a 100 million word corpus (a large collection of ‘real life’ language) of modern-day British English.This corpus is used by researchers to understand more about how language works and how it is evolving. Educators, dictionary compilers and the interested public will also be able to access it to find usage examples of modern published British English in different genres.

To collect samples of books for the Written BNC2014 we rely on the generosity of authors to give us access to their published texts to incorporate them into the corpus. We are asking authors to provide us with extracts of any books published in between 2013 and 2018 from as wide a spread of genres as possible.

You can submit these extracts (about 5,000 words each but we also accept longer or shorter extracts) as word documents (doc, docx, rtf etc.) or any other common electronic format. Your contribution to this world-leading language resource will be fully credited in the corpus documentation.

Thank you very much for your contribution.

The Lancaster team.

email: a.hawtin@lancaster.ac.uk

Title of book:………………………………..

Date of publication:…………………………

Publisher:……………………………………

Genre:……………………………………….

Name of author(s):………………………….

Author gender: Female/male/prefer not to say/other.

Please submit an extract (or multiple extracts) of your work which is roughly equivalent to 5,000 words. We also accept shorter or longer extracts. Ideally, these extracts would be a continuous excerpt from the book. You can submit this in any widely used file format.

Which part of the book does the extract come from? Beginning/middle/end/is a compilation of multiple parts/other.

Appendix H: Instructions for workshop participants regarding the selection and scanning of books

Data Collection - Books

What is the BNC2014?

The British National Corpus 2014 is a major project led by Lancaster University to create a 100 million word corpus (a large collection of ‘real life’ language) of modern-day British English. This corpus will be used by researchers to understand more about how language works and how it is evolving. Educators, dictionary compilers and the interested public will also be able to access the corpus to find usage examples of modern British English in different genres.

You can find more about the BNC2014 project at http://cass.lancs.ac.uk/.

Focussing on books

What do we need?

- Books written by **British** authors
- Genre:
    - Fiction: poetry and prose (general, children, teens, fantasy, crime, romance)
    - Non-Fiction
- Data of publication:
    - Fiction: **first** published in or after 2010
    - Non-fiction: books published before 2010 are acceptable as long as the selected edition was published in or after 2010

How much to get from each book?

Ideally, we want a number between 20,000 and 50,000 running words from each book. One single page contains around 400 words, but that of course depends on the book layout and font.

Some of you will get those pages from the beginning of the book, others from the middle, and others from end of the text.

NOTE: For copyright reasons, we can only use approximately 30% of a book.

Task 1: Selecting books

Step ❶: Find a book fits the criteria specified in “What we need?”:

- **Lancaster student/staff**: please borrow a book from the library to carry out the steps below.
- **If you do not have a Lancaster library card**: let us know and we will make sure we get a book for you.

- Check whether the author(s) is **British**
  - You can find a list of British authors at [link]
  - If the author is not listed, check the book blurb
  - Google is also an alternative

- Genre:
  - You will find Fiction and Non-Fiction books in the library’s “Leisure Reading” section (A Floor – next to “Reserved books”).

- Data of publication:
  - You will find the data of publication in the first pages

Step ❷: Fill in the metadata form

Task 2: Scanning

To carry out this task, you will use one of the photocopy machines on campus. A map of LU printer locations is attached. Alternatively, please refer to website: http://www.lancaster.ac.uk/iss/info/IThandouts/printing/printer-locations-map.pdf

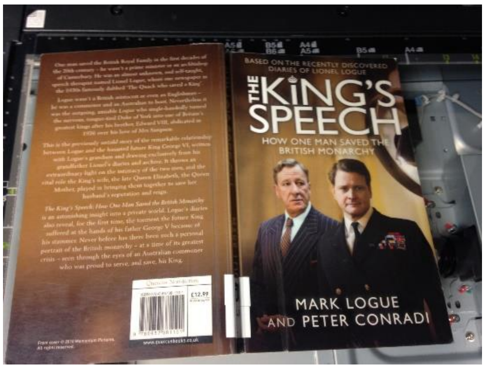
- **If you do not have a Lancaster library card**: let us know and we will be able to help

What do we want?

☑ A full set of all scanned pages from one single book, all in one single file, with no missing bits.

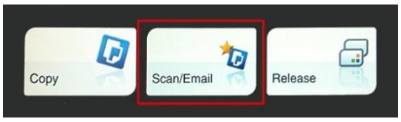
☒ No bibliographical references
☒ No index pages or glossaries
☒ No picture-only pages

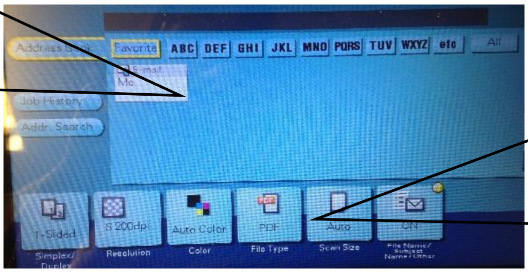
1. Place the book page(s) face down on the glass.
   ☺ **Tip**: Make sure the text is displayed horizontally. This is to minimize errors in the final stages, when we will have to convert the image into a Word file.


BASED ON THE RECENTLY DISCOVERED DIARIES OF LIONEL LOGUE
THE KING'S SPEECH
HOW ONE MAN SAVED THE BRITISH MONARCHY
MARK LOGUE AND PETER CONRADI

2. Login to the printer. When the "**Account Confirmation**" screen shows, press OK to continue.

3. Press the **Scan/Email** option.

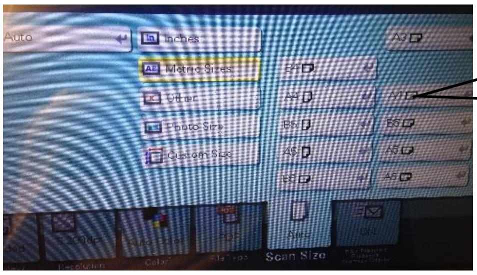




4. To email the scanned document to yourself, select the **Me** option and confirm your email address.

5. Press the "Scan Size" option in the menu bar in the lower part of the screen to adjust the format and size of the page accordingly

6. Press the "Metric Sizes" option and choose the format and page size that best match those of the book you are scanning.



For example, for *The King's Speech* (step one above), the best option is A4 landscape.

7. As you want to scan multiple pages, press on “Application” option (in the lower right corner) and make sure that the “Separate Scan” option is set as “ON”.
8. Press “Start” to scan the page(s).
9. When scanning is complete, you will see the following message on the screen: *Load the next original and press [Start]*. So follow these instructions.

*****IMPORTANT*****

To reach around 50,000 words from a single book, you are likely to need about 50 double pages (two single pages of the book). **WE HIGHLY RECOMMEND YOU TO SCAN THEM IN SETS OF 10.** This is to avoid ending up with a very heavy file to transmit to your email. The university system is very likely to block it and you will have to scan it all over again.

10. Once you have scanned all 10 pages you want, press “Finish”. You then press “Start” to begin transmission.

Task 3: What next

To be included in the corpus, we need to convert the image into a text file. This step requires specific software so we are not able to do it in this session.

We also need to:

- Remove any picture or template (i.e. title of the book at the top of the page)
- Correct OCR errors and join words divided in syllables
- Add the metadata to each text

Appendix I: The Google form where participants could submit a book scan

Book collection - submission form

Dear contributor,

Please use one form per book, to upload your scanned files.

Thank you very much for your contribution.

BNC2014 team

Title of the book:……………………………………………

Author(s):…………………………………………………

Year of first publication:…………………………………….

Scanned book: Year of publication:………………………….

Edition of scanned book:…………………………………….

Genre: Poetry/fiction/non-fiction

Sub genre: General fiction/children’s fiction/ teenage fiction/fantasy/romance/crime/travel/hobbies/history/technology/popular science/other.

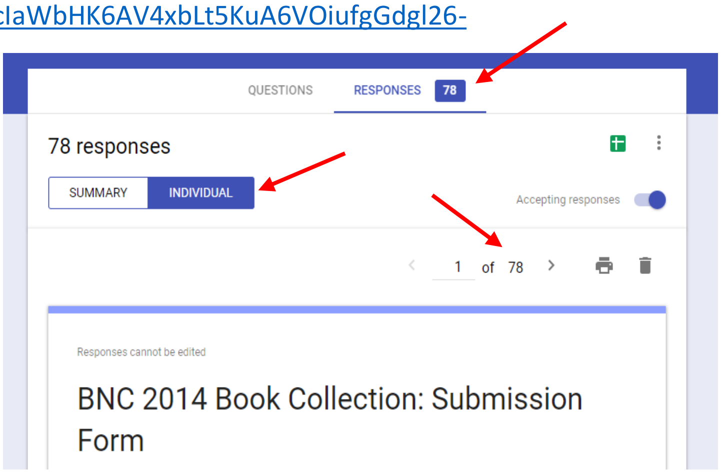
Sample from the: Beginning/middle/end

Name of contributor:…………………………………………

Appendix J: Instructions for converting a scanned book to text using Google OCR

Task 1: Convert PDF to text using OCR

1. Open the ‘BNC 2014 Book Collection: Submission form’ at the following link: https://docs.google.com/forms/d/18claWbHK6AV4xbLt5KuA6VOiufgGdgl26-hMnRbtlU4/edit (log in using BNC2014 details).
   Select ‘responses’ and then select ‘individual’.
   Use the arrows to navigate to a response which has not yet been converted using OCR.

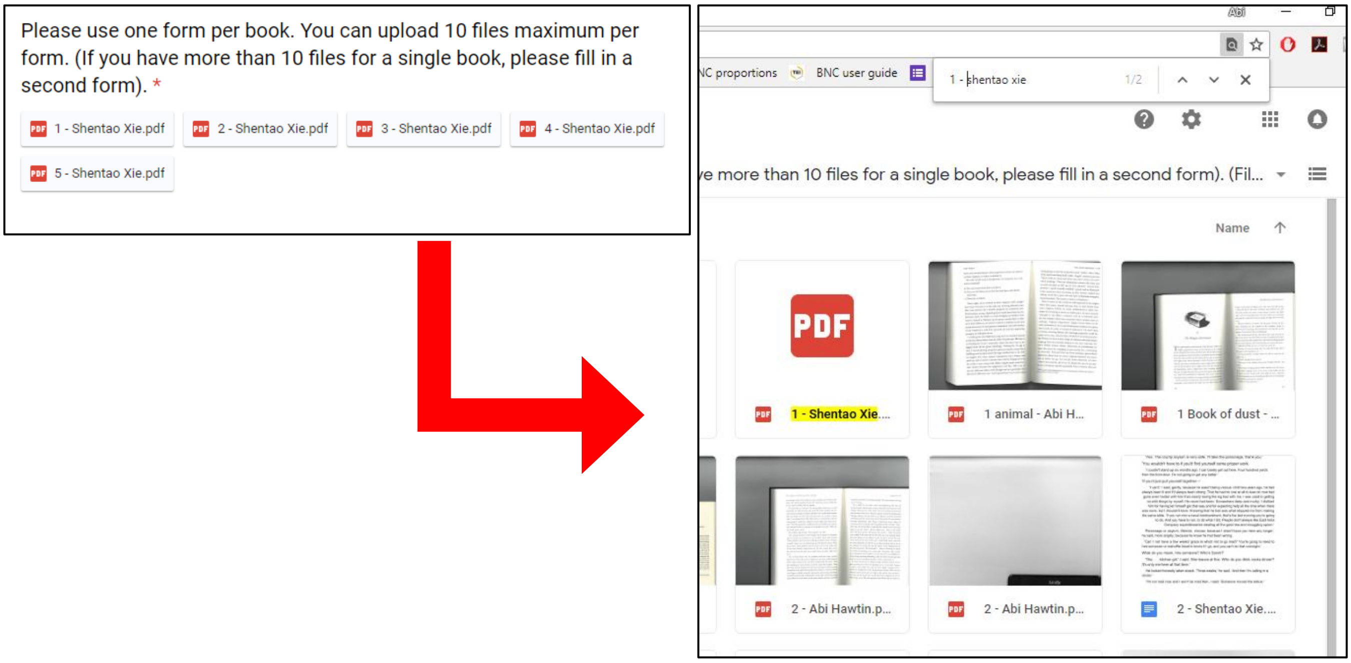


78 responses

Responses cannot be edited

BNC 2014 Book Collection: Submission Form

2. Open Google drive (https://drive.google.com/drive/u/0/my-drive) and log in using the BNC2014 details. Open the folder titled ‘BNC2014 Book Collection – Submission Form (File Responses)’ and then open the folder within. From here you can find the files which were uploaded to the Google form which you are currently working on. To do this, scroll to the bottom of the Google form to see the files which were uploaded, and then use Ctrl+F to search for the file names in the Google Drive.



Please use one form per book. You can upload 10 files maximum per form. (If you have more than 10 files for a single book, please fill in a second form). *

1 - Shentao Xie.pdf
2 - Shentao Xie.pdf
3 - Shentao Xie.pdf
4 - Shentao Xie.pdf
5 - Shentao Xie.pdf

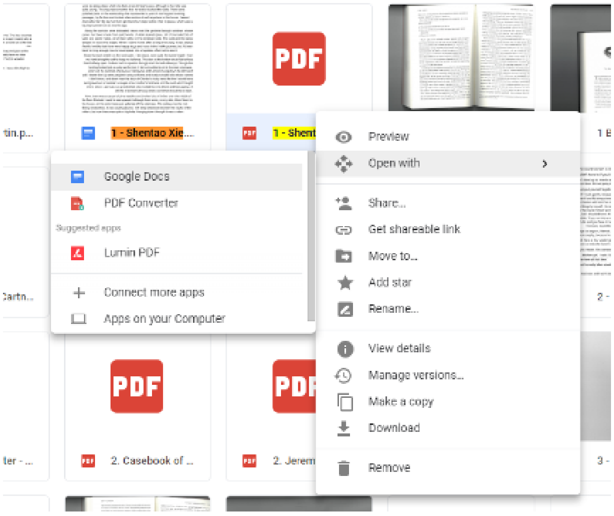
3. Open each file (by double clicking them) and check what order they belong in. You can usually do this by looking at the page numbers on the scanned books. At this point you may wish to rename the files to make it easier to remember the order

which they belong in e.g. if the book was called ‘The Bedlam Stacks’ you should rename the first file from the book ‘Bedlam Stacks 1’ and the second file ‘Bedlam Stacks 2’ etc. (You can find the title of the book on the Google form). You should also check for any files which don’t belong, for example, some people may have mistakenly uploaded a file from the wrong book, uploaded the same file twice, or uploaded something which isn’t from a book at all. Exclude these files from the next steps.

4. Right click on the first file and select ‘open with’ -> ‘Google Docs’. The file may take a while to open depending on its size.

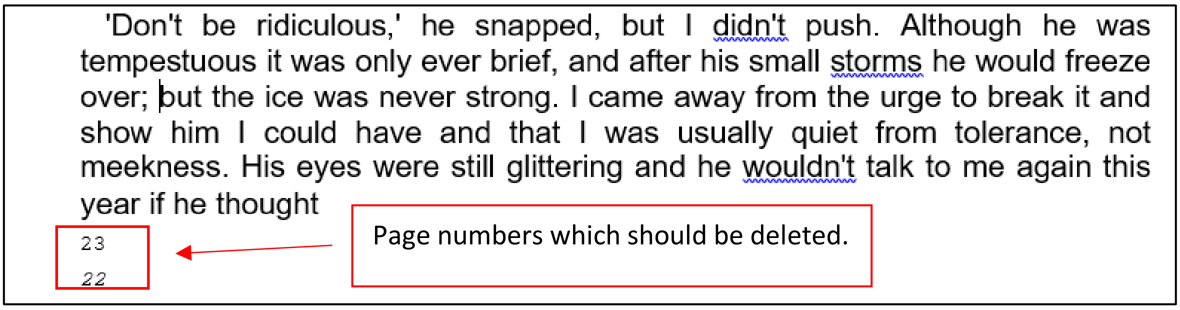
5. Copy the text from Google Docs to a Microsoft Word document.

6. Repeat step 4 and 5 for every file, until you have a Microsoft Word document which contains all of the texts from one Google form in the correct order.



Task 2: Clean the converted text

1. Look through the Word document for anything which isn’t part of the main content of the book, and delete these. These could be page numbers, chapter titles, book titles, author names etc.



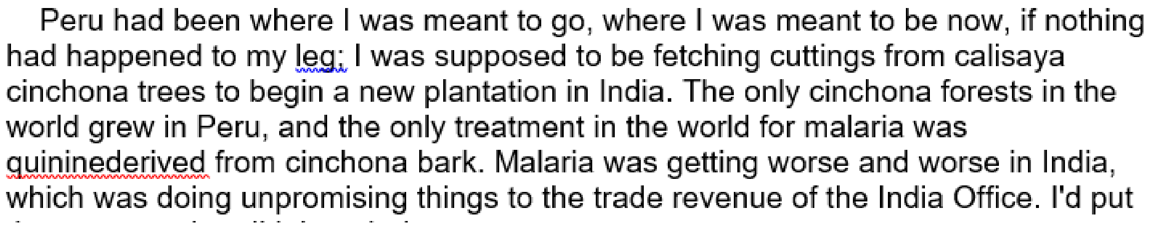
'Don't be ridiculous,' he snapped, but I didn't push. Although he was tempestuous it was only ever brief, and after his small storms he would freeze over; but the ice was never strong. I came away from the urge to break it and show him I could have and that I was usually quiet from tolerance, not meekness. His eyes were still glittering and he wouldn't talk to me again this year if he thought

23

22

Page numbers which should be deleted.

2. Also look for any instances where the OCR has introduced errors into the text and correct these – the OCR process may have resulted in typos, missing spaces etc. These will usually be highlighted by Microsoft Word’s spellchecker. NOTE: not all spellcheck ‘errors’ will be real errors, in fact most of them won’t be. Check the Word document against the original PDF if you’re unsure whether something is a mistake or not.
E.g. In the example below, after comparing with the original PDF, we see that “quininederived” should be corrected to “quinine-derived”.



Peru had been where I was meant to go, where I was meant to be now, if nothing had happened to my leg; I was supposed to be fetching cuttings from calisaya cinchona trees to begin a new plantation in India. The only cinchona forests in the world grew in Peru, and the only treatment in the world for malaria was quininederived from cinchona bark. Malaria was getting worse and worse in India, which was doing unpromising things to the trade revenue of the India Office. I'd put

3. Save the word doc using the name '[title of book]_OCR_cleaned'. E.g. If the book is titled 'The Bedlam Stacks', save the file as 'The bedlam stacks_OCR_cleaned'.

Appendix K: The questionnaire used to investigate people's feelings about the privacy of their e-language

1) How do you prefer to think of your gender? Male ☐ Female ☐ Other ☐

2) How old are you? ……………

3) Are you currently in education? Yes☐ No ☐

4) What is your highest completed level of education?

GCSE or equivalent ☐

A-Level or equivalent ☐

Undergraduate degree or equivalent ☐

Postgraduate degree or equivalent ☐

Doctoral degree or equivalent ☐

Other ☐ Please specify:…………………………………….

Prefer not to say ☐

5) How often to you send tweets on Twitter?

Never ☐ A few times a year ☐ Monthly ☐ Weekly ☐ Daily ☐

6) Do you have, or contribute to, a publicly accessible blog?

No ☐ Yes ☐ Not currently, but I have in the past ☐

7) How often do you contribute to online forums (e.g. mumsnet, The Student Room etc.)?

Never ☐ A few times a year ☐ Monthly ☐ Weekly ☐ Daily ☐

8) How often do you write online reviews (e.g. for Amazon products or on TripAdvisor)?

Never ☐ A few times a year ☐ Monthly ☐ Weekly ☐ Daily ☐

9) How often do you post public comments on things you read or watch online (e.g. news articles, blogs, YouTube videos, but NOT Facebook posts)?

Never ☐ A few times a year ☐ Monthly ☐ Weekly ☐ Daily ☐

10) If there is any other information that you would like to give about your online posting please write it below.

………

………

………

………

Please read the following:

Currently, anything which you post publicly online (i.e. those things mentioned in questions 5-9) can be used by researchers without seeking your permission and without any obligation to anonymise you.

11) Were you aware that this was the case?

Yes ☐ No ☐

12) What is your response to this?

……

……

……

……

……

……

13) If asked by a researcher, would you give permission for your online posts to be used in their research? Tick all that apply.

Yes ☐

Yes, if I was anonymised ☐

Some posts ☐

It would depend on the research ☐

No ☐

Other ☐ Please specify:…………………………………………………………………………

14) Why?

……

……

……

……

……

Please read the following:

A corpus is a large database (containing millions or billions of words) of real-life language. They are usually created by academics in order to research how people use language, to assist in the creation of dictionaries, or to help in the teaching of languages.

15) Would you be happy if your public online posts were used in the creation of a corpus without your knowledge?

Yes ☐ No ☐

16) Why?

………

………

………

………

17) Are there any other comments you would like to make about the issue of academics using people's online posts in their research?

………

………

………

………

………

18) Would you be willing to attend a follow-up interview or focus group based on your answers to these questions? (If you answered yes, please provide your name and email address below).

Yes ☐ No☐

19) Would you like to be informed of the findings of this study once it has been completed? (If you answered yes, please provide your name and email address below).

Yes ☐ No ☐

I give permission for my answers to this questionnaire to be used for research purposes.

Signed:…………………………

Parent/Guardian Signature (if under 18):………………………..

Name (optional):…………………………

Email address (optional):………………………………..

Thank you very much for taking the time to complete this questionnaire. Your responses are greatly appreciated.

Appendix L: The results of all comparisons in the investigation of the colloquialisation of academic British English

Table L1: Comparison of linguistic features in all academic writing.

	BNC1994 (freq per mill)	**BNC2014 (freq per mill)**	**Bootstrap % change**	**Significant? (p<0.05)**
First and second person pronouns	4,526.78	5,848.72	+19.226	NO
Present tense verbs	38,018.11	37,957.90	-8.244	NO
Verb contractions	347.22	713.9	+90.094	YES
Negative contractions	189.75	384.03	+86.775	YES
Questions (all)	780.8	805.65	-4.808	NO
Verb frequency	134,352.71	122,456.80	-16.154	YES
Genitives	3,664.52	3,675.00	-7.531	NO
Semi-modals	2,033.46	1,660.99	-24.624	YES (P<0.001)
Passive forms (all)	15,441.50	11,181.40	-33.176	YES (P<0.001)
Relative pronouns	8,420.04	6,010.71	-34.126	YES (P<0.001)
Noun frequency	254,176.85	273,095.00	-1.131	NO

Table L2: Comparison of linguistic features in academic books.

	Academic books BNC1994 (freq per mill)	**Academic books BNC2014 (freq per mill)**	**Bootstrap % change**	**Significant? (p<0.05)**
First and second person pronouns	4,669.37	6,787.20	+189.395	YES (p<0.001)
Present tense verbs	39,909.61	40,918.00	+104.141	YES (p<0.001)
Verb contractions	398.89	869.6	+334.906	YES (p<0.001)
Negative contractions	216.34	493.27	+353.998	YES (p<0.001)
Questions (all)	853.47	968.21	+125.878	YES (p<0.001)
Verb frequency	135,882.70	126,767.10	+85.753	YES (p<0.001)
Genitives	3,752.88	3,736.50	+98.243	YES (p<0.001)
Semi-modals	2,180.45	1,881.00	+71.788	YES (p<0.001)
Passive forms (all)	14,941.99	10,544.00	+40.504	YES (p<0.001)
Relative pronouns	8,907.69	6,827.40	+52.596	YES (p<0.001)
Noun frequency	248,857.91	261,682.30	+109.371	YES (p<0.001)

Table L3: Comparison of linguistic features in academic journal articles.

	Academic journals BNC1994 (freq per mill)	**Academic journals BNC2014 (freq per mill)**	**Bootstrap % change**	**Significant? (p<0.05)**
First and second person pronouns	3,871.48	4,223.29	-52.827	YES (p<0.001)
Present tense verbs	29,324.98	32,825.40	-51.595	YES (p<0.001)
Verb contractions	109.75	444.32	+75.067	NO
Negative contractions	67.54	194.81	+24.73	NO
Questions (all)	446.8	524.07	-49.277	YES (p<0.001)
Verb frequency	127,321.05	114,990.90	-60.944	YES (p<0.001)
Genitives	3,258.44	3,568.29	-52.644	YES (p<0.001)
Semi-modals	1,357.27	1,275.80	-59.343	YES (p<0.001)
Passive forms (all)	17,737.15	12,285.55	-70.048	YES (p<0.001)
Relative pronouns	6,178.85	4,596.32	-67.832	YES (p<0.001)
Noun frequency	278,622.07	292,862.90	-54.546	YES (p<0.001)

Table L4: Comparison of linguistic features in the humanities and arts genre.

	humanties_arts BNC1994 (freq per mill)	**humanities_arts BNC2014**	**Bootstrap % change**	**Significant? (p<0.05)**
First and second person pronouns	5,325.73	6,893.79	-42.433	YES
Present tense verbs	33,549.48	38,202.00	-49.359	YES (p<0.001)
Verb contractions	331.3	984.3	+32.131	NO
Negative contractions	235.02	445.16	-15.761	NO
Questions (all)	751.04	1,017.43	-39.752	YES
Verb frequency	131,596.44	118,641.90	-59.904	YES (p<0.001)
Genitives	5,777.38	6,944.88	-46.539	YES (p<0.001)
Semi-modals	2,052.18	1,349.90	-70.737	YES (p<0.001)
Passive forms (all)	12,633.86	9,361.00	-67.047	YES (p<0.001)
Relative pronouns	9,707.57	7,062.51	-67.653	YES (p<0.001)
Noun frequency	243,192.74	264,838.70	-51.567	YES (p<0.001)

Table L5: Comparison of linguistic features in the medicine genre.

	medicine BNC1994 (freq per mill)	**medicine BNC2014 (freq per mill)**	**Bootstrap % change**	**Significant? (p<0.05)**
First and second person pronouns	3,013.33	6,751.52	+98.399	NO
Present tense verbs	27,471.65	33,352.62	-42.042	NO
Verb contractions	84.87	1,878.53	+956.599	NO
Negative contractions	60.45	890.00	602.846 (+)	NO
Questions (all)	466.5	1,293.67	+32.385	NO
Verb frequency	126,908.36	118,237.60	-55.523	YES
Genitives	2,495.53	2,609.60	-50.08	NO
Semi-modals	1,127.17	1,554.30	-34.171	NO
Passive forms (all)	19,202.86	12,529.90	-68.851	YES (p<0.001)
Relative pronouns	4,719.96	5,346.54	+0.304	NO
Noun frequency	288,048.97	274,576.60	-54.495	YES

Table L6: Comparison of linguistic features in the natural science genre.

	nat_science BNC1994 (freq per mill)	**nat_science BNC2014 (freq per mill)**	**Bootstrap % change**	**Significant? (p<0.05)**
First and second person pronouns	3,500.44	5,904.34	-14.726	NO
Present tense verbs	44,824.11	36,166.10	-53.631	YES (p<0.001)
Verb contractions	28.92	281.1	+391.451	NO
Negative contractions	84.42	302.85	+16.604	NO
Questions (all)	335.35	580.46	-12.479	NO
Verb frequency	126,099.56	115,640.40	-53.631	YES (p<0.001)
Genitives	918.49	1,387.36	-23.626	NO
Semi-modals	1,067.02	1,396.50	-33.825	NO
Passive forms (all)	18,159.62	12,060.00	-66.421	YES, (p<0.001)
Relative pronouns	5,355.41	4,772.70	-54.945	YES
Noun frequency	271,374.15	282,428.80	-47.378	YES, (p<0.001)

Table L7: Comparison of linguistic features in the politics, law and education genre.

	polit_law_edu BNC1994 (freq per mill)	polit_law_edu BNC2014 (freq per mill)	Bootstrap % change	Significant? (p<0.05)
First and second person pronouns	3,631.52	4,976.07	+215.846	YES, p<0.001
Present tense verbs	36,723.23	36,640.30	+130.014	YES, p<0.001
Verb contractions	365.64	510.6	+221.94	YES
Negative contractions	153.82	245.33	+267.677	YES
Questions (all)	898.04	736.00	+88.937	YES
Verb frequency	138,390.55	127,014.30	+111.584	YES, p<0.001
Genitives	3,761.39	4,054.64	+148.509	YES, p<0.001
Semi-modals	2,324.67	1,879.20	+86.384	YES, p<0.001
Passive forms (all)	16,079.72	10,584.00	+51.743	YES
Relative pronouns	8,964.99	6,684.77	+71.875	YES
Noun frequency	254,804.12	268,010.30	+142.484	YES, p<0.001

Table L8: Comparison of linguistic features in the social science genre.

	soc_science BNC1994 (freq per mill)	soc_science BNC2014 (freq per mill)	Bootstrap % change	Significant? (p<0.05)
First and second person pronouns	5,510.77	6,739.17	+139.531	YES
Present tense verbs	43,106.53	40,811.00	+85.463	YES
Verb contractions	549.16	734.2	+161.907	NO
Negative contractions	285.21	482.75	+231.571	YES
Questions (all)	928.61	859.17	+81.246	YES
Verb frequency	136,018.53	130,349.30	+87.733	YES, p<0.001
Genitives	3,348.54	3,136.32	+83.48	YES
Semi-modals	2,228.75	2,031.90	+78.593	YES, p<0.001
Passive forms (all)	14,218.94	10,850.20	+49.484	YES
Relative pronouns	9,189.35	6,323.89	+34.793	NO
Noun frequency	245,133.25	264,276.60	+111.191	YES, p<0.001

Table L9: Comparison of linguistic features in the technology and engineering genre.

	tech_engin BNC1994 (freq per mill)	**tech_engin BNC2014 (freq per mill)**	**Bootstrap % change**	**Significant? (p<0.05)**
First and second person pronouns	5,993.29	4,153.00	-79.174	YES, decrease
Present tense verbs	51,314.82	40,203.80	-76.446	YES, p<0.001, decrease
Verb contractions	26.62	485.9	+448.786	NO
Negative contractions	30.17	207.95	+107.216	NO
Questions (all)	548.39	579.82	-68.214	NO, decrease
Verb frequency	140,946.26	124,099.70	-73.53	YES, p<0.001, decrease
Genitives	1,217.47	2,723.62	-32.745	NO, decrease
Semi-modals	2,353.30	1,787.10	-77.17	YES, p<0.001, decrease
Passive forms (all)	21,211.69	12,731.40	-81.956	YES, p<0.001
Relative pronouns	6,048.31	5,377.02	-45.379	YES
Noun frequency	259,889.29	285,727.20	-66.948	YES, p<0.001