Vaclav Brezina*, Abi Hawtin and Tony McEnery

The Written British National Corpus 2014 – design and comparability

https://doi.org/10.1515/text-2020-0052
Received April 15, 2020; accepted July 16, 2021; published online August 12, 2021

Abstract: The British National Corpus 2014 is a major project led by Lancaster University to create a 100-million-word corpus of present day British English. This corpus has been constructed as a comparable counterpart of the original British National Corpus (referred to as the BNC1994 in this article), which was compiled in the early 1990s. This article starts with the justification of the project answering the question of 'Why do we need a new BNC?'. We then provide a general overview of the construction of the Written British National Corpus 2014 (Written BNC2014); we also briefly discuss some issues of data collection before looking in detail at the design of the corpus. Compiling a large general corpus such as the Written BNC2014 has been a major undertaking involving teamwork and collaboration. It also required generosity on the part of the many individuals and organisations who contributed to the data collection.

Keywords: balance; BNC1994; comparability; corpus description; design principles; representativeness; sampling; Written BNC2014

1 Introduction

A synchronic corpus such as the British National Corpus 1994 (hereafter BNC1994) and the British National Corpus 2014 (hereafter BNC2014) provides a snapshot of language, and a window into social history, at the time of its compilation. Language and society, however, change over time. Consider the examples below. Example (1) is taken from the Written BNC1994, while example (2) represents a contribution to the Electronic language (E-language) subcorpus of the Written BNC2014.

***Corresponding author: Vaclav Brezina**, The Department of Linguistics and English Language, Lancaster University, County South, LA14YD, Lancaster, UK, E-mail: v.brezina@lancaster.ac.uk
Abi Hawtin, Research and Impact Services, University of Warwick, Coventry, UK
Tony McEnery, The Department of Linguistics and English Language, Lancaster University, Lancaster, UK; and Xi'an Jiaotong University, Xi'an, China. https://orcid.org/0000-0002-8425-6403

(1) Currently (1993) there are some twenty million users worldwide of the most commonly used electronic network, Internet, a figure that is growing at a rate of ten percent per month. Merely imagining the contexts in which we encounter electronic systems offers some ideas about the diversity of electronically stored information. [J0V]

(2) icymi. why i think our #cosmonauts exhibition celebrates a critical landmark for humanity http://t.co/xncvepkqlj [ElanSocTwi4][1]

It can be seen that, while example (1) is descriptive in nature, highlighting the vision for the emerging technology of the Internet as seen from the perspective of the year 1993, example (2) is a tweet, demonstrating in practise the language of social media, something that would not have been encountered in 1993. Indeed, social media evolved in the context of Web 2.0 after the turn of the millennium (Van Dijck 2013) and thus is not represented in the Written BNC1994 at all. Neither are other new and emerging genres/registers[2] which are common today. In sum, social changes and changes in technology over the past twenty or so years have transformed, among other things, communication and access to language data. For this reason, it is important to create a comparable counterpart to the BNC1994 reflecting these changes and taking advantage of new methods of data collection.

The aim of this article is to present methodological decisions that led to the creation of the Written BNC2014. It is the pursuit of specific methodological points that leads to the focus in the rest of this article on a detailed exploration of the Written BNC2014 design. Our guiding questions throughout the article are:

i) What methodological decisions were taken when designing and compiling the Written BNC2014?
ii) How does the Written BNC2014 compare to the Written BNC1994?

The article is organised as follows: Section 2 discusses the context of data collection for the Written BNC2014 and compares it to the situation of the original team who created the BNC1994. Section 3 starts with a theoretical discussion of corpus sampling (Section 3.1) and an overview of the structure of the Written BNC1994 (Section 3.2) as a baseline for the creation of the broadly comparable current British English counterpart, the BNC2014. The Written BNC2014 is then described in terms

1 The files in the BNC2014 are organised according to genre/register groups such as E-language (Elan) and subgroups such as Social media (Soc) and Twitter (Twi). The genre/register membership is also reflected in the file name.

2 In this article we use the term 'genre/register' to refer to both the situational aspect and the functional aspect of a category of texts used in corpus sampling (see Section 3.3).

of its structure (Section 3.3), sampling frame (Section 3.4) and the proportions of the genres/registers (Section 3.5). Section 4 offers a brief conclusion.

2 Data collection for the Written BNC2014

When the Written BNC1994 was constructed, the team building it had the support of a broad consortium of publishers who owned a wide range of electronic texts suitable for inclusion in the corpus; they also had access to substantial funding that they could use to reformat and annotate texts as part of the process of corpus construction.

The team building the Written BNC2014 had neither of those advantages – we began with no commercial partners and with only a slender budget drawn from the ESRC Centre for Corpus Approaches to Social Science (CASS) at Lancaster University. We did, however, have advantages that the original team did not have and these offset, to greater or lesser degrees, the lack of the commercial partners and funding that the Written BNC2014 had. Our advantages were four-fold.

Firstly, corpus construction is now routine. In 1994, corpus construction was a much more research driven process – character encoding, corpus markup languages and corpus annotation software were either in their infancy or were being actively developed while the construction of the BNC1994 was in progress (for example, with part-of-speech tagging, Leech et al. 1994). Similarly, widely accepted standards had not yet become available (for example Unicode for the encoding of characters) or were developing in the context of competing standards. On the later point, the BNC1994 used standard generalized markup language (SGML) to encode the corpus within a framework being developed by the Text Encoding Initiative (TEI). It was far from obvious that this was the 'right' choice at the time as corpora were being produced in a range of formats e.g. the IBM/Lancaster Treebank (Garside and McEnery 1993). So, while the BNC1994 had access to greater resources than we did, the team was required to undertake development work in areas where we could utilise a mature software package, The Constituent Likelihood Automatic Word-tagging System (CLAWS), and widely accepted standards (Unicode and XML) to process the corpus.

Secondly, the legislative environment in the UK has changed since the early 1990s. Importantly, UK copyright legislation has been updated to encourage the construction and exploitation of datasets such as the Written BNC2014. Copyright law can be a major impediment to corpus building projects, as can be seen from the example of the American National Corpus (Ide 2008). The creators of that 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 significant stumbling block for them in the project. While the current legislation which governs copyright in the UK predates the BNC1994 (the Copyright, Designs

and Patents Act 1988), several exceptions to copyright protection, introduced in 2017, altered the legislative context in the UK substantially. The exceptions relate to 'Non-commercial research and private study' and 'Text and data mining for non-commercial research' (Gov.uk 2017).

The first provision allows for the use of copyright materials when it is for the purpose of private study and does not lead to commercial gain. The second specifically allows data to be gathered for data mining purposes as long as it does not lead to commercial gain. Given that we were constructing a corpus for the purpose of research identified by the act as data mining and, moreover, given that we intended to make the corpus available free of charge for use in non-commercial research, the amendments greatly facilitated the collection of material for the Written BNC2014. The amendments to the 1988 Act are hugely important for corpus linguists in general and for corpus builders in particular. Of course, the nature of the legal system in the UK is such that it is only when cases are brought which test the law that the boundaries of permissions like these become clear. However, we are aware of no relevant challenges to uses of these permissions which would revise our view that they allow the construction and sharing of corpora, such as the Written BNC2014, for research purposes. We also note that the provisions of the law are being used routinely by computer scientists and others engaged in so-called 'big data' research, to construct and share large datasets scraped from the web.

A third, and related change which helped our project is the rise of the open access publishing movement – for example, through schemes such as the UK Research Councils' open access scheme, which encourages and in some cases mandates open access for publicly funded research outputs.[3] Similarly, through the creation of open access journals such as PLOS ONE, and platforms offering open access peer-reviewed books, such as Open Access Publishing in European Networks (OAPEN), researchers are making articles and books available to readers, free of charge, with levels of permission that run beyond what copyright legislation may permit. For our project, works published with the Creative Commons (CC) licenses, in particular CC BY 4.0 and CC BY-NC-ND 4.0, were very helpful as 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 2021). This meant that, for academic writing in particular, we were able to have access to a large number of texts that we could use for corpus building.

A fourth advantage, which relates to advantages 2 and 3 above, lies in the fact that we were able to persuade some publishers to contribute to our corpus in part because we could argue that what we were doing was permissible and could be done anyway – we were simply asking them to help speed the process. A few

3 See https://www.ukri.org/files/legacy/documents/rcukopenaccesspolicy-pdf/.

publishers agreed to help because, unlike in the early nineteen nineties, they had become very familiar with what we were doing and understood the advantages of supporting such work. Some of them, in particular Elsevier, had large teams working on the data mining of their content anyway, so it was relatively easy and nearly cost free for them to support us. Consequently, as the project developed, we gained substantial support from Cambridge University Press, Dunedin Academic Press, Elsevier and John Benjamins.

While the lack of the two advantages of the original Written BNC team seemed completely debilitating to begin with, we realised, as our project proceeded, that the context in which the work was being undertaken had changed so substantially that we neither needed substantial resource nor industrial partners to begin our work, though acquiring industrial partners as the work progressed certainly facilitated the construction of the corpus.

There is one further advantage that the original BNC team had which we did not. They were doing something for the first time. While there had been smaller, publicly available balanced 'snapshot' corpora produced prior to the BNC1994, notably members of the Brown family of corpora (Francis and Kučera 1979 *inter alia*), the BNC1994 was of a scale that had not been attempted before and, as a consequence, the team were able to take decision without having to consider, for example, comparability to a previously produced dataset. We could not do this. As has happened with the Brown family, we wanted the Written BNC2014 to be a point of comparison with the Written BNC1994. So, to an extent, we were hemmed in by decisions previously made. However, while accepting that, we decided that we would, wherever possible, not replicate what we believed were mistakes, often born from necessity, that the original project made, nor would we be blind to the changing context of written English in the two decades since the BNC1994 was produced.

3 The design of the Written BNC2014

When proceeding by design to generate a corpus that is comparable to the written component of the BNC1994, we felt we had to understand how that corpus was designed and to re-evaluate that design. As noted, the original corpus designers were working in something of a vacuum – we were not. We therefore had to take into account not merely what was done in creating the original BNC, but also to critically evaluate these decisions and modify the procedure to reflect the best present day practise in the field.

3.1 Corpus sampling

In statistics, a probability sample, i.e. sample taken from a population (*U*) based on measurable criteria,[4] is defined by a sampling design, which is a probability distribution (Ardilly and Tillé 2006). Each unit in the population (*u*) has a known non-zero probability to appear in the sample: $p(u) > 0$ for all *u* taken from *U* (Henry 1990: 25). In general terms, this framework allows us to weigh and balance the sample according to the needs given by the research questions in any particular project. We have to realise that the overall aim of sampling is to take a subset of the population (universe) that is useful for answering questions about the population (universe). A similar principle of sampling applies to recruiting participants in large-scale quantitative studies – in our case, texts instead of human subjects are selected as units *u* of the sample. A good sample is one which allows us to answer these questions reliably. For example, an extremely skewed sample of people taken from a general population of the UK (*U*), which would include only men, would not allow us to reliably discover people's preferences, opinions or habits, because only half of the population, all male, would be represented in the sample.

In addition, in corpus design, the crucial starting point is to realise which type of research questions the corpus is intended to answer and which it is not. As mentioned earlier, a distinct advantage of designing an updated British National Corpus is the fact that we can rely on the now more than 20 year experience with the BNC1994 and the types of research questions which were addressed in the more than 3,000 pieces of published research based on this corpus (Love et al. 2017). For example, the BNC1994 has been used to describe the grammar (e.g. Gardner and Davies 2007; Hoffmann 2007) and the vocabulary of British English (e.g. Leech et al. 2001). It has also been used to investigate social as well as situational (register) variation in English (e.g. Kennedy 2002; McEnery and Xiao 2004). In numerous studies on second language acquisition (e.g. Ellis et al. 2008; Nation 2004), the BNC1994 has been used as a reference corpus against which performance in second language has been evaluated. These are merely a few examples demonstrating the broad range of uses of the BNC1994 – uses which likely go beyond what the BNC1994 team could have anticipated at the time of its compilation. However, with the benefit of the experience with and reflection upon the typical uses of the BNC1994, we made a decision to design a large well-structured corpus representing a broad range of genres/registers. Such a corpus can reflect major features of the target language – British English – without necessarily reflecting all societal and

4 There are also nonprobability samples such as a convenience sample, which are not considered here.

Table 1: Types of bias in corpus linguistics.

Type of bias	Brief explanation
Text sample bias	Overrepresentation of certain parts of texts e.g. beginnings or ends of text in the corpus.
Topic bias	Overrepresentation of texts containing a small topic range. This leads to specific topic-related vocabulary items being overrepresented in the corpus.
Non-coverage bias	Coverage of a small range of 'visible' texts and non-coverage of other texts.
Traditional text type bias	Coverage of only text types traditionally included in corpora.
Legal considerations bias	Selection of texts to which copyright does not apply (older out of copyright texts, texts under creative commons licenses etc.)
Practicality bias	Overrepresentation of easily obtainable texts.
Self-selection bias	Inclusion of texts and spoken material only by contributors who volunteer to participate; this leads to overrepresentation of texts by highly-motivated individuals, which may not be representative of the whole population.

linguistic processes in the UK. The corpus is thus designed as a linguistic, not a sociological, sample with a typical range of linguistic research questions in mind.

Corpus sampling is a systematic and controlled process used to avoid different types of bias. Bias describes a situation, where "a statistic based on a sample systematically misestimates the equivalent characteristic (parameter) of the population" (Cramer and Howitt 2004: 13). Due to a bias a sample fails to accurately represent the population and thus has a limited value. Brezina (2018: 16–17) lists the following types of bias, which are relevant for corpus designers and need to be avoided when building a corpus (see Table 1).

While in random sampling a sampling error, i.e. the variability of a sample from the characteristics of a population, is a natural part of the sampling process (indeed statistics can account for and model this), deviations of a sample from the population due to various types of bias need to be carefully controlled. Following this principle, in designing the Written BNC2014, we have regularly evaluated our selection criteria against each of the seven types of bias. For instance, we sampled texts on a wide range of topics from a wide range of sources and included both traditional as well as emerging genres/registers (see Section 3.3). We used both whole texts and text samples. If text samples were used, we made sure that beginnings, middles and ends of texts were equally represented. We also dealt with the issue of copyright (see Section 2) to avoid legal considerations bias.

Overall, the Written BNC2014 has been designed as a stratified sample, comprising major genres/registers, represented according to the range of categories listed in Table 6. The proportions of each genre/register are given by weighting coefficients *c* and were established in order to sufficiently cover major genres/registers in the BNC1994 (see Section 3.2); new registers/genres were added to reflect changes in British English over the last 20 years (see Section 3.3).

Within a genre/register there are still choices to be made, and in making those choices we encounter some of the classic problems of sampling in corpus linguistics. For example, in the E-language genre/register, should we sample proportionately (see Biber 1993: 247 for a discussion) and if so how? To sample proportionately based on the actual distribution of genres/registers in a population may mean that some types of E-language effectively crowd out other forms of E-language simply by virtue of the brevity of some electronic communications or because of the high volume of certain types of communications. Typically, this is not what linguists would want; as noted by Bauer and Aarts (2000: 29), proportional representation of language in a corpus has the potential to squeeze "the rare event" as "representative sampling would suggest ignoring it". Yet such rare events are of interest to linguists. Biber (1993) instead rightly suggests that researchers actually require corpora which are representative in the sense that the full range of linguistic variation is adequately represented – it is in this sense of representative that we undertook stratified sampling within our genres/registers. We focussed on the stratification to try to ensure that linguistic variation was adequately represented. For example, if we had one genre/register category covering all blogs then chance would dictate whether any travel blogs were included in the corpus; but because a separate genre/register category has been identified for travel blogs, at least some such texts were included. This is just a small example of further stratification of the sample inside broader genre/register categories, which was systematically applied in order to include even rarer items, which are, however, of linguistic interest (Biber 1993).[5] Table 2 summarises the general framework for the Written BNC2014 discussed above.

Table 2: Basic sampling properties of the Written BNC2014.

Universe	recorded uses of British English
Design	stratified, random, weighted
Sample size	90M tokens
Sample unit (u)	text/text sample

5 Though see Váradi (2001) for arguments against this approach.

3.2 The structure of the Written BNC1994

For the sake of comparability and continuity with the BNC1994, we paid close attention to the structure of the original corpus. At the highest level, the *BNC User Reference Guide* (Burnard 2007) describes the Written BNC1994 as being divided into i) informative texts and ii) imaginative texts, categories labelled as 'domains'. The former domain represents a majority with 73% of texts in this category, while the latter forms the remaining 27%.[6] The domains can be further categorised as i) imaginative, ii) arts, iii) belief and thought, iv) commerce, v) leisure, vi) natural science, vii) applied science, viii) social science and ix) world affairs. Aston (2001: 73) points out that a decision was made to keep the framework of the categories broad enough to allow the designers to guarantee "at least 100 texts in most principal categories".

In terms of the written medium, 58% of texts come from books, 30% from periodicals and 12% from miscellaneous sources (published, unpublished, written to be spoken).[7] The selection procedure involved choosing target proportions (percentages) in the domain and medium categories. These were, however, set independently of each other, so there were no targets in the corpus design for specific domain/medium cross-sections (e.g. imaginative books). Random sampling was used for half of the books based on *Whitaker's Books in Print 1992*. Other sampling considerations included availability of texts (texts which were "easier to obtain in useful quantities", Burnard 2007: 'Sampling basis: production and reception'), the relative importance of the texts in terms of their reception and the range of different texts in terms of their production. The consideration for selecting prominent published written texts (based on "statistics about books and periodicals that are published, bought or borrowed", Burnard 2007: 'Sampling basis: production and reception') was counterbalanced with the need not to overrepresent a small number of highly popular items and to include a variety of texts. The proportions of unpublished written texts were based on "intuitive estimates" (Burnard 2007: 'Sampling basis: production and reception').

From this description we can see that the design decisions for the Written BNC1994 were made at a high level (domain and medium, Aston 2001) with the need to find a balance between multiple considerations (e.g. production and reception). The texts in the Written BNC1994 can be further classified according to e.g. author type, audience type, sampling type (whole text, sample, composite) and also, importantly, the genre/register. However, we need to realise that these

6 Note that the design of the corpus presupposed a 75/25% split.

7 Note that the design of the corpus presupposed a 60/30/10% split.

classifications, based on the corpus meta-data, are categorisations *ex-post* and did not play a role in the sampling of the Written BNC1994.

An important genre/register classification of the BNC1994 is provided by Lee (2001).[8] The scheme gives 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 prefers the term 'genre' to refer to the traditional text categories based on external (rather than internal – linguistic) criteria. However, as Lee (2001: 46) acknowledges, 'genre' and 'register' refer to the same concept from two different perspectives, the former from the perspective of the form, the latter from the perspective of language function: "I contend that it is useful to see the two terms genre and register as really two different angles or points of view, with register being used when we are talking about lexico-grammatical and discoursal-semantic patterns associated with situations (i.e. linguistic patterns), and genre being used when we are talking about memberships of culturally-recognisable categories." Lee's classification is available in Table 3. We have transposed genres and super genres to better represent the hierarchy of the classification.

Overall, the Written BNC1994 splits into 46 genre categories with 24 major super genre distinctions. It is important to realise that this categorisation is descriptive and was applied to the Written BNC1994 after its compilation and not as part of the corpus design. Indeed, Lee (2001) felt that the original BNC1994 classification system had many problems, which he aimed to solve by creating this new genre scheme. However, as is apparent from Table 3, the frame developed by Lee is uneven in terms of how it provides super genres only for certain genre categories. Moreover, as Aston (2001: 74) points out, many of the genre categories are "poorly represented in the corpus, both numerically and in terms of their variance." The Written BNC1994 thus cannot be used to investigate genre/register variation at this detailed level of granularity. In addition, many texts cannot be uniquely classified under a single genre because one text can contain, for example, both prose and poetry. Despite these limitations, Lee's (2001) categorisation serves as a useful inventory of the types of texts included in the original corpus; it helped us establish a design which promoted continuity with the original dataset (see Section 3.3).

8 Lee had explored the variation in speech and writing in his Doctoral dissertation (Lancaster 2000), where he critically applied multidimensional analysis to the British National Corpus 1994 data.

Table 3: Written BNC1994 genres classification scheme (Lee 2001: 57–58, super genres and genres transposed).

Super genres/genres	Genres
1. Academic prose	i) medicine; ii) natural sciences; iii) politics law education; iv) social & behavioural sciences; v) technology computing engineering
2. Administrative and regulatory texts in-house use	
3. Print advertisements	
4. Biographies/autobiographies	
5. Commerce & finance economics	
6. Email sports discussion list	
7. School essays	
8. University essays	
9. Excerpts from modern drama scripts	
10. Single- and multiple-author poetry collections	
11. Novels & short stories	
12. Hansard/parliamentary proceedings	
13. Official/governmental documents/leaflets company annual reports etc.; excludes Hansard	
14. Instructional texts/DIY	
15. Personal letters	
16. Professional/business letters	
17. Miscellaneous texts	
18. TV autocue data	
19. Broadsheet national newspapers	i) arts/cultural material; ii) commerce & finance; iii) personal & institutional editorials & letters-to-the-editor; iv) miscellaneous material; v) home & foreign news reportage; vi) science material; vii) material on lifestyle leisure belief & thought; viii) sports material; ix) arts; x) commerce & finance; xi) home & foreign news reportage; xii) science material; xiii) material on lifestyle, leisure, belief & thought; xiv) sports material
20. Regional and local newspapers	i) arts; ii) commerce; iii) reports; iv) science; v) social; vi) sports
21. Tabloid newspapers	
22. Non-academic	i) humanities; ii) medical/health matters; iii) natural sciences; iv) politics law education; v) social & behavioural sciences; vi) technology computing engineering
23. Popular magazines	
24. Religious texts excluding philosophy	

3.3 The structure of the Written BNC2014

When looking at actual subdivisions of the original BNC categories, we note a large fluctuation in the terminology used. 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 (Trosberg 1997). The definitions of such terms are often unclear, overlap, and are used differently by different linguists as noted by Biber and Conrad (2009: 21): "the terms register, genre, and style have been central to previous investigations of discourse, but they have been used in many different ways" largely because "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). In this article, we use the composite term genre/register to refer to the main subdivisions in the BNC1994 and the BNC2014 design. This term points to the need to recognise the external situational/contextual features, which helps us identify important categories of texts as well as the linguistic motivation to study the variation within these categories (e.g. Biber 1989).

In the design of the Written BNC2014 we referred to Lee's (2001) BNC1994 genre classification scheme to make sure we did not omit an important category. At the same time, following the principles of the original compilers (cf. Aston 2001 and Burnard 2007), we kept the Written BNC2014 genre/register categories fairly broad to allow sufficient amount of data per each category. We also used recent research on genres/registers available online (Biber et al. 2015; Egbert et al. 2015) to establish how the original set of genres/registers developed over the period of more than 20 years since the BNC1994 was constructed. Table 4 provides an overview of genre/register categories that occur on the web.

In terms of changes in the language, we have added a new genre/register to the BNC2014 – E-language, which was almost absent from the BNC1994 but which, had it been left out of the BNC2014, would have represented a serious omission. Accordingly, we split the Written BNC2014 into six major genres/registers (academic prose, fiction, newspapers, magazines, E-language and Other). This in turn required us to consider the subgenres/subregisters of E-language and to produce a framework for sampling them. We were able to offer further subdivision of the individual major genre/register categories. Table 5 shows the composition of the E-language medium. More generally, the table summarises the categorisation of the Written BNC2014 texts we have arrived at, having considered Lee's (2001) scheme, its criticism (Aston 2001) as well as changes to the English language over the period of 20 years between the BNC1994 and the BNC2014 (Biber et al. 2015; Egbert et al. 2015).

Table 4: Genre/register categories of web documents (Biber et al. 2015; Egbert et al. 2015).

Genre/register	Subgenre/subregister
Narrative	News report/blog; Sports report; Personal/diary blog; Historical article; Travel blog; Short story; Novel; Biographical story/history; Magazine article; Obituary; Memoir; Other narrative
Opinion	Opinion blog; Review; Religious blog/sermon; Advice; Letter to the editor; Self-help; Advertisement; Other opinion
Informational description/ explanation	Description of a thing; Informational blog; Description of a person; Research article; Abstract; FAQ about information; Legal terms and conditions; Course materials; Encyclopedia article; Technical report; Other informational description/explanation
Interactive discussion	Discussion forum; Question/answer forum; Reader/viewer responses; Other interactive discussion
How-to/instructional	How-to; Recipe; Instructions; FAQ about how-to; Technical support; Other how-to/instructional
Informational persuasion	Description with intent to sell; Persuasive article or essay; Editorial; Other informational persuasion
Lyrical	Song lyrics; Poem; Prayer; Other lyrical
Spoken	Interview; Transcript of video/audio; Formal speech; TV/movie script; Other spoken

Table 5: Written BNC2014 genres/registers and subgenres/subregisters.

Genre/register	Subgenre/register
1. Academic prose	i) arts and humanities; ii) medicine; iii) natural science; iv) politics, law and education; v) social science; vi) technical and engineering
2. Fiction	i) poetry; ii) general prose; iii) prose for children and teenagers; iv) science fiction and fantasy; v) crime; vi) romance
3. Newspapers (serious, mass market and regional)	i) arts and entertainment; ii) commerce; iii) editorial; iv) reportage; v) science; vi) lifestyle; vii) sports
	i) arts and entertainment; ii) commerce and business; iii) editorial; iv) reportage; v) science; vi) lifestyle; vii) sports
	i) arts and entertainment; ii) commerce and business; iii) editorial; iv) reportage; v) science; vi) lifestyle; vii) sports
6. Magazines	i) lifestyle; ii) mens' lifestyle; iii) TV and film; iv) motoring; v) food; vi) music; vii) science and technology
7. E-language	i) tweet; ii) facebook post, iii) blogpost; iv) discussion forum; v) email; vi) SMS instant message; vii) online review
8. Other	i) television script; ii) drama script; iii) Hansard; iv) miscellaneous

The scheme in Table 5 thus offers continuity with the Written BNC1994, allowing comparability between the 2014 and 1994 versions of the BNC to be more clearly achieved at the level of the six major genre/register categories (Table 6). While we relabel some genres/registers to match changes in production practises in the UK, notably with regard to newspapers, the core categories remain the same. These core categories, although, strictly speaking, not considered in the design of the 1994 dataset (see Section 3.2) are broad enough to be represented by a substantial number of texts in both the Written BNC 1994 and the Written BNC2014. As an example of relabelling, consider how the practise of serious newspapers being printed in broadsheet format while the popular press being produced in tabloid form has now almost wholly passed. Rather than echoing this past practise, we use *serious* and *mass market* as the new descriptors for these genres/registers, though the mapping of these genres/registers to *broadsheet* and *tabloid* in Lee's (2001) scheme of the Written BNC1994 is perfect.

Deciding the classification system for the texts was a crucial step in the design of the corpus. Resolving that issue allowed us to begin to design the corpus itself in a way which permitted comparability[9] to the Written BNC1994 where that is possible, but which also reflected changes in the production of written English and the range of data available in 2014 relative to 1994. In doing this, we reflected on the important aspects of corpus design highlighted in the literature (e.g. Baker 2009; Biber et al. 1994; Jakubíček et al. 2013; Křen et al. 2016; Leech 2007; Lüdeling 2011; Mair 1997). Yet the corpus design was only the first step towards creating the Written BNC2014. Having developed the corpus design, we then needed to take the next step and define the population to be sampled.

3.4 Population definition and sampling frame for the Written BNC2014

The population (universe) for the Written BNC2014 can be defined quite simply as 'all texts written using British English around the sampling point of 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

9 Comparability of corpora is a complex issue. In one sense, all corpora are trivially comparable (can be compared); in another sense, comparability can be problematic: when corpora are compared, the sources of variation between two corpora are not always apparent. With current corpus tools, however, it is possible to trace the frequencies of linguistic features in different parts of a corpus (texts, subcorpora) and thus better identify the sources of variation.

population as the definition is very broad but has specific boundaries in terms of the time it was produced and the language of the producer. Nonetheless we needed to clarify what we meant by 'written around the sampling point of 2014' – we wished not merely to collect texts published in 2014; we also wanted the texts to be ones which we might presume had been written in or around 2014.

To further specify the somewhat imprecise word 'around', we collected data in the time window of 2010–2019, with 2014 being roughly the midpoint. We wished the corpus to be as contemporary as possible, hence we pulled the sampling window forward in time as far as we could before resorting to material prior to 2014. Our choice of 2014 as the anchor year for our corpus texts was principled – this was the year when collection of the Spoken BNC2014 (Love et al. 2017) began and we wanted to facilitate a meaningful synchronous comparison of written and spoken present day English, hence keeping the date ranges similar between the spoken and written corpora was clearly desirable.

Another area we had to define for the purposes of the Written BNC2014 design is British English. This, in the context of the corpus, is understood as English either produced by a British English L1 (native) speaker or produced at a place or institution where the use of British English would be expected. Thus, for instance, all authors of fiction books in the Written BNC2014 are L1 speakers of the British variety of English; this was established by checking their author profiles. On the other hand, British English in the academic prose genre/register was established by sampling published work, whose author was found to be associated with a British University. Similarly, for open/anonymous E-language genres/registers such as reviews, we sampled sites and forums within the .uk domain.

In an ideal case, the sampling frame consists of a complete list of units u in the population U. Such a sampling frame would include all texts written using British English between 2010 and 2019. For some genres/registers, obtaining such a sampling frame was relatively easy. For example, we used a complete list of dates within the time window to randomly sample the newspapers published within our three categories (serious, mass market and regional). Similarly, we could rely on fairly comprehensive lists of books and academic journals to provide the selection for the corpus. On the other hand, other genres/registers such as E-language did not lend themselves so easily to this approach; in these cases, we used a strict bias monitoring protocol (see Section 3.1) to ensure the quality of the sample. This brings us to arguments made by Hunston (2008), Bauer and Aarts (2000) and Atkins et al. (1992) that delimiting the population to be represented by a corpus is often impossible because there are no lists of items within a population. While the availability of many texts and their indexes online makes the compilation of a sampling frame much more feasible than it used to be for the designers of the BNC1994, there is still a long way to go. A corpus designer thus often needs to strike a balance between what is desirable (e.g. a stratified random sample) and what is

practicable (a balanced and unbiased sample). When collecting texts for the Written BNC2014 we thus applied stratified random sampling where possible, supplemented by targeted data gathering for categories where the population definition was not available.

3.5 Proportions of genres/registers in the Written BNC2014

In building the Written BNC2014 we balanced comparability with the Written BNC1994 with the goal of making the corpus as representative of present day written English as possible. Where these goals came into conflict, we prioritised the goal of representing present day English, realising that users could subsample the final corpus to make it directly comparable to the Written BNC1994 if they wished. So, for example, they could exclude the E-language component if they wished in order to force greater comparability. However, if their goal was to explore present day English, they could not easily, or perhaps at all, get access to a balanced E-language subcorpus, which the Written BNC2014 provides. Overall, our goal was to build a corpus where users can study written British English at the beginning of the 21st century; they can run different comparisons with the Written BNC1994 at the level of comparable genres/registers in order to investigate how English has developed over more than 20 years which lie between the two corpora.

Consequently, the proportions of genres/registers in the Written BNC2014 are not the same as in the Written BNC1994, both due to the inclusion of new genres/registers and also due to a number of practical considerations. The inclusion of the E-language genre/register in particular has meant that the proportions of other genres have had to be re-calibrated compared to the Written BNC1994. We also intended to build a corpus, which will be more balanced overall in terms of the equal representation of the main genres/registers.

In terms of the proportions of the genres/registers within the Written BNC1994 and the Written BNC2014 corpus design, some have increased, some have decreased and some have stayed much the same. Table 6 shows this comparison.

Note that for this comparison we have merged some of the BNC1994 genres/registers (Lee 2001) into larger genres/registers in order to make the data comparable, e.g. we have combined multiple genres/registers in the 'Other' category. The proportions are calculated on the basis of the whole 100-million-word dataset, which in both cases includes a 10-million-word spoken component.

As can be seen from Table 6, the proportion of fiction texts has increased slightly in the 2014 corpus design to represent their "influential cultural role." (Burnard 2000: 7). Academic prose has also increased for similar reasons. The proportion of newspapers (including broadsheet, regional & local and tabloid) and

Table 6: Proportions of genres/registers in the Written BNC1994 and the Written BNC2014.

Genre/register	Proportion (BNC1994)	Proportion (BNC2014)
1. Academic prose	16.5%	20%
2. Fiction	16.6%	20%
3. Newspapers (Serious, mass market and regional)	9.6%	20%
4. Magazines	7.6%	20%
5. E-language	0.2%	5%
6. Other	39.5%	5%
7. [Spoken	10%	10%]

magazines has more than doubled in the new corpus design. This is to address the imbalance of newspaper and magazine 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 latter comprised only 0.7%). In BNC2014, we included equal proportions of serious, mass market and regional newspaper data. The category 6 (Other) has decreased due to more meta-data being available for the 2014 corpus; in addition, some categories (e.g. personal letters) were no longer relevant and were therefore excluded due to the streamlining process whereby only substantial genre/register categories were sampled.

E-language has, of course, increased in the 2014 corpus design because there were only a handful of texts in the 1994 corpus which could be categorised as E-language.[10] As a consequence of the desire for each genre/register within the corpus to be useful as an object of study in its own right, we made sure that each genre/register comprises sufficient amount of linguistic evidence for a meaningful analysis.

4 Conclusion

This article has introduced the design of the Written BNC2014 corpus and has discussed the impact 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 far as is practically possible, and this is reflected in the design of the corpus. The population has been clearly defined, anchored to the date

10 There are merely seven emails included in the Written 1994 (text IDs: J1C, J1D, J1E, J1F, J1G, J1H and J1J), all coming from the Leeds United email list.

of 2014 as an approximate mid sampling point. Yet while the definition provides clear boundaries for what is and is not included in the population, this has limited use in reality because, as many other linguists have pointed out, it is impossible to create an exhaustive list of members of the population. This impacts on the sampling methods which were employed in the creation of the corpus. Where possible, in approximately 60% of the corpus, we endeavoured to use a stratified random sampling 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. We have clearly documented all design decisions, which will be available in the corpus manual.

The decisions taken in the creation of the corpus design ensure that the Written BNC2014 is broadly comparable to the Written BNC1994. The Written BNC1994 and the Written BNC2014 corpus design contain mostly the same major genres/registers, with the notable addition of the 'E-language' section in the 2014 corpus. The proportions of the genres/registers in the 1994 corpus and the 2014 corpus design vary somewhat due to the addition of these new genres/registers and our desire to provide a better balance of data across the different categories, as shown in Table 6. Nonetheless, the proportions of data in the different subdivisions of the corpora are broadly similar. The design of the corpus also allows users, if they wish, to fashion a more comparable subcorpus from the Written BNC2014 if comparability with the Written BNC1994 is their primary aim.

Acknowledgements: We would like to thank a number of researchers who helped collect data for the Written BNC2014, in particular Carmen Dayrell, who collected large amounts of data across different genres/registers. In addition, we would like to thank Zoe Broisson, Mathew Gillings, Andressa Rodrigues Gomide, Vasiliki Simaki, Matt Timperley, Isolde van Dorst and Pierre Weill-Tessier. William Platt, Hanna Schmueck and Maggie Mi assisted with data processing.
Research funding: This work was supported by the Economic and Social Research Council (grant number EP/P001559/1, ES/K002155/1 and ES/R008906/1).

References

Ardilly, Pascal & Yves Tillé. 2006. *Sampling methods: Exercises and solutions*. New York: Springer.
Aston, Guy. 2001. Text categories and corpus users: A response to David Lee (commentary). *Language, Learning and Technology* 5(3). 73–76.
Atkins, Sue, Jeremy Clear & Nicholas Ostler. 1992. Corpus design criteria. *Literary and Linguistic Computing* 7(1). 1–16.
Baker, Paul. 2009. The BE06 corpus of British English and recent language change. *International Journal of Corpus Linguistics* 14(3). 312–337.

Bauer, Martin & Baas Aarts. 2000. Corpus construction: A principle for qualitative data collection. In Martin Bauer & George Gaskell (eds.), *Qualitative researching with text, image and sound*, 19–37. London: Sage.

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

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

Biber, Douglas & Susan Conrad. 2009. *Register, genre, and style*. Cambridge: Cambridge University Press.

Biber, Douglas, Edward Finegan & Dwight Atkinson. 1994. ARCHER and its challenges: Compiling and exploring A Representative Corpus of Historical English Registers. In Udo Fries, Peter Schneider & Gunnel 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, Douglas, Jesse Egbert & Mark Davies. 2015. Exploring the composition of the searchable web: A corpus-based taxonomy of web registers. *Corpora* 10(1). 11–45.

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

Burnard, Lou. 2000. Reference guide for the British National Corpus (World Edition). Available at: http://www.natcorp.ox.ac.uk/archive/worldURG/urg.pdf.

Burnard, Lou. 2007. BNC User Reference Guide. Available at: http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html.

Cramer, Duncan & Dennis Howitt. 2004. *The Sage dictionary of statistics: A practical resource for students in the social sciences*. London: Sage.

Egbert, Jesse, Douglas Biber & Mark Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology* 66(9). 1817–1831.

Ellis, Nick C., Rita Simpson-Vlach & Carson Maynard. 2008. Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *Tesol Quarterly* 42(3). 375–396.

Elsevier. 2021. Open access licenses. https://www.elsevier.com/about/policies/open-access-licenses (accessed 15 July 2021).

Gardner, Dee & Mark Davies. 2007. Pointing out frequent phrasal verbs: A corpus-based analysis. *Tesol Quarterly* 41(2). 339–359.

Garside, Roger & Tony McEnery. 1993. Treebanking: The compilation of a corpus of skeleton-parsed sentences. In Ezra Black, Roger Garside & Geoffrey Leech (eds.), *Statistically-driven computer grammars of English: The IBM/Lancaster approach*, 17–35. Amsterdam: Rodopi.

Gov.uk. 2017. Intellectual property – guidance. Exceptions to copyright. https://www.gov.uk/exceptions-to-copyright (accessed 22 January 2021).

Henry, Gary T. 1990. *Practical sampling*, vol. 21. Newbury Park: Sage.

Hoffmann, Sebastian. 2007. *Grammaticalization and English complex prepositions: A corpus-based study*. London: Routledge.

Hunston, Susan. 2008. Collection strategies and design decisions. In Anke Ludeling & Merja Kytö (eds.), *Corpus linguistics: An international handbook*, 154–167. Berlin: Walter De Gruyter.

Ide, Nancy. 2008. The American national corpus: Then, now, and tomorrow. In Michael Haugh, Kathryn Burridge, Jean. Mulder & Pam Peters (eds.), *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages*, 108–113. Sommerville, MA: Cascadilla Proceedings Project.

Jakubíček, Miloš, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý & Vít Suchomel. 2013. The TenTen corpus family. In Andrew Hardie & Robbie Love (eds.), *Corpus linguistics 2013 abstract book*, 125–127. Lancaster: UCREL.

Kennedy, Graeme. 2002. Variation in the distribution of modal verbs in the British National Corpus. In Randi Reppen, Susan M. Fitzmaurice & Biber Douglas (eds.), *Using corpora to explore linguistic variation*, 73–90. Amsterdam: John Benjamins.

Křen, Michal, Václav Cvrček, Tomáš Čapka, Anna Čermáková, Milena Hnátková, Lucie Chlumská, Tomáš Jelínek, Dominika Kováříková, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Michal Škrabal, Petr Truneček, Vondřička Pavel & Adrian Jan Zasina. 2016. SYN2015: Representative corpus of contemporary written Czech. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), 2522–2528. Portorož, Slovenia: ELRA.

Lee, David Y. W. 2001. Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language, Learning and Technology* 5(3). 37–72.

Leech, Geoffrey. 2007. New resources, or just better old ones? The Holy Grail of representativeness. In Marianne Hundt, Nadja Nesselhauf & Carolin Biewer (eds.), *Corpus linguistics and the web*, 133–150. Amsterdam: Rodopi.

Leech, Geoffrey, Roger Garside & Michael Bryant. 1994. CLAWS4: The tagging of the British National Corpus. (Proceedings of COLING 1994, Volume 1), 622–628. https://www.aclweb.org/anthology/C94-1103.pdf (accessed 22 January 2021).

Leech, Geoffrey, Rayson Paul & Andrew Wilson. 2001. *Word frequencies in written and spoken English: Based on the British National Corpus*. London: Routledge.

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina & Tony McEnery. 2017. The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics* 22(3). 319–344.

Lüdeling, Anke. 2011. Corpora in linguistics: Sampling and annotation. In Karl Grandin (ed.), *Going digital. Evolutionary and revolutionary aspects of digitization*, 220–243. New York: Science History Publications.

Mair, Christian. 1997. Parallel corpora: A real-time approach to language change in progress. In Magnus Ljung (ed.), *Corpus-Based Studies in English: Papers from the Seventeenth International Conference on English-Language Research Based on Computerized Corpora* (ICAME 17), 195–209. Amsterdam: Rodopi.

McEnery, Tony & Richard Xiao. 2004. Swearing in modern British English: The case of fuck in the BNC. *Language and Literature* 13(3). 235–268.

Nation, Paul. 2004. A study of the most frequent word families in the British National Corpus. Paul Bogaards & Batia Laufer. In *Vocabulary in a second language: Selection, acquisition, and testing*, 3–13. Amsterdam: John Benjamins.

Nunan, David. 2008. Exploring genre and register in contemporary English. *English Today* 24(2). 56–61.

Taavitsainen, Irma. 2001. Changing conventions of writing: The dynamics of genres, text types, and text traditions. *European Journal of English Studies* 5(2). 139–150.

Trosberg, Anna. 1997. Text typology: Register, genre and text type. In Anna Trosberg (ed.), *Text typology and translation*, 3–23. Amsterdam: John Benjamins.

Van Dijck, José. 2013. *The culture of connectivity: A critical history of social media*. Oxford: Oxford University Press.
Váradi, Tamás. 2001. The linguistic relevance of corpus linguistics. In Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie & Shereen Khoja (eds.), *Proceedings of the Corpus Linguistics 2001 Conference* (UCREL Technical Papers 13), 587–593. Lancaster: Lancaster University.

Bionotes

Vaclav Brezina
The Department of Linguistics and English Language, Lancaster University, Lancaster, UK
v.brezina@lancaster.ac.uk

Vaclav Brezina is a Senior Lecturer at the Department of Linguistics and English Language and a member of The ESRC Centre for Corpus Approaches to Social Science, Lancaster University. His research interests are in the areas of corpus design & methodology and statistics.

Abi Hawtin
Research and Impact Services, University of Warwick, Coventry, UK

Abi Hawtin completed her PhD research at the ESRC Centre for Corpus Approaches to Social Science, Lancaster University. Her thesis focused on the design of the written BNC2014.

Tony McEnery
The Department of Linguistics and English Language, Lancaster University, Lancaster, UK
Xi'an Jiaotong University, Xi'an, China
https://orcid.org/0000-0002-8425-6403

Tony McEnery is Distinguished Professor at the Department of Linguistics and English Language and the Founding Director of The ESRC Centre for Corpus Approaches to Social Science, Lancaster University and a Changjiang Professor at Xi'an Jiaotong University. He has been working for over 30 years to help pioneer new ways to use computers to analyse very large collections of language data.