RiCL Research in Corpus Linguistics

Addressing comparability and retrieval issues in conversation corpora: A case study on the *Spoken British National Corpora* (1994 and 2014), using the past perfect

Nicholas Smith[a] – Cristiano Broccias[b] – Cathleen Waters[c]
University of Leicester[a] / United Kingdom
University of Genoa[b] / Italy
Independent researcher[c]

Abstract – This paper addresses issues in comparison and analysis of conversation corpora. We focus on the demographically-sampled spoken portions of the *British National Corpora* (BNC), representing British English in 1994 and 2014, for the purposes of studying recent language change and sociolinguistic variation. Issues of comparability and representativeness of the two BNCs have been raised before (see Love 2020), with several measures taken to ensure backwards compatibility of the Spoken BNC2014 with its 1994 counterpart. However, we believe further considerations and solutions merit attention, relating to sampling, transcription, annotation, and corpus querying. The BNClab subcorpus (Brezina *et al.* 2018a), a sociolinguistic judgment sample derived from the parent BNCs, provides a very promising basis for analysis, although arguably its mixed geographical representativeness affects cross-time comparability. To address this, we make some proposals for modifying the BNClab subcorpus to improve comparability. Then, we use the modified sample to address issues in retrieval and quantification of grammatical constructions in the spoken BNCs, namely a) determining an appropriate frequency metric, b) retrieving a comprehensive but manageable set of examples from 'messy' spoken data, and c) handling transcription inaccuracies. Finally, we discuss the case study findings and wider methodological implications for users of these corpora.

Keywords – spoken BNCs; corpus comparability and representativeness; grammatical retrieval; precision and recall; past perfect

1. INTRODUCTION[1]

This paper addresses two related kinds of challenge: a) comparability issues in spoken language corpora, and b) issues in retrieval and analysis of grammatical constructions in such corpora. While these issues are discussed in relation to the conversational

[1] We thank the anonymous reviewers and the editor for their comments; and Sebastian Hoffmann, Vaclav Brezina, Agneta Svalberg, and Julie Norton for useful discussions.

components of the *British National Corpus* (BNC)[2] from 1994 and 2014, they could potentially apply to any pair or set of corpora separated in time.

Both iterations of the BNC offer countless opportunities for anyone interested in the English language (researchers, teachers, students, or laypeople) to explore patterns of authentic British English speech, variation across registers and speakers, and changes over time. Each corpus has been planned and documented in detail (see Burnard 2007, for BNC1994, and Love *et al.* 2017 or Brezina *et al.* 2021, for BNC2014). The demographically sampled conversation components of the corpora, widely known as BNC1994DS and BNC2014S, are among the largest collections of naturally-occurring spoken discourse currently available for free public use, and they support investigations into associations between language use and speakers' social characteristics.[3] Studies exploiting the affordances of BNC1994DS include Anderwald (2002) and Rühlemann (2007), while comparisons between BNC1994DS and BNC2014S appear in McEnery *et al.* (2017) and Brezina *et al.* (2018b).

While both conversational BNCs offer impressively large quantities of material from across the social spectrum (e.g., by region, gender, age, occupation), their compilers have readily admitted that the representation of these groups is somewhat uneven (Burnard 2007; Love *et al.* 2017; Love 2020). This is understandable given the unprecedented nationwide scale of each corpus, yet the limited resources to build them, the difficulty of implementing a strict sampling procedure at the outset, and the prioritization of each project to represent users and uses of British English at their respective time, and to adopt the latest standards of corpus categorization (cf. Crowdy 1993 on BNC1994DS and Love *et al.* 2017 on BNC2014S). However, anyone using these corpora needs to be aware of their limitations in terms of comparability and representativeness. We are not the first to raise this point (see Axelsson 2018; Love 2020), but we expand on these comparability issues and highlight new ones, alongside proposals for mitigating them.

Corpus comparability is, on the face of it, much simpler than representativeness, but no less important. As Gablasova *et al.* (2017: 137) state,[4] it refers to "the degree to which two corpora are similar… [in] represent[ing] different genres of a language or

[2] http://www.natcorp.ox.ac.uk/
[3] The Spoken BNC1994 also contains a 5–million-word non-conversational, context-governed component (Burnard 2007), which we leave aside for comparability reasons.
[4] Gablasova *et al.* (2017) treat 'genre' as equivalent to what other studies, including ours, call 'register'.

speakers." The extent to which we can meaningfully interpret differences in results between corpora clearly hinges on an equal footing in these parameters. Comparability is arguably a more acute issue for spoken than written corpora since, in general, spoken language changes more rapidly than written language (Leech *et al.* 2009). As to representativeness, Biber's (1993: 243) much-cited definition —"the extent to which a sample includes the full range of variability within a population" — is non-controversial, but open to different implementations. It can refer, for example, to situationally-defined registers (e.g., job interviews or job advertisements), speaker demographics, or the distributions of linguistic features (see section 2.1). To instill more confidence in any reported results, Egbert *et al.* (2022) exhort corpus compilers and users to be more explicit in estimating the extent to which their corpus represents what it purports to represent. Moreover, concerns for representativeness and comparability can easily clash. For instance, in striving to make a corpus representative of the registers or demographics of its time or in infusing it with state-of-the-art design features and standards, the risk grows of limiting opportunities for direct comparison with a corpus from another point in time.

Fortunately, the compilers of BNC2014S have taken measures to support backward compatibility with BNC1994DS, such as issuing a list of mappings between the respective social class categories and age categories (see section 2.1). Moreover, the size of each spoken BNC and the detailed metadata provided on speaker and register characteristics afford innumerable ways to subsample and suit different research purposes (see Love 2020 and Brezina *et al.* 2021), including boosting comparability. In this paper, we evaluate one such subsample, the BNClab subcorpus (Brezina *et al.* 2018a), and present some proposals for enhancing diachronic comparability within it.

Two further issues affecting comparisons between the two spoken BNCs, which (to our knowledge have not been discussed, are differences in a) transcription quality and, potentially at least, b) grammatical annotation. With alignment of the transcriptions in BNC1994DS to its speech recordings now available (Coleman *et al.* 2011), a surprisingly large number of transcription errors in this corpus can be found. Quality control measures in BNC2014S inspire greater confidence in its transcription accuracy, despite its audio files being publicly unavailable (see section 2.2). One therefore has to decide whether discarding false positives from BNC1994DS will undermine comparisons with BNC2014S. In this paper, we present a strategy that mitigates the impact of this problem. Regarding grammatical annotation, although both corpora use the same part-of-speech

(POS) tagging software, the final output may differ in the delicacy of the tags displayed, depending on the version being used.

To illustrate challenges in spoken corpus comparability and analysis and ways to overcome them, we provide a case study on the past perfect, as in (1):

(1) That's the first time you *'d met* her? (BNC2014 S6HP:S0303)[5]

The past perfect may seem an odd choice for a case study, as it is relatively uncommon in the tense and aspect system of English, as well as particularly infrequent in conversation (Mindt 2000). Yet recent studies report that the past perfect is undergoing a dramatic change, with significant declines in both spoken (Bowie *et al.* 2013; Smith and Waters 2019) and written English (Yao and Collins 2013). However, currently, there is no specific evidence of change in the register of conversation. Bowie *et al.*'s (2013) study examines the registers of the *Diachronic Corpus of Present-Day Spoken English* (DCPSE)[6] collectively rather than individually, and is limited to the late twentieth century. Using a corpus of biographical interviews from the popular BBC Radio programme *Desert Island Discs*, Smith and Waters (2019) find a small but significant decline of the past perfect between the 1980s and the early 2000s. Moreover, they note that the construction is socially stratified, with older and more highly educated speakers being more conservative in its use. An investigation of the past perfect in the conversational BNCs also illustrates typical challenges in retrieval and quantification of grammatical constructions in spoken discourse, notably:

a) Determining the unit of frequency measurement for the target construction and accounting for competitor constructions (see section 4.1).

b) Designing corpus queries for acceptable recall and precision of the target construction in sometimes 'messy' spoken data (see section 2.3).

c) Filtering out superficially similar vernacular forms. For instance, in informal British English, *have got* is commonly used with stative meaning, as in *she'd got a family* (BNC1994: PS25A), and this is easily mistaken for a past perfect (see section 4.4).

d) Addressing errors in corpus transcription (see section 2.2).

[5] Corpus references are to the parent BNC filename and speaker identifier.
[6] https://www.ucl.ac.uk/english-usage/projects/dcpse/

Our methodology addresses general areas of change and social variation in the past perfect. The research questions underpinning the case study are:

1. Has the frequency of the past perfect changed in recent everyday conversational British English?
2. What patterns of sociolinguistic change and variation are evident in recent British conversational use of the past perfect?

By offering an up-to-date picture of the frequency of the construction in British English, our research also has potential implications in applied linguistics, particularly in English language teaching (ELT). If, for example, our results support earlier studies on spoken British English by finding a substantial decline in conversational use, there is arguably a case for reducing attention to the past perfect in teaching materials and curriculum development. Conversely, if the past perfect is found to be dramatically expanding in contemporary use, it would seem worthwhile to share this discovery with ELT publishers, teachers, and learners (cf. Curry *et al.* 2022). Learners might also benefit from awareness-raising of the prevalence of alternatives to the past perfect (see section 4.1) and how to locate them in spoken corpora.

The paper is organized as follows. We first expand on key concepts, including representativeness, comparability of corpora, precision, and recall (section 2). We then describe how we negotiated the challenges summarized in points i) to iv) above, in pursuit of a more level playing field to compare the two conversational BNCs (sections 3 and 4). Finally, we discuss preliminary corpus findings on the past perfect, including their implications and limitations, and comparison to previous studies (section 5).

2. Representativeness, comparability, precision, and recall

In this section, we review theoretical and practical aspects of representativeness and comparability, including previous attempts to address them in the two conversational BNCs. We then consider concepts relevant to retrieval of linguistic features from corpora, namely precision and recall, and technological means to boost them.

2.1. Representativeness and comparability

As suggested above, representativeness and comparability are key concepts that need to be considered in any cross-time comparison of corpora, not just the spoken BNCs. In corpus linguistics, where written data has generally been prioritized, representativeness typically refers to the extent to which the corpus reflects *situational* variation (e.g., in communicative purpose or level of interactivity) within and/or across its component text registers (Biber 1993). It tends to be only in the register of conversation that corpus linguistic studies shift focus to *demographic* representativeness (Smith and Waters 2019). In sociolinguistics, by contrast, demographic representation is a prime concern, with sampling of speakers designed to reflect the social diversity in the community investigated (Sankoff 2005), but typically on a local rather than a national scale. Thus, both disciplines use a form of stratified sampling, one focused on texts as the sampling units, the other on speakers. One further kind of representativeness to note is *linguistic* (or distributional) representativeness, that is, the extent to which the corpus "includes the range of linguistic distributions in the population" of texts or speakers (Biber 1993: 243). Egbert *et al.* (2022) lament that many corpus linguistic projects fail to evaluate the representativeness of their corpus relative to their research goals. At the same time, they argue that representativeness is a matter of degree rather than an all-or-nothing construct, and that full representativeness is an idealized target and unattainable in practice (Egbert *et al.* 2022: 12).

Comparability is less explicitly discussed than representativeness in either the sociolinguistic or the corpus literature. Gablasova *et al.* (2017) identify comparability as a major issue in corpus-based second language acquisition research, where direct comparisons have been routinely drawn between the language use of L2 and L1 speakers but without paying attention to potentially confounding factors, such as type of elicitation task and L1 speakers' language proficiency. The tension between representativeness and comparability has arguably received more attention in diachronic corpus studies. Leech and Smith (2005) describe the challenge of extending the design model of the Brown and the LOB corpora (sampling date: 1961; Hofland *et al.* 1999) back to the 1930s and earlier, when genres such as science fiction and academic subdisciplines, e.g., sociology, were far less established. Baker (2023) encounters the reverse challenge in extending the Brown corpus model to British English published in 2021 and, to optimize comparability, he excludes new genres such as horror fiction, which did not exist in the 1960s.

In diachronic corpus studies, a balance also needs to be struck between comparability and contemporaneity of standards. Advances in computer hardware and corpus software, and improved standards of metadata, can lead the creators of a newer corpus to depart from the best practice of an earlier corpus. Annotation standards, e.g., the kinds of distinctions used in part-of-speech (POS) tagging, can also vary from one corpus to another, sometimes even when the same annotation software is used (see section 2.3 and section 4.2).

The representativeness and comparability of BNC1994DS and BNC2014S are discussed in Love (2020: 186–189). He notes, for example, that neither corpus uses strict stratified sampling. In BNC1994DS, only the speakers with recording responsibility were sampled in advance (by a random method), but the other speakers, like all speakers in BNC2014S, were selected opportunistically. Constraints of budget and time made it impossible to obtain balanced representation of social groups. For example, male speakers in BNC1994DS are over-represented in relation to females and to their proportions in the UK population as a whole. Conversely, females in BNC2014S are over-represented. Regarding age, in BNC1994DS speakers aged 25–59 are over-represented relative to the UK population, while in BNC2014S, those aged 19–29 proliferate. In both corpora, speakers from England are represented far better than those from Scotland, Wales, and Northern Ireland. Clearly, there are significant differences in the composition of the two spoken BNCs, and if due attention is not given to these differences, there is a risk of drawing naive conclusions from cross-time comparisons.

Love *et al.* (2017) describe measures taken to support backwards compatibility, and therefore comparability, of metadata categories between the newer and older corpus. For example, they provide a mapping list between the more fine-grained age bands of BNC2014S into those of BNC1994DS. Similarly, they translate the nine socioeconomic class categories (NS-SEC) in BNC2014S into the four social grade categories used in BNC1994DS. In terms of annotation, however, the corpora differ in that BNC2014S was tagged using a more fine-grained set of POS-tags than BNC1994DS. Further differences in POS-tagging are described in section 4.2.

2.2. Previous comparative studies of the conversational BNCs

In two edited collections of papers on the conversational BNCs (McEnery *et al.* 2017 and Brezina *et al.* 2018b), several contributors discuss comparability issues between BNC1994DS and an early sample release of BNC2014S, including age and class regroupings. Axelsson (2018), for instance, highlights a possibly greater awareness among speakers in BNC2014S of being recorded (because of more stringent ethical requirements for prior consent), which may have led to the conversations acquiring a more focused and formal character. If this is indeed the case, it is potentially problematic for diachronic comparison, yet difficult to see how it can be overcome.

The BNClab subcorpus (Brezina *et al.* 2018a) seeks to enhance comparability by deriving a judgment sample from the two parent BNCs. Judgment sampling involves, as Schilling-Estes (2007: 169) states,

> using one's judgment to decide in advance what types of speakers to include in the study and then obtaining data from a certain number of each type.

Using the BNClab subcorpus, Reichelt (2021) uncovers changing patterns in the pragmatic markers *kind of* and *sort of* across time and social groups. Given its potential for investigating spoken language change and variation in relatively controlled conditions, we evaluate the comparability and representativeness of the BNClab subcorpus (in section 3.1), and a modified version of it (in section 3.2) used in our own study.

To our knowledge, no studies comparing the two BNCs have yet addressed the issue of transcription quality in the 1994 corpus. The issue is more pervasive and concerning than the several instances of incorrect speaker assignment noticed by Axelsson (2018). It includes numerous cases where the content roughly matches the audio but the linguistic forms are incorrect, as illustrated in (2), and cases where neither content nor form match the recording, as shown in (3). Such anomalies have come to light following a project to align the BNC1994DS transcriptions with the original sound files (Coleman *et al.* 2011), and the subsequent implementation of audio playback of concordance hits in the *BNCweb* tool (Hoffmann and Arndt-Lappe 2021).

(2) Then I **bought**, yeah, I said I feel as if I've gone deaf [Correction: Then I **thought**…]. (BNC1994 KB2:PS01U).

(3) **but, it was a pity** he was able to speak on the telephone [Correction: **well considering** he was able...]. (BNC1994 KBW:PS087).

User access to recordings of BNC2014S is not yet possible. However, there are good reasons to believe that transcription of this corpus was done far more carefully. Each transcription went through careful rounds of checking (Love 2020) and transcribers were thoroughly trained on the transcription protocols and formally registered their level of confidence in identifying the speaker of a given utterance. Likewise, the recording devices used in the mid-2010s (smartphones) were far superior to the devices used in BNC1994DS.[7]

2.3. Precision, recall, POS-tagging, and retrieval software

When querying a corpus for a linguistic feature, an important consideration is finding the right balance between recall and precision. Recall is a measure (expressed as a percentage) of the extent to which a query retrieves all valid instances of a target item in the data. In practice, recall is difficult to quantify since it requires a fully hand-edited dataset. Precision, which is also expressed as a percentage, refers to the proportion of retrieved instances that are actually valid (see Jucker *et al.* 2008). In corpus studies, it is generally agreed that low precision is more tolerable than low recall, since automated results are likely to be hand-checked, and having a near full set of examples is key to a thorough analysis (see Hoffmann *et al.* 2008). While there is no consensus as to what constitutes acceptable precision thresholds, Jucker *et al.* (2008: 277) suggest that

> precision errors are not a serious problem, until the number of hits exceeds what is possible to scan manually, and until precision falls below a certain threshold: one tends to overlook positive examples if precision is much lower than 1%.

An excessive number of hits is an important issue in our study, not for the past perfect, but for the far more prevalent past non-perfect (e.g., *took*) with which it competes (see section 4.2.3). As for precision, we typically boost precision scores well beyond one percent (even in corpora of spontaneous speech) by using a grammatically annotated version of each corpus and sophisticated corpus query tools to exploit the annotations. In our study of the past perfect in the spoken BNCs, using this combination of tools obviates the need to manually sift through 45,087 cases of *had/'d* for instances containing a trailing participle.

[7] BNC2014S also allows users to investigate individual transcriber consistency by identifying them in the metadata.

Nevertheless, two factors need to be acknowledged as affecting recall under these conditions. The first one is that automated POS-taggers make errors. While error rates are generally reassuringly low, at around 3–4 percent, they will be higher for multiply ambiguous words (e.g., *left* is a noun, an adjective, a past tense verb, or a past participle, depending on context). The second consideration is that phrasal constructions like the past perfect can be used discontinuously (Trask 1993), that is, between the auxiliary and the participle, one or more words may intervene (e.g., *she had erm already gone*). It is therefore imperative to work out a strategy for optimizing recall in discontinuous uses of a construction, particularly in the unpredictable environment of spontaneous conversation. We address this issue in section 4.2.1.

3. Obtaining a sociolinguistically-balanced dataset from the BNCs

3.1. Our starting point: The BNClab subcorpus

Developed at Lancaster University, the BNClab subcorpus samples 250 speakers from BNC1994DS and 250 speakers from BNC2014S. Unlike most sociolinguistic judgment samples, the assignment of speakers to groups for BNClab was performed *post-hoc*, after the BNCs themselves had been created, which could make the judgment harder than when selecting during data collection. The subcorpus covers all nations of the UK, an unusually large area for a judgment sample. It has near-equal gender balance (126 females and 124 males in each of 1994 and 2014), and good representation across age cohorts (Brezina *et al.* 2018a; Reichelt 2021), as also shown in the Appendix. The age groups are sufficiently fine-grained to allow studies of apparent-time change within each period.

The basis for determining social class in the BNClab subcorpus (as in the full BNCs) is the speaker's occupation. This is in line with typical sociolinguistic research (Milroy and Gordon 2003), although classifying occupations is notoriously complex, and a different approach is taken in each parent corpus. In BNC1994DS a social grade scheme is used, based on categories in the UK's *National Readership Survey* (see Love *et al.* 2017). BNC2014S, meanwhile, uses the UK government's official *National Statistics Socio-economic Classification* (NS-SEC) scheme (Love *et al.* 2017), which broadly relates to the type of contract the occupation typically involves (Atkinson 2015).[8] Love

[8] Notably, jobs involving higher specificity (and scarcity) of skills and lower ease of monitoring are found at the top of the scale (Atkinson 2015).

et al. (2017: 332) helpfully provide a mapping table between the two classification schemes. In BNClab, speakers with occupations rated as NS-SEC classes 1 to 4, or social grade AB to C1, are assigned to middle class, while speakers in NS-SEC classes 5 to 8, or social grades C2 to E, are treated as working class, although some manual adjustments were made to improve consistency.

As can be noticed in the Appendix, the BNClab subcorpus has a class imbalance, with the 2014 component markedly under-representing working-class speakers and over-representing middle-class speakers relative to the 1994 subcorpus. To some extent, these differences reflect the changing nature of British society. By the mid-2010s, an increasing portion of the UK population was university-educated and engaged in higher-status professional occupations than in the 1990s. Even so, the middle-class numbers in BNClab exaggerate the scale of this upward mobility.

Likewise, an arguably important social variable for representativeness that is not included in the BNClab subcorpus is ethnicity. An increase in ethnic diversity is another major area of change in UK society.[9] However, the lack of data about ethnicity in BNC1994DS makes the omission from BNC2014S understandable.

Finally, the creators of the BNClab subcorpus excluded all speakers who produce fewer than 1,000 words. This is the minimum that Biber (1993) reports as sufficient to profile most grammatical features in a given text, suggesting adequate distributional representativeness (see section 3.2).

3.2. Modifications to the BNClab subcorpus

While the BNClab subcorpus provides a very promising platform for sociolinguistic inquiry, the fairly balanced numbers of speakers for single social variables (e.g., gender) sometimes mask sizable differences at the granular level, for instance, at the intersection of categories such as region and social class. In the 2014 data, for example, just one of the eight speakers from Scotland is categorized as working class, with six being middle class, and one unknown. Wales has nine middle-class but just two working-class speakers from 2014. By including all four component countries of the UK, plus Ireland, the chances of getting balanced representation at the granular level are greatly reduced. This can be

[9] https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/ethnicity/articles/2011censusanalysisethnicityandreligionofthenonukbornpopulationinenglandandwales/2015-06-1

problematic when investigating linguistic features known or suspected to exhibit marked social stratification, such as the past perfect. To mitigate this and other comparability issues and to balance the three social dimensions of gender, class, and age, we adapted the BNClab subcorpus into a new modified version, hereafter BNClab-M. Six main changes were made.

Firstly, we limited the speakers to those from England. Naturally, excluding Scotland, Wales, and Northern Ireland makes this revised sample unrepresentative of the UK as a whole, but the more focused geographical selection concentrating on the largest national population of speakers gives us more scope to balance the social variables. Unfortunately, this modified sample is no more capable than the original BNClab subcorpus of evenly reflecting regional variation in England (see Beal 2010), although wherever possible we included speakers across the five English regions recognized by BNClab (North, Midlands, Southeast, London, and Southwest).

Secondly, only speakers categorized in BNClab as either working class or middle class were considered. Those with uncategorized social class were discarded. Retired people were omitted because they constitute a socially opaque group, with little in common beyond their senior age. In view of the importance of age for stratification of the past perfect (see section 5), it would be desirable to find a way to incorporate retirees at a future point, but ideally incorporating their former occupations. Students were also omitted because they are a similarly problematic group for social class assignment since, in most cases, they are not in the labor market. Similarly, trainees (e.g., trainee engineers, nurses, and typists), whose incorporation into the labor market is unknown, were discarded.

Thirdly, regarding age, we excluded children (i.e., under-18s). The number of children in the BNClab subcorpus is patchy across regions (e.g., just two children from northern England). As for adult speakers, these were categorized into two age cohorts, namely under-45s and over-45s (the latter including 45-year-olds). This is based on observations in the sociolinguistic literature, that "the speech of middle-aged adults tends to be highly conservative, often more conservative even than that of older speakers" (Milroy and Gordon 2003: 39).

Fourthly, we reclassified cases that we considered to be errors and removed speakers whose classifications seemed uncertain. To do this, we drew on various sources of information, notably the BNC1994 and BNC2014 header files (showing speaker

occupations and relationships between speakers), and a set of social grade reclassifications of BNC1994 prepared at Oxford University.[10] We moved some speakers from working class to middle class where we judged their occupation to be similar to other, well-established, middle-class roles. Examples include a chartered engineer (speaker PS1BT) and a consultant engineer (S0179). Less clear-cut but still arguably middle class are clerks, administrators, and other office workers, who in the BNC metadata tend to be assigned social grade C1 or higher, that is, (lower) middle rather than working class. Our exclusions included those in an occupation typically placed as working class but who hold a degree (e.g., a graduate chef, speaker S0603), and those with a close family member assigned to a different class (e.g., a childminder, speaker PS14B, originally recorded as working-class but whose husband is a teacher).[11] However, it must be acknowledged that it is difficult to be fully consistent across 1994 and 2014 in applying such exclusions, since the earlier BNC did not record speakers' educational level.

Fifthly, in a few cases, information from the transcription files themselves helped inform a decision on inclusion and categorization of a speaker. For example, in BNC2014S, speaker S0463 is listed as a taxi driver, with social grade C1 and NS-SEC class 4, but as middle class in the BNClab documentation. In the transcription, the speaker refers to his previous work in finance, casting sufficient doubt on his class designation for him to be excluded.

Sixthly, we targeted at least five speakers for each combined set of social characteristics, as this is a common minimum target in sociolinguistic studies (Horvath 2013: 12), although we did not always manage to reach this. For cells in short supply, we turned to the full versions of BNC1994DS and BNC2014S, specifying the relevant demographic criteria to locate nine additional speakers.

The composition of the BNClab-M sample is detailed in Table 1.

[10] We gratefully acknowledge Katie Henley's work in this area. http://www.phon.ox.ac.uk/files/docs/SpokenBNCoccupationsubgroups.xlsx

[11] Both speakers are assigned social grade AB in the BNC metadata.

			1994		2014	
Gender	**Class**	**Age**	**Speakers**	**Words**	**Speakers**	**Words**
Female	Working class	Under 45	11	39,424	5	14,852
		Over 45	4	48,052	7	49,896
	Middle class	Under 45	16	69,150	24	166,296
		Over 45	4	15,159	17	141,724
Male	Working class	Under 45	12	25,943	6	12,526
		Over 45	4	10,032	5	52,381
	Middle class	Under 45	13	37,881	21	136,705
		Over 45	10	27,262	15	121,092
Total			**74**	**272,903**	**100**	**695,472**

Table 1: Speaker numbers and word counts in the BNClab-M sample

Reflecting on the comparability and representativeness of the BNClab-M subcorpus, we are aware that we have made significant compromises to the latter in order to boost the former. Our attempts to rebalance the social group sizes has yielded some success, bringing us closer to our target minimum of five speakers in each cell. The numbers of working-class speakers are now more balanced across the periods (31 in 1994, 23 in 2014), although middle-class speakers are still significantly over-represented in 2014 versus 1994 (77 and 43 respectively). We still have near-parity of females and males (88 and 86 respectively). Age groups are reasonably balanced, the biggest discrepancy being the relatively low figure of 22 over-45s in 1994, versus 44 in 2014.[12] However, our binary classification of age, as below or above 45, hinders analysis of change in apparent time. Overall, speaker numbers are somewhat low for analyzing individual variation within social factor groups (cf. Brezina and Meyerhoff 2014). Moreover, as Sönning and Krug (2022) observe, marked differences in word count between individual speakers may skew feature frequencies by social group. The latter issue was mitigated by retaining the minimum threshold from BNClab of 1,000 words per speaker, although we did not apply an upper limit. These characteristics need to be kept in mind when reviewing the results.[13].

[12] Average ages in these groups are comparable: under-45s' mean age is 35 in 1994 and 32 in 2014; for over-45s, the corresponding means are 51 and 55.

[13] The list of speakers in BNClab-M, and their characteristics, is openly available at https://doi.org/10.25392/leicester.data.25594368.

4. GAUGING THE FREQUENCY OF THE PAST PERFECT

Accurately establishing the frequency of the past perfect in the BNClab-M sample involves several steps, notably, determining an appropriate unit of measurement, setting an effective search strategy to retrieve occurrences of the past perfect (and any relevant competing constructions) from the corpus, and manually correcting the results. We describe each of these steps in turn.

4.1. Frequency and competitors of the past perfect

In corpus studies and variationist sociolinguistic studies, the choice is whether to relativize the raw number of occurrences of a linguistic feature to:

a) the corpus size (in words), i.e., normalized frequency: for example, past perfect instances per million words; or

b) the superordinate category the construction belongs to: for the past perfect, this would be the total number of finite verb phrases; or

c) the set of variants (choices) available for conveying the same or a similar discursive function (e.g., past perfect and other expressions of past time). In variationist sociolinguistics, this set of choices is called the 'linguistic variable' (Tagliamonte 2006).

For several reasons, the third type of measurement is preferable. As Bowie *et al.* (2013) and Smith and Waters (2019) point out, it is the measure that reflects the opportunity of occurrence of the past perfect most accurately. While the first metric is relatively easy to compute, since only the past perfect needs to be counted, it is clear that not every word in the corpus provides an opportunity for the construction to occur (Ball 1994). Also, the past perfect is a multi-word (rather than a single-word) construction, and makes normalized frequencies problematic. Using the superordinate category (finite verb phrases) reduces these problems by narrowing the field to more plausible contexts but misses the fact that the past perfect is restricted to past time. The third metric addresses this problem. It also handles the problem of transcription errors in BNC1994S better (see section 2.2), since any losses of examples due to faulty transcription should affect the past perfect and its competitors equally. At the same time, it must be acknowledged that circumscribing the variable context is far from straightforward, particularly at the level

of syntax (see Lavandera 1978). It entails reviewing the functions and uses of the construction in question and its putative competitors, as well as determining their degree of functional equivalence (Tagliamonte 2006).

The basic function of the past perfect is to express the anteriority of a past situation to a reference time earlier in the past, the latter being mentioned explicitly —as in (4)— or recoverable from the context —as in (5)— (cf. Declerck 2006). Example (4) illustrates that the verb may additionally be marked for progressive aspect.

(4) By the time I arrived, everyone else *had* already *left*. (Depraetere and Langford 2019: 198)

(5) Jane got that job she interviewed for. – I*'d been wondering* about that. (Depraetere and Langford 2019: 198)

In certain conditions it is possible to replace a past perfect with a past non-perfect (that is, a past simple or a past progressive), provided that the anteriority relationship can still be inferred, as illustrated in (6).

(6) After we *finished/had finished* the meeting, we all went out for a drink. (Depraetere and Langford 2019: 198)

However, the past perfect is by no means always substitutable by the non-perfect, and because the former has the specialized meaning of anteriority built in, substitution in the opposite direction is far less feasible. Ideally, we would test the acceptability of replacing each corpus example with a non-perfect but, given the many thousands of past non-perfects in BNClab-M, this would be prohibitively time-consuming. As in Bowie *et al.* (2013), we have pragmatically opted for a looser notion of the linguistic variable than in typical variationist studies, namely all past-marked verbs. Another pragmatic decision was to exclude the present perfect from this set of choices, as shown in (7). In British English, the present perfect tends not to occur in narrative past situations. The example in (8), with a definite time specifier, is a rare exception.

(7) I*'ve given* up smoking.

(8) (…) and then Saturday I*'ve put* that one up again. (BNC2014 SF8D:S0152)

However, in the specific context of unreal past conditionals introduced by *if* or *wish*, we do need to account for non-standard perfects, for instance, the double perfect, as in (9), which is alleged to be spreading in British English (Huddleston and Pullum 2002: 151).

(9) (…) if he *had*n't *have left* our command she was gonna make a formal complaint. (BNC2014 SVD6:S0256)

Other variants of the non-standard perfect include *would* or the elided form *'d*, as in *if he wouldn't have left*, *if he'd have left*. A survey by Ishihara (2003) suggests these forms are increasingly acceptable in colloquial American English.

4.2. Retrieving a comprehensive, comparable, and manageable set of examples

We considered three corpus tools, *BNClab* (Brezina *et al.* 2018a), *CQPweb* (Hardie 2012), and *BNCweb* (Hoffmann *et al.* 2008). The latter is very similar to *CQPweb* but hardwired to BNC1994. Each of these tools permits queries based on POS-tags, either individually or in sequence. The source of those tags in each case is Lancaster's CLAWS4 automatic tagger (Garside and Smith 1997), which in spoken texts has an estimated precision rate of 97 per cent and a recall rate of 98.8 per cent (Leech and Smith 2000). However, there are some differences in the tagging implementation and search software functionality that might affect equivalent retrieval of linguistic features from the spoken BNCs. The differences are summarized in Table 2.

	BNC1994DS in a) *CQPweb* b) *BNCweb*	BNC2014S in *CQPweb*	Subcorpora of BNC1994DS, BNC2014S in *BNClab* tool
Corpus scope	Full corpus or customized subcorpora	Full corpus or customized subcorpora	BNClab subcorpus only
CLAWS tagset	C5	C6	C6
Ambiguity tags	Yes	No	No
CLAWS Spoken mode	Yes	Yes	No
Template Tagger used	Yes	No	No
Flexibility of queries	High	High	More limited
Audio playback	a) No b) Yes	No	No

Table 2: Tagging implementation across the spoken BNCs and retrieval tools

The POS-tags in BNC1994DS are from the CLAWS C5 tagset, whereas those in BNC2014S are from the more granular CLAWS C6 tagset. Personal pronouns, for instance, are represented in C5 by just one tag (PNP), whereas in C6 they have ten tags, depending on person, number, and case (e.g., PPIS2: first person plural, nominative).[14] BNC1994DS also includes ambiguity tags (Burnard 2007), which signal where the probabilities of two competing tags were estimated by CLAWS to be too close to call.

[14] A mapping list between C6 and C5 is available at https://ucrel.lancs.ac.uk/claws/mapC7toC5.txt.

The ambiguity tag VVD-VVN, for example, is non-committal about whether a given word (e.g., *looked*) is past tense (VVD) or a past participle (VVN), although the order indicates that VVD is more probable. Moreover, both BNC1994DS and BNC2014S in *BNCweb*/*CQPweb* were tagged with CLAWS run in spoken mode. This means that the disambiguation of POS-tags was improved by the use of training data extracted from previous, hand-corrected spoken corpora. In the case of BNC1994DS, though not BNC2014S, a supplementary software named *Template Tagger* (Smith 1997) was used, affording marginal improvements in tagging accuracy.

As for retrieval capabilities, *CQPweb* supports queries with flexible pattern-matching, including optional and repeatable elements. At the time of writing, the *BNClab* tool has less advanced search functionality than *CQPweb*/*BNCweb*, but does support queries based on POS. Finally, only *BNCweb* currently supports audio playback of concordance hits.[15]

Given these circumstances, and the need to optimize comparability of the two spoken BNCs, we used the *BNClab* tool with its more consistent POS-tagging for the retrieval of less complex structures in the BNClab-M sample – that is, adjacent (i.e., non-discontinuous) past perfects (with *had/'d* immediately followed by a past participle), as well as past non-perfects (e.g., *took*, or *was* in *was sleeping*). Meanwhile, for the more complex types of retrieval (i.e., discontinuous past perfects and non-standard perfects), we set up the BNClab-M subcorpus in *BNCweb* and *CQPweb*.[16] *BNCweb* additionally allowed us to verify the transcription of most examples from 1994 by listening to the audio recordings. Thus, a combination of tools helped us to optimize the recall of these structures, as we detail below.

4.2.1. Non-discontinuous past perfects

A simple *BNClab* query sufficed for retrieving straightforward past perfects, where the auxiliary (*had*/*'d*) and past participle (tagged V*N) are adjacent, as illustrated in (10).

[15] Space prevents exhaustive coverage of BNC-compatible retrieval tools. However, *Lancsbox X* (Brezina and Platt 2024) now offers a promising alternative by tagging BNC1994DS in the same (C6) POS-tags as BNC2014S. The issue remains of inability to play back audio to verify transcriptions.

[16] More precisely, we made a 1994 BNClab-M subcorpus in the *BNCweb* area of the Lancaster server, and a 2014 subcorpus in the *CQPweb* area. To maximize recall, any differences in results noticed between the tools were added to the pool of hits.

(10) VHD V*N

One difficulty in retrieving the past perfect in speech is elliptical uses, where the trailing past participle is understood but does not appear, as shown in (11).

(11) Had you done it? – Yes, I *had* (invented example).

We could not devise a query to find such cases, and so none are included in our results. For consistency, even though our query for past non-perfects picked up elliptical cases (e.g., *yes, I did*), we discarded them.

4.2.2. Discontinuous past perfects

Since we did not know all the forms of intervening material in spoken past perfects in advance, our retrieval strategy had two heuristic steps: a) determine the maximum interval between the auxiliary and the participle, and b) list all POS sequences that appear within that interval. For the first step, by repeated experiments, we found that queries containing *had/'d* and a participle separated by more than five words (and up to ten words) yielded no valid cases of the past perfect in either period.[17] Note that the query in the 1994 data allowed for ambiguity tags in either order, VVN-VVD (i.e., ambiguous, but past participle more likely) and VVD-VVN (ambiguous, but past tense verb more likely). In the data retrieved, the participle tended to be part of a separate verb phrase from the one containing *had/'d*, as exemplified in (12)

(12) (…) if they **had** an accident <pause> the people would get **killed.** (BNC1994 KCT: PS0FP)

For the second step, we ran a pair of queries in *BNCweb*/*CQPweb* (on the 1994 and 2014 parts, respectively, of the BNClab-M sample) with a five-word maximum interval between auxiliary and participle and used the *Frequency Breakdown* tool to list the types of intervening POS-tag sequence that occur. The top ten POS-sequences in each period are shown in Table 3. Note that the POS-tags listed under 1994 are the simpler C5 tags, while those under 2014 are the finer-grained C6 tags. The only ambiguity tags under 1994 have tag VVN listed first (i.e., ambiguous, but past participle deemed more likely). This

[17] The queries we ran to check this were:
a) for BNC1994: [pos="VHD"] [pos!="PUN|TO0|V.*N"]{6,10} [pos="V[BDHV]N.*|.*V[BDHV]N"]
b) for BNC2014: [pos="VHD"] [pos!="PUN|TO|V.*N"]{6,10} [pos="V[BDHV]N.*|.*V[BDHV]N"].

reassures us that the 2014 recall rate is not disadvantaged by the omission of ambiguity tags.

	BNClab-M sample: 1994				BNClab-M sample: 2014		
Rank	**POS sequence type**	**Cases**	**% of types**	**Rank**	**POS sequence type**	**Cases**	**% of types**
1	VHD XX0 VVN	69	24.0%	1	VHD XX VVN	208	23.8%
2	VHD AV0 VVN	49	17.0%	2	VHD RR VVN	111	12.7%
3	VHD AV0 VBN	15	5.2%	3	VHD XX VBN	42	4.8%
4	VHD XX0 VBN	14	4.9%	4	VHD XX RR VVN	26	3.0%
5	VHD PNP VVN	9	3.1%	5	VHD RR VBN	24	2.7%
6	VHD AV0 VHN	7	2.4%	6	VHD RR VHN	22	2.5%
7	VHD UNC VVN	5	1.7%	7	VHD RR RR VVN	17	1.9%
8	VHD PNP VDN	5	1.7%	8	VHD XX VHN	16	1.8%
9	VHD XX0 VHN	5	1.7%	9	VHD XX VDN	14	1.6%
10	VHD XX0 AV0 VVN-VVD	5	1.7%	10	VHD PPH1 VVN	11	1.3%

Table 3: Query breakdown for past perfect candidates separated by up to five words (top ten items)

To optimize recall of discontinuous past perfect, concordances of all candidate cases were hand-checked. In what follows we discuss three notable patterns.

The dominant pattern is that of negation and/or adverb modification, i.e., items containing tag XX0 and AV0 in the 1994 subcorpus, and items containing XX and RR in the 2014 subcorpus. Examples are provided in (13) and (14), the latter also including the discourse marker *like*, which is treated by CLAWS as an adverb.

(13) He probably *had***n't** *paid* that much anyway. (BNC1994 KBX:PS1DW)

(14) (…) she *had* **like literally just** *pressed* submit on her assignment. (BNC2014 SUVL:S0598)

Another pattern is that of inverted subject-verb word order in questions and conditionals, e.g., the sequences ranked 5th and 8th under the 1994 subcorpus, and 10th under the 2014 subcorpus. Typically, the subject is pronominal, as in (15).

(15) (…) *had* **I** *had* more time yesterday. (BNC2014 SGMM:S0483)

A further pattern relates to spoken disfluency features which include hesitation markers, generally transcribed as either *er* or *erm* in the BNC but are tagged differently in BNC1994DS (with C5 tag UNC, for unclassified item) and BNC2014S (C6 tag UH, for

interjection). Examples are (16) and (17).[18] Another disfluency feature the POS-tag breakdown finds is truncated words, where the speaker breaks off mid-word). These are tagged as unclassified items (UNC in C5, FU in C6), and illustrated in (18) and (19).[19]

(16) when we *had* **er <pause>** *ordered* it (BNC1994 KBW:PS08A)

(17) (…) we*'d* **erm** *exchanged* some Tesco vouchers on a couple of occasions (BNC2014 S64H:S0257)

(18) (…) cos the jacket *had* **s-** <pause> *fallen.* (BNC1994 KBF:PS04V)

(19) (…) she*'d* actually **w-** *gone* past the turning. (BNC2014 SA69:S0262)

4.2.3. Past non-perfects

Specifying a query for the past non-perfect is relatively simple in that only one item, a past-marked verb form (e.g., *took*, *was*), need be identified. The *BNClab* query in example (20) below retrieves all such verbs by conflating the C6 tags VBD, VHD, VDD, VVD, VBDM and VBDR.

(20) (V*D OR VBD*)

However, the vast number of hits this query returned (99,698) from the BNClab-M subcorpus necessitated a reduced sample for manual analysis.[20] We opted to sample one in 50 non-perfects from each period, selecting them systematically to minimize bias. The precision rate for the queries was 71 per cent for 1994 and 80 per cent for 2014.

4.2.4. Non-standard perfects

Based on the literature (e.g., Denison 1993; Huddleston and Pullum 2002), the general structure of non-standard perfects appears to be the one shown in Figure 1.

1	**2**	**3**	**4**	**5**
If/	Subject	*had*	*have*	Past participle
wish		*would*	*'ve*	
		'd	(*of*)	

Figure 1: Structure of non-standard perfect

[18] Adding to this complexity, hesitation markers are assigned the C6 tag FU (unclassified) rather than UH (interjection) in the BNClab interface.
[19] In the BNClab interface, queries skip over truncated items.
[20] We acknowledge the help of Loveen Dyall in extracting and verifying the query results.

As with the discontinuous past perfect, we can use test queries to discover the types of interpolating POS-sequences within this structure in conversation and the maximum length of the subject in words. By this process, we found that the maximum interval between *if/wish* and *had/'d* in slot 3 was four words, as illustrated in (21), while for *would* it was just two words, as shown in (22).

(21) (…) **if** Bolton and Blackburn yesterday ***had***n't *have been* such a high-profile game. (BNC2014 S9B9:S0152)

(22) (…) **if** the police **would've** raided your nan's. (BNC2014 S4YQ:S0253)

Also, it is quite common in BNC1994DS for the auxiliary verb *have* to be transcribed as *of* in representing spoken British English. Our queries took account of this, as can be seen in (23).

(23) If it *had*n't **of** *been* for Steve's car breaking down … (BNC1994 KCX:PS1FC)

Despite the non-standard character of these perfects, our queries on BNClab-M achieved good precision: 96.8 per cent on the 1994 section and 87.9 per cent on the 2014 section.[21]

4.3. Managing transcription inaccuracies

To address the issue of transcription quality in BNC1994DS (see section 2.2), we checked all retrieved examples of past perfect, past non-perfect and non-standard perfect for which audio is available (approximately 80% of cases). We removed all false positives, which included a) clear errors, where something other than the target structure can clearly be heard or the utterance is clearly attributed to the wrong speaker, and b) inaudible cases, where the target structure could not be heard in repeated listening.[22] For the putative past perfect, illustrated in example (24), what looks in the transcription like a hesitation marker (*er*) sounds on closer listening far more likely to be a reduced auxiliary (*'ve*), and therefore part of a double perfect. In (25), the official transcription has *wasn't the a big one* and is attributed to a female, yet the voice is almost certainly that of a male, and *it mustn't be a big one* is clearly audible.

[21] Query for 1994: "if|wish.*"%c []{1,4} [word="\'d|had|would"%c] [pos="XX0|AV0|UNC|ITJ"]{0,} [pos="VHI|VHB|PRF"] [pos="XX0|AV0|UNC|ITJ"]{0,} [pos="V[BDHV]N|.*V[BDHV]N"]
Query for 2014: "if"%c []{1,4} [word="\'d|had|would"] [pos="XX|R.*|FU|UH"]{0,} [pos="VHI|VH0|FU"] [pos="XX|R.*|FU|UH"]{0,} [pos="V[BDHV]N"]

[22] We did not discard cases where the audio and the transcription were irrecoverably misaligned and impossible to verify. These are equivalent to the BNC2014S cases, which lack audio.

(24) (…) if our Margaret had **er** been working. (BNC1994 KB1:PS01B)

(25) I'm almost frightened to put a crescendo in because it **wasn't the,** a big one. (BNC1994 KBH:PS05B)

Errors attributable to transcription issues in the 1994 part of BNClab-M totaled 64 for the past perfect (7.3% of candidate cases) and 28 for the past non-perfect (5.2% of candidate cases). While it is disconcerting to find so many transcription errors, it is reassuring that they affect the two target constructions in similar proportions.

4.4. Exclusions

Among the categories of hits that we excluded were errors resulting from speech disfluencies. If the target construction (past perfect/past non-perfect/non-standard perfect) verb phrase was deemed to be incomplete because of a false start, we excluded it. This is shown in (26), where the false start is highlighted.

(26) (…) **they*'d booked*** they*'d booked* a four-wheel drive (BNC2014 SMEB:S0238)

Another common source of error is the stative, simple past use of *had got*, illustrated in (27), which is actually more frequent than the use of past perfect *had got*. Such cases were discarded.

(27) Yeah but was she a woman living on her own or *had* she *got* a husband? (BNC1994 KCT:PS0FX)

5. RESULTS AND DISCUSSION

With the adjustments to sample selection, transcription, retrieval, and frequency calculation described above, we now present provisional results from the BNClab-M sample, beginning with the overall frequency of past perfects and past non-perfects. Table 4 extrapolates figures for the past non-perfect by multiplying the total number of non-perfect hits or candidates (filtered for working/middle-class speakers from England) by the precision rates calculated by manual analysis of the one in 50 subsets (see section 4.2.3).

1994			2014			Change
Past perfect	Past non-perfect (extrapolated)	Past perfect %	Past perfect	Past non-perfect (extrapolated)	Past perfect %	Significance
634	19,053	3.2%	2,224	57,977	3.7%	**

Table 4: Overall frequencies of past perfect and past non-perfect (* p<.05; ** p<.01; *** p<.001)

We see that overall usage of the past perfect in the BNClab-M data significantly increases from 3.2 percent to 3.7 percent of past-marked tense forms. At the same time, we observe social differentiation and patterns of change according to age, gender, and social class, although they are not necessarily significant, as illustrated in Table 5.

			1994		2014		Change
Gender	**Class**	**Age**	**Past Perfect**	**Non-perfect[a]**	**Past Perfect**	**Non-perfect[a]**	**Significance**
Female	Working class	U-45	113 3.5%	3076	17 1.9%	895	**
		O-45	89 3.2%	2699	61 2.7%	2194	n. s.
	Middle class	U-45	163 3.4%	4695	629 3.4%	17881	n. s.
		O-45	26 3.1%	818	989 5.2%	18131	**
Male	Working class	U-45	74 2.4%	3045	20 2.9%	674	n. s.
		O-45	15 3.2%	451	22 1.5%	1410	*
	Middle class	U-45	86 3.3%	2519	233 2.5%	9069	*
		O-45	68 3.7%	1749	253 3.2%	7723	n. s.

Table 5: Results across social groups in the BNClab-M sample ([a] extrapolated figures). * p<.05; ** p<.01; *** p<.001, n.s.=not significant

At this granular level, a more complex picture emerges. The overall increase of past perfects seems mainly attributable to female middle-class speakers over 45 years of age. Recall, however, that this is the group with the highest discrepancy in speaker numbers between 1994 and 2014 (see Table 1). Meanwhile, declining proportions of past perfect are found in five of the eight social groups, including both age cohorts of female working-class speakers and male middle-class speakers. But only a few of these changes are statistically significant.

Our conclusions about sociolinguistic variation are similarly tentative. At the aggregate level, there is a general tendency for the past perfect to be used more by middle-class than working-class speakers and, in 2014, by females than males. At the granular level we find, for instance, that rates of past perfect in female working-class and female middle-class are higher among younger speakers in the 1994 corpus, but in the 2014 corpus they are higher among older speakers. These figures need further investigation and corroboration.

Regarding past unreal contexts, we also see a proportional increase of the past perfect, with non-standard variants becoming less popular, as shown in Table 6.

	1994		**2014**		**Change**
	Cases	**% of total**	**Cases**	**% of total**	**Significance**
Past perfect	75	55.1%	195	76.2%	*
Non-standard perfect	61	44.9%	57	22.3%	***
Double	9	6.6%	14	5.5%	n.s.
Would	5	3.7%	5	2.0%	n.s.
Elided (*'d*)	47	34.6%	42	16.4%	***
Total	**136**	**100.%**	**256**	**100%**	

Table 6: Unreal past conditionals in BNClab-M subcorpus (* p<.05; *** p<.001, n.s.=not significant)

Further work is needed to understand why our overall results do not support those of Bowie *et al.* (2013) and Smith and Waters (2019), who both found a significant decline of the past perfect in recent spoken British English. Recall that the DCPSE corpus used in Bowie *et al.*'s (2013) focuses on the late twentieth century (1960s-1990s) and is not stratified by sociodemographic variables. Moreover, their results do not specify distributions across spoken registers, making direct comparisons with our results problematic. The *Desert Island Discs* BBC radio study by Smith and Waters (2019) is closer in timeframe to the present study, and includes similar social variables (e.g., age, gender, occupation, and education) for speakers from England. Smith and Waters (2019) found that age and education, rather than gender and occupation (as in the present study), correlated most significantly with the frequency of the past perfect. However, they operationalized occupation not as a socioeconomic index but as an estimate of the speaker's occupational need to use standard English (cf. Sankoff and Laberge 1978).

Another factor worth recalling is the possible effect of speakers in BNC2014S being more aware of being recorded (cf. Axelsson 2018). This may have led some speakers to be more careful to a) use the past perfect in contexts where the non-perfect would suffice

(e.g., with temporal clauses), and b) avoid non-standard perfects in unreal past conditionals. Further investigation of this possibility seems appropriate.

6. CONCLUSION

Our paper has methodological implications for the diverse users of spoken corpora, in particular users of the two conversational BNCs, primarily researchers, but increasingly also teachers and students in linguistics, language teaching, and beyond. Like the creators of these corpora, we wish the affordances to be embraced widely. Our concern has been to increase awareness of issues that arise when comparing these corpora across time, especially regarding changes in grammar.

Addressing the need for closely comparable data, we took advantage of Brezina *et al.*'s (2018a) BNClab subcorpus. This judgment sample affords a closer balance of social group representation in the two periods than that of the parent BNCs, although its geographical scope (all nations of the UK) affects representativeness and comparability at the granular level. We partially improved comparability by limiting the subcorpus to working-class and middle-class speakers from England but narrowed the demographic representativeness in the process. Further refinement of the sample is no doubt possible.

The case study on the past perfect navigated challenges in investigating grammar in spontaneous conversation, namely deciding on the field of competition and units of measurement, and devising queries to retrieve constructions with satisfactory recall and precision. A heuristic approach allowed us to capture variability among discontinuous past perfects and non-standard perfects, facilitating recall. The mixed quality of transcriptions in BNC1994DS can also affect comparability with BNC2014S. Our decision to discard hits from BNC1994DS containing inaccurate transcription should not disrupt comparability with BNC2014S, since we made equivalent corrections for rival constructions and used proportional frequencies in both periods.

The implications of our results to date are less clear-cut. We find consistently higher frequencies of the past perfect by female and middle classes speakers. The finding that the past perfect is spreading in conversational use appears to contradict previous reports of a decline in spoken British English. While the discrepancy may be related to unbalanced recruitment of some categories of speaker in the two periods, it seems premature to suggest that the past perfect is on the wane, at least in English conversation.

Our investigation has highlighted some contemporary spoken usage characteristics associated with the past perfect that receive scant attention in English language teaching materials and curricula. One example, occurring in unreal past conditionals, is what we have labeled the 'non-standard' perfect', although it seems widely accepted in colloquial contexts (Huddleston and Pullum 2002; Ishihara 2003). These forms are almost entirely overlooked in ELT resources (Ishihara 2003) and, despite a proportional decline, their use in informal conversation seems frequent enough to draw the attention of advanced learners, at least as receptive knowledge (cf. Timmis 2005). Similarly, the simple past, stative use of *had got* in British English is almost absent from ELT coursebooks, and yet it is more common than its past perfect homograph.

In the future, we aim to follow up this study through detailed, selective exploration of the parent BNCs where social categories —particularly age— have more differentiated breakdowns.

References

Anderwald, Lieselotte. 2002. *Negation in Non-standard British English: Gaps, Regularizations and Asymmetries.* New York: Routledge.

Atkinson, Will. 2015. *Class*. Cambridge: Polity Press.

Axelsson, Karin. 2018. Canonical tag questions in contemporary British English. In Vaclav Brezina, Robbie Love and Karin Aijmer eds, 96–119.

Baker, Paul. 2023. A year to remember? Introducing the BE21 corpus and exploring recent part of speech tag change in British English. *International Journal of Corpus Linguistics* 28/3: 407–429.

Ball, Catherine. 1994. Automated text analysis: Cautionary tales. *Literary and Linguistic Computing* 9: 295–302.

Beal, Joan. 2010. *An Introduction to Regional Englishes: Dialect Variation in England*. Edinburgh: Edinburgh University Press.

Biber, Douglas. 1933. Representativeness in corpus design. *Literary and Linguistic Computing* 8/4: 243–257.

Bowie, Jill, Sean Wallis and Sebastian Aarts. 2013. The perfect in spoken British English. In Sebastian Aarts, Joanne Close, Geoffrey Leech and Sean Wallis eds. *The Verb Phrase in English: Investigating Recent Language Change with Corpora.* Cambridge: Cambridge University Press, 318–352.

Brezina, Vaclav and Miriam Meyerhoff. 2014. Significant or random? A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics* 19/1: 1–28.

Brezina, Vaclav and William Platt. 2024. *#LancsBox X*. http://lancsbox.lancs.ac.uk/ (accessed 5 May 2023.)

Brezina, Vaclav, Dana Gablasova and Susan Reichelt. 2018a. *BNClab*. http://corpora.lancs.ac.uk/bnclab (accessed 5 May 2023.)

Brezina, Vaclav, Robbie Love and Karin Aijmer eds. 2018b. *Corpus Approaches to Contemporary British Speech: Sociolinguistic Studies of the Spoken BNC2014*. New York: Routledge.

Brezina, Vaclav, Abi Hawtin and Tony McEnery. 2021. The written *British National Corpus* 2014 – design and comparability. *Text and Talk* 41/5–6: 595–615.

Burnard, Lou. 2007. *Reference Guide for the British National Corpus (XML edition)*. http://www.natcorp.ox.ac.uk/docs/URG/ (accessed 5 May 2023.)

Coleman, John, Mark Liberman, Greg Kochanski, Lou Burnard and Jiahong Yuan. 2011. Mining a year of speech. In *Proceedings from the Workshop of New Tools and Methods for Very-Large-Scale Phonetics Research*, 16–19. http://www.phon.ox.ac.uk/jcoleman/MiningVLSP.pdf (accessed 5 May 2023.)

Crowdy, Steve. 1993. Spoken corpus design. *Literary and Linguistic Computing* 8/4: 259–265.

Curry, Niall, Robbie Love and Olivia Goodman. 2022. Adverbs on the move: Investigating publisher application of corpus research on recent language change to ELT coursebook development. *Corpora* 17/1: 1–38.

Declerck, Renaat. 2006. *The Grammar of the English Verb Phrase. Volume 1: The Grammar of the English Tense System: A Comprehensive Analysis*. Berlin: Mouton de Gruyter.

Denison, David. 1993. *English Historical Syntax: Verbal Constructions*. London: Longman.

Depraetere, Ilse and Chad Langford. 2019. *Advanced English Grammar: A Linguistic Approach*. London: Bloomsbury.

Egbert, Jesse, Douglas Biber and Bethany Gray. 2022. *Designing and Evaluating Language Corpora: A Practical Framework for Corpus Representativeness*. Cambridge: Cambridge University Press.

Gablasova, Dana, Vaclav Brezina and Tony McEnery. 2017. Exploring learner language through corpora: Comparing and interpreting corpus frequency information. *Language Learning* 67/1: 130–154.

Garside, Roger and Nicholas Smith. 1997. A hybrid grammatical tagger: CLAWS4. In Roger Garside, Geoffrey Leech and Anthony McEnery eds., 102–121.

Garside, Roger, Geoffrey Leech and Anthony McEnery eds. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.

Hardie, Andrew. 2012. CQPweb – Combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17/3: 380–409.

Hofland, Knut, Anne Lindebjerg and Jørg Thunestvedt. 1999. *ICAME Collection of English Language Corpora*. Bergen: The HIT Centre.

Hoffmann, Sebastian, Stefan Evert, Nicholas Smith, David Lee and Ylva Berglund-Prytz. 2008. *Corpus linguistics with BNCweb – A Practical Guide*. Frankfurt: Peter Lang.

Hoffmann, Sebastian and Sabine Arndt-Lappe. 2021. Better data for more researchers: Using the audio features of BNCweb. *ICAME Journal* 45: 125–154.

Horvath, Barbara. 2013. Ways of observing: Studying the interplay of social and linguistic variation. In Christine Mallinson, Becky Childs and Gerard Van Herk eds. *Data Collection in Sociolinguistics: Methods and Applications.* New York: Routledge. https://doi.org/10.4324/9780203136065

Huddleston, Rodney and Geoffrey Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Ishihara, Noriko. 2003. “I wish I would have known!”: The usage of *would have* in past counterfactual *if*- and *wish*-clauses. *Issues in Applied Linguistics* 14/1: 21–48.

Jucker, Andreas, Gerold Schneider, Irma Taavitsainen and Barb Breustedt. 2008. Fishing for compliments: Precision and recall in corpus-linguistic compliment research. In Andreas Jucker and Irma Taavitsainen eds. *Speech Acts in the History of English*. Amsterdam: John Benjamins, 273–294.

Lavandera, Beatriz. 1978. Where does the sociolinguistic variable stop? *Language in Society* 7/2: 171–82.

Leech, Geoffrey and Nicholas Smith. 2000. *Manual to Accompany the British National Corpus (Version 2) with Improved Word-class Tagging*. https://ucrel.lancs.ac.uk/bnc2/bnc2postag_manual.htm (accessed 5 May 2023.)

Leech, Geoffrey and Nicholas Smith. 2005. Extending the possibilities of corpus-based research on English in the twentieth century: A prequel to LOB and FLOB. *ICAME Journal* 29: 83–98.

Leech, Geoffrey, Marianne Hundt, Christian Mair and Nicholas Smith. 2009. *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press.

Love, Robbie. 2020. *Overcoming Challenges in Corpus Construction: The Spoken British National Corpus 2014*. New York: Routledge.

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

McEnery, Tony, Robbie Love and Vaclav Brezina. 2017. Introduction: Compiling and analysing the Spoken British National Corpus 2014. *International Journal of Corpus Linguistics* 22/3: 311–318.

Milroy, Lesley and Matthew Gordon. 2003. *Sociolinguistics: Methods and Interpretation*. Oxford: Blackwell.

Mindt, Dieter. 2000. *An Empirical Grammar of the English Verb System*. Berlin: Cornelsen.

Reichelt, Susan. 2021. Recent developments of the pragmatic markers *kind of* and *sort of* in spoken British English. *English Language & Linguistics* 25/3: 563–580.

Rühlemann, Christoph. 2007. *Conversation in Context: A Corpus-driven Approach*. London: Continuum.

Sankoff, David. 2005. Problems of representativeness. In Ulrich Ammon, Norbert Dittmar, Klaus Mattheier and Peter Trudgill eds. *Sociolinguistics: An International Handbook of the Science of Language and Society*. Berlin: Walter de Gruyter, 998–1002.

Sankoff, David and Susan Laberge. 1978. The linguistic market and the statistical explanation of variability. In David Sankoff ed. *Linguistic Variation: Models and Methods*. New York: Academic Press, 239–250.

Schilling-Estes, Natalie. 2007. Sociolinguistic fieldwork. In Robert Bayley and Ceil Lucas eds. *Sociolinguistic Variation: Theories, Methods, and Applications*. Cambridge: Cambridge University Press, 165–190.

Smith, Nicholas. 1997. Improving a tagger. In Roger Garside, Geoffrey Leech and Anthony McEnery eds., 137–150.

Smith, Nicholas and Cathleen Waters. 2019. Variation and change in a specialized register: A comparison of random and sociolinguistic sampling outcomes in Desert Island Discs. *International Journal of Corpus Linguistics* 24/2: 169–201.

Sönning, Lukas and Manfred Krug. 2022. Comparing study designs and down-sampling strategies in corpus analysis: The importance of speaker metadata in the BNCs of 1994 and 2014. In Ole Schützler and Julia Schlüter eds. *Data and Methods in*

Corpus Linguistics: Comparative Approaches. Cambridge: Cambridge University Press, 127–160.
Tagliamonte, Sali. 2006. *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press.
Timmis, Ivor. 2005. Towards a framework for teaching spoken grammar. *ELT Journal* 59/2: 117–125.
Trask, R.L. 1993. *A Dictionary of Grammatical Terms in Linguistics*. New York: Routledge.
Yao, Xinyue and Peter Collins. 2013. Recent change in non-present perfect constructions in British and American English. *Corpora* 8/1: 115–135.

Corresponding author
Nicholas Smith
University of Leicester
School of Education
21 University Road
LE1 7HR
Leicester
United Kingdom
Email: ns359@leicester.ac.uk

received: July 2023
accepted: April 2024

APPENDIX: COMPOSITION OF THE ORIGINAL BNCLAB SUBCORPUS, REPRESENTING ALL UK REGIONS AND CLASSES (BASED ON BREZINA *ET AL.* 2018A)

		1994	**2014**
Gender	Female	126	126
	Male	124	124
Age	0–14	16	12
	15–24	47	56
	25–34	50	60
	35–44	47	31
	45–59	46	50
	60–74	31	31
	75–95	13	10
Social class	Middle class	62	113
	Working class	63	36
	Retired	35	27
	Student	41	49
	Unknown	49	25
Region	England: London	31	28
	England: Midlands	41	28
	England: North	47	62
	England: Southwest	30	25
	England: Southeast	34	45
	Scotland	13	8
	Ireland	22	1
	Wales	18	14
	Other	14	39
Total		**250**	**250**