From broadcast archive to language corpus: Designing and investigating a sociohistorical corpus from *Desert Island Discs*

Nicholas Smith and Cathleen Waters
University of Leicester

Abstract
The aims of this paper are twofold: i) to present the motivation and design of a sociohistorical corpus derived from the popular BBC Radio show, Desert Island Discs *(*DID*); and ii) to illustrate the potential of the* DID *corpus (DIDC) with a case study. In an era of ever-increasing digital resources and scholarly interest in recent language change, there remains an enormous disparity between available written and spoken corpora. We describe how a corpus derived from* DID *contributes to redressing the balance. Treating* DID *as an example of a specialized register, namely, a 'biographical chat show', we review its attendant situational characteristics, and explain the affordances and design features of a sociolinguistic corpus sampling of the show. Finally, to illustrate the potential of DIDC for linguistic exploration of recent change, we conduct a case study on two pronouns with generic, impersonal reference, namely* you *and* one*.*

1 Introduction

Research on recent and current language change in standard English is one of the most dynamic and popular areas of inquiry in contemporary linguistics. As illustrated in, for example, Hundt and Mair (1999), Leech *et al.* (2009), Millar (2009) and Aarts *et al.* (2013), the field has expanded rapidly in the last two decades, particularly in respect of uncovering evolving areas of grammar and lexicogrammar in standard American, British and other Englishes. A key element of such progress has undoubtedly been accelerated development of corpus resources and methodology.

Amid the massive expansion of recent-change corpora – from the Brown family (Leech *et al.* 2009) to mega-corpora such as the British National Corpus (BNC),[1] Corpus of Historical American English (COHA)[2] and Corpus of Contemporary American English (COCA)[3] – one much less impressive fact is inescapable, namely the skew towards written language. Researchers wishing to investigate changes in spoken registers, either as a central focus or as part of a broader view of change alongside written registers, continue to have very few options readily available. While there are a few, well-designed spoken corpora that partially address the major gaps, they tend either to lack balanced sociolinguistic sampling or to be focused exclusively on conversation.

In this paper we argue that some of the empty spaces in recent diachrony can be addressed by extracting corpora from broadcast forms of English. Radio shows in particular offer a very diverse menu of specialist programming, spanning numerous registers/genres, some of which have been running for a considerable time. Such programmes are, moreover, increasingly accessible to the public in electronic form in online speech archives. Although collecting samples of broadcast speech for a corpus is generally done 'after the fact', in the sense that we lack prior access to the people who speak, we can often benefit from the public renown of participants on radio/TV shows to glean useful demographic information about them.

The present paper illustrates some ways of capitalizing on broadcast archives for diachronic corpus research, and more specifically of considering the perspectives of register and social variation in sampling a particular programme. We have selected the BBC Radio 4 *Desert Island Discs* (*DID*) for these purposes. This show is one of the longest-running in broadcasting history, and remains popular to this day. In it, guests imagine they have been marooned on a desert island, and discuss their lives in relation to eight significant records they bring with them. While it has had to keep pace with the times, the structural changes *DID* has undergone have been remarkably few. These characteristics make it an attractive prospect for investigating variation and change in recent decades. In addition, the possibilities to use demographic data to sample the people who have appeared on the show facilitate a sociolinguistic perspective on the evolution of a specialized register. Thus we could call our *DID* corpus (DIDC) a sociohistorical corpus, that is, a methodically collected set of real-time data that takes into account the social/demographic characteristics of the speakers.

The case study we use to illustrate the potential of our corpus centres on two pronouns (*you* and *one*) that previous research on broadcast discourse has found to be used in diverse ways regarding deictic and generic reference. We were interested to see what sociolinguistic patterns of variation and change occur in the corpus, e.g. which speaker groups use *you* and *one* more? In a world where broadcast discourse is said to have become increasingly informal (Scannell 1989, Hendy 2007), to what extent has generic *one* given way to *you*?

2 Existing corpus provision and affordances of a DID *corpus*

2.1 Sociohistorical and/or diachronic spoken corpora

Among sociohistorical corpora, only a few examples include speech from across the last century. The Origins of New Zealand English archive (ONZE) contains a mixture of material: a) from the earliest periods, oral history interviews with members of the public collected for radio broadcast, and b) in the more recent data, sociolinguistic interviews. The use of the radio archives resulted in demographic information about those speakers being difficult to obtain, and the balanced social sampling of the later corpus was not possible in the earlier data (Gordon *et al.* 2007). Meanwhile the Diachronic Electronic Corpus of Tyneside English (DECTE; Corrigan *et al.* 2012) contains dialect speech from the northeast of England, from the 1960s to the present. Speaker characteristics (age and gender) are accessible, although because the corpus is an amalgamation of different projects, the sampling procedure varies over time, and thus diachronic comparisons need to be handled with care.

The BNC was initially conceived as a synchronic corpus. However, the recent creation of BNC2014 (Love *et al.* 2017), including (at the time of writing) 11 million words of demographically-sampled conversational speech, provides a diachronic counterpart to the Spoken Demographic BNC1994, with 5 million words. The newly-expanded BNC is extremely useful for sociolinguistically-informed diachronic analysis of spoken language, although it is restricted to casual conversation.

The use of early broadcast archives for studying language change is viewed rather pessimistically by Bauer (1994: 123). He argues that early recordings represent a stilted kind of language that, where it is not scripted, is heavily based on a formal, written-like style. However, rather than a hindrance, it is possible to embrace style shifts as phenomena to be investigated, and for which corpus methods are well suited. A few studies in English and other languages have already demonstrated how

broadcast talk over time can be investigated. Van de Velde *et al.* (1997), for example, extracted an age- and dialect-balanced corpus of speakers from sports commentaries and royal ceremonies in a Dutch radio archive. Their findings include shifts in vowel qualities in northern Dutch which they relate to 'deformalization' of Dutch broadcast media. In Australia, Price (2012) conducted a panel study of Australian newsreaders over time, by phonetically analysing original broadcasts from the 1950s and 1980s and inviting the same individuals to rerecord their newscasts during the early 2000s. While this study has the great advantage that external variables are largely controlled, the range of participants, as with Van de Velde *et al.* (1997), is demographically very limited. Other diachronic linguistic studies using broadcast speech recordings are summarized in Van de Velde *et al.* (1997: 386); see also Jucker and Landert (2015), cited below.

Two important multi-register corpora are the Diachronic Corpus of Present-Day Spoken English (DCPSE) and COCA. DCPSE includes several spoken registers of British English speech from the late 1950s to the early 1990s.[4] COCA includes more than 80 million words of transcripts from diverse TV and radio shows, from 1990 to the present. While the transcripts are said to be generally accurate (https://corpus.byu.edu/coca/), it is unclear how consistent they are. However, neither of these corpora samples by speaker characteristics (e.g. age, education).

In summary, despite some excellent resources, it is clear that a wide range of gaps or shortages exist in corpus coverage of recent spoken English. These include a paucity of real-time spoken resources in comparison to written ones, and, outside of conversation, a shortage of register-specific corpora that use sociolinguistically-balanced design, or supply metadata on speaker characteristics.

2.2 Affordances and representativeness of a sociohistorical DID *corpus*

DID has the the potential to address some of the shortcomings just mentioned. For example, as noted by Jucker and Landert (2015), the longevity of the show, the long service of three of its hosts, and online accessibility of its archive,[5] all present rich opportunities for exploring changes in a particular discourse type (radio talk show/chat show) diachronically. Jucker and Landert (2015: 37) identify a number of changes in interaction style between host and guest in *DID*, which they attribute to increasing "language immediacy in public contexts" over time.

We acknowledge these opportunities, while noting also that the material Jucker and Landert select from *DID* does not constitute a corpus, in the sense of sampling according to an explicit notion of representativeness. Among possible interpretations of representativeness, one option is to focus on capturing variation in the register (e.g. by sampling programme episodes randomly within different periods); another is to represent variation across the guest speakers. Our main interest in this paper is in the latter notion of representativeness. The fact that the guests tend to be well-known public figures makes it possible to find data on their social group characteristics (e.g. their formal education, their age at the time of recording, etc.) and use this to design as socially balanced and consistent a sample as the history of the show permits. This conceptualization of a corpus of *DID* should allow users to explore language use in the show with less risk that any variation and change they find is an artefact of changing proportions of speakers from certain demographic groups, such as females or younger speakers.[6]

From a variationist sociolinguistic perspective, the DIDC is one of the few corpora to permit analysis of sociolinguistic variation in more specialized registers than sociolinguistic interviews, with their traditional focus on vernacular usage (cf.

Biber and Conrad 2009: 264). While we acknowledge the critical role that studying the vernacular plays in understanding language variation and change, an examination of other registers allows us insight into other contexts of language use, such as the diachronic material of *DID* that we now briefly characterize.

3 A register profile of Desert Island Discs

What register is *DID*? Castell (1999: 392) calls it a chat show, and says that "the interview focuses upon a gentle, entertaining revelation of [the guest's] humanity". While we agree with Castell's comments about entertainment and revelation, we would describe *DID* more specifically as an example of the *biographical chat show* register, since a retrospective focus on the guest's life is routinely prominent. Through their music choices and prompts from the host, guests engage in reflective discussion about themselves, from childhood onwards.

Using Biber and Conrad's (2009) taxonomy, we list other situational characteristics of *DID* in Table 1, and clarify some of the more salient points below.

Table 1: Summary of situational characteristics of *DID*

Situational characteristic	Realization in *DID*
Communicative purposes	Entertainment; personal disclosure; intellectual edification
Participants	A host and generally one guest
	Large, absent but targeted audience
Relations among participants	Usually unfamiliar; direct interaction
Channel	Spoken, radio broadcast
Setting	Host and guest in shared space and time
	Audience removed in space and time
Production circumstances	Planned question topics, but unscripted
	Editing to fit programme length
Topics	Guest's life story and career, formative influences, attitudes/emotional responses to life events, significance of music choices
Other	Talk is interspersed by musical pieces

Almost invariably the only immediate participants are the host and one guest. Guests are chosen who have reached a position of some standing in their field, and "lived a rich and interesting life" (Magee 2012: 11). Bell and van Leeuwen (1994) suggest further that all chat show guests must fulfil at least one of three attributes: *news value, entertainment value* and *symbolic value*, i.e. to be closely connected not just to their specialist field, but also to discourse about that field. We would argue that the nature of the BBC Radio 4 audience – three-million strong,[7] well-educated, professional – imposes another selection criterion, namely articulacy.

The role of the host is to facilitate talk and ask questions on the audience's behalf (Bell and van Leeuwen 1994, Magee 2012). Although turnover of hosts has been low, there has been a clear development in interviewing approach and personality, from the genteel, formulaic and factually-focused Roy Plomley to the more probing, dialogic approaches of subsequent hosts (see Jucker and Landert 2015).

4 Methodology: Corpus design

4.1 Constructing a sociolinguistic **DID** *corpus: desiderata and constraints*

Labov (2001: 39) argues that random sampling in sociolinguistics allows the researcher to "capture the regular structure of variation within a large community".

However, obtaining a truly random sample can be problematic for sociolinguistic studies (Tagliamonte 2006: 22-23). Moreover, the somewhat exclusive nature of guest-selection on *DID* precludes using the show to represent the *entirety* of British society. Instead, we sought to reflect the demographic diversity that *DID* does afford. Our sample therefore follows what Tagliamonte (2006: 30-32) describes as a "stratification schema": we identified the viable social characteristics of interest and our sampling strategy aimed for consistency in the number of speakers in each cell (see Table 5).

To determine which demographic characteristics to include in DIDC, we reviewed the social characteristics included in previous corpora used for quantitative sociolinguistics. Synchronic and diachronic sociolinguistic corpora compiled in recent years have continued the use of well-established speaker categories such as sex, age, ethnicity, social class and education: see e.g. DECTE and ONZE (Section 2.1), the Corpus of Early English Correspondence (CEEC; Nevalainen and Raumolin-Brunberg 2003) and the York Corpus (Tagliamonte 2002). Hoffman and Walker (2010: 37-38) provide a succinct overview of previous sociolinguistic findings about (and some criticisms of the traditional sociolinguistic treatment of) age, sex, social class and ethnicity. We acknowledge that a more performative, non-binary conceptualisation of the categories we describe below (e.g. gender rather than sex, cf. Butler 1990) can provide a more nuanced understanding of some linguistic choices, particularly the behaviour of individuals. For historical data, however, an ethnographic approach is not possible. Moreover, for a macro-level study such as this, where the goal is to identify trends over time and to disentangle demographic and register change, we are not focussed on individuals. Accordingly, our sociolinguistic sampling is based on similar methods to those of contemporary and historical sociolinguistic corpora.

We initially explored a wide range of social characteristics including age, (binary) sex, education, occupation, social class, ethnicity and region of origin (e.g. the Midlands). Similarly to Nevalainen and Raumolin-Brunberg's (2003: 45-49) strategy for compiling social information for CEEC, we extensively investigated the backgrounds of all the guests on *DID* in our selected periods (see 4.2.1), using resources such as the *Oxford Dictionary of National Biography,* magazine profiles and newspaper obituaries. It soon became apparent that we would be unable to sample by ethnicity, due to the limited ethnic diversity of guests in the selected periods. (Ethnic diversity improves after 2005, and so ethnicity may be an interesting avenue for future extension of the corpus.) In terms of regional provenance, most guests had travelled or lived outside their region of origin for extended periods of time. We therefore have not attempted to sample or study regional origin; instead, we have limited the sample to those born in and resident in England, including those whose parents were born elsewhere (e.g. writer and actor Meera Syal). Given the highly complex nature of operationalizing social class (see e.g. Milroy and Gordon 2003), we instead concentrated on two specific categories widely associated with social class, namely education and occupation, which we discuss below. We have also considered age and sex in our sampling plan, which we turn to now.

4.2 Sampling variables

4.2.1 Period

In selecting periods to sample, we sought to take advantage of continuity not only of the show but also of the two longest-serving hosts, to aid comparability; cf. Table 2.

Table 2: Periodization of *Desert Island Discs* based on host interviewers

Period label	**1960s**	**1980s-A**	**1980s-C**	**2000s**
Years	1960-69	1980-85	1988-90	2000-2006
Host	Roy Plomley	Roy Plomley	Sue Lawley	Sue Lawley
Gender	Male	Male	Female	Female
Nationality	English	English	English	English
Age at time	46-55	66-71	41-44	56-60

Note: '1980s-A' and '1980s-C' distinguish early and late-1980s respectively. Another host, Michael Parkinson, presented the show in the mid-1980s.

However, the fragmentary character of surviving recordings from the 1960s led us to exclude this period in the present study, as it could impair comparability of results.

4.2.3 Gender

As we noted earlier, we used a binary categorisation for speaker gender (i.e. sex). The proportion of female guests (with the regional and ethnic characteristics mentioned above) increased over time, but continued to be less than a third of the guest total (see Table 3).

Table 3: Proportion of female guests on *DID,* by time period

Period	No. of women, of total	%female
1980s-A	26 out of 118	22.0%
1980s-C	28 out of 87	32.2%
2000s	38 out of 130	29.2%

The percentage of women included in our sociolinguistic sample is consistently 40 per cent across all the time periods sampled (see Table 5). We chose to include a proportion of women in the sample that was higher than the average proportion of women on the programme to allow us to better explore speaker gender. Although synchronic sociolinguistic corpora tend to aim for the same number of male and female speakers, our approach is consistent with that used in for the historical data in *CEEC* in which men outnumbered women (Nevalainen and Raumolin-Brunberg 2003:45). The use of a consistent number of male and female interviews across each time period ensured that we had comparable datasets across the time periods.

4.2.4 Age

As perhaps might be expected given the nature of *DID* (interviews with individuals who have "lived a rich and interesting life" (Magee 2012: xi), the average age of the guests in each of the time periods we examined was consistently over 50:

Table 4: Average age of all guests on *DID,* by time period

Period	Average age of guests
1980s-A	53
1980s-C	57
2000s	59

Moreover, very few guests were under the age of 30. Although we would have liked to examine a range of age groups (as is common in synchronic sociolinguistic

studies), the preponderance of older guests made more granular age distinctions impossible. Thus, a balanced sample of over-50s and under-50s seemed to us a reasonable compromise.

4.2.5 Education

As formal education is a means of transmitting prestige forms (Labov 2001:512), it is a widely-used sampling variable in sociolinguistic studies. Education has been operationalized using a variety of criteria such as attendance in secondary or further education (Ito and Tagliamonte 2003), or study beyond the legally compulsory stage of education (Waters 2013). However, neither of those distinctions effectively captured the range of educational experiences we observed among guests on *DID*. Almost all the guests attended school to at least age 13. Distinctions based on secondary school attendance were not sufficient either. Although we were generally able to determine whether or not a guest had undertaken any education beyond secondary school, we noticed that some of the guests whose education ended with secondary school had attended secondary institutions that select based on academic ability, notably grammar schools. Moreover, a recent study by Ndaji et al. (2016) reports that independent schooling in the UK gives an academic advantage equivalent to two years of additional schooling by the age of 16. Therefore, type of educational institution may be as important as number of years of study. To reflect this, we created a bespoke strategy for categorizing guests' educational backgrounds. We grouped those who had attended independent schools and grammar schools (regardless of whether they had subsequently undertaken higher education) together with those who had attended university; we call this group Educational Group 2 (henceforth, [edu2]). Guests who had attended non-grammar state schools and who had also not undertaken higher education were considered together in a group that we call Educational Group 1 (henceforth, [edu1]).

4.2.6 Occupation

While the importance of occupation is widely recognized, there is no consensus as to the most appropriate way to classify different occupations (Milroy and Gordon 2003). Occupational classification in *DID* was complicated by several issues. First, the range of occupations represented on *DID* is in some respects more diverse than reflected in occupational classification schemes. The NS-SEC (National Statistics Socio-Economic Classification), for example, has a single category called 'Actors, entertainers and presenters': many *DID* guests would fit into this category, but our intuitions suggested that it would obscure linguistic variation among them. On the other hand, using a very wide range of occupational categories would seem to spread our results too thinly and important commonalities of speech pattern among speakers might go undetected. Moreover, sometimes the occupation listed next to each guest in the *DID* Archive only partly reflects what the guest did for a living. Alan Titchmarsh, for example, is listed as a horticulturalist, but by the time of his interview his main occupation would be better described as a TV/radio presenter or broadcaster. He had also written several books. Our solution to these issues was to review the job(s) the guest was performing at the time of interview and rate them collectively according to an index of the 'linguistic market' (Sankoff and Laberge 1978, Sankoff *et al.* 1989). This is a measure of a speaker's relative need to use the standard language variety in their working life. We used two values: [occ1] for speakers with a relatively low occupational demand for standard English, and [occ2] for speakers with a relatively high demand. In the first group we include, for instance, Arthur English (a former

music hall entertainer, and soap opera actor at the time of interview) and Mollie Harris (a 'salt-of-the-earth' character in a country soap opera, *The Archers*). In the second group we include 'character' actors such as John Hurt, Jenny Agutter, and Kristin Scott Thomas. There are less clear-cut cases, however, such as the conceptual artists Cornelia Parker and Tracey Emin.

4.3 Composition of the Socio sample

Table 5 gives the structure of the Socio sample of *DID*.[8]

Table 5: Guest speakers in the Sociolinguistic sample

	1980s-A	**1980s-C**	**2000s**	**Overall**
Guests total	20	20	20	60
Under-50	10	10	10	30
Over-50	10	10	10	30
Female	8	8	8	24
Male	12	12	12	36
[edu1,occ1]	5	5	4	14
[edu1,occ2]	5	5	5	15
[edu2,occ1]	2	2	3	7
[edu2,occ2]	8	8	8	24
Words[9]	27,109	27,411	27,509	82,029

The number of guests on *DID* to choose from in the [edu1] and [occ1] categories was low, particularly in the early 2000s. This suggests a somewhat elitist bias in the show's guest selection.

4.4 Composition of the Random sample

As a comparator to our sociolinguistic strategy of sampling, we also created a parallel, random sample of guests. Although in this study the Random sample is not a main focus (cf. Smith and Waters, under review), we briefly outline its composition (see Table 6), and touch on some of the first findings from it in Section 5.3. Speakers of English provenance were selected using the random-number generator at www.random.org.

Table 6: Distribution of guest speakers in the Random sample (number of speakers overlapping with Socio sample in parentheses)

	1980s-A	**1980s-C**	**2000s**	**Overall**
Guests total	20 (2)	20 (6)	20 (4)	60 (12)
Under-50	8 (1)	12 (4)	3 (1)	23 (6)
Over-50	12 (1)	8 (2)	17 (3)	37 (6)
Female	5	9 (3)	6 (2)	20 (5)
Male	15 (2)	11 (3)	14 (2)	40 (7)
[edu1,occ1]	0	4 (1)	1 (1)	5 (2)
[edu1,occ2]	2 (1)	3 (2)	1 (1)	6 (4)
[edu2,occ1]	2	1 (1)	2 (1)	5 (2)
[edu2,occ2]	16 (1)	12 (2)	16 (1)	44 (4)
Words	27,085	27,402	27,466	81,953

With a few exceptions (e.g. gender in 1980s-C), the Random sample is skewed towards male, over-50s speakers in higher educational and occupational groups. These characteristics make it more representative of the guest profile of the show than the Socio sample, but problematic for social group analysis.

4.5 Transcription format and length

We developed a simple orthographic transcription scheme to identify speaker turns and other salient features of speech in DIDC. This is summarized in the Appendix. The starting point for transcription was a random point between one and ten minutes into the recording, where consistently the early part of the guest's life is prominently covered.

5 Case study on generic pronouns

5.1 Motivation and research questions

Previous research on broadcast talk includes a number of studies on the use of personal pronouns. Chang (2002) and O'Keeffe (2005), for instance, highlight ways in which speakers on radio and television chat shows and phone-ins exploit flexibility and sometimes ambiguity in the reference of the pronouns *we*, *you* and *they*, according to their intention to include or exclude particular groups and individuals in their referential scope. In our data, *you* was selected for exploration as it was frequent enough in generic function to permit a quantitative analysis.

Although we are not aware of any diachronic studies of pronouns in broadcast talk, historical development of generic *you* in British and American English registers is reported in Haas (forthcoming).[10] From the second half of the 17th century to the second half of the 20th century, Haas finds a dramatic increase in the proportion of cases of *you* that have impersonal reference, with particularly strong gains in more speech-based or speech-like registers, notably drama, prose fiction, and diaries/ personal journals. While ARCHER does not provide finer-grained periods with which to trace development over the latter half of the 20th century, we might expect the frequency of generic second person in the oral chat of *DID* to be similarly increasing. Analysis of *DID* by Jucker and Landert (2015) identifies several respects in which its discourse has evolved to become more conversation-like. They characterize more recent episodes, in the post-Plomley era, as more informal, less factually-focused exchanges, with turn length becoming more equal between host and guest, and dysfluencies increasing in number.

As is conventional in a sociolinguistic analysis, we consider an accountable variable context. That is, we include both generic *you* and other pronouns that refer to people in general. We reviewed instances of *you, one, we* and *they* in the data, but only uses of *you* and *one* appeared interchangeable in generic function. Therefore, our discussion focuses on alternation between *you* and *one.* As part of the increasing personalization described above, we might expect an expansion in the more 'inclusive', involved generic pronoun *you*, and a corresponding decline of the less involved, more formal pronoun *one*.

Thus our exploration of DIDC is guided by three research questions:

1. What is the distribution of a) second person overall, b) generic *you*, and c) generic *one*, and what is the impact of quotation?
2. What is the diachronic development of generic *you* and generic *one*?
3. What evidence is there of social group effects on the generic pronoun results?

5.2 Analytical method

Our analysis combines corpus and variationist methods. We examine frequencies per million words (pmw) as well as proportional use. We included both the subject and object forms of *you* and *one*, as well as the possessive forms *your* and *one's*. In addition, the data included the archaic singular forms *thee* and *thou*; although these appear in the overall frequency summary (in the rows labelled 'second person'), they do not occur in the non-quoted contexts that we subsequently focus upon. We excluded idioms (e.g. *thank you*, *mind you*, *I beg your pardon*, *if you like*), the nominal possessive *yours* (as it has no equivalent with *one*), and pronouns within repetitions and false starts, e.g. (1):

(1) *You* don't *you* don't get here by just being lazy (Tracey Emin, 2000s)

To indicate our interpretation of the reference of *you* and *one* as specific, generic or ambiguous, we use the following coding scheme: g = generic use, e.g. (2) (also the second *you* in (1) above); g/s = generic or specific use (ambiguous reference), e.g. (3); s = specific, deictic use (either singular or plural addressee), cf. (4):

(2) I had all the insecurities and anxieties that *one* does when *you*'re a teenager (Jane Asher, 1980s-C)
(3) as *you* know nowadays <pause> we have 19-year-olds who drive BMWs (Sir Bobby Robson, 2000s)
(4) people used to say to him <quote>Why don't *you* switch off sometime?</quote> (1980s-C, Ernie Wise)

Coding pronominal reference is sometimes challenging. Part of the problem is, as Biber *et al.* (1999: 331) state, that "[w]hen *we*, *you*, and *they* are used with reference to people in general … they tend to retain a tinge of their basic meaning". We used any available textual indicators (e.g. collocation with a vocative) and replayed the recording as necessary. The first author coded the data initially, then with the second author discussed and examined all uncertain cases until agreement was reached. The outcome in many cases left the ambiguous code 'g/s' intact: in other words, there seemed no way of determining the reference. We discuss such cases in the results below.

A further distinction we make is between quoted and non-quoted use. While we are mainly interested in non-quoted cases, as they represent the speaker's own usage, we noticed that *you* in quoted speech and thought is surprisingly frequent.

Finally, to investigate social group effects on the use of *you* versus *one*, we use a mixed-effect statistical model, with speaker as a random effect, and the social variables age, gender, education and occupation as fixed effects. We briefly compare results from the sociolinguistically balanced version of DIDC with the Random sample outlined in 4.4.

5.3 Results and discussion

Table 7 presents the frequencies of personal pronouns *you* and *one* in three periods of our *DID* Socio sample, considering overall use, and specific and generic reference.

Table 7: Personal pronouns *you* and *one* in the Socio sample: overall use, specific and generic reference

	1980s-A		**1980s-C**		**2000s**	
	n.	pmw[11]	n.	pmw	n.	pmw
Generic *one*	18	664	14	511	4	145
All second person	226	8,337	385	14,045	291	10,578
- outside quotation	146	5,721	281	11,063	223	8,407
- inside quotation	80	50,220	104	51,793	68	63,129
Outside quotation						
- Specific *you*	6	4.1%	11	3.9%	11	4.9%
- Generic *you*	140	95.9%	270	96.1%	212	95.1%
Inside quotation						
- Specific *you*	80	98.8%	103	99.0%	64	94.1%
- Generic *you*	1	2.3%	1	1.0%	4	5.9%

Regarding research question 1, it is clear that generic *one* is far less frequent than generic *you* in all periods. We also see the outcome of distinguishing quoted from non-quoted usage: outside of quotation, second person pronouns occur in the range of 5,000-11,000 times pmw, and consistently over 95 per cent of cases are either clearly generic, e.g. (5), or highly probably generic, cf. (6):

(5) jazz is something that unless *you* look for it *you* find it very difficult to hear
(John Surman, 1980s-A)
(6) and if *you* watch Gardeners World an= and that every Friday is filmed in my garden what *you* see there is is is my lump of <pause> my bleeding piece of earth if you like
(Alan Titchmarsh, 2000s, male, age2, edu1, occ2)

In sharp contrast, inside quotation the frequency of second person pronouns is over 50,000 pmw in each period, and nearly all cases are addressed to a specific individual or individuals, cf. (7):

(7) my grandchild said <quote>That's like *you* Nana</quote>
(2000s, Sheila Hancock, female, age2, edu2, occ2)

When *one* is used as a generic pronoun in our data, it invariably occurs outside of quoted contexts, cf. (8) and (9):

(8) it's lovely to feel the grandeur of something <pause> and to be moved in the way that Beethoven can move *one*
(1980s-A, Jenny Agutter, female, age1, edu2, occ2)
(9) and that is <pause> I suppose <pause> the ability of a great actor <pause> that *one* sees <pause> the spirit overcome erm the frailties of the body
(2000s, Adrian Noble, male, age1, edu2, occ2)

To some extent we can compare these frequencies with Biber *et al.*'s (1999) results for everyday conversation in the Longman Spoken and Written English corpus, circa early 1990s. The authors report the frequency of *you* as 30,000 pmw – apparently several times higher than second person pronouns in DIDC overall. Unfortunately,

Biber *et al.* do not distinguish quoted and non-quoted use, nor specific and generic cases. For generic *one* more direct comparison is possible: the frequencies in DIDC (664, 511 and 145 pmw in 1980s-A, 1980s-C and the 2000s, respectively) far surpass Biber *et al.*'s (1999: 354) figures for conversation (less than 25 pmw) and even, in the 1980s, for academic writing (approximately 400 pmw). The comparative data appear to suggest that until recently *DID* speakers have favoured a feature far more associated with detached expository style than with casual conversation. In our corpus cases of *one* in object function, cf. (8), seem particularly formal.

Regarding research question 2, focusing on generic uses of the two pronouns we discard the few quoted instances to enable a cleaner comparison. Proportional use of each pronoun across time is reported in Table 8.

Table 8: Proportional use of generic *one* vs. *you* across periods (excluding quotations)

	1980s-A		**1980s-C**		**2000s**	
	n.	%	n.	%	n.	%
one	18	11.4%	14	4.9%	4	1.9%
you	140	88.6%	270	95.1%	212	98.1%

It is clear that *you* has gained proportionally in each period as *one* has declined. The shift is statistically significant across the sample as a whole ($p<.001$, chi-square 15.00) and from 1980s-A to 1980s-C ($p<.05$, chi-square 6.31). Between 1980s-C and the 2000s the change is not significant (chi-square 3.35).

The results broadly support Haas' (forthcoming) findings on rising use of generic *you* in British speech-based registers. They also seem to concur with reports of informalization, in broadcast talk in general (cf. Scannell 1989) and in *DID* in particular (Jucker and Landert 2015).

Interestingly, the prevalence and spread of generic *you* appears to be a reflection of a (growing) convention in chat shows for guests to 'invite audience members to see the celebrity as like themselves' (Bell and Van Leeuwen 1994: 189). It is used, perhaps, to reduce the social distance between guest and audience. Example (10) would appear to illustrate this.

(10) the first thing *you* have to do when *you* er going on go on a stage any <trunc>t</trunc> anywhere <pause> because *you* get a bit excited and er hyped-up <pause> is *you* have to control *your* hands (Ken Dodd, male, age2, edu1,occ1)

You arguably carries inclusive overtones even when there is almost zero possibility that the host/audience could feasibly be a member of the group that is generalized over, cf. (11):

(11) Well I didn't know what a fit-up company was but <pause> *you* did six plays a week
(Ray Cooney, 1980s-A, male, age2, edu1, occ2)

(12) He'd just sort of look right through *you*
(Jacqueline Wilson, 2000s, female, age2, edu1, occ2)

Gast *et al.* (2015) call such cases 'simulated' reference, and Haas (forthcoming) suggests they are a factor in the increase of generic *you* in ARCHER. In our data, however, we found simulated-reference too indeterminate to quantify with confidence.

We turn now to research question 3, to consider the impact of social characteristics (age, gender, education and occupation) on the results. By using a mixed effects analysis to examine the relationship between generic pronoun choice and social characteristics of the speakers in the dataset as a whole, we found education alone to be selected as statistically significant ($p<.01$). A higher use of *one* is associated with more prestigiously educated speakers. We also tested within each period, but none of the social factor groups was significant; we believe this may be a result of low frequencies in each individual period. (In the 2000s there is a tendency for *one* to be used more by speakers over 50, but it is not statistically significant.) Generally speaking, however, the results again accord with expectations. Given the role of the education system in transmitting prestige forms (see 4.2.5), the association of *one* with more educated speakers serves to emphasize its connotations of formality. It also highlights the fact that even within a specialized register, where we can expect chat show norms (cf. Section 3) to influence speakers to talk in similar ways, language variation and change is mediated by social group factors.

The benefits of a balanced sociolinguistic sample design are reinforced when we compare the results for generic pronouns with those in the Random sample (cf. Section 4.4). Again, we focus on non-quoted generic use only, cf. Table 9.

Table 9: Proportional use of generic *one* vs. *you* across periods, in the Random sample (excluding quotations).

	1980s-A		**1980s-C**		**2000s**	
	n.	%	n.	%	n.	%
one	22	12.6%	15	6.4%	15	8.2%
you	152	87.4%	218	93.6%	169	91.8%

In marked contrast to the Sociolinguistic sample, in the Random sample the proportional change overall (increase of *you* by 5.1%, decrease of *one* by 35.5%) is *not* statistically significant (chi-square=1.95). The change from 1980s-A to 1980s-C is significant at $p<.05$ (chi-square=4.64), while that from 1980s-C to the 2000s is not significant (chi-square=0.45). In summary, from these data we do not get such a clear sense – unlike in the Socio sample – of the ground shifting from *one* to *you*. Moreover, because of its limited demographic diversity, the Random sample does not allow us to explore social factor groups, and detect that the lack of confirmed change is likely because of over-representation of highly educated guests. For a more detailed comparison of random versus sociolinguistic sampling in *DID*, see Smith and Waters (under review).

6 Conclusion

Our brief case study exemplifies the potential of DIDC for exploring recent linguistic change in a specialized register, and combining corpus-based and sociolinguistic methods. Although some imperfections remain in the Socio sample, it appears to have sufficient balance and consistency to examine generic pronouns and reveal significant patterns of change and sociolinguistic variation. Qualitatively, the data also reveal ways in which the referential ambiguity of generic *you* appears to be increasingly exploited in chat shows. Undoubtedly the study could be extended, for example by analysing additional implied meanings of generic *you* and *one*, including earlier and later periods of *DID*, and taking more data from each period. As always with spoken data, transcription time and costs need to be taken into account.

We should also bear in mind, as Jucker and Landert (2015) do, that *DID* is just one example of broadcast speech, in a particular institutional context. We see a corpus derived from *DID* archives as an important stepping-stone towards a more comprehensive coverage of recent language change across spoken registers.

Acknowledgements

In developing the DIDC, we thank the University of Leicester for periods of study leave, as well as the transcribers, especially Sarah Creer, Rebecca Hings, Naomi Obeng and Adam Percival. For discussion of the corpus design, we thank Doug Biber, David Denison, Gregory Garretson, Terttu Nevalainen, Emma Smith, Amy Wang, and audiences in Uppsala, Leicester and Hong Kong. Sebastian Hoffmann kindly filtered the corpus word counts for quotation/non-quotation.

Notes

[1] http://www.natcorp.ox.ac.uk/
[2] https://corpus.byu.edu/coha/
[3] https://corpus.byu.edu/coca/
[4] http://www.helsinki.fi/varieng/CoRD/corpora/DCPSE/index.html
[5] http://www.bbc.co.uk/programmes/b006qnmr
[6] We hope one day to be able to share our DID corpus with other researchers.
[7] Based on RAJAR figures in 2013, see http://www.bbc.co.uk/blogs/radio4/entries/421f245c-259b-3f2e-9234-e089d3290d3a.
[8] A full list of selected speakers and their social categorizations is available at http://hdl.handle.net/2381/41039.
[9] All word counts are according to (Rayson 2008) and exclude talk by the host.
[10] Haas's study does not include generic use of *one*. However, he does find slightly higher frequencies of generic *you* in American than British English by the late 20th century.
[11] Frequencies per million words are based on filtered word counts, within and outside quotation.

References

Aarts, Bas, Joanna Close, Geoffrey Leech and Sean Wallis (eds.). 2013. *The Verb Phrase in English: Investigating Recent Language Change with Corpora*. Cambridge: Cambridge University Press.

Bauer, Laurie. 1994. *Watching English Change: An Introduction to the Study of Linguistic Change in Standard Englishes in the Twentieth Century*. London: Longman.

Bell, Philip and Theo van Leeuwen. 1994. *The Media Interview: Confession, Contest, Conversation*. Kensington: University of New South Wales Press.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8: 243-257.

Biber, Douglas and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Butler, Judith. 1990. *Gender Trouble: Feminism and the Subversion of Identity*. London: Routledge.

Castell, Sarah. 1999. Phone-ins and chat shows. In Peter Childs and Mike Storry (eds.) *Encyclopaedia of Contemporary British Culture*, 392-292. London:

Routledge.
Chang, Phoebe. 2002. Who's behind the personal pronouns in talk radio? Cartalk: a case study. In Antonia Sánchez-Macarro (ed.) *Windows on the World: Media Discourse in English*, 115-152. Valencia: University of Valencia Press
Corrigan, Karen, Isabelle Buchstaller, Adam Mearns and Hermann Moisl. 2012. *The Diachronic Electronic Corpus of Tyneside English* (DECTE). Newcastle University. http://research.ncl.ac.uk/decte.
Gast, Volker, Lisa Deringer, Florian Haas and Olga Rudolf. 2015. Impersonal uses of the second person singular: A pragmatic analysis of generalization and empathy effects. *Journal of Pragmatics* 88: 148-162.
Gordon, Elizabeth, Margaret Maclagan and Jennifer Hay. 2007. The ONZE Corpus. In Joan Beal, Karen Corrigan and Hermann Moisl (eds.), *Creating and Digitizing Language Corpora*, vol. 2, Diachronic databases, 82-104. Basingstoke: Palgrave.
Haas, Florian. Forthcoming. "You can't control a thing like that": Genres and changes in Modern English human impersonal pronouns. In R.J. Whitt (ed.), *Diachronic Corpora, Genre and Language Change*. Amsterdam: Benjamins.
Hendy, David. 2007. *Life on Air: A History of Radio Four*. Oxford: Oxford University Press.
Hoffman, Michol and James Walker. 2010. Ethnolects and the city: Ethnic orientation and linguistic variation in Toronto English. *Language Variation and Change,* 22: 37-67.
Hundt, Marianne and Christian Mair. 1999. 'Agile' and 'uptight' genres: The corpus-based approach to language change in progress. *International Journal of Corpus Linguistics* 4: 221-242.
Ito, Rika and Sali Tagliamonte. 2003. *Well weird*, *right dodgy*, *very strange*, *really cool*: Layering and recycling in English intensifiers. *Language in Society* 32: 257–79.
Jucker, Andreas and Daniela Landert. 2015. Historical pragmatics and early speech recordings: Diachronic developments in turn-taking and narrative structure in radio talk shows. *Journal of Pragmatics* 79: 22-39.
Labov, William. 2001. *Principles of Linguistic Change*. Vol. 2: Social Factors. Blackwell.
Leech, Geoffrey, Marianne Hundt, Christian Mair, and Nicholas Smith. 2009. *Change in Contemporary English: A Grammatical Study.* Cambridge: Cambridge University Press.
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.
Magee, Sean. 2012. *Desert Island Discs: 70 Years of Castaways from one of BBC Radio 4's Best-loved Programmes*. London: Bantam.
Millar, Neil. 2009. Modal verbs in TIME: Frequency changes 1923–2006. *International Journal of Corpus Linguistics* 14(2): 191-220.
Milroy, Lesley and Matthew Gordon. 2003. *Sociolinguistics: Methods and Interpretation*. Oxford: Blackwell.
Ndaji, Francis, John Little and Robert Coe. 2016. A comparison of Academic Achievement in Independent and State Schools: Report for the Independent Schools Council. Centre for Evaluation and Monitoring, University of Durham. http://www.isc.co.uk/media/3140/16_02_26-cem-durham-university-academic-value-added-research.pdf (accessed 2017-12-20)

Nevalainen, Terttu and Helena Raumolin-Brunberg (eds.). 2003. *Sociolinguistics and Language History: Studies Based on the Corpus of Early English Correspondence*. Amsterdam: Rodopi.
O'Keeffe, Anne. 2005. *Investigating Media Discourse*. Abingdon: Routledge.
Oxford Dictionary of National Biography. Oxford University Press. http://www.oxforddnb.com/, accessed 2017-12-20.
Price, Jenny. 2012. Old news: Rethinking language change through Australian broadcast speech. In Terttu Nevalainen and Elizabeth Closs Traugott (eds.) *The Oxford Handbook of the History of English*. Oxford: Oxford University Press.
Rayson, Paul. 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4), 519-549.
Sankoff, David and Susan Laberge. 1978. The Linguistic Market and the Statistical Explanation of Variability. In David Sankoff (ed.). *Linguistic Variation: Models and Methods* (pp. 239-250). New York: Academic Press.
Sankoff, David, Henrietta Cedergren, William Kemp, Pierette Thibault and Diane Vincent. 1989. Montreal French: Language, Class and Ideology. In Ralph Fasold and Deborah Schiffrin (eds.) *Language Change and Variation*, 107-118. Amsterdam: Benjamins.
Scannell, Paddy. 1989. Public service broadcasting and modern public life. *Media, Culture and Society*, 11(2): 134-166.
Smith, Nicholas and Cathleen Waters. (Under review). Variation and change in a specialized register: A comparison of random and sociolinguistic sampling in *Desert Island Discs*.
Tagliamonte, Sali. 2002. Variation and change in the British relative marker system. In Patricia Poussa (ed.), *Relativization in the North Sea Littoral: Proceedings of the North Sea Littoral Conference.* Umeå, Sweden: Lincom Europa. 147-165.
Tagliamonte, Sali. 2006. *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press.
Van de Velde, Hans, Roeland van Hout and Marinel Gerritson. 1997. Watching Dutch change: A real-time study of variation and change in standard Dutch pronunciation. *Journal of Sociolinguistics* 1(3): 361-391.
Waters, Cathleen. 2013. Transatlantic variation in English adverb placement. *Language Variation and Change*, 25: 179-200.

Appendix

Transcription scheme for the *DID* Corpus

Code	Feature
<guest id="name">…</guest>	Turn uttered by guest
<host id="name">…</host>	Turn uttered by host
<pause>	Unfilled pause, of any length
<O>…</O>	Overlapping speech
<trunc>…</trunc>	Truncated word, e.g. a <trunc>re</trunc> rebellion
<laugh>	Laughter
<music duration="…">	Musical piece, with duration