5 Comparing Study Designs and Down-Sampling Strategies in Corpus Analysis

The Importance of Speaker Metadata in the BNCs of 1994 and 2014

Lukas Sönning and Manfred Krug

5.1 Introduction

When using corpus data to address questions of linguistic interest, we extract from a collection of texts a set of observations, or hits, which then form the basis of our studies. These observations can be enriched with different types of metadata, the nature of which will depend on their availability and the objectives of the investigation. The most fundamental type of metadata, which is arguably relevant for any type of corpus-based work, is the source of a particular incidence. By 'source', we refer to the language user who produced the token: In writing, this is the author; in spoken corpora, it is the speaker. It is the relevance of this type of metadata, the link between language events and their sources, that the present chapter addresses. We focus on two aspects that we consider to be of broader relevance for corpus-based work. First, we discuss the consequences of ignoring the source of data points for the statistical and linguistic conclusions that we distil from our data. We illustrate that failure to account for the source may lead to qualitative shifts in our results. Factoring in the link between corpus hits and speakers, we obtain more intuitively plausible findings, which coincide with our background knowledge and expectations.

Our second aim is to demonstrate how this type of metadata may usefully inform the design stage of a study. We focus on down-sampling strategies, that is, systematic procedures to select from a large body of corpus hits a smaller subset for detailed study. We illustrate how knowledge about the source of observations can be used to obtain an optimized sub-sample. For illustration, we present two case studies on *actually*, an adverb that has, over the past 200 years or so, developed into a discourse marker. Our linguistic interest lies in traces of this development in contemporary British speech, and we focus on sociolinguistic usage patterns documented in the spoken parts of the BNC corpora.

https://doi.org/10.1017/9781108589314.006 Published online by Cambridge University Press

The remainder of this chapter is structured as follows. Setting the scene for our methodological discussion, Section 5.2 introduces the linguistic background for our case studies and describes the structure of our data in the light of our principal concern: the grouping of tokens at the speaker level. We then move on to the issue of metadata accessibility and identify informational gaps in the output offered by current web interfaces for the British National Corpus (BNC), such as the popular CQPweb (Hardie 2012). Section 5.4 then turns to our first case study, which highlights the sensitivity of statistical conclusions to this feature of the data. Following this, Section 5.5 shifts the focus to the design stage of a study. We discuss the role of down-sampling and illustrate how the grouping of corpus hits by speaker may inform the choice of procedure and help us optimize the efficiency of our down-sampling design. Section 5.6 then draws some general conclusions and offers recommendations for interface design in corpus work.

5.2 Case Study on *Actually*: Linguistic Background and Data Structure

We begin by taking a closer look at the linguistic context and objectives of our two-pronged case study. Our illustrative analyses focus on distributional patterns of *actually*: (i) the overall frequency of this form in speech, and (ii) its position in the clause (initial, medial, final). Our descriptive aims, then, include two recurrent quantities in corpus-based research: the frequency of an item, which is usually expressed in relative terms (e.g. per million words), and the share of a category among a set of alternatives, usually expressed as a percentage. Despite their ancillary role in the present discussion, distributional patterns of *actually* are also of linguistic interest, as we will discuss in Section 5.2.1. Following this, Section 5.2.2 then describes the structure of our data.

5.2.1 Background and Research Questions

Actually ranks among the most frequent adverbs in present-day spoken British English. According to recent analyses (Traugott & Dasher 2002), it was originally an adverb indicating realness, which has developed the additional meaning of an epistemic adversative and, most recently, the pragmatic function of a discourse marker. Existing real-time and apparent-time studies suggest an ongoing, steady increase in frequency in Present-Day English (Waters 2011; Aijmer 2013). A rough sketch of the diachronic development of *actually* is given in Figure 5.1 (adapted from Krug & Sönning 2018), which portrays its semantic history and the recent rise in frequency, which is chiefly linked to its emergence as a discourse marker.

https://doi.org/10.1017/9781108589314.006 Published online by Cambridge University Press

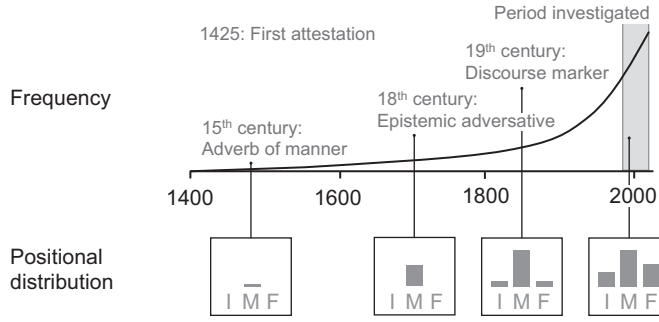

Figure 5.1 A sketch of the diachronic development of *actually* from an adverb of manner to an epistemic adverb and a discourse marker: frequency and positional distribution (initial, medial, final) at different stages. [1]

Our case studies look more closely at Present-Day English variation and change in progress, highlighting the sociolinguistic dimension of variation. We restrict our attention to the categories age and gender. First, we investigate the prevalence of *actually* in different demographic groups. We interpret higher rates as signalling that semantic (or pragmatic) change has progressed further. Our expectation (based on the analyses already mentioned; Krug 1998; Oh 2000; and Mair 2006) for apparent-time patterns is that younger speakers should show a higher usage rate. As for gender, considering that, in variationist sociolinguistic terms, we are looking at a change from below (i.e. emerging from within the language and, at least initially, below the level of consciousness), Labovian principles would point to women as the leaders of change (Labov 1972, 1990).[1]

Our second line of investigation targets the position of *actually* in the clause. As a discourse marker, it is typically found in the periphery, that is, in initial or final position (Aijmer 2002: 253). The distribution of *actually* in the clause should therefore also change diachronically, as illustrated in Figure 5.1: The share of initial and final occurrences is expected to increase over time. Estimates for the share of non-medial occurrences of *actually* in contemporary British speech range between 30% and 60% (Lenk 1998; Aijmer 2002, 2015). We anticipate that this figure will vary systematically across demographic subgroups. Accordingly, we expect younger speakers and female speakers to show a higher proportion of peripheral occurrences.

[1] Images with the symbols in the figure caption have been published under the Creative Commons Attribution 4.0 licence (CC BY 4.0, http://creativecommons.org/licenses/by/4.0) in the accompanying OSF project (https://osf.io/v3byq/).

https://doi.org/10.1017/9781108589314.006 Published online by Cambridge University Press

5.2.2 Data Structure

In order to combine a real- and apparent-time approach to the usage of *actually* in contemporary British speech, we turn to the demographically sampled spoken part of the BNC1994 (Crowdy 1995) and the Spoken BNC2014 (Love et al. 2017). The compilation of these data sources proceeded along similar lines and both represent conversational speech. Throughout this paper, we will refer to these corpora using the shorthand labels 'Spoken BNC1994DS' and 'Spoken BNC2014'. The data analysed in the present chapter are available via the *Tromsø Repository of Language and Linguistics* (*TROLLing*) (Sönning & Krug 2021).

Table 5.1 reports the size of these corpora. Since we are interested in sociolinguistic usage patterns, we need to exclude cases from speakers of unknown sex and age. Further, we restrict our attention to speakers whose overall word count exceeds 100.[2] These exclusion criteria had little effect on the figures for the Spoken BNC2014, but for the Spoken BNC1994DS, the number of speakers went down from 1,408 to 886, and the token count for *actually* decreased from 3,310 to 2,688. A comparison of the totals for speakers and words shows that the 1994 material includes more informants and fewer words per speaker (roughly 5,000 vs 17,000 in the 2014 set).

Let us first inspect how the speakers in our data are distributed across sociolinguistic sub-groups. Here, we concentrate on graphical data summaries. For a more detailed breakdown of the number of speakers, words, and instances of *actually* by age and gender for each corpus, please refer to Table 5.A1 in the Appendix. Figure 5.2 shows, for each data set, the age distribution for male and

Table 5.1 *Corpus size before and after applying our exclusionary criteria*

	Spoken BNC1994DS			Spoken BNC2014		
	Speakers	Words	*Actually*	Speakers	Words	*Actually*
Before exclusion	1,408	5.0 m	3,310	668	11.4 m	17,525
After exclusion[†]	886	4.3 m	2,688	662	11.3 m	17,431

[†] Criteria: information on age and sex of speaker, more than $n = 100$ words in total

[2] Speakers contributing fewer than 100 words ($n = 6$ in the Spoken BNC1994DS, $n = 6$ in the Spoken BNC2014) add little insight to our present investigation. Furthermore, since we will also analyse the number of speakers representing different demographic sub-groups (see Appendix 5.1), they would unduly increase these tallies.

https://doi.org/10.1017/9781108589314.006 Published online by Cambridge University Press

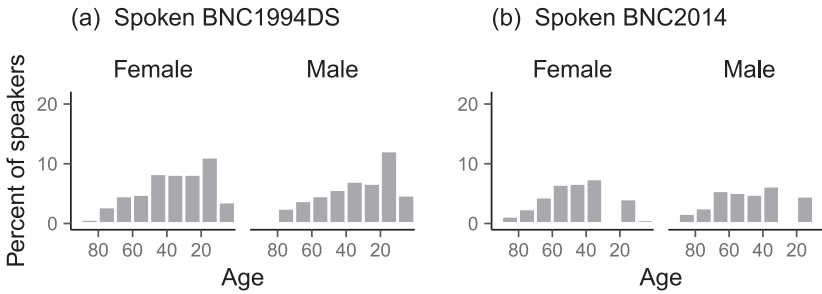

Figure 5.2 Distribution of speakers in the corpora by age and gender.

female speakers.[3] Speakers are arranged from oldest (far left) to youngest (far right). This setup will later facilitate our interpretation of apparent-time patterns as (partially) reflecting real-time change. Both profiles indicate an over-representation of younger speakers: In the Spoken BNC1994DS, the 10- to 20-year-olds account for the largest share; in the 2014 set, about a third of the informants is between 20 and 30 years of age. In both sub-corpora, therefore, younger age groups are over-represented.

The Spoken BNC1994DS and Spoken BNC2014 were composed by systematically recruiting informants who then recorded conversations over a certain period of time. The number of words produced by speakers is therefore free to vary and, as can be seen in Figure 5.3, this data collection regime actually yields a highly uneven distribution of word counts per speaker. We see remarkably similar profiles for the two data sets: Most informants occupy the left corner of the histogram. In the 1994 part, most speakers (75%) contribute fewer than 5,000 words to the corpus; 15 speakers exceed the 40,000-token mark. In the 2014 part (which is roughly three times as big as the older BNC, see Table 5.1), 81% of all speakers contribute fewer than 20,000 words; 10 speakers exceed the 150,000-token mark. To relate these word counts to the total corpus size, we have added percentage marks. The individual contributing most to the Spoken BNC2014, for instance, produced more than 360,000 words and accounts for more than 3% of the total corpus.

These uneven word counts matter when drawing sociolinguistic comparisons. Thus, for our present purposes, we will break down the corpus by age and gender. Certain individuals may boost the word count for a given sub-corpus and have a disproportionate influence on its usage patterns. To get a better

[3] *R* scripts for the data-based graphics can be found in the online supplementary files (https://osf.io/u4v6f/). For data visualization, we relied on the *R* packages 'lattice' (Sarkar 2008) and 'ggplot2' (Wickham 2016).

https://doi.org/10.1017/9781108589314.006 Published online by Cambridge University Press

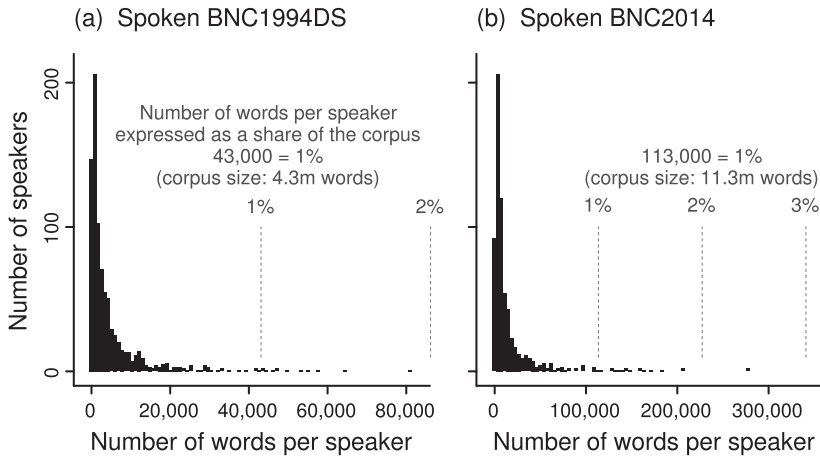

Figure 5.3 Distribution of word counts across speakers.
Note: The total number of words contributed by a speaker varies widely. The percentages indicate the share of the total corpus size in words.

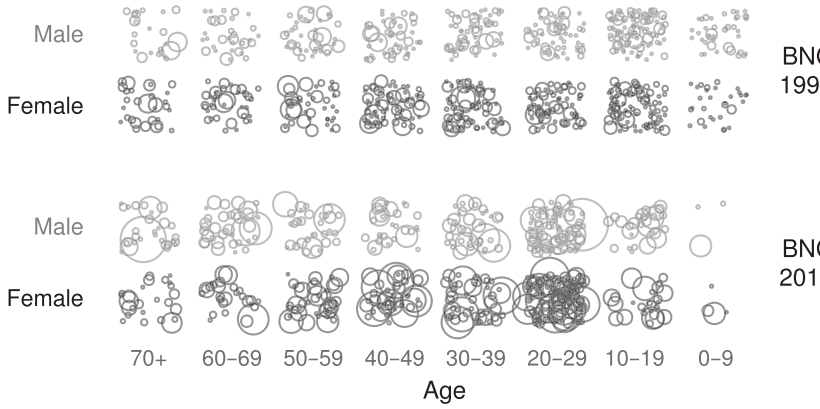

Figure 5.4 Distribution of word counts across speakers in the demographic sub-groups.
Note: Each circle denotes a speaker and the size is proportional to the number of words produced by the speaker.

insight, Figure 5.4 illustrates the word counts by speaker in the different sub-groups. The top two rows show the breakdown for the Spoken BNC1994DS, the bottom rows that for the Spoken BNC2014. Each circle denotes a speaker,

https://doi.org/10.1017/9781108589314.006 Published online by Cambridge University Press

and the size of the circles is proportional to the number of words produced by the speaker. Circles are, on average, bigger in the 2014 sub-groups, reflecting the size of the data set (11.4 m vs 4.3 m words). In the 2014 line-up, we can locate a 20- to 29-year-old male whose word count by far exceeds that of the other speakers in this group. It turns out that this individual makes up 35% of the words in this segment of the corpus. Overall, it seems that uneven distributions are typical of all sub-groups. An analysis of these data must be alert to the disproportionate influence these individuals may have on the quantities we attach to our socio-demographic groups. We keep these general insights into the structure of our data in mind as we discuss the use of software for corpus data retrieval.

5.3 Corpus Interfaces and Metadata Accessibility

There are different tools for processing textual data and retrieving hits from a corpus. Graphical user interfaces such as CQPweb (Hardie 2012) are powerful and convenient means to this effect. As a result, they currently seem to be the most popular choice among linguists. This section therefore takes a closer look at these interfaces with an eye to metadata accessibility. For illustration, we will use CQPweb[4] to retrieve from the Spoken BNC2014 instances of *actually* for the socio-demographic sub-groups of interest.

Figure 5.5 shows a screenshot of the CQPweb interface to the Spoken BNC2014, where we can enter our search string. The restricted-query tab on the left-hand side allows us to restrict our inquiry to a particular part of the corpus, or target group of speakers – in our case, certain socio-demographic sub-groups. Thus, by checking boxes prior to running a corpus query, we can extract hits only for speakers that satisfy certain criteria, say, male speakers aged 60 or older (for whom relevant metadata – i.e. age and sex – are documented). In the Spoken BNC2014, this query for *actually* returns just over 1,000 matches, which we can download, with the option of adding meta-information to the corpus hits. While this allows us to enrich the set of observations with details about the speaker and the communicative situation, we cannot link tokens to speakers directly. Our restricted query therefore allows us to state that each token in this subset was uttered by *some* male speaker aged 60 or older.

Among the download options, we can choose to include the 'sub-text region boundary markers', which adds speaker IDs to the left and right context. Table 5.2 illustrates part of the exported concordances. Scanning the 'context before' column from right to left, we are then, in principle, able to detect the

[4] Both corpora were accessed via https://cqpweb.lancs.ac.uk/.

https://doi.org/10.1017/9781108589314.006 Published online by Cambridge University Press

Table 5.2 *Concordance lines exported from the Spoken BNC2014*

Hit	Text ID	Context before	Query item	Context after
1	S94Z	[S0565:] supposed to be eyes	actually	(.) don't forget it 's
2	S9RV	li- I quite like ouzo	actually	[S0266:] mm [S0309:] but
3	SXWR	it in about ten minutes	actually	(.) he got a bit cross
4	S5PW	[S0013:] yeah [S0008:] so they 're	actually	growing into the other thing
5	SP2X	a really bad idea but	actually	I'm not I was

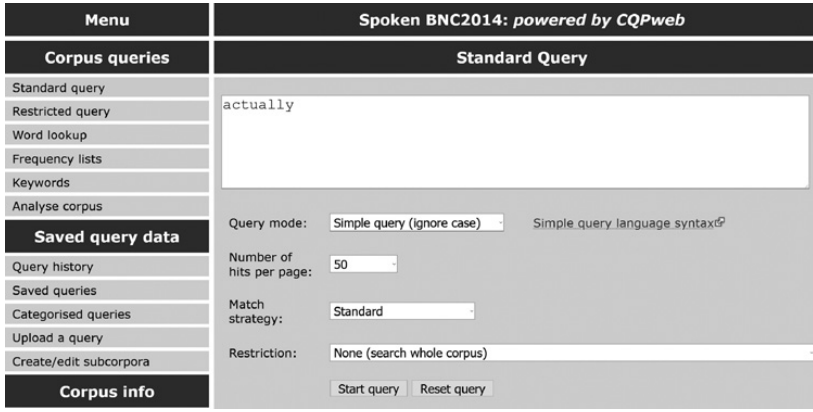

Figure 5.5 The CQPweb interface to the Spoken BNC2014.

speaker ID.[5] It is these identifiers that we need to unambiguously link each *actually* token to an individual.

In order to relate each hit to a speaker, we must invest some extra effort and tap into the plain text files that constitute the corpus. These are often stored in an XML format, and we can process these files using different software, for example *R* (see Gries 2017: 125–7), the *IMS Corpus Workbench* (Evert & Hardie 2011) and *R* packages that combine these resources (e.g. Desgraupes & Loiseau 2018). The command-line access to the raw text files allows us to extract the speaker ID linked to each *actually* token, along with biographical details (which are also available in the web interface) and useful additional information such as the total number of words produced by a speaker.

[5] This strategy fails, however, if the pre-*actually* part of a speaker's turn exceeds 500 words (currently the maximum size of context in the download options).

https://doi.org/10.1017/9781108589314.006 Published online by Cambridge University Press

The link between token and speaker forms the backbone of the ensuing discussion. As we will illustrate, connecting each token to a specific speaker allows us to side-step quantitative pitfalls when analysing corpus data, and bears potential in that it allows us to apply effective down-sampling strategies. Our case studies, outlined in the following sections, deal with these aspects in turn.

5.4 Metadata and Statistical Conclusions: The Usage Rate of *Actually* across Sociolinguistic Categories

Recall that we are interested in tracing the emergence of *actually* as a discourse marker across traditional sociolinguistic categories, which may offer some insight into which groups are leading this change. In other words, we interpret variation in Present-Day English as a reflex of not only ongoing but also historical change. We will juxtapose two analyses: one that does not account for the link between token and speaker, and one that does. Let us start, however, with an overview of the variation in usage rates across speakers.

5.4.1 The Rate of Actually *across Speakers*

We will use the term 'usage rate' (or 'rate', for short) to refer to a normalized frequency, the frequency of *actually* expressed as per million words of running text (pmw). We therefore tacitly assume that the overall word count for a speaker constitutes a valid baseline of comparison.[6] We can compute this rate for each speaker in the corpus. Figure 5.6 displays these estimates, along with a 50% confidence interval indicating the range of statistical uncertainty. Each point denotes a speaker, and speakers are arranged based on their overall word count, with individuals contributing only 100 words to the corpus (our cut-off) sitting at the far left of each panel. We have rescaled the word counts logarithmically for better resolution – the skew that was evident in Figure 5.3 is now 'hiding' in the scale. Likewise, the usage rates for *actually* are rescaled logarithmically.[7] Note how, as the word counts per speaker increase towards the right, the confidence intervals become shorter.

[6] See Chapter 4 of this volume for a discussion of the choice of different baselines. For *actually*, circumscribing a functionally defined envelope of variation appears to be challenging considering its status as a discourse marker (but see Waters 2011 for a study of *actually* in the variationist-sociolinguistic tradition). However, occurrences of *actually* could also be normalized based on the number of sentences or clauses.

[7] Zero (uses of *actually*) cannot be converted to logarithms, so the log-scaled pmw rate is undefined for speakers who did not say *actually*. To include such speakers in this display, we assigned to them an arbitrary single value (of $\exp(1) \approx 2.72$), which was chosen so as to dislocate them sufficiently far from the point clouds (i.e. rates greater than 50) without compromising the resolution of the non-zero rates.

https://doi.org/10.1017/9781108589314.006 Published online by Cambridge University Press

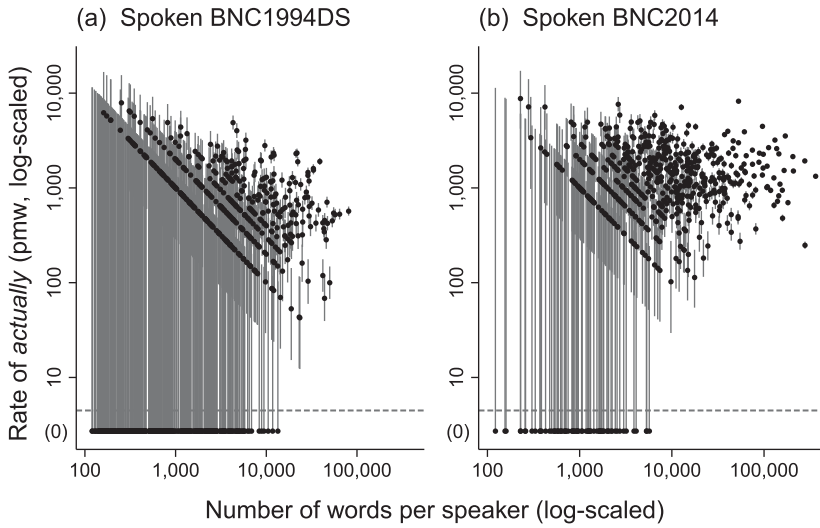

Figure 5.6 Variation in the usage rate of *actually* across speakers: the rate of *actually* plotted against the total word count for each speaker. ⓒⓘ
Note: Error bars reflect 50% confidence intervals.

Rates in panel (b) appear to be higher, on average, which hints at real-time changes in our data. We also note that there is considerable between-speaker variation. In both data sets, the observed rates vary by several orders of magnitude: from 0 to just under 10,000 pmw. As we will show, this feature of the data has important consequences for their statistical treatment.

5.4.2 Analysis without Metadata

For our first approach to uncovering sociolinguistic nuances in the usage rate of *actually*, we retrieve instances via the CQPweb interface. We use restricted queries to extract occurrences of *actually* in each sub-group. For the Spoken BNC1994DS, our search yields the counts recorded in Table 5.3.

In what follows, we will refer to the pmw rates obtained from this query as *crude* rate estimates. This label reflects the fact that these figures do not take into consideration (i) the number of speakers underlying these tallies, which varies substantially (see Figure 5.2 and Table 5.A1 in the Appendix), (ii) the skewed word counts across speakers, a key feature of these data (see Figure 5.4), and (iii) the between-speaker variation in the usage rate of *actually*, which, judging from Figure 5.6, is certainly non-negligible.

https://doi.org/10.1017/9781108589314.006 Published online by Cambridge University Press

Table 5.3 *The rate of* actually *across sociolinguistic sub-groups in the Spoken BNC1994DS: results returned by an interface-based query*

	Male			Female		
Age group	*Actually*	Words	pmw	*Actually*	Words	pmw
0–14	112	247,560	452	91	187,726	485
15–24	205	212,977	963	437	383,136	1,141
25–34	202	287,983	701	308	528,041	583
35–44	291	317,356	917	321	508,501	631
45–59	205	321,379	638	229	538,357	425
60+	153	303,508	504	146	480,086	304

Given that we are interested in estimating the usage rate of *actually* in these sub-groups, our statistical data summaries require only simple procedures. These can be directly applied to the counts in Table 5.3 and no information beyond that given in the table is needed. In the terminology introduced by Evert (2006), the socio-demographic sub-corpora would be treated as random 'bags of words' – collections of observations (i.e. words) with no further information about their internal structure such as the distribution across speakers. To compute inferential data summaries on the basis of the tallies in Table 5.3, we would rely on procedures such as chi-square or likelihood-ratio tests for hypothesis testing. Our focus is on estimation, that is, we compute confidence intervals rather than *p*-values. We employ the Poisson distribution to this end, which allows us to estimate a rate parameter (i.e. a pmw rate) based on the number of events (i.e. *actually*) observed in a certain sub-corpus (measured in words of running text). It treats each sub-corpus as a bag of words without internal structure. In a first step we therefore use this distribution to derive a point estimate for the usage rate in each group (which is equal to the crude rates listed in Table 5.3) and a confidence interval for this estimate. The results are displayed in Figure 5.7a.[8]

The validity of the insights offered by a statistical analysis hinges on a set of assumptions about the data (see e.g. Amrhein, Trafimow & Greenland 2019). The critical assumption of the Poisson distribution is that, within our sociolinguistic sub-groups, the probability of observing *actually* (in other words: the usage rate of *actually*) is constant. This is to say that each speaker in a particular category (e.g. males over the age of 60) is presumed to use *actually* at exactly the same rate; between-speaker variation in the data merely reflects empirical sampling variation. It is up to the researcher to decide whether the simplifying

[8] The *R* scripts for running the analyses with the 'brms' package (Bürkner 2017) are deposited to the OSF (https://osf.io/cpshu/).

https://doi.org/10.1017/9781108589314.006 Published online by Cambridge University Press

assumption of a constant rate within each sub-group is acceptable. Giving some thought to this premise in light of our understanding of language (use), it does not seem appropriate to assume that speakers sub-classified according to age and sex will have even nearly the same likelihood of using *actually* in a conversation. There are other aspects underlying the use of this word, including situational, discourse-pragmatic factors (e.g. Aijmer 2013) and idiolectal features of speech.

Figure 5.6 afforded us the luxury of going beyond Table 5.3 to appreciate between-speaker variation: The large variation in usage rates also casts doubt on the match between data and assumptions. It is clearly unsatisfactory, then, that data in the form obtained via a web interface, as presented in Table 5.3, are not amenable to other analysis strategies. In order to relax the assumption of identical rates within each sub-group, these counts would need to be broken down by speaker. Only then can usage rates be measured at the speaker level and variation in these rates be incorporated into the estimation process.

Before we refine the statistical approach, let us consider the output of this crude analysis. We have visualized the estimates from both corpora in Figure 5.7a. Before we discuss these, we need to comment on the layout of these graphs. Since we are dealing simultaneously with apparent-time and real-time differences, we have decided to show estimates by year of birth. Figures based on the Spoken BNC2014 therefore reach out further to the right. These are graphed using filled circles and solid lines. There are 10 estimates, one for each age bin offered by the web interface. A floating axis at the top of the graph indicates the age groups (with 2014 as the assumed date of recording). For the Spoken BNC1994DS, which is shown using empty circles and dashed lines, there are six age bins, with the age ranges (assuming a recording date of 1992) added at the bottom.[9] The display allows us to discern apparent-time trends, which are reflected by the four curvilinear profiles running from left to right. These show differences between age groups at a single point in time, and we interpret these patterns as indirect hints at language change in real time. Further indications of real-time trends can be read from the display: Thus, vertical differences for the same cohort (i.e. at the same point along the horizontal scale) reflect how language use among speakers with roughly the same birth year changes from 1992 to 2014.

We will return to an interpretation of the crude estimates once we consider Figure 5.7b. For now, however, note the high level of statistical accuracy implied for each of the estimates. Consider, for instance, male children (0–10 years of age) in the Spoken BNC2014. Our analysis suggests that this sub-

[9] This is a slight simplification of the actual recording facts, which actually date from roughly 1991 to 1993, and 2012 to 2016, respectively.

https://doi.org/10.1017/9781108589314.006 Published online by Cambridge University Press

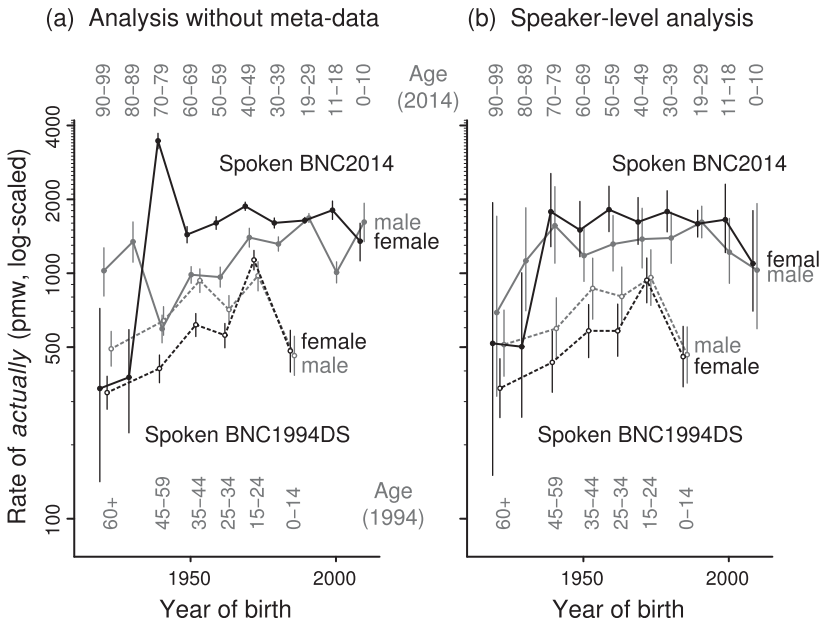

Figure 5.7 Estimates for the usage rate of *actually* in different sub-groups, graphed against year of birth.
Note: Error bars indicate 95% statistical uncertainty intervals.

group uses *actually* at a rate of about 1,600 pmw, the 95% uncertainty interval extending from roughly 1,400 to 1,900 pmw. This level of precision is questionable considering that there are only three speakers in this sub-group. The statistical uncertainty indications are clearly over-confident.

5.4.3 Speaker-Level Analysis

For comparison, let us now take an approach that makes use of the link between token and speaker to accommodate between-speaker variation that is not captured by our sociolinguistic cross-classification. While there is a variety of strategies we could pursue, all basically aim to relax the assumption of equal rates within conditions. This is done by explicitly factoring in the amount of between-speaker variation suggested by the data. We will opt for a fairly standard alternative, the negative binomial distribution (see Ehrenberg 1982: 59–63 for a concise discussion; Mosteller & Wallace 1984 for an application to word rates). This technique takes into consideration (i) the number of different speakers in a sub-group, (ii) the skewed word counts across speakers, and

https://doi.org/10.1017/9781108589314.006 Published online by Cambridge University Press

(iii) the extensive amount of between-speaker variation in the usage rate of *actually*.[10]

Figure 5.7b shows the point and uncertainty estimates returned by our speaker-level analysis. We instantly note the larger uncertainties associated with our estimates. The indications for male children, for instance, are now more cautious. Interestingly, the shape of our apparent-time profiles has changed noticeably, especially for the 2014 data. The age groups pattern more evenly from left to right, which intuitively makes more sense. The spike for 70-to-79-year-old women in the 2014 set has disappeared and we see a more regular pattern with fewer intersections between female and male speakers, whose curves become more parallel (except for the noticeable increase among females in their seventies, which marks the point when women take the lead in the 2014 data).[11] Likewise, the plunge for male 70-to-79-year-olds in the BNC2014 also shifts upwards, back in line, as it were. Figure 5.7b also throws into relief that usage rates in 2014 are consistently higher than 20 years earlier, thus clearly supporting real-time change. The erratic patterns in Figure 5.7a, then, turn out to be due to influential individuals (i.e. outliers with both an unusually high or low usage rate and a high word count). Thus, among 70-to-79-year-old females, one speaker with an extremely high rate of *actually* (8,200 pmw) makes up a quarter of the sub-corpus in terms of word count. The dip observed for 70-to-79-year-old males in Figure 5.7a is likewise due to a single influential speaker, this time with a very low rate of *actually* (about 250 pmw) and an enormous share (75%) of the total word count in this sub-group (cf. Figure 5.4).

To summarize, we have seen that the ability to connect each instance in the corpus to a speaker allows us to sensitize our data summaries to unequal word counts across speakers and between-speaker variation in the outcome quantity – in our case, the rate at which *actually* is used. The data obtained via the web interface forces us to make debatable assumptions, which are at odds with the data and our understanding of the linguistic phenomenon. Our comparison of two analytic approaches showed that incorporating the grouping structure can lead to different statistical conclusions. This not only concerns the validity of

[10] The negative binomial distribution basically considers each speaker as a data point. Based on two counts for each speaker, the total number of words and the number of *actually* tokens, it computes a usage rate for each individual. What we end up with is a distribution of rates (one rate per speaker) and the negative binomial builds the dispersion of rates among speakers into the estimates. Thus, even if a single speaker were to make up half the words in the corpus, the model would consider this a single data point and therefore on a par with the other speakers in the data. If the variation among speakers is large, summary estimates for the sub-groups become less precise.

[11] Notice that this observation is in line with (and a continuation of) the 1994 data, in which female speakers (at least in the refined analysis) of all cohorts have consistently lower usage rates for *actually* than their male counterparts, which is surprising given the tendencies identified by Labov (1990).

inferential uncertainty estimates (confidence intervals and *p*-values), but also the estimates of primary interest: the rate of *actually* in different sub-groups. This means that supposedly simple (or, in our earlier words, crude) data summaries may also be off-target. The fact that the current implementation of web interfaces does not allow us to side-step these issues in the BNC corpora is clearly unsatisfactory. We will return to this issue after the second part of our case study.

5.5 Metadata and Down-Sampling Design: The Positional Distribution of *Actually* across Sociolinguistic Categories

In this section, we take a closer look at the position of *actually* in the clause. Our discussion in Section 5.2.1 showed that the functional layering of *actually* in Present-Day English surfaces in its positional versatility: Medial placement is typically associated with adverbial usage, while as a discourse marker, *actually* is found in the periphery of the clause. The real- and apparent-time increase in the rate of *actually*, which we attribute to its spread as a discourse marker, should therefore be reflected in positional patterns.

5.5.1 Down-Sampling in Corpus-Based Work

This shift of focus requires us to code tokens in terms of their position: Each case must be inspected and assigned to one of three categories: initial, medial, or final. Corpus-based research often requires manual work of this kind, as false positives need to be eliminated or forms may require disambiguation. In some settings, observations are annotated for a large set of variables, not all of which can (or should) be coded automatically. If a case-by-case inspection is necessary for the research task at hand, this phase of the study is likely to consume a considerable, if not the largest, share of our efforts. In many cases, we do not have the resources to analyse all tokens returned by our query. Instead, we select a sub-sample for detailed annotation and analysis, a step that is often referred to as 'down-sampling'.

It is important to note at the outset that sampling and down-sampling are very different tasks. Sampling theory, which is primarily concerned with strategies for collecting samples that are representative of a certain population, is relevant for corpus compilation (see Vetter 2021: 19–32 for a discussion; also Chapter 3 in this volume). Down-sampling theory, on the other hand, which is yet to be established, is concerned with strategies for selecting from an existing sample a subset that is optimal for a particular purpose. It is relevant for corpus-based research under budgetary constraints, where only a fraction of the data can be analysed.

https://doi.org/10.1017/9781108589314.006 Published online by Cambridge University Press

In what follows, we will discuss and illustrate different down-sampling strategies. It turns out that the way in which we compose our sub-sample matters. Not all subsets of hits from a corpus are equally informative. We may consider a down-sampling strategy optimal if it tends to maximize the amount of information attainable under sample size constraints. In general, a cost-effective allocation of resources relies on metadata. As we will illustrate in this section, if the link between a corpus hit and its source (i.e. speaker) is missing, this limits the range of down-sampling schemes we can apply.

In our case, the number of *actually* tokens in our data sets (20,119 in total) is too large for manual annotation. If time constraints allow us to manually inspect a maximum of, say, 1,000 cases, we would like to invest our resources wisely, to maximize inferential information. What this means in practice is that, given our budgetary limitations, we aim for confidence intervals with the smallest possible widths, or hypothesis tests with the highest attainable statistical power for a certain amount of annotation time.

A question that arises in down-sampling concerns sample size: the larger the sample, the greater the yield in inferential information. We should make sure, however, that our effort is balanced against the gain in information. From a statistical perspective, a key insight is that scaling up the sample size yields diminishing returns. In other words, at some point, increasing the number of cases for manual annotation hardly affects the precision of our estimates. Here, recommendations are again sensitive to the research objectives. We discuss, within the context of our case study, reasonable limits for the size of a sub-sample.

Crucially, the choice of down-sampling design depends on our research objectives. Our interest is in the sociolinguistic categories age and gender. These variables are tied to speakers, so speakers should play a key role in our procedure. We adopt terminology from sampling theory (e.g. Thompson 2012) and refer to speakers as primary units, and *actually* tokens as secondary units. The primary units are speakers sampled from different demographic categories. These inform our estimates about usage patterns across age and gender. The secondary units are samples of *actually* tokens from a certain individual (and remember that sometimes there is only one secondary unit per speaker). These secondary units offer information about the language use of a particular speaker, the primary unit from which they are sampled. Data that are organized in this way are often called ‘hierarchical’ or ‘clustered’. This structure, that is, the distinction between primary and secondary units, plays a key role in the following discussion. We will restrict our focus to the Spoken BNC1994DS and assume that our resources allow us to manually code a maximum of 1,000 cases.

5.5.2 *Down-Sampling in the Absence of Metadata*

Currently, corpus-based work seems to default to *simple random down-sampling*, a term we adopt from sampling theory. This means that we randomly select 1,000 cases from the total number of hits returned by our corpus query. Each hit, then, has the same probability of being selected. This procedure is illustrated in Figure 5.8, where each square represents one *actually* token in our data for the Spoken BNC1994DS. The tokens are distributed unevenly across the sociolinguistic groups, with 15-to-24-year-old females accounting for the largest, and female under-14-year-olds for the smallest share. The black squares are the 1,000 cases that form our sub-sample. Since, in this down-sampling scheme, each token has the same probability of being selected – in this illustrative example, about one in three is randomly selected – the sub-sample mirrors the uneven distribution in the total set: Females aged 14 or younger, for instance, are under-represented.

If our research objectives do not foreground a particular sub-group, we may wish to allocate our resources more evenly across the conditions of interest (i.e. cross-classifications of age and gender). Then we might opt for *stratified random down-sampling*. In our case, the demographic sub-groups form the strata and we randomly select the same number of tokens from each stratum. This puts the sub-groups on an equal footing and aims to balance resources and inferential information more evenly. The application of random stratified down-sampling is illustrated in Figure 5.9, where the number of black squares is equal across the cells. Note that for the youngest female sub-group, shown at the bottom left, almost the entire set of tokens is selected.

These down-sampling schemes are applicable using web interfaces, where we can thin a list of corpus hits to a randomly selected subset of *n* observations. To carry out stratified sub-sampling, we need to use the restricted-query options

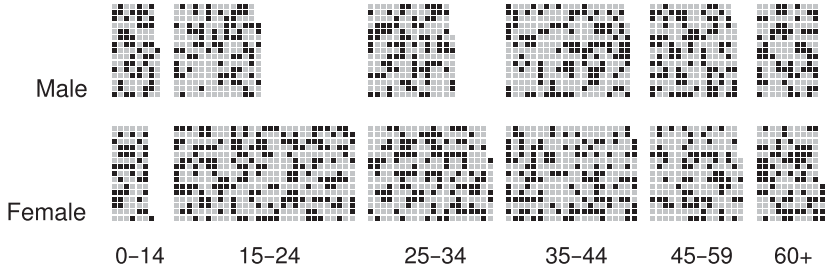

Figure 5.8 Illustration of simple random down-sampling.
Note: Each *actually* token in the Spoken BNC1994DS data has the same probability of being selected.

https://doi.org/10.1017/9781108589314.006 Published online by Cambridge University Press

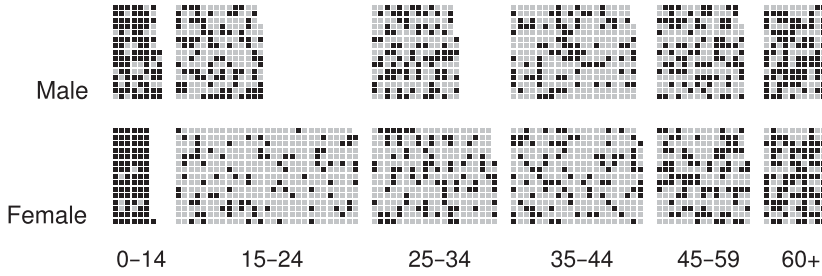

Figure 5.9 Illustration of stratified random down-sampling. 🅭🅯
Note: The number of *actually* tokens selected from the Spoken BNC1994DS is (approximately) balanced across the demographic groups.

to first partition the cases, on a one-by-one basis, into relevant sub-groups and then use the thinning option provided by the interface. Alternatively, we can use the facilities provided by spreadsheet software such as Excel.

5.5.3 Down-Sampling from Skewed Distributions

Upon reflection, we realize that this approach yields a sub-sample in which those speakers who contribute a large number of *actually* tokens are over-represented. It follows that these individuals might then have a disproportionate influence on our data summaries. To decide whether this concern is warranted, let us see how *actually* tokens are distributed across speakers in the Spoken BNC1994DS. We will disregard individuals with zero instances, as they do not inform this part of our case study. This leaves us with 471 speakers. Figure 5.10 arranges them by their token count for *actually*, in decreasing order from left to right. Each square denotes one *actually* token. The distribution is very uneven: headed by a speaker with 74 tokens at the far left, the number of instances quickly levels off towards the right. The annotations indicate the number of speakers with (at least) a certain number of tokens. Thus, there are 320 speakers with 2 or more *actually* tokens, and only 73 individuals contribute 10 or more tokens to our tally.

Our concern seems justified, and we should pause to consider whether a disproportionate representation of speakers in our sub-sample might be problematic. Recall that the focus of our study is to describe sociolinguistic patterns, that is, to break down the data by age and gender. The differences of interest are between speakers, so our down-sampling strategy should aim to maximize information on differences between speakers. In other words, we should include every speaker that is available in the corpus. By relying on simple

https://doi.org/10.1017/9781108589314.006 Published online by Cambridge University Press

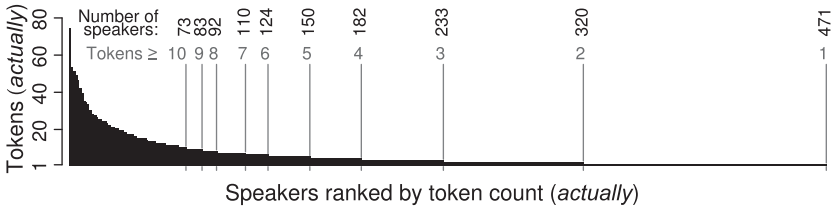

Figure 5.10 Distribution of token counts for *actually* across speakers in the Spoken BNC1994DS. CC BY
Note: Subjects with at least one instance of *actually* (n = 471) are ordered by the number of tokens they contribute to the data set.

and stratified random down-sampling, however, we run the danger of dropping speakers with few instances of *actually*.

5.5.4 *Down-Sampling from Hierarchical Data Structures: A Simulation Study*

While these considerations make sense intuitively, we can use simulation to understand how different down-sampling procedures affect our inferences (cf. Thompson 2012: 32). Our goal is to get an idea of the relative value of increasing the number of speakers in our sub-sample versus increasing the number of tokens per speaker. Consider, for instance, a setting where we have selected 10 speakers (the primary units) and, from each speaker, 10 tokens (the secondary units). This adds up to a total of 100 tokens. Assume that we choose to double our efforts: Is it better then to collect another 10 tokens from each speaker in our sample (i.e. a total of 10 speakers with 20 tokens each); or should we select 10 new speakers, and from each speaker, 10 tokens (i.e. 20 speakers with 10 tokens each)? In terms of annotation work, both constellations produce the same cost.

Our simulation exercise yields two key insights. First, given our study objectives, it turns out that a token from a new speaker is 'worth more'. Our sub-sample should therefore maximize the number of speakers. Further, sampling more than 10 tokens per speaker is probably not worth the effort, as the gain in accuracy levels off rather quickly. We defer details about our simulation study to the web appendix.[12]

To illustrate, consider a simplified setting: our target of estimation is a single percentage value. Our data structure is the following: we have a certain number

[12] https://osf.io/nj3yd/; in our simulation studies, we used the *R* package *aod* (Lesnoff & Lancelot 2019) to compute estimates.

https://doi.org/10.1017/9781108589314.006 Published online by Cambridge University Press

of speakers from the population of interest (the primary sampling units) and, from each speaker, we sample a certain number of tokens (our secondary sampling units). In the interest of simplicity, we assume a balanced design, that is, the same number of tokens per speaker.[13] We are interested in the inferential information offered by different sampling schemes.[14]

Figure 5.11 shows an 'efficiency map', where we can read off the statistical efficiency of different designs (i.e. combinations of speaker and token counts). The horizontal axis shows the total sample size, which represents the scale of effort (i.e. the amount of annotation work invested). The vertical axis shows the width of a 90% confidence interval. Values decrease (!) from bottom to top to indicate precision: The top of the graph shows more efficient designs, which produce narrower confidence intervals. The graph is rich in information and quite complex – we will therefore walk through it step by step.

Let us start with the curves in the graph. The grey dashed lines indicate the number of speakers, which increases from 10 (at the bottom) to 400 (at the top). The black profiles denote designs with the same number of tokens per speaker, starting with 1 at the far left.[15] Three grey vertical bands have been drawn into the graph. Each band marks combinations of speaker and token counts that add up to the same total sample size. For instance, 100 tokens could be from 10 speakers (10 tokens each), or 20 speakers (5 tokens each), and so on. The graph reveals differences in precision: If we choose a 10-by-10 design (i.e. 10 speakers, 10 tokens each), our confidence interval (CI) will be 32 percentage points wide. This design is marked as 'A1' in Figure 5.11. Precision increases as we increase the number of speakers: For design A2 (20 speakers, 5 tokens each) the confidence interval is about 25 points wide, and for A3 (100 speakers, 1 token each) it is at 16 points. This pattern holds in general: Fewer tokens from more speakers are better, and the maximum is attained with 1 token per speaker.[16]

[13] As we have argued and illustrated, corpus data are never balanced in this way. However, this feature of our simulation does not compromise the relevance of the insights for questions of corpus study design.

[14] One complication arises: The added value of sampling from a 'new' versus an 'old' speaker depends on (i) how close the estimated proportion is to 0 or 1 and (ii) how large the variation is between speakers. Variation here refers to how similar speakers are with respect to the quantity of interest: Do they vary greatly in the share of non-medial placement, or is this percentage rather similar across individuals? As we do not know the extent to which speakers vary in the positional distribution of *actually*, we need to consider the sensitivity of our insights to this unknown quantity. In the web appendix (https://osf.io/nj3yd/), we therefore implement a range of reasonable values in our simulations.

[15] The profiles are wiggly due to simulation variance and the fact that we chose only a limited number of representative values (number of speakers, number of tokens per speaker) for our simulation.

[16] If the variation between speakers is small), this pattern still holds. The 'new-speaker benefit', however, is less pronounced. See the web-appendix (https://osf.io/nj3yd/) and note 17.

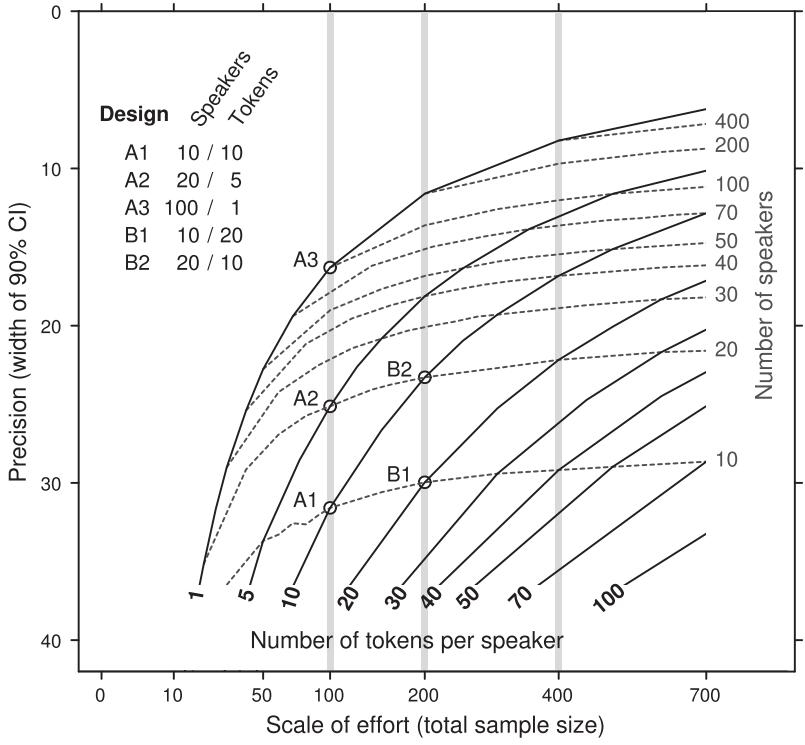

Figure 5.11 Results of our simulation study. CC BY
Note: The vertical axis shows precision, expressed as the width of a 90% confidence interval, where smaller values signal higher statistical precision. The horizontal axis indicates the amount of manual work (i.e. the number of cases in the sub-sample). The black lines reflect the number of tokens per speaker, the dashed grey lines the number of speakers in the sub-sample.

A second way in which we can read Figure 5.11 is to focus on a particular design and read off how much precision we gain when increasing the total sample size. Consider design A1, for instance, with a confidence interval width of 32 percentage points. Increasing the number of tokens per speaker to 20 (i.e. 10 speakers, 20 tokens each) means that we move rightward along the grey dashed line. We arrive at design B1 and the CI width goes down by 2 points, to 30. If we instead add 10 new speakers with 10 tokens each, we move along the black line. This brings us to design B2 and the confidence interval width goes down by 9 points, to 23. This differential effect of increasing the number of

https://doi.org/10.1017/9781108589314.006 Published online by Cambridge University Press

speakers versus the number of tokens per speaker holds in general: No matter what our current sub-sample looks like, increasing the number of speakers yields a greater pay-off in terms of precision. In other words, the difference between confidence intervals for a given number of additional observations analysed is always greater on the (steeper) black lines than on the dashed grey lines.[17]

Moreover, there is a third and final insight we gain from the grey dashed curves in Figure 5.11. Recall that each of these curves represents a fixed number of speakers, and moving rightward along the curve means that we are increasing the number of tokens per speaker. These curves quickly flatten out from left to right, which reflects the diminishing returns of increasing the token count per speaker. Thus, gains in precision level off rapidly and going beyond 5 to 10 tokens per speaker hardly seems worth the additional effort.[18]

Let us reiterate the conclusions we draw from this simulation: Given our research objectives, our sub-sample should include all speakers. Further, sampling more than 10 (if not 5) cases per speaker seems not worthwhile considering the diminishing returns.

5.5.5 Structured Down-Sampling Using Metadata

With the insights from the previous section, we would like to exert some control over the down-sampling procedure. Still operating under the same resource constraints, we would allocate the 1,000 cases as follows: We first rearrange the tokens according to the primary sampling units, the speakers. Then, we draw a random sample of *n* tokens from each speaker. We increase *n* within the leeway of our budgetary constraints, keeping in mind that our provisional upper boundary of tokens per speaker is around 5. In our case, setting *n* to 3 yields a sub-sample of n = 1,017 tokens, so we settle for this scheme.

Figure 5.12 illustrates this down-sampling strategy for the Spoken BNC1994DS. Each square again denotes one *actually* token, and tokens are again sub-divided according to sociolinguistic categories. Now, however, we introduce additional structure by grouping tokens according to speaker. Thus,

[17] Figure 5.11 is based on input values that clearly bring out this differential effect. The estimated proportion is set to 0.5, the variation among speakers to a standard deviation of 2 on the logit scale. Note that we can quantify the differential effect as the difference between designs B1 and B2. In Figure 5.11, this difference amounts to 6.7 percentage points. As the estimated proportion moves closer to 0 (or 1), this difference grows smaller: For proportions of 0.2 and 0.1, we observe a difference of 5.9 and 5.3 percentage points. The same is true if we reduce the variation among speakers: For a standard deviation of 1 and 0.5 (on the logit scale), it is at 3.9 and 1.7 points.

[18] If variation between speakers is small, this threshold seems to be at about 10 to 15 tokens per speaker.

https://doi.org/10.1017/9781108589314.006 Published online by Cambridge University Press

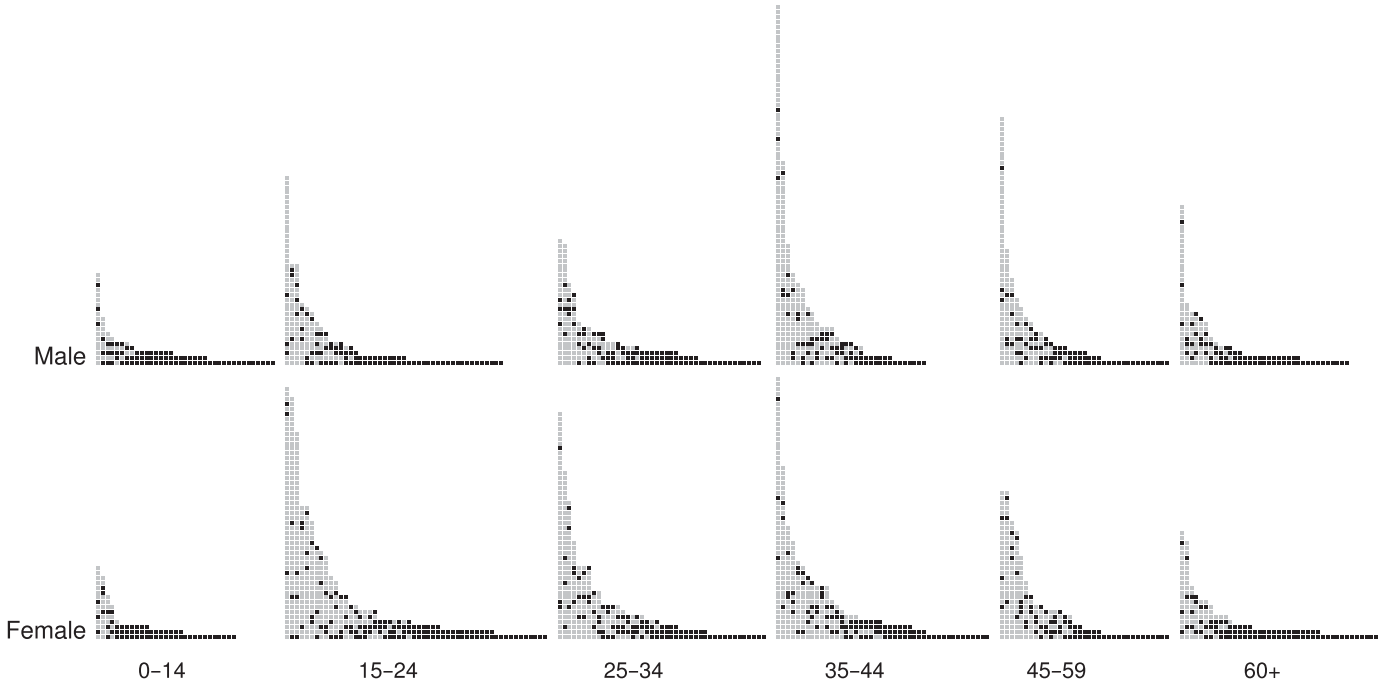

Figure 5.12 Structured random down-sampling: We (randomly) sample a maximum of three tokens per speaker.

within each group, speakers are arranged by the number of *actually* tokens. Note how the skewed token distributions emerge in each partition of the data. We see that the top-scoring individual with 74 tokens, the largest spike in the graph, is male and between 35 and 44 years of age. As in Figures 5.8 and 5.9, black squares indicate those tokens that are included in the sub-sample (i.e. a maximum of 3 per speaker). Where necessary, these are selected randomly from the full set of tokens for a speaker; for about half of the speakers, however, we select all available tokens as their total does not exceed 3 (cf. Figure 5.10).

5.5.6 The Efficiency of Different Down-Sampling Designs

We have discussed three down-sampling designs: (i) simple random down-sampling, (ii) stratified random down-sampling and (iii) structured random down-sampling. Let us now put them to the test. To this end, we annotated all 2,688 instances of *actually* in the Spoken BNC1994DS. For 20 tokens, we were unable to determine the clausal position, which leaves us with 2,668 cases. The inferential information included in this complete set of tokens is the maximum we can extract from our data. Our objective now is to determine how close to this ceiling we get using different down-sampling designs. Note that, in what follows, all analyses heed the hierarchical structure of the data – that is, they are carried out at the speaker level. Our insights therefore isolate the added value of using speaker metadata at the design (rather than analysis) stage of a study.

We first run an analysis with all cases to pin down the highest possible amount of statistical information contained in the data.[19] For each of the 12 conditions (i.e. sub-classifications of age and gender), we compute the percentage of *actually* tokens that occurred in the periphery of the clause (i.e. in non-medial position), and a confidence interval for this estimate. Figure 5.13 shows that the percentage of *actually* in non-medial position ranges between roughly 40% and 70%, which is consistent with previous studies. A comparison of age groups reveals a V-shaped pattern, with younger and older speakers showing a higher share of non-medial usage, on average. Differences between male and female speakers only surface among older age groups, where females tend to show a higher percentage of peripheral cases.

The confidence intervals in Figure 5.13 reflect the amount of statistical information at our disposal when extrapolating from the limited number of speakers in the BNCS1994DS to spoken British English in the early 1990s. For our present purposes, each interval expresses the maximum attainable precision in a sub-group. In what follows, the key quantity will be the width of these

[19] Details are given in the online supplementary materials (https://osf.io/dw6yg/).

https://doi.org/10.1017/9781108589314.006 Published online by Cambridge University Press

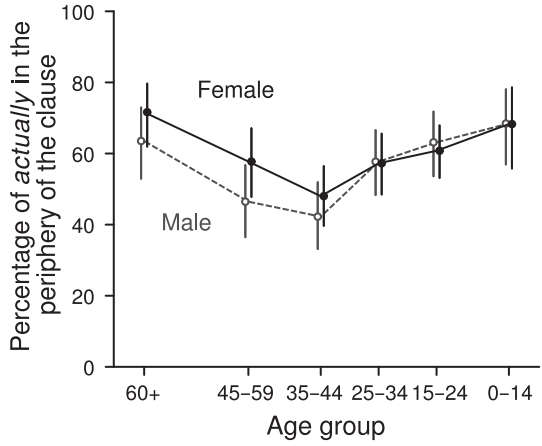

Figure 5.13 The percentage of non-medial (i.e. peripheral) occurrences of *actually* in the Spoken BNC1994DS, by age and gender. 🅭🅯
Note: Analysis of the complete set of n = 2,668 cases. Error bars denote 95% confidence intervals.

confidence intervals in percentage points, and the intervals in Figure 5.13 will serve as a benchmark.

We will assume that our resources allow us to annotate 1,000 tokens in total, and we fix the size of sub-samples to this value. Our interest, then, is to compare the levels of precision yielded by different down-sampling strategies. Since each technique utilizes randomization to some extent, there are different sub-samples we can draw. To balance out this random component, we take 1,000 sub-samples using each procedure and then average over the uncertainty estimates – that is, we compute, for each sub-group, the average width of the 1,000 95% confidence intervals. All analyses take into account the hierarchical structure of the data. A step-by-step documentation of this procedure can be found in the web appendix.[20]

Let us now compare the efficiency of the three schemes. In Figure 5.14, the vertical axis shows precision as we have decided to express it: the percentage point width of a 95% confidence interval. The left half of the display shows results for female speakers, the right half for male speakers. Note the grey area at the top of the graph, which shows the ceiling for our comparative assessment. It traces the precision yielded by an exhaustive analysis of all 2,668 cases (i.e. the width of the intervals in Figure 5.13).

[20] https://osf.io/dw6yg/

https://doi.org/10.1017/9781108589314.006 Published online by Cambridge University Press

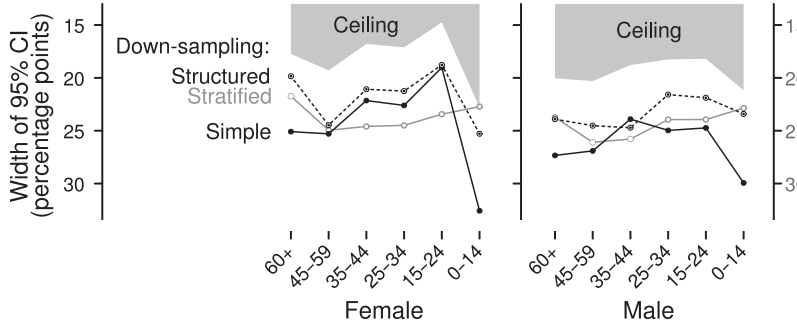

Figure 5.14 Efficiency of down-sampling designs: precision of estimates for each demographic sub-group in the Spoken BNC1994DS. ⓒⓘ
Note: The y-axis shows the width of a 95% confidence interval, decreasing from bottom to top to express precision. The grey area denotes the maximum attainable precision.

Figure 5.14 shows the performance of down-sampling schemes relative to this ceiling. The following points are noteworthy:

- On average, the structured design (i.e. up to three *actually* tokens for the maximum number of speakers, cf. Section 5.5.3) yields the highest levels of precision.
- The stratified scheme (i.e. balancing *actually* tokens across demographic sub-groups, cf. Figure 5.9) allocates precision most evenly across conditions.
- The simple procedure (i.e. a random sample of all *actually* tokens found in the data, cf. Figure 5.8) performs worst, on average, and yields highly uneven levels of precision across sub-groups.

Note how stratified random down-sampling outperforms the other schemes for the youngest speaker groups. As these groups showed the smallest number of *actually* tokens, the stratified scheme, which fixed the number of tokens per condition to 1,000/12 (i.e. about 83), led to a near-exhaustive analysis of this sub-group. Performance is therefore at (or near) ceiling.

Another way of comparing these procedures borrows the notion of 'design effect' from sampling theory (see Lohr 2019: 309–12), where it expresses the additional uncertainty introduced by the choice of sampling design. We can extend this notion to down-sampling to express the additional uncertainty introduced by using down-sampling instead of an exhaustive analysis of all cases. The down-sampling design effect, then, expresses the factor by which the width of a confidence interval increases. A factor of 1.35, for instance, would indicate that the intervals for our sub-sample are 35% wider than those derived from the full set of corpus hits.

https://doi.org/10.1017/9781108589314.006 Published online by Cambridge University Press

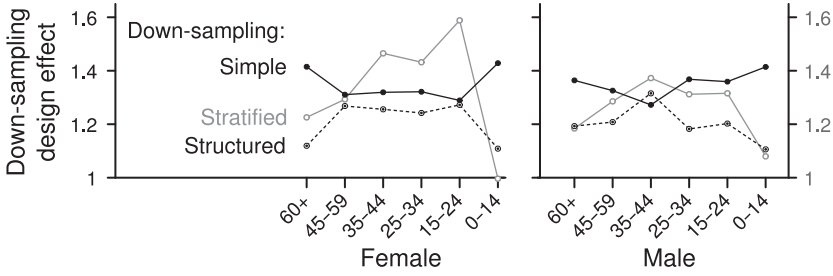

Figure 5.15 Down-sampling design effects for different schemes. CC BY
Note: The values are ratios of the down-sampled confidence interval width divided by the confidence interval width of the analysis using all cases in the Spoken BNC1994DS (i.e. the minimum attainable confidence interval width with the data at hand).

These factors, which are graphed in Figure 5.15, range from 1.00 to 1.60. With average ratios of 1.35 (simple), 1.30 (stratified), and 1.21 (structured), we again see that the structured design performs best for our data.[21] Thus, with the same amount of manual work, we arrive at more precise estimates.

In summary, the second part of our case study has shown that the ability to relate each observation to a speaker proves useful at the design stage of a study. We saw that there are different strategies for winnowing a set of corpus hits to a manageable fraction for detailed study. In the absence of metadata, we need to rely on simple and stratified down-sampling. Structured down-sampling, on the other hand, exploits the token–speaker links to produce an efficient sub-sample that yields the greatest return in terms of the precision of our inferential estimates and, thus, outperforms the other strategies.

5.6 Conclusions and Recommendations for Study and Interface Design

This chapter has been concerned with an elementary type of metadata in corpus-based work: information about the source of instances extracted from a corpus. Our case study drew on spoken corpora to illustrate the benefits of unambiguously linking corpus hits to a speaker or writer (i.e. their source). Using two recurrent outcome quantities in corpus research, frequency of

[21] The youngest male and female cohorts (speakers aged 0–14), as well as the oldest male cohort, are notable exceptions because here the stratified down-sampling method includes almost all tokens of the female speakers and a large proportion of the male ones (cf. Figure 5.9 in Section 5.2), which naturally leads to a factor close to 1.0.

https://doi.org/10.1017/9781108589314.006 Published online by Cambridge University Press

occurrence and the proportional share of certain categories among a set of alternatives (in our case study: peripheral vs medial position), we have illustrated how simple descriptive summaries and inferential uncertainty assessments may mislead us if our analysis does not relate corpus instances to their sources. We also argued that this type of metadata allows us to optimize our strategies for narrowing down a large set of corpus hits to a sub-sample that is amenable to detailed, qualitative analysis. We saw that not all sub-samples are equally informative, and that not all down-sampling designs yield an efficient selection of cases. Table 5.4 offers a concise summary of the points we have raised in this chapter.

Since we have only presented a single case study here, let us consider the generality of the issues we have identified. Based on what has been shown, it seems that there are two features of language data that determine the urgency of these issues. First, if language producers (i.e. speakers or writers) contribute to a data set vastly different numbers of tokens, the issue of disproportionately influential individuals seems particularly pressing. It is difficult to anticipate the amount of imbalance required to yield qualitative shifts even in descriptive summaries, but for data collection regimes that are bound to yield skewed token counts across informants (or text samples), we would expect these issues to be(come) a major concern. The amount of skew in token counts across speakers may also depend on the type of structure under investigation. Thus, the incidence rate of certain formal or functional categories may show greater variation between speakers. *Actually*, for instance, appears to be an item whose usage rate is particularly unstable across individuals.

The second decisive feature is the variability among speakers or writers with regard to the outcome quantity. If variation among informants is large, we need to be particularly alert to imbalance in token counts. Unfortunate combinations of skewed token counts and variability among speakers (or writers) can yield disoriented data summaries, as we observed when estimating the rate of *actually* for 70 to 79-year-old speakers in the Spoken BNC2014 (see Figure 5.7). Thus, the concerns we have raised apply chiefly when studying a variable phenomenon using data that are not balanced in terms of token counts across speakers. Such imbalanced data are common in linguistic analysis, though.

We are currently observing fruitful syntheses between the domains of sociolinguistics and corpus linguistics; corpus data are increasingly employed to address the social dimension of language variation. Surprisingly, however, the necessity of conducting analyses at the speaker level is not highlighted in current discussions of this development (cf. Baker 2010; Andersen 2010; Friginal 2018; Brezina, Love & Aijmer 2018). This neglect, by both practitioners and interface designers, produces a methodological stalemate. The fact that researchers need to draw on command-line interfaces to sensitize their

Table 5.4 *Comparative overview of the advantages and limitations of using speaker metadata at the design and analysis stage of a corpus study*

Analysis without speaker metadata	Analysis with speaker metadata
+ Comfortable data retrieval through popular corpus interfaces	− Retrieval of speaker ID currently requires advanced data processing/command-line corpus queries
− The distribution of tokens across speakers/texts cannot be inspected	+ Data screening for skewed counts across speakers/texts possible; appropriate data-analytic decisions can be made
− Data analysis must resort to crude procedures that ignore the grouping of observations at the speaker level	+ Analyses can be sensitized to the clustering of observations at the speaker level
− Statistical analysis must assume that observations are independent (e.g. that each token stems from a different speaker)	+ The distinction between primary and secondary sampling units can be integrated into the analysis
− Statistical inferences targeting speaker-level variables (e.g. social factors) are based on exaggerated sample sizes	+ Inferences extending across speakers (primary sampling units) are based on appropriate sample sizes
− Statistical uncertainty estimates (e.g. confidence intervals) can be too narrow, and *p*-values can be too small	+ The adequacy of uncertainty estimates is less doubtful, spurious inferences are less likely
− Biased estimates can arise if speakers with unusual behaviour contribute a disproportionate number of instances	+ Estimates can be adjusted for imbalance in sampling; speakers are treated on a par
− Problems are aggravated for (i) data collection regimes that are bound to yield widely varying tokens per speaker; and (ii) structures whose rate of occurrence is unstable across speakers or communicative situations	+ Comparisons between (sub-)corpora rest on a safer statistical footing; differences in corpus compilation can be overcome to some extent at the analysis stage
− Allows only simple or stratified down-sampling schemes	+ When studying variation between speakers, structured down-sampling can be used as a more efficient strategy
− If budgetary limits curtail the number of corpus hits that can enter a study, the allocation of manual work may not be optimal	+ Depending on the focus of a study, the amount of inferential information per token can be optimized using structured down-sampling

analyses to the structure in their data precludes a broader engagement of the corpus linguistic community with the issues we have raised in this chapter. To enable future work to produce more reliable insights into patterns of language use, we can offer the following suggestions for the design of corpus interfaces:

https://doi.org/10.1017/9781108589314.006 Published online by Cambridge University Press

- Output for spoken data should by default include a column indicating the speaker ID.
- To enable researchers to apply structured down-sampling schemes, a further column should be added, where the indexes 1 to *n* (with *n* being the number of tokens contributed by a speaker) are randomly permuted. Exported data can then be sorted by these values, which allows us to prioritize adding new speakers (over adding new tokens for a speaker) to our sub-sample.

It seems that down-sampling has so far not received the attention it may deserve in the corpus linguistic community. As such, it is not discussed in the methodological literature aimed at corpus linguists (e.g. Desagulier 2017; Brezina 2018), and it seems that currently, corpus studies resort to simple random down-sampling as a default strategy. We would argue that the development of expertise in this domain of research methodology holds potential for corpus-based work. Thus, the principled and economic selection of a subset of observations allows us to reallocate resources to other parts of our empirical work.

To conclude, let us put the insights that have emerged from this chapter into a broader methodological perspective. Natural language data, the key object of corpus-linguistic description, produces sets of observations that are organized in systematic ways. An inherent feature of corpus data is therefore their hierarchical structure. We have focused on one type of grouping: the clustering of corpus hits by speaker. By default, the structure in the data must be taken into account for any type of descriptive or inferential data summary. In other words, corpus analysis should always embrace the organization of tokens at the speaker (or text) level.

Further Reading

Ehrenberg, Andrew S. C. 1982. *A Primer in Data Reduction*. Chichester, UK: John Wiley & Sons. Chapter 7.

Johnson, Daniel E. 2014. Progress in Regression: Why Natural Language Data Calls for Mixed-Effects Models. Unpublished manuscript. www.danielezrajohnson.com/johnson_2014b.pdf.

Lohr, Sharon L. 2019. *Sampling: Design and Analysis*. Boca Raton, FL: CRC Press. Chapters 2, 3, 5 and 7.

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage. Chapter 8.

Winter, Bodo. 2020. *Statistics for Linguistics*. New York: Routledge. Chapters 14 and 15.

Winter, Bodo, and Martine Grice. 2021. Independence and Generalizability in Linguistics. *Linguistics* 59(5). 1251–77.

References

Aijmer, Karin. 2002. *English Discourse Particles: Evidence from a Corpus*. Amsterdam: John Benjamins.

Aijmer, Karin. 2013. *Understanding Pragmatic Markers: A Variational Pragmatic Approach*. Edinburgh: Edinburgh University Press.
Aijmer, Karin. 2015. Analysing Discourse Markers in Spoken Corpora: *Actually* as a Case Study. In Paul Baker and Tony McEnery, eds. *Corpora and Discourse Studies: Integrating Discourse and Corpora*. New York: Palgrave Macmillan. 88–109.
Amrhein, Valentin, David Trafimow and Sander Greenland. 2019. Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication. *The American Statistician*, 73(sup1). 262–70. https://doi: 10.1080/00031305.2018.1543137.
Andersen, Gisle. 2010. How to Use Corpus Linguistics in Sociolinguistics. In Anne O'Keeffe and Michael McCarthy, eds. *The Routledge Handbook of Corpus Linguistics*. New York: Routledge. 547–62.
Baker, Paul. 2010. *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.
Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.
Brezina, Vaclav, Robbie Love and Karin Aijmer, eds. 2018. *Corpus Approaches to Contemporary British Speech: Sociolinguistic Studies of the Spoken BNC2014*. New York: Routledge.
Bürkner, Paul-Christian. 2017. brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software* 80(1). 1–28.
Crowdy, Steve. 1995. The BNC Spoken Corpus. In Geoffrey Leech, Greg Myers and Jenny Thomas, eds. *Spoken English on Computer: Transcription, Mark-Up and Annotation*. Harlow: Longman. 224–34.
Desagulier, Guillaume. 2017. *Corpus Linguistics and Statistics with R: Introduction to Quantitative Methods in Linguistics*. Cham: Springer.
Desgraupes, Bernard, and Sylvain Loiseau. 2018. rcqp: Interface to the Corpus Query Protocol. R package version 0.5. https://CRAN.R-project.org/package=rcqp.
Ehrenberg, Andrew S. C. 1982. *A Primer in Data Reduction*. Chichester, UK: John Wiley & Sons.
Evert, Stefan. 2006. How Random Is a Corpus? The Library Metaphor. *Zeitschrift für Anglistik und Amerikanistik* 54(2). 177–90.
Evert, Stefan, and Andrew Hardie. 2011. Twenty-First Century Corpus Workbench: Updating a Query Architecture for the New Millennium. In *Proceedings of the Corpus Linguistics 2011 Conference*. Birmingham, UK: University of Birmingham.
Friginal, Eric. 2018. *Studies in Corpus-Based Sociolinguistics*. New York: Routledge.
Gries, Stefan Th. 2017. *Quantitative Corpus Linguistics with R: A Practical Introduction*. New York: Routledge.
Hardie, Andrew. 2012. CQPweb: Combining Power, Flexibility and Usability in a Corpus Analysis Tool. *International Journal of Corpus Linguistics* 173. 380–409.
Krug, Manfred. 1998. Progressive and Conservative Modern British Dialects: Evidence from the BNC. Paper presented at the 19th ICAME Conference, Newcastle, Northern Ireland.
Krug, Manfred, and Lukas Sönning. 2018. A Sociolinguistic Study of *Actually* in Spoken British English. Paper presented at ICAME39, Tampere. https://osf.Io/472dt.
Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia, PA: University of Philadelphia Press.
Labov, William. 1990. The Intersection of Sex and Social Class in the Course of Linguistic Change. *Language Variation and Change* 2. 205–54.

Lenk, Uta. 1998. *Marking Discourse Coherence: Functions of Discourse Markers in Spoken English*. Tübingen: Gunter Narr Verlag.

Lesnoff, Matthieu, and Renaud Lancelot. 2019. *aod: Analysis of Overdispersed Data*. https://CRAN.R-project.org/package=aod. R package version 1.3.1.

Lohr, Sharon L. 2019. *Sampling: Design and Analysis*. Boca Raton, FL: CRC 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). 319–44.

Mair, Christian. 2006. *Twentieth-Century English: History, Variation and Standardization*. Studies in English Language. Cambridge: Cambridge University Press.

Mosteller, Frederick, and David L. Wallace. 1984. *Applied Bayesian Inference: The Case of The Federalist Papers*. New York: Springer.

Oh, Sun-Young. 2000. *Actually* and *in Fact* in American English: A Data-Based Analysis. *English Language and Linguistics* 4(2). 243–68.

Sarkar, Deepayan. 2008. *Lattice: Multivariate Data Visualization with R*. New York: Springer.

Sönning, Lukas, and Manfred Krug. 2021. Actually in Contemporary British Speech: Data from the Spoken BNC Corpora. https://doi.org/10.18710/A3SATC. DataverseNO, V1, UNF:6:rp13HUEAY75735Bcul7eCg== [fileUNF].

Thompson, Steven K. 2012. *Sampling*. Hoboken, NJ: John Wiley & Sons.

Traugott, Elizabeth C., and Richard B. Dasher. 2002. *Regularity in Semantic Change*. Cambridge: Cambridge University Press.

Vetter, Fabian. 2021. Issues of Corpus Comparability and Register Variation in the International Corpus of English: Theories and Computer Applications. Doctoral Dissertation, University of Bamberg. doi: https://doi.org/10.20378/irb-52406.

Waters, Cathleen M. 2011. Social and Linguistic Correlates of Adverb Variability in English. PhD dissertation. Department of Linguistics, University of Toronto.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

Appendix 5.1

Table 5.A1 *Distribution of speakers (after application of exclusionary criteria), overall word counts and* actually *tokens across socio-demographic sub-groups*

	Speakers				Words				*Actually*			
	Female		Male		Female		Male		Female		Male	
Sub-group	%	N	%	N	%	N†	%	N†	%	N	%	N
BNC 1994												
0–9	3	31	5	42	1	46	2	98	1	15	1	26
10–19	11	100	12	108	9	372	7	299	13	343	8	219
20–29	8	75	7	58	8	344	5	236	11	288	8	203
30–39	8	72	7	62	13	548	5	237	11	309	8	210
40–49	8	74	6	52	11	478	6	267	9	236	8	213
50–59	5	42	5	41	8	355	6	248	6	162	6	158
60–69	5	41	4	34	6	245	3	136	4	101	3	72
70+	3	30	3	24	5	235	4	167	2	55	3	78
Total	52	465	48	421	61	2,623	39	1,688	56	1,509	44	1,179
BNC 2014												
0–9	1	4	0	3	1	80	1	65	1	105	1	108
10–19	4	27	5	30	5	511	4	459	6	1,006	3	507
20–29	21	137	14	91	24	2,708	11	1,199	25	4,327	12	2,016
30–39	7	49	6	41	9	1,038	6	632	10	1,665	5	834
40–49	7	44	5	32	11	1,295	3	338	14	2,432	3	472
50–59	7	43	5	34	6	710	4	457	7	1,139	3	441
60–69	4	29	5	36	4	432	6	633	4	622	4	623
70+	4	25	5	31	2	258	5	523	4	731	2	403
Total	55	358	45	298	62	7,032	38	4,306	69	12,027	31	5,404

Note: Percentages denote the share of total number of speakers/words/*actually* tokens in the corpus;
† Token counts are scaled by 1,000 (i.e. 80 refers to 80,000).

https://doi.org/10.1017/9781108589314.006 Published online by Cambridge University Press