

## Secondary Publication



Sönning, Lukas; Schlüter, Julia

### **Comparing Standard Reference Corpora and Google Books Ngrams : Strengths, Limitations and Synergies in the Contrastive Study of Variable h- in British and American English**

Date of secondary publication: 25.05.2023

Version of Record (Published Version), Bookpart

Persistent identifier: urn:nbn:de:bvb:473-irb-595434

#### **Primary publication**

Sönning, Lukas; Schlüter, Julia: Comparing Standard Reference Corpora and Google Books Ngrams : Strengths, Limitations and Synergies in the Contrastive Study of Variable h- in British and American English. In: Data and Methods in Corpus Linguistics : Comparative Approaches. Schützler, Ole; Schlüter, Julia (Hg). Cambridge ; New York : Cambridge University Press, 2022. S. 17-45. DOI: 10.1017/9781108589314.002.

#### **Legal Notice**

This work is protected by copyright and/or the indication of a licence. You are free to use this work in any way permitted by the copyright and/or the licence that applies to your usage. For other uses, you must obtain permission from the rights-holder(s).

This document is made available with all rights reserved.

# 1 Comparing Standard Reference Corpora and Google Books Ngrams

Strengths, Limitations and Synergies in the Contrastive Study of Variable *h-* in British and American English

---

*Lukas Sönning and Julia Schlüter*

## 1.1 Introduction

Corpus linguistics has recently witnessed an almost exponential increase in the size of the databases available. This chapter will explore resources that are wide apart in terms of word count: two standard reference corpora – the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA) – and the Google Books Ngram (hereafter GBN) database (Michel et al. 2010). The main attraction of the latter is, of course, its unparalleled size. However, big data tend to come with reduced monitoring of what goes into these collections and a limited amount of metadata available for the analysis (Hiltunen, McVeigh & Säily 2017).

The linguistic question we pursue in this chapter concerns the strength of the onset consonant in *h*-initial words. In Present-Day English, the phonetic realization of orthographically represented [h]-onsets is gradient, and lexemes form a cline from strong (e.g. *hand*, *high*) to weak or absent (e.g. *hour*, *honest*). In writing, this is reflected in the choice of indefinite article allomorph (i.e. *a hand* vs *an hour*), and lexemes in the middle of the cline occur with both variants (e.g. *hypothesis*, *historical*). Comparative accounts of British English (BrE) and American English (AmE) have pointed out that, in general, onset [h] appears to be weaker not only in many rural and urban British accents, but also in certain types of words in standard BrE (Cruttenden 2014: 207; Schlüter & Vetter 2020). This claim provides the linguistic setting of our methodological discussion.

The main focus of this chapter will be on two key differences between standard corpora and GBN – issues related to size and issues related to metadata – and their implications for statistical analysis. Size, in our case study, manifests itself in the number of types (i.e. the different lexical items that can enter the analysis) and in token frequency (i.e. the number of hits for

a certain lexical type). As for metadata, we will be concerned with information about the source of each data point (i.e. text file or book), and the genre represented by a text. The standard corpora allow us to trace each instance to a text sample and record relevant attributes of the source, including text category. The metadata provided by GBN, on the other hand, are limited to year of publication and British or American English.

Assuming that the indefinite article is a valid indicator of the presence or absence of an onset consonant, we rely on written texts to shed more light on British–American contrasts. The BNC and COCA provide natural starting points for our investigation. However, we would also like to query the GBN database to expand the range of *h*-initial lexemes for detailed study. Our enthusiasm, however, is curbed by a concern about the absence of metadata: comparisons between varieties (BrE vs AmE) and databases (corpora vs GBN) may be distorted, for instance, by differences in genre composition. The more richly annotated standard corpora allow us to adjust our comparisons for such (potentially) confounding variables. To illustrate, we capitalize on the BNC and COCA metadata to determine the sensitivity of our linguistic insights to these factors and assess the direction and amount of bias that arises. In a concerted effort, the feedback loop between corpora and big data resources then helps us put GBN figures into a more proper perspective and form some judgement as to the validity of our conclusions. Our case study therefore not only discusses relative merits and shortcomings of these two sources of language data, but also illustrates synergy opportunities arising from a joint analysis.

The present chapter is structured as follows. In [Section 1.2](#), we elaborate on the linguistic background and our research questions. [Section 1.3](#) describes data retrieval procedures for both sources and compares the data sets. Issues arising in drawing statistical comparisons between diverse text collections are considered in [Section 1.4](#). [Section 1.5](#) illustrates how corpus metadata can be leveraged to address concerns about the validity of comparisons. In [Section 1.6](#), we discuss the GBN data in the light of these insights and highlight the affordances and limits of a quantitative analysis of data from this resource. [Section 1.7](#) closes with a summary and discussion of the key points. Data, scripts and web appendices are archived as an OSF project.<sup>1</sup>

## 1.2 Introduction to Our Case Study

### 1.2.1 *Variation in Onset Strength in h-Initial Lexemes*

The consonant [h] has attracted ample attention as one of the weak sounds in English that are liable to variation and loss (cf. two chapters in [Minkova](#)

<sup>1</sup> <https://osf.io/47p6u/>

2009). In early Middle English, even word-initial [h] was no longer obligatorily realized, especially in Midland and Southern varieties, from which Standard English was to develop (Lass & Laing 2010: 348; Minkova 2014: 107). Its deletion marks the culmination of a long-term weakening of the consonant.

The re-establishment of initial [h] has commonly been regarded as driven by spelling pronunciation (Minkova 2014: 107). Besides the orthographic pull, which applies to *h*-initial lexemes across the board, recent research has pointed to two dimensions of systematic variability: (i) etymological provenance (native Germanic vs borrowed Romance words), with loanwords lagging behind due to their weak [h] in the donor languages; and (ii) the phonological prominence of the word-initial syllable, with more prominent onsets spearheading the re-emergence (Schlüter 2019; Schlüter & Vetter 2020).

A further factor that has been mentioned, but has remained largely unexplored (except in the methodologically oriented contribution by Schlüter and Vetter (2020)), is a lingering difference between the present-day standard varieties of BrE and AmE: Cruttenden (2014: 207) states that many speakers of British English still drop the [h] in words like *history* and *hotel*, at the same time pointing out that the use of *an* does not necessarily indicate the dropping of [h]. Peters (2004: 1) and Algeo (2006: 49) quote limited corpus evidence showing that BrE is more prone to use the long form of the indefinite article before certain *h*-initial words than AmE. Conversely, Peters (2004: 1) mentions the item *herb* as “the only distinctive case” where the *h* is commonly pronounced in BrE, but not in AmE. Dictionaries like the *Longman Pronunciation Dictionary (LPD)*, the *English Pronunciation Dictionary (EPD)* and the *Oxford English Dictionary (OED)* additionally list [h]-less forms for *heir*, *honour* (for BrE and AmE) and *homage* (for AmE) as well as for words derived from these stems. Overall, this amounts to a relatively poor state of research, which prompted the present study.

### 1.2.2 Research Questions

From a linguistic perspective, we set out to investigate differences between the two standard varieties in terms of onset strength in *h*-initial words. Based on claims in the literature about generally weaker [h]-onsets in BrE, we study the degree to which this cline materializes in the distribution of the indefinite article allomorphs in written language. Thus, we operationalize relative onset strength as the proportion of *a* (vs *an*): the higher the proportion of *a*, the stronger the consonantal onset of the *h*-word is assumed to be. The research questions guiding the following analyses are:

- Does the claim that [h]-onsets are, on average, weaker in BrE show in our data?

- If so, does the difference hold across lexemes or do we observe variation among lexemes?
- Is the difference more notable in some genres than in others?

### 1.3 Data from Google Books and Two Standard Corpora

In this section, we describe how we obtained our data and which criteria we applied to restrict the scope of our analysis. We first explain how we extracted tokens from the BNC and COCA (Section 1.3.1) and then move on to GBN (Section 1.3.2). Section 1.3.3 closes with a comparison of the two sets of data. All *R* scripts used for data collection and processing are available in the OSF repository.

#### 1.3.1 Data Retrieval from the BNC and COCA

The BNC was compiled in the 1990s according to a pre-defined sampling scheme. The written part, relevant for our study, contains around 3,000 text files and almost 90 million words (see Table 1.1). The bulk of the material comes from the early 1990s and text samples are generally limited to a maximum of 45,000 words (Burnard 2007). By contrast, the written part of COCA (in the offline version used for this study) comprises around 180,000 text files and almost 420 million words (discounting 5% of data unavailable for our search for copyright reasons).<sup>2</sup> It contains a similar cross-section of genres (academic texts, magazines, newspapers and fiction in approximately equal shares; Davies 2008–). The years of coverage (1990 to 2017) are represented with roughly identical corpus sizes.

A comparison of the average text file length (30,000 words in the BNC vs 2,300 words in COCA) indicates differences in structure. While COCA uses one file per text, and texts can mostly be traced back to their authors, the BNC files frequently consist of several texts from the same periodical or from a thematic newspaper section.<sup>3</sup> Thus, in a random sample of 100 text files from the BNC written domain, only 45% consisted of one text from beginning to end (mostly excerpts from fictional and informative books); the majority of text files include texts from different authors. The metatextual information provided by the BNC therefore does not allow us to link each instance to an individual author.

For data retrieval, processing and analysis, we relied on *R* (R Core Team 2019). We used the *R* package ‘rcqp’ (Desgraupes & Loiseau 2018) to query the

<sup>2</sup> See [www.corpusdata.org/limitations.asp](http://www.corpusdata.org/limitations.asp).

<sup>3</sup> For a list of the files making up the BNC, see [www.natcorp.ox.ac.uk/docs/URG/bibliog.html](http://www.natcorp.ox.ac.uk/docs/URG/bibliog.html); for COCA, see [www.english-corpora.org/coca/files/coca\\_2019\\_12.zip](http://www.english-corpora.org/coca/files/coca_2019_12.zip).

BNC and COCA for sequences of ‘a’, ‘A’, ‘an’ and ‘An’ followed by an *h*-initial word (upper or lower case).<sup>4</sup> We then converted all strings to lower case and manually inspected the list of unique instances to filter out false positives (e.g. initialisms such as *HIV*, *HBO*). Since our focus is on differences between BrE and AmE, we then excluded (i) those *h*-words that occurred in only one corpus and (ii) those that occurred categorically and with the same variant in both corpora, as these do not shed light on differences between the standard varieties. This left us with 154 candidate items, of which 4 (*hew*, *hid*, *hi*, *heigh*) had to be excluded since they failed to meet our inclusion criteria after sorting out typos (*hew* for *few*), abbreviations (*HID*, *HEW*, *HI*) and a quotation of a Middle English spelling variant (*heigh*).

### 1.3.2 Data Retrieval from Google Books

The Google Books Ngrams collection contains 468 billion words from over 4.5 million English books, contributed by over 40 university libraries and publishers around the world (Michel et al. 2010: 1). The raw GBN data can only be accessed in the form of *n*-gram files, which, of course, greatly reduces the types of structures that can be studied in the first place. These files document, for each 1-, 2-, 3-, 4- and 5-gram, how often it occurred in a given year (*match\_count*) and in how many different books (*volume\_count*). This limited amount of metadata is aggravated by lack of representativeness (over-sampling of books stocked in academic libraries) and the inclusion of multiple copies, editions and reprints of older works (Pechenick, Danforth & Dodds 2015). Further, *n*-grams need not correspond exclusively to existing (English) words, and complications arise from errors in optical character recognition (OCR). Thus, the same lexeme, or word form, may be represented by different character strings.

As we are interested in *h*-initial words following an indefinite article, we need to access the 2-grams beginning with the two indefinite article allomorphs.<sup>5</sup> Due to our focus on Present-Day English, we excluded all occurrences prior to 1975 to roughly align the GBN data with the BNC. At this stage, we counted 15,856 unique *h*-strings. To cope with the inaccuracies illustrated in Web Appendix 1,<sup>6</sup> we decided to exclude from consideration those items that contained (i) a digit, (ii) a punctuation mark, (iii) a sequence of two capital

<sup>4</sup> We thank Fabian Vetter for creating and maintaining the technical infrastructure that made it possible to run these queries and the computationally expensive statistical analyses.

<sup>5</sup> The files available from the repository at <http://storage.googleapis.com/books/ngrams/books/datasetsv2.html> are named after the first two characters in the *n*-gram. Thus, we used the 2-gram files labelled ‘an’ (for the long article allomorph) and ‘a\_’ (where ‘\_’ denotes a space) for the short article allomorph.

<sup>6</sup> <https://osf.io/n5gxx>

letters or (iv) no lower-case vowel (*a e i o u y*). This left us with 14,095 unique strings. We then transformed all characters to lower case and reduced the list to those items that occurred in both varieties. This brought the number down to 6,675. Next, we computed, for each item, the proportion of *an*-tokens among all instances and excluded lexemes that occurred categorically with the same indefinite article variant in both varieties. This reduced the set to 1,484 strings, which then needed to be cleaned to a set of actual lexemes (since, at this stage, the set of *h*-strings still included (nonce) items such as *hofmannsthal*, *hih* and *hrc*). Of these, 956 were either *h*-lexemes or variants of these (i.e. inflected forms,<sup>7</sup> OCR-induced deviations or obsolete spellings), which were manually assigned to the standard spellings. This resulted in a total of 827 unique types, or items, for analysis, which are listed in Web Appendix 2.<sup>8</sup>

### 1.3.3 Comparison of the Data Sets

Key characteristics of our data sources and the derived data sets are reported in Table 1.1. In the period under investigation (from 1975 onwards), the GBN database counts 221 billion words in the American part and 58 billion words in the British part (a ratio of about 4 to 1). As it happens, a similar imbalance holds between the corpora, with about 417 million written words in COCA and roughly 88 million in the BNC (a ratio of 5 to 1). The GBN word count (from 1975) therefore exceeds that in the corpora by a factor of over 500 to 1.

We would expect the 150 types in the corpus data to be a proper subset of the GBN set. This is largely the case, with the exception of five types that are absent from our GBN data: *hairpin*, *heparin*, *hexadecimal* and *hiking*, which occur

Table 1.1 Overview of the data sources: number of *h*-types and *h*-tokens in the corpora and GBN

Data source	General structure		Our data	
	Words	Text IDs	Tokens	Types
Corpora				
COCA, 1990–2017 (written part)	417,295,550	178,686	172,943	150
BNC, ~1975–1993 (written part)	87,903,571	3,141	41,564	150
GBN, 1975–2012				
American English	220,642,393,621		122,225,722	827
British English	57,623,773,839		28,770,980	827

<sup>7</sup> We treated *-ly* adverbs as variants of a type (e.g. *highly* was assigned to the type *high*).

<sup>8</sup> <https://osf.io/cnvxf>

categorically with *a* in the British and American parts, as well as *home-grown*, which did not pass the filter due to the punctuation sign.

Figure 1.1 shows how the items are spread out in terms of their frequency of co-occurrence with *a/an*. To establish comparability between the two data sets, the GBN frequencies are used as a basis for division into frequency bins – words that occur in both data sets then end up in the same bin. The types extracted from GBN range from about 0.0001 to 80 pmw, and those from the corpora from 0.008 to 80 pmw.<sup>9</sup> With their limited size, the corpora therefore fail to capture those types that rarely co-occur with *a/an*. Figure 1.1 also clearly demonstrates the greater number of types in GBN compared to the BNC and COCA.

It is instructive to look at Web Appendix 2, which lists, in rank order of occurrence, the items in the GBN data. Lexemes in bold print also occur with both article allomorphs in the corpus data. Among the high-frequency forms, the grey instances are those that found their way into this study via the GBN data, but not the corpora. Corpus instances of *hand*, for instance, which ranks ninth, categorically occurred with *a* in the BNC and COCA and were therefore filtered out by our exclusion criteria. A manual check using the Google Books Ngram Viewer revealed that sequences of *an hand* in

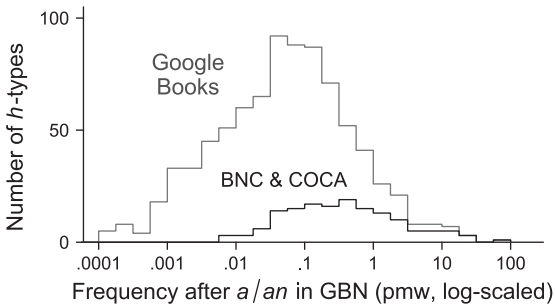




Figure 1.1 The distribution of lexemes in terms of their co-occurrence rate with *a/an* in the corpus data ( $n = 150$ ) and the GBN data ( $n = 827$ ). <sup>10</sup>

<sup>9</sup> Note that the pmw frequencies quoted here are those observed for the GBN data. Thus, the least frequent sequence in the corpus data (*a/an hydroxyapatite*) occurred at a rate of 0.008 pmw in the GBN data.

<sup>10</sup> 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/47p6u/>).



post-1975 GBN were in fact mostly found in German passages.<sup>11</sup> Errors of this kind, which we collectively refer to as ‘noise’, probably account for the surplus of high-frequency types in the GBN data we see in Web Appendix 2 (and Figure 1.1).

At this point, we clearly note the size advantage of GBN: the number of types and tokens available for analysis far exceeds that in the corpora. Before we use these data to quantify differences between BrE and AmE, we should reflect on the comparability of our data sources and the question of whether they allow us to draw valid statistical comparisons.

#### 1.4 The Validity of Statistical Comparisons: Illustration

Given our objective to describe pronunciation differences between the two standard varieties, we would ideally like to control for factors that may distort this comparison. Consider, for instance, the behaviour of *historic*. Table 1.2 lists the counts observed in the four data sets and Figure 1.2 shows them graphically. The error bars indicate 95% uncertainty intervals (UIs; cf. Gelman & Greenland 2019) as a crude first approximation to the statistical precision of the estimated percentages.

We note that the [h]-onset appears to be stronger in AmE. However, the corpus data suggest a greater difference between the varieties (36 percentage points vs 11 points in GBN), which may raise doubts about the comparability of the corpus and the GBN data. One explanation that comes to mind is genre differences. It might be the case that the cline between the varieties is more levelled in academic writing. If that were the case, we should, of course, also make sure that our comparisons between the BNC and COCA are not distorted

Table 1.2 *Observed counts for a/an historic in the four data sets*

Data set	<i>a</i>	<i>an</i>	Total	Share of <i>a</i>	95% UI
Corpora					
BNC	90	92	182	49.5%	[42.3%, 56.7%]
COCA	1,363	223	1,586	85.9%	[84.1%, 87.6%]
Google Books Ngrams					
BrE	50,981	31,477	82,458	61.8%	[61.5%, 62.2%]
AmE	261,348	95,405	356,753	73.3%	[73.1%, 73.4%]

<sup>11</sup> See <https://books.google.com/ngrams>. *An Hand*, in present-day German *anhand*, is a frequent sequence meaning ‘by means of’, and has grammaticalized into a preposition. Since we transformed our data to be case-insensitive, capitalization could no longer be applied as a criterion for exclusion.

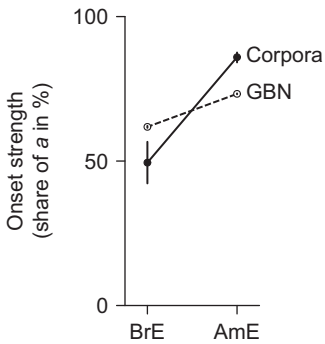


Figure 1.2 The estimated share of *a* for *historic* in the corpora and the GBN data set for BrE and AmE. ©<sup>1</sup>

Note: Error bars reflect 95% uncertainty intervals.

by an unbalanced representation of academic texts. In general, then, the percentages reported in Table 1.2 and Figure 1.2 may have failed to isolate contrasts between the varieties due to other systematic differences between the data sets, such as the constituent text categories.

We may also cast doubt on the validity of the uncertainty intervals. These are computed on the assumption that the occurrences of *a/an historic* in a given set of data are independent events. If several tokens stem from the same text and author, however, this assumption is unlikely to hold. Tokens are then said to be clustered by text. We would expect a writer, when faced with the sequence *a/an historic*, to be relatively consistent in the choice of *a* versus *an*. Consistency may also be due to editorial changes and/or spell check software. As a result, the error intervals reported so far may understate the uncertainty associated with these estimates.

As we illustrate in the next section, the metadata that comes with the BNC and COCA allow us to address these concerns and adjust percentages and uncertainty intervals accordingly. This will put our comparisons between the standard corpora on a firmer statistical footing and may caution us against interpreting the GBN data at face value.

### 1.5 Using Metadata to Adjust for Clustered Sampling and Genre Bias

Since our focus is on differences between the standard varieties, we need to demonstrate that the comparative figures we offer are robust to these justified concerns. We will first consider the issue of clustered observations and then address the sensitivity of our comparisons to genre differences.

### 1.5.1 *Adjustment for Clustered Sampling*

A typical feature of natural language data (and, by implication, corpus data) is their hierarchical structure (see Johnson 2014; Barth & Kapatsinski 2018; Speelman et al. 2018: 2–3; Winter 2020: 232–3; Winter & Grice 2021). By this we mean that observations are almost always clustered, or grouped. Commonly, for instance, a speaker (or author) contributes multiple data points to a study. Observations are then clustered by source, where ‘source’ refers to the producer of language. This is the case in our data, and in what follows, we will rely on the text IDs in the corpora as a proxy for the source of a linguistic event. Appropriate adjustment for clustered sampling requires that each text ID refer to a unique text. As we mentioned in Section 1.3, COCA offers a higher level of resolution and we primarily turn to this corpus to assess statistical consequences of clustering in the data.

Let us start by taking a closer look at the distribution of our observations across text files. For a given *h*-word, the ideal scenario would be for each text to feature only a single token. This would obviate the need for adjustments to our uncertainty intervals. If, however, there are authors who contribute multiple tokens for a given item, our analysis should take this into account.

Figure 1.3 provides a sketch of the distribution of the 150 types in COCA. The lexemes are ordered by frequency of co-occurrence with *a/an*: *high* ranks first, *hexadecimal*, *hydroxyapatite* and *heparin* last. The horizontal axis shows token counts per text, which range from 1 to 32. A square denotes that there is at least one text file in the corpus with the respective token count. Consider, for instance, *a/an historic*, which ranks 21st. The distribution of squares shows that there is at least one text in the corpus that contains 6 instances of this sequence. For *high*, at the top, there is at least one text contributing 30 tokens to our data. The exact number of texts represented by each square can be read from Web Appendix 3.<sup>12</sup> For our present purposes, we note that the COCA data do show clustering of observations at the text level. We should therefore examine whether adjustments to our uncertainty intervals are required.

To this end, we will juxtapose what we refer to as a naïve analysis, which ignores the structure in the data, and a(n) hierarchical analysis, which integrates the grouping structure into the estimation process. We use the label ‘hierarchical’ to describe an analysis that operates at the text level and therefore considers the texts as the primary sampling units. These units then constitute the relevant total ‘sample’ size and form the basis of our statistical inferences. This analysis strategy takes into account the intuitive fact that 100 observations from 100 different AmE authors would tell us more about this standard variety than 100 observations from a single author.<sup>13</sup> The quantity of interest will be

<sup>12</sup> <https://osf.io/hbgst/>

<sup>13</sup> There are different statistical procedures to account for clustered data structures. A method that is often used is mixed-effects regression modelling. For technical reasons, we use beta-

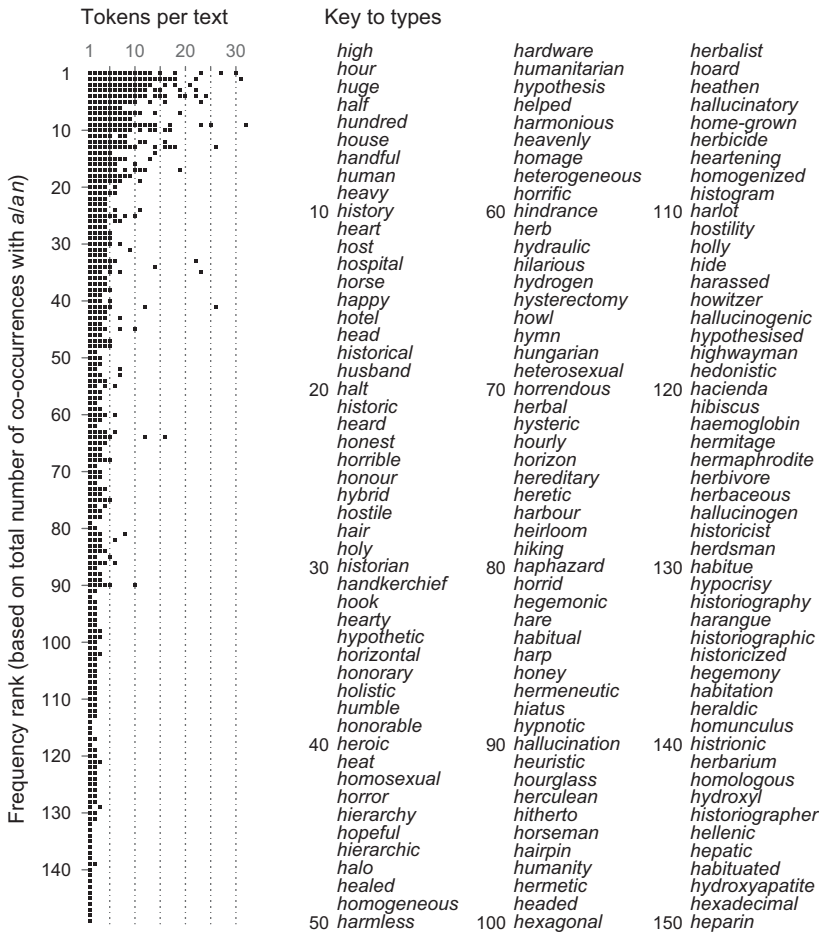


Figure 1.3 A sketch of the distribution of the 150 types in COCA. ©

Note: Squares denote observed token counts per text.

the lexeme-specific share of  $a$  (our indicator of [h]-strength) in COCA. We can then assess the degree of over-confidence in our estimates by comparing the width of the uncertainty intervals.

binomial regression (and, to double-check the results, an overdispersed binomial regression). For details, see the *R* script `sensitivity_analysis_clustering.Rmd` in the OSF repository (<https://osf.io/47p6u/>).

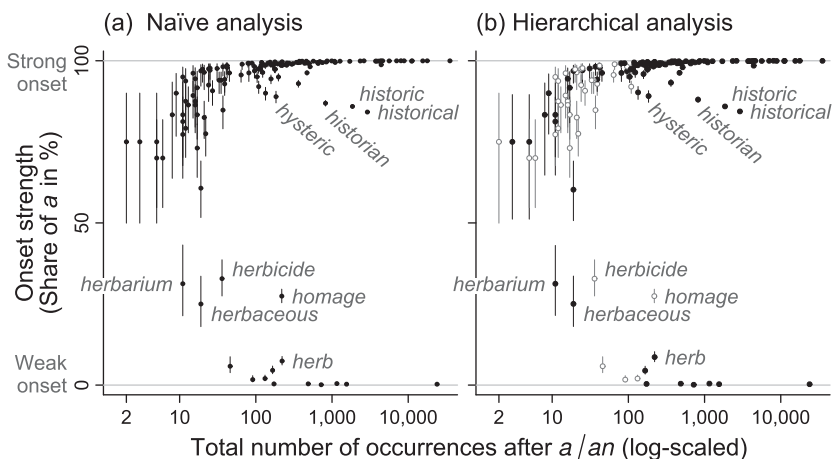


Figure 1.4 The share of *a* for each of the 150 types in the COCA data: (a) estimates based on a naïve analysis, ignoring the clustering by text file; (b) estimates from a hierarchical analysis. ©

Note: In panel (b), grey open circles denote *h*-words that never occur more than once per text file and therefore show no clustering; their estimates reduce to the naïve version. See Web Appendix 4 for a key to the lexemes.

Figure 1.4 shows, for the 150 items, the proportion of *a* we observed in COCA.<sup>14</sup> Panel (a) reports the naïve estimates and panel (b) those based on a hierarchical analysis. We have flagged a number of interesting items and refer to Web Appendix 4<sup>15</sup> for a full key to the lexemes. To avoid visual clutter, the error bars indicate (only) 50% uncertainty intervals.<sup>16</sup>

Figures 1.4a and 1.4b show virtually identical constellations. Accounting for the clustered data structure apparently yields no discernible changes in the estimates and their uncertainties. We followed up on this finding with more direct comparisons and computed differences between (i) the estimated percentages (i.e.

<sup>14</sup> For a BNC version of Figure 1.4, see Web Appendix 5: <https://osf.io/5n6j7/> (key to the lexemes: <https://osf.io/fsgjh/>; <https://osf.io/sbtkq/>).

<sup>15</sup> <https://osf.io/t39zu/> (Figure 1.4a); <https://osf.io/g9s3m/> (Figure 1.4b)

<sup>16</sup> For some *h*-words, our regression analysis confronts problematic data situations due to (i) (near-)categorical usage patterns and/or (ii) scant token counts. To obtain sensible results, we implemented what is known as Firth bias adjustment (Firth 1993; see also Greenland, Mansourina & Altman 2016). This involves adding, for our naïve analysis, a ‘count’ of ½ to each cell, and for our hierarchical analysis, an augmented ‘text’ with a ‘count’ of ½ each for *a* and *an*. While we would have preferred to sidestep data manipulations of this kind, this strategy allowed us to run parallel (i.e. naïve and hierarchical) analyses for all *h*-words. As a result, estimated percentages for low-frequency types are slightly deflected away from categoricity, especially for items with fewer than 10 tokens in total. For details, please refer to the R script available via the OSF project associated with this article (<https://osf.io/47p6u/>).

the share of *a*) and (ii) the widths of the uncertainty intervals. The distribution of these differences is supportive of the impression we get from [Figure 1.4](#).

For the BNC,<sup>17</sup> we likewise observe only minor differences between naïve and hierarchical analyses, and the same is true for comparisons between the corpora: estimates of the onset strength difference between BrE and AmE are largely immune to the analysis strategy. Overall, then, we conclude that, for the corpus data, accounting for the clustering of observations leads to negligible shifts in our statistical uncertainty assessments. We will therefore disregard this feature of our data as we turn to our next task, the appraisal of suspected biases due to genre differences.

### 1.5.2 Adjustment for Genre Differences

Given what we know from previous research about potential differences between genres (e.g. [Biber 1998: 135–71](#); [Biber & Gray 2013](#)), across-the-board comparisons between the BNC and COCA may not be purely reflective of differences between the standard varieties. [Table 1.3](#) shows that our corpora do not represent the same types of written language; further, shared text categories are weighted differently in terms of word count.

Our first step, therefore, is to streamline the corpora by narrowing down the BNC to a set of categories parallel to COCA. While this renders the selection of text types more comparable, the proportional share of these parts differs (cf. the percentages in bold face). By design, the four domains in COCA are balanced. The distribution in the BNC, on the other hand, is

Table 1.3 *Text categories in the BNC and COCA: word count and proportional share*

BNC (written)			BNC subset		COCA (written)		
Text category	Words	Share	Words	Share	Text category	Words	Share
Academic prose	17.8 m	18%	17.8 m	<b>24%</b>	Academic	111.3 m	<b>24%</b>
Fiction and verse	19.4 m	19%	19.4 m	<b>26%</b>	Fiction	119.1 m	<b>26%</b>
Newspapers	10.6 m	11%	10.6 m	<b>14%</b>	News	114.7 m	<b>25%</b>
Non-academic prose, biography	27.2 m	27%	27.2 m	<b>36%</b>	Magazine	112.7 m	<b>25%</b>
Other published written material	20.2 m	20%	-	-	-	-	-
Unpublished written material	5.0 m	5%	-	-	-	-	-

<sup>17</sup> We should note that, with many text IDs in the BNC not referring to unique texts, the underlying adjustments may not be fully effective.

uneven, with fewer words in News (14%) and a larger proportion in Prose (36%). Our adjustment needs to take account of the unequal weighting of the categories.

Our next task will be to sensitize our data summaries to the factor genre. To this end, we compute four estimates per corpus – one for each text category – and then take the simple average over the four percentages. Each category then receives the same weight, irrespective of the number of tokens observed. To illustrate the procedure, let us return to the item *historic*. Table 1.4 lists genre-specific token counts and percentages. Note that News accounts for the largest share of tokens (47% in COCA, 51% in the BNC) and therefore has the potential to disproportionately influence a naïve estimate. In addition, since the sequence *a/an historic* is distributed very unevenly across the text categories, the balanced design of COCA does not guarantee an even-handed treatment of the four text categories.

The percentages are shown graphically in Figure 1.5, where we observe differences between the text categories. In COCA, for instance, News and Prose appear to exhibit stronger *h*-onsets for *historic* (i.e. a larger share of *a*). Prima facie, the BNC percentages are suggestive of larger genre differences. However, as indicated by the error intervals, these estimates are less precise due to smaller token counts. Fiction, for instance, only offers four instances of *a/an historic*. The tilted lines in the middle of the display contrast the naïve across-the-board averages (grey) with adjusted estimates (black). For *historic*, our genre adjustment entails minor shifts in the estimated percentages.

We also note, however, that a new form of systematic error may arise. Thus, the imprecise BNC Fiction estimate for *historic* receives the same weight as the other percentages when computing an adjusted share. Distortions of this kind are sometimes referred to as sparse data bias (see Greenland, Mansourina &

Table 1.4 *Observed counts for a/an historic in the BNC and COCA, broken down by text category*

Text category	COCA				BNC			
	<i>a</i>	<i>N</i>	%	95% UI	<i>a</i>	<i>N</i>	%	95% UI
Academic	215	278	<b>77%</b>	[72%, 82%]	6	12	<b>50%</b>	[25%, 75%]
Fiction	67	86	<b>78%</b>	[68%, 85%]	1	4	<b>25%</b>	[5%, 70%]
News	641	730	<b>88%</b>	[85%, 90%]	42	93	<b>45%</b>	[35%, 55%]
Prose	440	492	<b>89%</b>	[86%, 92%]	41	73	<b>56%</b>	[45%, 67%]
Naïve estimate			<b>86%</b>	[84%, 88%]			<b>49%</b>	[42%, 57%]
Adjusted estimate			<b>84%</b>	[81%, 86%]			<b>44%</b>	[28%, 60%]

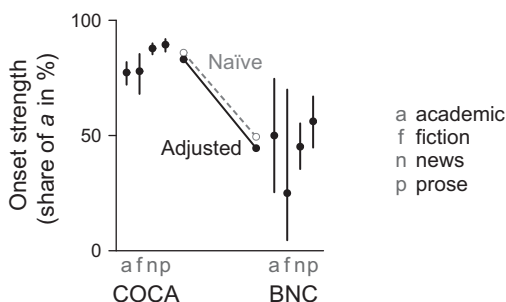


Figure 1.5 The estimated share of *a* for *historic* in COCA and the BNC, broken down by text category. ©<sup>1</sup>

Note: Comparison of naïve and adjusted estimates. Error bars indicate 95% uncertainty intervals.

Altman 2016). The new estimate of 44% might therefore be downwardly biased. We should thus keep an eye on the token counts for the text categories. Especially in the BNC, however, these level off rather quickly. In Web Appendix 6,<sup>18</sup> we provide cross-tabulations for all items.

Our sensitivity analysis will therefore have to be selective. As a first step, we will do spot checks on five further items with sufficient token counts and divergent shares in the BNC and COCA. The results are shown in Figure 1.6, where token counts are added in the lower part of the display. We observe no substantial changes in the estimated percentages, the only exception being *horrific*. Since this word occurs only twice in the Academic section of the BNC, however, our adjustment may be distorted due to data sparsity.

With small token counts in the BNC prohibiting sensible adjustments for the majority of our items, we ran further audits with the COCA data, comparing naïve and adjusted estimates for the 75 most frequent items. Most differences (82%) amounted to less than 1 percentage point, the maximum being just under 4 percentage points. For details, we refer to Web Appendix 7.<sup>19</sup>

Overall, then, our spot checks suggest that our naïve estimates do not change appreciably when factoring genre into the analysis. This alleviates our concerns about the comparability between the two corpora, and also suggests that the large discrepancies between these and the GBN figures are unlikely to result from differences in genre composition. We will return to this point in Section 1.7.

<sup>18</sup> <https://osf.io/rdk5/> <sup>19</sup> <https://osf.io/w8gse/>



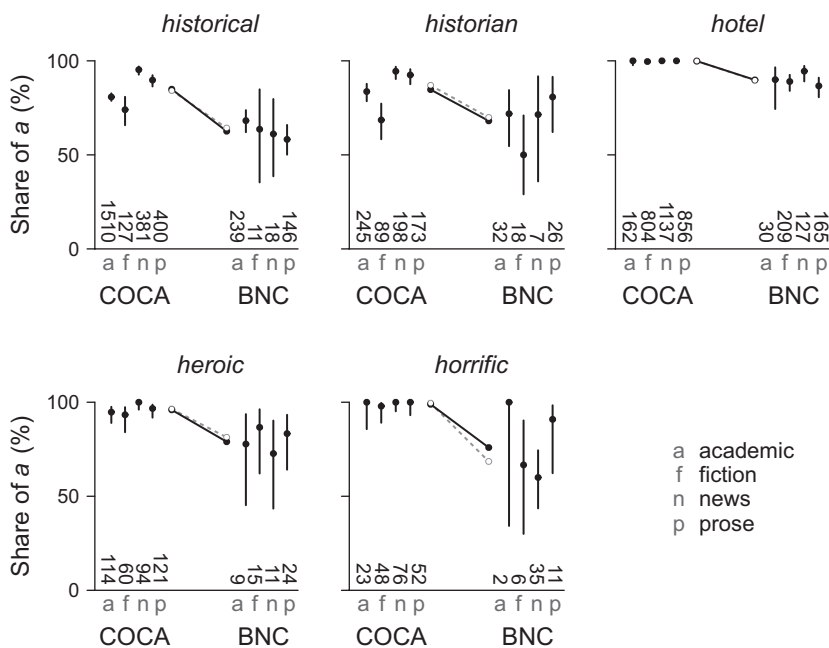


Figure 1.6 Spot checks on five further items: Estimated share of *a* in COCA and the BNC, broken down by text category. ©<sup>1</sup>  
*Note:* Comparison of naïve (dashed grey line) and adjusted estimates (solid black line). Error bars indicate 95% uncertainty intervals.

### 1.5.3 Sensitivity Analysis: Summary

The metadata offered by the BNC and COCA allowed us to appraise the robustness of our data summaries to two factors that, based on our background knowledge, sounded a note of caution. Our sensitivity checks suggest that our comparisons are largely immune to these potential disturbances. Before we put our concerns to rest, however, we should be reflecting on these findings, which certainly come as a surprise.

The fact that the clustering of tokens at the text level hardly affects our error intervals makes sense in the light of discourse-pragmatic factors. Thus, while entities are often introduced in the discourse with an indefinite article, once the referent is established, it is unlikely to be accompanied (again) by *a/an*. In other words, there are pragmatic constraints on the number of tokens per text file.

As for the minor variations across genres, we would argue that the choice of *a* or *an* is typically a subconscious one, occurring in on-line language production with *h*-initial words at the same level of automation as with other vowel-

consonant-initial words. While phonological *h*-dropping as such is heavily stigmatized as a non-standard feature of many British dialects, the written standard represented in our corpora and Google Books regularly retains initial ⟨*h*⟩. The selection of *a* or *an*, in contrast, may be below the radar of the writer's attention – prescriptivist or other – with the note by Peters (2004: 1) mentioned earlier being an exception.

We now turn to the second key difference between our data sets, their size, and discuss opportunities created by big data resources for the quantitative study of *h*-onsets.

## 1.6 Type and Token Frequency: Scope and Stability of Comparisons

Our comparison of the four data sets in Section 1.3 underscored the size advantage of the GBN data, which yielded five times the number of types (827 vs 150) and 700 times as many tokens (150 million vs 210,000). We will now look more closely at these two dimensions of frequency and illustrate how these differences affect the scope and stability of our data summaries. We will begin, however, with a side-by-side inspection of the onset strength estimates from the different data sources.

### 1.6.1 Comparison of Estimates from the Four Data Sets

Figure 1.7 brings together the full set of estimates gained from the standard corpora (left-hand panels) and GBN. In the right-hand panels, which display the GBN estimates, filled black circles denote types that also occur in the corpus data and are therefore also found in panels (a) and (c). As we have already seen in Figure 1.1, a large share of the additional items, which are here shown with grey empty circles, are rarely accompanied by *a/an*. Judging from Figure 1.7, it appears that GBN yields, for both BrE and AmE, a considerable number of interesting items (i.e. *h*-words that show intermediate levels of onset strength). For a key to the lexemes, please see Web Appendix 8.<sup>20</sup>

Overall, we observe that, due to the massive number of tokens for each lexeme, the error bars are vanishingly small for most GBN estimates. Only in the leftmost part of panels (b) and (d) do we begin to see visual indications of statistical uncertainty. This brings us to the first key aspect, the token frequency advantage of GBN.

<sup>20</sup> <https://osf.io/4jadf/> (panel a); <https://osf.io/zydpu/> (panel b); <https://osf.io/fbrcj/> (panel c); <https://osf.io/jxn27/> (panel d)

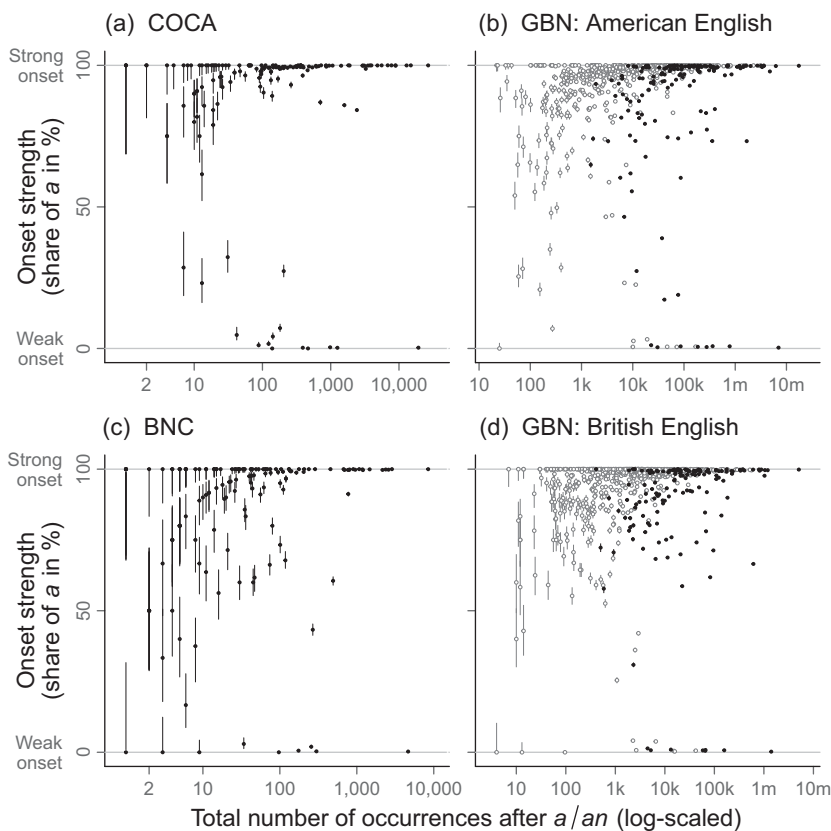


Figure 1.7 The proportion of *a* for different items in the corpora (150 types) and GBN (827 types). ©<sup>1</sup>  
*Note:* For the GBN data, filled circles denote items that also occur in the corpus data. Estimates are based on a naive analysis. Error bars indicate 50% uncertainty intervals. See Web Appendix 8 for a key to the lexemes.

1.6.2 *Token Frequency: Stability of Estimates*

To illustrate the effect of token counts on our data summaries, let us turn to a comparison of the two varieties. Since BrE is the variety with reportedly weaker [h]-onsets, we expect the share of *a* to be generally lower. Thus, if we subtract, for each lexeme, the AmE share of *a* from the BrE share, most differences should be negative, signalling a smaller proportion of *a* (or weaker [h]) in BrE.

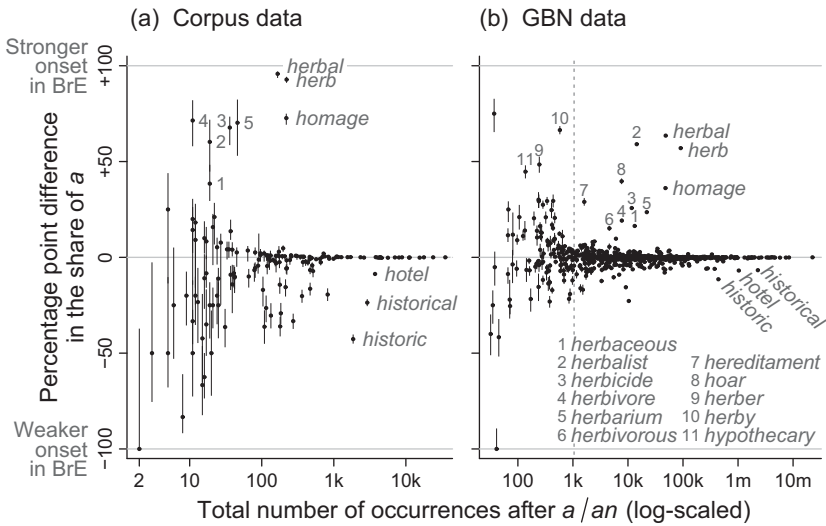


Figure 1.8 Percentage point difference in the share of *a* for each *h*-lexeme. ©

Note: Comparison of estimates from a (naïve) analysis of (a) the 150 items in the corpus data and (b) the 827 items in the GBN data. See Web Appendix 9 for a key to the lexemes.

Results are displayed in Figure 1.8a (corpora) and 1.8b (GBN), where points represent estimated differences and error bars show 50% uncertainty intervals. Points located below the horizontal line denote *h*-words with weaker onsets in BrE (the expectation, on average). The items are again arranged by frequency.

As in Figures 1.4 and 1.7, the width of the error bars decreases, overall, from left to right, reflecting the increased precision of larger samples. The expected trend for points to be located below the line marking zero appears to hold on balance, and there are various morphologically unrelated types to be found here (e.g. *hotel*, *historic(al)*, *hosteler*, *habitude*, *henrician*, *hypallage*, *homopterous*, *hypercriticism*, *hexangular*). Conversely, there are quite a few items above the line, pointing to stronger [h] in BrE, but limited to the items *homage* and *herb* and its derivatives, which inherit the [h]-‘strength’ of their root to various extents. A full key to the lexemes can be found in Web Appendix 9.<sup>21</sup>

A striking pattern in Figure 1.8 is the trumpet-like shape of the point cloud – difference estimates fan out towards the left end of the scale. Taken at face value, this seems to suggest that differences between the varieties are greater

<sup>21</sup> <https://osf.io/eaubm/> (panel a); <https://osf.io/krxzf/> (panel b)

for low-frequency items. It is, however, more likely that we are looking at a well-known statistical artefact. Estimates based on smaller samples are subject to greater sampling variation, i.e. another form of sparse data bias (see e.g. [Steel, Liermann & Guttorp 2019: 399](#)). What this means is that scores computed from only a handful of tokens will vary noticeably from sample to sample. For our analysis of differences between BrE and AmE, this means that we should not over-interpret the differences at the left end of the scale. We could, of course, decide to exclude types with few occurrences. However, this would require a fairly arbitrary cut-off and, more importantly, reduce the number of *h*-lexemes and narrow the scope of our investigation (see [Section 1.6.4](#)).

Panel (b) gives the estimates for the GBN data, with 50% uncertainty intervals based on a naïve analysis. Points also fan out somewhat at the far left end of the scale, though not as dramatically as in the left-hand panel. The much larger token numbers per type reduce random fluctuation of point estimates and therefore yield more stable indicators of [h]-strength. As a result, sampling variation is less of a concern when it comes to interpreting these scores. The dashed vertical line in panel (b) marks the lower limit of the corpus data on the frequency scale (cf. [Figure 1.1](#)). Thus, in relative terms, a count of two in the corpora corresponds to a count of about 1,000 in the GBN data (recall the size ratio of 1:500).

Before we turn to the type frequency advantage of GBN, we will follow up on a pattern that emerges from our comparison of difference estimates in [Figure 1.8](#).

### 1.6.3 *The Distinction of Standard Varieties in GBN*

Recall that in [Figure 1.2](#), we observed that the onset strength difference between BrE and AmE for *a/an historic* was attenuated in the GBN data. This also shows in [Figure 1.8](#), where the GBN estimate for *historic* diverges less from the horizontal line. If we take a closer look, we note that difference estimates appear to be generally smaller in the GBN data. In other words, the two varieties seem to be more similar.

In [Section 1.4](#), we reasoned that this might be due to genre differences, but our spot checks in the corpus data suggest that this is an unlikely explanation. However, it has been pointed out elsewhere that the GBN data lack tidiness, representativity and metadata, which hampers linguistically valid conclusions ([Pechenick, Danforth & Dodds 2015](#); [Hiltunen, McVeigh & Säily 2017](#); [Koplenig 2017](#)). By the same token, we must assume that ascriptions of *n*-grams to a British or a US source may be unreliable. In other words, the GBN data may to a certain extent offer a blend of the two varieties. If this were the case, GBN estimates would be biased towards the respective other variety.

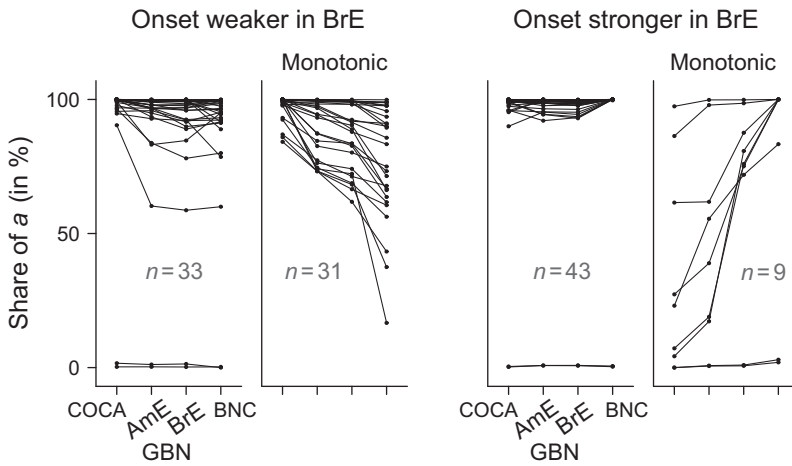


Figure 1.9 Comparison of corpus and GBN estimates for a subset of 116 items. 

It follows that if we line up our estimates, we would expect a monotonic cline (i.e. COCA > GBN AmE > GBN BrE > BNC, or vice versa). The items that occur in both data sets allow us to assess whether we encounter this pattern in our data. To avoid clutter due to imprecise estimates, we exclude items that occur fewer than five times in the BNC ( $n = 29$  types). [Figure 1.9](#) groups the remaining lexemes as follows: The left-hand panels show types that feature a weaker onset in BrE (i.e. appearing below the horizontal line in [Figure 1.8](#)), the right-hand panels those with a stronger onset in BrE. Each subset is further broken down into items that do not show a monotonic cline between the four data sets and those that do, that is, whose GBN estimates are intermediate between the corpus figures (i.e. COCA < GBN AmE < GBN BrE < BNC for panels on the left; COCA > GBN AmE > GBN BrE > BNC for panels on the right).

We find that this is the case for a third of the items (40 out of 116), a proportion that is higher than expected by chance (i.e.  $1/(4!/2) = 8\%$ ). The majority of items not yielding a steady cline shows (near-)categorical shares, however, with GBN figures likely to be distorted by noise in the data (cf. our discussion of *hand* in [Section 1.3.3](#)). We therefore interpret [Figure 1.9](#) as supporting the view that the GBN data blur the distinction between the varieties: Rates we obtain for individual items as well as for the group as a whole tend to be attracted towards the cross-varietal average. We would suggest that this effect may be tied to the quality of the metatextual information and that the

assignment of Google Books to Britain and America is an unreliable indicator for authors' linguistic allegiances.<sup>22</sup>

We must therefore conclude that the GBN data offer partially neutralized figures for the key comparison we wish to draw (viz. pronunciation differences between BrE and AmE), as the percentages are systematically distorted and therefore differentiate less clearly between the standard varieties.

#### 1.6.4 *Type Frequency: Scope of Analysis*

To understand gradience in onset strength, we have directed attention to etymological and phonological features of the *h*-lexemes (Schlüter 2019; Schlüter & Vetter 2020). While we do not pursue this line of investigation further in this chapter, let us nevertheless take a look at relevant cross-classifications of items in the corpus and GBN data. This is of interest for the present discussion, as it highlights how the size advantage of GBN can be brought to bear on questions of linguistic interest. To this end, we group items according to (i) etymology (Germanic vs Romance) and (ii) the prominence of the first syllable, distinguishing between 'primary stress', 'secondary stress' and 'unstressed'.<sup>23</sup>

Table 1.5 shows how these sub-groups are represented in the two data sets in terms of type and token counts. In the Germanic group, the distribution is very uneven across the phonological conditions. Germanic *h*-types almost categorically carry primary stress on the initial syllable. Scaling up the number of types in the GBN data does not change this distributional pattern. However, GBN offers at least a handful of items in the underrepresented cells<sup>24</sup> and thereby allows us to compare stress levels among Germanic types. Among Romance words, the distribution of lexemes across stress levels is much more even. Particularly the GBN data strike a balance between the three phonological conditions. Looking at Table 1.5, we recognize the type frequency advantage of GBN as an attractive feature: It permits systematic investigation of sub-groups with few representatives in relevant contexts (i.e. following *a/an*).

<sup>22</sup> Thus, a reliance on places of publication as a substitute for author nationality could have led to erroneous assignments since books may have been written by authors on the other side of the Atlantic. Many publishers nowadays have various legal business locations in both countries. Moreover, the Google Books project draws heavily on academic libraries, and increasing shares of academic books are written by non-native speakers of English (especially if we consider data from 1975 and after).

<sup>23</sup> Types of uncertain origin or with variable stress patterns are excluded here. Type counts therefore do not add up to 150 and 827, respectively.

<sup>24</sup> For the category with secondary stress, these are *halfhour*, *halfmoon*, *hamiltonian*, *haphazard*, *hereafter*, *heretofore*, *hobgoblin* and *hudibrastic*; for the category with an unstressed initial syllable, these are *hadronic*, *haversian*, *henrician*, *hogarthian*, *hurrah* and *huzza*.

Table 1.5 *Etymological and phonological sub-groups in the corpus and GBN data: type and token frequencies*

Sub-group	Corpora		GBN	
	Types	Tokens	Types	Tokens
Germanic				
Primary stress	45	121,321	295	99,479,883
Secondary stress	1	154	8	159,134
Unstressed	0	0	6	8,990
Romance				
Primary stress	34	75,251	142	36,670,873
Secondary stress	17	7,290	125	5,378,186
Unstressed	50	10,554	161	7,193,897

## 1.7 Discussion

Let us look back and reflect on the strengths and limitations of the data sources we have compared, and the insights that emerged from a conjoint analysis. Table 1.6 summarizes key contrasts.

The size advantage of GBN featured in two ways. First, we were able to collect more tokens for each lexeme. From a statistical perspective, the amplification of token counts yields more stable estimates. Thus, for items producing 30 or fewer tokens in the corpora, augmenting the sample size is clearly beneficial: Taking another look at Figure 1.8a, we see that it is below this mark (which, on the log scale, is half-way between 10 and 100) that difference estimates begin to form a trumpet bell. When confronted with this graph, we might run the danger of trying to attach a linguistic, perhaps frequency-related, interpretation to a statistical artefact. With ‘small data’, we must constantly be on the lookout for deceptive patterns of this kind. Comparing the right and left panels of Figure 1.8, we see that the GBN estimates remain more stable, which safeguards against erroneous interpretations.

The second frequency boost concerns the number of types, both in total as well as for sub-classifications, or conditions, of linguistic interest. Thus, if the research focus rests on systematic differences between certain groups of lexemes, increasing the number of specific types instantiating these conditions yields two advantages. On the one hand, certain cross-classifications may occur rarely or not at all in small data sets, which precludes their linguistic study or increases uncertainty estimates out of proportion. Table 1.5 revealed that, in the domain of Germanic lexemes, the scope of our corpus data is in effect limited to types carrying stress on the initial syllable. This makes it difficult to disentangle the relative contribution of phonological and etymological features to the behaviour of this group, which, by contrast, is possible in the GBN data. Second, in a similar fashion to the discussion in the previous paragraph, a higher number of types per condition produces more



Table 1.6 *Comparative overview of characteristics of the GBN database and the corpora (BNC and COCA)*

Criterion	GBN	Corpus data
Token counts for lexical types	+ High token counts yield stable estimates	- Token counts below 20 or 30 become subject to considerable sampling variation, risking over-interpretation of estimates
Type counts for conditions of interest	+ Sparsely populated conditions can be studied + Higher type counts yield more stable estimates for a condition	- Rare conditions may be sparsely represented or absent; scope of analysis limited - Estimates for sparsely populated conditions may be unstable
Data quality	- OCR errors handicap variable-slot queries of the <i>n</i> -gram lists - Metatextual information (country and year of publication) unreliable	+ Error rate much lower + Metatextual information more ample and generally reliable
Data handling	- Reliance on automatized string processing and pragmatism	+ Reliance on linguistically motivated inclusion/exclusion criteria
Accessibility of linguistic context	- Context-sensitive screening/disambiguation not possible - Data contaminated by irrelevant tokens	+ Context-sensitive screening/disambiguation possible + Validity of data points can be verified
Information on the source of data points	- Source of tokens unknown - Language-external clustering by source cannot be factored into the analysis - Uncertainty intervals may be too narrow - Bias in point estimates cannot be corrected	+ Tokens can be linked to text files + Clustering by source can be represented - Level of individual author is unavailable for some text files (in the BNC) + Uncertainty intervals more accurate + Disproportional-representation bias in point estimates can be controlled

stable estimates. In Table 1.5, this benefit surfaces in the sub-groups of Romance lexemes: The number of types with *h*-onsets in syllables carrying secondary stress, for instance, goes up from 17 in the corpora to 125 in GBN. In general, then, a boost in type frequency may allow us to broaden the linguistic scope of our analysis and may stabilize estimates for sparsely populated conditions.

Turning to problems with the GBN data, let us first recapitulate our data retrieval and processing strategies and issues of data quality. To obtain relevant tokens from the *n*-gram files, we had to rely heavily on automatic character string processing and pragmatism in data selection. Most problems arose because we

issued a wildcard-type query, extracting all character strings starting with <h>. This variable-slot search of the 2-gram files confronted us squarely with the messiness of these lists and the scale at which OCR errors and non-English text passages materialize in the extracted elements.<sup>25</sup> To get a grip on the vast pool of *h*-initial strings, we had to resort to practicable, rather than clearly defined linguistic criteria to weed out non-lexical material. To avoid a manual screening of close to 16,000 character strings, we implemented pragmatic exclusion criteria and quantitative criteria connected to the outcome. Thus, we excluded types that occurred categorically with the same variant in both varieties. While we were able to motivate this choice given our interest in BrE-AmE contrasts, this data selection strategy will disturb analyses if differences between groups of lexemes are of interest. This is because the excluded types also offer relevant information on lexical sub-groups, whose behaviour would be misrepresented – in our case, biased away from zero differences. In the causal inference literature this is referred to as a form of selection bias (see [Elwert & Winship 2014](#)).

At the data screening stage, we recognized the value of being able to inspect the context of occurrence in the corpus data. Doubtful cases, such as *hew*, *hi*, *hid*, and *heigh*, could be disambiguated manually, which resulted in their exclusion. With the GBN data, this is not possible due to the lack of contextual information. However, we should keep in mind that even if we were able to access the context, manual screening would not be feasible due to the massive token counts. We would need to resort to down-sampling strategies, and big data would become small data.

We further illustrated the added value of the BNC and COCA metadata, which allowed us to address reasonable concerns about the validity of our quantitative comparisons. In our case study, we focussed on two aspects that may lead us astray in our conclusions: genre differences and the clustering of data points at the text level. The corpus metadata allowed us to appraise the amount of systematic error that may be due to these factors, and, if necessary, adjust for underlying disturbances. For the structure under investigation here, the suspected biases did not materialize. Our robustness checks on the COCA and BNC data only produced negligible shifts in our data summaries, and, importantly, these findings appeared to make sense linguistically. Whether we can extrapolate these insights in a direct manner to the GBN data is open to debate and to a certain extent unverifiable. Nevertheless, we have illustrated how sensitivity analyses that tap into the corpus metadata allowed us to form some judgement about our GBN figures. In general, a close exchange between different data sources may help us assess (anticipated) objections to the reliability and validity of big data resources. While we were able to demonstrate the stability of our conclusions to genre effects and the clustered structure of

<sup>25</sup> See Web Appendix 1 (<https://osf.io/n5gxh/>) for an illustration.

our data, we would generally recommend safeguarding against these (and other) potential disturbances – it appears overly optimistic to expect sensitivity checks to be this comforting for other linguistic structures.

At first sight, the GBN data seemed to hold great promise for the study of BrE–AmE contrasts since the distinction between British and American books is one of the few metatextual categories that are consistently available for each data point. However, in the context of the present study, we have observed that descriptive GBN measures differ systematically from corpus-based figures: The share of *a* was, on average, lower in the BNC and higher in COCA compared to GBN data for the same variety. As a result, the difference between the varieties was downwardly biased in the GBN data. While it appears plausible that a mistaken attribution of books to varieties is responsible for these disturbances, we are unable to verify, let alone correct for, this form of data contamination given the information in the *n*-gram database.

Reservations apply not only to this binary assignment, but also to the origin of book authors more generally: The increasing use of English as an international lingua franca has led to a situation where non-native users by far outnumber native users, which has to be taken into consideration when differences between national varieties come into focus. The more recent data in the Google Books archive, being supplied by academic libraries, may be more representative of a worldwide levelling of academic English than of national standards.

Another attraction of the GBN database that has been backgrounded for present purposes, but might inspire research on language change, is its maximal diachronic resolution: Every data point comes with a unique publication year, and the entire database covers five centuries. However, the accuracy of OCR processing can be expected to decrease with increasing time depth, and even the set of bigrams used for our analysis of Present-Day English was flawed by OCR errors. What is worse, unsystematic spot checks have revealed that the GBN database contains numerous reprints and re-editions of works whose originals date back several decades or even centuries and represent older states of the language. Analysts will thus have to reckon with a tilt towards conservative usage.

In general, then, our comparison has highlighted strengths and limitations of both data sources. Certainly, the concerns we have raised about GBN data are more fundamental since they touch on the issue of data quality. If questions of linguistic interest are to receive careful study, we would therefore advise against (solely) relying on GBN for empirical evidence. Nevertheless, we hope to have illustrated that insights gained from richly annotated corpora can be leveraged to carefully navigate our research efforts in the era of big data.

## Further Reading

- Cochran, William G. 1983. *Planning and Analysis of Observational Studies*. New York: John Wiley & Sons. [Chapters 1, 2, 3, 4, 6](#).
- Greenland, Sander, Mohammad Ali Mansourina and Douglas G. Altman. 2016. Sparse Data Bias: A Problem Hiding in Plain Sight. *British Medical Journal* 352(i1982). <https://doi.org/10.1136/bmj.i1981>.
- Johnson, Daniel E. 2014. Progress in Regression: Why Natural Language Data Calls for Mixed-Effects Models. Unpublished manuscript. [www.danielezrajohanson.com/johnson\\_2014b.pdf](http://www.danielezrajohanson.com/johnson_2014b.pdf).
- Winter, Bodo. 2020. *Statistics for Linguistics*. New York: Routledge. Chapters 14 and 15.

## References

- Algeo, John. 2006. *British or American English? A Handbook of Word and Grammar Patterns*. Cambridge: Cambridge University Press.
- Barth, Danielle, and Vsevolod Kapatsinski. 2018. Evaluating Logistic Mixed-Effects Models of Corpus-Linguistic Data in Light of Lexical Diffusion. In Dirk Speelman, Kris Heylen and Dirk Geeraerts, eds. *Mixed-Effects Regression Models in Linguistics*. New York: Springer. 99–116.
- Biber, Douglas. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Biber, Douglas, and Bethany Gray. 2013. Being Specific about Historical Change: The Influence of Sub-Register. *Journal of English Linguistics* 41(2). 104–34. <http://eng.sagepub.com/cgi/doi/10.1177/0075424212472509>.
- Burnard, Lou, ed. 2007. Reference Guide for the British National Corpus (XML edition). British National Corpus Consortium & Research Technologies Service at Oxford University Computing Services. [www.natcorp.ox.ac.uk/docs/URG/BNCdes.html](http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html).
- Cruttenden, Alan. 2014. *Gimson's Pronunciation of English*. 8th ed. London: Arnold.
- Davies, Mark. 2008–. The Corpus of Contemporary American English (COCA): 600 Million Words, 1990–Present. [www.english-corpora.org/coca](http://www.english-corpora.org/coca).
- 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>.
- Elwert, Felix, and Christopher Winship. 2014. Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. *Annual Review of Sociology* 40. 31–53. <https://doi.org/10.1146/annurev-soc-071913-043455>.
- Firth, David. 1993. Bias Reduction of Maximum Likelihood Estimates. *Biometrika* 80 (1). 27–38. <https://doi.org/10.2307/2336755>.
- Gelman, Andrew, and Sander Greenland. 2019. Are Confidence Intervals Better Termed ‘Uncertainty Intervals’? *British Medical Journal* 366(15381). <https://doi.org/10.1136/bmj.l5381>.
- Greenland, Sander, Mohammad Ali Mansourina and Douglas G. Altman. 2016. Sparse Data Bias: A Problem Hiding in Plain Sight. *British Medical Journal* 352 (i1982). <https://doi.org/10.1136/bmj.i1981>.
- Hiltunen, Turo, Joe McVeigh and Tanja Säily. 2017. How to Turn Linguistic Data into Evidence? In Turo Hiltunen, Joe McVeigh and Tanja Säily, eds. *Big and Rich Data in*

- English Corpus Linguistics: Methods and Explorations*. Studies in Variation, Contacts and Change in English 19. [www.helsinki.fi/varieng/series/volumes/19/introduction.html](http://www.helsinki.fi/varieng/series/volumes/19/introduction.html).
- Johnson, Daniel E. 2014. Progress in Regression: Why Natural Language Data Calls for Mixed-Effects Models. Unpublished manuscript. [www.danielezrajohnson.com/johnson\\_2014b.pdf](http://www.danielezrajohnson.com/johnson_2014b.pdf).
- Jones, Daniel. 2011. *English Pronouncing Dictionary (EPD)*. Edited by Peter Roach, Jane Setter and John Esling. 18th ed. Cambridge: Cambridge University Press. CD-ROM edition.
- Koplenig, Alexander. 2017. The Impact of Lacking Metadata for the Measurement of Cultural and Linguistic Change Using the Google Ngram Data Sets: Reconstructing the Composition of the German Corpus in Times of WWII. *Digital Scholarship in the Humanities* 21(1). 169–88. <https://doi.org/10.1093/lc/fqv037>.
- Lass, Roger, and Margaret Laing. 2010. In Celebration of Early Middle English ‘H’. *Neuphilologische Mitteilungen* 111(3). 345–54.
- Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden et al. 2010. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* 331(6014). 176–82. <https://doi.org/10.1126/science.1199644>.
- Minkova, Donka, ed. 2009. *Phonological Weakness in English: From Old to Present-Day English*. Basingstoke and New York: Palgrave Macmillan.
- Minkova, Donka. 2014. *A Historical Phonology of English*. Edinburgh: Edinburgh University Press.
- OED (Oxford English Dictionary Online)*. 2000–. Oxford: Oxford University Press. <http://dictionary.oed.com/> (accessed 3 March 2020).
- Pechenick, Eitan, Christopher M. Danforth and Peter Sheridan Dodds. 2015. Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution. *PLoS ONE* 10(10), e0137041. <https://doi.org/10.1371/journal.pone.0137041>.
- Peters, Pam. 2004. *The Cambridge Guide to English Usage*. Cambridge: Cambridge University Press.
- R Core Team. 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. [www.R-project.org/](http://www.R-project.org/).
- Scherer, Ralph. 2018. PropCIs: Various Confidence Interval Methods for Proportions. R package version 0.3–0. <https://CRAN.R-project.org/package=PropCIs>.
- Schlüter, Julia. 2019. Tracing the (Re-)Emergence of /h/ and /j/ through 350 Years of Books: Mergers and Merger Reversals at the Interface of Phonetics and Phonology. *Folia Linguistica* 40(s1). Special issue on diachronic phonotactics. Edited by Nikolaus Ritt, Andreas Baumann and Christina Prömer. 177–202. <https://doi.org/10.1515/flih-2019-0009>.
- Schlüter, Julia, and Fabian Vetter. 2020. An Interactive Visualization of Google Books Ngrams with R and Shiny: Exploring a(n) Historical Increase in Onset Strength in a(n) Huge Database. *Journal of Data Mining and Digital Humanities* 21. Special issue on visualizations in historical linguistics. Edited by Benjamin Molineaux, Bettelou Los and Martti Mäkinen. <https://jdmhd.episciences.org/7000>.
- Speelman, Dirk, Kris Heylen and Dirk Geeraerts, eds. 2018. *Mixed-Effects Regression Models in Linguistics*. New York: Springer.

- Steel, E. Ashley, Martin Liermann and Peter Gutterop. 2019. Beyond Calculations: A Course in Statistical Thinking. *The American Statistician* 73. 392–401. <https://doi.org/10.1080/00031305.2018.1505657>.
- Wells, John. 2008. *Longman Pronunciation Dictionary (LPD)*. 3rd ed. Harlow: Pearson Longman. CD-ROM edition: *Longman Pronunciation Coach*.
- Winter, Bodo. 2020. *Statistics for Linguistics*. New York: Routledge.
- Winter, Bodo, and Martine Grice. 2021. Independence and Generalizability in Linguistics. *Linguistics* 59(5). 1251–77.