# An interactive visualization of Google Books Ngrams with R and Shiny

Julia Schlüter, Fabian Vetter

Journal of Data Mining and Digital Humanities

# An interactive visualization of Google Books Ngrams with *R* and *Shiny*:

# Exploring a(n) historical increase in onset strength in a(n) huge database

**Julia Schlüter, Fabian Vetter***

University of Bamberg, Germany

*Corresponding author: Fabian Vetter fabian.vetter@uni-bamberg.de

**Abstract**
Using the re-emergence of the /h/ onset from Early Modern to Present-Day English as a case study, we illustrate the making and the functions of a purpose-built web application named **(*an*:*a*)-lyzer** for the interactive visualization of the raw n-gram data provided by Google Books Ngrams (GBN). The database has been compiled from the full text of over 4.5 million books in English, totalling over 468 billion words and covering roughly five centuries. We focus on bigrams consisting of words beginning with graphic <h> preceded by the indefinite article allomorphs *a* and *an*, which serve as a diagnostic of the consonantal strength of the initial /h/. The sheer size of this database affords us the possibility to attain a maximal diachronic resolution, to distinguish highly specific groups of <h>-initial lexical items, and even to trace the diffusion of the observed changes across individual lexical units. The functions programmed into the app enable us to explore the data interactively by filtering, selecting and viewing them according to various parameters that were manually annotated into the data frame. We also discuss limitations of the database, of the app and of the explorative data analysis. The app is publicly accessible online at https://osf.io/ht8se/.

**Keywords**
Data visualization, corpus linguistics, historical phonology, historical linguistics, R, Shiny, n-grams, Google Books, Google Books Ngrams

## INTRODUCTION

With the release of databases such as the Corpus of Global Web-based English (GloWbE; 1.9 billion words; https://www.english-corpora.org/glowbe/), the News on the Web Corpus (NOW; 8.7 billion words and growing; https://www.english-corpora.org/now/; accessed 26 February 2020) and the Intelligent Web-based Corpus (iWeb; 14 billion words; https://www.english-corpora.org/iweb/), linguists have witnessed the latest peak in a trend towards ever larger databases becoming available free of charge for linguistic investigation.[1] Even in the area of historical English linguistics, which by its very nature has to rely on texts that happen to have survived from former centuries and that are for that reason typically more restricted in scope and quantity, huge databases have been accumulated. Early English Books Online (EEBO; https://eebo.chadwyck.com/home or https://www.english-corpora.org/eebo/), for example, contains scans of several tens of thousands of books printed in the English-speaking

[1] Our thanks go to two anonymous reviewers for their pertinent and constructive comments on an earlier version of this contribution.

world plus works in English printed elsewhere from 1473 to 1700 (https://eebo.chadwyck.com/about/about.htm).[2] Taking word counts to an unprecedented dimension (see the visualization in [Jenset and McGillivray, 2017: 75]), Google has digitized over 15 million books stocked in university libraries around the world, over 8 million of which have been included into the Google Books Ngrams database (https://books.google.com/ngrams). Google Books Ngrams (henceforth GBN) represent the text of about 6 % of all books ever published from the 1500s to 2008, and over 50 % of this text is in English. The size of the English data runs to around 468 billion words (cf. [Lin et al., 2012: 169–170]; see also [Michel et al., 2010] for details on the data sampling process). While this shift away from small and carefully compiled corpora to large (and often messy) databases has been discussed critically (e.g. [Mair, 2006]; [Mair, 2013]), one of the major benefits of such massive datasets is certainly that they provide us with new empirical perspectives that were denied us before.

On the negative side, besides issues of lacking metatextual information, what sets these large datasets apart from smaller corpora from a technical point of view is that they usually come with their own specialized software, e.g. the English Corpora interface (https://www.english-corpora.org/) maintained by Mark Davies, or the Google Books Ngram Viewer (https://books.google.com/ngrams). This becomes problematic when the interface does not provide the function(s) necessary for a specific methodological approach. Where small corpora can easily be accessed with a range of established programs that presuppose little technical knowledge (e.g. *AntConc*, *Wordsmith Tools*) or even can be screened manually, large databases require either the use of technically more demanding software (e.g. *The IMS Open Corpus Workbench* + *CQPweb*) or scripting languages (e.g. *Python* or *R*). Even though scripting languages are the preferable option as they do not limit the researcher to a predefined set of functions, they are usually not part of linguistic training. While we do not wish to engage in the discussion of whether programming languages should be part of a linguistics curriculum and all that this would entail (see [Gries, 2011: 92–94] on that matter), it is true enough that programming lets us pursue avenues of research that ready-made tools often preclude. However, writing one's own scripts requires a non-negligible amount of training and, if used for the exploration of highly complex datasets, it is certainly mentally more taxing than using a ready-made interface.

A viable solution to this latter problem would be using programming languages to write an interactive interface for the specific task at hand, which then facilitates interaction with a complex dataset. This way, it is possible to combine the power and flexibility of programming languages with the ease of use and agility of ready-made interfaces. An interactive visualization specifically designed for a given task would allow the researcher to quickly and intuitively explore the data and follow up on emerging patterns. Research in the field of information visualization has convincingly shown that visualizations that the analyst can query interactively promote scientific inquiry, in that they facilitate the extraction of data, the testing and refinement of hypotheses and the sharing of insights (e.g. [Pike et al., 2009]). It is precisely for this reason that we opted for programming a web-based, interactive data visualization app, nicknamed **(*an*:*a*)-lyzer**, in our exploration of the intricate evolution of the strength of certain word-initial onset consonants in English. The application and its source code as well as an appendix table and high-resolution versions of the figures generated for this contribution are accessible via the Open Science Framework (OSF) at https://osf.io/ht8se/.

[2] Out of the total of around 132,600 titles, as of the time of writing, 25,363 titles are publicly accessible as full-text transcriptions.

The remainder of the article will proceed as follows. In Section I, we will sketch the issue of increasing consonantal strength of initial <h> and the theoretical significance of this phenomenon. The focus will however be on the methodological approach to the data through the configuration and functionality of the app, explaining the options that were programmed into it with their relevance for the phenomenon under study. Thus, the first part of Section II introduces the GBN data, while the second part details the pre-processing and linguistic coding we performed on the data. Section III describes the general design and technical aspects of the app itself and portrays its features and options. The concluding section provides a discussion and critical assessment of the affordances and remaining weaknesses of the interactive visualization.

## I THE CASE STUDY

[Schlüter, 2009] demonstrates how an interactive visualization of big databases such as GBN can be exploited in novel ways to advance insights in historical phonology. The article presents a use case where the specifically designed web app described in Section III allowed tracking the development of the initial sounds of three sets of lexical items (words beginning with graphemic <u>, <eu> and <h>) in English. Schlüter finds a diachronic pattern that could simplistically be described as moving from vowel to consonant. The diagnostic that is used to this end is the allomorphy that has characterized the indefinite article since late Old English times: Before a word beginning with a vowel, the article appears in its full form *an*, while before a word beginning with a consonant, it appears in its reduced form *a*. Previous research exploiting the same diagnostic based on collections of literary texts has shown that before <u>-, <eu> and <h>-initial words, the proportion of the full form has generally been decreasing over the past five centuries [Schlüter, 2006]. This has been interpreted to reflect the gradual emergence of the /j/-onset in words like *unit*, *use*, *eulogy* and *euphemism*, and the even slower re-emergence of the /h/-onset in words like *habit*, *historical* and *hysteric*. These had a mute initial <h> in the languages from which they were borrowed, but also in native Germanic words like *hand*, *healthy* and *hundred*, which appear to have been /h/-less during most of the Middle English period. To illustrate the functionality of the **(*an*:*a*)-lyzer** app, we will concentrate on <h>-initial words (while the full dataset remains accessible at the web location).

In the analysis, it has been argued that the strength of initial /h/ and the rate at which it re-appears depend on various intersecting etymological and phonological factors and that the choice of *an* and *a* provides a plausible measure of the variable strength of the onsets (for detailed discussion of the nature of the transition, see [Schlüter, 2019]). Crucially, the granularity of the resulting picture (both in terms of categories and subcategories that we can distinguish and in terms of temporal resolution) hinges on the amount of data that feeds into it, and this is where the mass of data available from GBN comes in handy. The availability of the **(*an*:*a*)-lyzer** app does not only enable a speedy and interactive management of the big dataset, but also a user-friendly and flexible way of grouping, sub-grouping and re-grouping lexemes based on the phonological and etymological factors annotated into the data frame. Details of the factors involved can be gleaned from Section 2.2. What is more, any potential effects that are displayed can be underpinned with exact absolute and relative figures and traced back to the individual lexemes contributing to each data point. Some exemplary visualizations will be included in Section 3.3 along with the technical demonstrations of the app's functions. On the theoretical side, the in-depth analysis sheds new light on longstanding issues such as asymmetries between phonetic detail and phonological categories in an exemplar-based, functional approach and the logical problem of the unmerging of purported phonemic mergers.

Though it seems a long way from big and messy collections of centuries-old prints (automatically converted into machine-readable text by modern software) to phonetic distinctions too minute for listeners to perceive (near-mergers), Schlüter shows that the large amount of data allows for an analytical precision for historical data that is comparable to the acoustic measurements of near-mergers obtained in modern phonetics labs (despite being of a completely different nature). Since major theoretically relevant insights have been discussed in [Schlüter, 2019], we will refrain from reiterating these.

## II DATA

### 2.1. Raw n-gram data

The English GBN data (version 20120701; [Lin et al., 2012]) were extracted from around 468 billion words of text. The data are available for download as 1-grams, 2-grams, 3-grams, 4-grams and 5-grams (http://storage.googleapis.com/books/ngrams/books/datasetsv2.html). Only n-grams that appear over 40 times in total made it into the database. There are separate files for British and American books as well as a version containing all English books. Due to the amount of data and the resulting file sizes, the data were split by Google into individual files based on the two initial letters of the n-grams. As our analysis draws on the allomorphs of the indefinite article preceding <u>-, <eu>- and <h>-initial words as a diagnostic, only the files containing bigrams starting with *a* + space and with *an* as their first elements were downloaded (i.e. filenames "a_" and "an"). The files consist of separate lines for each n-gram and year and specify the number of individual tokens of the n-gram and the number of volumes in which the tokens occur for that publication year. In subsequent steps, the bigrams relevant for the case study had to be extracted from these lists. In total, our dataset for <h>-initial lexemes contains the impressive number of 233 million bigram tokens; however, the metadata that are available are limited to the year and country for each token.

The methodology thus inherits the limitations that come with a big but messy dataset (cf. [Hiltunen et al., 2017], Section 4, and sources quoted therein), the most relevant ones for the present analysis being: lack of representativeness (due to oversampling of books stocked in scientific libraries), lack of bibliographical metadata (with the exception of publication year and country), inclusion of duplicates (multiple copies, re-editions and reprints of the same work, often published several decades or even centuries after the original publication date) and a seeming misalignment between country of publication and variety used by the author (cf. [Sönning and Schlüter, to appear]). What is more, when using the downloaded n-gram raw data, the information we gain is completely decontextualized: The researcher has no way of viewing individual n-grams in their original context. In addition, like other databases that have not been compiled according to a stratified sampling scheme (e.g. those in [Schlüter, 2009]), the Google Books library represents (a large selection of) written standard English and thus pastes over dialectal diversity, of which there is a lot in connection with variable /h/ (cf. [Wells, 1982: 255–256]; [Ihalainen, 1994: 217]). There is however no hope of ever accumulating dialectally differentiated materials for the past 500 years that would be ample enough to trace changes in mid- and low-frequency lexemes or specific groups of lexemes.[3] We have no choice but to trade off metatextual detail for depth of phonetic-phonological analysis.

For present purposes, however, the affordances of the database clearly outweigh its limitations: GBN cover the whole period from 1500 to 2008, offering a maximal diachronic resolution. For

[3] A study of the national varieties of British and American English is however feasible based on GBN, and is presented in [Sönning and Schlüter, to appear].

obvious reasons, the density of data increases exponentially with time and reaches sufficient numbers for the present study from 1650 onwards. Context-free bigrams provide sufficient information on most of the factors relevant in the present context, and due to the unparalleled amount of text, even low-frequency words can be included in a quantitative analysis. Similar arguments had already been marshalled in favour of large full text collections as compared to small and neat historical corpora (see [Schlüter, 2013]). For more discussion of the statistical implications of the GBN dataset compared to commonly used mega-corpora, see [Sönning and Schlüter, to appear].

### 2.2. Data pre-processing, filtering and annotation

Broadly speaking, the raw GBN bigram lists are processed in two stages: Following the download of the two data files, some pre-processing, filtering and manual annotation had to be completed before the data were imported into the app. Once imported, the data frame is processed in real time and visualization is performed interactively inside the app. The latter functions will be dealt with in Section III.

At the first stage, we used *GNU grep* to filter out all bigrams in the lists where the second constituent did not start with <h>. In order to identify all the most relevant lexemes with variable onset strength, the list of n-grams with *an* was rearranged, aggregating n-gram frequencies for both varieties and all years and sorting them in descending order of total frequency. We selected the most frequent *an* <h>-combinations plus some important obsolete or obsolescent lexemes that had not made it to the top of the frequency list because the bulk of the n-gram data comes from the most recent decades.[4] Second, manual processing involved linking the most common spelling variants (e.g. *honour/honor*, *heavy/heavie*, *heretic/heretick/heretike*) as well as obvious character recognition mistakes (e.g. *hofpital*, *hoftile*) with the entries for one head lexeme. We also incorporated some morphological variants and derivatives, provided they did not differ in terms of the (phonological) variables to be investigated (e.g. *honours/honors* under *honour*; *hyperbole* under *hyperbola*; *historical* and *historically* under *historic*; *hydraulics* and *hydraulically* under *hydraulic*). Entries, including any orthographic or morphological variants, will be referred to as *lemmas* for our purposes. Loanwords with a markedly foreign ring were excluded (e.g. *homme*, *hombre*, *habeas (corpus)*, *honnete (homme)*, *haute (couture)*). This procedure left us with 399 lemmas. Based on this list, the selected GBN bigram data and their co-occurrence frequencies with *an* and *a* per year were then imported into *R* as data frames and the entries for variants were merged with those for the main entries.

Further annotation involved enriching the entry for each lemma with five categorical variables known to influence the strength of initial /h/ from previous research ([Schlüter, 2006]; for further discussion of the categorization, see [Schlüter, 2019]). These variables are hierarchically arranged, thereby delimiting groups and subgroups that are relevant for the linguistic analysis.[5]

- The first criterion was the type of onset of the lemma. Disregarding <eu>- and <u>-types for present purposes, lemmas beginning with <h> can be split up into a large group of more or less prototypical items and one or two subtypes that show an exceptional pull towards mute initial <h>. For British English, these are the words *hour*, *honour*, *honest*, *heir* and their derivatives. For American English, the group additionally includes *homage* as well as *herb* and its derivatives. In these cases, the /h/-less realization has

[4] Selecting lemmas from the top of the frequency list of bigrams with *an* produces a strong bias in favour of loanwords, but in this case the imbalance favouring loanwords serves the purpose of the study.
[5] As will be shown in Section 3.3.2, the hierarchy implemented in the app is not fixed, but can be reshuffled by the user.

remained or has returned in the standard pronunciation (as can be displayed in **(*an*:*a*)-lyzer**; see Figures 1 and 9 below).

- Second, in some <h>-words the re-emerging /h/ has not remained the only onset consonant, but was joined by a newly developing /j/, which was often (but not always) an outcome of the adoption of loanwords with the rounded front vowel /y/ (as in *human*, *humanity*, *humour*). Among the 399 lemmas included in the study, there is also one native word with a developing /hj/-onset, namely *hue*. The combination of /h/ and /j/ has as a rule added to the consonantal force of the onset, which can be shown when comparing this group to /j/-less words in **(*an*:*a*)-lyzer** (see Figure 10, top right panel).
- Third, we subdivided lemmas according to their etymological origins, based on information in the *OED* [Oxford University Press, 2019]. Among the 399 lemmas selected, 133 are of native Germanic origin (e.g. *height*, *holy*, *hound*). 251 are loanwords, mostly of Latin or French origin (e.g. *hotel*, *horticultural*, *hospitable*) and some of Greek origin (e.g. *hypochondriac*, *heuristic*, *hydrated*), which have typically come into the language via Latin and/or French. In addition, there are 15 lemmas that are originally Germanic (Frankish) words that were borrowed into continental (Norman) French and after the Norman invasion reborrowed into English (e.g. *hardy*, *heraldic*, *helmet*, *habergeon*, *heinous* etc.). Etymological factors have played a role insofar as Germanic <h>-words have arguably never been 100 percent /h/-less, but have preserved a barely perceptible onset that has gradually re-emerged over the past 500 years (see the discussion in [Schlüter, 2019]), while Romance (and Greek) loanwords were integrated into this pattern with a considerable delay. Reborrowed Germanic-Romance-English words are assumed to have been realized more like native Germanic words (see [Pope, 1973: 15, 227]; [Minkova, 2014: 106]), which is confirmed by the data in the app (see Figure 2 and Figure 10, top left panel).
- Fourth, items were assigned to classes according to the stress level of their initial syllables. Those with lexical stress on their initial syllable were classified as 'primary stress' (e.g. *húge*, *híerarchy*, *hístory*); items with a rhythmically prominent initial syllable at a distance of one or two unstressed syllables from the primary stress were assigned to the category 'secondary stress' (e.g. *hùmanitárian*, *hòmoséxual*, *hèsitátion*); items beginning with an unstressed syllable, which is usually adjacent to a stressed syllable, were considered as having 'zero stress' in initial position (e.g. *humáne*, *hypérbole*, *históric*). The effect of stress level can be shown to differentiate between groups of lemmas and individual lemmas derived from the same stem: The more prominent the initial syllable, the stronger appears its consonantal onset (see Figure 7, the bottom left panel of Figure 10, and Figure 13).
- The final, fifth important distinction exploited to maximize the granularity of the analysis holds between vocalic nuclei of different quantities, which cuts across different stress levels of <h>-initial words: Types with long monophthongs or diphthongs (e.g. *héro*, *hỳpothétical*, *hermétic*) possess a more prominent initial syllable than types with a short monophthong within the same stress category (e.g. *hórrible*, *hìstriónic*, *habítual*). Pronunciations were checked against the *EPD* [Jones et al., 2011], *LPD* [Wells, 2008] and *OED*; in case of disagreement or attested variation in stress levels or vowel quantities, words were assigned to the category 'variable stress' (e.g. *harassment*, *hostess*, *hydrated*, *hegemony*).[6] The effect of vowel quantity on the prominence of the

[6] For words that would be assigned to different categories in the Received Pronunciation (RP) and in General American (GA), we opted for the British version due to its greater historical depth, which fits the long-term perspective of the present dataset better. Thus, *hair*, *half*, *hereafter* etc. were categorized as having long rather than short initial vowels, and vice versa for *horrible*, *hospital*, *hot* etc.

preceding onset consonant can be witnessed by comparing relevant subgroups with identical stress levels in the app (see Figure 10, bottom right panel, and Figure 11).

As a final step in data preparation, the individual lemmas in the list were then combined with the information about the five categorical variables just outlined. The final data frame contains one line per year and lemma, each annotated with the categorical grouping variables, the number of occurrences with the two allomorphs of the definite article and the number of volumes this combination occurs in. Table 1 illustrates a section of the data frame, selecting the group with mute <h> in American English (*homage*, *herb* and derivatives).

| lemma | variants | year | count_a_year | volumes_a_year | count_an_year | volumes_an_year | onset | j_glide | origin | initial_stress | v_quantity | variety | words | volumes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| … | | | | | | | | | | | | | | |
| herb | hearbe | 1721 | 1 | 1 | 2 | 2 | mute < h > in US only | no /j/ | borrowed | primary stress | long vowel | GB | 15215468 | 162 |
| herb | hearbe | 1723 | 0 | 0 | 1 | 1 | mute < h > in US only | no /j/ | borrowed | primary stress | long vowel | GB | 16270489 | 183 |
| herb | hearbe | 1724 | 0 | 0 | 2 | 1 | mute < h > in US only | no /j/ | borrowed | primary stress | long vowel | GB | 12029449 | 134 |
| herb | hearbe | 1725 | 1 | 1 | 8 | 8 | mute < h > in US only | no /j/ | borrowed | primary stress | long vowel | GB | 17281390 | 178 |
| herb | hearbe | 1726 | 1 | 1 | 98 | 3 | mute < h > in US only | no /j/ | borrowed | primary stress | long vowel | GB | 18176371 | 165 |
| herbal | | 1726 | 0 | 0 | 2 | 1 | mute < h > in US only | no /j/ | borrowed | primary stress | long vowel | GB | 18176371 | 165 |
| homage | hommage | 1726 | 1 | 1 | 0 | 0 | mute < h > in US only | no /j/ | borrowed | primary stress | short vowel | GB | 18176371 | 165 |
| herb | hearbe | 1727 | 0 | 0 | 11 | 1 | mute < h > in US only | no /j/ | borrowed | primary stress | long vowel | GB | 21815858 | 208 |
| homage | hommage | 1727 | 1 | 1 | 1 | 1 | mute < h > in US only | no /j/ | borrowed | primary stress | short vowel | GB | 21815858 | 208 |
| herb | hearbe | 1728 | 0 | 0 | 1 | 1 | mute < h > in US only | no /j/ | borrowed | primary stress | long vowel | GB | 16979039 | 198 |
| herb | hearbe | 1729 | 0 | 0 | 1 | 1 | mute < h > in US only | no /j/ | borrowed | primary stress | long vowel | GB | 18058610 | 188 |
| … | | | | | | | | | | | | | | |

Table 1. An extract from the data frame for items with mute <h> in American English.

## III THE APPLICATION

### 3.1. General design

At its core, **(*an*:*a*)-lyzer** is based on the free, platform-independent scripting language *R*. *R* was specifically designed and optimized for statistical calculations and consequently includes a wide range of built-in functions for this purpose. As is the case for many other programming languages, *R* can be further extended with a variety of pre-built functions known as libraries or packages. These can be downloaded either from a central internet repository (CRAN; https://cran.r-project.org/) or directly from third-party developers. In our scenario, we mainly relied on libraries for increasing the processing speed (e.g. *data tables*), for customizing the visualization (*ggplot2*) and for achieving interactivity.[7] With regard to the latter, we rely mainly on two packages: *Plotly* and *Shiny*. *Shiny* is an *R* package that was designed to allow users to

[7] A full list of packages used can be found in the project description in the OSF repository.

create web-based applications for interactive data visualization and exploration. The look and feel of *Shiny* apps can be further customized with additional *R* libraries (such as *shinydashboard* or *shinyTree*). *Shinydashboard* organizes the user interface in a common dashboard-like fashion and *shinyTree* integrates the javascript *jsTree* library that allows the creation of tree-like, hierarchical selection modules. *Plotly* is a library that enables interactive manipulation of the otherwise static plots produced with *ggplot2*. *Plotly*'s main affordance consists in enabling the user to zoom in and choose which groups of data points should be displayed without having to re-render the plot. A direct consequence of the interactive nature of the app is that when the user performs a request, the data are filtered and grouped based on the selected settings and all visualizations are rendered in real time. Thus, if made accessible (as in our case), the app allows readers of published work not only to reproduce the results, but also to pursue approaches to the data that previous publications may have missed.

Apps created with *Shiny* can be run either locally with one's own *R* installation or distributed over the internet. When working locally, the user needs to have an installation of *R* with all necessary libraries installed. A potential pitfall here is that not all versions of individual packages are compatible with one another. A workaround consists in bundling the source code of an app with a portable version of *R* where all the necessary packages are installed in suitable versions. If deployed on a server (which is the more common way of distribution), apps can either be hosted on a commercial platform (https://www.shinyapps.io) or on a self-managed webserver using *Shiny Server* (https://www.rstudio.com/products/shiny/shiny-server/). The appeal of commercial hosting is certainly that it requires very little technical knowledge as no server needs to be maintained. Self-hosting on the other hand offers more flexibility and functionality (e.g. access to apps can be controlled via on-premises user authentication systems).

With the above in mind, we opted for a mixed deployment strategy: The app can be downloaded as a standalone version via the Open Science Framework (https://osf.io/ht8se/).[8] In addition, while the app is still being actively developed, it is also available via browser (using the link available in the OSF repository).[9] Additionally, the repository contains the source code of the app.

### 3.2. In-app data processing

The data frame described in Section 2.2 forms the basis for all in-app processing. Running the app involves the following calculations: Based on the selected grouping variables, the data are aggregated using the *R* function *summarySE()* from the package *Rmisc*. In the default scenario (when viewing proportions of *an* : *a* or *a* : *an*), the app first aggregates the raw frequencies of each individual lemma + indefinite article combination per time interval, then calculates the proportions of *a* and *an* for each aggregated value + lemma pair, and if the grouping option is applied, it finally calculates the mean of the resulting proportions per group. This way, each lemma is weighted equally within a group and we can prevent highly frequent lemmas from dominating the proportion of the entire group. An example of the aggregated data just before plotting is given in Table 2.

[8] The standalone version runs on Windows only. To run the app on a Mac or in Linux, users have to install R and all the required packages which are specified in the source code of the app.
[9] The online version of the app automatically shuts down after a few minutes of idleness or if the connection to the server is lost (e.g. due to network problems).

| Lemma | Time interval | Frequency per time interval | Proportion |
|---|---|---|---|
| … | | | |
| herb | (1720,1729] | 127 | 0.976377953 |
| herb | (1730,1739] | 298 | 0.993288591 |
| herb | (1740,1749] | 208 | 0.985576923 |
| herb | (1750,1759] | 289 | 0.923875433 |
| herb | (1760,1769] | 237 | 0.957805907 |
| herb | (1770,1779] | 607 | 0.970345964 |
| herb | (1780,1789] | 367 | 0.948228883 |
| herb | (1790,1799] | 997 | 0.942828485 |
| homage | (1790,1799] | 202 | 0.603960396 |
| herb | (1800,1809] | 1507 | 0.869940279 |
| homage | (1800,1809] | 453 | 0.600441501 |
| herb | (1810,1819] | 1574 | 0.888182973 |
| homage | (1810,1819] | 586 | 0.392491468 |
| herb | (1820,1829] | 2254 | 0.791925466 |
| homage | (1820,1829] | 983 | 0.399796541 |
| herb | (1830,1839] | 1864 | 0.806330472 |
| herbal | (1830,1839] | 122 | 0.442622951 |
| herbarium | (1830,1839] | 350 | 0.722857143 |
| herbivorous | (1830,1839] | 127 | 0.74015748 |
| … | | | |

Table 2. An extract from the aggregated data for items with mute <h> in American English after in-app processing, just before plotting.

The *R* function *summarySE()* additionally outputs the standard deviation, the standard error and 95 % confidence intervals for the resulting values, which can optionally be visualized as error bars or error bands. The options just mentioned (**[Proportion]**, **[Group items by]** and **[Show error bars]**) will be introduced along with the other functions of the app in the following section.

### 3.3. Interface design and available options

This section details the functions that have been programmed into the **(*an*:*a*)-lyzer** app, at the same time illustrating their respective importance for the visualization output and the interpretation of the results. We proceed from the relevant linguistically informed decisions on data organization and selection to the various display options.

Figure 1. Interface of the web app **(*an*:*a*)-lyzer**.

The interface itself is divided into two parts: the **options sidebar** on the left of the screen and the **plot panel** on the right (see Figure 1). The x-axis represents the timeline (with the labels indicating the first year of the respective interval); the y-axis represents the proportions of frequencies of bigrams fulfilling the criteria selected in the options sidebar. For both axes a number of choices can be set. The options sidebar contains all settings, including advanced controls for selecting and filtering data as well as options related to the graphical representation of the data. The individual settings are grouped by the functions they serve and are described in detail in the following subsections.

The first two functions are relevant for the basic running of the app. **[Update plot]** is the button that has to be clicked each time settings in the options sidebar have been changed. It redraws the plot with the altered settings. **[Load data]** asks the user to select a pre-defined (sub-)set of bigrams. As the app was designed for the purposes of more than one study, this drop-down menu lets the user choose which part of the dataset should be loaded. The appropriate set will be pre-selected according to the web link from which the app is accessed, but users can also switch to the complete dataset named **[All data]**, which provides access to (native and borrowed) <h>-initial words as well as to <u>- and <eu>-initial words investigated in other publications (e.g. [Schlüter, 2019]).

### *3.3.1. General*

The most important panel where the user can select data to be visualized and analyzed is labelled **[General]**. The radio buttons can be set to **[Show proportion of]** *an* : *a* or of *a* : *an*; the setting that is chosen is rendered on the y-axis of the plot. Together, *a* and *an* always add up to 100 %, so that either setting provides exactly the same information. In the default setting, which is *an* : *a*, a falling proportion along the time axis can be interpreted as indicating a reinforcement of the onset consonant, and an increasing proportion as a weakening of the onset consonant. If we are witnessing a diachronic increase in onset strength of <h>-initial items, we should thus see a decrease of *an* in favour of *a* for the average of lemmas, which is indeed the case. Figure 2 exemplifies both ways of looking at the data for the three major groups of borrowed, native and reborrowed <h>-words.

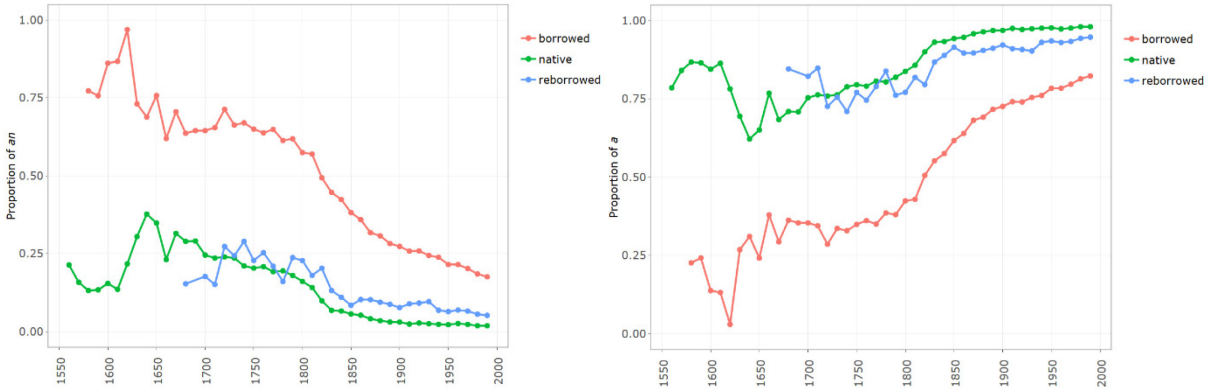

Figure 2. Proportions of article allomorphs for the groups of borrowed, native and reborrowed <h>-initial words. Left panel: Proportion of *an* : *a*. Right panel: Proportion of *a* : *an*.[10]

For a few exceptional lemmas, however, the data indicate a fall followed by a rise in the proportion of *an* : *a*, thus a reversal of the consonantal strengthening (see Figure 3, left panel). A further subdivision of the data into British and American English[11] reveals that the renewed rise is largely restricted to the American English data (labelled US), where *herb* and its derivatives and *homage* are nowadays frequently /h/-less (see Figure 3, right panel). In British English (labelled GB), *homage* also seems to be affected by this trend, to the extent that the geographical division of the data is reliable.

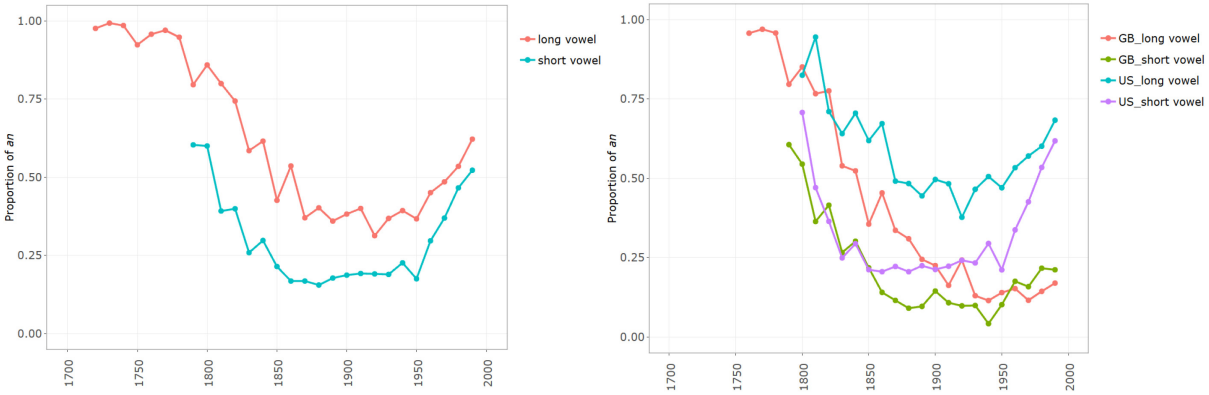

Figure 3. Proportion of *an* : *a* for the group *herb*, *herbal*, *herbalist*, *herbarium*, *herbed*, *herbicide*, *herbivore* (labelled 'long vowel') and the item *homage* (labelled 'short vowel'). Left panel: All data. Right panel: GB and US separated.

Before any diagram can be created, the data to be displayed have to be selected. This is done with the help of the **[Lemma selection tree]** in the **[General]** panel in the sidebar. By default, the tree classifies the items hierarchically by onset type, presence of a /j/-glide, loanword status, initial syllable stress and vowel quantity (cf. Section 2.2). Users can view the classes and subclasses of lemmas by clicking on the nodes to expand all relevant branches of the tree. Figure 3 singles out the group around *herb* and the individual lemma *homage*. To take the example of *herb*, this can be found in the onset class **[mute <h> in US only]**, in the class **[no /j/]** on the second level, in the class **[borrowed]** on the third level, in the class **[primary stress]** on the fourth level and in the class **[long vowel]** on the fifth level. By ticking the boxes for individual

[10] To ensure replicability of the figures generated for the present study, all settings for the options sidebar are listed in the table "Appendix_(ana)-lyzer_settings_for_figures" online at https://osf.io/ht8se/. There, the reader can also find high-resolution versions of all the figures.
[11] This can be achieved by setting the function **[Group items by]** to **[Variety]**, as explained in Section 3.3.4.

items or for whole super- or subordinate groups, users can select and deselect (single or multiple) lemmas or groups of lemmas to be included in the plot. A visualization is generated upon clicking **[Update plot]**.

*Plotly* comes with some interesting and useful display options that can be applied to the plots to facilitate visual inspection and data analysis. The plots can be viewed as interactive plots (the default) or as non-interactive images, which tend to be less error-prone than the former in the case of complex searches. For that, click the option **[Non-interactive plot]** above the plot panel. In the **[Interactive plot]**, running the mouse over the graphs pops up labels for each data point, which specify the name of the group or individual lemma and the exact proportion of the indefinite article allomorph displayed. In case a single lemma is graphed, they also indicate the total number (value) of tokens for that item for the respective year(s), including both variants of the article. In case a group of lemmas is graphed, value means the average number of tokens across all lemmas in that group, which allows the user to get a rough idea of the amount of data making up a data point. At the top of the interactive plot, users find a range of viewing options that can be activated (see Figure 4).

Figure 4. The display options in *Plotly*'s interactive plot panel.

Among other functions, these allow zooming in and out, manually selecting parts of the plot, scaling the axes, indicating spike lines and simultaneously comparing the levels of all the curves displayed for a particular time interval. The latter two options are shown in the screenshot in Figure 5: The spike lines help the viewer identify the exact values of a data point on both axes, and the function described as "Compare data on hover" (the second icon from the right) displays multiple pop-up labels that specify the y-axis levels of all curves drawn for a particular time interval (by default, a decade). In addition, individual curves can be turned off and on by clicking on their labels in the legend to the right of the plot. This is illustrated in Figure 11 below. In the case of many superimposed curves, hiding selected curves can help disentangle the display.

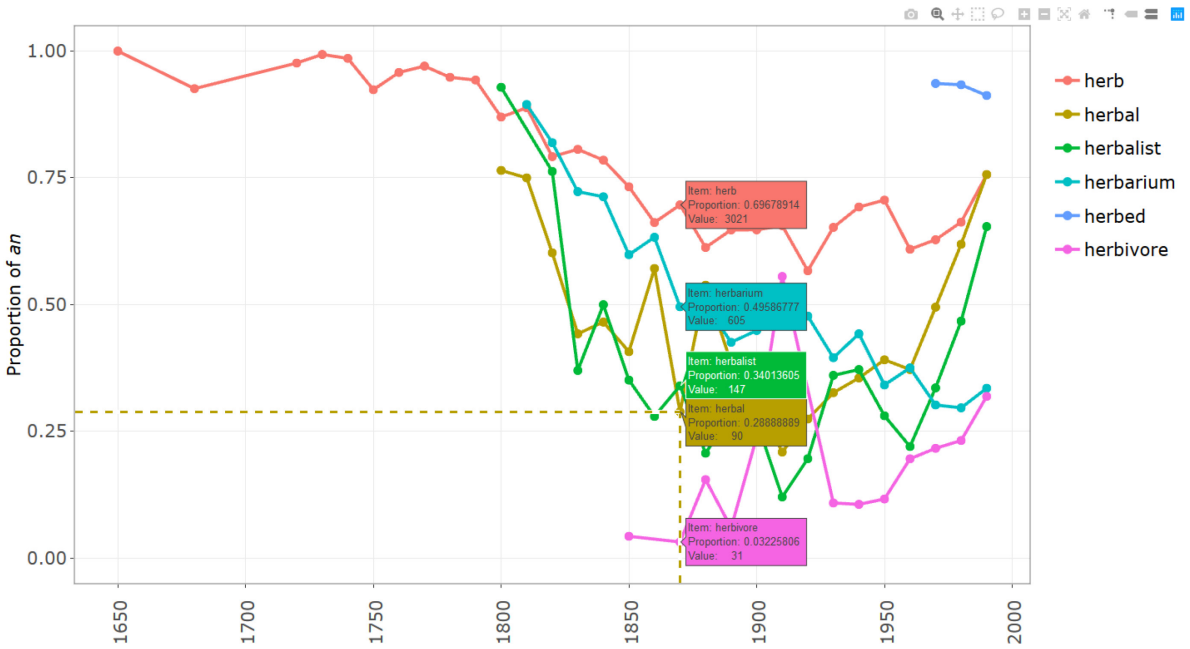


Figure 5. Interactive plot with individual items plotted and display options **[Toggle spike lines]** and **[Compare data on hover]** activated. Proportion of *an* : *a* for *herb* and its derivatives.

### *3.3.2. Reorder lemma selection tree*

Using the panel **[Reorder lemma selection tree]** in the options bar allows the user to flip the hierarchical order of the tree, which is pre-set to onset type > presence of a /j/-glide > loanword status > initial syllable stress > vowel quantity. For example, if a specific analysis should wish to foreground the distinction between primary, secondary and zero stress and to relegate that between native and borrowed words and presence/absence of /j/ to the background, the user can drag and drop the grouping criteria from **[Available classifiers]** to **[Order]** and click **[Update order of item tree]**. After the tree has been re-arranged, the user can select all items with primary stress at a single click, which percolates to all subordinate levels of the adapted hierarchy. Figure 6 shows the default tree and the adapted tree, as well as the boxes that need to be checked to select all lemmas with primary stress on their initial syllables.

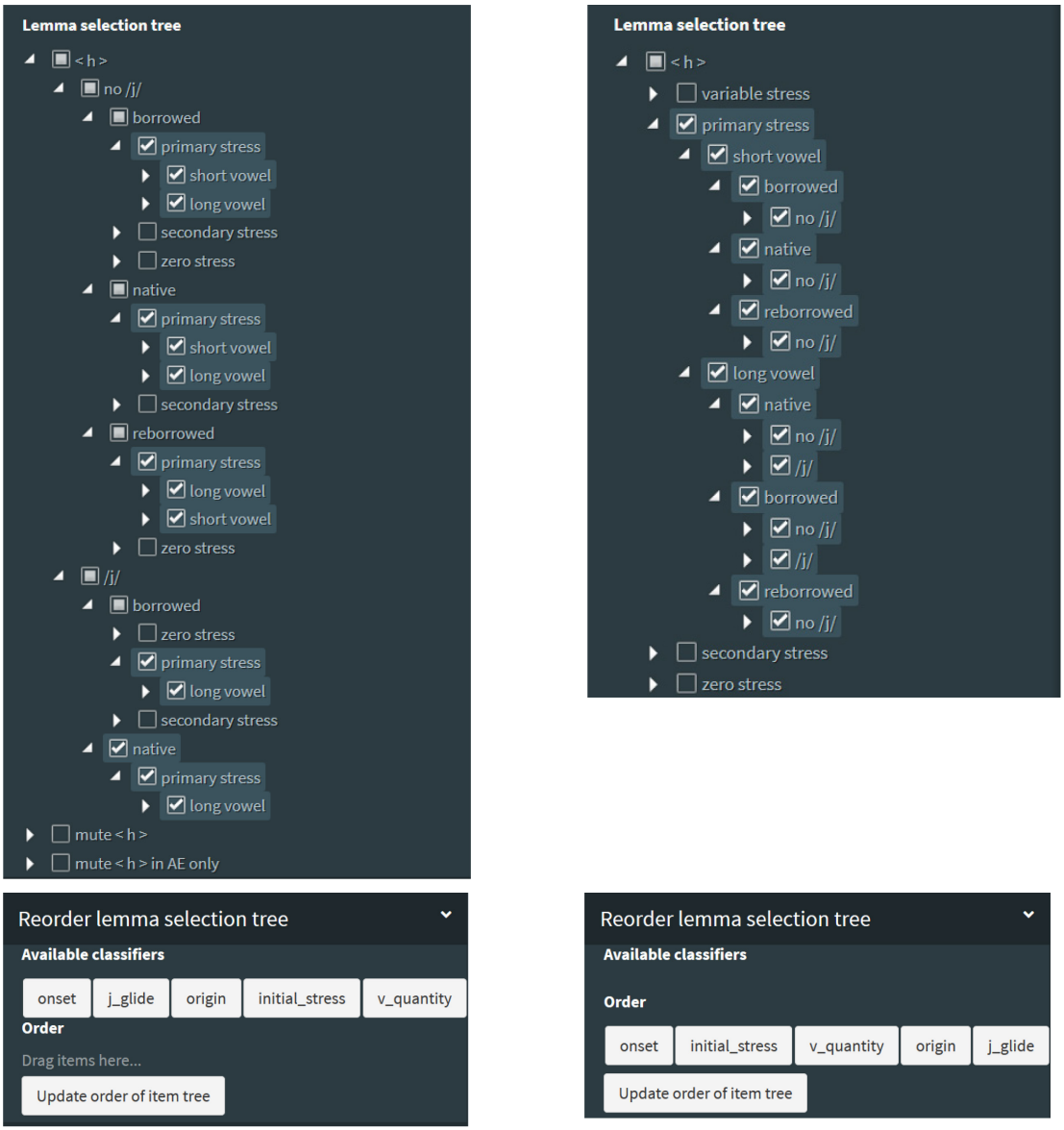


Figure 6. Lemma selection tree for capturing all items with primary stress and interactive reordering function. Left panel: Default tree order. Right panel: Reordered tree prioritizing initial stress and vowel quantity over origin and /j/-glide.

If fewer than the complete set of five categories are used, the resultant tree becomes flatter; if only onset is used as a classifier, all the words of a common onset type are listed in alphabetical order.

### *3.3.3. Filters*

Applying the **[Filters]** enables the user to automatically include or exclude parts of the underlying n-gram data based on the distinction between British and American publications, frequencies and years of attestation. For one thing, **[Include data from varieties]** distinguishes between books published in Britain (labelled GB) and the United States (US). The reliability of this distinction as a proxy for the variety used by the author should be taken with a large pinch of salt (see [Sönning and Schlüter, to appear]), and this is where the diatopic information that can be extracted from GBN ends. Even so, in the case at hand, studying British and American data separately produces the expected results: With the notable exception of the lemmas

*homage*, *herb* and derivatives of the latter (seen in Figure 3, right panel), American English is generally more advanced in the re-emergence of initial /h/. A close-up comparison of the proportions in the left and right panel of Figure 7 shows that the difference has for some decades been larger with primary and secondary stress, but is now perpetuated mostly in items with zero stress.

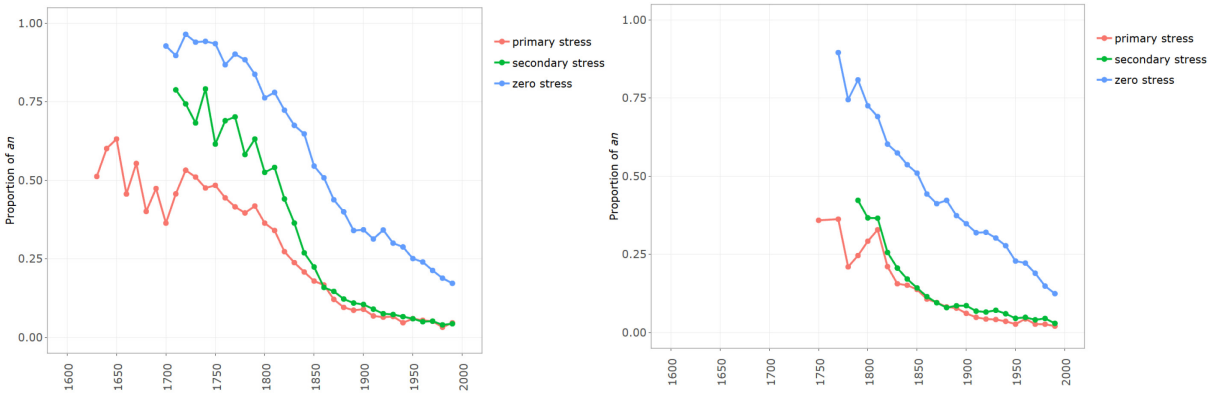


Figure 7. Proportion of *an* : *a* for borrowed <h>-initial lemmas without /j/ grouped by initial stress. Left panel: Filter set to GB. Right panel: Filter set to US.

The function **[Basis of quantification]** draws on the information provided by the GBN files about the number of volumes containing the bigrams in question. By default, the app counts the bigram tokens per time interval, irrespective of the number of works in which they are contained. For statistical reasons (see [Sönning and Schlüter, to appear]), it may sometimes be desirable to include only one or two records per volume: one in case the volume consistently selects either *a* or *an* with a specific <h>-initial word; two in case the volume oscillates between *a* and *an* (in which case each combination would represent one record). The effect that this has on the proportions of *an* and *a* can be explored by setting the function to **[Volumes]**.

To reduce noise from low-frequency items, it may also make sense to include only items above pre-determined thresholds per time interval. Note that low-frequency items are given the same weight in the aggregation as high-frequency items. Thus, in the default settings, we exclude any lemmas that appear less than 10 times per selected time interval (since calculating their proportions of *an* : *a* makes little sense) and we do not visualize groups that are not supported by at least 100 bigram tokens. In addition, since the focus of the present contribution is on the pronunciation of individual items and on diachronic shifts in the realization of phonemes, in the graphs selected for illustration we discount all lemmas that made their first appearance in a bigram with *a* or *an* after the year 1900 (thus excluding items like *hallucinogen*, *histogram*, *holistic*, *hypoxic*, etc.). The latter step was taken to discard highly technical and rarely pronounced words. This focus can however be changed by setting the slide bars for the filters **[Minimal frequency per time interval and item]**, **[Minimal frequency per time interval per group]**, **[Process only data for years]** and **[Process only items where data exists before]** to different cut-off points, for instance if the research should target the phonological integration of recently borrowed loanwords in the 20th century.

### *3.3.4. Group data*

Turning to another important panel of the options sidebar, we now illustrate the various functions subsumed under **[Group data]**. To get an impression of the amount of data and detail

included in the set, consider Figure 8, which displays the ungrouped curves for all 399 <h>-initial lemmas.

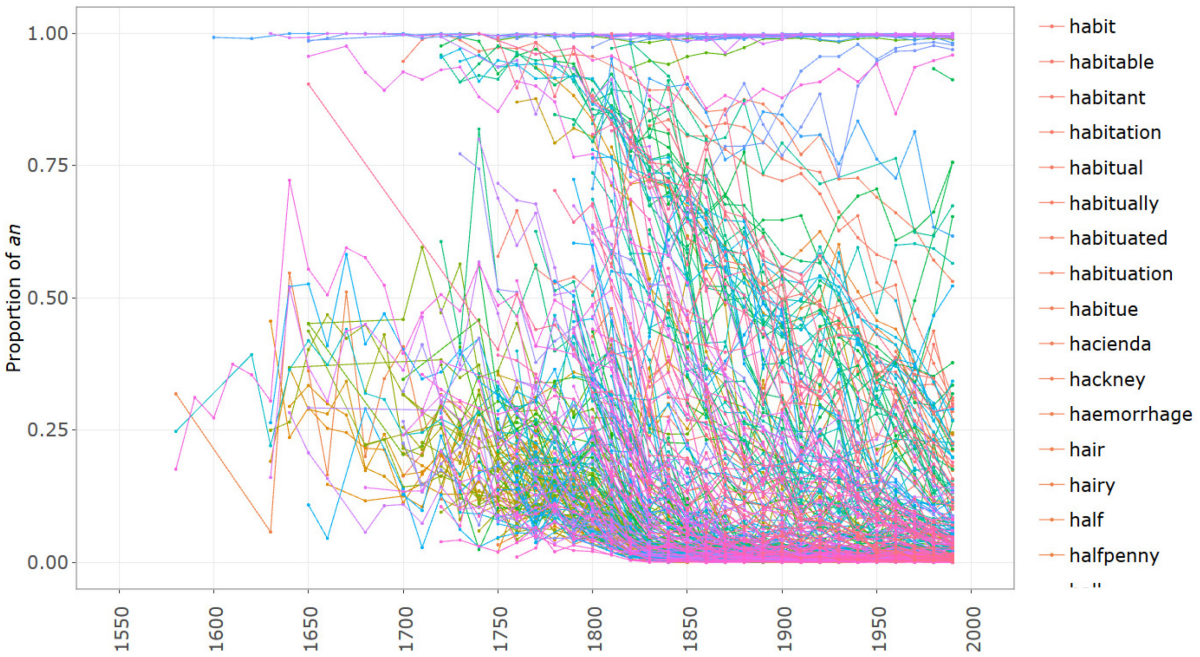


Figure 8. Proportion of *an* : *a* for all <h>-initial lemmas, ungrouped.

Visual inspection and/or information available from previous research on the topic can help discern three or more groups that cluster together. As explained in Section 2.2, factors that have been shown to distinguish between different degrees of initial consonant strength are subtypes of <h>-onsets (variably pronounced, generally mute or frequently mute only in American English), presence of a /j/-glide, loanword status, initial syllable stress and vowel quantity, and each lemma in the underlying dataset has been annotated with these features. In addition, the n-grams divide between Google Books from GB and the US. We can now apply interactive groupings according to these factors by making appropriate (single or multiple) selections through the **[Group items by]** function and assess differences between the proportions of *an*-selection for lemmas per group. One result of grouping by **[Variety]** has already been shown in Figure 3 (right panel), another one of grouping by **[Initial Stress]** has appeared in Figure 7. Figure 9 is produced via grouping by **[Onset]**, which assigns three different colours to the three onset types.

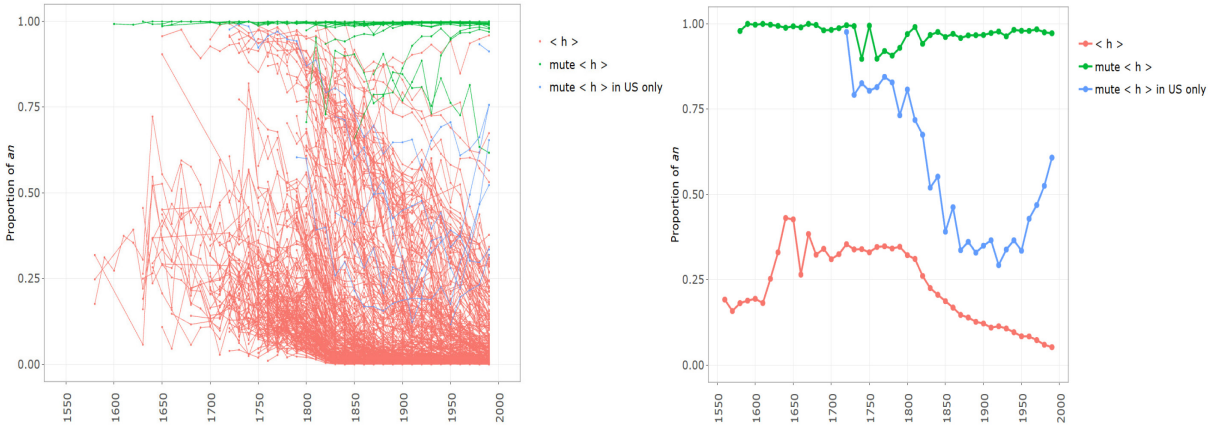


Figure 9. Proportion of *an* : *a* for all <h>-initial lemmas, grouped by onset type. Left panel: Plotting individual items. Right panel: Plotting aggregated data per group.

As shown in Figure 9, left panel, multiple curves (one per lemma) are plotted if the radio button for the function **[Plot individual items]** is set to **[Yes]**. The grouping function is then only used for colouring. By default, this function is set to **[No]**, which reduces the number of curves to the number of groups, based on the proportions of *an* : *a* averaged across all members of that group figuring in the relevant time interval (Figure 9, right panel).

Further interesting groupings are depicted in Figure 10. We now restrict the analysis to the largest and most diverse group of <h>-initial words in which the pronunciation of /h/ has tended to re-emerge to various degrees and at various speeds across the past five centuries. Again, this development is discussed in detail elsewhere [Schlüter, 2019], which is why we can restrict ourselves to exemplifying the visualization options and their affordances for exploratory data analysis. The distinction between lemmas with an emerging /j/-glide and those without it can only be usefully applied to borrowed words, since among the Germanic words in the dataset there is only one (*hue*) pronounced with /j/ (and none among the reborrowed Germanic-Romance words). Figure 10, top left panel, selects only loanwords excluding mute <h> (via the **[Lemma selection tree]**) and groups them according to the presence or absence of /j/ (via the **[Group items by]** function). The resultant picture shows that the combination of a re-emerging /h/ and an emerging /j/ reinforces the consonantal onset and reduces the proportion of *an*. The top right panel selects only words without the /j/-glide and subdivides them according to their etymological origins, indicating that native Germanic words have quite consistently had the strongest realization of initial /h/, followed at a surprisingly short distance by originally Germanic words that had been borrowed into French and from there reborrowed into English, followed at a large distance by ordinary Romance loanwords. We can thus assume that etymological distinctions have persisted insofar as Germanic <h>-words have demonstrably preserved a subliminal trace of their onset, while Romance (and Greek) loanwords were integrated into the native pattern of pronounced <h> with a considerable delay.

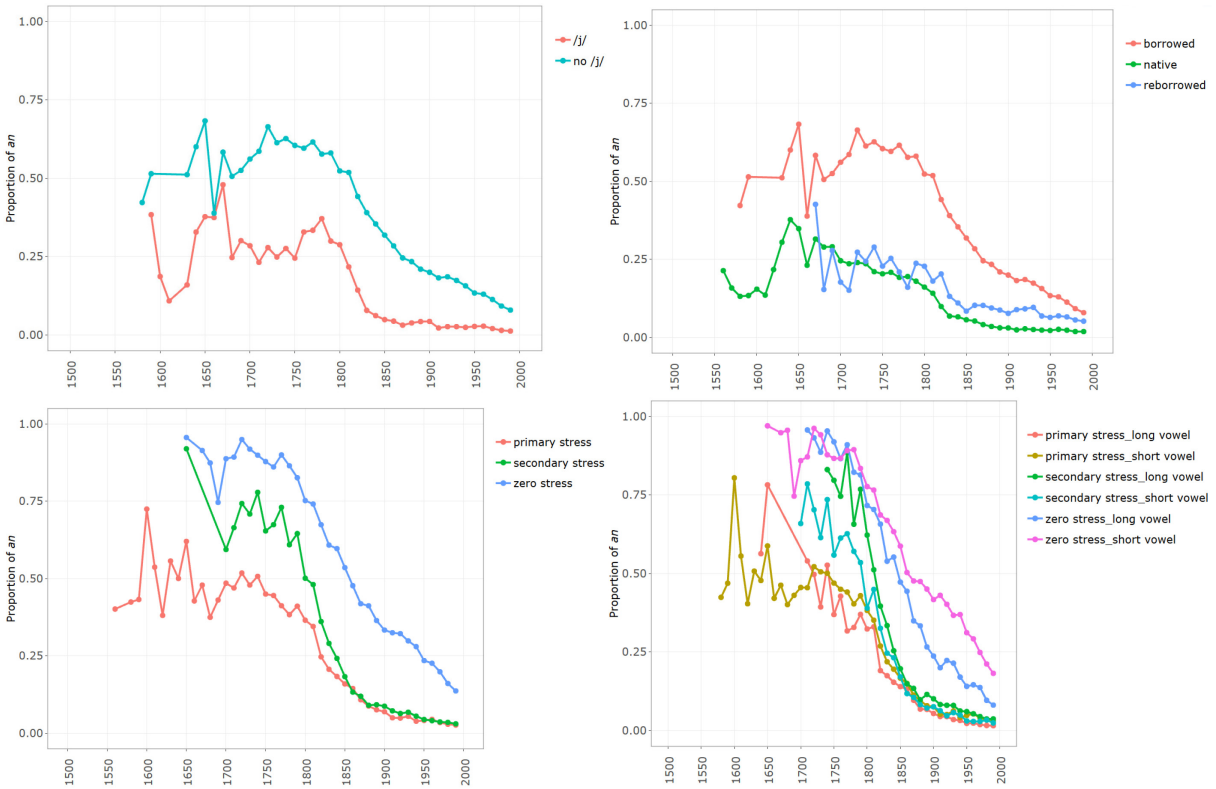


Figure 10. Proportion of *an* : *a* for <h>-initial lemmas. Top left panel: Borrowed words, grouped by presence/absence of /j/. Top right panel: Words without /j/, grouped by etymological origin. Bottom left panel: Borrowed words without /j/, grouped by initial stress. Bottom right panel: Borrowed words without /j/, grouped by initial stress and vowel quantity.

The two panels at the bottom restrict the set of lemmas selected to the most conservative group identified so far, the regular (Greek-, Latin- and French-derived) loanwords. This group still consists of 188 individual lemmas, which form a heterogeneous set because they differ with regard to stress placement and quantity of their initial vowel. The bottom left panel of Figure 10 groups these items according to three degrees of stress on their initial syllables, while the bottom right panel introduces an additional division by vowel quantity. The threefold division in the bottom left panel illustrates very clearly that the average strength of an /h/-onset has always correlated with the stress level of the syllable in which it occurs: Up to the 18th century, lexemes with a secondary stress were roughly intermediate between those with primary and zero stress. Since the mid-19th century, they have all but converged with those with primary stress, so that combinations with *an* have become virtually restricted to words with completely unstressed initial syllables. The bottom right panel paints an even more fine-grained picture, where we can see that all three stress levels fall into two subgroups. Diachronically, subtypes with short vowels in their initial syllables tend to have even weaker onset consonants than those with long vowels. While lemmas with secondary and primary initial stress converge towards the present day, the difference remains noticeable among items with zero stress. We can thus assume that the perception of a word's onset consonant depends on the prominence of its first syllable measured in terms of stress as well as vowel quantity.

The seemingly consistent picture does, however, exhibit one unexpected configuration: Among the secondary-stressed lemmas, those with a long vowel nucleus seem to be conspicuously slower to implement the /h/ than those with a short nucleus. At this point, the analyst can resort to the option **[Plot individual items]** to figure out which group members are responsible for this unexpected finding. Switching off the curves for primary and zero stress by clicking on

their labels in the legend leaves us with the plot for items with secondary stress only, shown in Figure 11.

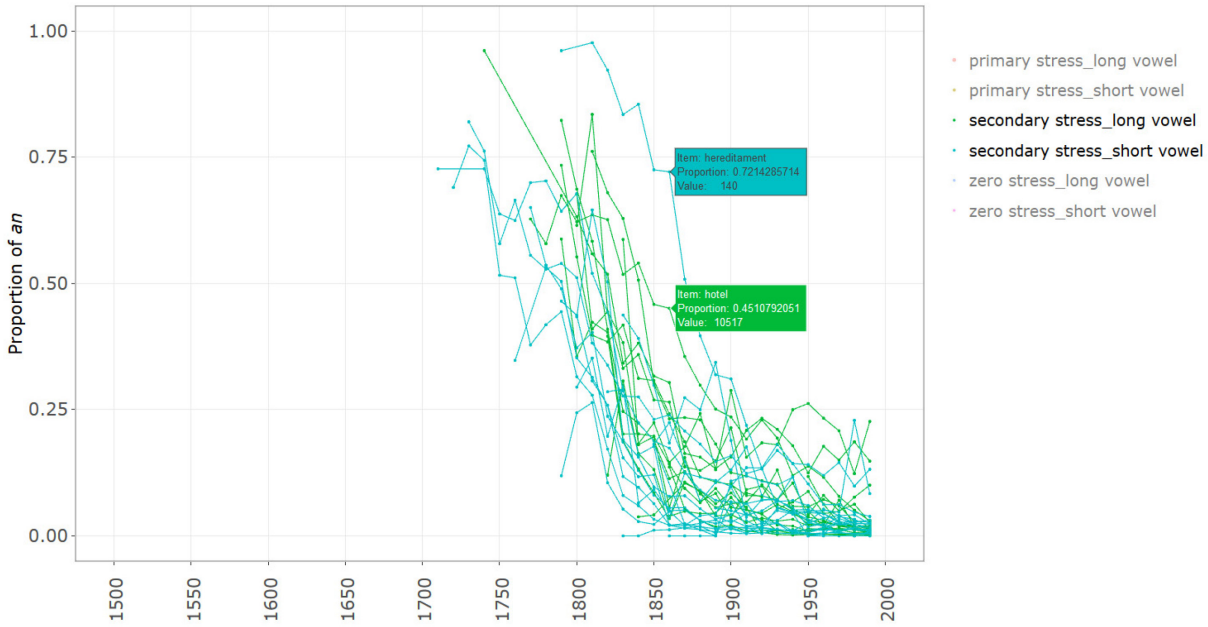


Figure 11. Proportion of *an* : *a* for borrowed <h>-initial lemmas without /j/. Plotting individual items, grouped by initial stress level (primary and zero stress deselected) and vowel quantity.

This representation shows that apart from one or two particularly conservative items (identified by the pop-up labels as *hereditament* and *hotel*), the behaviour of the whole group is relatively homogeneous, with a slight but stable skewing of items with short vowels to the progressive (*a*-selecting) side and of items with long vowels to the conservative (*an*-selecting) side. The choices offered by **(*an*:*a*)-lyzer** in the next panel to be discussed cannot resolve this riddle, but perhaps they can take us one step further in its exploration.

### *3.3.5. Display options*

Let us first sketch the **[Display options]** to be found under the label **[Basic]** in the options bar. These do not come automatically with *Shiny*, but have been implemented because they allow the user some flexibility when creating plots for different purposes. Their names are relatively straightforward: The function **[Draw]** can be set to **[Points]**, **[Lines]** or **[Smoothed lines]**.[12] Combining these options is possible.

The drop-down menu **[Show error bars]** offers three different uncertainty measures. Users can supplement the plotted averages with bands indicating 95 % confidence intervals, standard errors or standard deviations. In this context, it should be pointed out that we have deliberately refrained from showing error bars in the diagrams extracted for this paper. Deriving uncertainty estimates for the data turns out to be far from straightforward since the tokens are not all independent of each other (see [Sönning and Schlüter, to appear]): Many authors contributed more than one token and more than one lemma and can be expected to show some internal consistency. What is more, the curves for groups of lemmas are aggregates of subgroups, which

[12] **[Smoothed lines]** employs *ggplot2*'s *geom_smooth()* function.

can be shown to diverge more or less substantially from the aggregate average, and subgroups are again aggregates of lemmas with potentially divergent behaviour.

By default, the display option **[Autoscale X-axis]** is activated, so that the timeline is limited by the amount of data available for an individual plot. To ensure a constant scale, irrespective of the spread of the data (as for multiple plots shown next to each other, e.g. Figures 3 and 7), this can be deselected. If deselected, the **[Limits of X-axis]** slide bar appears and can be manipulated if the user wants to focus on a particular stretch of time. This has been done for the graphs in Figure 12. If the preselected function **[Set scale of Y-axis to 0-100 %]** is de-activated, the web app will autoscale the Y-axis depending on the range of proportions displayed.

The **[Advanced] [Display options]** can take us one step further in the exploration of the dataset of Figure 11. There are two functions that integrate indications of frequency into the display. Selecting the same data and groupings as above, but activating the additional display function **[Scale shape according to frequency]**, we obtain the visualization in the left panel of Figure 12. The size of the circles surrounding the points (provided that the function **[Draw]** is set to **[Points]**) is now determined by the total frequency of a lemma (in bigrams with *a* and *an*) per time period, which helps find out at a glance how much evidence is behind each data point. We can see that the lemma *hotel* stands out from the group with long vowels for being much more frequent.[13] Activating the function **[Opacity of lines according to total frequency]** correlates the shading of the lines with the total number of n-grams in the group, not of the individual data points. This is useful to quickly compare aggregated frequencies of groups or to identify highly frequent items (if plotted individually). The right panel of Figure 12 shows this for the entire six groups defined by stress level and vowel quantity. Taken together, these depictions suggest that the dominant patterns in terms of frequencies are provided by lexemes with primary stress and long vowels, and by lexemes with secondary or zero stress and short vowels. In contrast, words with secondary stress and a long vowel have low token frequencies, the single most frequent item among them being the lemma *hotel*. This item appears to be particularly conservative (possibly a tribute paid to its French connotations), but deselecting *hotel* in the lemma tree does not change the proportions for the remaining group substantially. The other group members (many of them containing the Greek-derived combining forms *hiero-*, *hydro-*, *hyper-* and *hypo-*) also tend to combine with *an* more often than expected based on their rather prominent vocalic nucleus. We have no conclusive explanation for the unexpected reversal of the curves for secondary stress. It would be desirable to trace the n-grams that have entered the dataset back to the contexts, authors, books and genres in which they appear, but the GBN database does not offer this kind of information.

[13] An alternative basis for the size of circles, involving a visualization of frequencies based on the size of the GBN database for a specific period (per million words), remains on our to-do-list.

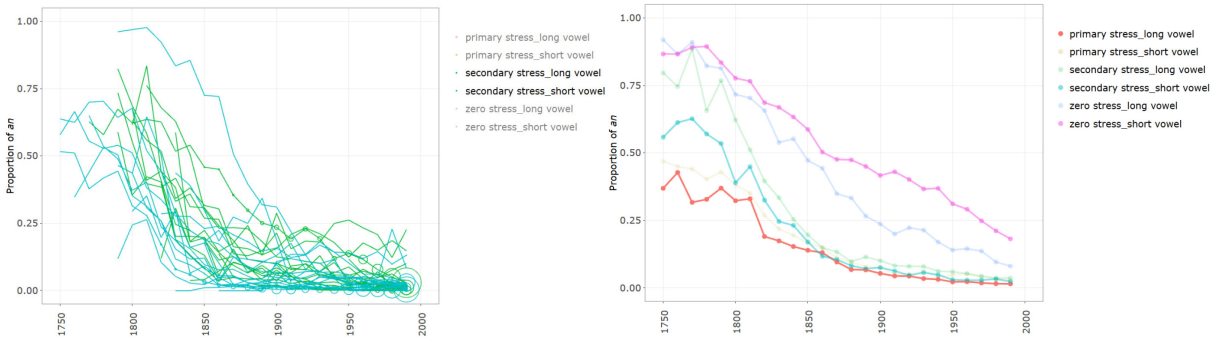


Figure 12. Proportion of *an* : *a* for borrowed <h>-initial lemmas without /j/ grouped by initial stress level and vowel quantity. Left panel: Plotting individual items and scaling shape according to frequency, primary and zero stress deselected. Right panel: Plotting groups and adjusting opacity according to total frequency.

The remaining **[Advanced] [Display options]** are self-explanatory and serve to adapt the display (as well as exported PNG or PDF files) for special requirements (publication, projection etc.). **[Disable colours]** serves to turn plots into black and white or greyscale. **[Stroke width]** controls the width of the lines, **[Point size]** controls the size of the shapes, and **[Font size]** controls the size of the labels and legends.

### *3.3.6. Aggregation and transformation*

The functions in the panel **[Aggregation & transformation]** of the options bar perform some calculations on the data that allow the user to view the data from different perspectives. The drop-down menu for **[Time interval for aggregation in years]** is set to decades by default. Alternatively, the intervals over which the data are aggregated can be set to various values between 1 and 100 years. In the case of low-frequency lemmas, small values can lead to considerable zigzagging of the lines and to gaps for intervals with fewer tokens than the minimal frequency per time interval and item; but setting the span too long would paste over gaps in the data, miss out on incomplete intervals (such as the interval 2000-2049) and obscure the high temporal resolution afforded by the underlying dataset (see the two visualizations of three morphologically related words that differ in terms of initial syllable stress in Figure 13).

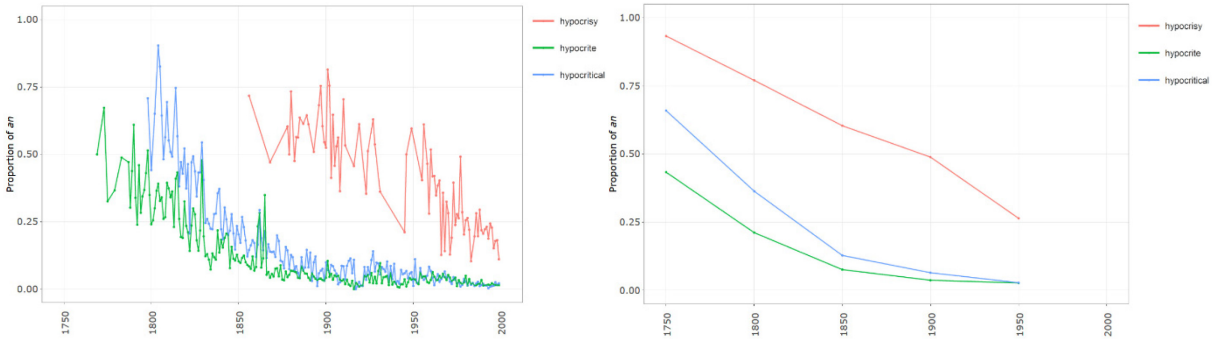


Figure 13. Proportion of *an* : *a* for *hypocrisy*, *hypocrite* and *hypocritical*. Left panel: Aggregation baseline set to 1 year. Right panel: Aggregation baseline set to 50 years.

Finally, the options under **[Transformation]** determine whether the data are visualized as **[Proportions]** of *an* and *a* (which is the default setting used for the above plots) or as **[Absolute frequencies]**, normalized frequencies **[Per million words]** or log-transformed frequencies **[Log10(+1)]**, all in bigrams involving one of the indefinite article allomorphs. The latter three options can be employed to determine changes in the frequencies with which a lexeme or group of lexemes occurred in different periods. In this way, the analyst can follow up on the question

if usage frequencies have an influence on the strength of onset consonants. Note that if one of the latter three frequency measures is selected, the numbers of *an* and *a* are totted up and proportions of *an* and *a* can be viewed only in the pop-up labels of the interactive plot.

### *3.3.7. Export options*

Among the **[Export options]**, the two functions **[Width of exported diagram (cm)]** and **[Height of exported diagram (cm)]** determine the resolution of the plot for export. Based on the plot currently displayed, the function **[Export high-res plot]** produces a high-quality PNG or PDF file (e.g. for publishing and printing), dependent on which **[File format]** is selected.

## IV DISCUSSION AND EVALUATION

### 4.1. Accessing big data

Undoubtedly, GBN supplies an unprecedented amount of data, covers five centuries (with an exponential rise in density of coverage) and comes with an ideal temporal resolution. This combination of features offers linguists new and promising opportunities. On the downside, an important impediment precluding many possible research objectives consists in the limitation of n-gram length to the maximum of five elements, i.e. 5-grams. Any study that requires access to a larger context is seriously hampered by this restriction. What is more, even in studies like the present one, recourse to contextual and metatextual information would be important to analyze unexpected results (such as the reversal of proportions of article variants with loanwords having a secondary stress on their initial syllables; see Figures 11 and 12). All that GBN provides us with is the publication year and whether the book was printed in Britain or in the US (for what this piece of information is worth as a proxy to the variety in which the book was written). We have no way of knowing whether the book was a first edition or a re-print, which genre it represents, how many n-grams were contributed by one and the same book or author etc.

### 4.2. Developing the app

The development of **(*an*:*a*)-lyzer** itself proved to be more time consuming than initially expected. One of the reasons was that new potential avenues of research revealed themselves at different stages of this iterative process, which in turn required the programming of new functions or sometimes even the rewriting of core functions. The decision to make the app available to the public further increased the necessary development efforts: A wider user base necessitates the implementation of error handling and a more careful design of the interface. In view of the hierarchical structure of the lemma selection tree, for example, we deemed it necessary to also implement an option for reordering this hierarchy. The rationale behind this is that since the data are not hierarchically organized per se, implementing a predetermined and strictly hierarchical presentation might limit potential alternative approaches or, worse, distort the perception of the data.

It should be noted at this point that the development of the app has not reached a final state yet. Two features that are left on our to-do-list are an option to scale the size of data points according to normalized frequencies based on the size of the GBN data for a specific period (per million words), and a feature allowing researchers to create and share reproducible links to the settings for particular graphs.

### 4.3. Using the app for research

Once the huge amount of information has been extracted from the raw data and the different levels of annotation have been added, interactive visualization offers the researcher a convenient approach to constructing, re-constructing and comparing groupings of lexemes. It is fascinating to immerse oneself in the data and view them from ever changing angles. Zooming in and out, from the general picture to the individual lemma and back, allows one to see patterns that would formerly have gone unnoticed, but it also exposes divergences between group members that may come as a surprise and lend support to linguistic models that capitalize on lexeme-specific mental representations (such as exemplar-based phonology) and differential change (such as lexical diffusion). As a result, we can assert that visualizing big data in an interactive app imparts us with knowledge on a new dimension and allows us to answer questions that we could not have answered when we had only linguistic reference corpora as databases.

Of course, there remain questions that we cannot answer with the help of exploratory data visualization alone. We cannot specify the exact weight of the linguistic factors involved (variety of English, onset type, presence of a /j/-glide, loanword status, initial syllable stress, vowel quantity) for the realization strength of initial /h/: Indeed, each data point is simultaneously subject to all of these effects and there are important interactions between factors. For instance, Romance loanwords have mute <h>, a /j/-glide and non-initial stress significantly more often than Germanic words. To assess factor weights and interactions, we would need to conduct a hierarchical mixed-effects logistic regression analysis with individual lexemes as random effects. Moreover, the model would have to include individual authors (or texts) as random effects, but such an enterprise is seriously hampered by the lack of metatextual information in the GBN database (see [Sönning and Schlüter, to appear]).

### 4.4. Sharing the app

Considering the manifold visualizations that are not only feasible, but also linguistically informative, making the app available to readers and to the public is an attractive option. The relevant comparisons at different levels of granularity have still not been exploited to the full, and in this way, users can explore the data according to their own interests. What is more, by replicating the visualizations presented here, users can zoom in further to see which selection of lemmas the curves are based on, or they can zoom out to view the larger picture. This relieves us as authors from the need to provide lengthy appendices, such as lists of coded lemmas, or to stretch the credulity of our readers by performing what may look like magic on our data. The app should remain accessible on a permanent basis thanks to the alternative options for sharing that we use: Hosting the app on a server provides quick and easy access to the public. Making it available as a standalone version for download, which is no longer subject to change in the software packages it integrates, and uploading it to a platform like OSF should ensure that the app does not require constant maintenance and can be run independently on the user's own machine.

### 4.5. Evaluating the interactive visualization

As developers of the tool and as researchers interested in diachronic change and synchronic influences affecting consonantal onsets in English, we do of course value the affordances of the interactive visualization offered by the app. Undoubtedly, it does meet its key requirement – making a huge dataset amenable to linguistic investigation. Thanks to the unprecedented GBN data, the manual pre-processing and annotation that have gone into the data frame, the attractive interface using *Shiny* and *Plotly* and the various options specially programmed into the app,

exploring the data offers instantaneous results for any combination of parameters that the researcher decides to set. What is more, visualization makes the results easy and fast to grasp and compare, thereby increasing the gratification that the user may experience.

On the negative side, the GBN data – while inviting further exploration – come with major drawbacks. Although developed and documented in a transparent way that makes recycling (parts of) it possible, the visualization app described in this contribution is currently restricted to data on consonantal onset strength of certain categories of words preceded by an indefinite article. Besides <h>-initial lexemes, which have taken centre stage here, the app includes <u>- and <eu>-initial words, but these three sets exhaust the data that have so far been extracted from the GBN lists of bigrams. An extension that could be implemented rather easily would include the determiners *my* : *mine*, *thy* : *thine* and *no* : *none*, which remained variable during the Early Modern English period, with the short forms coming to prevail in prenominal position. More demanding extensions could be designed for linguistic phenomena in which proportions of two alternative options in a 2-, 3-, 4- or 5-gram are of interest, such as dual form adverbs before certain adjectives (e.g. *real good* vs. *really good*), *a quite* vs. *quite a* in attributive structures (e.g. *a quite bad idea* vs. *quite a bad idea*), variable determiner use (e.g. *to school* vs. *to the school*) or alternative prepositions (e.g. *different from* vs. *different to*). There is no doubt that each example of linguistic variation will come with its own challenges, but nothing ventured, nothing gained.

Time will tell – and we hope that readers of this article and users of the app will tell us – if the effort will pay off and inspire more research. The OSF platform includes a comments section where we ask for feedback, answer any queries about the app, the data or their interpretation, post new updates and would be glad to hear about any uses to which the resource has been put.

## References


Gries, S. Methodological and interdisciplinary stance in corpus linguistics. In V. Viana, S. Zyngier and G. Barnbrook (eds.). *Perspectives on corpus linguistics* (Studies in corpus linguistics 48). Benjamins (Amsterdam). 2011: 81–98.

Hiltunen, T., McVeigh, J. and Säily. T. How to turn linguistic data into evidence? In T. Hiltunen, J. McVeigh and T. Säily (eds.). *Big and rich data in English corpus linguistics: Methods and explorations* (Studies in variation, contacts and change in English 19). 2017.

Ihalainen, O. The dialects of England since 1776. In R. W. Burchfield (ed.). *The Cambridge history of the English language*: *Origins and development*. Cambridge University Press (Cambridge, New York). 1994: 197–274.

Jenset, G. B. and McGillivray, B. *Quantitative historical linguistics*: *A corpus framework* (Oxford studies in diachronic and historical linguistics; 26). Oxford University Press (Oxford). 2017.

EPD. Jones, D., Roach, P., Setter, J. and Esling, J. H. *Cambridge English pronouncing dictionary.* 18th edn. Cambridge University Press (Cambridge, New York). 2011.

Lin, Y., Michel, J., Aiden, E. L., Orwant, J., Brockman, W. and Petrov, S. Syntactic annotations for the Google Books Ngram corpus. In *Proceedings of the 50th annual meeting of the Association for Computational Linguistics*. 2012: 169–174.

Mair, C. Tracking ongoing grammatical change and recent diversification in present-day standard English: The complementary role of small and large corpora. In A. Renouf and A. Kehoe (eds.). *The changing face of corpus linguistics* (Language and computers: Studies in practical linguistics no. 55). Rodopi (Amsterdam, New York). 2006: 355–376.

Mair, C. Using 'small' corpora to document ongoing grammatical change. In M. Krug and J. Schlüter (eds.). *Research methods in language variation and change*. Cambridge University Press (Cambridge, New York). 2013: 181–194.

Michel, J., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., and Aiden, E. L. Quantitative analysis of culture using millions of digitized books. *Science*. 2010; 331(6014): 176–182.

Minkova, D. *A historical phonology of English* (Edinburgh textbooks on the English language. Advanced). Edinburgh University Press (Edinburgh). 2014.

OED. Oxford University Press. *OED Online*. 2019. https://dictionary.oed.com (3 December, 2018).

Pike, W. A., Stasko, J., Chang, R. and O'Connell, T. A. The science of interaction. *Information visualization*. 2009; 8(4): 263–274.

Pope, M. K. *From Latin to modern French with especial consideration of Anglo-Norman*: *Phonology and morphology*. 2nd edn. Manchester University Press (Manchester). 1973.

Schlüter, J. A small word of great interest: The allomorphy of the indefinite article as a diagnostic of sound change from the sixteenth to nineteenth centuries. In N. Ritt, H. Schendl, C. Dalton-Puffer and D. Kastoysky (eds.). *Medieval English and its heritage*: *Structure, meaning and mechanisms of change* (Studies in English medieval language and literature 16). Lang (Frankfurt am Main). 2006: 37–59.
Schlüter, J. Consonant or 'vowel'? A diachronic study of the status of initial ‹h› from early Middle English to nineteenth-century English. In D. Minkova (ed.). *Phonological weakness in English*: *From Old to Present-Day English* (Palgrave studies in language history and language change). Palgrave Macmillan (Houndmills, Basingstoke, Hampshire & New York). 2009.
Schlüter, J. Using historical literature databases as corpora. In M. Krug and J. Schlüter (eds.). *Research methods in language variation and change*. Cambridge University Press (Cambridge, New York). 2013: 119–135.
Schlüter, J. Tracing the (re-)emergence of onset consonants through 500 years of books: Big data on a detail of historical English phonetics and phonology. *Folia Linguistica Historica*. 2019; 40.
Sönning, L. and Schlüter, J. Comparing strengths and limitations of linguistic corpora and big data resources: Evidence for the pronunciation of initial *h* in the BNC, COCA and Google Books Ngrams. In O. Schützler and J. Schlüter (eds.). *Data and methods in corpus linguistics*: *Comparative approaches*. Cambridge University Press (Cambridge). to appear.
Schützler, O. and Schlüter, J. (eds.). *Data and methods in corpus linguistics*: *Comparative approaches*. Cambridge University Press (Cambridge). to appear.
LPD. Wells, J. *Longman pronunciation dictionary.* 3rd edn. Pearson Longman (Harlow). 2008.
Wells, J. C. *Accents of English I*: *An introduction*. Cambridge University Press (Cambridge). 1982.