6 Using historical literature databases as corpora

JULIA SCHLÜTER

1 Introduction

The present chapter introduces a set of historical literature collections (available on CD-ROM or online) and their use as historical corpora for linguistic research. Despite the fact that the evolution of the English language is documented in a considerable body of written texts and is remarkably well represented in historical corpora, studies of earlier stages of English often suffer from a serious lack of data. Indeed, for many quantitative questions, the field of historical linguistics is hindered by the limits of electronically stored, computer-readable material.

As a backdrop to the present chapter, the most important diachronic corpora of English will be used and compared with the literature databases (Section 2). Issues of the representativeness of fictional writing with regard to other historical registers of writing in English will also be addressed. As a next step, a few technical tips on the computer-assisted exploitation of literature collections will be given (Section 3). To illustrate their use as corpora, three example studies from widely disparate areas will be outlined, thereby aligning data from standard diachronic corpora with such from the literature databases under discussion (Section 4). In the conclusion, the advantages and disadvantages of their use as corpora will be summarized (Section 5).

2 Comparison with historical reference corpora

Since historical literature databases can be used to supplement the purpose-built corpora available to and employed by linguists, some comparative facts and figures are of interest here.

2.1 Historical reference corpora

From the variety of historical reference corpora, three (groups of) corpora have been picked that are roughly comparable in their division into

https://doi.org/10.1017/CBO9780511792519.009 Published online by Cambridge University Press

diachronic parts and date ranges. The first two, the *Helsinki Corpus* (HC) and the set of *Penn Parsed Corpora of Historical English*, which are partly derived from the former, are available to individuals and institutions at relatively modest prices of a few hundred USD/GBP. The third one, *A Representative Corpus of Historical English Registers* (ARCHER), is still under construction and not yet generally available.

Table 6.1. *Details of three exemplary historical reference corpora*

Corpus name	Date range: Number of words	Genres	Tagging	Software
Helsinki Corpus (HC)	OE –1150: 413,250 ME 1150–1500: 608,570 EModE 1500–1710: 551,000 Total: 1,572,820	law document, handbook, science, philosophy, homily, sermon, rule, religious treatise, preface/ epilogue, history, travelogue, biography, fiction, romance, Bible, diary, drama, educational treatise, letters, proceedings	plain text	conventional concordancers (e.g. Oxford Concordance Program, Wordcruncher, Lexa, Wordsmith)
Penn Parsed Corpora of Historical English	*Penn-Helsinki Parsed Corpus of Middle English* (PPCME2) 1150–1500: 1,155,965 (based on HC ME) *Penn-Helsinki Parsed Corpus of Early Modern English* (PPCEME) 1500–1710: 1,794,010 (based on HC EModE) *Penn Parsed Corpus of Modern British English* (PPCMBE) 1700–1914: 948,895 Total: 3,898,870	law document, handbook, science, philosophy, homily, sermon, rule, religious treatise, preface/epilogue, history, travelogue, biography, fiction, romance, Bible, diary, drama, educational treatise, letters, proceedings	plain text, POS-tagging, syntactic parsing	plain text version usable with conventional concordancers; distributed with CorpusSearch 2 for using POS-tagged and parsed versions

Table 6.1. (*cont.*)

Corpus name	Date range: Number of words	Genres	Tagging	Software
A Representative Corpus of Historical English Registers (version 3.1) (ARCHER-3.1)[(a)]	BrE 1650–99: 180,189 1700–49: 177,726 1750–99: 178,675 1800–49: 180,793 1850–99: 181,026 1900–49: 176,907 1950–90: 178,241 AmE 1750–99: 180,268 1850–99: 176,707 1950–90: 178,777 Total: 1,789,309	drama, fiction, sermons, journal/diaries, legal, medicine, news, science, letters	plain text	conventional concordancers (e.g. Oxford Concordance Program, Wordcruncher, Lexa, Wordsmith)

[(a)] Pending its completion, use of ARCHER is limited to users at the participating institutions (see full database reference). At the time of publication, ARCHER-3.2 is under way, which will total 3,298,080 words and include POS-tagging and syntactic parsing.

As can be seen from Table 6.1, the corpora have a more or less fine-grained subdivision into diachronic subsections. The finer subdivisions of the HC and the *Penn* corpora are omitted for lack of space (see 'Online resources' in the Further Reading section for more detailed information). Each corpus also represents a variety of different text types or genres, some of which (in particular private letters, journals/diaries, court proceedings, drama and sermons) share certain characteristics of spoken language. A major asset of the *Penn Parsed Corpora of Historical English* is that they are enriched with part-of-speech (POS) tagging as well as syntactic parsing. Thus, they allow for corpus searches aimed at whole word classes rather than lexical instantiations of these and at specific syntactic structures. For these functions, the corpora come along with special search software (CorpusSearch). The plain text versions of all three (groups of) corpora are, however, accessible with any of the commonly used concordancing programs, thereby offering users convenient facilities for searching, sorting and storing their data. Finally, since corpus size is often a critical issue, the table indicates the number of words in the corpus subsections, which are in the order of one to four million words per corpus (group) for the entire time spans covered.

2.2 Historical literature databases

In today's world of increased access to information of all kinds, improved data storage and processing facilities and global information flow, linguists have become accustomed to the availability of huge datasets and

https://doi.org/10.1017/CBO9780511792519.009 Published online by Cambridge University Press

refined possibilities for analysis (see Kretzschmar, Chapter 3, Hoffmann, Chapter 10, and Smith and Seoane, Chapter 11, this volume). In historical linguistics, the increasing need for data is most difficult to satisfy since sources are limited, hard to obtain and laborious to transform into computer-readable format. A promising way out of this quandary is the use of historical literature databases produced by the commercial provider Chadwyck-Healey and distributed by ProQuest. These collections cover fiction, drama and poetry from the sixteenth century onwards and are thus of major interest to students of literature as well as linguists (who will generally be less interested in versified drama and poetry). Among the most suitable for linguists are the six collections listed in Table 6.2.

The databases can be purchased on CD-ROM and/or as yearly subscriptions. The prices for permanent acquisition range from a few thousand to around 20 thousand USD/GBP, with variable pricing conditions dependent on license type, country, acquisition of database packages etc. The considerable cost is doubtless the main reason why the databases have not made it into many libraries or

Table 6.2. *Details of six historical literature databases*

Database name	Texts	Date range: Number of words	Genres	Software
Early English Prose Fiction (EEPF)	211	1518–1700: 9,562,865[(a)]	fiction	KWIC enabled
Eighteenth-Century Fiction (ECF)	96 (– 3)[(b)]	1705–1780: 11,206,534 (– 1,503,835)	fiction	KWIC enabled
Nineteenth-Century Fiction (NCF)	250 (– 1)[(c)]	1782–1903: 37,589,837 (– 78,110)	fiction	KWIC enabled
English Prose Drama (EPD)	1651	1540–1700: 6,751,673 1701–1780: 6,334,892 1781–1903: 12,916,935 1904–1965: 413,740 Total: 26,417,240	drama (only prose)	KWIC enabled
Early American Fiction (EAF)	567	1789–1875: 34,634,666	fiction	
American Drama (AD)	1558	1714–1915: 22,027,683	drama (verse/ prose)	

[(a)] These dates disguise the fact that the first decades are sparsely represented in the database.
[(b)] Three works figure twice in ECF: Jonathan Swift's *Gulliver's Travels* is contained in the Motte edition of 1726 and in the Faulkner edition of 1735; Samuel Richardson's *Pamela* in the 1st edition of 1741 and in the 6th edition of 1742; the same author's *Clarissa* in the 1st edition of 1748 and in the 3rd edition of 1751. It is suggested that only the earlier editions should be included in a linguistic analysis.
[(c)] Mary Wollstonecraft Shelley's *Frankenstein* is contained both in the original edition of 1818 and in a corrected and revised edition of 1831.

linguistics departments; nonetheless, their editorial accuracy and quality significantly exceed that of less costly or freely accessible resources.[1]

The databases introduced in this chapter include four British and two American ones. There is one British and one American collection of dramatic texts. Of the former, the verse and prose parts (which will be the focus here) can be purchased separately; the latter contains an option to restrict searches to verse or prose drama only. The other databases all represent fictional prose of various sub-genres. The three British collections of fiction are chronologically arranged, so that EEPF covers the Early Modern English era (sixteenth and seventeenth centuries), ECF covers the larger part of the eighteenth century, and NCF focuses on the nineteenth century. The drama collections, in particular, have a more extended coverage. Therefore, the EPD database has, for current purposes, been subdivided into periods matching those of EEPF, ECF and NCF, but the user can specify any date range s/he wants for any search in all the databases.

One major advantage of the databases over the reference corpora is immediately apparent from the word counts: Even the smallest databases (EEPF and ECF) contain around 10 million words, with the largest (NCF) almost reaching 40 million. What is more, these fiction databases can be supplemented by drama databases if the amount of data is crucial, thereby adding another 6 to 13 million words to the three British databases and even more to the American one.

The major disadvantage is also obvious: While the compilation of linguistic corpora aims at maximizing the number of text types sampled, the literary databases represent only two genres and contain full texts rather than balanced samples. This fact has to be kept in mind when evaluating results. It is, however, useful to know that in Biber and Finegan's (1989) study, fictional prose from the seventeenth to twentieth centuries turned out to occupy a fairly middle ground between essays and letters on the continuum from literate to oral styles (while all three genres tended to drift towards the oral extreme in the course of time). Moreover, fictional prose data can be usefully compared with dramatic prose, which exemplifies language that has been written to be spoken and can be assumed to be imitative of contemporary spoken usage, at least to a certain extent.

Another disadvantage of the databases is that their contents do not come as simple text files, but that they include their own search interfaces, which are primarily tailored to the needs of literary scholarship (some of them allow users to search for keywords in the title of a work, for authors, sub-genres, publishers or for characters within a play). This precludes certain amenities that linguistic concordancing tools offer. However, upon request, ProQuest is generally prepared to provide raw data (in XML-coded format) that can be made accessible to concordancers (see Section 3.3 below).

[1] Websites hosting literary texts that are freely downloadable and fully searchable include the Oxford Text Archive (http://ota.ahds.ac.uk), Project Gutenberg (www.gutenberg.org/wiki/Main_Page) and ManyBooks (http://manybooks.net).

3 Technical tips

Using the literature databases is largely self-explanatory and in line with ordinary search interfaces. Thus, only some details will be mentioned here that are of particular interest to linguists.

3.1 Search syntax

The most important field in the *Standard Search* window, illustrated in Figure 6.1, is the *Keyword*. By clicking on the downward arrow on its right, an alphabetical keyword browse list opens, which allows the retrieval and selection of variant forms of a word, at the same time indicating the number of occurrences of a particular spelling in the database. Two or more keywords can be connected by the Boolean operators *and*, *or* and *and not*.

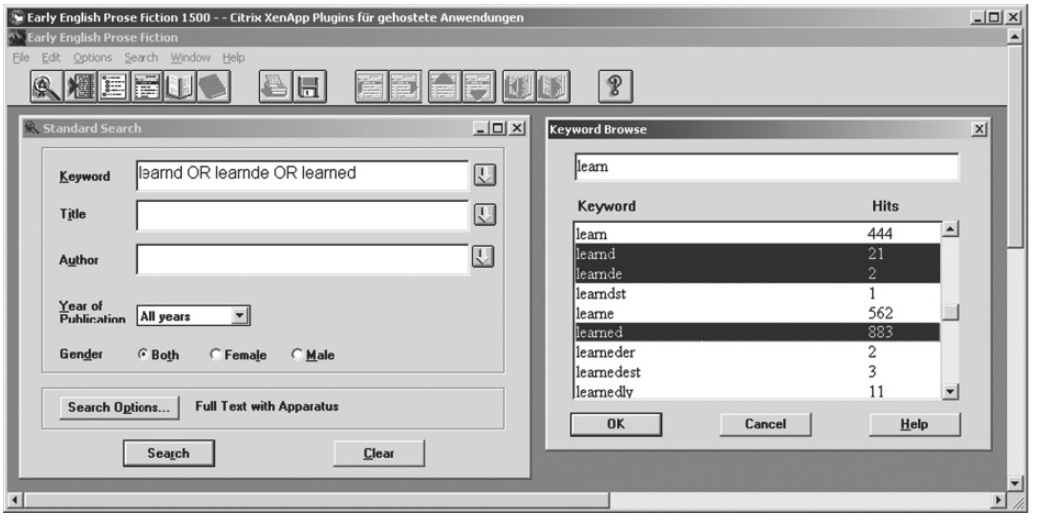

Figure 6.1. *Standard Search window and Keyword Browse window in the EEPF database*

It is possible to execute proximity searches by entering two keywords and defining the maximum distance between them. The proximity operators have to be typed into the keyword field, as follows: *[keyword 1] within N words of/after/before [keyword 2]*, for instance: *learned within 3 words before man* or *person within 9 words after learnt*.

Due to their function as operators, the following stop words cannot be searched: *after, and, before, cont, containing, directly, in, inside, not, of, or, with, within, word, words*. However, when enclosed in double quotes, they can be included, e.g. *person "of" quality* or *"person of quality"*.

Orthographic variants can be searched in several ways: Square brackets enclose alternative characters, e.g. *v[ie]rtue* or *p[iy]racy*. The wildcard *?*

https://doi.org/10.1017/CBO9780511792519.009 Published online by Cambridge University Press

represents any character, thus *s?ng* finds *sing*, *sang*, *sung* and *song*. The wildcard * represents any number of characters or, when surrounded by spaces, any word, e.g. *person** finds *person*, *persons*, *personal*, *personally*, *personification*, etc. and *person * quality* finds *person of quality*, *person and quality*, etc.

Further useful functions of the search interface, partly depending on the particular database, include restriction to a certain date range, nationality or ethnicity of the author or to male or female authors (but note that the earlier data include a significant share of anonymous authors).

3.2 Displaying, saving and sorting results

After a search has been carried out through the integrated search interface, the (*Brief*) *Summary of Matches* is displayed. Unfortunately, only the British databases, which are marked with *KWIC* (*Key Word in Context*) in Table 6.2, offer the option of viewing all hits in one window, similarly to linguistic concordancers. In both types of databases, processing the hits is moderately to extremely laborious within the customary interface. The 'KWIC enabled' databases, however, offer a convenient bypass: it is possible to select all relevant works at a time and to view all matches in context, as illustrated in Figure 6.2. From the *Context of Matches* window, it is possible to save the entire concordance or selected entries in a text file, which facilitates further processing. When transformed into a table (in text processing or spreadsheet software), examples can be deleted, categorized and sorted similarly to the facilities offered by concordancing software. In the American databases, which do not enable KWIC display, it is merely possible to enter the full text display and jump from one hit to the next. Unfortunately, this version of the search interface offers no option for saving matches in context.

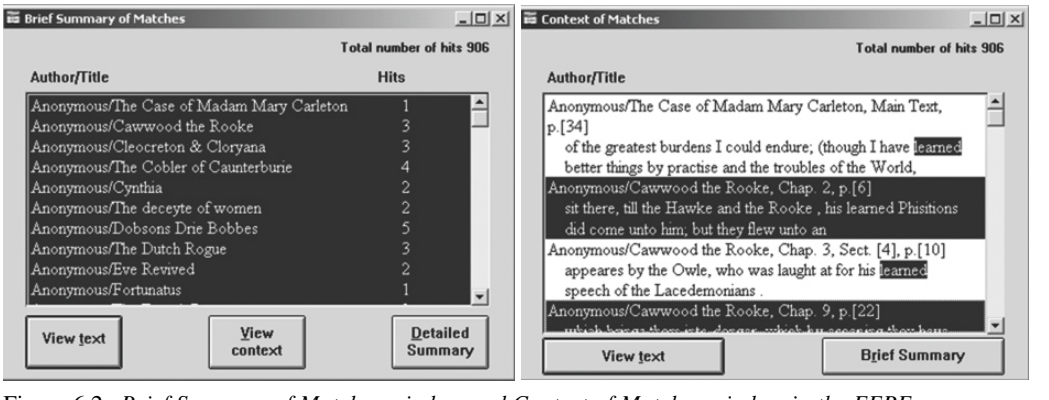

Figure 6.2. *Brief Summary of Matches window and Context of Matches window in the EEPF database*

https://doi.org/10.1017/CBO9780511792519.009 Published online by Cambridge University Press

3.3 Using raw data

As already mentioned, it is possible to obtain the raw data of the literature collections from the provider at no extra cost. These data basically come as one comprehensive XML-coded text file or, depending on the database, as one such file per author. Thus, some time and effort has to be expended dividing this into separate files corresponding to the individual works included in the database, removing unwanted XML tags and converting the files into plain text. This can, however, be partly automatized using a programming language or the appropriate functions of linguistic tools (the Wordsmith package, for instance, offers the Splitter and Text Converter tools). The advantages of this conversion are obvious: the resulting plain text files can be accessed with linguistic concordancers, which allow for easy searching, displaying, categorizing, deleting, (re-)sorting, thinning, saving, printing etc.

When selecting text files for searches, the structure of the file name will be of particular interest. One possible way of coding the most important information within a few characters was chosen in a research project on *Determinants of Grammatical Variation in English* at the University of Paderborn, Germany:[2] the file name *6400112f.689* from the EEPF database, for instance, codes the following information: The first three digits represent the year of birth of the author (omitting the initial *1*); the next two digits indicate that s/he is the first (or only) author born in this year; the next two count the number of the work by this author included in the database; the letter *f* indicates that the author is female, and the three digits of the extension represent the year of publication of the work (again, omitting the initial *1*). The full bibliographical information appears in the header section of each of the split-up files; in this example, it is Aphra Behn (1640–89): *The Lucky Mistake: A New Novel* (1689). As a consequence of this coding, the *Choose Texts* function of Wordsmith, for instance, allows one to sort files according to their names (= the year of birth of their author) or their file name extensions (= the year of publication of the work), so that it is easy to search within predefined date ranges.

4 Example analyses

To illustrate the possibilities opened up by the use of literature databases in addition to standard reference corpora, this section briefly sketches three case studies illustrating three different levels of linguistic description and phenomena from distinct frequency ranges. The results from three of the linguistic corpora in Table 6.1 will be compared with those from the literature databases in Table 6.2. For convenience, the raw data versions of the databases

[2] Thanks are due to the German Research Foundation (DFG) for funding this project (grant RO 2271/1–3), of which I was a member from 2000 to 2006.

have been used, and searches have been run through Wordsmith's *Concord* function. Table 6.3 summarizes the three case studies. Due to limitations of space, readers will be referred to the relevant literature for further details.

Table 6.3. *Survey of the three case studies sketched in Sections 4.1–4.3*

Example	reintroduction of initial <h>	variant inflection of past tense and past participle forms	restrictions on negated attributive adjectives
Area	phonology	morphology	syntax
Frequency	high	intermediate	low
Corpora/ databases	PPCEME vs. EEPF	ARCHER vs. EEPF, ECF, NCF	PPCMBE vs. ECF, NCF, EPD

4.1 The reintroduction of initial <h>

As shown in Schlüter (2009a, 2000b), the pronunciation of the initial letter <h>, which had become virtually mute in early Middle English, was reintroduced in a slow and differentiated process of phonological change beginning in late Middle English. The progress of the change can be traced by comparing the choice of variant forms of determiners before <h> and fully fledged consonants and vowels. One example of such a determiner is the first person possessive pronoun *min(e)*, which shed its final <n> in unstressed (i.e. prenominal) position. Figure 6.3 illustrates the two simultaneous, but independent, developments on the basis of the Middle and Early Modern English parts of the *Penn Parsed Corpora of Historical English*.

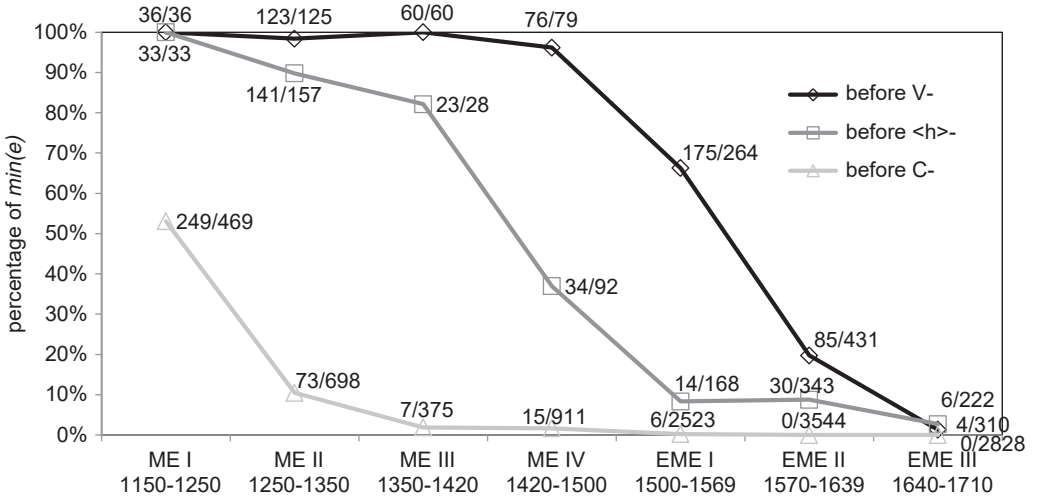

Figure 6.3. *The distribution of* min(e) *and* my *in prenominal position as a function of the initial sound of the following word in PPCME2 and PPCEME*

https://doi.org/10.1017/CBO9780511792519.009 Published online by Cambridge University Press

Besides the increasing replacement of *min(e)* by *my* (which, predictably, was faster before consonants than before vowels), Figure 6.3 shows that <h> behaved more like vowels up to the subperiod ME III (1350–1420), after which it gradually adopted a more consonant-like behavior, i.e. its articulation and perception were strengthened. The *Penn Corpora* are fully sufficient to document this changeover, as can be seen from the absolute number of hits retrieved for each subperiod indicated for each data point (for instance, 249/469 means that 249 out of a total of 469 examples contained *min(e)*). Indeed, the data from the much larger EEPF database, replacing the PPCEME data in the right-hand half of Figure 6.4, paint a very similar picture.

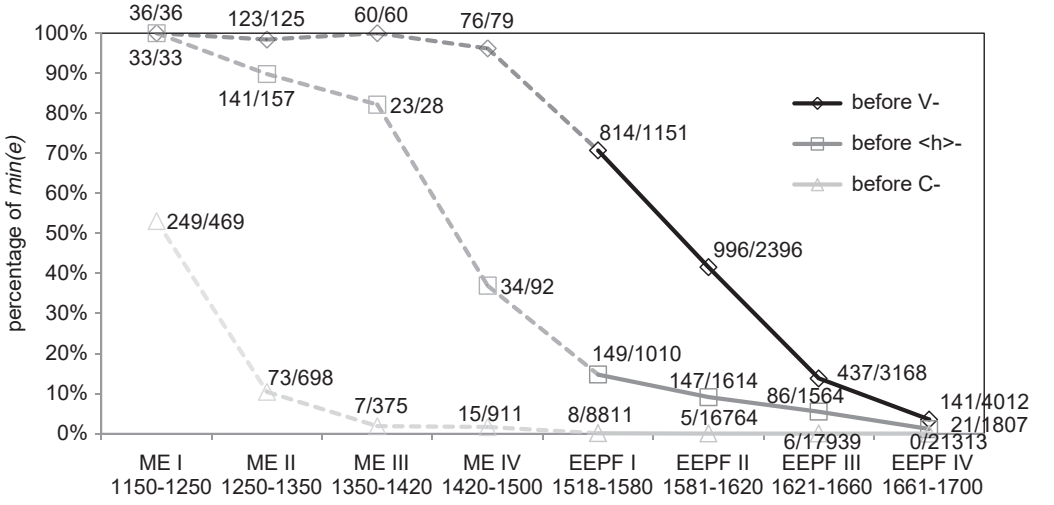

Figure 6.4. *The distribution of* min(e) *and* my *in prenominal position as a function of the initial sound of the following word in (PPCME2 and) EEPF*

Thus, one might conclude that a linguistic study derives no advantage from the more onerous study of the literature database. Prenominal *min(e)*/*my* and <h>-initial lexemes are, after all, high-frequency phenomena. However, as Schlüter (2009a) has shown, a large number of hits for initial <h> allows for a much more fine-grained analysis. In effect, the realization strength of <h> depends on factors such as the etymological source of the word (e.g. *my house*, but *mine host*), the amount of stress on its initial syllable (e.g. *my history* but *mine historic victory*), its overall textual frequency (e.g. *my hypocrisy* but *mine host*) and some others. These can only be isolated if the number of examples is statistically sufficient. Besides, the close parallels between the multi-genre corpus PPCEME and the single-genre database EEPF suggest that fictional prose is an acceptable representative of written usage generally.[3]

[3] The different division into subperiods is due to the fact that the EEPF data are taken from Schlüter (2009b), which employed four subperiods of 40–60 years, while the PPCEME uses three subperiods of 70 years each.

https://doi.org/10.1017/CBO9780511792519.009 Published online by Cambridge University Press

4.2 Variant inflection of past tense and past participle forms

The second case study concerns a group of verbs that have variable inflections for the past tense and the past participle (namely, *burn*, *dwell*, *learn*, *smell*, *spell*, *dream*, *kneel*, *lean*, *leap*, *spill* and *spoil*): They can take either the regular *-ed* or the irregular *-t* inflection. This group of verbs has been investigated in present-day databases (Levin 2009) as well as in ARCHER (Hundt 2009a: 24–27). Yet, their history and current trends have remained somewhat obscure, which is doubtless owed to the insufficient size of the databases. Take, for instance, the data displayed in Figure 6.5, which are based on the British part of ARCHER. The survey is limited to the more frequent among the verbs under consideration and charts the share of irregular *-t* forms (e.g. *burnt*) against the sum of irregular plus regular forms.

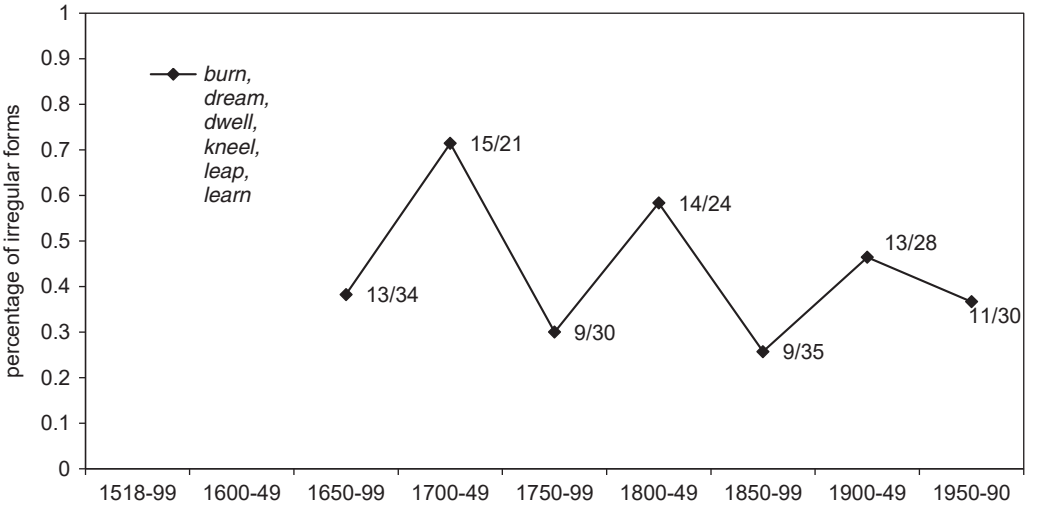

Figure 6.5. *The distribution of regular and irregular past tense and past participle forms of the verbs* burn, dream, dwell, kneel, leap *and* learn *in ARCHER (BrE only)*

On the basis of similar (though less restricted and manually checked) results from *ARCHER*, Hundt (2009a) concludes that we are witnessing a regularization process here, but Figure 6.5 shows that there is in fact more of a zigzag movement than a recognizable trend. A critical look at absolute numbers of examples suggests the reason: even when treated as a group, the verbs occur too rarely to warrant reliable quantitative conclusions. Figure 6.6, which is based on the EEPF, ECF and NCF databases in chronological sequence, fills this gap and at the same time keeps individual verbs separate. For easier comparison, the subperiods are matched with those of ARCHER.

As can be seen from this dataset, there is actually no point in treating these six verbs as a homogeneous group if one aims to unearth diachronic trends. Individual verbs show widely discrepant tendencies, which only a data-rich study can disentangle. Only *burn* and *dream* develop in parallel, though at

https://doi.org/10.1017/CBO9780511792519.009 Published online by Cambridge University Press

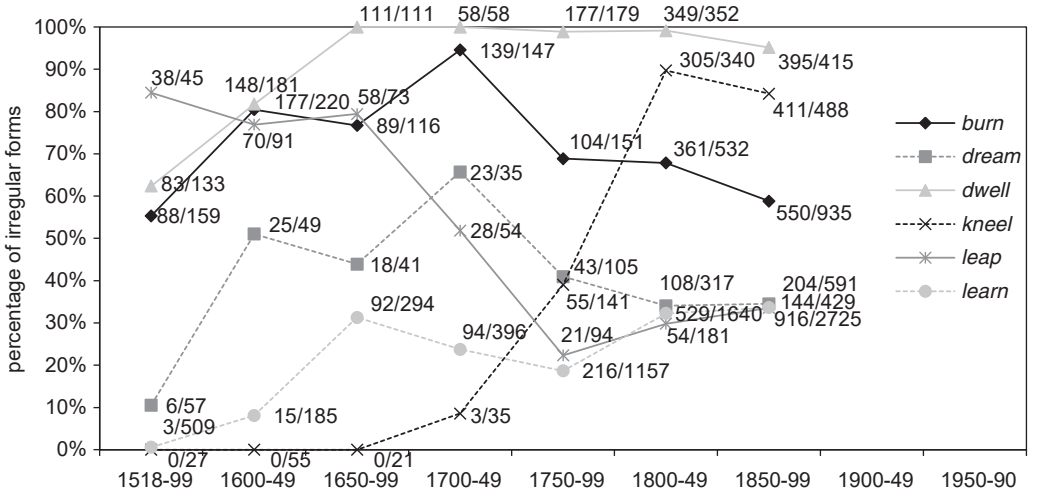

Figure 6.6. *The distribution of regular and irregular past tense and past participle forms of the verbs* burn, dream, dwell, kneel, leap *and* learn *in EEPF, ECF and NCF*

different levels, showing irregularization followed by regularization, with a turning point in the first half of the eighteenth century. In the case of *dwell* and *kneel*, we can observe irregularization rather than regularization, starting at very different points in time. *Leap* is a truly exceptional case, with the early prevailing irregular form being gradually ousted by the regular one. The inflection of *learn*, finally, has not undergone much change since 1650.[4] What is more, trends for each verb are fairly reliable since the number of examples is sufficient even in the earlier subperiods.

It can thus be shown that a graph like the one in Figure 6.5 is no more than an artefact of the unpredictable frequencies of occurrence of individual members of the group of *-ed*/*-t* verbs, blurred by extremely heterogeneous developments characterizing each verb. In addition, the most frequently occurring member(s) of the group (in this case, *learn*) tend(s) to distort the overall picture. To sum up, in the case of a genre-independent, mid-frequency phenomenon such as the variant inflection of the past tense and past participle forms in question, a large (set of) database(s) not only prevents false conclusions but also affords a much more informative picture than does a standard reference corpus.

4.3 Restrictions on negated attributive adjectives

The third and last case study follows up an observation by Bolinger (1980) about why *a not happy person* is generally judged unacceptable, while *a*

[4] Note, however, that among these raw data there are many instances of the invariant participal adjective *learned* meaning 'highly educated', as in *learned gentleman*, which would have to be excluded from a more rigorous analysis.

https://doi.org/10.1017/CBO9780511792519.009 Published online by Cambridge University Press

not unhappy person and *a not very happy person* are both acceptable. His intuition, according to which rhythmic preferences play a major role, has been confirmed on the basis of a collection of British newspapers from recent years examined in Schlüter (2005: 129–143). The historical dimension of the phenomenon has so far not been covered, but a legitimate question might be whether the same preferences played a role in earlier centuries. Replicating Schlüter's query (*a*/*the* immediately followed by *not*) on the entire PPCMBE yields merely four examples, which can only serve to illustrate the rhythmic difficulty, but not to substantiate it:

(1) There was not one, they said, like a Nurse of the not modern Schools. (nightingale-189X)

(2) It must have been a not uncommon experience of all of us that after severe and unwonted muscular effort general tremor of the muscles has set in, . . . (poore-1876)

(3) . . . it is necessary, by some means or other, to disabuse them of a not unnatural delusion, much encouraged by commentators, that . . . (benson-1908)

(4) . . . I have unwittingly passed by upon the roadside a not very noticeable country house, . . . (bradley-1905)

Example (1) will probably strike the reader as rather jarring, at least when quoted out of context. Following Bolinger's hunch, this is largely due to the adjacency of two stressed syllables in *nót módern*.[5] Examples (2) and (3) are unproblematic because the adjectives negated by *not* both lack initial stress. Although *nóticeable* in example (4) is initially stressed, the unaccented intervening adverb *very* steps into the breach to avert a threatening stress clash. Thus, attributive adjectives (at least in Present-Day English) can be negated by *not* if either they carry no initial stress or a semantically weak and unaccented adverb intervenes as a buffer. However, to ascertain this rule for the eighteenth and nineteenth centuries, the data from the PPCMBE are totally inadequate. (Note that, when negated by *never*, no similar restrictions should apply to initially stressed attributive adjectives, because *néver* has a second, stressless syllable that helps to prevent a stress clash. The PPCMBE contains no more than two examples, which happen to point in this direction: *a néver-fáiling Remedy* and *a néver-fáiling resource*.)

In view of the low frequency of negated attributive adjectives, it seems advisable to use plays dating from 1701 to 1903 from the drama collection EPD as a supplement to ECF and NCF. The results of the search for *a*/*the* followed by *not* and *never*, after exclusion of irrelevant hits, are shown in Figure 6.7.

[5] The context of this example is a discussion of modern schools of nursing, so that the negator *not* in this sentence may actually carry a strong contrastive stress, compared to which *modern* is given information and relatively unstressed.

https://doi.org/10.1017/CBO9780511792519.009 Published online by Cambridge University Press

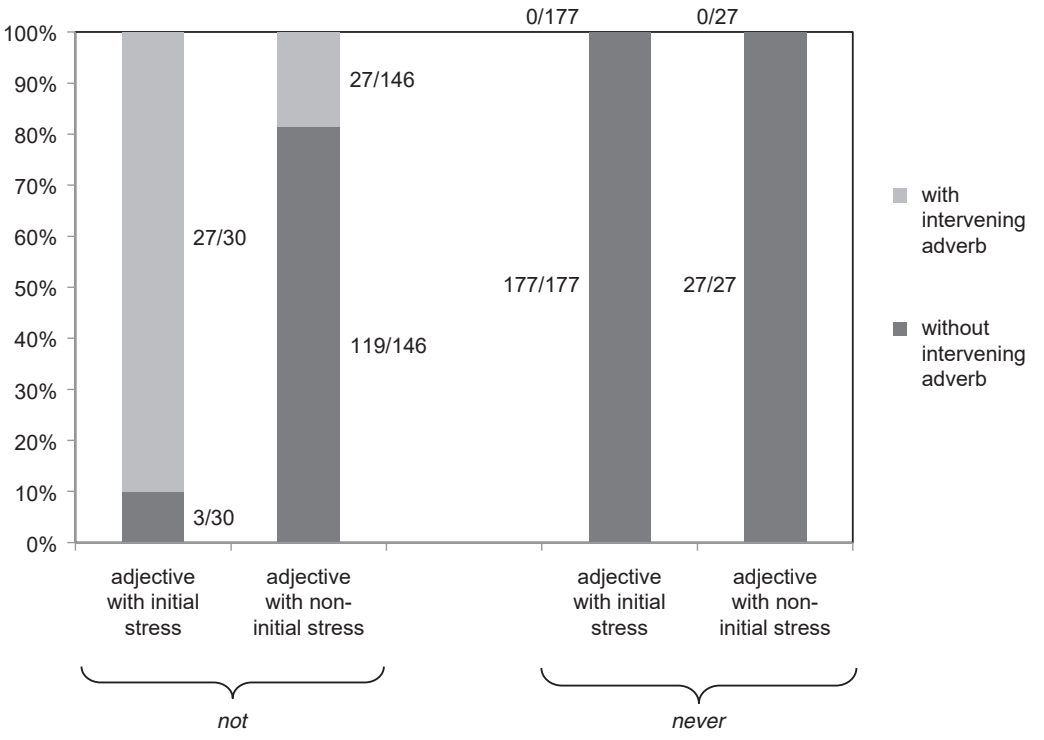

Figure 6.7. *The occurrence of adverbs intervening between the negators* not *and* never *and attributive adjectives in ECF, NCF and EPD (1701–1903)*

Despite the size of the database (over 68 million words), the data are far from ample, but the striking contrast between initially and non-initially stressed adjectives negated by *not* is statistically highly significant ($\chi^2 = 58.83$; df = 1; $p = 1.03 \cdot 10^{-14}$): initially stressed attributive adjectives seldom occur without an intervening buffer element. In contrast, a buffer is only rarely inserted before non-initially stressed ones. Thus, Bolinger's explanation in terms of stress clash avoidance receives strong support from eighteenth and nineteenth century data (and in fact is a potential linguistic universal, which should not be subject to diachronic change). The results for negation with *never* support the rhythmic account, considering that initially stressed adjectives show completely different behaviour here.

It remains to be added that, if we were to include data from the American literature databases EAF and AD for the eighteenth and nineteenth centuries, we would find at least 77 additional examples involving *not* plus 160 involving *never*. Thus, the considerable amount of material contained in the literature databases is just enough to shed light on a low-frequency syntactic construction such as the negation of attributive adjectives.

https://doi.org/10.1017/CBO9780511792519.009 Published online by Cambridge University Press

5 Conclusion

Drawing on the three case studies sketched above, the concluding part of this chapter summarizes the benefits and limitations of using literature databases as corpora.

5.1 Disadvantages

The one major disadvantage of the databases is their enormous price. Buying several databases at a time reduces the overall cost considerably, and linguistics departments interested in the purchase should ask for a contribution from literature departments, but when compared to ordinary linguistic corpora, the commercial interests of the provider remain an incontrovertible obstacle that will remain in place as long as Chadwyck-Healey/ProQuest retain their monopoly over professionally edited collections of literature.

Another setback compared to POS-tagged and parsed corpora such as those from the *Penn* family is, of course, the absence of such additional grammatical information. Admittedly, such markup can facilitate morphological and syntactic analyses, for instance when looking for past tense and past participle forms or for attributive adjectives. However, in many cases getting around such problems takes only a little ingenuity (e.g. searching for a representative set of inflected forms or entering determiners as part of the target expression).

In contrast to well-balanced linguistic corpora, the databases introduced here represent only one or two genres of written texts, and their authors all come from relatively privileged social classes. Thus, comparisons between genres are not possible, apart from those between fictional prose and prose drama, and sociolinguistic differences are difficult to discover, apart from those between female and male authors. Yet, as the study by Biber and Finegan (1989) and the juxtaposition of parallel data from PPCEME and EEPF in Section 3.1 have shown, fiction can be taken as a good representative of written language, since it is located somewhere between the most literate and the most oral styles. What is more, previous studies have indicated that historical drama may anticipate developments that also manifest themselves in present-day corpora of spoken language (see Schlüter 2005: 112–124, 195–196). Thus, by comparing fictional prose with dramatic prose, we can at least get an impression of potential divergences between written and spoken usage of the day.

Finally, since the selection of works for inclusion was based on literary rather than linguistic criteria, the smaller the database (or subsection of a database) that is used, the more likely the data are to be biased towards individual writers. For instance, the smallest among the databases considered, ECF, contains only 93 works (discounting three double editions), of which 7 are by Penelope Aubin, 8 by Daniel Defoe, and as many as 15 by Eliza F. Haywood. What is worse, 2.3 of the 8.7 million words come from four works by Samuel Richardson alone. This imbalance is most problematic in ECF, but less extreme in the other collections.

https://doi.org/10.1017/CBO9780511792519.009 Published online by Cambridge University Press

5.2 Advantages

Since some of the disadvantages of using literature databases instead of corpora do not, after all, seem overly prohibitive, the advantages clearly dominate. If the collections are purchased rather than only subscribed to and the raw data are made accessible to concordancing software, they can be searched and processed in the same comfortable way as linguistic corpora.

Their overriding asset is without any doubt their considerable size. Thus, in cases where corpora yield too little data or would force analysts to construct groups of items that perhaps show no homogeneous behaviour at all, the databases can still provide statistically viable results. Thanks to their size, they allow for refined analyses of the idiosyncrasies of individual items within a group and for fine-grained chronological subdivisions. What is more, in addition to observing diachronic developments from one period to another, sufficiently large sets of results enable users to analyse in great detail the influence of language-internal factors (such as effects of the presence or absence of stress, etymological distinctions, lexical frequency and many more).

In a nutshell, with the large quantities of data made available in the literature databases, it is hoped that historical linguistics can achieve the same depth and quality of analysis as has become the norm in the study of Present-Day English.

Using historical literature databases as corpora

Pros and potentials	Cons and caveats
• raw data are available and can be accessed with linguistic concordancing tools	• high retail price of literature databases
• considerable size of datasets allows for detailed investigations of frequent and statistically solid studies of rare structures	• absence of POS-tagging and syntactic parsing
• density of data is sufficient for fine-grained chronological or other subdivisions	• databases represent only one or two text genres
	• imbalance in favour of authors of great literary importance

Further reading

Biber, Douglas and Finegan, Edward 1989. 'Drift and the evolution of English style: a history of three genres', *Language* 65: 487-517.

Lindquist, Hans 2009. *Corpus linguistics and the description of English*. Edinburgh University Press.

Lüdeling, Anke and Kytö, Merja (eds.) 2009. *Corpus linguistics: an international handbook*. 2 volumes. Berlin and New York: Mouton de Gruyter.

McEnery, Tony, Xiao, Richard and Tono, Yukio 2006. *Corpus-based language studies: an advanced resource book*. London etc.: Routledge.

Schlüter, Julia 2005. *Rhythmic grammar: the influence of rhythm on grammatical variation and change in English*. (Topics in English Linguistics 46.) Berlin and New York: Mouton de Gruyter.

Useful online resources

ARCHER homepage: www.alc.manchester.ac.uk/subjects/lel/research/projects/archer

Chadwyck-Healey literature databases: http://collections.chadwyck.co.uk/marketing/list_of_all.jsp

Helsinki Corpus: www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus

Penn Parsed Corpora of Historical English: www.ling.upenn.edu/hist-corpora

https://doi.org/10.1017/CBO9780511792519.009 Published online by Cambridge University Press