CHAPTER 2

Corpus and Methodology

2.1 Introduction

This chapter deals with three areas: the corpora, the genres analysed, and the methodology. It opens with a thorough description and characterisation of the main corpora used as a source of data for the present work (2.2), namely ARCHER (2.2.1), the OBC (2.2.2), and the ICE-GB (2.2.3), and a brief discussion of secondary sources (2.2.4). Given that ARCHER is the primary corpus used, and the source for the majority of examples, a discussion is also given of its suitability (2.2.1.1) and limitations (2.2.1.2). Section 2.3 presents a detailed account of the genres analysed for the study, concluding with a reclassification of these on the speech–written and the formal–informal continua (2.3.8). The chapter closes with a description of the methodology employed for the extraction of examples from the various corpora, including information about the software used, systems for data storage and classification, and methods for statistical analyses (2.4).

2.2 The Corpora

2.2.1 ARCHER: A Representative Corpus of Historical English Registers

Although numerous corpora have been consulted for this study, most of the data are in fact drawn from ARCHER, the OBC, and the ICE-GB, the former being the principal source. ARCHER is a multi-genre historical corpus of British and American English covering the period 1650 to 1999. It was initially created by Douglas Biber and Edward Finegan (see Biber et al. 1994a, 1994b) and since then has been used for a large number of studies, with several additional versions of the corpus having been created (Yáñez-Bouza 2011). From December 2008 a consortium of fourteen
universities¹ from seven countries around the world worked on the compilation of a new unified and extended version of the corpus, which was finally released in 2013.² Unfortunately, the new version was not available for my own research, so the data here are drawn from version 3.1 (2006), which totals almost 1.8 million words distributed over 955 files. The British component in this version is divided into seven fifty-year subperiods, with eight different genres represented (drama, fiction, journals/diaries, letters, medicine, news, science, and sermons), these amounting to a total of more than 1.2 million words distributed over 674 files (Table 2.1)³.

The American component also includes eight genres, but is divided into three fifty-year subperiods, these corresponding to the second halves of the eighteenth, nineteenth, and twentieth centuries, and amounts to a total of just over half a million words (Table 2.2).

According to Biber et al. (1994a: 3), the genres in ARCHER can be classified into two major groups in terms of two criteria: written vs. speech-based genres and formal vs. informal genres. However, clear-cut distinctions are difficult to make; for this reason, the defining parameters with which texts are usually characterised (e.g. formal vs. informal) are better described as dimensions on a continuum rather than poles (see Biber 1988: 9). As Biber and Finegan (1989: 488) note, dimensions are continuous parameters of variation, such that each parameter comprises a group of co-occurring linguistic features […] identified empirically […].

Table 2.1 Number of words of British English across genres in ARCHER

<table>
<thead>
<tr>
<th>Period</th>
<th>Drama</th>
<th>Fiction</th>
<th>Letters</th>
<th>Journals and diaries</th>
<th>Medicine</th>
<th>News</th>
<th>Science</th>
<th>Sermons</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1650–99</td>
<td>26,648</td>
<td>41,112</td>
<td>12,659</td>
<td>21,374</td>
<td>23,117</td>
<td>22,292</td>
<td>21,441</td>
<td>11,146</td>
<td>180,189</td>
</tr>
<tr>
<td>1700–49</td>
<td>25,177</td>
<td>44,021</td>
<td>12,093</td>
<td>21,443</td>
<td>21,936</td>
<td>21,612</td>
<td>20,780</td>
<td>10,664</td>
<td>177,726</td>
</tr>
<tr>
<td>1750–99</td>
<td>23,962</td>
<td>45,056</td>
<td>12,091</td>
<td>21,843</td>
<td>21,003</td>
<td>23,087</td>
<td>20,465</td>
<td>11,068</td>
<td>178,675</td>
</tr>
<tr>
<td>1800–49</td>
<td>26,267</td>
<td>44,946</td>
<td>12,576</td>
<td>21,740</td>
<td>20,278</td>
<td>22,903</td>
<td>20,994</td>
<td>11,089</td>
<td>180,793</td>
</tr>
<tr>
<td>1850–99</td>
<td>24,679</td>
<td>43,289</td>
<td>10,705</td>
<td>22,686</td>
<td>22,143</td>
<td>23,066</td>
<td>21,715</td>
<td>10,953</td>
<td>181,026</td>
</tr>
<tr>
<td>1900–49</td>
<td>23,048</td>
<td>45,274</td>
<td>12,434</td>
<td>22,066</td>
<td>20,204</td>
<td>21,975</td>
<td>21,337</td>
<td>10,569</td>
<td>176,907</td>
</tr>
<tr>
<td>1950–90</td>
<td>24,450</td>
<td>45,095</td>
<td>11,259</td>
<td>22,225</td>
<td>20,794</td>
<td>22,920</td>
<td>21,308</td>
<td>10,190</td>
<td>178,241</td>
</tr>
<tr>
<td>Total</td>
<td>176,021</td>
<td>309,193</td>
<td>83,817</td>
<td>153,777</td>
<td>149,475</td>
<td>157,855</td>
<td>148,140</td>
<td>75,679</td>
<td>1,253,557</td>
</tr>
</tbody>
</table>

¹ Northern Arizona, Southern California, Michigan (USA), Freiburg, Heidelberg, Bamberg, Trier (Germany), Helsinki (Finland), Uppsala (Sweden), Manchester, Lancaster, Salford (United Kingdom), Zürich (Switzerland), and Santiago de Compostela (Spain).
² The new 3.2 version starts as early as 1600 for drama, fiction, and legal opinions and contains a wider variety of genres (see Yáñez-Bouza 2011).
³ Note that, unlike the American section, the British section includes texts dated only until 1990.
linguistic variation in English; rather, a multidimensional analysis is required. Dimensions have both linguistic and functional content. The linguistic content is defined by a group of linguistic features (such as nouns, attributive adjectives, and prepositional phrases) that co-occur with a markedly high frequency in texts. On the assumption that co-occurrence reflects shared function, these co-occurrence patterns are interpreted functionally. Each dimension thus characterises the situational, social, and cognitive functions most widely shared by the co-occurring linguistic features.\(^4\)

Within the group of written registers, the corpus contains samples of ‘personal styles of communication (journals/diaries and personal letters), prose fiction, popular exposition represented by news reportage, and specialist expository registers, represented by [...] medical prose, and scientific prose’ (Biber 2001: 94). The corpus also includes various speech-based registers, such as dramatic and fictional dialogues, these as reflections of casual face-to-face conversation, and sermons, which serve as examples of planned monologue styles (see Biber 2001: 94, Biber & Finegan 1997: 255–7, Biber et al. 1994a: 3).

Regarding the formal vs. informal dimension, journals/diaries, letters, together with dramatic and fictional dialogues, occupy the informal end of the continuum, whereas the formal end is represented by medicine, science, and sermons. In turn, news and fictional prose stand halfway between the two poles of the formal–informal continuum.

The variety of genres or text types\(^5\) over three and a half centuries provided by ARCHER allows us to investigate the ‘diachronic relations among oral

\(^4\) Biber (1988) distinguishes five dimensions: (1) Involved vs. Informational Production; (2) Narrative vs. Non-Narrative Discourse; (3) Situation-Dependent vs. Elaborated Reference; (4) Overt Expression of Argumentation; (5) Non-Impersonal vs. Impersonal Style. For a complete description of the linguistic features which make up each of the dimensions see, among others, Biber (1988, 1995, 2001, 2003a) and Biber and Finegan (1989, 1997).

\(^5\) The terms ‘genre’ and ‘text type’ will be used interchangeably to refer to the various types of textual materials contained in the corpus. The distinction is noted, for example, in Biber (1989: 5–6) and Taavitsainen (2001: 140), who use the term ‘genre’ to identify language-external criteria, whereas

### Table 2.2. Number of words of American English across genres in ARCHER

<table>
<thead>
<tr>
<th>Period</th>
<th>Drama</th>
<th>Fiction</th>
<th>Letters</th>
<th>Journals and diaries</th>
<th>Medicine</th>
<th>News</th>
<th>Science</th>
<th>Sermons</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1850–99</td>
<td>24,214</td>
<td>44,224</td>
<td>11,253</td>
<td>22,534</td>
<td>20,424</td>
<td>21,992</td>
<td>21,326</td>
<td>10,740</td>
<td>176,707</td>
</tr>
<tr>
<td>1950–99</td>
<td>23,810</td>
<td>44,214</td>
<td>11,611</td>
<td>22,131</td>
<td>22,473</td>
<td>23,072</td>
<td>21,343</td>
<td>10,123</td>
<td>178,777</td>
</tr>
<tr>
<td>Total</td>
<td>75,355</td>
<td>130,855</td>
<td>33,920</td>
<td>66,774</td>
<td>66,330</td>
<td>67,335</td>
<td>63,333</td>
<td>31,850</td>
<td>535,752</td>
</tr>
</tbody>
</table>
and literate registers based on the framework of multi-dimensional analysis, whereby distributional patterns among linguistic features are investigated’ (Yáñez-Bouza 2011: 206), as has been successfully demonstrated in previous research (see, among others, Biber 1988, 2001, Biber & Finegan 1988, 1989, 1992, 1994, 1997, 2001). In other words, ARCHER ‘was designed for a specific major research agenda: to analyse historical change in the range of written and speech-based registers of English from 1650 to the present’ (Biber & Finegan 1997: 255).

2.2.1.1 Suitability of ARCHER
ARCHER constitutes a very useful resource for the diachronic study of linguistic phenomena thanks to both the ample time span covered and the wide selection of genres it contains, which allow comparisons between written registers and those varieties closer to the spoken language.

As mentioned in Chapter 1, much of the work on the historical development of phrasal verbs has focused on the OE, ME, and EModE periods and on contemporary English, whereas the LModE period and Twentieth-century English have received far less attention. With this in mind, ARCHER is particularly useful, in that it covers my target time span, LModE and Twentieth-century English, as well as a fifty-year subperiod corresponding to the late part of EModE (1650–99), thus allowing for comparisons with earlier stages of the language. In addition, given that PDE phrasal verbs tend to be associated with informal and colloquial styles, the range of text types included in ARCHER allows for both synchronic and diachronic cross-genre comparisons that can shed light on the status of phrasal verbs in previous stages of the language. This task is undoubtedly facilitated by the numerous multivariate analyses that Biber and associates have carried out with ARCHER which have shown that text types vary over time and that the linguistic conventions of a particular genre in PDE do not necessarily apply to earlier periods of the language (see, among others, Biber 1988, 1995, 2001, Biber & Finegan 1988, 1989, 1992, 1997, 2001).

‘text type’ refers to purely linguistic features. Although there have been various attempts to clarify the distinction between the two categories (see, e.g., Lee 2001), no consensus has been reached in the literature about the terms in the English tradition. For further discussion, see, among others, Biber (1988), Jucker (2000), Lee (2001), and Görlach (2004).

6 Thim (2006a) argues that in the period 1500–1700 the use of phrasal verbs in a particular text type is not motivated by its degree of formality, but rather by the content of the texts herein, which may encourage the use of phrasal verbs in conveying certain meanings. This question will be addressed in more detail in Chapter 6.
2.2.1.2 *Corpus Limitations*

Although ARCHER 3.1 constitutes a very useful resource for the diachronic study of linguistic change, the corpus also has a number of limitations which must be borne in mind.

The first and most significant drawback here relates to the time gaps in the American component of the corpus (see Table 2.2). The primary aim of my research is to investigate the diachronic evolution of phrasal verbs in the recent history of English. As noted above, the American English sub-corpus covers only three fifty-year subperiods (1750–99, 1850–99, and 1950–99), and thus any analysis of the evolution of phrasal verbs in this variety would necessarily be incomplete. For this reason, my focus has been on the British section of the corpus. A diachronic comparison between the two major standard varieties of English is certainly worth considering, especially for a phenomenon such as phrasal verbs, which have been often described as typical of American English (see Konishi 1958: 124, 127, Live 1965: 429, Traugott 1972: 173, Pelli 1976: 43, McArthur 1989: 42).

The second major problem of ARCHER relates to the quality of the texts included. Although ARCHER 3.1 intends to be a revised and unified version of previous (slightly different) iterations of the corpus (Yáñez-Bouza 2011), a number of inconsistencies and errors remain. For example, whereas editorial notes in some of the texts are included in caret brackets (< >), others have been annotated with square brackets ([ ]). Similarly, dialogues in some of the fictional texts have been included within brackets, whereas in others no distinction has been made between narration and dialogue. These inconsistencies are, however, minor drawbacks for the study of phrasal verbs, especially if one is aware of them. One further problem, though, is indeed significant for the present study. It concerns the repetition of a fragment of a scientific text from the period 1650–99: text 3 of the file 1674an01.s2b, which contains 126 words. This is in fact a fragment of the file 1676coxe.s2b. Such duplication is relevant here since it contains an instance of the phrasal verb *make out*. It would

---

7 Note, however, that Martin (1990: 100) finds no evidence of such a difference when comparing personal letters written in both varieties. Similarly, no difference is found by Wild (2010) in ARCHER, 'where the frequency of phrasal verbs in British and American English is very similar in each period' (2010: 228). Martin (1990: 178) provides a tentative explanation for the widely held view that phrasal verbs are more common in American than in British English. For her, the difference between the two varieties is a qualitative one, since American English personal letters make use of aspectual combinations slightly more often than British English ones. Given that aspectual combinations are often perceived as redundant, they become particularly noticeable, thus giving the impression that phrasal combinations are more common in American English.

8 This repetition has been amended in ARCHER 3.2.
of course have been possible to remove these 126 words from my data to avoid any possible skewing of the statistical analysis. However, since version 3.1 of ARCHER is a fixed release, and one which has been used extensively throughout the world with all its pros and cons, I decided to keep it as such. For this reason, the instance below is repeated in my database of examples.

(1) The former part of this Position may be thus made out. (1676coxe.s2b/1674ano1.s2b)

The third limitation of ARCHER has to do with the genres it contains. The number of text types is sufficient to carry out a complete diachronic analysis of any linguistic feature. However, given that phrasal verbs have often been associated with the colloquial, spoken language, it seems necessary to include an additional genre in which the spoken language is best represented. Although the importance of speech in the development of languages is self-evident (Rissanen 1989), the language historian must often rely on evidence from written texts or texts which either represent (e.g. dramatic dialogues) or approximate to the oral language (e.g. diaries, sermons). Some of the genres in ARCHER show varying degrees of speech-likeness (diaries, drama, letters, sermons, and, to some extent, journals), yet none can be said to represent the spoken language of the time per se, and for this reason I also analysed the use of phrasal verbs in the trial proceedings of the OBC (see Section 2.2.2).

One additional drawback which relates not only to ARCHER but to any linguistic corpus is conveniently summarised in Matti Rissanen’s phrase ‘God’s true fallacy’ (1989: 17), and implies that corpora tend to create the erroneous impression that they are accurate reflections of the entire reality of the language that they represent. Moreover, when working with historical corpora, one must also consider that the data are preserved only randomly and that they are ‘only indirectly related to everyday spoken communication’ (Nevalainen 1999b: 499). This has been described by Labov (1994: 11) as the ‘bad-data problem’, relating to the fact that the linguist can only rely on written data to draw conclusions, and thus can only aim at ‘making the best use of bad data’ (Labov 1994: 11). This will be my concern in the following sections.

2.2.2 The Old Bailey Corpus

The OBC is a 14-million-word corpus based on the *Proceedings of the Old Bailey*, London’s central criminal court, which were published between
2.2 The Corpora

Since the proceedings were taken down in shorthand by scribes in the courtroom, the corpus texts can be taken as the most faithful representation of the spoken language of the time available to us (see Section 2.3.7).

Given the huge size of the OBC, the results presented in this study are based on the analysis of a sample of the corpus. Since my intention was to draw comparisons with the genres in ARCHER, I selected a number of files and organised them into four different periods which approximate to the fifty-year divisions in ARCHER (see Table 2.3).

The aim was to establish around 100,000 words of direct speech per subperiod, but given that the last subperiod was longer than the others, the target number of words for it was 150,000. This process was made easier because the corpus is XML-tagged, so that only the parts tagged as <speech> were considered in the selection of files, that is, those which reproduce the actual words of those people involved in trials. As shown in

---

Table 2.3 Number of files and words per subperiod in the OBC sample

<table>
<thead>
<tr>
<th>Period</th>
<th>Filename</th>
<th>Number of words</th>
<th>Year</th>
<th>Total words per subperiod</th>
</tr>
</thead>
<tbody>
<tr>
<td>1720–1749</td>
<td>OBC-17250407</td>
<td>7,150</td>
<td>1725</td>
<td>105,453</td>
</tr>
<tr>
<td>1720–1749</td>
<td>OBC-17250630</td>
<td>6,273</td>
<td>1725</td>
<td>1725</td>
</tr>
<tr>
<td>1720–1749</td>
<td>OBC-17250827</td>
<td>8,277</td>
<td>1725</td>
<td>1725</td>
</tr>
<tr>
<td>1720–1749</td>
<td>OBC-17260420</td>
<td>8,695</td>
<td>1726</td>
<td></td>
</tr>
<tr>
<td>1720–1749</td>
<td>OBC-17300116</td>
<td>6,624</td>
<td>1730</td>
<td></td>
</tr>
<tr>
<td>1720–1749</td>
<td>OBC-17320906</td>
<td>29,529</td>
<td>1732</td>
<td></td>
</tr>
<tr>
<td>1720–1749</td>
<td>OBC-17400116</td>
<td>25,833</td>
<td>1745</td>
<td></td>
</tr>
<tr>
<td>1720–1749</td>
<td>OBC-17471014</td>
<td>25,833</td>
<td>1747</td>
<td></td>
</tr>
<tr>
<td>1750–1799</td>
<td>OBC-17520218</td>
<td>14,206</td>
<td>1752</td>
<td>101,729</td>
</tr>
<tr>
<td>1750–1799</td>
<td>OBC-17620210</td>
<td>14,493</td>
<td>1762</td>
<td></td>
</tr>
<tr>
<td>1750–1799</td>
<td>OBC-17780603</td>
<td>18,887</td>
<td>1778</td>
<td></td>
</tr>
<tr>
<td>1750–1799</td>
<td>OBC-17900224</td>
<td>54,143</td>
<td>1790</td>
<td></td>
</tr>
<tr>
<td>1800–1849</td>
<td>OBC-18020602</td>
<td>25,569</td>
<td>1802</td>
<td>100,853</td>
</tr>
<tr>
<td>1800–1849</td>
<td>OBC-18306615</td>
<td>35,577</td>
<td>1835</td>
<td></td>
</tr>
<tr>
<td>1800–1849</td>
<td>OBC-18430703</td>
<td>41,707</td>
<td>1843</td>
<td></td>
</tr>
<tr>
<td>1850–1913</td>
<td>OBC-18640229</td>
<td>44,510</td>
<td>1864</td>
<td>146,230</td>
</tr>
<tr>
<td>1850–1913</td>
<td>OBC-18930501</td>
<td>56,449</td>
<td>1893</td>
<td></td>
</tr>
<tr>
<td>1850–1913</td>
<td>OBC-19130304</td>
<td>45,271</td>
<td>1913</td>
<td></td>
</tr>
<tr>
<td>Total words</td>
<td></td>
<td>454,265</td>
<td></td>
<td>454,265</td>
</tr>
</tbody>
</table>

---

9 The most recent version 2.0 of the OBC contains a larger sample of texts (10 million words larger than the initial version). See https://fedora.clarin-d.uni-saarland.de/oldbailey/.
Table 2.3, the sample also includes texts from a wide variety of years, so that the data were equally distributed over the whole subperiod.

Since the proceedings are verbatim representations of the speech of the trial participants, they are, in the words of Magnus Huber, ‘arguably as near as we can get to the spoken word of the period’ (Huber 2007). They constitute an excellent source of comparison with the ARCHER texts, in which the spoken dimension is only represented by texts which show a certain degree of speech-likeness but which do not actually reproduce the literal words of the speakers. However, as a note of caution, we might add that ‘corpora including trial proceedings and studies based on such corpora have to take account of the fact that what looks like language variation and change may in fact be due to the influence of scribes and printers’ (Huber 2007). Further details about trial proceedings as a genre are provided in Section 2.3.7.

2.2.3 The International Corpus of English, Great Britain

Since the OBC does not contain data beyond 1913, in order to trace the evolution of phrasal verbs in the spoken language over the course of the twentieth century the legal cross-examinations of the British component of the ICE-GB were examined. The ICE-GB is a one-million-word corpus of written and spoken English from 1990 to 1993 inclusive. It contains 500 tagged and parsed texts of approximately 2,000 words each. Among the spoken texts, the corpus contains dialogues (public and private), monologues (scripted and unscripted), and mixed discourse; in turn, the written texts include printed (academic writing, non-academic writing, reportage, instructional writing, and creative writing) and non-printed (non-professional writing and correspondence) materials. The subgenre of legal cross-examinations is included among the public dialogues of the spoken section and contains ten texts which together total 21,352 words. Although colloquial, spoken language in PDE is probably best represented by casual face-to-face conversation, for my present purposes the legal cross-examinations of the ICE-GB provide a more reliable comparison to the data from the OBC, given that both record the direct testimony of participants in trials. It must be acknowledged, however, that my corpus data present a gap from 1913 to 1990 for this subgenre.

10 For further discussion about cross-genre differences, and the relation of genres to the spoken/written medium and their degree of (in)formality, see Section 2.3.
2.2 The Corpora

2.2.4 Further Sources of Data

The core sources for quantitative and qualitative data, then, were ARCHER and a sample of the OBC, the latter supplemented by data from the legal cross-examinations from the ICE-GB. However, other linguistic corpora and dictionaries have also been used. The HC, the CLMETEV, and the database of examples of the OED provided further evidence of particular uses of combinations, especially those which have developed idiomatic meanings from their literal ones (see Section 4.5.3.1). Both the HC (730–1710; 1,572,800 words) and the CLMETEV (1710–1920; 15 million words) are large multi-genre corpora and are very useful as sources of lexical items such as phrasal verbs. Likewise, the database of examples of the OED, which contains three million quotations over 1,000 years of English, proved to be of considerable use, not least due to its ‘sheer mass of material’ (Mair 2004: 124), which provides the linguist with ‘a wealth of useful information’ (Hoffmann 2004: 26). From the BNC, a 100-million-word multi-genre corpus covering the late part of the twentieth century, and also from the World Wide Web, I collected illustrative examples of contemporary uses of phrasal verbs. The immense amount of language information available in these two sources meant that they were especially useful in finding evidence of current uses of phrasal verbs, the web being particularly rich in the most colloquial and spontaneous forms. Finally, for Section 3.3.2.3, where I argue for the existence of a special type of particle which I have called ‘emphatic’, several other corpora were consulted in order to find diachronic and synchronic evidence of combinations with this particle, especially those which are not registered in dictionaries and which have more restricted use. To this end, the following additional corpora were occasionally consulted: A Standard Corpus of Present-Day Edited American English (BROWN), A Corpus of Contemporary American English

11 For a full description, see the respective entries at Corpus Resource Database (CoRD):
HC (www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/)
CLMETEV (www.helsinki.fi/varieng/CoRD/corpora/CLMETEV/)
12 For my purpose, I employed the online version available at http://corpus.byu.edu/bnc/. For a full description of the corpus, see the entry at CoRD: www.helsinki.fi/varieng/CoRD/corpora/BNC/.
13 For the use of the World Wide Web as a corpus, see, among others, Kilgarriff and Grefenstette (2003), Meyer et al. (2003), Hundt et al. (2007), Kilgarriff and Sharoff (2012). Growing interest in the web as a source of linguistic data can also be seen in the existence of regular conference series, such as the Web as a Corpus Workshop (www.sigwac.org.uk/), and the recent creation of several web-based corpora, such as the Corpus of Global Web-based English (GloWbE; see http://corpus.byu.edu/glowbe/).
Corpus and Methodology

1990–2010 (COCA), Corpus of Historical American English (COHA), Freiburg-LOB Corpus of British English (FLOB), Freiburg-Brown Corpus of American English (FROWN), and Lancaster-Oslo/Bergen Corpus (LOB). 14

2.3 A Note on the Genres

As noted in Section 2.2, ARCHER is divided into eight different text types, namely drama, fiction, journals/diaries, personal letters, medicine, news, science, and sermons. The corpus is described by its compilers as containing samples of personal styles of communication (journals/diaries and personal letters), prose fiction, popular exposition (news reportage), and specialist expository registers (medical and scientific prose). In addition, dramatic and fictional dialogues can be said to represent casual face-to-face conversation, whereas sermons are a reflection of planned monologue styles (see Biber et al. 1994a: 3, Biber & Finegan 1997: 255–7, Biber 2001: 94). In turn, the OBC contains trial proceedings whose sections of direct speech are described by their compilers as close representations of the spoken language of the time.

The genres in ARCHER can in principle be divided according to two main parameters, written vs. speech-based and formal vs. informal (see Biber et al. 1994a: 3), although in practice the defining parameters are better described as dimensions, because rather than representing poles they constitute a continuum (see Biber 1988: 9). Previous work has raised the possibility that the characteristics of the spoken language in earlier periods could be approximated through the analysis of ‘speech-based’ registers. Although there are no records of spoken registers from earlier stages of English, those genres situated towards the informal end of the continuum (diaries, drama, letters, and, to some extent, journals) can be said to be closer to the oral language, and thus they become the only available means for a historical linguist to approximate early spoken English. Moreover, since diaries, journals, and letters represent the most ‘personal styles of communication’ (Biber & Finegan 1997: 255), it is here that we must look for typically colloquial features such as phrasal verbs, because

14 For further information about these and other corpora, see their entries at CoRD:

BROWN (www.helsinki.fi/varieng/CoRD/corpora/BROWN/)
COCA (www.helsinki.fi/varieng/CoRD/corpora/COCA/basic.html)
COHA (www.helsinki.fi/varieng/CoRD/corpora/COHA/)
FLOB (www.helsinki.fi/varieng/CoRD/corpora/FLOB/)
FROWN (www.helsinki.fi/varieng/CoRD/corpora/FROWN/)
LOB (www.helsinki.fi/varieng/CoRD/corpora/LOB/)
it is in private, intimate contexts, as well as in close relationships, that ‘people are more likely to be at their ease and drop their guard, also linguistically’ (Tieken-Boon van Ostade 2005: 131). The distinction between written vs. speech-based genres, on the one hand, and formal vs. informal text types, on the other, seems clear and has often been used in historical corpus linguistics to characterise linguistic phenomena by extra-linguistic parameters. Nevertheless, a number of specifications in this respect seem in order before embarking on the analysis of phrasal verbs across genres in the corpora. These relate mainly to the changing conventions of text types over time, as well as to particular characteristics which concern primarily the texts included in ARCHER and the OBC.

### 2.3.1 Journals/Diaries

Journals and diaries present a series of common features which somehow justify their incorporation into one single category. As pointed out above, both represent personal styles of communication and are generally placed at the informal end of the continuum. Moreover, they are both non-interactive,\(^{15}\) they are usually written in the first person, and typically narrate private matters, such as routines and personal reflections, not generally intended for publication (see, among others, Nevalainen & Raumolin-Brunberg 1989, Biber et al. 1994a, 1994b, Biber & Finegan 1997, Biber 2001). Despite these similarities, a number of characteristics also set these two genres apart, as suggested by Nuria Yáñez-Bouza. In collecting the data for her doctoral dissertation (Yáñez-Bouza 2007; see also Yáñez-Bouza 2015a) she noted important differences in form and style between them, and therefore decided to split the two genres and carry out a pilot study to assess whether significant differences could be detected in the use of stranded prepositions. Her findings showed that in most cases there were statistically significant differences between them (see Yáñez-Bouza 2007: 333–41, 2015a: 145–7, 2016).

The splitting of diaries and journals in Yáñez-Bouza (2007) was carried out by following four main principles, which have also served as the basis for the classification of the two genres in version 3.2 of ARCHER. Yáñez-Bouza (2015b) applied Biber and Conrad’s (2009) framework for analysing register variation, focusing on the situational characteristics that describe

\(^{15}\) An exception is constituted by epistolary journals, with a real or imaginary addressee, which became popular during the LModE period; see, e.g., Fanny Burney’s diary (Tieken-Boon van Ostade 1991) and Sterne’s *A Sentimental Journey* (Oldireva Gustafsson 2002: 134).
Corpus and Methodology

registers, namely participants, the relation between participants, channel/mode, production circumstances, setting, communicative purpose(s), and topic. Following the original design of ARCHER, the first and foremost criterion is topic or subject matter, followed by its communicative purpose. Yáñez-Bouza (2015b) concludes that, although a number of parameters are shared by both diaries and journals, they also display important differences which favour their classification as two distinct registers.

As regards their topic or subject matter, diaries mostly deal with private affairs related to daily life, such as love relationships, personal worries, daily routines and tasks (breakfast, lunch, praying, visiting friends), and domestic activities (cooking, looking after children, going to the market). Journals, in turn, are mostly concerned with keeping documentary accounts of business, military, or exploratory journeys, and family archives (Oldireva Gustafsson 2002: 134). They are generally written by those whose profession required such documents to be kept, and for this reason their content is of a very different nature: travel journals, sea journals, war journals, and scientific and political journals. Moreover, as noted by Yáñez-Bouza (2015b), since journals are often commissioned, the objective/descriptive approach very commonly overtakes the subjective, first-person narration.

In terms of communicative purposes, most journals, and especially diaries, were kept for private purposes, and were not meant to be read or shared with others. However, it must be noted that the notion of privacy in earlier times (especially the LModE period) may differ from our present-day one, given that formerly ‘private space was [often] exposed to public control’ (Nevala 2004a: 273). Family letters, for instance, were commonly read out to friends and relatives in private gatherings (see Tieken-Boon van Ostade 2000: 447, 2005: 129). Moreover, although dealing with private matters, journals and less often diaries were occasionally written for publication; Captain James Cook’s sea journals are a clear example of this. Percy (1996) shows how Cook’s language becomes more ‘correct’ in the journals of his third voyage (1776–1779), probably influenced by his knowledge that his journals were destined for publication. Similarly, some diaries were written with an eye to publication, and are thus likely to show a more careful type of writing and vocabulary (see Tieken-Boon van Ostade 2009: 128). Agnes Porter, for example, sent her private diary to her former pupil, Lady Mary Talbot, in 1812 hoping that it would ‘amuse’ her (see Nurmi & Nevala 2010: 173).

In sum, although diaries and journals are related text types, in terms of Biber and Conrad’s (2009) framework, their analysis as two separate genres seems justified. For this reason, along the lines in Yáñez-Bouza
2.3 A Note on the Genres

(2007, 2015a) and thanks to her generous help, I decided to carry out the analysis of phrasal verbs in the two genres separately. Table 2.4 sets out the total number of files and words corresponding to the two text types after the splitting of journals and diaries. 16

The figures in this table bring to light one of the methodological problems stemming from the division of the genre journals/diaries with the files available in ARCHER 3.1, namely the fact that the distribution of text types in the corpus turns out to be quite irregular. 17 As can be seen, in some subperiods one of the text types is clearly overrepresented, with samples for the other being relatively scarce. This is the case, for example, with the first subperiod, in which there is one single diary (The Diary of the Revd Henry Newcome, 1661), as opposed to nine journals. The data for the diaries from this period are clearly biased by the particular idiolect of Revd Newcome, so that any conclusions derived from this analysis must

---

### Table 2.4 Total files and number of words of diaries and journals in ARCHER 3.1

<table>
<thead>
<tr>
<th></th>
<th>Diaries</th>
<th></th>
<th>Journals</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total files</td>
<td>Total words</td>
<td>Total files</td>
<td>Total words</td>
</tr>
<tr>
<td>1650–99</td>
<td>1</td>
<td>2,192</td>
<td>9</td>
<td>19,182</td>
</tr>
<tr>
<td>1700–49</td>
<td>6</td>
<td>12,475</td>
<td>4</td>
<td>8,968</td>
</tr>
<tr>
<td>1750–99</td>
<td>4</td>
<td>8,688</td>
<td>6</td>
<td>13,155</td>
</tr>
<tr>
<td>1800–49</td>
<td>5</td>
<td>10,667</td>
<td>5</td>
<td>11,073</td>
</tr>
<tr>
<td>1850–99</td>
<td>7</td>
<td>15,713</td>
<td>3</td>
<td>6,973</td>
</tr>
<tr>
<td>1900–49</td>
<td>4</td>
<td>8,971</td>
<td>6</td>
<td>13,095</td>
</tr>
<tr>
<td>1950–99</td>
<td>6</td>
<td>13,236</td>
<td>4</td>
<td>8,989</td>
</tr>
<tr>
<td>Total</td>
<td>33</td>
<td>71,942</td>
<td>37</td>
<td>81,435</td>
</tr>
</tbody>
</table>

---

16 The files in ARCHER 3.1 are named with the structure nnnnabcd.gpv, where nnnn corresponds to the year of publication, abcd to a four-letter string representing an abbreviation of the author’s name, and gpv refer to genre, period and variety respectively. The genre codes are as follows: d = drama, f = fiction, h = sermons, j = journals/diaries, m = medicine, n = news, s = science, x = letters. Periods are coded according to the formula: 2 = 1650–99, 3 = 1700–49, 4 = 1750–99, 5 = 1800–49, 6 = 1850–99, 7 = 1900–49, 8 = 1950–90. The variety variable is coded as b = British and a = American. Thus, for example, a file named 1692cong.f2b dates from 1692 and corresponds to the genre fiction, subperiod 1650–99 and British variety. As a result of the separation of diaries and journals into two categories for the present study, some modifications were necessary in the filenames for these two genres. Following the guidelines for the compilation of ARCHER 3.2 (see Yáñez-Bouza 2011), I replaced the genre code for diaries with a ‘y’ and preserved the original ‘j’ for journals.

17 Balanced word counts were one of the aims of version 3.2 of the corpus (see Yáñez-Bouza 2011).
be taken with caution. I will return to this issue in my discussion of the corpus findings (see Section 6.2.1).

2.3.2 Personal Letters

It is generally assumed that private letters contain linguistic features and patterns that often correspond to the informal and oral dimensions of language (see Biber 1988). As noted by Romaine (1998: 18), personal letters ‘are among the most involved and therefore oral of written genres’, with family correspondence being more informal and involved than those letters written to more distant acquaintances (Palander-Collin et al. 2009: 12). Biber and Finegan (1989, 1997) show that letters display characteristics typical of face-to-face conversation and spontaneous language. They are interactive in character and generally addressed to a specific individual whose temporal and physical surroundings are familiar to the writer, so that ‘a letter writer is free to refer directly to personal feelings and situations’ (Biber & Finegan 1989: 497). Similarly, Fitzmaurice points out that ‘[a]lthough the letter is patently not conversation on paper, epistolary discourse does imitate some of conversation’s characteristics’ (2002: 233).

Elspass (2012: 157), in turn, argues that letters, diaries, and other ego-documents are close to speech and can therefore be used to cast light on the history of natural languages. However, in analysing letters from earlier periods a number of observations seem in order. Although LModE letters might in principle seem to belong to the private sphere, during the eighteenth and nineteenth centuries they were very often read out to friends and relatives (see Tieken-Boon van Ostade 2000: 447, 2005: 129). As argued by Bell (1984: 161ff), and as shown by numerous studies (see, among others, Nevala 2004b, Auer 2015, Hernández-Campoy & Conde-Silvestre 2015, Conde-Silvestre 2016, Sairio 2009, Hernández-Campoy & García-Vidal 2018a), style variation may respond to an effort by the letter-writer to accommodate not only to the intended audience, but also to a non-intended one. In a recent study, Hernández-Campoy and García-Vidal (2018b) demonstrate that the social status of the addressee and the potential presence of third-person auditors were crucial in the language standardness in some LME letters. John Paston I, for example, used more exaggeratedly the prestigious th spelling (instead of the old runic þ) when addressing his wife, Margaret Paston, than with other addressees of the same social rank, in a probable attempt to converge with the scribes who would read aloud the letters to his wife. Moreover, LModE letters were the result of a very conscious process of writing, which followed a series
of conventions, typically acquired at home (see Austin 1998: 323), at grammar schools, or during apprenticeship (see Nevalainen 2001: 219–20) and, for this reason, they cannot be considered mere ‘thoughtless outpourings’ (Anderson & Ehrenpreis 1966: 273). These conventions were laid out in a vast number of letter-writing manuals which appeared as early as the EModE period (see Nevalainen 2001) and became very popular especially during the nineteenth century (Tieken-Boon van Ostade 2009: 2–3). Letter-writing manuals typically gathered a series of conventions related to grammar and forms of address, accompanied by model letters. One of these conventions was that letters should sound natural and spontaneous, just like a normal conversation (Tieken-Boon van Ostade 2009: 121). Despite the fact that LModE letters tend to follow conventions, it has been noted that ego-documents produced by semi-literate writers have ‘the highest potential to render authentic sources of historical orality’ (Elspass 2012: 159). Nevertheless, it must be borne in mind that literacy can affect a person’s writing habits to some extent, so that they cannot really be said to reflect everyday speech in a straightforward way (Montgomery 1995: 5, Schneider 2002: 76).

Letters have therefore evolved as a text type from the eighteenth century to the present day, as has been shown in a number of multivariate analyses by Biber (2001) and Biber and Finegan (1989, 1997), in which eighteenth-century letters were found to be ‘expository, descriptive, or argumentative in purpose’ as opposed to PDE letters, which are ‘personally involved and interactive’ (Biber 2001: 105). For this reason, many eighteenth-century letters ‘avoided the use of overtly speech-based features, such as contractions and discourse particles’ (Biber 2001: 104). This feature of eighteenth-century letters is possibly related to the fact that, as noted above, they were often written following a series of conventions, whereas other personal documents such as diaries and, to some extent, journals tend to follow a freer style.

Shifting tendencies in letter writing throughout the LModE period may have been prompted by external factors such as the increase in literacy,

---


19 Note, however, that the original spellings in ARCHER have been transcribed in various ways by the compilers of the corpus: in some texts the original spellings have been kept, whereas in others they have been included in caret brackets (see Section 2.2.1.2). For this reason, measuring the number of contractions of a text does not constitute a reliable form of ascertaining its speech-like character in this corpus. The correction of these and other undesirable practices in a historical corpus was indeed one of the main aims of ARCHER 3.2 (see Yáñez-Bouza 2011).
with which the writing of letters and indeed written production in
general ‘grew explosively’ (Tieken-Boon van Ostade 2009: 3, 8), as well
as the introduction of the Penny Post in Britain in 1840. With the latter,
postage became far cheaper (Tieken-Boon van Ostade 2009: 2) and this
favoured the use of letters as an accessible means of communication for all
social classes, even uneducated, lower-class people, whose language was
characterised by ‘colloquialisms’, ‘incorrectness’, and ‘old-fashionedness’
(McIntosh 1986: 12).

Other important factors that must be taken into account in dealing
with letters relate to the writer’s social background, as well as to his/her
relationship with the addressee (intimate/distant), since the latter ‘is likely
to cause linguistic register variation’ (Nurmi & Palander-Collin 2008: 21).
For this reason, letters are often characterised by heterogeneity and idio-
syncrasy (Görlach 2001: 52–3), as shown by their varying linguistic features
(see Nurmi & Palander-Collin 2008).

2.3.3 Drama and Fiction

Fiction differs from other text types in that it is ‘produced and edited
carefully and directed towards a large, specific but unbounded and
unindividuated audience […] for purposes of aesthetic enjoyment’ (Biber &
Finegan 1989: 495). Both drama and fiction are regarded by Biber as speech-
based registers, on the grounds that the dialogues in both text types are a
representation of casual face-to-face conversation (see Biber et al. 1994a: 3,
Biber & Finegan 1997: 255–7, Biber 2001: 94). However, such a claim must
be taken carefully, for one main reason: dialogue and narration, or asides
in the case of plays, are not separated in ARCHER 3.1. This means that the
corpus has no specific mark-up that distinguishes dialogue from narration,
and thus conclusions about the oral character of, for example, the genre
drama must be viewed with caution.

It would certainly be interesting to carry out an analysis of phrasal
verbs in dialogue and narration separately, among other things to ascer-
tain whether differences can be perceived in the use of these constructions
when comparing the written vs. the oral dimension. However, this has
not been attempted in the present work, where the findings for drama
and fiction make no distinction between the narrative sections and asides,
on the one hand, and fictional dialogue, on the other. These results can,

---

20 According to Görlach, ‘[p]rivate letters became a major text-type in the 18th century’ (2001: 211).
nevertheless, be compared safely with the multivariate analysis carried out by Biber and associates, in which this distinction is also disregarded.

Moreover, as noted by Yáñez-Bouza (2007: 342), another crucial factor that must be taken into account when dealing with the analysis of fictional texts over time is their heterogeneity, largely because of the changing linguistic conventions undergone by this genre during the EModE and LModE periods, especially in terms of their purpose and content (see especially McIntosh 1998).

### 2.3.4 News

Like fiction, news reportage represents ‘non-interactive written registers, carefully crafted for larger audiences’ (Biber & Finegan 1997: 269), also referred to as ‘popular written registers’.

Just like other text types, news reportage has undergone a number of changes in stylistic conventions over time. Newspapers first appeared in England in the seventeenth century (Percy 2012: 195). At the time newspapers differed from their modern counterparts in that, for example, they did not contain editorials, news reports, and commentaries, and were not divided into thematic sections (Nevalainen 2002: 67). The forerunners of newspapers were called ‘corantsos’ or ‘newsbooks’ depending on the size of the publication, and consisted basically of sequences of letters with collections of dispatches from correspondents (Jucker 2005: 11). In fact, early newspapers can be compared in many senses to early newsletters or the letters ‘written and published to communicate the news of the day (common in the later part of the seventeenth and beginning of the eighteenth century)’ (*OED s.v. newsletter*). For this reason, basic news items formerly bore a ‘structural resemblance to chronicle writing, a major genre with a long history of its own’ (Nevalainen 2002: 67). After the introduction of the printing press ‘mass communication [was made] possible, first with pamphlets, later with regular news sheets and then with daily newspapers’ (Jucker 2003: 130). Until then, ‘domestic news passed by word of mouth or private letter more quickly than it could be printed’ (Nevalainen 2002: 68).

The first newspaper in England was the *Oxford Gazette* of November 1665, which subsequently became *The London Gazette* (Claridge 2010: 600, Fries 2012: 1065). Newspapers began to gain popularity in the LModE period. According to Görlach (2001: 207), whereas by 1712 there were twelve London newspapers, the figure had increased to fifty-two by 1801. In 1855 the tax on newspapers was lifted and they became within reach of
the working classes (Tieken-Boon van Ostade 2009: 142). The first penny paper was *The Daily Telegraph (and Courier)*, which appeared in 1855 (Fries 2012: 1066). With the development of communication technology (mail services, telegraph) the amount of circulating news increased to an unprecedented scale. In the nineteenth century the role of journalists changed, in the sense that they were not only charged with passing on information but also with deciding which events were newsworthy (Jucker 2005: 12–13). This gave rise to the well-known top-down structure generally followed by news reporting in LModE times in which the most salient aspect of the story usually appears at the beginning of the text (Ungerer 2002, Jucker 2005: 13).

Apart from those crucial structural differences just discussed, the linguistic features of PDE newspapers differ dramatically from their earlier counterparts. In general, the texts in early news discourse took the form of narratives arranged in chronological order (Jucker 2005: 18, Fries 2012: 1070) and ‘remained more “informational” and less “interpersonal”’ (Percy 2012: 196; see also Claridge 2009: 92–3). According to Biber and Finegan (1997: 269), newspaper language evolved in a direction similar to that of fiction in that both ‘became popular registers, appealing to an increasingly wider readership across the centuries’ (1997: 269). In fact, Tieken-Boon van Ostade (2009: 8, 142–3) links the increasing demand for newspapers, novels, and magazines with the spread of literacy during the LModE period. However, whereas the change towards oral styles in fiction can be appreciated as early as the eighteenth century, the drift in newspaper language started later (Biber & Finegan 1997: 269). As observed by Mair (1997: 203), the colloquialisation process in news seems to have taken place only over the last thirty years of the twentieth century. Rühlemann and Hilpert (2017) also observe an increased colloquialisation of journalistic writing towards the last decades of the twentieth century. Cotter (2003), for example, accounts for the change in the use of pragmatic connectives in terms of a shift from a text-oriented to an audience-oriented mode of discourse. For her, the decreased use of temporal connectives and the increase in coordinating connectives in newspapers are directly related to ‘the trajectory from a more formal, distant, text-centered written norm to a more informal, interactional, audience-centered norm over time’ (2003: 67). In spite of this, it must be borne in mind that in general ‘daily newspaper

21 The top-down structure is no longer the prototypical format of news reporting in PDE, in which the large number of articles which start on the front page and continue elsewhere in the newspaper has given rise to the expression ‘jigsaw journalism’ (Jucker 2005: 15).
reporters and editors are highly aware of language and linguistic usage’ (Cotter 2003: 45) and that they usually ‘follow prescribed parameters of form’ (2003: 51). As a result, their language tends to be more formal in the sense that it follows the standard/ normative prescriptions of grammar. This is an important point of divergence with the newspapers of the nineteenth century, which, according to Görlach (1999: 146), underwent a considerable linguistic decay due to the pressures of the medium: texts had to be written quickly and published without major revision. Moreover, journalists were often criticised for lacking sufficient competence in English (Görlach 2001: 207). In fact, as noted by Tieken-Boon van Ostade, ‘[t]he negative term journalese was first recorded in the OED for the year 1882, as the opposite meaning to “plain English”’ (2009: 144). This difference between early and present-day newspaper language probably relates to a change in the way of producing mass media communication. As argued by Jucker (2009: 2–3), whereas in PDE it is practically impossible to ascribe the exact wording of an article to one single individual, since the text usually undergoes several editorial revisions, earlier news publications were dominated by single individuals and did not follow style manuals. 22

2.3.5 Scientific and Medical Discourse

According to Biber and Finegan (1997: 269), science and medical prose tend to use a more specialised language and are accessible only to a narrow audience. For this reason, they are described as specialist expository genres (1997: 255).

All the scientific texts and the earliest medical texts (1650–99) in ARCHER belong to the Philosophical Transactions, ‘the first longstanding scientific journal in England […] designed by the Royal Society members to discuss their current interests and share scientific news’ (Taavitsainen 2010: 32–3). Later medical texts are extracted from other specialised journals, but they all include medical reports of the type of those published in the Philosophical Transactions. 23

The Philosophical Transactions were first published in 1665 as a means for the members of the Royal Society ‘to communicate their scientific

22 Apparently, the first style guide was drawn up by The Times in 1913 (Claridge 2009: 92n).
23 ARCHER does not include other medical-related documents (e.g. recipes, regimens, or health guides) which are part of more specialised medical corpora, such as the Corpus of Early English Medical Writing (see Taavitsainen et al. 2005, 2014 Taavitsainen & Pahta 2010a) and The Málaga Corpus of Late Middle English Scientific Prose (see www.helsinki.fi/varieng/CoRD/corpora/SciProse/).
findings and opinions’ (Taalvitsainen 2010: 49). They represent the top
level of scientific discourse during the last decades of the seventeenth
century as opposed to the older commentary styles retained in other
writing (Taalvitsainen 2010: 50–1). In fact, most of the texts in the
Philosophical Transactions were dialogic in character, ‘often representing a
kind of extended interaction among researchers’ (Gray et al. 2011: 224),
probably due to the fact that many of the earlier publications were written
in the form of letters (Atkinson 1999: 81) in response to prior publications.
Moreover, as noted by Valle (1999: 110), the type of language adopted
by the Philosophical Transactions is better described as ‘plain style’ as
opposed to the so-called ‘rhetoric style’ replete with stylistic flourishes
and decorations irrelevant to the substance of the text. This is probably
related to the fact that, although the contributors to the journal were
highly educated specialists, the Philosophical Transactions were not only
designed as a channel of information for specialists, but also intended to
attract other educated and informed readers interested in the field (Gotti
Biber and Finegan (1997) have noted that medical and scientific texts
have evolved in a direction opposite to that of popular written genres, poss-
sibly due to a fundamental difference in purpose and readership. As they
put it, the ‘specialist registers have followed an essentially steady develop-
ment towards ever more “literate” styles’ (1997: 269; see also Biber & Clark
2002), becoming more informational and impersonal and acquiring more
elaborate forms of reference (1997: 262–9). Similar results are reported
by Atkinson (1996, 1999), who applied Biber’s (1988) multivariate ana-
lysis to texts of the Philosophical Transactions from 1675 to 1975. Atkinson
notes that medical and scientific discourse became less involved and more
informational over time, in accordance to a shift from author-centred to
object-centred style (1999: 76–7, 110ff). Whereas in the seventeenth and
eighteenth centuries articles were characterised by a prominent authorial
persona, probably due to the fact that they were often presented in epis-
tolary form, from the early nineteenth century onwards they became more
object-centred. Moreover, the articles in the Philosophical Transactions
seem to have evolved from a low level of narrative to an even lower one
(decrease in the use of the past tense, the perfect aspect, public verbs (e.g.
say, mention) and third-person pronouns) and from a low to an even
lower level of explicit reference (infrequent use of place, time, and other
adverbs). Atkinson also points out that scientific and medical writings
tend to become less persuasive (1999: 123ff) and more abstract (1999: 125ff)
over time. Abstractness here involves scientific discourse becoming highly passivised, probably related to the emergence of object-centred rhetoric; as observed by Oldireva Gustafsson, the passive is ‘an ideal grammatical form for discourse associated with objectivity and non involvement’ (2006: 110). Other typical features of scientific and medical writing include complex sentences, extended noun phrases (NPs) including nominalisations, and the low frequency of first- and second-person pronouns (Taaivtsainen & Pahta 2010b: 551).

### 2.3.6 Sermons

Other text types included in ARCHER, such as sermons, can be described as occupying an intermediate position between the oral and the written language (see Görlach 2001: 204). The *OED* defines a sermon as ‘a discourse, usually delivered from a pulpit and based upon a text of Scripture, for the purpose of giving religion instruction or exhortation’ (*OED* s.v. sermon 2a). Hence, sermons can be ‘oral, or written, or a mixture of both, making a clear characterisation of the genre as a whole very problematic’ (Claridge & Wilson 2002: 25). As noted by Biber, although sermons are addressed to a specific audience and intended to be spoken, they are essentially planned monologues (2001: 94). In a way, sermons are similar to political speeches in that their objective is usually one of persuasion (Görlach 2001: 204). Moreover, sermons belong to the religious discourse, which ‘has traditionally been a more formal, more conservative register than others, one slower to change and thus exhibiting more archaic features’ (Claridge & Wilson 2002: 25). Sermons are generally detached from ordinary, everyday language and often contain full biblical passages which are very formal in tone and include conservative linguistic features (see, e.g., Görlach 2001: 202). It must also be borne in mind that ‘priests are naturally good speakers, so that carefully preparing their sermons is the obvious thing for them to do’ (Claridge & Wilson 2002: 29). Therefore, from a stylistic point of view, sermons are better defined as formal, whereas in terms of register they are generally described as speech-based texts, although they do not really represent natural speech (Biber 2001: 94, Biber & Finegan 1997). Nevertheless, ‘these are some of the most “spoken-like” registers available from earlier historical periods, and as such they provide useful comparative data to the analysis of written registers’ (Biber et al. 1998: 252–3). A finer-grained distinction is made by Culpeper and Kytö (2010), who classify sermons as speech-purposed because, in contrast to plays, for example,
which are ‘designed to produce real-time spoken interaction’, sermons ‘are designed to produce monologue (they are “read out”)’ (2010: 17).

As with most genres, sermons have undergone changes in their stylistic conventions over time. By applying the multidimensional analysis developed by Biber (1988) to a number of sermons, Claridge and Wilson (2002) demonstrate that from the seventeenth to the twentieth century sermons have become less informational and more involved, thus following the general trend already observed in other genres (see Biber & Finegan 1989, 1992). The change in tendencies along Dimension 1 (Involved vs. Informational Production) in sermons can be explained as evidence of a gradual abandoning of the classical rhetorical style during the seventeenth century and the development of a more charismatic and plain style (generally advocated by sermon manuals) showing a higher degree of involvement. Claridge and Wilson (2002) do not observe any remarkable developments in sermons as far as Dimension 4 (Expression of Persuasion) and Dimension 6 (Informational Elaboration) are concerned, although they also acknowledge that individual idiosyncrasies can affect the particular characterisations of texts.

2.3.7 Trial Proceedings

Since the trial proceedings analysed in the present study are drawn from the Proceedings of the Old Bailey, the description of this genre is focused mainly on these.

Trial proceedings are written accounts of the actions, events, and testimony of the participants in a trial. The first published reports of trials held at the Old Bailey date back to 1674 and were published under the title News from Newgate: OR, An Exact and true Accompt of the most Remarkable, TRYALS OF Several Notorious Malefactors: At the Gaol delivery of Newgate, holden for the CITY of LONDON, and COUNTY of MIDDLESEX. In the Old Baily … (Huber 2007). The proceedings were cheap and soon became popular. Initially, they were published periodically eight times a year following each session of the court, then ten to twelve times a year until April 1913 when publication came to a sudden halt. As noted by Huber (2007), since true crime sold well, publishers used to send scribes

Further discussion is provided in Section 2.3.8.

Apart from the various works mentioned in this section, the information about the Old Bailey Proceedings comes from the official web page (www.oldbaileyonline.org/) maintained by Robert Shoemaker (Department of History, Humanities Research Institute, University of Sheffield) and Tim Hitchcock (Department of History and Social Sciences, University of Hertfordshire).
2.3 A Note on the Genres

to the Old Bailey to record the trials in shorthand, and for this reason early proceedings tend to be rather sensationalist and judgemental. However, from 1679 onwards the City of London gained control over publication, so the tone of the proceedings becomes more objective. In 1775 it was established that the proceedings should be published under the authority of the chief judge at the Old Bailey and authenticated under his name, and three years later it was determined that the proceedings should ‘contain a true, fair, and perfect narrative of the whole evidence upon the trial of every prisoner’ (Shoemaker 2008: 561). Thus, the proceedings were no longer provided for the purpose of entertainment and became a means of keeping an accurate public record of events in the courtroom.

Early editions of the *Proceedings* were rather short and included brief summaries of trials. A key development was the change from publishing third-person summaries of testimonies (as happens in witness depositions) to first-person accounts around 1712. From this point onwards, the *Proceedings* started to include verbatim testimonies of prosecutors, witnesses, and defendants, as well as judges’ comments and questions.

The reliability of the proceedings as a linguistic and historical source has been extensively analysed in the literature (see, among others, Culpeper & Kytö 2000, 2010, Kytö & Walker 2003, Huber 2007, Shoemaker 2008), with one of the major criticisms being the extent of scribal or editorial intervention. The proceedings at the Old Bailey were probably recorded in shorthand from the 1670s. Shorthand techniques were not new at the time. In fact, as noted by Culpeper and Kytö (2010: 51), ‘ever since the earliest printed English shorthand manual, Timothy Bright’s *Characterie* (1588), appeared, there was a growing interest in developing shorthand systems into a more efficient direction’. Huber (2007) analyses the shorthand method developed by Thomas Gurney, the scribe who recorded the proceedings at the Old Bailey from 1749 to 1770. He comes to the conclusion that the recording system and the publication of the proceedings immediately after each session are arguments in favour of an accurate portrayal of the spoken language, although he also acknowledges that being faithful to the spoken word was not always possible. Manipulation by scribes and publishers has also proved to be a crucial factor in the recording of bad language in the proceedings. Widlitzki and Huber (2016) note that offensive language is used less and less in the proceedings over the two-hundred-year span covered by the OBC, and that the increase

---

26 On the importance of the role of the recorder and methods to ascertain whether dialogues are constructed or reconstructed by scribes, see Grund (2007) on the Salem witch trials.
of municipal control over the proceedings between 1750 and the second half of the nineteenth century could be one of the reasons behind this. Similarly, the presence of examples in the proceedings in which the scribe makes use of non-standard spelling in an attempt to imitate, for example, the Irish accents of speakers indicates that trial reports ‘are not a “clean” record of a speech event, linguistically speaking, but, rather, an interpretation of that speech event’ (Archer 2012: 148). Alleged verbatim records do not always grant a faithful representation of a speech event, and one might expect that certain features typical of spoken language (e.g. slips of the tongue, false starts) would be left out. Yet these are indeed present in some proceedings (see Kytö & Walker 2003: 225). In any case, the omission of these features would not affect a study on the development of phrasal verbs in any substantial way.

Huber (2007) further analyses the internal consistency and external fit of the Old Bailey Proceedings. Internal consistency is defined by Schneider (2002: 86) as the degree to which variable features are portrayed consistently across large corpora, whereas external fit refers to the extent to which the results of the analysis of a specific corpus concur with those of other studies. Regarding their external fit, Huber (2007) shows that scribes differentiated systematically between speech and prose in the use of negative contractions, an argument in favour of the proceedings’ portrayal of the spoken language. However, in comparing a sample trial with an alternative case from the same court, notable differences (omissions, verbal, morphological, and syntactic divergences) were found, which shows that the proceedings cannot be taken as exact representations of the spoken language. Moreover, the analysis of their internal consistency suggests that the representation of linguistic features can be greatly affected by scribal and/or printer interference.

Culpeper and Kytö (2000) compare trial proceedings, witness depositions, prose fiction, and drama in terms of their reliability as faithful representations of speech. They test these genres for features which are generally assumed to be strongly associated with spoken interaction, namely lexical repetitions, turn-taking features (question-answer adjacency pairs, and interruptions), and single-word interactive features (first- and second-person pronouns, private verbs, and demonstrative pronouns). In their data they find that, although there is evidence in favour of drama being close to real speech, trial proceedings can also be described as linguistically close to spoken face-to-face interaction. The advantage of the latter over drama is that they are based on real, not imagined, speech.

According to Culpeper and Kytö (2010: 63–4), trial proceedings and witness depositions are not entirely speech-like in a number of
respects: speech is regulated, the order of topics is predetermined, linguistic forms are prescribed, and language is always produced for a public audience, often on behalf of a collective one (e.g. the court). Even so, trial proceedings present two main characteristics which make them particularly speech-like: they are highly interactive and are not planned. In Culpeper and Kytö’s study, comedy drama scores higher than trial proceedings on all speech-like aspects (2010: 63–4, 142–57). However, it must be noted that Culpeper and Kytö’s (2010) research refers to A Corpus of English Dialogues 1560–1760 (CED), which differs from my corpus in two important respects. First, the trial proceedings contain some (though not many) explicit interventions by the scribe, such as the description of non-verbal behaviour, whereas the selection of texts from the OBC used in the present research contains only direct speech (see Section 2.2.2). Second, the dramatic works in CED are exclusively comedy drama, whereas those in ARCHER represent a larger range of types.

In spite of the obvious drawbacks mentioned above, the main conclusion drawn from Huber’s (2007) analysis is that, although the proceedings cannot be described as faithful word-by-word transcriptions, ‘they are still among the few and best sources we have of spoken language before the advent of mechanical recording’ (Huber 2007). In sum, although we cannot claim that speech-related texts are exact equivalents to real speech, in the absence of better data they constitute one of the best sources for the study of the development of speech.

2.3.8 A Reclassification Proposal for the Genres in ARCHER and the OBC

As shown in the preceding section (2.2.1), ARCHER includes text types which are representative of formal (e.g. sermons) and informal (e.g. diaries) registers, as well as others which are better seen as neutral in this sense (e.g. news). On the other hand, although all the texts in ARCHER are recorded in the written medium, some can be described as closer to the oral language. These have been termed speech-based texts in various multivariate analyses, where they are defined as ‘registers that have their origin in speech, even though they are preserved in writing’ (Biber & Finegan 1997: 253). Nevertheless, Biber and associates rightly note that none of these registers really represents natural speech: ‘[d]ramatic and fictional dialogue are shaped by author’s intuitions and stereotypes about conversation, while sermons are often written and then read. Thus, any claims
about actual speech based on these samples must be extremely cautious’ (Biber et al. 1998: 252–3). Moreover, the definition of speech-based texts as registers that have their origin in speech is probably not the most appropriate one for genres such as sermons, with a tradition of written composition reaching back to late Classical Antiquity.27

For my purposes, I will establish a classification similar to that in Yáñez-Bouza (2007: 133, inspired by Kytö (1996)),28 but using the terminology employed by Culpeper and Kytö (2010). In ARCHER and the OBC we can distinguish between two major groups of texts, speech-related (diaries, drama, letters, sermons, and trial proceedings) and writing-based and writing-purposed texts (fiction, journals, news, medicine, and science). The former are further subdivided into three subgroups (see Culpeper & Kytö 2010: 17):

(a) **Speech-like texts** (letters and diaries). These are defined in terms of a scale consisting of features of communicative immediacy. Personal correspondence, for example, ‘is neither based on nor designed to be like speech; its claim to being “oral” or “colloquial” is solely that it contains features that are speech-like’ (Culpeper & Kytö 2010: 17).

(b) **Speech-based texts** (trial proceedings) are those ‘based on an actual “real-life” speech event’ (Culpeper & Kytö 2010: 17). They do not necessarily involve an accurate reproduction of a speech event, but are in most cases reproductions enhanced by notes. However, in the absence of audio-visual equipment or complete and accurate shorthand systems, they are the closest we can get to the spoken language of the time. Since ARCHER does not contain any speech-based texts, the trial proceedings of the OBC constitute a very convenient source of data to be compared with other texts which show a certain degree of speech-likeness.

(c) **Speech-purposed texts** (drama and sermons) are designed to be articulated orally, either as monologues (sermons) or reproducing real-time interaction (drama).

The dichotomy speech-related texts vs. writing-based and writing-purposed texts does not constitute a clear-cut set, as both categories are often interrelated and overlapping. Similarly, the degree of (in)formality of a given text is better measured in terms of a scale or continuum, with

---

27 I thank an anonymous reviewer of Rodríguez-Puente (2012c) for calling my attention to this.
28 The classification was reformulated in Yáñez-Bouza (2015a: 32–4) in terms of medium and setting.
2.4 Methodology

Searches for phrasal verbs in the corpora were carried out automatically using *WordSmith Tools* versions 3 and 6 (Scott 1999, 2012). Given that ARCHER 3.1 is not morphologically tagged, the procedure required searching for the individual particles by means of concordances. This was followed by further manual analysis to identify those examples in which
diaries, drama, letters, and the dialogues in trial proceedings approaching the informal end, sermons, medicine, and science towards the formal end, and with fiction, journals, and news standing halfway. The representation of the text types in ARCHER and the OBC conceived of in such a framework is represented in Figure 2.1.

The scale of (in)formality represented in Figure 2.1, however, must be taken with caution. If we bear in mind the changing conventions of genres over time described in Sections 2.3.1 to 2.3.7, we need to develop different configurations to represent the text types. Thus, for example, letters would approach the formal end of the continuum during earlier periods and move to the informal end by the twentieth century. These genre shifts towards more colloquial or literal styles have led many authors to talk about a colloquialisation process or the acquisition of features typical of speech in certain genres (see, among others, Biber & Finegan 1997, Mair 1997, 2006, Hundt & Mair 1999, Seoane & Loureiro-Porto 2005, Smitterberg 2008, Leech et al. 2009: 20), whereas the opposite phenomenon can be labelled as ‘decolloquialisation’ (see Rodríguez-Puente 2014, 2017). These changing conventions, although not represented in Figure 2.1, have been considered in the cross-genre analysis in Chapter 6.

![Figure 2.1 Distribution of genres in ARCHER and the OBC according to the dimensions of (in)formality and their speech-like vs. written characterisation](https://doi.org/10.1017/9781316182147.002)
the particles were part of phrasal verbs. For this purpose, and for the
sake of comparison with earlier periods, I made use of the list of EModE
particles in Claridge (2000: 46), which is based on lists provided in pre-
vious works in the field (see Bolinger 1971: 17ff, ODCIE 1975: lxxx, Fraser
1976: 5, Quirk et al. 1985: 1151), and which includes the following thirty-
four items (see Section 3.3):^{29}

\[
\text{aback, aboard, about, above, across, after, ahead, along, apart, around, aside,}
\text{astray, asunder, away, back, behind, by, counter, down, forth, forward(s),}
\text{home, in, off, on, out, over, past, round, through, to, together, under, up.}
\]

The list is far from complete, but the items included seem to constitute
an adequate representation of the category in question. I also excluded
eamples of the particle out followed by of (see Section 3.3), as well as
cases of particle combinations with the verb to be, following the common
practice in earlier studies on the topic (see, e.g., Pelli 1976, Martin 1990,
Claridge 2000).^{31}

The procedure of extraction of examples from ARCHER was as follows.
In the first instance I obtained a wordlist to check all the possible spellings
of the particles at issue, then carried out concordances for the individual
particles in the seven subperiods of the corpus. The extraction of examples
was a very long and tedious process, in that most phrasal verb particles have
homonymous prepositions. To give an example, if we make a concordance
for the particle in in the period 1950–90, WordSmith Tools provides 3,642
results of which only eighty-four are phrasal verb particles. The OBC and
the ICE-GB are POS-tagged, which greatly facilitates searches for phrasal
verbs therein. These were carried out automatically, although further

^{29} From Claridge's (2000) list, I excluded the particle ashore because it is restricted to nautical usage. See Section 3.3 for further details.

^{30} According to Palmer, 'to occurs as part of phrasal verb only in come to and bring to (return to consciousness) [...] To, moreover, in bring to can occur only after the noun phrase' (1988: 228). Note, however, the existence of the combination lay to ('come to a stationary position with the head towards the wind', OED s.v. lay to 3).

^{31} Claridge (2000) does not explicitly state that the verb be is excluded from her data, but none of the combinations in her list includes this verbal base. Given the high frequency of the verb be (see, e.g., Biber et al. 1999: 359), it seems highly improbable that the Lampeter Corpus would not include one single instance of be + particle combination, which suggests that Claridge actually excluded these structures. In order to check my intuitions, I performed a concordance in the Lampeter Corpus to look for combinations of the pattern be is/are/was/were/being + away/down/in/on/out up. This produced a total of 1,058 results, some of which included be + particle combinations, such as the one illustrated in (i) below:

(i) Papists are up, and Atheists and Infidels and Jews are up, and abundance of secret Apostates are up openly reproaching the Ministry, that privately deride Christ and Scripture, and the life to come.
manual analysis was also necessary. I began by looking for the tag <RP>, defined in the CLAWS7 tagset as ‘prep. adverb, particle (e.g. about, in)’. However, a search such as this would exclude adverbs such as away, behind, forward, past, and together which can also be used to form phrasal verbs. Therefore, the tag <RL> had to be also included, which in the CLAWS7 tagset stands for ‘Locative adverb (e.g. alongside, forward)’.

The relevant examples obtained from the corpora were then stored in a Microsoft Excel database and coded according to various linguistic parameters. This allowed for calculations and the creation of tables and graphs representing the tokens. It must be noted that cases of -ing nominalisations (e.g. the taking off) and of other derivatives (e.g. a cast-off nightgown, a setter forth) have not been counted as phrasal verbs (see Section 5.2.4), although all such examples found in the corpus were stored in a separate worksheet.

Finally, differences in the number of words between the corpora were of course taken into account for the analysis by normalising raw data per 10,000 words when appropriate. Moreover, I verified the statistical significance of the different variables whenever necessary by performing the Kruskal-Wallis test and the Wilcoxon test by means of the free software R version 3.3.1 (2016). These two tests were preferred to the Chi-square on the grounds that tests based on the assumption that all words occur independently (such as the Chi-square test) tend to overestimate significance in frequency comparisons, especially for poorly dispersed words (Lijffijt et al. 2016; see also Kilgarriff 2001).

---

32 See http://ucrel.lancs.ac.uk/claws7tags.html.
33 The Kruskal-Wallis one-way analysis of variance by ranks is a non-parametric method used for comparing more than two samples that are independent or not related. When the Kruskal-Wallis test leads to significant results, then at least one of the samples is different from the other samples. In such a case, the Wilcoxon test can be applied for a more fine-grained analysis. This test is used to compare two related samples, matched samples, or repeated measurements on a single sample, to assess whether the population mean ranks differently.