DE GRUYTER OPEN

ICAME Journal, Volume 41, 2017, DOI: 10.1515/icame-2017-0003

The *Corpus of Historical English Law Reports 1535–1999 (CHELAR)*: A resource for analysing the development of English legal discourse

Teresa Fanego,[1] Paula Rodríguez-Puente,[2] María José López-Couso,[1] Belén Méndez-Naya,[1] Paloma Núñez-Pertejo,[1] Cristina Blanco-García[1] and Iván Tamaredo[1]

[1] University of Santiago de Compostela

[2] University of Oviedo

1 Introduction

The importance of textual genres[1] for our understanding of the development of linguistic features has long been recognised in historical linguistics, as is reflected in the pioneering work of Görlach (1991, 1992) and in the availability, from the early 1990s onwards, of an increasing number of electronic corpora affording access to "various text types, levels of style and modes of expression" (Kytö 1996: 1). The *Helsinki Corpus of English Texts* (Rissanen *et al*. 1991), the first and arguably the most influential database of historical English, aimed at "generic coverage" (Kytö and Rissanen 1992: 12) through a representative sampling of texts from several major text categories. Yet despite its many merits, the *Helsinki Corpus*, due to its limited size (1.5 million words covering the years c. 850 to 1710), cannot come close to illustrating all the text types of English in a chosen period. This, as noted by Diller (2001: 30), was clearly one of the reasons why the *Helsinki Corpus* soon came to be "expanded in the direction of supplementary corpora concentrating on single genres." Among these we might mention the *Corpora of Early English Correspondence* (Nevalainen *et al*. 1993–ongoing), the *Corpus of Early English Medical Writing* (Taavitsainen *et al*. 1995– ongoing), the *Corpus of English Religious Prose* (Kohnen *et al*. 2003–ongoing), the *Coruña Corpus of English Scientific Writing* (Moskowich *et al*. 2004– ongoing), the *Corpus of English Dialogues 1560–1760* (Kytö and Culpeper 2006), the *Málaga Corpus of Late Middle English Scientific Prose* (Calle *et al*. 2012–2015), and the *Old Bailey Corpus* of trial proceedings (Huber *et al*. 2012).

In this article we report on the *Corpus of Historical English Law Reports 1535–1999* (*CHELAR*), a new specialised corpus of legal English which has been developed at the Research Unit for Variation, Linguistic Change and Grammaticalization (VLCG; http://www.usc-vlcg.es/) of the University of Santiago de Compostela. *CHELAR*, initiated by the VLCG team in 2011, was completed in 2016 and has recently been made available.[2]

The article is structured as follows. Section 2 provides a brief overview of the history of legal English and a more extended account and characterization of law reports, the raw material for *CHELAR*. Section 3 compares *CHELAR* to other existing synchronic and diachronic corpora of legal English. Section 4 explains the complex process in compiling *CHELAR* and its annotation system. The paper concludes with a section offering an outlook on *CHELAR*'s research possibilities.

2 Legal English

English legal discourse is a register with a long history (see especially Hiltunen 1990; Tiersma 1999: 9–47; Claridge 2012: 239–240; Scotto di Carlo 2015: 5–27). The oldest legal texts date back to Old English times, specifically to the laws promulgated by King Ethelbert of Kent (c. 558–635 AD) and various other legal codes which followed them, until Cnut's decrees of the eleventh century (1016–1035). With the Norman Conquest, however,

> English ceased to exist as a language of the law for about four centuries, until the 1362 Statute of Pleading re-established English as the oral legal language, and the first Act of Parliament to be written in English was passed in 1483. During the Middle English period, legal writing had used first 'Law Latin', later French, while pleading had taken place in French. [...] The full establishment of English in all spheres of law was gradually carried through during the Early Modern English period, involving translation of important texts into English. (Claridge 2012: 240)

From the sixteenth century legal texts of various kinds became more readily available: records and law reports written in running English prose are found from the early years of the modern period, as discussed below; more oral forms of legal discourse, such as trial records and transcripts, are scarce during the greater part of the sixteenth century, but become "more numerous from the mid-1600s onwards" (Culpeper and Kytö 2010: 50). The most iconic of English trial courts – the Old Bailey – dates from 1673; its proceedings were published from

1674 to 1913 and constitute a large body of texts containing almost 200,000 trials and providing "verbatim passages [which] are arguably as near as one can get to the spoken word of the period" (Huber *et al*. 2012; see also Archer 2014).

As mentioned above, law reports constitute the raw material for *CHELAR*. In English common law, reports are records of judicial decisions which are "cited by lawyers and judges for their use as precedent in subsequent cases" (*Encyclopædia Britannica Online* s.v. *law report*; see also *OED* s.v. *report* n. 2.b). Law reports are thus of fundamental importance in a legal system which, unlike the civil law of countries following the Roman legal tradition, is not based on pre-established legal codes, but rather "has grown in a way inductively, through individual cases and decisions" (Hiltunen 1990: 13). In their modern form, law reports are divided into distinct components and typically assume the form of "faithful records of all the facts of the case, the arguments of the judge, his reasoning, the judgment he arrives at and the way he does it, the kind of authority and evidence he uses and the way he distinguishes the present case from others cited as evidence" (Bhatia 1993: 119).

Bhatia's seminal typology (1987: 227–230) of British legal genres reflects the communicative purposes they tend to fulfil, the settings or contexts in which they are used, the communicative events they are associated with, the social or professional relationship between participants, and the background knowledge that such participants bring to the situation in question. Based on these and other related factors, Bhatia categorises the written language of the law under three major headings:

- *academic* writing: research journals and legal textbooks
- *juridical* writing: law reports, cases and judgments
- *legislative* writing: acts of parliament and statutory instruments, and also legal documents such as contracts, agreements, wills and insurance policies

Juridical and legislative writings correspond, respectively, to the categories of documents that Tiersma (1999: 139–141), in another well known classification of legal texts, has called *expository* and *operative*: expository documents "typically delve into one or more points of law with a relatively objective tone" (1999: 139); operative documents, by contrast, tend to modify or create legal relations and "to have direct and highly significant consequences" (1999: 141); it is in these latter that "the most notorious attributes of legal English tend to occur" (1999: 139). Law reports and judicial opinions, "to the extent that the judge expresses what the law is", are expository, but they typically "also contain

a judgment or order at the end that constitutes the actual disposition of the case; such an order is operative" (1999: 139).

Bhatia's and Tiersma's functional typologies are ultimately equivalent to the distinction between *prescriptive* (or normative) and *descriptive* (or non-normative) legal texts which other authors, such as Šarčević (2000) and Williams (2005), prefer to use. Between those two clearly defined groups they identify *hybrid* texts, which contain both prescriptive and descriptive features. When examined from this perspective, law reports and judicial opinions would rank primarily as descriptive according to Šarčević (2000: 11), and as hybrid according to Williams, since they combine prescriptive and descriptive features, though "it is the descriptive element – as opposed to the prescriptive element – that usually predominates" (Williams 2005: 29; see also López-Couso and Méndez-Naya 2012: 8–9).

To sum up, law reports are a type of legal text which is marked, in modern times, by distinctive functional features and "a typical discourse organisation" (Bhatia 1987: 230); such texts play a pivotal role in the UK judicial system, because "law courts follow their previous judgments within more or less well-defined limits" (Bhatia 1993: 118).

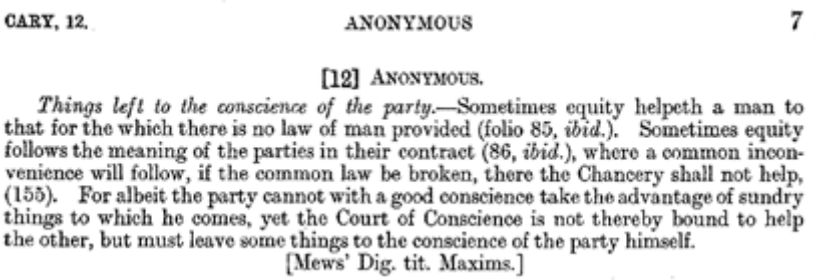
2.1 Stages in the history of English law reporting

The earliest reports in English common law were collected in the *Year Books*. These were brief manuscript notes of proceedings which were collected and published annually, whence their name *Year Books* (see Tiersma 1999: 22). The *Year Books* were produced between 1268 and 1535 and consisted of anonymous reports written either in Latin or French, though they were later on translated into English.

From the year 1535 onwards, the *Year Books* were superseded by published editions known as the *Nominate Reports*, because they were named after the reporter who attended the court as an observer and who then compiled and edited them. With the passage of time, the *Nominate Reports* became more expansive and introduced the style and approach which has become characteristic of modern law reporting (see further Cornish *et al*. 2010: 1211 ff). Reporting developed into a professional activity, and this often led to the publication of different versions of the same judgment in different sets of reports. To deal with this problem, in 1865 the Incorporated Council of Law Reporting for England and Wales (ICLR, <http://www.iclr.co.uk/>) was established as the only authorised publisher of the official series of law reports for the superior and appellate courts of England and Wales. The ICLR was also responsible for compiling the majority of the best copies of cases predating its founda-

tion, which were eventually published in the form of reprints known as the *English Reports*. These contain both the translations into English of the *Year Books* (1268–1535), and the *Nominate Reports* (1535–1865). Any reports published after 1865 and produced by the ICLR are known as the *Law Reports*, these constituting the third and final stage in the history of English law reporting.

Figure 1 below shows the reprint of a year book (*Anonymous*, 1468); Figure 2 gives an example of a nominate report (*Pawlinge v. Homfrey*, 1578).



CARY, 12. ANONYMOUS 7

[12] ANONYMOUS.

Things left to the conscience of the party.—Sometimes equity helpeth a man to that for the which there is no law of man provided (folio 85, *ibid.*). Sometimes equity follows the meaning of the parties in their contract (86, *ibid.*), where a common inconvenience will follow, if the common law be broken, there the Chancery shall not help, (155). For albeit the party cannot with a good conscience take the advantage of sundry things to which he comes, yet the Court of Conscience is not thereby bound to help the other, but must leave some things to the conscience of the party himself.

[Mews' Dig. tit. Maxims.]

Figure 1: Example of a year book (Anonymous, 1468)

the defendant was committed to the prison of the Fleet. And now upon motion he is ready to pay the fine, upon payment whereof he is discharged of his imprisonment. Blagrave plaintif, Wotton defendant. Anno 21 Eliz. [1578–79].

[131] BROWN *v.* BENION.

For want of a Bill costs is gotten and discharged, for that by the defendants means he was stayed to proceed by authority of the Councel of Wales.—The defendant got costs for want of a Bill, and after the plaintif shewed forth a commandment from the Councel of the Marches of Wales at the defendants suite, procured after the serving of the Subpœna, to shew cause why he should not stay his proceeding in this Court, whereupon he staid the putting in of his Bill. Therefore discharged of the costs. Brown Plaintif, Benion Defendant, Anno 21 Eliz. [1578–79].

PAWLINGE *v.* HOMFREY.

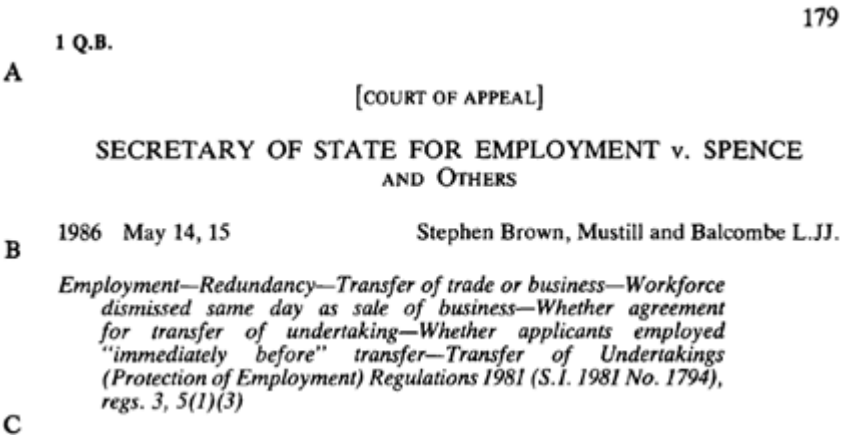
Jurisdiction of Oxford allowed.—Forasmuch as the Commissary of the University of Oxford hath certified under the seal of the said University, that the said Gilbard, one of the defendants, is a Master of Arts and Batchellor of Divinity within the said University: And that by the confirmation of our Sovereign Lady the Queens Majesty, and by the grant and confirmation of her Highness Noble Governours, none of the said University, which the said Commissary shall certifie to be a necessary member of the same, shall be compelled to answer to any suit or action whatsoever out of the same University, except it be for fellony, mayhem, or free-hold, as by the same Certificate more at large appeareth. And for that also it seemeth unto this Court, that the matter wherewith the said Nicholas Gilbard is charged, is for a supposed going about to get a coppy of a Court Roll out of the hands of the said Margaret Pawlinge, another of the said defendants: It is ordered that the said Gilbard be dismissed, and the Plaintifs referred to take their remedy before the Chancellor of Oxford, Vice-Chancellor or Commissary of the same University. Pawlinge Plaintif, Homfrey Sacræ Theolog. Professor, Gilbard and Pawling [132] widow defendants. Anno 21 Eliz. [1578–79].

Figure 2: An example of a nominate report (Pawlinge v. Homfrey, 1578)

As can be seen in Figures 1 and 2, in the top left-hand corner there is an abbreviation which refers to the original collection and the page on which the report can be found. In the year book (Figure 1) the abbreviation CARY in the top-left hand corner refers to the series *Cary's Chancery Reports*, whereas the nominate report (Figure 2) belongs to *Choyce Cases*. The number next to the name of the series indicates the page on which the report appeared in the original manuscript (page 12 in the year book and 131 in the nominate report). The original pagination also appears between square brackets within the body of the text. In the year book it can be seen before the title of the case, whereas in the nominate report it appears in the last line (note that page 131 indicates the beginning of the previous case, *Brown v. Benion*). The figure in the top right-hand corner, on the other

hand, indicates the page on which the report appears in the *English Reports* reprint (7 in the case of the year book and 79 in the case of the nominate report).

The third stage in the history of law reporting in England, as already noted, corresponds to the period extending from the foundation of the ICLR in 1865 to the present day. Since the second half of the nineteenth century, law reports have been "published according to the court where the case took place" (Kearns 2007: 31). They must follow a standard format and must be reported by a barrister-at-law who can vouch for the accuracy of the report (*Encyclopædia Britannica Online* s.v. *law report*). For this reason, although the ICLR does not belong to the UK government, the *Law Reports* are widely regarded as the most authoritative series of reports for England and Wales. Figure 3 below shows the front page of a law report (*Secretary of State for Employment v. Spence and Others*, 1987).



179

1 Q.B.

A

[COURT OF APPEAL]

SECRETARY OF STATE FOR EMPLOYMENT v. SPENCE AND OTHERS

B

1986 May 14, 15 — Stephen Brown, Mustill and Balcombe L.JJ.

Employment—Redundancy—Transfer of trade or business—Workforce dismissed same day as sale of business—Whether agreement for transfer of undertaking—Whether applicants employed "immediately before" transfer—Transfer of Undertakings (Protection of Employment) Regulations 1981 (S.I. 1981 No. 1794), regs. 3, 5(1)(3)

C

Figure 3: Front page of a law report (Secretary of State for Employment v. Spence and Others, 1987)

As can be seen, the information on the front page of a law report is much more complete than that in the *English Reports*. The abbreviation in the top left-hand corner indicates that this report belongs to the Queen's Bench Cases. The court where the case was judged (Court of Appeal) appears centred at the top of the page in square brackets, followed by the parties (*Secretary of State for Employment v. Spence and Others*). Underneath the parties, the dates on which the case was judged are shown on the left-hand side, with the names of the presiding

judges on the right. Moreover, law reports always include key words or words related to important concepts which are dealt with in the report. This introductory information is followed by a summary of the case which is being reported, then a list of the cases referred to in the report, and finally the body of the text. As a result, and as this brief description suggests, the information in the *Law Reports* is far more accurate and thorough than in the *English Reports*.

Figure 4 displays graphically the history of law reporting, as described in this section:

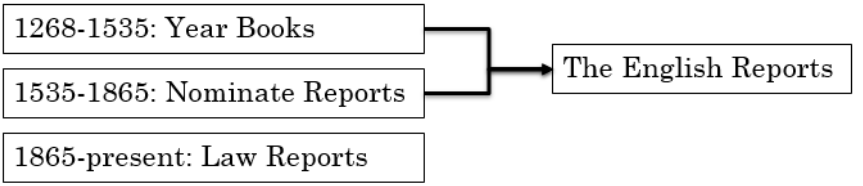


Figure 4: The history of English law reporting

3 Corpora of legal English

According to Bhatia, the expansion of interest in legal English is relatively recent, and can be seen as a result of developments in three disciplines, namely:

> (1) in linguistics proper, where the inclusion of pragmatics in the study of language has encouraged linguists to look for the use of language in real life settings, (2) in applied linguistics, where the main concern has been to design and teach language support courses for academic as well as professional legal courses, and (3) in social science disciplines, where legal language [...] is increasingly being recognised as the vehicle for social action. (Bhatia 1987: 227)

These combined trends can undoubtedly help to explain the appearance, over the past few years, of several electronic corpora which contain legal documents of various kinds, both synchronic and diachronic.

Probably the best-known corpus of legal English is the *Cambridge Legal English Corpus*, owned by Cambridge University Press; it contains 20m words of contemporary books, journals and newspaper articles relating to the law and legal processes, but is not commercialised or accessible to the general public. Other synchronic corpora are the *House of Lords Judgments Corpus* (*HOLJ*, 3m

words; Grover *et al*. 2004), which covers the years 2001–2003 and includes judgments delivered by the judges of the House of Lords, and the *British Law Report Corpus* (*BLaRC*), a 8.85m word database compiled by Marín Pérez and Rea Rizzo (2012); it includes judicial decisions issued between 2008 and 2010 by British courts and tribunals. For American English the most comprehensive corpus to date is the *American Law Corpus* (*ALC*, 5,578,393 words), compiled by Goźdź-Roszkowski (2011) for his research on (mainly) lexical aspects of Contemporary American Legal English. *ALC* contains legislation, contracts, US Supreme Court opinions, briefs, textbooks, professional articles and journal articles, and is thus representative of the three overarching categories (legislative, juridical and academic) of written legal English distinguished by Bhatia (1987).

Turning now to diachronic corpora, the researcher interested in the development of legal English since Middle English times can, in the first place, make use of the legal components in the five main multi-genre historical corpora currently in existence, namely the *Helsinki Corpus of English Texts* (c. 850–1710; see Kytö and Rissanen 1992), the *Penn-Helsinki Parsed Corpus of Early Modern English* (1500–1710; Kroch, Santorini and Delfs 2004), the *Penn-Helsinki Parsed Corpus of Modern British English* (1700–1914; Kroch, Santorini and Diertani 2010), the *Lampeter Corpus of Early Modern English Tracts* (1640–1740; Schmied, Claridge and Siemund 1999), and *ARCHER 3.2 – A Representative Corpus of Historical English Registers*, *version 3.2* (1600–1999; Biber *et al*. 1990–1993/2002/2007/2010/2013). As regards the *Helsinki Corpus*, the legal texts for the Early Modern English period comprise statutory writings (36,750 words) and trial proceedings (43,960 words). The *Penn-Helsinki Parsed Corpus of Early Modern English* includes the *Helsinki Corpus* files plus supplementary material, with a total of 115,863 words of statutory writings and 105,090 words of trial proceedings. Its successor, the *Penn-Helsinki Parsed Corpus of Modern British English*, sought to keep its genre composition as close as possible to that of its predecessor, and thus only statutes (65,748 words) and trial proceedings (58,973 words) were included. Statutory writings, as mentioned in Section 2 above, fall within the category of prescriptive legal texts and hence are very different from the legal texts which make up *CHELAR*. Trial proceedings, in turn, are "a paradigmatic example of speech-based genres" (Culpeper and Kytö 2010: 18), in the sense that they "are based on an actual 'real life' speech event" (2010: 17); they can therefore be assumed not to have much in common with the *CHELAR* texts. In turn, the *Lampeter Corpus of Early Modern English Tracts* contains a total of 1,193,385 words, and includes samples from six different domains: religion, politics, economy and trade, science, law, and miscellaneous; law writings are represented by 20 texts pertaining to a somewhat diverse range

of legal categories, such as laws and ordinances, petitions, law reports, trial proceedings, and even essays on legal issues.

Finally, in the case of *ARCHER 3.2*, the legal samples (166,343 words in total) date from 1600 to 1999 and consist of law reports, like *CHELAR*. This is no coincidence, as the VLCG team was responsible for the compilation of *ARCHER*'s legal component (for details, see López-Couso and Méndez-Naya 2012). The process of text selection, however, was carried out with care in order to avoid any textual overlapping, so that the two corpora can complement each other nicely. Tables 1 and 2 in Section 4.2 below give comparative figures on the number of words and files per subperiod in both *CHELAR* and *ARCHER*.

English legal documents are also represented in three other important diachronic corpora of a more specialised nature, namely the *Corpus of Early Modern English Statutes (1491–1707)* (Lehto 2013), the *Corpus of English Dialogues 1560–1760* (*CED*; Kytö and Culpeper 2006) and the *Old Bailey Corpus* (*OBC*; Huber *et al*. 2012). The *Corpus of Early Modern English Statutes (1491–1707)* was compiled by Anu Lehto at the University of Helsinki for her doctoral research; it contains 214,000 words in four categories of legislative writing: parliamentary acts, proclamations, royal orders and Privy Council orders (Lehto 2013: 239); at the time of writing this article, the corpus was not yet accessible. *CED* (1,2m words) is conceived, to some extent, as a *special purpose* corpus (on this term, see Bowker and Pearson 2002: 12), since it focuses on spoken interaction in the past and contains only "reliable speech-related texts" (Culpeper and Kytö 2010: 23). Three of the five genres included in *CED* – drama comedy, didactic works and prose fiction – exemplify dialogue 'constructed' by an author, while the other two genres – trial proceedings (285,660 words) and witness depositions (172,940 words) – purport to be written records of real speech events, since "they preserve the substance of the utterances exchanged between interlocutors" (Culpeper and Kytö 2010: 60). In turn, the *OBC* is a very large body of legal texts (14m words) in a fully searchable edition; it consists of trial proceedings held at London's central criminal court from 1720 to 1913 and thus offers, like *CED*, "the rare opportunity of analyzing spoken everyday language" (Huber 2007: 1) in past periods.[3]

So it is clear from the preceding account that *CHELAR* differs from all other diachronic corpora of legal English in a number of fundamental respects, as follows:

(a) Coverage of an extensive time span of nearly five centuries, from 1535 to 1999. The text samples in the *Penn-Helsinki* corpora date from 1500 to 1914 and thus leave out most of the 20th century; the *Corpus of*

Early Modern English Statutes (1491–1707), the *Corpus of English Dialogues* (1560–1760) and the *Old Bailey Corpus* (1720–1913) cover only two centuries each; the *Lampeter Corpus of Early Modern English Tracts* (1640–1740) covers one; *ARCHER 3.2* covers four centuries, from 1600 to 1999.

(b) Most importantly, *CHELAR* enables researchers to investigate a legal text type, law reports, which is not included in any other existing diachronic corpus – apart from *ARCHER 3.2* and, to a far lesser extent, the *Lampeter Corpus of Early Modern English Tracts*. Yet law reports, as already mentioned (Section 2), "stand at the very core of the common law system, acting as the main source of law [...], and thus hold[ing] a prominent position in legal ESP" (Marín Pérez and Rea Rizzo 2012: 135).

(c) *CHELAR* also compares favourably with the legal section in *ARCHER 3.2*, also consisting of law reports. *CHELAR* is not only much bigger in terms of the total number of words (463,009 versus 166,343), but, crucially, each of its nine time periods (see Section 4.2 below) is represented by a sample of approximately 50,000 words or slightly more, compared to just 20,000 words in the case of the time periods in *ARCHER 3.2*. Fifty thousand words is often considered to be the (minimum) desirable size for syntactic research, since in this domain the utility of a corpus depends on the number of clauses that it contains (see Kroch, Santorini and Delfs 2004: Background).

4 Creating the Corpus of Historical English Law Reports 1535–1999

4.1 Text selection

The *Law Reports* (1865 onwards) and the reprints of the *Nominate Reports* (1535–1865) are the documents used for our corpus. As already mentioned (Section 2.1), the *Year Book* reprints (1268–1535) are translations of earlier reports written in French or Latin; they contain scant bibliographical information and are often difficult to situate in time, so they are highly unsuitable for a historical corpus. The year 1535, when the *Year Books* were superseded by the *Nominate Reports*, was therefore chosen as the starting date for *CHELAR*; the latest reports included in the corpus date from 1999.

In July 2009, when we were about to start the compilation of the British English legal texts for *ARCHER 3.2*, we purchased an annual subscription to

Justis Publishing Limited (<http://www.justis.com>), an online legal library. The *Law Reports* and the *English Reports* (i.e., the reprints, up to 1865, of both the *Year Books* and the *Nominate Reports*) were then downloaded, together with other relevant information: dates of the trials, date of publication, judges, courts that judged the cases, parties, source, etc. In some of the oldest cases, however, it was impossible to identify either one or indeed both participants. In two of the cases from 1557 the second party appears with a slash indicating that it is an unknown person, whereas two texts from the same year are labelled as "anonymous".

The list of texts and word counts can be found in the Appendix to *CHELAR*, which can be downloaded from the VLCG homepage (<www.usc-vlcg.es/>).

4.2 Corpus structure

Although a balanced corpus is hard to achieve (Atkins, Clear and Ostler 2007: 111), in an attempt to make *CHELAR* as balanced and representative as possible, we divided the corpus into eight fifty-year subperiods going from 1600 to 1999, plus a longer subperiod from 1535 to 1599. Our aim was to include 20 different files of approximately 2,500 words each per subperiod, which would amount to circa 50,000 words per fifty years. The final corpus structure is set out in Table 1; Table 2 gives comparative figures for the legal section in *ARCHER 3.2*.

Table 1: Structure of *CHELAR*

Subperiod	Number of words	Number of files	Number of texts
1950–99	50,662	20	20
1900–49	50,816	20	21
1850–99	51,447	20	24
1800–49	52,350	20	32
1750–99	51,084	20	21
1700–49	50,465	20	22
1650–99	51,019	20	30
1600–49	52,185	22	112
1535–99	54,337	23	87
Total	**463,009**	**185**	**369**

Table 2: Structure of the legal section in *ARCHER 3.2*

Subperiod	Number of words	Number of files
1950–99	20,721	10
1900–49	21,160	10
1850–99	20,757	10
1800–49	20,531	10
1750–99	20,367	10
1700–49	21,315	10
1650–99	20,466	10
1600–49	21,026	10
1535–99	–	–
Total	**166,343**	**80**

In some cases, a single report sufficed to reach the 2,500 words required for the file, but occasionally it was necessary to include two or more texts per file, especially when dealing with the *Nominate Reports*, which tend to be shorter than the *Law Reports*. Thus, whereas in subperiod 1950–99 one single law case was enough to attain the target number of words, for earlier subperiods we had to include various cases from the same year in a single file. Moreover, in subperiod 1600–49 we had to extend the number of files to 22, as the available material from 20 different years was not enough to cover our target number of words. We also attempted to neutralise the effects of sampling bias by selecting reports of cases judged in the various existing courts and written by different authors.[4]

4.3 Text processing

The downloaded reports (scanned images) were typed into a word processor, saved as plain text (.txt), and revised to correct typos and any other errors. Then they were named following the structure *year*x*xxx*10–2, where *year* indicates the year of publication, *xxxx* corresponds to four random letters of the participants' names, and a number ranging from 2 to 10 indicates the specific subperiod to which the text belongs (2=1535–99, 10=1950–99). So the first corpus file is named 1999regi10; it corresponds to a text published in 1999 which belongs to the final subperiod (10) and whose title is *Regina v. Woolin* (hence the string *regi* in the file name). Note that the first subperiod (1535–1599) in *CHELAR* is num-

bered 2, to allow for the potential addition in the future of a tenth, earlier subperiod; this would consist of samples from the *Year Books*.

4.4 Text edition

The process of typing up the texts was far from simple. With older texts in particular we faced a number of problems relating to the contents and the quality of conservation of the original documents, for which editorial decisions were needed.

Punctuation, for example, was complicated in some texts because it was partially blurred; occasionally, what seemed a colon was actually a semi-colon or an apparent stop was in fact a comma, a semi-colon or a colon with a blurred lower part. In such cases, we had to infer from the context what these dots represented.

Regarding the contents of the texts, some of the documents initially downloaded from Justis contained paragraphs in Latin. Although Latin expressions and sentences could not be totally excluded because of the formulaic nature of legal language, we attempted to keep them to a minimum by downloading new texts. Footnotes were also eliminated from the texts, as they are very lengthy (sometimes even longer than the body of the text itself) and mainly contained cross-references to other cases. We considered that they were irrelevant to any linguistic analysis and their inclusion would only bias the target number of words. Further details about these and other necessary editorial decisions are provided in Rodríguez-Puente (2011).

4.5 Annotation

The corpus mark-up is being carried out at two levels: part of speech (POS) and Extensible Markup Language (XML; Bray *et al*. 2008). An account of these is given in the following sections.

4.5.1 POS tagging

Part-of-speech mark-up facilitates the linguistic analysis of any corpus. For our purposes, we employed the CLAWS-7 tagger (Constituent Likelihood Automatic Word-tagging System; see Garside 1987) developed by the University Centre for Computer Corpus Research on Language (UCREL) at the University of Lancaster. The texts had to undergo a process of adaptation in order to add the tags, as CLAWS does not recognise non-ASCII characters.[5] The list of adapted characters and their adaptations is shown in Table 3 below:

Table 3: List of characters not recognised by CLAWS-7 and adaptations made for POS tagging

List of special characters in CHELAR not recognised by CLAWS-7	Adaptations made for POS tagging
£	pounds
à, â, á, ä, etc.	same vowel without diacritics
æ	ae
œ	oe
’ (curved apostrophe)	' (straight apostrophe)
½, ¾, etc.	1/2, 3/4, etc.
º	degrees
§	Section
º (for ordinal numbers), e.g. 13º	-*th*, e.g. 13th

The addition of POS tags in *CHELAR* has already been concluded. CLAWS claims to have achieved 96–97% accuracy, although the degree of precision varies according to the type of text. We have checked the accuracy of the POS tagging in one thousand word samples from the different subperiods. The results obtained are displayed in Table 4:

Table 4: Degree of accuracy of CLAWS-7 in the different subperiods of *CHELAR*

Subperiod	**Accuracy**
1950–99	97.6%
1900–49	96.7%
1850–99	96.2%
1800–49	97.3%
1750–99	98.5%
1700–49	97.9%
1650–99	96.5%
1600–49	95.7%
1535–99	95.5%

Although the degree of accuracy was higher than expected, especially in the earlier subperiods, we can see that the tagger is less precise with earlier texts. An exception to this rule is the text examined for the second half of the eighteenth century, for which CLAWS provided the highest degree of accuracy. Nevertheless, the precision of the tagger was above 95 percent in all cases. It must be noted, however, that these figures are approximations because the degree of accuracy increases or decreases according to the type of text. Legal texts tend to be very repetitive and repetitions are a hindrance for the precision of the tagger: if there is one tagging mistake in a word repeated several times in a text, that text will yield a lower degree of accuracy. Similarly, the tagger tends to be imprecise with Latin formulae which are frequent in some texts but not in others.

Most of the tagging errors made by CLAWS correspond to those potential mistakes typical in any kind of text, such as confusion between *after*, *as*, *before*, *since* and *until* as prepositions or subordinating conjunctions, *that* as determiner or subordinating conjunction, *–ed* forms of verbs as participles or past forms, *–ing* nouns (e.g., *proceedings*) as adjectives or verbs, or confusion between homographic words (e.g., *fine* noun/adjective).

Other errors, however, are a consequence of the complexity of sentence structure in many of the reports. In (1) below, for example, the noun *indenture* was tagged as the base form of a verb probably because of the difficulty of the relative construction with continuative *which* (i.e., "the uses specified in an indenture, [...] *which indenture* they find in *hæc verba* [...]"):

(1) […] `Howard_NP1 of_IO the_AT other_JJ part_NN1 :_: which_DDQ indenture_VV0 they_PPHS2 find_VV0` […] (*Berry v. White*, 1662)

Example (2) below shows a case number tagged as a formula:

(2) `The_AT requirement_NN1 of_IO section_NNi 277(8)_FO` […] (*South Lakeland District Council v. Secretary of State for the Environment and Another*, 1992)

Another common error has to do with the identification of Latin expressions, which are not recognised as foreign words (`FW`) by the tagger, as shown in (3):

(3) […] `de_NP1 proedict'_JJ decimis_NN1 garbarum_NN1` […] (*Simms v. Bennet*, 1579)

Although we are satisfied with the accuracy of the tagger, we believe that there are several ways in which its precision can be improved. Manual correction

might be an option in the near future, but resorting to (semi-)automatic tools such as the VARiant Detector (VARD; Baron and Rayson 2009)[6] to normalise variant spellings in the earlier texts might prove more useful. VARD is designed to assist corpus compilers in dealing with spelling variation, particularly in EModE texts. Following the guidelines established in Lehto *et al*. (2010), Hiltunen and Tyrkkö (2013) increased the accuracy of the tagger from 80 percent to over 90 percent in two texts from the sixteenth century which are part of the *Corpus of Early Modern English Medical Texts* (see Taatvitsainen and Pahta 2010). Archer *et al*. (2015), who have also applied spelling normalization to the *Corpus of English Dialogues* (Kytö and Culpeper 2006), advocate the creation of normalization guidelines that can be generalised to other historical corpora.

4.5.2 XML tagging

CHELAR will also be annotated using Extensible Markup Language (XML; Bray *et al*. 2008), following the *TEI P5 Guidelines for Electronic Text Encoding and Interchange* developed by the Text Encoding Initiative Consortium (TEI Consortium 2016). TEI XML encoding has become the standard practice adopted in digitally based humanities research for Present-day English corpora and is beginning to be implemented in historical corpus compilation, such as *ARCHER*, the *Helsinki Corpus* and the *Corpus of Late Modern English Medical Texts* (see Taavitsainen *et al*. 2014).

The TEI XML encoding of *CHELAR* is already in progress. The twentieth and nineteenth centuries are almost finished and we hope to conclude the process in the very near future.

5 Research possibilities

CHELAR was compiled with the explicit aim of complementing the legal material in *ARCHER 3.2*, and serving at the same time to fill a gap in the vast field of legal English corpora. As discussed earlier in this paper (Section 3), the electronic resources existing to date for the study of diachronic legal English have privileged genres such as statutes, whose function is prescriptive and regulatory, or have focused on oral forms of historical legal discourse such as trial transcripts and witness depositions. By making available a large body of texts representing a very different genre of legal writing – law reports and judicial decisions –, *CHELAR* will facilitate research in three of the four major 'trajectories' of corpus-based research on legal language identified by Biel (2010: 4–5), namely:[7]

Trajectory 1. External variation: how does legal language differ from general language and other languages for special purposes?

Trajectory 2. Internal variation: how do legal genres differ from each other?

Trajectory 3. Temporal variation: how does the current legal language differ from a historic one?

CHELAR is currently being piloted in work on the history of English (Trajectory 3) and on variation across the domain of legal genres (Trajectory 2; see below in this section). The corpus is clearly too small for the analysis of low-frequency phenomena, but other than this it can be fruitfully employed for investigating the development across time of numerous lexical, morphosyntactic and discoursal features.

Lexical features of various kinds – e.g., Latin words, Old French and Anglo-Norman words that have not found their way into general currency, heavy use of compound adverbs such as *hereof*, *whereof*, *hereinafter*, *heretofore*, etc., technical vocabulary unfamiliar to non-specialists, use of empirical verbs such as *find* or *determine* in place of propositional attitude verbs such as *think* or *believe* (cf. Alcaraz Varó 2007 [1994]: 77), etc. – constitute some of the manifestations of 'legalese' most frequently commented upon; they are the topic, for instance, of Mellinkoff's seminal monograph (1963) on the language of law. However, perhaps more interesting as suitable research topics, because less conspicuous, are certain features of grammar which have also been claimed to be distinctive of Present-day English written legal discourse: an extremely high rate of nominalization (e.g., *the provisions for the recovery of possessions* instead of *the provisions for recovering possessions*; quoted from Gotti 2003: 78), binomial and multinomial expressions (e.g., *within Singapore or elsewhere*; *by the Government or by government, public or local authority or by any person other than the person claiming relief*; quoted from Bhatia 1993: 108), lexical bundles and phraseological units (e.g., *the benefit of*, *as a matter of*, *it is clear that*, *on the basis that*, *be regarded as*, etc.; cf. Goźdź-Roszkowski 2011: 109–144), intricate patterns of coordination and subordination, impersonal style and frequent use of passive constructions, conditional constructions, overuse of certain modal verbs (e.g., *shall*, *may*) or, alternatively, of verb groups in the simple present or present perfect, rather than in the simple past or past perfect (Williams 2005: 150–156), long sentences (50 words on average; cf. Trosborg 1997: 13), heavy use of *any* as determiner (e.g., *any such underwriter for any legal and any other expenses*; cf. Goźdź-Roszkowski 2011: 13), etc. These and other features appear to be inextricably linked to the language of the law, and some, such as conditionals

and binomials, have been attested since Old English times (see Tiersma 1999: 15–16, Scotto di Carlo 2015: 13–17). It has to be said, however, that in the relevant literature exemplification of all such features draws heavily on prescriptive, legislative texts; this is in fact the only legal genre usually discussed in both classic studies (Crystal and Davy 1969: 193–217; Gustafsson 1975; Finegan 1982; Bhatia 1993: 105–117; Trosborg 1997; Williams 2005: 31–38, 113–165) and more recent treatments (Scotto di Carlo 2015: 29–55). Legislative texts constitute, after all, "the hard core" (Bhatia 1987: 230) of all the written varieties of legal language, and those with the most distinctive style.

Some of the morphosyntactic features just mentioned have also been examined with reference to the legal language of the past, either in general overviews of the genre such as Hiltunen (1990) or in more specialised studies. Among the latter we can mention Rissanen (1999), who uses the Statutes of the Realm (1488–1699) to examine the conjunctive use of *except* and compactness of expression at the noun phrase level; Kohnen (2001), who discusses frequency developments in one type of subordinate construction (the postmodyfying participial construction) in Late Middle English and Early Modern English petitions and statutes; Gotti (2001), on the semantic and pragmatic values of *shall* and *will*, also in Early Modern English statutes; Facchinetti (2001), on conditional constructions in a corpus of British and American English legal texts ranging from 1500 to 1800 and comprising both statutory writings and law reports; Bugaj (2006), on binomials in Scots and English burgh records, acts of parliament and statutes over the period 1500–1570; Kopaczyk (2009, 2013), on binomials and various kinds of lexical bundles and formulaic patterns in Scots legislative writing from 1380 to 1560; Lehto (2013), on coordination, subordination, binomials, and other complexity features in parliamentary acts and proclamations dating from 1491 to 1707.

Most of the above analyses rely on corpora such as the *Helsinki Corpus of English Texts* (Rissanen 1991) and the *Helsinki Corpus of Older Scots* (Meurman-Solin 1995), or on extended versions of these built on the same principles and with similar materials, a factor which has no doubt determined the emphasis on legislative writing.[8] One of the obvious advantages of *CHELAR*, therefore, is that it enables research along the lines that we identified previously as Biel's (2010) Trajectory 2; in other words, *CHELAR* will make it possible to compare the linguistic and textual findings in the above mentioned studies, and in other similar studies that might be published in the future, with the findings obtained in the very different category of legal writing (expository, juridical writing) which law reports represent. In connection with this, Facchinetti's (2001) analysis of conditional constructions already reveals the extent of the differ-

ences that may be found when comparing earlier legislative writing and juridical writing. For her study she employs two different samples, one drawn from the statutes in the *Helsinki Corpus of English Texts*, and the other from *ARCHER 1*, this consisting of reports of law cases discussed in American courts. Conditional constructions figure prominently in both corpora (2001: 147), but Facchinetti finds dramatic differences in usage between the conditionals in the *Helsinki Corpus* and those in *ARCHER 1*: conditionals in the *Helsinki Corpus* are practically all 'normative' conditionals containing the modal *shall* in the apodosis (and also often in the protasis) and are introduced by performative formulas like *be it enacted* / *ordained that*, as in (4). Conditionals in *ARCHER 1*, by contrast, are mostly non-normative, express the speaker's point of view, and show very high percentages of the modal *would* in the apodosis, as in (5):

(4) Provided and *bee it enacted* by the Authority aforesaid That *if* such person [...] *shall* not happen to be the Goaler or Keeper of such Goal or Prison [...] that then the said Justice [...] *shall* administer and give to such Person [...] an Oath to the Effect following (*CELAW3*)

(5) For the Commonwealth, it was answered, that *if* the present attempt was successful, it *would* prostrate the authority of the individual states (*ARCHER* 1798)

Finally, another promising line of research for which *CHELAR* seems particularly well suited is Multi-Dimensional (MD) analysis (Biber 1988, 1995, 2001, 2013, etc.; Biber and Finegan 2001). This is clearly not the place for a full discussion of Biber's model, whose background concepts and methodology are well known. The framework proposed by Biber (1988) and further developed in many subsequent publications examines register variation in terms of six *dimensions* conceived as groupings of linguistic features (a total of 67 features, in the original 1988 MD model) that co-occur with a markedly high frequency in texts. The dimensions have since been applied statistically to analyze register variation in specialised discourse domains (e.g., university spoken and written registers, Biber 2006; conversational text types, Biber 2008; written legal registers, Goźdź-Roszkowski 2011; 19th century fiction, Egbert 2012, etc.) and in an extensive set of languages other than English (for a complete listing of these, see Biber 2013: xxxii). MD analysis has also been used for diachronic research, tracing the evolution of registers in English (e.g., Biber 1995: 283–300; Biber and Finegan 2001; Claridge and Wilson 2002; Geisler 2002) and other languages (e.g., Somali; see Biber 1995: 300–311).

Biber's analyses of English diachronic registers cover the seventeenth to the late nineteenth centuries, and are based on *ARCHER*'s earliest version (*ARCHER 1*). The focus is on eight different genres, namely essays, fiction, letters, dialogue in drama, dialogue in novels, medical research articles from the *Edinburgh Medical Journal*, scientific research articles from the *Philosophical Transactions of the Royal Society of London*, and 'legal opinions' (i.e., law reports); the latter, however, are represented only by a relatively small sample of American English texts from the Pennsylvania Supreme Court dating from 1750 to 1899 (see Biber 1995: 88).[9] Biber's initial findings suggest that with respect to the dimensions relevant for narrative discourse (Dimension 2: 'Narrative vs Non-Narrative Concerns') and for the oral/literate continuum (Dimensions 1: 'Involved vs Informational Production', 3: 'Situation-Dependent vs Elaborated Reference', and 5: 'Non-impersonal vs Impersonal Style'), legal opinions "have followed a consistent course" (Biber and Finegan 2001: 82) towards a progressively less narrative and more literate style. They acknowledge, however, that "further research is required to test and refine these generalizations" (2001: 82). In connection with this, ongoing work at the VLCG Research Unit suggests that *CHELAR*, with its extensive diachronic coverage of nearly five centuries, is an ideal resource to examine variation in law reports over time, as well as variation relative to other legal genres, both writing-based and speech-based, for which diachronic corpora are now accessible, as discussed in Section 3 above. Mega-corpora and big databases have become increasingly available to linguists, but small and 'beautiful' corpora like *CHELAR* still "have some life left in them, and interesting new data to offer" (Mair 2013: 193).

Availability

For the conditions of use of *CHELAR*, interested researchers can contact Teresa Fanego (teresa.fanego@usc.es) or Paula Rodríguez-Puente (rodriguezp-paula@uniovi.es).

Reference line and copyright

Corpus of Historical English Law Reports 1535–1999 (CHELAR). 2016. Compiled by Paula Rodríguez-Puente, Teresa Fanego (Project Director), María José López-Couso, Belén Méndez-Naya and Paloma Núñez-Pertejo. University of Santiago de Compostela: Research Unit for Variation, Linguistic Change and Grammaticalization, Department of English and German. ISBN: 978–84–608–8006–6.

Acknowledgements

For generous financial support, we are grateful to the European Regional Development Fund and the Spanish Ministry of Economy and Competitiveness (grants HUM2007–60706, FFI2014–52188–P, FFI2014–51873–REDT, BES–2012–057555 and BES–2015–071233). Thanks are also due to Merja Kytö and two anonymous reviewers of *ICAME Journal* for valuable comments on an earlier version.

Notes

1. The distinction first formulated by Biber (1988: 70, 170 and elsewhere) between *genre* as a category of text identified on the basis of external criteria such as subject-matter, author's purpose or the relation between the communicative participants, and *text type* as a grouping of texts that are similar in their linguistic form, is not observed in this article. *Genre*, *text type* and the related term *register* ("a variety associated with a particular situation of use", Biber and Conrad 2009: 6) are therefore used largely interchangeably throughout the following pages. For discussion of these various labels, the reader is referred to Diller (2001), Moessner (2001), Biber and Conrad (2009), Culpeper and Kytö (2010: 21–23) and Claridge (2012), among many others.
2. *CHELAR* arose as a result of the initiative of the third author (López-Couso) while she and Méndez-Naya were compiling the British English legal texts for *ARCHER 3.2*. During the early stages (2011–2012) of work on *CHELAR* López-Couso and Rodríguez-Puente, the latter an FPI postgraduate researcher at the VLCG Research Unit at the time, coordinated a team of students comprising Zeltia Blanco-Suárez, Eduardo Coto, Tania de Dios, Iria Pastor, Alba Pérez-González, Paula Rodríguez-Abruñeiras, Iria Gael Romay and Vera Vázquez. Work on the corpus was resumed in July 2015 and has since been coordinated by the first and second authors (Fanego and Rodríguez-Puente), who have profited from the close collaboration of Cristina Blanco-García, Iván Tamaredo, Noelia Castro-Chao and Daniela Pettersson-Traba as research assistants. Fanego and Rodríguez-Puente were also chiefly responsible for writing this article.
3. A few samples of legal documents, such as wills, witness depositions, and court records of defamation cases, can also be found in Cusack's (1998) useful collection of 64 non-literary texts from the Early Modern English period.

4. An anonymous reviewer enquires whether additional textual resources would be available to extend the current size of *CHELAR*, so as to make it less prone to statistical noise. We are aware that its relatively modest size renders the corpus inadequate for the study of low-frequency phenomena, as more fully discussed in Section 5 below; however, for the analysis of a good many morphosyntactic and discoursal features, half a million words seems sufficient. In addition, the available textual materials for the earlier part of the Early Modern English period are scanty and would make it difficult to build a balanced corpus, as was our goal.
5. See CLAWS Input/Output Format Guidelines at <http://ucrel.lancs.ac.uk/claws/format.html>.
6. See <http://ucrel.lancs.ac.uk/vard/about/>.
7. A fourth trajectory, cross-linguistic variation, would involve the use of comparable corpora with components in at least two languages.
8. Due to limitations of space, we cannot refer here to the many important studies that have looked at the language of speech-based genres such as trial proceedings and witness depositions; see, among many others, Archer (2005), Grund (2007), Huber (2007), Culpeper and Kytö (2010), Kytö, Grund and Walker (2011), Widlitzki and Huber (2016), etc. Needless to say, these and other publications have afforded insights into oral legal discourse which could, in many cases, be compared and contrasted with data drawn from *CHELAR*.
9. *ARCHER* has a complex textual history: the original version (*ARCHER 1*) contained only American English legal texts. British English legal texts were added by the VLCG team during the compilation of *ARCHER 3.2* over the period 2009–2013; for details, see Yáñez-Bouza (2011) and López-Couso and Méndez-Naya (2012: 9).

Corpora and electronic resources

ARCHER 3.2. A Representative Corpus of Historical English Registers, version 3.2 (1990–1993/2002/2007/2010/2013). Originally compiled under the supervision of Douglas Biber and Edward Finegan at Northern Arizona University and University of Southern California; modified and expanded by subsequent members of a consortium of universities. Current member universities are Bamberg, Freiburg, Heidelberg, Helsinki, Lancaster, Leicester, Manchester, Michigan, Northern Arizona, Santiago de Compostela, Southern California, Trier, Uppsala, Zurich. <http://www.manchester.ac.uk/archer/>

Bray, Tim, Jean Paoli, C.M. Sperberg-McQueen, Eve Maler and François Yergeau (eds.). 2008. *Extensible Markup Language (XML) 1.0*. Fifth edition. W3C Recommendation 26 November 2008. <https://www.w3.org/XML/>

Cambridge Legal English Corpus. Official website: <http://www.cup.es/>

Corpora of Early English Correspondence. 1993– ongoing. Project leader: Terttu Nevalainen.
<http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/>

Corpus of Early English Medical Writing. 1995– ongoing. Project leader: Irma Taavitsainen.
<http://www.helsinki.fi/varieng/CoRD/corpora/CEEM/>

Corpus of English Dialogues 1560–1760. 2006. Compilers: Merja Kytö and Jonathan Culpeper.
<http://www.helsinki.fi/varieng/CoRD/corpora/CED/>

Corpus of English Religious Prose. 2003– ongoing. Compilers: Thomas Kohnen, Tanja Rütten, Ingvilt Marcoe, Kirsten Gather, Dorothee Groeger, Anne Döring, Stefanie Leu.
<http://www.helsinki.fi/varieng/CoRD/corpora/COERP/>

Coruña Corpus of English Scientific Writing. 2004– ongoing. Project leader: Isabel Moskowich.
<http://www.helsinki.fi/varieng/CoRD/corpora/Coruna/>

Encyclopædia Britannica Online. Available at <http://www.britannica.com>

Helsinki Corpus of English Texts. 1991. Project leader: Matti Rissanen.
<http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/>

Helsinki Corpus of Older Scots. 1995. Compiler: Anneli Meurman-Solin.
<http://www.helsinki.fi/varieng/CoRD/corpora/HCOS/>

Lampeter Corpus of Early Modern English Tracts. 1999. Compilers: Josef Schmied, Claudia Claridge and Rainer Siemund.
<http://www.helsinki.fi/varieng/CoRD/corpora/LC/>

Málaga Corpus of Late Middle English Scientific Prose. 2012–2015. Project leaders: Javier Calle-Martín and Antonio Miranda-García.
<http://www.helsinki.fi/varieng/CoRD/corpora/SciProse/>

OED = Oxford English Dictionary Online. <http://www.oed.com/> (accessed 20 April 2016).

Old Bailey Corpus = Huber, Magnus, Magnus Nissel, Patrick Maiwald and Bianca Widlitzki. 2012. *The Old Bailey Corpus*. *Spoken English in the 18th and 19th centuries*. <http://www.uni-giessen.de/oldbaileycorpus>

Penn-Helsinki Parsed Corpus of Early Modern English. 2004. Compilers: Anthony Kroch, Beatrice Santorini and Lauren Delfs. <http://www.helsinki.fi/varieng/CoRD/corpora/PPCEME/>

Penn-Helsinki Parsed Corpus of Modern British English. 2010. Compilers: Anthony Kroch, Beatrice Santorini and Ariel Diertani. <http://www.helsinki.fi/varieng/CoRD/corpora/PPCMBE/>

TEI Consortium. 2016. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. <http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf>

References

Alcaraz Varó, Enrique. 2007 [1994]. *El inglés jurídico: textos y documentos.* Barcelona: Ariel.

Archer, Dawn. 2005. *Historical sociopragmatics: Questions and answers in the English courtroom (1640–1760).* Amsterdam: John Benjamins.

Archer, Dawn. 2014. Historical pragmatics: Evidence from the Old Bailey. *Transactions of the Philological Society* 112: 259–277.

Archer, Dawn, Merja Kytö, Alistair Baron and Paul Rayson. 2015. Guidelines for normalising Early Modern English corpora: Decisions and justifications. *ICAME Journal* 39: 5–24.

Atkins, Susan, Jeremy Clear and Nicholas Ostler. 2007. Corpus design criteria. In W. Teubert and R. Krishnamurthy (eds.). *Corpus linguistics: Critical concepts in linguistics,* Vol. II, 99–133. London: Routledge.

Baron, Alistair and Paul Rayson. 2009. Automatic standardization of texts containing spelling variation, how much training data do you need? In M. Mahlberg, V. González-Díaz and C. Smith (eds.). *Proceedings of the Corpus Linguistics Conference, CL2009, University of Liverpool, UK, 20–23 July 2009*. Liverpool: University of Liverpool.
Available at <http://ucrel.lancs.ac.uk/publications/cl2009/>

Bhatia, Vijay K. 1987. Language of the law. *Language Teaching* 20(4): 227–234.

Bhatia, Vijay K. 1993. *Analysing genre. Language use in professional settings.* Harlow, Essex: Pearson Education.

Biber, Douglas. 1988. *Variation across speech and writing.* Cambridge: Cambridge University Press.

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic perspective.* Cambridge: Cambridge University Press.

Biber, Douglas. 2001. Dimensions of variation among 18th century registers. In H-J. Diller and M. Görlach (eds.). *Towards a history of English as a history of genres*, 89–110. Heidelberg: C. Winter.

Biber, Douglas. 2006. *University language: A corpus-based study of spoken and written registers.* Amsterdam: John Benjamins.

Biber, Douglas. 2008. Corpus-based analyses of discourse: Dimensions of variation in conversation. In V. Bhatia, J. Flowerdew and R. Jones (eds.). *Advances in discourse studies*, 100–114. London: Routledge.

Biber, Douglas. 2013. Multi-dimensional analysis. A personal history. In T. Berber Sardinha and M. Veirano Pinto (eds.). *Multi-dimensional analysis, 25 years on. A tribute to Douglas Biber*, xxix–xxxviii. Amsterdam: John Benjamins.

Biber, Douglas and Edward Finegan. 2001. Diachronic relations among speech-based and written registers in English. In S. Conrad and D. Biber (eds.). *Variation in English: Multi-dimensional studies*, 66–83. Harlow, Essex: Longman/Pearson Education.

Biber, Douglas and Susan Conrad. 2009. *Register, genre, and style*. Cambridge: Cambridge University Press.

Biel, Łucja. 2010. Corpus-based studies of legal language for translation purposes: Methodological and practical potential. In C. Heine and J. Engberg (eds.). *Reconceptualizing LSP. Online proceedings of the XVII European LSP Symposium 2009*. Aarhus: Aarhus University. [No pagination]

Bowker, Lynne and Jennifer Pearson. 2002. *Working with specialized language: A practical guide to using corpora.* London: Routledge.

Bugaj, Joanna. 2006. The language of legal writings in 16th century Scots and English: An etymological study of binomials. *ESP Across Cultures* 6: 7–22.

Claridge, Claudia. 2012. Linguistics levels: Styles, registers, genres, text types. In A. Bergs and Laurel J. Brinton (eds.). *English historical linguistics. An international handbook.* Vol. 1, 237–253. Berlin: Mouton de Gruyter.

Claridge, Claudia and Andrew Wilson. 2002. Style evolution in the English sermon. In T. Fanego, B. Méndez-Naya and E. Seoane (eds.). *Sounds, words, texts and change. Selected Papers from 11 ICEHL, Santiago de Compostela, 7–11 September 2000*, 25–44. Amsterdam: John Benjamins.

Cornish, William, J. Stuart Anderson, Ray Cocks, Michael Lobban, Patrick Polden and Keith Smith. 2010. *The Oxford history of the laws of England.* Vol. XI. Oxford: Oxford University Press.

Crystal, David and Derek Davy. 1969. *Investigating English style*. London: Longman.

Culpeper, Jonathan and Merja Kytö. 2010. *Early Modern English dialogues. Spoken interaction as writing.* Cambridge: Cambridge University Press.

Cusack, Bridget. 1998. *Everyday English 1500–1700. A reader.* Edinburgh: Edinburgh University Press.

Diller, Hans-Jürgen. 2001. *Genre* in linguistic and related discourses. In H-J. Diller and M. Görlach (eds.). *Towards a history of English as a history of genres*, 3–43. Heidelberg: C. Winter.

Egbert, Jesse. 2012. Style in nineteenth century fiction: A multi-dimensional analysis. *Scientific Study of Literature* 2: 167–198.

Facchinetti, Roberta. 2001. Conditional constructions in Modern English legal texts. In M. Gotti and M. Dossena (eds.). *Modality in specialized texts*, 133–150. Bern: Peter Lang.

Finegan, Edward. 1982. Form and function in testament language. In R.J. Di Pietro (ed.). *Linguistics and the professions. Proceedings of the Second Annual Delaware Symposium on Language Studies*, 113–120. Norwood, N. J.: Ablex.

Garside, Roger. 1987. The CLAWS word-tagging system. In R. Garside, G. Leech and G. Sampson (eds.). *The computational analysis of English: A corpus-based approach*, 30–41. London: Longman.

Geisler, Christer. 2002. Investigating register variation in nineteenth-century English: A multidimensional comparison. In R. Reppen, S.M. Fitzmaurice and D. Biber (eds.). *Using corpora to explore linguistic variation*, 249–271. Amsterdam: John Benjamins.

Görlach, Manfred. 1991. Text types and the linguistic history of Modern English. In C. Uhlig and R. Zimmermann (eds.). *Anglistentag 1990 Marburg. Proceedings*, 195–215. Tübingen: Max Niemeyer.

Görlach, Manfred. 1992. Text type and language history: The cookery recipe. In M. Rissanen, O. Ihalainen, T. Nevalainen and I. Taavitsainen (eds.). *History of Englishes: New methods and interpretations in historical linguistics*, 736–761. Berlin: Mouton de Gruyter.

Gotti, Maurizio. 2001. Semantic and pragmatic values of *shall* and *will* in Early Modern English statutes. In M. Gotti and M. Dossena (eds.). *Modality in specialized texts*, 89–111. Bern: Peter Lang.

Gotti, Maurizio. 2003. *Specialized discourse. Linguistic features and changing conventions*. Bern: Peter Lang.

Goźdź-Roszkowski, Stanisław. 2011. *Patterns of linguistic variation in American legal English. A corpus-based study.* Frankfurt am Main: Peter Lang.

Grover, Claire, Ben Hachey and Ian Hughson. 2004. The HOLJ corpus: Supporting summarisation of legal texts. In S. Hansen-Shirra, S. Oepen and H. Uszkoreit (eds.). *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora (LINC-04), Geneva, Switzerland*. Geneva: University of Geneva. [No pagination]

Grund, Peter. 2007. From tongue to text: The transmission of the Salem witchcraft examination records. *American Speech* 82: 119–150.

Gustafsson, Marita. 1975. *Some syntactic properties of English law language.* Turku: Department of English, University of Turku.

Hiltunen, Risto. 1990. *Chapters on legal English. Aspects past and present of the language of the law.* Helsinki: Suomalainen Tiedeakatemia.

Hiltunen, Turo and Jukka Tyrkkö. 2013. Tagging Early Modern English medical texts (1500–1700). Paper presented at CANS 2013 ('Corpus Analysis with Noise in the Signal'), a workshop held within the 7th Corpus Linguistics Conference (CL2013), Lancaster University, 22nd July 2013.

Huber, Magnus. 2007. The *Old Bailey Proceedings*, 1674–1834. Evaluating and annotating a corpus of 18th- and 19th-century spoken English. In A. Meurman-Solin and A. Nurmi (eds.). *Annotating variation and change* (Studies in Variation, Contacts and Change in English, Vol. 1). Helsinki: Research Unit for Variation, Contacts and Change in English, University of Helsinki.

Kearns, Martin. 2007. *Legal English*. Madrid: Colex.

Kohnen, Thomas. 2001. On defining text types within historical linguistics: The case of petitions/statutes. In L. Moessner (ed.). *Early Modern English text types*. Special issue of *European Journal of English Studies* 5: 197–203.

Kopazcyk, Joanna. 2009. Multi-word units of meaning in 16th-century legal Scots. In R.W. McConchie, A. Honkapohja and Jukka Tyrkkö (eds.). *Selected Proceedings of the 28th Symposium on New Approaches in English Historical Lexis (HEL-LEX 2)*, 88–95. Somerville, MA: Cascadilla Proceedings Project.

Kopazcyk, Joanna. 2013. *The legal language of Scottish burghs. Standardization and lexical bundles (1380–1560)*. Oxford: Oxford University Press.

Kytö, Merja. 1996. *Manual to the diachronic part of the Helsinki Corpus of English Texts. Coding conventions and lists of source texts.* 3rd edition. Helsinki: Department of English, University of Helsinki.

Kytö, Merja and Matti Rissanen. 1992. A language in transition: The Helsinki Corpus of English Texts. *ICAME Journal* 16: 7–27.

Kytö, Merja, Peter J. Grund and Terry Walker. 2011. *Testifying to language and life in Early Modern England*. *Including CD-ROM: An electronic text edition of depositions 1560–1760 (ETED)*. Amsterdam: John Benjamins.

Lehto, Anu. 2013. Complexity and genre conventions. Text structure and coordination in Early Modern English proclamations. In A.H. Jucker, D. Landert, A. Seiler and N. Studer-Joho (eds.). *Meaning in the history of English. Words and texts in context*, 233–256. Amsterdam: John Benjamins.

Lehto, Anu, Alistair Baron, Maura Ratia and Paul Rayson. 2010. Improving the precision of corpus methods: The standardized version of Early Modern English Medical Texts. In I. Taavitsainen and P. Pahta (eds.). *Early Modern English Medical Texts: Corpus description and studies*, 279–290. Amsterdam: John Benjamins.

López-Couso, María José and Belén Méndez-Naya. 2012. Compiling British English legal texts: A contribution to ARCHER. In N. Vázquez (ed.). *Creation and use of historical English corpora in Spain*, 5–19. Newcastle upon Tyne: Cambridge Scholars Publishing.

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

Marín Pérez, María José and Camino Rea Rizzo. 2012. Structure and design of the British Law Report Corpus (BLRC): A legal corpus of judicial decisions from the UK. *Journal of English Studies* 10: 131–145.

Mellinkoff, David. 1963. *The language of the law.* Boston, MA: Little, Brown and Company.

Moessner, Lilo (ed.). 2001. *Early Modern English text types*. Special issue of *European Journal of English Studies* 5: 131–256.

Rissanen, Matti. 1999. Language of law and the development of Standard English. In I. Taavitsainen, G. Melchers and P. Pahta (eds.). *Writing in nonstandard English*, 189–203. Amsterdam: John Benjamins.

Rodríguez-Puente, Paula. 2011. Introducing the Corpus of Historical English Law Reports: Structure and compilation techniques. *Revista de Lenguas para Fines Específicos* 17: 99–120.

Šarčević, Susan. 2000. *New approach to legal translation*. The Hague: Kluwer Law International.

Scotto di Carlo, Giuseppina. 2015. *Diachronic and synchronic aspects of legal English: Past, present, and possible future of legal English.* Newcastle upon Tyne: Cambridge Scholars Publishing.

Taavitsainen, Irma and Päivi Pahta (eds.). 2010. *Early Modern English Medical Texts: Corpus description and studies*. Amsterdam: John Benjamins.

Taavitsainen, Irma, Turo Hiltunen, Anu Lehto, Ville Marttila, Päivi Pahta, Maura Ratia, Carla Suhr and Jukka Tyrkkö. 2014. *Late Modern English Medical Texts 1700–1800*: A corpus for analysing eighteenth-century medical English. *ICAME Journal* 38: 137–153.

Tiersma, Peter M. 1999. *Legal language*. Chicago, IL: The University of Chicago Press.

Trosborg, Anna. 1997. *Rhetorical strategies in legal language. Discourse analysis of statutes and contracts.* Tübingen: Gunter Narr Verlag.

Widlitzki, Bianca and Magnus Huber. 2016. Taboo language and swearing in eighteenth and nineteenth century English: A diachronic study based on the *Old Bailey Corpus*. In M.J. López-Couso, B. Méndez-Naya, P. Núñez-Pertejo and I.M. Palacios-Martínez (eds.). *Corpus linguistics on the move: Exploring and understanding English through corpora*, 313–336. Amsterdam: Brill/Rodopi.

Williams, Christopher. 2005. *Tradition and change in legal English. Verbal constructions in prescriptive texts.* Bern: Peter Lang.

Yáñez-Bouza, Nuria. 2011. ARCHER past and present (1990–2010). *ICAME Journal* 35: 205–236.

Appendix

CHELAR: Source texts and word counts. 2016. Compiled by Paula Rodríguez-Puente. Available online on the VLCG homepage (http://www.usc-vlcg.es/).