Creation and Use of Historical English Corpora in Spain

Creation and Use of Historical English Corpora in Spain

Edited by

Nila Vázquez

CAMBRIDGE SCHOLARS PUBLISHING

Creation and Use of Historical English Corpora in Spain,
Edited by Nila Vázquez

This book first published 2012

Cambridge Scholars Publishing

12 Back Chapman Street, Newcastle upon Tyne, NE6 2XX, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Copyright © 2012 by Nila Vázquez and contributors

All rights for this book reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the copyright owner.

ISBN (10): 1-4438-4251-6, ISBN (13): 978-1-4438-4251-8

TABLE OF CONTENTS

PREFACE

TERESA FANEGO
UNIVERSITY OF SANTIAGO DE COMPOSTELA

The establishment of English Studies in Spain dates back to 1952, when the Spanish Ministry of Education and Science sanctioned the first university syllabus of English Philology, initially at the University of Salamanca (1952), and from 1954 also at the Complutense University of Madrid. From the very beginnings, the initial connections of English Studies with departments of Romance languages, where the philological component was strong, ensured that the English curricula offered at Spanish universities comprised several compulsory courses on the history of English. Interest in the field was also assisted by the creation of a panel on English historical linguistics at the conferences of AEDEAN, the Spanish Association for Anglo-American Studies, and by the founding, in the 1980s, of the Spanish Society for Medieval English Language and Literature (SELIM) and the Spanish and Portuguese Society for English Renaissance Studies (SEDERI). These organisations served as convenient fora for the presentation of historical research by Spanish scholars.

The outcome of these various developments is that over the last twenty or thirty years innovative research on English historical linguistics has been carried out in Spain, an area which had traditionally been outside the centre of gravitation of the discipline. A new generation of scholars has emerged who have published a significant number of high-quality books and articles, both in the more traditional fields of English historical linguistics (historical lexicology and lexicography, dialectology, textbooks and teaching materials, textual editing) and in new research fields, such as grammaticalization and mechanisms of change, text types and specialized discourse, historical sociolinguistics, historical pragmatics, and corpus linguistics.

With regard to the area of corpus linguistics in particular, Spanish scholars were among the first to make use in their investigations of the Helsinki Corpus of English Texts (1991) and other pioneering historical corpora of the English language; witness in this regard early research such as Seoane (1993), Fanego (1996), López-Couso (1996) and Núñez-Pertejo (1996), all based on data retrieved from the Helsinki Corpus. However, the

establishment in 2008 of the Spanish Association for Corpus Linguistics (AELINCO) has no doubt greatly contributed to enhancing the visibility of corpus-based research in Spain and giving it a new dimension, in that AELINCO has actively promoted the compilation of both historical and non-historical corpora.

The present volume, edited by Nila Vázquez, testifies to this fact. The essays in Part I ('Creation of historical English corpora') detail the compilation of seven historical corpora remarkably diverse: British legal texts, Early English recipes, Middle English scientific prose, Northern English, astronomy texts, Late Modern English literary dialects, and Late Modern English British and American Prose. No less specialized are the studies comprised in Part II ('Use of historical English corpora'), which approach topics such as the complementation of adjectives in OE, verbs of saying in Late Modern English, relativization in New Englishes across the diachronic dimension, lexis and semantics in Middle English medical texts, gender distinctions in earlier English, evidentials in medical texts, determiners in Lancashire English, etc. One can only hope that such concerted activity will continue, serving to consolidate and strengthen Spanish research into historical corpora and historical linguistics for the future.

References:

Fanego, Teresa. 1996. The gerund in Early Modern English: Evidence from the Helsinki Corpus. *Folia Linguistica Historica* 17: 97-152.

Kytö, Merja. 1991. *Manual to the diachronic part of the Helsinki Corpus of English Texts: Coding conventions and lists of source texts.* Helsinki: Department of English, University of Helsinki.

López-Couso, María José. 1996. On the history of *methinks*: From impersonal construction to fossilised expression. *Folia Linguistica Historica* 17: 153-169.

Núñez-Pertejo, Paloma. 1996. On the origin and history of the English prepositional type *a-hunting*. *Revista Alicantina de Estudios Ingleses* 9: 105-117.

Seoane, Elena. 1993. The passive in Early Modern English. *Atlantis* 15/2: 191-213.

INTRODUCTION

NILA VÁZQUEZ[1]
UNIVERSITY OF MURCIA

After the creation of the Spanish Association for Corpus Linguistics (AELINCO) and the celebration of its first international conference at the University of Murcia, I realised that there were several research groups in Spain either working in the compilation of historical English corpora or using these new corpora and previous material compiled in other countries for their own research. However, their work had not reached all the attention it deserved, so I thought that it was necessary to bring them all together into a special volume to be widely distributed. I contacted these scholars and all of them agreed to be part of this project. They also thought that the volume would be a great contribution to our field as there was nothing like this available for our scholarship and it was necessary to put all this information together. Even before the *Helsinki Corpus* was published, Spain had a good amount of Historical English researchers interested in the use of corpora in their research, as for instance the group directed by Teresa Fanego in Santiago de Compostela. In the last decades, the number of scholars working on the field of Historical Corpus Linguistics has increased and, nowadays, there are some attractive projects in progress that will result in the publication of valuable material for scholars of the whole world.

The aim of this volume is twofold. On the one hand, the first section describes the work of the aforementioned research groups who have compiled corpora such as the Colmobaeng (*Corpus of Late Modern British and American English Prose*, University of Santiago de Compostela), CoER (*Corpus of Early English Recipes*, University of Las

[1] Nila Vázquez is part of the research groups *Variation, Linguistic Change and Grammaticalization* (U. Santiago de Compostela) and *Filología Inglesa y Lingüística Histórica* (U. Murcia). She is currently involved in the research project *Constructionalization and Grammaticalization in English* (grants HUM2007-60706 and FFI2011-26693-C02-01 are hereby gratefully acknowledged).

Palmas de Gran Canaria) or the *Coruña Corpus* (a corpus of scientific documents that will help in diachronic studies of specific types of discourse in English, University of Coruña), among others. The output of these projects will also throw light on the dialectology of early periods of the English Language and will help, for instance, to correct some mistakes in dating words in the *OED* (some have already been suggested by members of the *Salamanca Corpus*).[2] With this new material, scholars in Europe and other continents who work on this field will have a great amount of new information for their research.

The second part of the volume is devoted to present the work of some of the Spanish scholars who work with Historical English Corpora, either the Spanish ones or foreign material, such as the *Helsinki Corpus*, the *Corpus of Early Correspondence* or the *Corpus of Late Modern English Texts*, among others.

With these two complementary parts, the volume aims at offering a comprehensive view of the studies on Historical English Corpora in Spain. The contents of the volume are as follows:

Part 1: Creation

Research Groups (RG) working in the compilation of corpora in Spain or independent scholars who have compiled their own corpus (alphabetical order according to the institutional affiliation or main researcher):

International
- RG University of Santiago de Compostela (María José López Couso and Belén Méndez Naya)
Compiling British English Legal Texts: A Contribution to ARCHER

National
- RG University of A Coruña (Isabel Moskowich Spiegel Fandiño)
A 'Smooth Homogeneous Globe' in CETA: Compiling Late Modern Astronomy Texts in English

[2] Ruano-García (2010c: 381, 441) mentions some examples where the *Salamanca Corpus* adds an interdating to the *OED* for the word *gloppen* and an antedating for *rannel-bawke*. As for the latter, the first evidence shown in the *OED* online dates to 1790 and the one offered by the *Salamanca Corpus* is 1673, found in *A Yorkshire dialogue between an awd wife a lass and a butcher*. (Cf. References, Chapter 20, this volume).

- RG University of Las Palmas de Gran Canarias (Francisco Alonso Almeida, Ivalla Ortega Barrera and Elena Quintana Toledano)
Corpus of Early English Recipes: Design and Implementation

- RG University of Málaga (Antonio Miranda and Javier Calle)
Compiling the Malaga Corpus of Late Middle English Scientific Prose

- RG University of Salamanca (Mª Fuencisla García-Bermejo Giner)
The Online Salamanca Corpus of English Dialect Texts

- RG University of Seville (Julia Fernández Cuesta and José Gabriel Amores Carredano)
The SCONE Corpus of Northern English

- Teresa Fanego (U. of Santiago de Compostela)
COLMOBAENG: A Corpus of Late Modern British and American English Prose

Part 2: Use

Researchers using historical English corpora, either compiled in Spain or abroad (alphabetical order according to the main researcher)

- Alejandro Alcaraz Sintes (U. Jaén)
Complementation of Noun-modifying Adjectives in Old English

- Javier Calle and Antonio Miranda (U. Málaga)
Annotating the Corpus of the Federalist Papers*: A Resource for Attributing Authorship*

- Juan Camilo Conde (U. Murcia)
Diachronic Corpora as Sources for the Study of Variation in the History of Languages: Strengths and Weaknesses

- Begoña Crespo García (U. A Coruña)
Communicating Astronomy in the 19th Century: Verbs of Saying in CETA

- Isabel de la Cruz Cabanillas (U. Alcalá)
On the Evolution of Some Old English Prefixes: The Case of SAM-

- Javier Enrique Díaz Vera (U. Castilla la Mancha)
Metaphors of Pain in Middle English Medical Texts: A Corpus-based Approach

- Trinidad Guzmán González (U. León)
Assigned Gender in 18th-century English Prose: A Corpus Study

- Juan Manuel Hernández Campoy (U. Murcia)
Gender, Polarity and Mood in Late Medieval England: Evidence from The Paston Family

- María José López Couso and Belén Méndez Naya (U. Santiago)
On Comparative Complementizers in English: Evidence from Historical Corpora

- Javier Martín Arista (U. Rioja)
Old English Lexical Primes: Corpus Analysis and Database Compilation

- Ivalla Ortega Barrera (ULPGC) and Elena Quintana Toledo
Good *as an Early Lexical Evidential Marker in Medical Texts*

- Javier Pérez Guerra (U. Vigo)
A Corpus-based Analysis of It-*clefts in the Recent History of English: What is't You Lack?*

- Javier Ruano García (U. Salamanca)
Th'Monn, Twoman *and* T'felly*: On the Definite Article in Traditional Lancaster English*

- Pilar Sánchez García (U. Salamanca)
Depicting Northern English Dialects through Spelling. A Study Based on the Salamanca Corpus

- Cristina Suárez Gómez (U. Illes Balears)
The Study of New Englishes as a Window to the History of English: a Corpus-based Study of Relativization

CHAPTER ONE

COMPILING BRITISH ENGLISH LEGAL TEXTS: A CONTRIBUTION TO ARCHER[1]

MARÍA JOSÉ LÓPEZ-COUSO
BELÉN MÉNDEZ-NAYA
UNIVERSITY OF SANTIAGO DE COMPOSTELA

1. Introduction

ARCHER, *A Representative Corpus of Historical English Registers*, is a multi-purpose diachronic corpus which represents different text-types of British and American English between the seventeenth century and the present day. It thus serves as a suitable complement to the diachronic part of the *Helsinki Corpus of English Texts*, which contains material from different genres from Early Old English to the year 1710.[2] Over the years, ARCHER has therefore become one of the most widely used resources for the computer-driven analysis of register-based variation in the Late Modern English period.

The aim of the present chapter is to report on the contribution made by the research group *Variation, Linguistic Change and Grammaticalization* (VLCG) to the compilation of a new version of ARCHER by collecting British English legal texts from the early seventeenth century to the end of

[1] We gratefully acknowledge funding from the following institutions: Spanish Ministry for Science and Innovation (grants HUM2007-60706 and FFI2011-26693-C02-01) and Autonomous Government of Galicia (grants CN2011/011 and CN2012/012).

[2] For more information on the *Helsinki Corpus*, see Kytö (1996) and the Corpus Resource Database (CoRD) at http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/index.html.

the twentieth century. We begin with a general overview of the ARCHER project, then describe the kind of legal texts included in the corpus, before discussing various issues and problems that the project team encountered during the collection and editing of material.

2. The ARCHER Project[3]

The original version of ARCHER, now known as ARCHER-1, was compiled in the early 1990s by Douglas Biber (Northern Arizona University) and Edward Finegan (University of Southern California). Its main purpose was "to enable analysis of historical change in the range of written and speech-based registers of English from 1650 to the present. The general design goal has thus been to represent as wide a range of variation as possible" (Biber et al. 1994: 3; cf. also Biber and Finegan 1997: 255). ARCHER-1 is divided into 50-year sub-periods, starting from 1650 for the British English material. The coverage for American English, however, is less exhaustive, since this version of the corpus only includes texts corresponding to the second half of the eighteenth, nineteenth, and twentieth centuries. As a multi-genre corpus, ARCHER-1 contains a wide variety of registers: drama, fiction (both fiction prose and fiction dialogue), journals/diaries, legal opinions, letters, medicine, news, science, and sermons. The corpus thus includes texts ranging from the more formal kinds of writing (e.g. science) to the more informal (e.g. letters), and from written registers (e.g. legal opinions) to those showing a higher degree of speechlikeness (e.g. drama). In all, the corpus consists of 1,037 texts and around 1.7 million words, with a target sampling of ten texts of approximately 2,000 words per genre, variety, and sub-period.

A second edition of the corpus, ARCHER-2, was completed in 2004-2005. Built on ARCHER-1, this version included an extension of the corpus, which now contained 1,074 files and around 2,290,000 words. The original make-up of ARCHER was modified by the addition of some new texts: (i) American English material from drama, fiction, and news reportage for the periods 1800-49 and 1900-49; (ii) American English texts from 1750 to 1990 representing a new written register, advertising; and (iii) British English material from drama and fiction from the first half of the seventeenth century.

[3] For detailed information on the ARCHER project, the reader is referred to the ARCHER Consortium website at http://www.llc.manchester.ac.uk/research/projects/archer/, and to Nuria Yáñez-Bouza's 2011 article on the (hi)story of ARCHER.

The current version, known as ARCHER 3.1, was completed in summer 2006. Among other changes, in the 3.1 version some of the errors and inconsistencies detected in the original files were corrected, some duplicate texts were deleted, some new files were added, and some texts were shortened or lengthened in order to attain a more balanced distribution of texts per genre and/or sub-period. Moreover, the additional files in ARCHER-2 and the American legal texts from ARCHER-1 were excluded. The resulting version of the corpus contains about 1.8 million words from 955 files, distributed across eight different genres.

In addition to these three complete versions of ARCHER, a new edition of the corpus is about to be launched. This enhanced and updated version, known as ARCHER 3.2, involves the expansion of the corpus to the first half of the seventeenth century, the incorporation of legal texts (both British and American), the addition of texts representing the language of advertising, which were present for the first time in ARCHER-2, and the division of the category journals-diaries into two separate registers. It also aims at the revision, correction, and morpho-syntactic tagging of the texts included in the corpus. This phase of the project has been carried out by a consortium of fourteen universities in seven different countries (cf. Table 1 below), coordinated by David Denison and Nuria Yáñez-Bouza at the University of Manchester. Our own research group, the VLCG, has been a member of the ARCHER Consortium since May 2009. Regarding access to the corpus, at present ARCHER can only be consulted on-site at the member universities, on acceptance of the terms and conditions of the ARCHER user agreement form.[4]

[4] Interested scholars who wish to consult the corpus at Santiago de Compostela should contact María José López-Couso (mjlopez.couso@usc.es).

Table 1. The ARCHER Consortium

- Department of English, Northern Arizona University
- Department of Linguistics, University of Southern California
- Englisches Seminar, Albert-Ludwigs-Universität Freiburg
- Anglistisches Seminar, Ruprecht-Karls-Universität Heidelberg
- Department of English, University of Helsinki
- Department of English, Uppsala University
- Department of English, University of Michigan
- Department of Linguistics and English Language, University of Manchester
- Department of Linguistics and English Language, Lancaster University
- Lehrstuhl für Englische Sprachwissenschaft einschließlich Sprachgeschichte, Otto-Friedrich-Universität Bamberg
- Englisches Seminar, Universität Zürich
- Fachbereich Anglistik, Universität Trier
- School of English, Sociology, Politics & Contemporary History, University of Salford
- Research Unit on Variation, Linguistic Change and Grammaticalization, Departamento de Filología Inglesa y Alemana, Universidad de Santiago de Compostela

3. A Note on Legal Opinions

As noted above, legal texts, more specifically legal opinions, was one of the registers included in the initial architecture of the ARCHER corpus. Legal or judicial opinions, also known as *consilia*, are accounts written by a judge or a group of judges which accompany an order or ruling in a case, explain the facts of the case, and clarify the rationale and the legal principles of the ruling. Some legal opinions introduce a new rule or modify an existing rule, and can therefore serve as a precedent or can overturn a precedent.

Legal texts have two primary communicative functions, the prescriptive or regulatory, and the informative or descriptive (Williams 2005: 28). While texts containing rules or norms, such as laws, regulations, codes, and treaties, are prescriptive, in legal opinions and law reports the informative function clearly predominates. They are therefore regarded by some authors as descriptive legal texts (Šarčević 2000: 11) or as hybrid legal texts (Williams 2005: 29), since they combine prescriptive and

descriptive features, though the latter prevail.

The oldest reports in English law are found in the so-called *Year Books*, written in Anglo-Norman and containing reports of cases from the late thirteenth century to 1535. From the sixteenth century onwards, a number of judges and lawyers prepared publications of reports, known as the *Nominate Reports*, which are named after the court reporter who compiled and edited them. In 1865 the Incorporated Council of Legal Reporting (ICLR) for England and Wales was established in order to avoid the multiplicity of such reports decided in the higher English courts and also to avoid the delay between a judgement and the publication of its report. The ICLR has compiled most of the best copies of cases predating its foundation in the so-called *English Reports*. Reports published after 1865 and produced by this institution are known as the *Law Reports*. Although the Council does not have the monopoly on law reporting in Britain, the *Law Reports* are considered to be the most authoritative and those which should be cited whenever possible, since they are issued by a semi-official organism (cf. Barker 2007: 26-27; ICLR special issue and ICLR webpage: http://iclr.co.uk).

4. British Legal Opinions in ARCHER

The original version of the ARCHER corpus (ARCHER-1) contained 57 files comprising around 147,000 words of American English legal texts corresponding to the periods 1800-49 and 1900-49 (cf. Yáñez-Bouza 2011). As mentioned above, these files were excluded from the 3.1 release of the corpus. In version 3.2 of the corpus, this American legal material has been restored, and a new batch of British English legal texts has been added. When our research unit, the VLCG, joined the ARCHER Consortium, our assigned task within the ARCHER 3.2 project phase was the compilation of British English legal texts from 1600 to the late twentieth century, effectively filling the gaps in coverage of earlier versions of ARCHER. In what follows, we give a brief description of the major issues concerning the compilation of the British legal texts sampled in the corpus, documenting the most important decisions which were taken in the compilation process, and focusing on areas of data collection and coding conventions.

4.1. Collecting the Data

The texts we have contributed to the ARCHER project are law reports drawn from the ICLR for England and Wales, through Justis Publishing Limited (http://www.justis.com), an online resource for legal texts. In July 2009 our research group purchased an annual subscription to Justis, which was activated on acceptance of the licence agreement.

In accordance with the general guidelines for the compilation of new texts in ARCHER 3.2, the material comprises ten files per half century of approximately 2,000 words each. Thus, the sub-corpus of British English legal opinions comprises a total of roughly 160,000 words of running text. Although "absolute representativeness is an unattainable ideal" (Mukherjee 2004: 114) in corpus linguistics, we tried to ensure balancing representativeness by including at least two texts per decade, in this way providing an even coverage for each century. Some reports have been taken in *toto*, other files contain excerpts from longer reports, and, whenever the report was too short, the file has been completed with material from one or more different reports until the average length of 2,000 words was reached.

Justis provides the *English Reports* (up to 1865) as PDFs, which replicate exactly the pagination and the appearance of the original hard copies. The *Law Reports* (1865-present), in turn, are offered in three different electronic formats: as PDF files, as .doc files, and as HTML. Some of the older texts (*English Reports*) have been copied manually, while others have been converted to Word documents by means of the software ABBYY Fine Reader. For the later texts (*Law Reports*) we have used the Word versions, which have then been carefully collated with the PDF files and corrected where necessary. Once the Word documents were typed and edited (see Section 4.2.2 below on the editing process), they were stored as WordPad files, and underwent several rounds of proofreading. For this task we had the assistance of a host of postgraduate students, working under the careful supervision of the FPI researcher Paula Rodríguez-Puente.[5] The final versions of files were then sent to Manchester, where the texts were revised again to check whether they fully conformed to the common guidelines of the ARCHER project.

[5] In addition to Paula Rodríguez-Puente, we would like to thank the group of postgraduate students who have collaborated with us during different stages of the project: Zeltia Blanco-Suárez, Eduardo Coto-Villalibre, Tania de Dios-Miguéns, Iria Pastor-Gómez, Alba Pérez-González, Paula Rodríguez-Abruñeiras, Iria-Gael Romay-Fernández and Vera Vázquez-López.

4.2. Issues of Corpus Compilation

4.2.1. Headers

Following the ARCHER conventions (cf. Yáñez-Bouza 2011), in each of the files the text itself is preceded by a header which, as with all other material inserted by the compilers, appears between caret brackets. In the case of British English legal texts the information in the header includes the following (see Appendix for illustration):

(a) The name of the file. The current convention for filenames was introduced in ARCHER 3.1 with the formula nnnnacbd.gpv, where "nnnn" stands for year; "abcd" for the author abbreviation; "g" for genre; "p" for period; and "v" for variety (cf. Yáñez-Bouza 2011). Instead of the author abbreviation, the filenames of our legal texts normally show the first four characters in the name of the first party in the case. The only exceptions are those cases in which the first party is *Regina* "the Queen", where the abbreviation of the second party has been used instead, thus avoiding repetitions in the filenames. The abbreviation for legal texts is "l", which is then followed by the period and variety, British English in our case.[6]
(b) Word count in Perl script.
(c) Court which judged the case.
(d) Parties.
(e) Date of trial.
(f) Judges.
(g) Bibliographic information.
(h) Year of publication of the report. Although this may coincide with the date of trial, this is not invariably so.

All the information in the headers was also stored in a database which accompanied the files sent to the Manchester team, responsible for the coordination of the ARCHER project.

[6] The periods in ARCHER 3.2 are as follows: 1 = 1600-49; 2 = 1650-99; 3 = 1700-49; 4 = 1750-99; 5 = 1800-49; 6 = 1850-99; 7 = 1900-49; 8 = 1950-99. The two varieties are marked b = British English, and a = American English.

4.2.2. Editing the Texts

A set of guidelines for compilation and annotation of texts has been drawn up and shared by all the ARCHER Consortium members in order to guarantee consistency and coherence throughout the corpus (cf. Yáñez-Bouza 2011). Naturally, these guidelines and coding conventions have been followed in the edition of our British English legal opinions. One general principle in preparing the material was to guarantee the readability of the text, while trying at the same time to be as faithful to the original as possible, without altering characteristic features of the register at issue. Thus, editorial interference was kept to a minimum, indentation and spacing between paragraphs was maintained as in the original, and special (non-ASCII) characters and symbols, such as †, à, â, $ for the dollar, and fractions (e.g. ¼), were also retained. The most important editorial decisions taken in the process of editing the texts include the following:

(a) Punctuation in some of the older texts was not always clearly set in the PDF files provided by Justis, with the result that documents showed some blanks where a punctuation mark was to be expected.[7] In such cases, we added the corresponding punctuation mark (period, colon, semi-colon) as appropriate.
(b) In law reports it is common to find very lengthy footnotes, mostly providing information on related cases. The decision was made to leave all footnotes out. However, notes found within the main text, and which introduce cross-references to other cases, with or without further comments by the reporter, were maintained. For illustration, see Samples 1, 2, and 3 below, where notes are indicated in bold type.

This will not prevent the respondents, if they find it necessary, from bringing a fresh action for recovery of the legitimate elements in the contributions sought to be recovered: **Gillespie v. Russell; Waterson v. Murray & Co.**

Sample 1. Excerpt from file 1937ferr.l7b

[7] Similarly, the dots of small case <i> were also missing in places.

A statutory hypothesis is to be treated with reserve and not pressed beyond its intended purposes: **see Hill v. East and West India Dock Co. (1884) 9 App.Cas. 448, 454; cf. Murphy v. Ingram [1974] Ch. 363, 370.**

Sample 2. Excerpt from file 1989inla.l8b

It would be unlawful for the United Kingdom to ratify the Protocol because the Protocol provides for an increase in the powers of the European Parliament. **[Reference was made to Defrenne v. Sabena (Case 43/75) [1976] I.C.R. 547.]**

Sample 3. Excerpt from file 1994rees.l8b

(c) The reports make extensive use of italics. Since, in accordance with the ARCHER conventions, the WordPad files had to be saved as plain text documents, the issue was raised as to whether italics should be marked or ignored. Given that retaining all italics present in the original reports would greatly affect the readability of the resulting text, since an editorial comment has to be added in caret brackets both before and after the italicized string, we decided to be very selective as to which italics to retain. Italics were therefore silently removed except in the following cases:

(i) Foreign words and expressions italicized in the original text:

ORIGINAL: The respondents having put in no appearance, the case was heard *ex parte*.
CORPUS: The respondents having put in no appearance, the case was heard <italics in the original>ex parte</italics in the original>.

Sample 4. Excerpt from file 1868muss.l6b

It must be noted, however, that not all foreign expressions, especially Latin legal terms, are italicized in law reports, as is the case with *nisi* and *certiorari* in the following sample:

RULE NISI for a writ of certiorari calling upon the Minister of Health to show cause why a certain order dated November 12, 1926.

Sample 5. Excerpt from file 1927king.l7b

(ii) When used for emphasis. In those cases where italics are used in order to emphasize a particular word or words in the report, they have been kept:

ORIGINAL: In considering whether a licence of the old type authorized the use of an omnibus on what is now *the* route to be considered under the Ordinance of 1942, a licence is good if it made it lawful for the operator to use his omnibus along the whole of what is now called *the* route, even if it had also authorized more than that.
CORPUS: In considering whether a licence of the old type authorized the use of an omnibus on what is now <italics in the original>the</italics in the original> route to be considered under the Ordinance of 1942, a licence is good if it made it lawful for the operator to use his omnibus along the whole of what is now called <italics in the original>the</italics in the original> route, even if it had also authorized more than that.

Sample 6. Excerpt from file 1946kela.l7b

ORIGINAL: Again, if as was stated in Wakefield's Case (1827) 2 Lew. 279, 287, it is a privilege that the wife can give evidence against her husband in cases where she is personally involved why must she be *compelled* to give evidence?
CORPUS: Again, if as was stated in Wakefield's Case (1827) 2 Lew. 279, 287, it is a privilege that the wife can give evidence against her husband in cases where she is personally involved why must she be <italics in the original>compelled</italics in the original> to give evidence?

Sample 7. Excerpt from file 1979hosk.l8b

(iii) In eighteenth-century texts when used to mark direct speech, thus contrasting the quotation with the main body text (cf. Barber 1976: 19; Bringhurst 2002: 86), as shown in Sample 8.

ORIGINAL: Robert the son 26 Decemb. 1669, died, leaving Robert the grandson his son and heir; Robert the father afterwards, 16 March, 1671, makes a codicil to his will in these words: *And also I the aforenamed Robert Berrier, do give to my grandchild Judith Berrier and her heirs, more than I have formerly given her by this my last will and testament, a little close I purchased of Robert Greenshaw in Fitling, this I will shall be added as a codicil and taken as part of my last will*.
CORPUS: Robert the son 26 Decemb. 1669, died, leaving Robert the grandson his son and heir; Robert the father afterwards, 16 March, 1671, makes a codicil to his will in these words: <italics in the original>And also I the aforenamed Robert Berrier, do give to my grandchild Judith Berrier and her heirs, more than I have formerly given her by this my last will and testament, a little close I purchased of Robert Greenshaw in Fitling, this I will shall be added as a codicil and taken as part of my last will</italics in the original>.

Sample 8. Excerpt from file 1702berr.l3b

Retention of italics in the above three cases is indicated by the following editorial note after the header: <Original italics retained if for emphasis or marking foreign expressions or indicating direct speech; any others ignored.>

Some of the contexts in which italics occurring in the original have been removed include the following:

I. Proper names (e.g. *Sir Arthur Irvine Q.C.*, *Igor Judge* and *D. Hale* for the appellant).
II. Names of cases (e.g. … the next case (*Stuart* v. *Nixon and Bruce*) raises a point …).
III. Abbreviations for currencies (e.g. 4*s*. 6*d*).[8]
IV. Titles of books, journals, etc. (e.g. "Evasion" is defined in the *Shorter Oxford English Dictionary* …).
V. Months (e.g. ... upon the 10th of *September* 1685).

(d) The original reports also used italics in the old abbreviation for the pound sterling, *l.*, which was placed immediately after the quantity (cf. OED s.v. *pound*, n[1] 2a). Given that setting the abbreviation in

[8] On the use of the old abbreviation for the pound sterling, see below.

roman typeface, as done with other currencies (see above), would result in confusion with the number '1', we decided to replace the old abbreviation for the new symbol of the pound £.

ORIGINAL: at the reduced rateable value of 399*l*.
CORPUS: at the reduced rateble value of 399£

Sample 9. Excerpt from file 1913metr.l7b

This replacement is acknowledged in an editorial note where appropriate: <Original italics retained if for emphasis or marking foreign expressions or indicating direct speech; any others ignored; l. replaced by £>.

5. Outlook

As mentioned in Section 4.2 above, the sub-corpus of British legal opinions included in ARCHER 3.2 comprises around 160,000 words. The size of this sub-corpus is probably too limited for an in-depth analysis of the characteristics of legal language from the early seventeenth century to the present day, especially for the description of low-frequency phenomena. In view of the obvious limitations of the legal material in ARCHER 3.2, the research group *Variation, Linguistic Change and Grammaticalization* intends to complement the ARCHER data by compiling a specialized corpus of British English legal opinions. The corpus, known as the *Corpus of Historical English Law Reports* (*CHELAR*) (cf. Rodríguez-Puente 2011), is currently under preparation and will allow research into the most salient features of this particular type of language and its development over the last four centuries of the history of English.

References

Barber, Charles. 1976. *Early Modern English*. London: André Deutsch.

Barker, David L.A. 2007. *Law made simple*. 12th edn. Oxford: Elsevier.

Biber, Douglas & Edward Finegan. 1997. Diachronic relations among speech-based and written registers in English. In Terttu Nevalainen & Lena Kahlas-Tarkka (eds.), *To explain the present: Studies in the changing English language in honour of Matti Rissanen*, 253-275. Helsinki: Société Néophilologique.

Biber, Douglas, Edward Finegan, Dwight Atkinson, Ann Beck, Dennis Burges & Jena Burges. 1994. The design and analysis of the ARCHER corpus: A progress report [A Representative Corpus of Historical English Registers]. In Merja Kytö, Matti Rissanen & Susan Wright (eds.), *Corpora across the centuries: Proceedings of the First International Colloquium on English Diachronic Corpora, St Catharine's College Cambridge, 25-27 March 1993*, 3-6. Amsterdam & Atlanta: Rodopi.

Bringhurst, Robert. 2002. *The elements of typographic style*. Vancouver: Hartley & Marks.

CHELAR = *Corpus of Historical English Law Reports*. 2011-. Compiled by María José López-Couso, Belén Méndez-Naya, Teresa Fanego, Paloma Núñez-Pertejo, Paula Rodríguez-Puente, Zeltia Blanco-Suárez, Eduardo Coto-Villalibre, Tania de Dios-Miguéns, Beatriz Mato-Míguez, Paula Rodríguez-Abruñeiras, Iria-Gael Romay-Fernández & Vera Vázquez-López. Research Unit 'Variation, Linguistic Change and Grammaticalization', Department of English and German, University of Santiago de Compostela.

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

Mukherjee, Joybrato. 2004. The state of the art in corpus linguistics: Three book-length perspectives. *English Language and Linguistics* 8/1: 103-119.

OED = *The Oxford English Dictionary on CD-ROM*, ed. John A. Simpson & Edmund S.C. Weiner. 2nd edn. Oxford: Oxford University Press.

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 (Special Issue on *Diachronic English for Specific Purposes*).

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

Williams, Christopher. 2005. *Tradition and change in legal English: Verbal constructions in prescriptive texts*. Linguistic Insights 20. Bern: Peter Lang.

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

Appendix: Text Sample

<1897fred.l6b. 2,081 words. Court: House of Lords. Parties: Frederick Bloomenthal v. James Ford (the Liquidator of Veuve Monnier et ses Fils, Limited). Date: February 23 1897. Judges: Lord Halsbury L.C., Lord Herschell, Lord MacNaghten, Lord Morris and Lord Shand. Source: The Law Reports, Appeal Cases ([1897] A.C. 156), compiled by Justis (www.justis.com). Original source: The Incorporated Council of Law Reporting for England and Wales. Date of publication: 1897>

<Footnotes ignored; original italics retained if for emphasis or marking foreign expressions; any others ignored; l. replaced by £>

THE company, Veuve Monnier et ses Fils, Limited, was registered in 1890 under the Companies Acts, as a company limited by shares. In February 1894 the appellant, a stationer in the city of London, who had supplied the company with stationery and printed for them, was asked to lend the company 1000£, and verbally agreed with the secretary and the managing director to lend that sum upon the terms that he should have the company's acceptance for 1000£, and as collateral security 10,000 of the company's fully paid shares of 1£ each, and that if the company should wish to pay off any part he should return a proportionate part of the shares. The appellant advanced the company 1000£, and afterwards received a letter from the company enclosing the company's acceptance for 1000£, and ten certificates. Each certificate stated that the appellant was the registered proprietor of 1000 7 per cent. cumulative preference shares of 1£ each in the company, numbered ---- <blank space in the original> to ----, <blank space in the original> and "that on each of such shares the full amount has been paid". In April 1894 the appellant made a further advance of 600£ and received 6000 shares under precisely similar circumstances. In November the company sold by auction 200 of the above shares at 17s. 6d. each. The appellant returned to the company one of the certificates, executed a transfer of the 200 shares to the purchaser, authorized the auctioneers to pay the purchase-money to the company, and received from the company 20£ in reduction of the loan. In 1895 the company was ordered to be wound up, and the appellant was placed by the liquidator on the list of contributories in respect of 16,000 shares. Upon the hearing before Vaughan Williams J. of an application by the appellant to strike his name out of the list, he was cross-examined upon his affidavit which related the facts above stated. He said that he did not know that the shares handed to him were part of the unissued shares for which the public had not applied, and that he always believed they were fully paid up until December 1894, when he was told by a friend that they were not; that he then gave notice of an application to remove his name from the register of shareholders, but the winding-up supervened.

The secretary was also called and said that the shares issued to the appellant were part of the company's unissued preference shares, that he did not know whether he told the appellant this, but that he assumed that the appellant

understood it.

Vaughan Williams J. refused the application, and this decision was affirmed by the Court of Appeal (Lindley, Lopes and Rigby L.JJ.).

From these decisions the present appeal was brought.

[…]