A multidimensional analysis of late Modern English scientific texts from the *Coruña Corpus*

Autora: Leida Maria Monaco

Tese de doutoramento UDC / 2017

Directora: Isabel Moskowich-Spiegel Fandiño

Titora: Isabel Moskowich-Spiegel Fandiño

Programa Oficial de Doutoramento en Estudos Ingleses Avanzados: Lingüística, Literatura e Cultura

UNIVERSIDADE DA CORUÑA

Isabel Moskowich-Spiegel, profesora do Departamento de Filoloxía Inglesa da UDC, en calidade de directora da tese de doutotamento

A multidimensional analysis of late Modern English scientific texts from the Coruña Corpus

escrita pola doutoranda Dna. Leida Maria Monaco*, estudante do Programa Oficial Interuniversitario de Doutoramento en Estudos Ingleses Avanzados: Lingüística, Literatura e Cultura.*

FAGO CONSTAR

Que a devandita tese de doutoramento reúne os requisitos formais e técnicos necesarios para a súa lectura e defensa pública, e que cumple tamén cos requisitos para optar á mención internacional

A Coruña, 7 de xaneiro de 2017

MOSKOWICH-SPIEGEL
FANDIÑO ISABEL SOFIA -
DNI 32763137M

Firmado digitalmente por MOSKOWICH-SPIEGEL FANDIÑO ISABEL SOFIA - DNI 32763137M
Nombre de reconocimiento (DN): c=ES, o=UNIVERSIDAD DE A CORUÑA, ou=certificado electrónico de empleado público, ou=FILOLOXIA INGLESA, ou=8867, serialNumber=32763137M, sn=MOSKOWICH-SPIEGEL FANDIÑO, givenName=ISABEL SOFIA, cn=MOSKOWICH-SPIEGEL FANDIÑO ISABEL SOFIA - DNI 32763137M
Fecha: 2017.01.04 12:18:35 +01'00'

ASDO. Isabel Moskowich-Spiegel Fandiño

Acknowledgements

First, I would like to thank my supervisor, Dr. Isabel Moskowich, for her continuous guidance and enormous support. While she constantly encouraged me to find and follow my own research path, her invaluable insight would always lead me to the light at the end of the tunnel. Thank you for being the most patient, kind and caring mentor one could possibly wish for, and for teaching me not to be afraid to take risks, even if it means going back and trying again and again.

I am also very grateful to Prof. David Banks for his help and counselling while hosting me at the Université de Bretagne Occidentale at the initial stage of my research. Likewise, I am deeply thankful to Andrew Hardie, who hosted me at Lancaster University twice, for his priceless contribution to this dissertation by tagging our corpus, as well as for his angelical patience while teaching me programming from scratch. And it was in Lancaster where I met Ghada Mohamed, to whom I also owe a lot. Thank you, Ghada, for giving me a very friendly hand with the methodology and for all our inspirational chats around the campus. In line with this, I would also like to thank Nick Smith (University of Leicester) and Costas Gabrielatos (Edge Hill University) for kindly solving my doubts over certain grammatical structures that I was finding rather challenging; Prof. Pascual Cantos Gómez (Universidad de Murcia) and Jack Grieve (Aston University) for patiently walking me through the steps of factor analysis; and Moisés López Caeiro for generously guiding me through the literature on the history of science.

At my *alma mater*, I am indebted to Celso Álvarez Cáccamo for first introducing me into the world of discourse studies and awakening in me a love for pragmatics, that dark and twisty branch of linguistics. My gratitude extends to all the professors in the department of English Philology of the University of A Coruña, as well as to those of the departments of Spanish Philology and Linguistics, under whose guidance I have been fortunate to grow as a researcher and a teacher. On the other hand, this project simply would not have been possible without the help of the Spanish Ministry of Science and Innovation (FPU Grant AP2009-3206), who has supported me financially for a significant part of my research period.

I also owe many thanks to the members of the research group MuStE by whose side I have had the chance to both work and learn. I am especially grateful to Inés Lareo, Begoña Crespo and Emma Lezcano for their caring advice and encouragements, and to Fanny Sánchez, Paula Lojo and Iria Bello for all those memorable moments that we have enjoyed ‘down there’ (i.e. in the research room) and whenever we organised conferences together. And, above all, to my closest PhD-mate Luis Puente-Castelo, who not only has repeatedly read, proofread and given me feedback on my academic ramblings, but also cheered me up every time this postgraduate life felt like too much of a challenge.

Last but not least, I would like to express my gratitude to my family (Mum, Grandma, Enrico, you are my stars) and all those who have been close to me, in one way or another, and have supported me through this time. Thanks a million.

Resumen

Los siglos XVIII y XIX constituyen un período crucial en el desarrollo del inglés como lengua de la ciencia y la conseguinte formación de un "inglés científico". A lo largo de dicho período, cada disciplina científica y cada género utilizado para su transmisión en la escritura, tanto con fines profesionales como didácticos, desenvolvieron su propio registro, llegando a formar el hoy en día llamado "registro científico" que, a pesar de contener características comunes para todas las ciencias, presenta una importante variación interna. El objetivo principal de esta tesis doctoral es estudiar tanto la variación como el cambio lingüísticos en textos pertenecientes a tres disciplinas científicas (astronomía, filosofía y ciencias de la vida), publicados por autores anglohablantes entre 1700 y 1900, y que forman parte del Coruña Corpus of English Scientific Writing, el corpus electrónico recopilado por el grupo de investigación MuStE en la Universidad de A Coruña. La metodología utilizada en este estudio consiste, por un lado, en la utilización de programas de concordancia, tales como Coruña Corpus Tool y CQPWeb, para la recuperación de diversas categorías léxicas y gramaticales en el corpus y, tras una exhaustiva desambiguación manual, del recuento de sus frecuencias de aparición en cada uno de los textos; y, por otro lado, en el empleo de una técnica estadística multivariada, el análisis factorial, para poder observar dimensiones de variación de las distintas disciplinas científicas entre ellas y a lo largo del tiempo. Dicha metodología sigue el método utilizado en el análisis multidimensional (Multidimensional Analysis) de Biber (1988), que se utilizó en distintos estudios de variación intertextual a lo largo de los últimos veintisiete anos.

Resumo

Os séculos XVIII e XIX constitúen un periodo crucial no desenvolvemento do inglés como lingua da ciencia e a conseguinte formación dun “inglés científico”. Ó longo de dito periodo, cada disciplina científica e cada xénero utilizado para a súa transmisión na escritura, tanto con fins profesionais coma didácticos, desenvolveron o seu propio rexistro, chegando a formar o hoxe en día chamado “rexistro científico” que, a pesar de conter características comúns para todas as ciencias, presenta unha importante variación interna. O obxectivo principal desta tese de doutoramento é estudar tanto a variación como o cambio lingüísticos en textos pertencentes a tres disciplinas científicas (astronomía, filosofía e ciencias da vida), publicados por autores anglofalantes entre 1700 e 1900, e que forman parte do Coruña Corpus of English Scientific Writing, o corpus electrónico compilado polo grupo de investigación MuStE na Universidade da Coruña. A metodoloxía utilizada neste estudo consiste, por unha banda, na utilización de programas de concordancia, tales como Coruña Corpus Tool e CQPWeb, para a recuperación de diversas categorías léxicas e gramaticáis no corpus e, tras unha exhaustiva desambiguación manual, do reconto das súas frecuencias de aparición en cada un dos textos; e, por outra banda, no emprego dunha técnica estadística multivariada, a análise factorial, para poder observar dimensións de variación das distintas disciplinas científicas entre elas e ó longo do tempo. Dita metodoloxía segue o método utilizado na análise multidimensional (Multidimensional Analysis) de Biber (1988), que se empregou en distintos estudos de variación intertextual ó longo dos últimos vintesete anos.

Abstract

The eighteenth and nineteenth centuries constitute a key period in the development of English as the language of science and the consequent formation of a "scientific English". Throughout this period each scientific discipline and each genre adopted for their transmission through writing, both with professional and didactic purposes, developed their own particular registers, evolving to the nowadays so-called "scientific register" which, despite presenting certain characteristics common to all sciences, also shows important internal variation. The main aim of this doctoral dissertation is the study of both linguistic variation and change in texts belonging to three scientific disciplines – astronomy, philosophy and life sciences – published by English-speaking authors between 1700 and 1900, and belonging to the Coruña Corpus of English Scientific Writing, an electronic corpus compiled by the MuStE Research Group at the University of A Coruña. The methodology used in this study consists, at a first stage, in the automated retrieval of various lexical and grammatical features from the corpus with the help of concordance programs such as the Coruña Corpus Tool or CQPWeb, and, after an exhaustive manual disambiguation, the recount of their frequencies of appearance in each text. After that, a multivariate statistical technique known as factor analysis is used in order to establish dimensions of variation among the three scientific disciplines and the different genres used within them, and across time. This methodology was first used by Biber (1988) and called Multidimensional Analysis, and has been used in a large number of textual variation studies along the past twenty-seven years.

Table of Contents

List of Tables

List of Figures

Introduction

This study analyses English register variation in three scientific disciplines across the eighteenth and nineteenth centuries, often referred to as the Late Modern Period. More specifically, it looks at how scientific language evolved in the fields of astronomy, philosophy and life sciences along the two hundred years comprised between 1700 and 1900. The choice of dates has not been arbitrary in that they loosely demarcate a key period starting with the so-called Scientific Revolution and culminates with the publication of Einstein's theory of special relativity in 1905. Although the foundations of Western science were laid down more than two millennia ago by natural philosophers, mathematicians and engineers such as Pythagoras, Archimedes or Aristarchus of Samos, it would not be until the seventeenth century that a keen desire of embracing the mechanisms of Nature, would become stronger than faith, politics, or tradition. This desire, which had empowered so many fearless scientists along history to sacrifice their own life for knowledge, and others to accept the ordeal of continuous persecution, ripened into a generalised phenomenon among men of science by 1660, pushing observation and experience to step over the borders of prejudice and religious loyalty. Thus, the two following centuries saw the rise of Empiricism, the creation of scientific societies and academia, the growth of universities and the professionalisation of science, as well as its progressive availability to less privileged classes and women. During that time also, eighteenth- and nineteenth-century philosophers debated over the existence of causality, natural order, or the immortality of human soul, discussing, in addition, certain unorthodox and largely controversial subjects at that moment, such as women's rights or the suitability of marriage.

This paralleled the development of the vehicle for the transmission of science, the scientific language, also called scientific register[1]. During the Scientific Revolution, Latin was being gradually yet steadily replaced by the vernacular, not without a conscious effort on the part of lexicographers, grammarians and scientists themselves, such as Boyle, Sprat, or Wilkins, who established their own rules for writing science. It will appear therefore reasonable to frame a diachronic study of the English scientific register within the limits of the Late Modern Period, starting at its birth, and stretching to the border between the nineteenth and twentieth century, when Present-Day English is considered to begin. This said, it should be noted that by no means do we imply by this that the English scientific register has suffered no changes for the last hundred years. It is nevertheless a fact that the foundations of the theory of special relativity and quantum mechanics in the early twentieth century marked the beginning of a new type of science, breaking, up to a point, with the physical models established to date. On the other hand, the First World War would undermine the pillars of many Western socio-cultural standards, whether moral, political, or religious, bringing forth a shift in philosophical thought and, quite consequently, in philosophical language. In fact, language itself would become a subject of constant debate in the fields of semiotics and theoretical linguistics, first inspired by Saussure's analysis of linguistic signs and their meanings. We consider therefore that the 1900s may be likewise regarded as a boundary for the study of scientific register, demarcating the birth of a new period in its history.

In a study of register variation, the sample analysed has to contain a wide enough range of subjects that would allow for variation to emerge on the maximum number of levels. In this particular case, samples belong to the *Coruña Corpus of English Scientific Writing* (henceforth, *Coruña Corpus*), a corpus of scientific texts which were written by English-speaking authors and published throughout the

[1] For a discussion on the definitions of *register*, *subregister* and *scientific register*, as well as other

eighteenth and nineteenth centuries. This corpus, which is currently under compilation, contains several subcorpora, each containing texts belonging to a scientific discipline, written in a variety of genres (such as essays, treatises, textbooks, etc.). As outlined earlier, this research focuses on the disciplines of Astronomy, Philosophy and Life Sciences. Thus, three of the subcorpora of the *Coruña Corpus* have been selected: *CETA* (the *Corpus of English Texts on Astronomy*), *CEPhiT* (the *Corpus of English Philosophy Texts*) and *CELiST* (the *Corpus of English Life Sciences Texts*), respectively.

The reason for our choice lies in the UNESCO (1988) classification of sciences, according to which Astronomy and Life Sciences belong to the so-called natural sciences, while Philosophy is classified among those called the humanities. The aim here, therefore, is to provide sufficiently different disciplines to look at, both in what concerns their subject matter and, presumably, their language. The sciences contained in both fields have been constantly evolving across time – along the recent years, decades, centuries, even millennia. For a considerable time there was no clear-cut distinction between the two tendencies. Rather, the different sciences were grouped in the Western world under the Greek label *episteme* (knowledge), and, later, Natural Philosophy, from the medieval organisation in Trivium and Quadrivium. However, during the Late Modern period the distinct branches of science culminated the process of their consolidation as separate scientific disciplines such as chemistry, physics, biology, zoology, philosophy, and so on. In what regards language, several studies (Bazerman 1988; Biber and Finegan 1989, 1997, 2001a; Atkinson 1999) have shown that the different subregisters, each common to a particular scientific discipline (e.g. experimental research articles), or literary genre (e.g. sermons, plays, expository prose), have evolved in slightly different ways. Scientific subregisters in particular (Atkinson 1999) have been observed to gradually tend towards a standard scientific register, which, by the start of the twentieth century, already presented many of the features that characterise the language of science today, namely: the use of passive structures, a general lack of personal pronouns, and an abundance of logical connectors and nominsalisations, all of which create a rather impersonal, object-centred kind of discourse. All in all, despite the fact that these characteristic features are nowadays common to most scientific disciplines, each particular scientific subregister must have evolved towards this standard in its own way.

The primary goal of this study, therefore, is to look at the evolution of the three selected subregisters – astronomy, philosophy and life sciences – in order to spot both synchronic variation (i.e. similarities and differences among disciplines at a given point in time) and diachronic change (i.e. the directions they take across time) at a maximum number of linguistic levels. For this purpose we are going to use Biber's (1988) Multidimensional Analysis, a methodology that has been successfully used in a large number of variation studies along the past twenty-eight years (see Chapter 1). Multidimensional Analysis consists in the selection of a sufficiently large number of linguistic features from texts and in the classification their frequencies of occurrence with a multivariate statistic technique of data reduction, known as factor analysis. With factor analysis, linguistic features are grouped, based on their co-occurrence in the texts, into factors, each factor having an underlying communicative function common to those features, and being thus interpreted as a "dimension" of variation. If each dimension corresponds to a communicative function (such as, for instance, informational density, or persuasiveness), then each text – and, likewise, each subregister – can be characterised with respect to that dimension at any point of the time span analysed. As shall be forwarded at the end of the present introductory chapter, a full account of the methodology followed in this study is given in Chapters 4 and 5, whereas its theoretical background and its impact on the study of variation is provided already in Chapter 1.

A secondary goal of this study is to spot variation and change among the different genres contained in each subcorpus. The sample of the *Coruña Corpus* selected for our analysis contains a total of eight genres: treatise, textbook, essay, lecture, article, letter, dialogue, and dictionary. Not all the genres are present in each subcorpus, and some genres appear to be more characteristic of some scientific disciplines than others. For instance, textbooks abound in the Astronomy subcorpus, whereas treatises comprise more than half of the Philosophy subcorpus and almost two thirds of Life Sciences in the beta version used for this work. Essays, in turn, are more commonly found in Philosophy texts. Genres are normally used with specific communicative purposes. Although some of them can be inferred *a priori*, from the very definition of the genre, as shall be discussed in Chapter 3, it is nonetheless our intention to unveil some of their communicative features with the help of factor analysis.

Two research questions stem from these two goals. The first one concerns the degree of mutual influence between two variables, namely: To what extent are the scientific disciplines and genres interdependent? It is expected that some genres will behave differently in different scientific disciplines, while it is likewise expected that scientific discipline characteristics will extend to more than one genre. As shall be seen in Chapter 1, any given sample can be analysed as a (sub)register, either a subregister of the category "astronomy" or as a subregister of the category "treatise", and sometimes it is not that easy to identify which characteristics define which category. The second question, in turn, establishes a relation between the resulting model of register variation and the socio-historical context of the study, namely: How do the resulting patterns of variation and change reflect the changes occurring in the scientific discipline in question and/or in science as a whole at the time? Although scientific breakthroughs are first of all reflected as lexical innovations (see Camiña-Rioboo 2013; Bello 2014), it is expected that some of the changes in the communicative patterns of scientific subregisters should have an extra-linguistic basis. An outline of the evolution of Western science and its language in the Modern Period is given in Chapter 2.

On the other hand, given that the nature of this study is largely methodological in that it consists in the application of Biber's (1988) Multidimensional Analysis, it also contains goals specific to this methodology. First and foremost, it seeks to complement other variation studies where the Multidimensional Analysis has been applied, in particular those analysing scientific discourse across time (Biber and Finegan 1989, 1997, 2001a; Atkinson 1999) and across registers and subregisters (Conrad 1996, 2001; Csomay 2000; Carkin 2001; Gray 2011). This goal is both oriented to adding to the panorama of register variation in English and to trying out one more time Biber's (1988) Multidimensional method, introducing some tentative modifications suggested at its trial-and-error stage. Secondly, despite the fact that all corpora are nowadays machine-readable, which infinitely eases their search for linguistic features, the actual process of retrieval and counting is more than merely mechanical work. Thus, while in some cases manual disambiguation is the only way to select the desired features, in other cases the construction of complex algorithms is required for their retrieval, which is a challenging task. The aim here is to improve certain algorithms that have been previously published in order to maximise precision and recall of a given linguistic feature, or else to obtain an entirely new feature from the corpus that has not been attempted to be retrieved before (see Chapter 4). Still, at the stage of factor analysis, even if all the desired linguistic features have been retrieved, some will inevitably “drop” during the trial-and-error stage. With regard to this, and thirdly, factor analysis itself has, at the first stage of its application, a purely exploratory goal, which consists in determining the right number of independent variables (linguistic features) to be included in the analysis, the right type of rotation method and, eventually, of factors to be extracted (see Chapter 5).

Finally, it is also the aim of this dissertation to contribute to the study of the scientific register and its history in general, both in order to find evidence that would, once more, support the findings of those who have provided invaluable perspectives on this topic (to cite some, Bazerman 1984, 1988; Halliday 1988, 1989, 1990; Montgomery 1996; Valle 1996; Gotti 1996, 2001, 2003, 2008, among many others), and also in the hope of shedding some light on some other aspects of scientific writing that have not yet been investigated in depth by providing (yet) another model of cross-register variation. On the other hand, our purpose is to explore the *Coruña Corpus* from a macro-analytical perspective for the first time, offering a “bird’s view” of its communicative patterns, aspiring thus to add to the research carried out so far on its different subcorpora, including studies in morphology and specialised lexicon (Bello 2010, 2014; Camiña-Rioboo 2010, 2012, 2013; Camiña-Rioboo et al. 2010) grammatical and lexico-semantic bundles (Lareo 2011a, 2011b, 2012; Sánchez Barreiro 2010a, 2010b, forthcoming; Moskowich 2012; Gray and Biber 2012; Alonso and Lareo 2016), as well as discourse structure and pragmatics (Lareo and Montoya 2007; Moskowich 2011, 2013, 2016a, 2016b; Crespo 2011, 2012, 2014; Bello 2016; Puente-Castelo 2014, 2016a, 2016b, forthcoming, among others).

The structure of this dissertation, determined by the nature of its goals and methodology, is as follows:

Chapter 1 starts by reviewing the literature on variation at different linguistic levels and comparing different approaches of tackling corpus research, following, in the second part, with a discussion on terminology, focusing on categories such as *register*, *genre*, *text-type* and *style* in order to agree on a definition for the term *register* as used in our study. The third part of this chapter introduces Biber’s (1988) Multidimensional Analysis of register variation, exploring its antecedents and providing an outline of its methodology, and offers a review of relevant variation studies based on Biber’s approach. The last part of the chapter is devoted to discussing the characteristics of the scientific register, which is the object of our analysis.

Chapter 2 deals with the socio-historical context of science and its language. After setting the chronological boundaries for the late Modern period, it offers an insight into the history of Western science, contextualising what we know as the Scientific Revolution, and analyses the role that certain institutions, such as the Royal Society of London, and some of its founders played in the emergence and development of the English scientific register. This chapter, thus, traces a timeline of the evolution of science and the language used for its communication from the seventeenth to the late nineteenth century.

Chapter 3 begins with a review of the use of corpora along the past decades and provides some definitions of the term *corpus* in linguistics, following with a presentation of the *Coruña Corpus* and its characterisation with respect to its research scope and its compilation principles. Each of the three subcorpora included in this particular study are described in detail, with a focus on their composition in what regards scientific disciplines and genres. The second part of this chapter is devoted to the technical aspects of the *Coruña Corpus*, such as its annotation for part-of-speech categories and the concordance programs used for its handling.

Chapters 4 and 5 both deal with methodology, albeit with different parts. Although the methodology was initially intended to be described as a continuum, it was eventually decided to present the statistical part as a chapter on its own, given its complexity and rather technical nature. Thus, while Chapter 4 details the process of selection, retrieval and counting of linguistic features in the corpus and provides the algorithms used to develop each query, Chapter 5 is entirely devoted to factor analysis. The first part of this chapter offers an introduction to the procedure and focuses on certain key aspects in statistics such as sampling adequacy, the notions of variance, communality and uniqueness, and different methods of extracting and rotating factors. The rest of the chapter provides a step-by-step description of the application of factor analysis to the actual data and discusses the problems arising at the trial-and-error stage.

Finally, Chapter 6 analyses and interprets the patterns of register variation resulting from the factor analysis, characterising the different subregisters and genres with respect to each dimension of variation, according to their respective communicative functions. Findings are supported with examples from the texts, which are contrasted and discussed. Conclusions are presented in the last section, in an attempt to answer the research questions posed earlier in this introductory chapter and to analyse the findings in the light of the socio-historical context of this study, presented in Chapter 2. Likewise, the problems encountered at the different stages of the study are analysed and improvements for its possible replications and/or further research are suggested in this section.

Four Appendices are included at the end. Appendix I lists all the linguistic features included for analysis and their abbreviated names. Appendix II contains a table with the descriptive statistics (mean, maximum and minimum values, range and standard deviation) per dimension and subregister (discipline/century). Appendix III presents the factor scores, or dimension scores, for each text analysed. Finally, Appendix IV provides the output several of the factor analyses that have been run during the trial stage and which are discussed as relevant in Chapter 5.

Chapter 1

Exploring register variation and change

1. Introduction

If we think about language as we experience it, regardless of concepts, definitions, or any theorisation whatsoever, we will probably have in mind a variety of linguistic situations we face every day. For instance, when we say "hello" when we arrive at work; when we bend over our laptop to compose a draft of what should at some point become a research paper; when we write a letter to a friend who lives in a distant place; or when we gather for an after-work drink, or for a family meal. Likewise, we may also think of a novel, a poem, or a book of recipes, all of which contain language in different shapes. From such a perspective, which is the perspective of an average speaker, we could agree with Atkinson (1992: 3) that "there is no such thing… as 'pure language' – all language is language in context, so to speak". Our use of language varies depending on the context in which we use it. And simple and practical as it may sound, language use can vary on many levels: morphological, syntactic, lexical, semantic, pragmatic, etc., and sometimes on more than one level at once. When we chat with a friend on the phone, we may often use contractions (*can't, won't*), short sentences (*Sure! Why not?*), and a colloquial type of vocabulary (*yeah, whatever*) which we would be unlikely to use in a more formal context, such as when we write a legal document or a research report for a specialised journal. The different varieties of language that we use in different situations will be here called *registers* (although sometimes the labels *genre*, or *style*, are preferred by some authors[2], while the choice to use one or another register in different contexts is usually referred to as *register variation* (Halliday et al. 1964; Halliday 1985; Ure 1982; Biber 1995; Biber and Conrad 2009; Neumann 2013). The history of these terms will be briefly outlined below.

Variation in language has been a primary focus of research in different fields of linguistics over the past sixty years. Regional varieties and language contact have been approached from different perspectives, combining dialectology, language typology, sociolinguistics and/or historical linguistics (Chambers & Trudgill 1998; Berns & Van Marle 2002; Kortmann 2004, among many others). In the nineteen-sixties and seventies, a number of North-American and British sociolinguists, linguistic anthropologists and sociologists of language began to consider the different varieties of language in the light of the surrounding society and culture, taking into account a wide range of social factors, such as the speakers' social status, level of education, ethnicity, religion, or gender, and their impact on the speech community (Labov 1963, 1965, 1966, 1972a, 1972b; Gumperz and Hymes 1972; Gumperz 1982a, 1982b; Hymes 1962, 1976; Trudgill 1972, 1978, 2000, 2002; Milroy 1980, 1992, 1997). Weinreich et al. (1968: 100-101) defended the idea that language is characterised by an "orderly" or "structured heterogeneity", seeing variation as part of its very description (Kiesling 2011: 7), and linguistic and social factors as "closely interrelat[ing]" (Weinreich et al. 1968: 188). This was the birth of the study of discourse in its broad definition as "language in use" (Brown and Yule 1983; Stubbs 1983; see also Schiffrin et al. 2001), which meant that language was no longer considered on its own, no longer as an abstract entity separated from its speakers. Halliday (1978, 1985) and Halliday & Hasan (1980, 1989) went further to distinguish between language *users* and language *use*. They separated geographical and social variation, where the resulting dialects and sociolects are a consequence of the

[2] See discussion of terminology in Section 2.

speakers' socio-geographic backgrounds, from functional or diaphasic variation, where the choice to use of one or another language variety, or *register*, is determined by a particular situation or activity. Halliday's theory of register, based on three principal components – field, tenor and mode of discourse – has evolved into what is known today as Systemic Functional Linguistics (SFL; see Martin 1985, 2001; Matthiessen 1993; Banks 2002, 2004, 2005). Although we are not adopting this theoretical framework in the present study, our use of the term *register* largely coincides with its use in SFL (see Biber and Conrad 2009: 20; see also Section 2).

On the other hand, the time dimension can add perspective to the study of variation, and it is historical linguistics that looks at the bigger picture of a constantly changing language across its different periods. It has been suggested that variation and change may be considered as "two facets of the same phenomenon" (Fried 2010: 5) in that synchronic variation can sometimes be regarded as a potential initiator of what may later be perceived as diachronic change (Hoenigswald 1960: 55; Bright 1976: 36). Andersen (2006: 65) offers a metaphorical vision of the two sides: "[i]n the synchronic perspective… the 'language' that changes is a 'practice of speaking'", whereas "[i]n the diachronic perspective, the 'language' is a 'tradition of speaking'". Similarly to variation, language change occurs at different linguistic levels; but the clearest view of its dynamics can be obtained through a sociolinguistic lens, which brings to light socio-historical factors external to language that may have very likely triggered that change. Such factors would include language contact as a result of migrations, invasions, or military conquests; the acceptance of a foreign culture as more 'worthy' or prestigious; sociolinguistic pressure within the speech community; or waves of political, philosophical and cultural innovations, all of which appear reflected on the linguistic landscape when contemplated from a distance (see Labov 1963, 1965; Smith 1996; Romaine 2000; Moskowich 1995, 2012; Beal 2004, 2012; Camiña-Rioboo 2013). Still, when it comes to analysis, variation and change are rarely looked at as interconnected parts. More often, linguists will either focus on patterns of synchronic variation, or look for traces of diachronic change, and register variation can be tackled from both perspectives.

The compilation of large electronic corpora along the past fifty years made it possible to study register variation on a large scale, comparing different registers in different languages, both synchronically and diachronically. Present-day English alone has proved to be a great source of written and spoken texts belonging to its different regional varieties. For instance, the *Lancaster-Oslo-Bergen Corpus of British English* (*LOB Corpus*; see Johansson et al. 1978, Johansson 1982) and the *London-Lund Corpus of Spoken English* (Svartvik & Quirk 1980; Johansson 1982) were used for the first comprehensive analysis of spoken and written English registers (Biber 1988), which was later compared to register variation analyses in three other languages (Biber 1995; see Section 3.3). These and other corpora were inspired by the ca. one-million-word *Standard Corpus of Present-Day Edited American English*, compiled in the 1960s (widely known as the *Brown Corpus*, see Francis & Kučera 1964). In the 1990s, the *Frown* (Hundt et al. 1999a) and the *F-LOB* (Hundt et al. 1999b) corpora were built at the University of Freiburg – wherefrom the initial "F" – as "younger" counterparts of *Brown* and *LOB*, respectively, matching their size and composition, and permitting to trace the most recent evolution of British and American English along a period of forty years (Lovejoy 1995; Hundt 1997; Mair 1995, 1997, 2002). Other large-scale projects include the *British National Corpus* (*BNC*; see Aston & Burnard 1998; Hoffman et al. 2008), which has been the source of several register variation and text typology studies within contemporary British English (Lee 1999; Takahashi 2006; Mohamed 2011); the *Corpus of Contemporary American English* (*COCA*; see Davies 2008-, 2009), which, like the *BNC*, contains a variety of registers (classified as spoken, fiction, popular magazines, newspapers, and academic), ranging across a period of twenty-two years (1990-2012); and the *International Corpus of English* (*ICE*; see Nelson 2002a, b), which provides both written and spoken materials for the study of the different regional varieties of the English language across the world (Lange 2012; Deuber 2014).

On the other hand, diachronic corpora, such as the *Helsinki Corpus of English Texts* (Kytö and Rissanen 1992), *ARCHER* (*A Representative Corpus of Historical English Registers*; see Biber *et al.* 1994), or the *Corpus of Historical American English* (*COHA*, the diachronic counterpart of *COCA*; see Davies 2010-), allow the study of register variation and change across several centuries. Biber & Finegan (1989; 1997; 2001a) analysed the diachronic evolution of different specialised and non-specialised English registers, while Atkinson (1996, 1999, 2001) traced the development of the scientific reports published by the Royal Society of London from the seventeenth to the twentieth century. Similarly, Culpeper & Kytö (2010) looked at historical change in four "spoken" (i.e. dialogical) English registers. On the other hand, corpora specialising in one particular register, such as the *Corpus of Early English Correspondence* (*CEEC*; see Nevalainen & Raumolin-Brunberg 1994), the *Corpus of Early English Medical Writing* (*CEEM*; see Taavitsainen et al. 2010; Taavitsainen & Pahta 2010), or a sample of news reportage articles from *Time* Magazine published along the twentieth century (Davies 2007), permit to analyse the development of some characteristics specific to that register, such as formulas of politeness (Nevalainen & Raumolin-Brunberg 1995) or code-switching structures (Pahta 2003, 2004). Moreover, specialised diachronic corpora can help to detect subtler patterns of variation and change within the limits of a single register, sometimes uncovering smaller sub-registers with their own evolution dynamics (Biber & Gray 2013a). This, precisely, is the aim of the present study: to spot variation across three changing sub-registers of English scientific prose along the eighteenth and nineteenth century, using three subcorpora from the *Coruña Corpus of English Scientific Writing* (Moskowich & Crespo 2007, 2012; Lareo 2009; Moskowich et al. 2016; see Chapter 3).

As has been noted earlier, it should be borne in mind that the term *register* is sometimes replaced by *genre* in the literature, whereas on other occasions both terms appear together but with different meanings. Section 2 presents a brief overview of the terms *register*, *genre*, *style* and *text-type* in the literature, which will allow us to set some terminological boundaries. On the other hand, given the methodological nature of this study, Section 3 introduces Biber's (1988) Multidimensional Analysis as its theoretical-methodological basis, although a detailed analysis of the methodology applied in this piece of research will be given in Chapters 4 and 5. Finally, Section 4 focuses on the object of our analysis – the scientific register – discussing some of its characteristics.

2. Register, genre, text-type, and style

Although we have thus far discussed variation among registers, some authors prefer to talk about *genres*, such as *spoken* and *written genres* (e.g. Biber 1988) or *academic* and *research genres* (Bhatia 1993, 1996, 2002). While an exhaustive debate over the suitability of either term falls out of the scope of this study, it may be useful to revise their different definitions in the literature so that any further confusion can be avoided. Biber and Conrad (2009: 21) explain that *register* and *genre* are often used indistinctively, referring to "varieties associated with particular situations of use and particular communicative purposes", and give a classification of different authors according to their preference for one or another term:

> GENRE: Biber (1988), Bhatia (2002), Samraj (2002), Bunton (2002), Love (2002), and Swales (1990, 2004). Biber (1988) referred to "pervasive linguistic patterns"[3], while Swales (1990) referred to structural organisation;
>
> REGISTER: Ure (1982), Ferguson (1983), Hymes (1984), Heath and Langman (1994), Bruthiaux (1994, 1996), Biber (1995), Conrad (2001), and Biber et al. (1999). (And Bhatia (1993: 6) also cites Gregory (1967), Crystal and Davy (1969), Ellis & Ure (1969), Hasan (1973) and Gregory & Carroll (1978), who also use the term register.)

Besides and including the above-mentioned studies, there have been many attempts to establish a relationship between *register* and *genre*, *register* and *style*, or *genre* and *text-type* (e.g. Biber and Finegan 1986, 1994; Swales 1990, 2004; Bhatia 1993, 1996, 2004; Biber 1995, 2006; Biber et al. 1998; Johnstone 2002;), and their relatively large number seems to suggest that these concepts form a fuzzy set of categories with no clear-cut limits among them. In a(nother) recent theorisation about *genre*, Giltrow (2010: 29) reflects on the vagueness of the word:

> In the language disciplines, *genre* has been a term both easily summoned and easily displaced. Easily summoned, it can name what people recognise as broad similarities in ways of thinking or it can name much narrower formations – predictable wordings or familiar collocations. Easily displaced, it can give way to *discourse* in the broader perspectives or *style* or *register* in the narrower ones.

For Swales (1990: 58), a genre reflects the specific communicative purposes shared by the members of a discourse community[4]; those communicative purposes "shape" the schematic structure, content, and style of the discourse, forming genres. Following this line of thought, Bhatia (1993: 16) defines genre as "an instance of a successful

[3] This term is from Biber and Conrad (2009). Biber (1988) uses a different definition of genre (see below).

[4] Swales (1990) maintains that the members of a discourse community "own" genres and use them to communicate their knowledge.

achievement of a specific communicative purpose using conventionalized knowledge of linguistic and discourse resources.” In a revision of the history of the term, Bhatia (1996) identifies three different perspectives on genre that had been used up to that moment, in chronological order of appearance:

(a) genre as a typified rhetorical action;
(b) genre as a staged, goal-oriented social process, and
(c) genre as a conventionalised, communicative event.

The first perspective, developed in the North-American tradition for essentially didactic purposes (Bitzer 1968; Miller 1984; Berkenkotter and Huckin 1995), regards genres as conventional discourse categories built to respond to recurring rhetorical situations (Bhatia 1996: 41). For Berkenkotter and Huckin (1995: 3), genres are “inherently dynamic rhetorical structures” that change over time “in response to their users’ sociocognitive needs”. The evolution of the scientific journal article along the twentieth century, studied by Bazerman (1988, 1993, 1994), is an example of a dynamic genre that gradually adapts to the needs of the scientific community. The second perspective corresponds to the semiotic orientation of register and genre analysis within the SFL framework, developed in Australia (Martin 1985; Couture 1986; Martin, Christie and Rothery 1987; Kress 1987, 1993). Martin (1985) considers that register and genre are on different “semiotic planes”, the former being the “expression-plane” of genre, while the latter refers to the conventional organisation of texts (Couture 1986, in Martin 1985: 80, cited in Biber & Conrad 2009: 22). Just like the American authors, Kress (1987: 44) considers genres as dynamic, cultural constructs that “change historically” along with the social groups. Finally, the third perspective on genre is the one developed in the United Kingdom within the field of applied linguistics by Dudley-Evans (1986), Swales (1990), and Bhatia (1993), according to which genre is seen as a communicative event that has specific communicative purposes. In an attempt to find a common ground for these three perspectives, Bhatia (1996: 54) describes genre analysis as “narrow in focus but broad in vision”, comparing it to “a diamond with a number of carefully crafted facets; the more facets it has, the more insightful and illuminating the analytical activity and more exciting the results”.

In an attempt to reconcile different perspectives on register and genre, Johnstone (2002: 158) defines them as two complementary and interrelated parts of language (see Table 1.1):

Table 1.1

REGISTER	GENRE
Definition: a variety of language (or "style") associated with a recurrent communicative situation or set of communicative roles.	Definition: a recurrent verbal form (or "text-type") associated with a recurrent purpose or activity; "genre knowledge" is the procedural competence required to produce a form and use it.
Examples: testamentary language, **scientific discourse**, teacher-talk, medical discourse.	Examples: wills, **research reports**, essay questions, medical consultations
To describe a register, you need to describe: 1. the situation which calls for the register; 2. the linguistic features that constitute the register	To describe a genre, you need to describe: 1. the form of texts of the genre; 2. the contexts in which the genre is relevant, in which participants may use the genre to organize, explain what they are doing and why; 3. the activities by which people create and share the knowledge required to produce texts in the genre; 4. how the genre works in interaction: how people draw on the generic conventions in creating new text, how they use genre to categorize situations, how the genre serves to maintain the status quo and/or make change possible.

Difference between register and genre in Johnstone (2002: 158; my emphasis)

As can be seen in Table 1.1, Johnstone (2002: 158) defines *register* as a "variety of language", or style, and *genre* as "a recurrent form" that this variety of language adopts, or text-type. The different registers may come in the shape of different genres, just as the scientific discourse often comes in the shape of a research article. Because genres are dynamic constructs, conditioned by the situational context in which they are created, by the communicative purpose they serve, and by the way participants use them, the research article may be now considered a relatively fixed genre, associated to a particular (i.e. the scientific) register. However, the definition of genre as conventional shape means that genres can also "cut across registers" (Bhatia 1996: 45). For instance, the communicative purposes of a textbook are essentially the same whether the textbook is on social sciences, anatomy, history, or linguistics – i.e., to instruct in a particular discipline. A textbook would be therefore a didactic genre that can be used for diverse academic registers.

For Johnstone (2002: 158), register is the same as *style*, while genre is the same as *text-type*. The duality *genre*/*text-type* is another widely discussed point, trapped in-between genre theory and text typology. As it also happens with genre and register, some authors prefer to use the term *genre* (Paltridge 1996, 1997) in those occasions where others prefer to use *text-type* (Kinneavy 1971; Werlich 1982; Hatim & Mason 1990; Görlach 2004). For Taavitsainen (2001: 140) genre is a "mental frame in people's minds which gets realized in texts for a certain purpose in a certain cultural context", and she considers text-types to be the linguistic realisations of genres (Bello 2014: 151). Biber and Finegan (1986: 20) adopt a less theoretical and more empirical perspective: they use cluster analysis to identify *speech styles*, which are "sets of texts that are similar in linguistic form", and distinguish them from *register* (which groups texts according to "the relations among participants and other characteristics of the communicative situation") and *genre* (which groups texts "according to topic and purpose"). Biber (1988: 68) defines genres as "text categorizations made on the bases of external criteria relating to author-speaker purpose", while text-types are, in contrast, "classes of texts that are grouped on the basis of similarities in linguistic form, irrespective of genre classifications" (Biber 1988: 206; see also Biber 1989). Genre, therefore, would be a category based on external (non-linguistic, rhetorical) features, while text-type would be a category based on internal (linguistic) features (Mohamed 2011: 3). Biber and Finegan (1994: 52-53) explain that

> regardless of purpose, topic, interactiveness, or any other non-linguistic factors, text-types are defined such that the texts within each type are maximally similar with respect to their linguistic characteristics (lexical, morphological, and syntactic), while the types are maximally distinct with respect to their linguistic characteristics… After the text types are identified on formal grounds, they can be interpreted functionally in terms of the purposes, production circumstances, and other situational characteristics shared by the texts in each type.

Thus, certain texts belonging to different genres (such as, for instance, official letters, academic prose and legal documents) may present similarities on a strictly linguistic level, and would be grouped into a single text-type on the basis of these linguistic similarities. Following Biber & Finegan (1986) and Biber (1988, 1989), and using cluster analysis, Mohamed (2011) proposed a textual typology of the spoken and written registers in the *BNC,* identifying six main text-types – persuasion, narration, informational narration, exposition, scientific exposition and literary exposition – based on the clustering of texts which have similar linguistic features.

In a later study, Biber (1995: 9) replaces what he earlier called *genre* by *register* "as a general cover term", suggesting that the two of them can be used interchangeably. Biber et al. (1999), Biber (2001, 2006), Biber & Finegan (2001a, b) and Conrad & Biber (2001) use the term *register* exclusively. Eventually, Biber & Conrad (2009) revisit this problematic terminology, setting somewhat clearer boundaries between register and genre, and bringing the third category *style* to the picture (while totally disregarding the category *text-type*). They define register as "a variety associated with a particular situation of use (including particular communicative purposes)" and which can be described with regard to "three major components: the situational context, the linguistic features, and the functional relationships between the first two components" (Biber & Conrad 2009: 6). A situational variety, or register, therefore, can be described in function of the lexical and grammatical features that characterise it. These linguistic features are "pervasive" of that situational variety, or register, which means that they are very common in that particular register, where they appear much more frequently than in any other register. Likewise, a register is also described with respect to its situational context, such as whether it is written or spoken, and what communicative purposes it serves. From a *register perspective*, the relationship between the linguistic features and the situational context is functional – i.e. such that the former always have a function in the latter. In Biber & Conrad's (2009: 6) words, "linguistic features tend to occur in a register because they are particularly well suited to the purposes and situational context of the register".

The *genre perspective*, on the other hand, focuses on the rhetorical organisation of texts, rather than on their lexical and grammatical component. The characteristic features of a genre are not pervasive, but of a more structural character, occurring only once or twice in a text, often at the beginning or/and at the end, and usually have a conventional, rather than a functional relationship with the situational context (Biber & Conrad 2009: 7, 16-17). This means that, while registers can be described by analysing text samples, genres can only be identified by analysing complete texts. For instance, a register study of scientific prose can analyse a corpus of two-thousand-word samples of research articles by describing certain linguistic features that are typical of scientific writing, such as certain types of passive constructions, adverbial subordinators and nominalisations (see Section 4). A genre analysis, conversely, would need to consider the research articles in their totality in order to take into account their formal characteristics, such as the abstract at the very beginning of the article, the conventionally established Introduction-Method-Results-Discussion (hereafter IMRD) sections, and the summary or conclusions of the findings.

Finally, a *style perspective* is another way of describing situational varieties or texts. As we had seen earlier on Table 1.1, Johnstone (2002: 158) considers register and style to be essentially the same. Biber & Conrad (2009: 22) mention some earlier studies from the field of descriptive linguistics (Joos 1961, Crystal and Davy 1969) where style is treated similarly to register, referring to "general situational varieties", while for Labov (1966, 1972a) "sociolinguistic styles" were a reflection of the production circumstances which required speakers to adjust their speech accordingly. For Biber and Conrad (2009), the style perspective – like the register perspective – can be applied to a representative sample of texts from a variety in which certain frequent linguistic features, typically associated with a particular style, can be analysed. However, the difference between the register perspective and the style perspective lies "in their interpretation – that is, in the underlying reasons for the observed linguistic patterns", because those "associated with styles are not functional. Rather, these are features associated with aesthetic preferences, influenced by the attitudes of the speaker/writer about language." (Biber & Conrad 2009: 18). Thus, a particular style may be associated with a particular author, or, sometimes, with a particular group of authors, belonging to a particular historical period, and the linguistic patterns associated with that author or authors are *stylistic*, aesthetic, but they do not define a register. The style perspective can be therefore applied to compare texts within a register or a genre (Biber & Conrad 2009: 18, 72), which may be useful when comparing novels by different authors and/or from different historical periods.

Table 1.2 is an extended version of a table in Biber & Conrad (2009: 16) which summarises the defining characteristics of registers, genres and styles. I have added a fifth column for the category text-type as considered in Biber (1989) and Mohamed (2011), and a sixth row describing a previously tested method for applying each of the four perspectives. As can be seen in this table, texts and/or situational varieties can be analysed from four different perspectives: from a register, genre, or style perspective (Biber & Conrad 2009), and also from a text-type perspective (Biber 1989; Mohamed 2011). Full texts have to be used in order to apply the genre perspective, because text samples "do not necessarily represent the linguistic conventions that define the genre" (Biber & Conrad 2009: 18). Conversely, registers, styles and text-types are described by their pervasive linguistic patterns and therefore can be identified through an analysis of representative text samples.

Table 1.2

Defining characteristic	**Register**	**Genre**	**Style**	**Text-type**
Textual focus	sample of text excerpts	complete texts	sample of text excerpts	sample of text excerpts
Linguistic characteristics	any lexico-grammatical feature	specialised expressions, rhetorical organisation, formatting	any lexico-grammatical feature	any lexico-grammatical feature
Distribution of linguistic characteristics	frequent and pervasive in texts from the variety	usually once-occurring in the text, in a particular place in the text	frequent and pervasive in texts from the variety	frequent and pervasive in texts from the variety
Interpretation	features serve important communicative functions in the register	features are conventionally associated with the genre: the expected format, but often not functional	features are not directly functional; they are preferred because they are aesthetically valued	features serve important communicative functions in the text-type, regardless of register/genre classifications
Method	analysis of co-occurring linguistic features in a register (Biber 1988, 1995)	rhetorical analysis (Atkinson 1999)	close reading of the texts; microanalysis	clustering of texts similar in their linguistic features (Biber & Finegan 1986; Biber 1989; Mohamed 2011)

Summary of the characteristics of registers, genres and styles (Biber and Conrad 2009: 16), and text-types, as well as methods of analysing each category (my addition)

From a register perspective, the identified pervasive linguistic patterns have an underlying communicative function, whereas those analysed from a style perspective have an aesthetic function. From a text-type perspective, on the other hand, varieties are identified *a posteriori*, i.e. after certain pervasive linguistic patterns are spotted in certain texts. These texts are grouped according to their common linguistic patterns, and the relationship between those patterns and the resulting groupings (or text-types) is then considered functional, responding to some communicative purposes shared by the texts in the type.

In order to agree on a terminology for the present study, we shall adopt the perspectives on register and genre from Biber & Conrad (2009) as they seem to provide the clearest definitions for each term and to establish a sharp boundary between the two perspectives. Moreover, their approach to register and genre variation compels us to consider our text samples from a register perspective, discarding the possibility of a genre analysis due to the lack of complete texts (see Chapter 3), which permits us to avoid any confusion resulting from the genre classifications of our samples. Finally, Biber & Conrad's (2009) terminological framework is established on an empirical basis, stemming from Biber's (1988) Multidimensional Analysis of register[5] variation. This theoretical-methodological approach, which shall be the one used in the present study, is explained in what follows.

3. The Multidimensional (MD) Analysis of register variation

As we have seen at the beginning of this chapter, the study of situational varieties such as registers usually requires a relatively large amount of study materials, such as a minimally comprehensive corpus of texts, belonging to one or more registers. There are two different perspectives to approach the analysis of a corpus: a top-down (or corpus-based) perspective, and a bottom-up (or corpus-driven) perspective (Tognini-Bonelli 2001; Biber 2009; Gray 2011). The top-down, or corpus-based approach, consists of searching in a corpus for the occurrences of one or more linguistic features, usually selected on the basis of previous research, which are later analysed.

[5] As was mentioned in Section 2, Biber (1988) speaks about variation in spoken and written *genres*. At that time, however, he used the term *genre* in the same way as, in later studies (Biber 1995, Biber & Conrad 2009) he uses the term *register* – i.e. referring to a situational variety identified for its typical linguistic patterns. As we have agreed to adopt Biber & Conrad's (2009) definitions for *register* and *genre*, I will hereafter stick to the term *register* in the sense specified above, regardless of the terminologies used in previous studies.

The researcher knows what (s)he is looking for, although (s)he may or may not eventually find it in the corpus (that is, one may look for a set of words but only some of them could be present in the corpus). An example of a corpus-based approach would be an analysis of pre-selected formulaic expressions (e.g. Moon 1998) or specific lexical items from a closed list. In contrast, a bottom-up, or corpus-driven approach is "more inductive" in that "the linguistic constructs themselves emerge from analysis of a corpus" (Biber 2009: 276), with minimal to no *a priori* assumptions made. In this case, the researcher may have an idea of what (s)he is going to find, but, mostly, the search in the corpus is a discovery process. In principle, examples of corpus-driven research would include analyses of collocations (Sinclair 1991; Kennedy 2003; Lareo 2006, 2008, 2011b), syntactic patterns (e.g. the pattern *N* + *PREP* + *V-ing*, as in Gray & Biber 2012), lexical bundles (Biber et al. 2004) or keywords (Leech et al. 2001; Lee 2008; Groom 2010). However, in practice the two approaches represent "a continuum, rather than a dichotomy" (Gray 2011: 23) in that most of the corpus studies mentioned above are hybrid, rather than strictly corpus-driven (Biber 2009: 281), combining elements from both approaches. As we will see later in this section, Biber's (1988) Multidimensional Analysis of register variation is also a hybrid approach, being partially corpus-based and partially corpus-driven.

On the other hand, following Conrad (2002) and Biber et al. (1998), Gray (2011: 19-22) classifies corpus analyses in terms of the "comprehensiveness of the linguistic features" that are investigated. According to this parameter, studies would fall into three categories: those investigating a single linguistic feature (Type 1); those investigating a few linguistic features that fulfil a common communicative function (Type 2); and those which investigate a large set of linguistic items or structures in order to provide a comprehensive description of language or a linguistic variety (Type 3). Earlier, Biber (1988) distinguished between microscopic and macroscopic analyses of textual variation, with the former broadly corresponding to Gray's (2011) Type 1. A microscopic study "provides a detailed description of the communicative functions of particular linguistic features", while "[m]acroscopic analysis attempts to define the overall dimensions of variation in a language" (Biber 1988: 61). Such studies can provide functional information on certain linguistic constructions, but this information alone is not sufficient to describe a register. On the other hand, in order to be truly comprehensive, macroscopic analyses need to be complemented with microscopic studies that will provide the necessary information about the communicative functions fulfilled by individual linguistic features. For Biber (1988: 62-63), linguistic variation can only be analysed fully through an approach that combines both the macroscopic and the microscopic perspectives.

Until the late 1980s, most variation studies were microscopic analyses, focusing either on the functions of a single word (Aijmer 1986; Stenström 1986) or a particular grammatical structure (Thompson 1983). Conversely, macroscopic approaches were uncommon, although Biber (1988: 62) cites a few studies which unveiled "dimensions of variation" in fictional and non-fictional prose (Carroll 1960; Marckworth & Baker 1974). At the same time, theoretical approaches to linguistic variation (Ervin-Tripp 1972; Hymes 1974; Brown & Fraser 1979; Chafe 1982; Halliday 1988) defended that registers, genres and text-types needed to be described by looking at co-occurrence patterns of different linguistic features. A few sociolinguistic studies focused on a particular parameter of register variation and analysed a group of linguistic features associated with that parameter. Such studies would correspond to Type 2 in Grey's (2011) classification. For instance, Irvine (1979) analysed formal and informal registers, while Ochs (1979) analysed planned versus unplanned discourse. Taking a step further, Longacre (1976) and Chafe & Danielewicz (1986) suggested that certain linguistic co-occurrence patterns were related to certain parameters of variation, or communicative functions, in a relationship of opposition.

In an attempt to combine the microscopic and macroscopic perspectives in one study, and considering that neither a single linguistic feature, nor a single set of linguistic features associated to a particular communicative function can provide a comprehensive description of register variation, Biber (1988)[6] developed his Multidimensional Analysis of variation in a large corpus[7] of spoken and written English. His approach presents three major differences with respect to previous variation studies (Biber & Conrad 2001: 8). The first one is the assumption that one single parameter, or dimension, cannot capture the full range of variation among the registers of a language. Instead, there are several dimensions, each conveyed by a set of co-occurring linguistic features which share an underlying communicative

[6] Although fully developed in Biber (1988), the Multidimensional approach was first used in Biber (1984, 1985, 1986).

[7] See description of Biber's (1988) corpus in Section 3.1.

function. This appears to agree with Biber & Conrad's (2009: 9) postulate that "registers differ in their characteristic distributions of pervasive linguistic features, not the single occurrence of an individual feature". The second difference, also related to the notion of dimension, is that variation is considered along a continuum, rather than through dichotomous distinctions (e.g. formal versus informal register). More accurately, dimensions indicate a range between "very" and "not at all", and the different registers vary from "more" to "less" (e.g. more or less formal with respect to each other). Finally, unlike Longacre's (1976) and Chafe & Danielewicz's (1986) parameters, which were identified on an intuitive basis, Biber's (1988) dimensions of variation are identified in the MD approach empirically, by means of a multivariate statistical technique (factor analysis; see below)[8], combining both quantitative and qualitative research.

In practice, Biber's (1988) MD analysis has shown that some linguistic features – whether lexical items or grammatical constructions – tend to appear in texts in the company of other linguistic features, forming groups, or sets of co-occurring features, and that each set of features has an underlying communicative function. Different sets of co-occurring features have been found to appear in different texts and in different registers. Moreover, it has also been demonstrated that some of those sets of co-occurring features appear in complementary patterns (i.e. the presence of a particular set of features entails the absence of another particular set of features) because they fulfil mutually excluding functions in the discourse. For instance, first and second person pronouns, present tense verbs, contractions and general emphatics tend to appear together, or co-occur, because they all contribute to an involved, interactive type of discourse. Conversely, nouns, attributive adjectives and prepositions (which also tend to appear together) characterise a dense, highly informational kind of discourse. Thus, texts or registers presenting a high frequency of features from the first set would normally show little or no presence of features from the second set, and vice versa. This complementary distribution shows how registers can be distinguished with regard to their pervasive linguistic patterns.

While Biber's (1988) study constitutes in itself a model of register variation in contemporary English, its methodological technique has proved to be fully replicable

[8] Biber (2014: xxxi) accredits Carroll (1960), mentioned earlier in this section, to be the first variation study to use factor analysis in order to identify patterns of linguistic co-occurrence, although Carroll's (1960) paper does not appear to have had the same impact as Biber's (1988) MD analysis.

and can therefore be applied to any other corpus, resulting in a new model of register (or sub-register) variation (see other MD studies in Section 3.2). A description of the methodology used in Biber's (1988) MD analysis is given in the following paragraphs.

3.1. Methodological outline of the MD analysis

As summarised in Conrad & Biber (2001: 13-14), a complete MD analysis consists of eight methodological steps. A more detailed description of each step as applied in Biber (1988) follows the list.

1. First of all, a corpus is designed. Texts are first collected (spoken texts, if used, are transcribed) and eventually input into the computer to be later processed through a specialised corpus process software.
2. Research is conducted in order to identify the linguistic features that are to be included in the analysis and the functional associations of those linguistic features.
3. Usually, a computer programme is developed to annotate the texts with part-of-speech (POS) tags.
4. The whole corpus is annotated with POS tags so that the linguistic features included in the analysis can be automatically retrieved through another computer programme.
5. Additional computer programmes (such as concordancers) are developed so that frequency counts of each linguistic feature can be computed for each text in the corpus.
6. A multivariate statistical technique, known as factor analysis, is applied to derive co-occurrence patterns from the feature counts in each text. The co-occurrence patterns are seen from the factors, resulting from the factor analysis.
7. Each factor is interpreted functionally as an underlying 'dimension of variation'. Thus, in principle, number of factors = number of dimensions.
8. Dimensions scores are computed for each text with regard to each dimension. Eventually, mean dimension scores are computed for each register and compared in order to analyse linguistic similarities and differences among the registers.

In Biber's (1988) analysis, a corpus of twenty-three spoken and written present-day English registers was used. The study included a sample from the *Lancaster-Oslo-*

Bergen Corpus of British English (*LOB Corpus*; see Johansson et al. 1978, Johansson 1982), which contained 500 texts of ca. 2,000 words, and from the *London-Lund Corpus of Spoken English* (Svartvik & Quirk 1980; Johansson 1982), plus a few texts, with 100 spoken British English texts of ca. 5,000 words. The former contains fifteen written registers, including editorials and press reportage, academic prose, legal documents and different types of fiction. The latter contains six spoken registers: private and public conversations, telephone conversations, radio broadcasts, and spontaneous and prepared speeches. In addition, Biber's (1988) corpus included a sub-corpus of letters that were collected separately.

The linguistic features for the analysis were selected on the basis of previous research which focused on particular lexical items or grammatical constructions and their communicative functions. This is the part where the microscopic perspective is integrated into the study. After a careful research, Biber (1988) selected a total of sixty-seven relevant linguistic features, which were grouped according to sixteen major grammatical and functional categories (Biber 1988: 73-75; Conrad & Biber 2001: 17):

1) tense and aspect markers
2) place and time adverbials
3) pronouns and pro-verbs
4) questions
5) nominal forms
6) passives
7) stative forms
8) subordination features
9) prepositional phrases, adjectives and adverbs
10) lexical specificity
11) lexical classes
12) modals
13) specialised verb classes
14) reduced forms and dispreferred structures
15) coordination
16) negation

A comprehensive description of the linguistic features used in the study, as well as the different studies analysing their communicative functions, are given in Biber (1988: Appendix).

Computer programs were developed in order to annotate the corpus for parts of speech (POS tagging). Biber (1988) used his own 'tagger' (see Biber 1988: Appendix II; Biber 2014: xxxi). Linguistic features could be retrieved from the corpus automatically, either with the POS tag searched for, or by directly entering lexical items. In principle, these steps in the study reflect a corpus-based approach in that the researcher looks for particular linguistic features, identified through previous research. However, when searched by POS tags, the resulting grammatical structures were not lexically determined *a priory* (e.g. the actual verbs, nouns or prepositions conveyed by the structure), and such searches might therefore be regarded as part of a hybrid, rather than solely corpus-based approach.

After that, another computer program was built to count each feature in each text of the corpus. Raw frequency counts were normalised per 1,000 words of text so that they could be directly comparable among themselves, regardless of the total number of words in the samples. Although normalised frequency counts already permit to compare the different texts with regard to the frequencies of each linguistic feature, this comparison does not provide information about co-occurrence patterns. For instance, we may find out that first person pronouns are relatively much more frequent in telephone conversation text than in legal documents or in academic prose texts. However, how this particular linguistic feature co-occurs with other linguistic features in these and other texts from the corpus cannot be known unless each of those linguistic features are compared to each other in every single text. Roughly, this is the work of factor analysis, a multivariate statistical technique of data reduction. The application of a factor analysis corresponds to the corpus-driven part of this study. Factor analysis reduces a large number of original variables (in this case, linguistic features) into a smaller set of derived variables, or *factors* (also called latent variables or latent constructs; see Biber 1988: 79; Tabachnick & Fidell 1996: 636). In Biber's (1988) study, a factor analysis was run on a dataset of sixty-seven linguistic features, seven factors being extracted. Although virtually all sixty-seven features have a particular weight, or *loading*, on each factor that is extracted, some features have larger loadings on one factor, while other features load more strongly on another, and so on. The features which load more strongly on a particular factor are said to correlate to each other, which means that they normally occur together in the texts. In other words, each of the seven factors extracted in Biber's (1988) study is, in fact, a set of co-occurring linguistic features that share an underlying communicative, or discursive, function. [9]

Based on each group of features, Biber (1988) interpreted his seven factors as "dimensions of variation". For example, Biber's (1988) Factor 1 (*ibid*: 89; Conrad & Biber 2001: 21-23) has a number of linguistic features with large loadings: first person pronouns, second person pronouns, present tense, place adverbs, and pronoun *it*, among others. Conversely, other linguistic features do not have much weight (i.e. have smaller loadings) on Factor 1 but load strongly on other factors. The features with large loadings on Factor 1 are correlated, which means that they share an underlying communicative function, relative to a particular parameter of register variation. On the other hand, another group of features – nouns, word length, type/token ratio, prepositions and attributive adjectives – also load strongly on Factor 1, but have sub-zero values. The fact that one group of features has 'positive' loadings and another group has 'negative' loadings on a factor indicates that the two groups of features occur in the texts in a complementary pattern, which means that they have opposite communicative functions. Having carefully considered each group of features, Biber (1988: 104-108) distinguished between them as "involved" (positive features) and "informational" (negative features). On this basis, Biber (1988: 107) labelled Factor 1 as Dimension 1 "Informational versus Involved Production" (see Figure 1.1 on the next page).

[9] This description of factor analysis has an introductory character and does not explain relevant concepts in statistics, such as *correlation*, or *variance*, which are central to factor analysis. A full technical description of factor analysis is offered in Chapter 5.

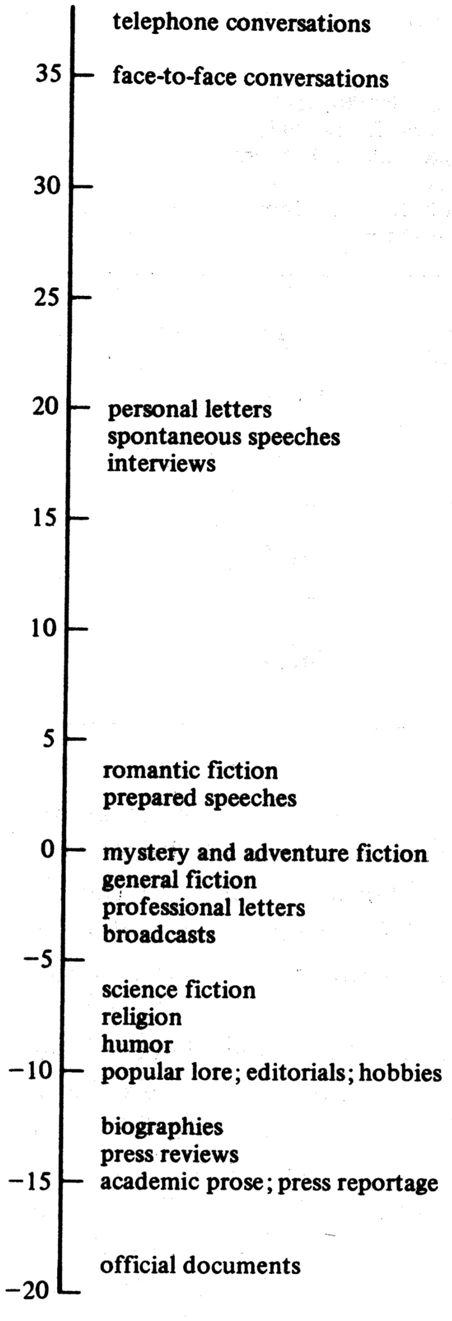


Figure 1.1
Mean scores for different genres on Dimension 1 “Involved versus Informational Production” (taken from Biber 1988: 128)

Out of the six remaining factors, five more dimensions of variation were identified in Biber's 1988 study, with the following functional labels: Dimension 2 "Narrative vs. Non-narrative Concerns"; Dimension 3 "Elaborated vs. Situation-Dependent Reference"; Dimension 4 "Overt Expression of Persuasion"; Dimension 5 "Abstract vs. Non-Abstract Information", and, finally, Dimension 6 "On-line Informational Elaboration" (whereas Factor 7, as explained in Biber (1995), did not offer a convincing explanation and was subsequently dropped out of the model). The two 'poles' or extremes of a factor represent opposite discursive functions (such as involved vs. informational; narrative vs. non-narrative), and each factor is thus interpreted as a 'dimension', or continuous scale of variation, along which the different texts and registers can be situated in function of the distribution of their linguistic features (Biber 1988: 79-97). For this purpose, *factor scores* are computed for each text by adding up the standardised frequencies[10] for each linguistic feature that corresponds to a particular factor. Likewise, mean factor scores can also be computed for each register by adding up the factor scores of all the texts in a register and dividing the resulting number by the number of texts (see Chapter 5). As explained in Conrad & Biber (2001: 24), the interpretation of factors as functional dimensions of variation does not only depend on the analysis of the communicative function(s) shared by each set of co-occurring features, but also on an analysis of the similarities and differences among the different registers with respect to each factor. Thus, when compared along each dimension of variation, texts and registers can be described as more interactional than others, more narrative than others, and may present a more abstract or impersonal style than others, in function of the sets of co-occurring features characterising each of them.

The six dimensions of variation identified in Biber's (1988) MD analysis provide a comprehensive description of the relationship among the different spoken and written registers in present-day English. This model has served as a baseline to a wide range of corpus-based variation studies which described a particular register or set of registers with respect to the dimensions of variation established in Biber's (1988) study. However, its statistical methodology can also be replicated "from scratch", giving way to new MD analyses. The following paragraphs present some of

[10] Normalised frequencies are standardised to a mean of 0.0 and a standard deviation of 1.0 before computing factor scores. See Chapter 5 Section 4 for a detailed explanation of the several methodologies that may be followed in this process.

the numerous MD studies that have been conducted since Biber (1988), either using the latter as a reference, or conducting a separate factor analysis which results in new dimensions of variation. Likewise, alternative statistical methodologies to carry out a comprehensive analysis of register variation will be mentioned.

3.2. Types of MD and other comprehensive analyses of register variation

As explained in Gray (2011: 25-29), MD analyses can be either corpus-based or hybrid, whereas completely corpus-driven comprehensive studies of register variation are not normally conducted.[11] Those studies that apply Biber's (1988) six-dimensional model are essentially corpus-based in that both the linguistic features selected for analysis and their co-occurrence patterns are adopted from a previous study in order to describe a new register or language. Thus, instead of running a factor analysis – skipping steps 6 and 7 in the MD methodology – researchers compute their frequency counts for each of the previously identified dimensions, sometimes comparing them *against* the scores of Biber's (1988) registers.

Often, such studies focus on a specialised discourse domain. For instance, Conrad (1996, 2001) compares two academic disciplines – biology and history – within the academic register, and, subsequently, analyses research articles and textbooks in each academic discipline, analysing subregisters within subregisters. Csomay (2000, 2002) looks at variation within the subregister of academic lectures, while Carkin (2001) analyses pedagogic language in textbooks and lectures in Biology and Macroeconomics, and Biber & Finegan (2001b) analyse the IMRD subsections in research articles. From a historical perspective, Atkinson (1996, 1999, 2001) applies Biber's (1988) dimensions to scientific discourse from 1675 to 1975, balancing his MD analysis with an additional rhetorical analysis. Biber & Finegan (1989, 1992, 1997, 2001a) use three of the 1988 dimensions to analyse the evolution of several specialised and non-specialised English registers along four centuries. Similarly, González-Álvarez & Pérez-Guerra (1998) analyse the same registers as Biber & Finegan (1989, 1992, 1997) in earlier stages of the language (from the fifteenth to the seventeenth century). On a relatively shorter time-scale, Westin &

[11] See Gray (2011, Chapter 3) for an exhaustive classification of corpus-based, corpus-driven and hybrid analyses of register variation, compared with regard to the comprehensiveness of the linguistic features analysed.

Geisler (2002) track diachronic change in twentieth-century British newspaper editorials, comparing the *Guardian*, the *Daily Telegraph* and *The Times*.

Other studies target demographic variation, such as Biber & Burges' (2000, 2001) analysis of male and female dramatic speech, belonging to male and female authors, or Helt's (2001) comparison of conversational registers in British and American dialects. Recent studies applying the 1988 model of register variation include Berber Sardinha (2014), who compares Internet and pre-Internet registers by comparing Biber's (1988) corpus with a representative corpus of online communication, as well as Condi de Souza's (2014) diachronic analysis of variation in *TIME* magazine, which, in fact, uses both Biber's (1988) dimensions and carries out a new MD analysis, allowing a richer perspective. In addition, some researchers focus on one particular dimension from Biber's (1988) MD, rather than on the whole set. For instance, Crespo (2012) analyses the linguistic features of Dimension 4 "Overt Expression of Persuasion" in eighteenth-century Astronomy texts, comparing persuasive strategies used by male and female authors. Moskowich (2013) looks at linguistic features conveying abstractness as represented in Biber's (1988) Dimension 5 "Abstract vs. Non-Abstract Style" in scientific texts written by women in the 1700s, whereas Moskowich & Monaco (2014) extend this analysis to the nineteenth century.

In contrast, studies applying the MD methodology in order to carry out new MD analyses are, like Biber's (1988) study, hybrid in their approach, combining top-down and bottom-up research. Although previous research is used to decide what linguistic features should be retrieved for the analysis, factor analysis entails new information in the form of co-occurrence patterns resulting in new dimensions of variation. According to Gray (2011: 29), a hybrid approach can sometimes be more useful than an entirely top-down approach because particular linguistic features identified in previous research, such as a previous MD analysis carried out on a particular register or set of registers, may not have the same function(s) in the new register that is being investigated.

Lee (2000) carried out a MD analysis on a different multi-register corpus of contemporary English, the *BNC*, proving that Biber's (1988) dimensions of variation are largely replicable on a similar corpus and, as such, constitute a valid model of register variation research. In order to explore the possibilities of a MD analysis on a smaller sample from the *BNC*, Gómez Guinovart & Pérez Guerra (2000) also ran a factor analysis on a set of forty-eight linguistic features, partially different from those analysed in Biber (1988), with a resulting bi-dimensional model and with two additional dimensions which were difficult to explain with regard to the differences among the texts analysed. Many other studies are listed in Gray (2011: 28) and Biber (2014: xxxii). Among those, Reppen (1994, 2001) researches elementary school spoken and written registers, while White (1994) investigates variation in the field of job interviews. Biber (2001) conducted a separate MD analysis on a set of eighteenth-century speech-based and written registers, with four new dimensions of variation. Some time later, university spoken and written registers were exhaustively analysed in Biber (2006), with four resulting dimensions of variation. Hämäläinen (2008) discovered three dimensions of variation by looking at variation in software engineering articles. On the other hand, Gray (2011) examined academic writing by comparing different types of research articles across six scientific disciplines, coming up with four dimensions of variation specific to the academic register. Other hybrid MD studies revealed new dimensions of variation in conversation (Biber 2008), call-centre discourse (Friginal 2009), spoken and written registers in World English (Xiao 2009), blogs (Grieve et al. 2011), regional variation in American English (Grieve 2014, 2016), as well as variation in North American movies (Veirano Pinto 2014) and in pop songs (Bértoli-Dutra 2014).

Furthermore, from its early stages, the MD approach has also been applied to languages other than English (see Biber 2014: xxxii), such as Nukulaelae Tuvaluan (Besnier 1988), Somali (Biber & Hared 1992, 1994), or Korean (Kim & Biber 1994). These three studies are extensively described in Biber (1995). Other studies conducted MD analyses in Taiwanese (Jang 1998), Spanish (Biber et al. 2006; Parodi 2007; Asención-Delaney & Collentine 2011; Asención-Delaney 2014), Czech (Kodytek 2008), Bagdani (Purvis 2008), as well as Brazilian Portuguese (Berber Sardinha et al. 2014).

We have thus seen that, whether we wish to apply previously established dimensions of variation to compare a new register or language against a set of previously analysed registers, or whether we decide to conduct a new factor analysis “from scratch” that will result in new dimensions of variation, the MD approach is a useful tool in that it offers an comprehensive description of the relationship among different registers by looking at their pervasive linguistic features. These linguistic features usually appear in the texts in co-occurrence patterns, indicating that they share a basic communicative function, and are therefore represented as dimensions of variation along which the different texts, and registers, are situated.

However, the MD analysis is not the only statistically quantitative method that is currently used to explore variation in language. For instance, Xiao & McEnery (2005) argue that keyword analysis – specifically, through Tribble's (1999) keyword function in the WordSmith corpus software program[12] (Scott 1999) – can achieve similar results to those in Biber's (1988) MD analysis, providing a "low effort" alternative (Xiao & McEnery 2005: 68). Takahashi (2006) studies variation in the written *BNC* with respect to several variables (such as domain, age group, place of publication and gender), using the Extended Hyashi's Qualification Method Type III, which is equivalent to correspondence analysis (Takahashi 2006: 124) and, to some extent, similar to factor analysis (Mohamed 2011: 150-151). Croft & Poole (2008) use another multivariate statistical technique, known as multidimensional scaling (MDS), for a cross-linguistic typological analysis. Mohamed (2011), in turn, uses cluster analysis (following Biber 1989) on the *BNC* to study the different text types in spoken and written English.

Although it is not our intention to dwell on the description of these studies any further, it is important to bear in mind that the MD approach is one of the several different possible ways currently available to analyse variation and change among different text varieties, even if it has proved to be particularly appropriate for looking at register variation. In the present study, a MD analysis will be carried out in an attempt to reveal dimensions of variation in a corpus of late Modern English scientific register. Although a description of the evolution of science and its language along the past four centuries will be given in Chapter 2, the next – and last – section briefly introduces the category *scientific register* as it is conceived of in variation studies.

4. The scientific register

If, sticking to Biber & Conrad's (2009) terminological framework, we define register as a situational variety that can be identified and described by its pervasive linguistic features, we should then describe the scientific register in function of the typical linguistic features that characterise it. In the previous section we have mentioned a number of MD studies that analyse scientific register in its different domains and sub-

[12] Currently, other corpus-processing software such as AntConc (Lawrence 2009) or CQPWeb (Hardie 2012) also have the keyword analysis function key.

domains, such as specific academic disciplines and research subregisters (Conrad 1996, 2001; Csomay 2000, 2002; Carkin 2001; Hämäläinen 2008; Gray 2011). These studies show that the scientific register is quite a complex type of discourse that presents internal variation on many levels. For example, Conrad's (2001: 98-99) analysis of history and ecology research articles and textbooks reveals that these subregisters are all very informational along Biber's (1988) Dimension 1, although the former are even more informational than the latter; but that they differ along Dimension 2 in that history texts are considerably more narrative than ecology texts. On the other hand, Gray's (2011: 164-165) extensive comparison of humanities, social sciences and hard sciences reveals, among other findings, that hard sciences present very few finite complement clauses, reflecting little or no use of in-text citations and statements of purpose compared to social sciences and humanities. Biber (1988) also explored variation within the register of academic prose, looking at the distribution of the different subregisters (e.g. humanities, medical, mathematics, technology/engineering) along his six dimensions of variation. While academic prose is characterised in Biber's (1988) study to be "considerably abstract" on Dimension 5 "Abstract vs. Non-Abstract Style" (with a mean dimension score of 5.5), it was discovered that there is a considerable difference between humanities (3.0) and technology/engineering (9.8) academic prose, meaning that the latter is characterised by a much more abstract, impersonal linguistic style than the former (Biber 1988: 189). Although these examples only account for a tiny proportion of the variation found in these and other studies, they are sufficient to demonstrate that the scientific register is very far from being a homogeneous, stable category.

Apart from these MD studies, a wide range of research has been conducted on the internal and formal characteristics of the scientific register, sometimes referring to it as "academic register" or "academic discourse" (Ulijn 1989; Gerzymisch-Arbogast 1993; Taavitsainen and Pahta 2004; Biber 1995, 1998, 2006; Hyland & Bondi 2006), sometimes as "scientific writing" (Bazerman 1984, 1988) or "scientific discourse" (Swales 1974, 1990; Atkinson 1999; Gotti 2001, 2003, 2008), and sometimes as "the language of science" (Halliday 1998). Scientific discourse may be regarded as a type of specialised discourse, which in turn can be defined as "the specialized use of language in contexts which are typical of a specialized community stretching across the academic, the professional, the technical and the occupational areas of knowledge and practice" (Gotti 2003: 24). Specialised discourse, in turn, has a number of characteristic features that, in view of the above definition, should also stretch to the academic domain; these features include monoreferentiality, lack of emotion, precision, transparency, conciseness, and conservatism (Gotti 2008: 33-41). In fact, scientific discourse has been often described as impersonal, detached and faceless (Besnier 1994, Hyland 1995, in Moskowich 2013). In the second half of the twentieth century, the term *scientific discourse* has developed an immediate association with the rather rigid organisational scheme of the IMRD subsections which normally frames empirical research (Swales 1990; Sollaci & Pereira 2004). This is an example of the strict discourse conventions that allow scientific reading and writing to be "performed with maximal efficiency" (Atkinson 1999: 6-7). Biber (1988) and Atkinson (1996, 1999) demonstrated that the impersonal style is conveyed to the scientific discourse – disregarding internal variation – through a high frequency of passive constructions and nominalisations. All in all, in his famous article "On the Language of Physical Science", Halliday (1988: 162) defines *scientific English* as "a useful label for a generalised functional variety, or register, of the modern English language", considering it "a semiotic space within which there is a great deal of variability at any one time, as well as continuing diachronic evolution", but insisting that this spatial-temporal variation "in no way distinguishes scientific English from other registers". Bazerman (1988: 8), on the other hand, described "the canvas of scientific writing" as "vast and growing". This allows the scientific discourse (or writing, or register) to be seen as a dynamic, continuously changing variety.

Along the past fifty years, scientific discourse has become a popular object of variation studies. Functional values of diverse lexico-grammatical features, such as verb tense or noun phrases, were explored in Oster's (1981), Channel's (1981) and Swales & Najjar (1987) research on diverse scientific genres, such as chemistry texts, biomedical journal articles, or research article introductions, while Halliday (1989, 1990) concentrated on grammatical functions and problems in scientific English. In a study of hedging in the scientific discourse, Myers (1989) highlights the importance of the scientific community which is always present for the scientist when (s)he is stating his/her claims. This idea was later developed in Salager-Meyer (1994), Hyland (1995, 1996, 1998a,b), Crompton (1997), Lewin (2005) and Alonso-Almeida (2012), among others, who analyse hedging in scientific writing both on a lexical and on a grammatical level. On the other hand, Bazerman (1988) and Hunston (1993) insist on the persuasive function of scientific writing, which entails that scientists seek to persuade the reader (i.e. the scientific community) of the validity of their claims (Hunston 1993: 58). Allen et al. (1994) and Montgomery (1996) demonstrate that the scientific language has had a persuasive character since its earliest stages, dating back to the seventeenth century, even though the ways of conveying persuasion have evolved with time. In fact, a large number of sociohistorical studies (Atkinson 1996, 1999; Biber and Finegan 1989, 1997, 2001a; Taavitsainen and Pahta 2004, among others) show that the scientific register has been constantly adapting to continuously changing socioeconomic and rhetorical conventions, while some of its early characteristics have gradually consolidated into what we now regard as inherent to the language of science.

Considering scientific discourse "a culturally contingent rhetoric, one that is dependent on cultural norms and historical periods", Zerbe (2007: 19) explains the nature of its changing dynamics in how they are both caused and perceived by the scientific community:

> A wide range of discourse can be described as "scientific", and, over time, many different forms of scientific discourse have appeared. As scientists have developed more sophisticated techniques to conduct their investigations, recognized biases and conflicts of interests, and noticed flaws of logic and mismatches between evidence and hypotheses, they have criticized earlier forms of scientific discourse as naive and sadly misinformed.

In Chapter 2, the evolution of the scientific register along the seventeenth, eighteenth and nineteenth centuries will be analysed as inseparable from the evolution of science and scientific thought.

Chapter 2

Scientific English in late Modern times

> There is much in common between the mediaeval monkish scholar and the present-day research worker: books and manual work, quiet concentration on a small field, corporate action.
>
> H. T. Pledge, *Science since 1500* (1959)

1. Introduction

In the previous chapter we have seen that the *scientific register*, while by no means a monolithic one, is generally agreed to have a number of distinguishable characteristics. These characteristics, whether strictly linguistic (lexical, grammatical) or discursive (genre/text-type), or situational (epistemic community, type of readership, etc.), in their different combinations and permutations, may appear as merely conventional nowadays if we look at the different subtypes of scientific register that are used in the different subfields of science. An example of this apparent conventionality is the widely accepted *IMRD* structure of research articles (Introduction, Method, Results and Discussion), which is normally followed in experimental studies, regardless of their discipline. This structure has been found to reveal micro-purposes conveyed through particular linguistic characteristics (Swales

1990: 133-137; Biber and Finegan 2001b: 108-109). Another example are the linguistic and stylistic guidelines for the submission of manuscripts to scientific journals, which tend to vary according to the scientific discipline dealt with. Although such conventions of language and style may appear as established *ad hoc*, most of them had been developing for decades, and some for centuries, before consolidating into what they are nowadays.

It is widely acknowledged that linguistic variation and change cannot be analysed without looking at the sociolinguistic background standing behind this variation and change (Smith 1996; Moskowich 2001). This chapter, therefore, aims to introduce the so-called scientific English register from a diachronic perspective in the light of its sociohistorical context and provide an insight into some of the key stages in the development of English as a language of science. Section 2 delimits the period in the English language covered in the present study, referred to here as "late Modern". In turn, Sections 3 and 4 offer a brief account of the emergence of Western science and its language, respectively, while Sections 5 and 6 deal with subsequent stages in their development along the eighteenth and nineteenth centuries.

2. The late Modern English period: scope and characterisation

If we look at the history of the English language along a timeline in the literature, we will find that the boundaries demarcating subperiods are generally vague, and that a time span labelled as *late Modern* (as opposed to *early Modern*) English may, or may not, appear. It was in the late nineteenth century that the phonetician Henry Sweet proposed to divide the English language into Old, Middle and Modern, "based mainly on the inflectional characteristics of each stage", where Old English would be "the period of full inflections", Middle English "the period of levelled inflections", and Modern English "the period of lost inflections" (Sweet 1873-4: 620, in Beal 2004: 1 and in Matthews 2000: 53). Although this model has been successfully adopted, further subdivisions have been deemed necessary. Particularly in what concerns Modern English, bearing in mind that the loss of inflections occurred along the fifteenth century, it appears that so straightforward a classification would fail to do justice to the changes that occurred in the English language along the past five hundred years (Moskowich 2001: 625). Hence, the terms early Modern English (eModE) and late/later Modern English (lModE) are often used nowadays for convenience in order to separate two main stages within the Modern English period.

The starting date for the eModE subperiod is generally agreed to be the late fifteenth century, often rounded to 1500 (Barber 1976; Millward 1989; Görlach 1991, 1994; Burnley 2014). The *Cambridge History of the English Language* (Lass 1999) establishes very precise dates both for the start of eModE (1476, when Caxton introduced the printing press in England) and for its end (1776, with the American Declaration of Independence). Considering that language does not change overnight, and that, therefore, the abovementioned dates are to be read loosely, authors usually highlight a number of linguistic and extra-linguistic happenings, dated from the mid-fifteenth to the mid-sixteenth century, as essential to the transformation of Middle English (ME) into eModE. Apart from the loss of inflections, the gradual rise of a new written standard after 1430 and the major change in phonology, known as the Great Vowel Shift, caused a definitive separation between English orthography and pronunciation. Printers were established in London and books were published in the London dialect (Millward 1989: 124). Literacy started to increase, not only as a consequence of the availability of printed materials, but also thanks to the secularisation of education and its comparatively fast expansion among the population, which was further strengthened after the replacement of Latin by English as the language of the Church with the Reformation in the sixteenth century (Bourcier 1981: 178-182; Mele-Marrero and Martín-Díaz 2001: 574-578). All these factors, along with the end of the medieval feudal system after the Wars of the Roses in 1471, the coming of the Tudors to the throne in 1485, and the discovery of America in 1492, characterise the fifteenth century as "a transitional period, both linguistically and culturally" (Görlach 1991: 10).

In what concerns the duration of the eModE period, some authors stretch it as far as to 1800 (Millward 1989; Burnley 2014), although others, such as Baugh and Cable (2000), make a clear division between "The Renaissance, 1500-1650", a time of commercial, technological and cultural expansion which gave rise to an unprecedented growth of literacy and enrichment of the English language, but also of political agitation that erupted into the English Civil War in 1642, and "The Appeal to Authority, 1650-1800", a period of obsession over security, stability and correctness which extended to all fields, including language (Moskowich 2001: 624). The middle date, which broadly coincides both with the Restoration and the foundation of the Royal Society of London in 1660, is often regarded as the separating point between the eModE and lModE periods (Bailey 2003; Beal 2004). On the other hand, distinctions based on primarily linguistic criteria focus on a number of phonological, grammatical and lexical changes that emerge along the sixteenth and seventeenth centuries and consolidate in the eighteenth century (Barber 1976; Görlach 1991), and set therefore 1700 as the closing date of the eModE period. In agreement with this line of thought, Moskowich (2001: 625) considers that the start of lModE[13] takes place in the eighteenth century, at the time when authors prefer to use the standard variety of the language, independently from their dialectal origin, and when academics, grammarians and dictionary makers try what is in their power to "fix" and "correct" the English language (see Section 3).

As for the duration of the lModE period, it should be borne in mind that, to some linguists and historians of the language, *late Modern* equals "contemporary", and thus whenever it is specified that lModE spreads over an undefined period "since 1700" (e.g. Barber 1976), it could be considered to overlap to some extent with the label Present-day English (PrE). PrE in turn, when first used by Wyld (1920), was intended to cover a period ranging "from the beginning of the eighteenth century to the time of writing" (Beal 2004: 1-2). Indeed, 1800 – if not 1700 – may not have appeared so far-reaching a point when looked at from a little more than a hundred-years distance. However, the farther we move away from the object of observation, the better the perspective becomes (Beal 2004: xii), and it is evident in our early twenty-first century that the English we write and speak today *has* evolved with respect to the English that was written (and presumably spoken) three hundred years ago, even if most of the changes it underwent were lexical and stylistic, rather than morphological or syntactic (Bailey 1996; Denison 1998).

In order to establish the terminological boundaries for the time span contained in the present study (1700-1900), and taking into account that language change necessarily goes hand in hand with certain extra-linguistic events which, in one way or another, trigger that change (Smith 1996; Moskowich 2001), I will adopt the term lModE from Beal (2004: 2) as starting in 1660 and covering the so-called "long" eighteenth and nineteenth centuries, up to the end of World War I in 1918. My choice

[13] Moskowich (2001) uses the term ModE (Modern English), rather than lModE, as opposed to eModE.

for this rather long temporal segment is based on the importance of transitional periods in language change, which I consider particularly relevant to the scope of this study, despite the fact that the actual samples analysed only cover the time span comprehended between 1700 and 1900. In particular, the boundaries demarcating the lModE period specified earlier coincide with a fundamental phase in the evolution of scientific English as a register, as will be seen in Sections 3 and 4 of the present chapter. On the other hand, doing justice to the concept *transitional period*, I also consider that the boundaries specified above are fuzzy, rather than sharp, and that, consequently, the initial and final days of the lModE period must overlap with the end and start of the eModE and PrE periods, respectively.

3. Science and language in the Early Enlightenment

Although the focus of this section is the long eighteenth century, this key period for science and its language cannot be understood without looking back in time, well before the Renaissance. In fact, Western science may be said to have been born in Ancient Greece, although since the division of the Roman Empire around A.D. 395 it would not arrive in Christian Europe until the twelfth century, when the continent began to experience a cultural, technological and economic revival that characterised the High Middle Ages and constituted "our First Renaissance of Learning" (Pledge 1951: 11). This was a time of various improvements and technical developments, such as the clock or the magnetic compass, as well as the growth of towns and trade and the spread of water power. In what concerns knowledge, it was the time of Saint Thomas Aquinas' scholasticism, which would characterise science and scientific practice for three hundred years to come. This period has been largely described as an epoch of endless, dogmatic study of classical texts which had to be interpreted from the confinement within the walls of a convent cell and the sacred covers of the Bible, and with little access to the knowledge of the natural world. Despite this fact, some historians (Crombie 1969; Cabezón 1998; Roy 1998) insist on the necessity to give the deserved credit to a period without which neither the second Renaissance, nor the Scientific Revolution would have taken place. The growing thirst for an improved scientific method was, to a large extent, owing to the efforts of those who devoted their lives to the Socratic pursuit of the truth for its own sake, and who attempted to accommodate an explanation to anything unknown, and inexplicable otherwise, within the – at that time highly malleable and even metaphysical – frames of logic. Logic, in turn, was intrinsically linked to grammar, which must at that time have been regarded as sacred as the Bible itself. And only Latin had at that time a perfect grammar (as opposed to vernacular languages, which were considered corrupt), and was therefore the only vehicle for the transmission of knowledge, as it also was the only vehicle for the transmission of the Word of God. Still, even though this highly conservative way of learning may seem sterile and even vicious (Dear 1995; Burke 2000), the exhaustive attempts of scholastic philosophers to interpret the Bible and reinterpret classic, patristic, and earlier scholastic "authorities" may in fact be regarded as the basis of modern scientific thought; for it is common to find this "problem-solving mentality", characteristic of scholasticism (Roy 1998: 24), in the young scientists of our time, who would often "solve problems arising in their elders' theories rather than in 'reality'" (Pledge 1959: 23).

Already in the thirteenth century Robert Grosseteste, and later his disciple Roger Bacon, argued in favour of experimentation, and the latter's detailed examination of the cause of the rainbow through its recreation in various natural objects (i.e. by looking at the colours in crystals or in dew drops), as well as through observation of the rainbow itself and taking of measurements was indicative of an accurate conception of the inductive method (Crombie 1959: 117). In the fourteenth and fifteenth centuries, the rise of long-distance navigation brought forth a practical interest in geography and astronomy that demanded the making of terrestrial and celestial maps and globes. Yet, it was the necessity to solve the so far unsolved problems of both classical and medieval science that must have inspired Galileo, Kepler, or Francis Bacon to combine mathematics, observation and experiment – the essence of the Scientific Revolution in the seventeenth century. And although it cannot be always affirmed with certainty to what extent the scholastic doctrines could have influenced the works of early Modern scientists, indebtedness to the ancient past may be seen in the use of language by some of them, whether it is their using certain old terms with new meanings, or certain established labels to describe new methodological procedures. This so-called "linguistic inertia", according to Crombie (1969: 95), "is evidence of continuity with earlier forms of thought, whatever changes the requirements of successful scientific practice may have brought about".

In line with this idea of continuity, Shapin (1996: 1) opens his analysis of the Scientific Revolution by stating that "[t]here was no such thing as the Scientific Revolution", but, rather, "some self-conscious and large-scale attempts to change belief, and ways of securing belief, about the natural world" (1996: 5). This reformulation – or, even, negation – of a widely accepted historical term reflects a more analytical, less anachronistic, and perhaps less enthusiastic perspective on what has traditionally been considered a groundbreaking change from "old" to "new" science (Hall 1954; Kuhn 1970). Moreover, it might be argued that *Scientific Revolution* is also an unsuitable term because, strictly speaking, there was no such *science,* as used in the modern sense of the word, in the sixteenth and seventeenth centuries (Shapin 1996; Park & Daston 2006; Camiña-Rioboo 2013). Instead, there were a variety of branches of knowledge, presenting a wide range of problems, theoretical and practical, which at that time belonged to what was called natural philosophy. Likewise, "proper sciences" and "pseudosciences" (Shapin 1996: 6) such as astronomy and astrology, or chemistry and alchemy, still coexisted and were not entirely distinguishable. However, Shapin (1996: 5) subsequently counteracts his earlier attack on the idea of a Scientific Revolution, defending that it was precisely at that time – in the late sixteenth and seventeenth centuries – that a group of natural philosophers from different corners of Europe decided to step forward and proposed new ways of approaching the study of the natural world; new ways that, in comparison to those of the "ancients", would revolutionise science. In essence, the decision to make experiment an indispensable part of scientific practice caused not only a notorious advancement in technology, with the growing need for innovative instruments such as the telescope and the microscope, but also a gradual change in the conception of scientific "experience" itself (Dear 1995: 13). If at the beginning of the seventeenth century experimentation was still conceived of largely in its Aristotelian sense – i.e. the statement of facts "known to the senses" –, at the end of the century experiments will be carried out expressly for the scientific study and will be meticulously described in the writings of natural philosophers. The modern conception of scientific method will thus have changed forever since.

On the other hand, it is also worth mentioning that science, during the sixteenth and seventeenth centuries, had likewise been inspired by the technical achievements of masons, craftsmen (or engineers, as we would call them today) which became possible thanks to the development of crafts. The use of instruments inspired the study of their laws and functioning, which in turn encouraged the creation of new projects. In fact, the seventeenth century became known as “The Age of Projects”, and some of them were so complicated that they were to remain on paper forever (Forbes & Dijksterhuis 1963: 306-307). Overall, by the late 1600s urban Europe had a dramatically different aspect after the Reformation and the Renaissance (Beal 2004; Park & Daston 2006). Education was visibly becoming independent from the Church, and the Church had lost much of its control over science. Latin became largely (though by no means entirely) replaced by the vernacular, while scientific practice had moved away from monasteries to universities, and was spreading out from universities to academies and independent (or quasi-independent) societies of learning. The learned European society was becoming enlightened and sought to transmit this enlightenment further on. In England, the foundation of the Royal Society of London in 1660 encouraged scientific communication, both direct and in the shape of public correspondence, which later gave way to the emergence of the research article as a new scientific genre (Valle 1996; Atkinson 1999; Moessner 2009). The following paragraphs attempt to describe the crucial role that the Royal Society would play in the formation of a new English scientific register along the early decades of the Enlightenment.

4. The Royal Society: the scientific method and the experimental essay

The seventeenth century, marked by the English Civil War in the 1640s, the execution of Charles I in 1649, and a ten-year period of tyranny under Cromwell’s Commonwealth, is often described as “one of the most turbulent in British history” (Banks 2008: 39). The following Restoration of the Monarchy in 1660 may be thus regarded “not so much as an attempt to reimpose the old social order, but as a last chance to restore *any* social order” (Atkinson 1999: 15). It was on that year when, with an earnest desire to cultivate and promote the empirical study of nature in a spirit of temperance, serenity and gravity, a group of learned gentlemen and noblemen decided to lay the foundations of what they had called the “Royal Society of London for Improving Natural Knowledge”. A parallel movement was taking place on the European continent. Already in 1609, the Academie dei Lincei had been created in Rome. Germany opened its Academia Naturae Curiosorum (later Leopoldina) in

1652, while the Académie des Sciences in Paris would be founded eight years later under Louis XIV. Just as the Royal Society, they had a common purpose: the development of scientific research within a scientific community. On the other hand, and in the particular case of the Royal Society, there was also an intention to institutionalise the practice of experimental science outside the universities – which were traditionally places of education, rather than research –, as well as to achieve the status and rights of a public corporation, independent from private patronage (Hunter 1989: 1-2), thus becoming "England's first true professional scientific group" (Montgomery 1996: 84). All in all, academies remained largely a privilege of nobility until well into the eighteenth century, and the Royal Society was no exception. Despite the claims of Thomas Sprat', early historiographer and one of the key fellows of the Society, that "the Soldier, the Tradesman, the Merchant, the Scholar, the Gentleman, the Courtier, the Divine, the Presbyterian, the Papist, the Independent, and those of Orthodox Judgment, have [in the Society] laid aside their names of distinction" (1667: 427), it was still "a society of gentlemen in the fullest sense – run by gentlemen, for genteel purposes, via genteel standards of conduct and communication, as part and parcel of a genteel form of life" (Atkinson 1999: 17), meaning also that they were ruled by a genteel – and therefore polite – code of conduct.

At the beginning of the seventeenth century, Francis Bacon was not only one of the fathers of the scientific method, but also a pioneer in what would soon become a widespread desire to improve the English language in order to make it suitable for the expression of science. In his works *The Advancement of Learning* (1605) and *Novum Organum* (1620), Bacon proclaimed that "in order to progress beyond medieval sophistry, knowledge would require a new type of speech, a plain and unadorned style of writing capable of carrying the truth of the world in as direct a manner as possible" (Montgomery 1996: 74). Bacon died in 1626, but the members of the Royal Society adopted him "as their linguistic messiah" (Montgomery 1996: 75), with John Wilkins and Robert Boyle being two of his most passionate followers. Wilkins's *Essay towards a Real Character and a Philosophical Language* (1668) condemns metaphor and polysemy because they make language vague, rather than precise, preventing it from expressing "the semantic characteristics of nature" (Camiña-Rioboo 2013: 57). A perfect language, instead, would maintain a one-to-one relationship between *words* and *things*, “laid on the superiority of *res* over *verba*” (Gotti 2001: 231). Boyle, in turn, shaped a genre for the transmission of the new science: the experimental essay, adopted – and adapted – from Montaigne. Gotti (1996, 2001) summarises five characteristics that an experimental essay should have, as stated in Boyle’s *Proemial Essay… with Some Considerations Touching Experimental Essays in General* (1661): brevity, lack of assertiveness, perspicuity, simplicity of form, and objectivity.

Bacon’s, Boyle’s and Wilkin’s efforts appear to be intrinsically connected in the Society’s engagement with improving and transmitting scientific research. The following explanation by Camiña-Rioboo (2013: 46) illustrates how this structure was intended to work:

> The ambitious enterprise to reform science and education purported by the members of the Royal Society was founded on three pillars: a) the methodology employed to deal with scientific facts, b) the vehicle to disseminate the results of the experiments performed and the knowledge acquired, and c) the language employed to communicate those experiments and knowledge. The scientific method, the experimental essay and the philosophical (scientific) language represented those pillars, respectively.

All in all, despite their fiery defense of the necessity of a plain and straightforward philosophical language, the abovementioned members of the Royal Society tend, in fact, to disregard their own precepts. Boyle, for instance, resorts sometimes to polysemy and other figurative devices in his use of terminology (Gotti 1996: 39; 2001: 232). And Sprat, while advocating for “a close, naked, natural way of Speaking”, chooses the metaphor of “putting in Execution the only Remedy” (as was indeed the use at those times!) to save language from “all the Amplifications, Digressions, and Swellings of Style” (1667: xx). Many of their writings, therefore, give the impression that “[a]ttacking and accusing metaphorical language for its wantonness, its antipathy to truth, became one of the only real rhetorical standards observed by the Society… this was often done in florid fashion, with magnificent self-negation” (Montgomery 1996: 86). However, the cause of verbosity lay sometimes in the fear of excessive brevity, which would in turn compromise clarity of exposition, and Boyle, being conscious of this problem, apologised for the occasional inconsistencies in his own writings (Gotti 2001: 227). In any case, this linguistic habit would change some years later with Newton. Compared to Boyle, who "still has his linguistic feet firmly planted in the Renaissance", Newton is considered to "ha[ve] a flatter linguistic footprint and is much easier to follow" (Montgomery 1996: 97). His sober style, untouched by flourishes and for the most part unaffected by emotion, has been found to be considerably influenced by Latin (Banks 2008: 59-63). This may, to a point, explain the frequency of passive constructions in his *Opticks* (1704), a linguistic construction usually associated with a detached, impersonal character. All in all, Newton's language is still very far from losing a personal voice of a "public and private self" (Montgomery 1999: 98), telling a story.

Apart from the somewhat polemical clarity and brevity, persuasion is agreed to be another characteristic of scientific prose, used as a rhetorical strategy to convince the members of the scientific community through writing (Bazerman 1988; Allen *et al*. 1994). This strategy would be widely exploited since 1665, when the Royal Society started the publication of its journal, the *Philosophical Transactions*, edited by the Society's first Secretary Henry Oldenburg and intended primarily for the communication of the latest discoveries among the members of the Society. First and foremost, experiments could not be trusted unless witnessed and carefully reported. In the beginning, scientists would engage in providing a detailed account of their experiments, ideally in a laboratory accessible to another scientist whenever that was an option, but also in the solitude of their houses; their stories would serve as a guarantee of the credibility of their scientific reports (Shapin 1988: 376). With time, however, personal discoveries gave way to a shift in focus towards "more universal grounds: the proof of a claim transcending the particulars of an investigation" (Bazerman 1988: 78). This gradual change can be seen in Atkinson's (1999: 76-81) rhetorical analysis of the *Transactions* since 1665 to 1975. He noticed that the rhetorical focus in the early publications of the Society's journal was mostly author-centered; that is, the emphasis was on the author relating his experiment to the reader (presumably, other members of the Society) in the first person. Atkinson (1999: 77) observes four characteristics that often accompany this author-centered approach, all of them being also typical of Boyle's writing (Gotti 2001: 227-237): *witnessing* (or giving the names of renowned fellow-scientists who were present during the experiment or observation; also, sometimes, inviting the reader to be a "virtual witness" in the absence of a real one); *indexes of modesty or humility* (i.e. lack of assertiveness, or hedging); a *tendency towards miscellaneity*, with common digressions and sometimes unconnected observations; and *elaborate politeness* explicitly addressed to fellow-members of the Society.[14] By the end of the eighteenth century, the author-centered approach started shifting towards an object-centered one (Atkinson 1999: 78-80), highlighting the observed, rather than the observer, the experiment, rather than the experimenter, processes, rather than actions. And it is only at that time that experimental articles would start offering claims and experimental proofs as a solid basis for scientific trust (Bazerman 1988: 78). This was accomplished through a progressive increase in the use of the passive voice, which largely substituted the use of the first person.

Looking back at the first half of the late Modern period, we may thus agree with Montgomery that, "[i]n the mid- to late 17th century… even at its Senecan extreme, what we witness is an attitude towards language, not its achievement" (1996: 98). Change in scientific language would arrive at a relatively slow pace, and not necessarily at a steady one, as it would parallel change in scientific thought and in science itself. In the next section, I will try to provide a glimpse of the three scientific disciplines included in the present study as they looked in the context of change outlined above and in the subsequent decades, until they became what we know today as astronomy, philosophy and life sciences.

5. From natural philosophy to science: a matter of evolution

At some point in the history of science, there was an "astronomy before the telescope", as there also was a "biology before the microscope" (Pledge 1959: 20, 31). Such a time extended from the origins of the study of nature – whether animal or human, terrestrial or celestial – to the early decades of the seventeenth century, when advances in the study of optics and experiments with different types of lenses made it possible for Galileo and Janssen to challenge the limitations of the human eye. Before that moment, the natural world, and in particular, that of the heavens, was an inexhaustible source of fascination, and sometimes, even still in the Renaissance, of superstition. As was mentioned earlier in this chapter, astronomy and astrology were

[14] It may be interesting to note that this last characteristic, as well as the acknowledgement of other scientists as authorities, would gradually evolve into the conventional forms of referencing in the modern research article, as explained in Allen *et al.* (1994: 279-310).

closely connected with one another, and despite the ‘strictly’ scientific – often mathematical – advances that had been made in the study of astronomy since the times of Eratosthenes, Aristarchus of Samos, or Ptolemy, any new phenomenon seen in the skies, be it an eclipse or a supernova, were regarded as a premonition or some other kind of divine sign.

Nonetheless, astronomy is considered to be “the first exact science to make decisive advances” (Pledge 1959: 31). In the sixteenth century, long-distance voyages to the colonies demanded a more technical application of astronomy, and navigational astronomers and cartographers made their own contributions in the form of maps, globes of the earth, floating magnets and compasses. And it was at that time, as well as in the seventeenth and eighteenth centuries, when astronomy, in an intimate connection with mechanics, arrived at its golden age, having its first “revival” with Copernicus, and later with Tycho Brahe, Kepler and Galileo (Bryant 1907; Dreyer 1953). Looking at this constellation of geniuses, it indeed appears extraordinary – how Copernicus provided the theory of a heliocentric universe; how Tycho gave Kepler access to his invaluable observational data (although he would not fully accept the Copernican system, but, instead, combined the latter and the Ptolemaic (geocentric) system, creating a new one: the Tychonic system); and how Kepler used Tycho’s observations to discover his four laws of planetary motion, while Galileo perfected the telescope, allowing a much closer and clearer view of the heavenly bodies and revolutionising astronomy for centuries to come. The culmination of this golden period came with Sir Isaac Newton, a member of the Royal Society, whose *Philosophiae Naturalis Principia Matematica* (1687) – made known thanks to another English astronomer, Edmond Halley – opened to the world his laws of motion and of universal gravitation. Many eighteenth-century astronomers, including James Ferguson, John Lacy, or William Nicholson (all of them included in the corpus of Astronomy analysed in this study), would refer to Newton and his works again and again.

Newton, though, was not an astronomer strictly speaking; neither was he a mathematician, nor a physicist. He was a *natural philosopher* – a person who, in the seventeenth and eighteenth centuries, “had a breadth of comprehension, perceived analogies and other irregularities, derived rules that explain phenomena, and predicted the future”, combining “accuracy of observations”, “precision of judgment”, and "speculative curiosity" (McCormmach 2004: 17). Astronomy, just as philosophy and mathematics, was one of the branches of natural philosophy (*philosophia naturalis*), which was regarded in the late Renaissance and early Enlightenment as a "superscience" (Camiña-Rioboo 2013: 39) that encompassed nearly everything else. Park & Daston (2006: 4) provide a detailed explanation of this fairly complex model:

> Natural philosophy examined change of all kinds, organic and physical, including motion, as well as the principles that produced the phenomena of the heavens (cosmology), the earth's atmosphere (meteorology), and the earth itself (such as minerals, plants, and animals, including human beings). The two topics of plants and animals fell generally under the study of the soul, understood as that which distinguishes living from nonliving beings (…). [It] also addressed questions that would now be seen as metaphysical, such as the nature of space and time and the relation of God to creation (…).

It becomes hence apparent that, unlike in the present-day classification of sciences, where astronomy and life sciences would be labelled as "Natural Sciences", and philosophy as "Humanities", the three aforementioned branches of knowledge would all go largely under the same roof in late Modern context, having in some cases very questionable boundaries that would allow them to be distinguished from one another. Still, a transition towards more clearly defined sciences was in progress, albeit a very gradual one. Already at the beginning of the seventeenth century, life sciences had changed radically with the invention of the microscope. Leeuwenhoek's discovery of bacteria and spermatozoa in 1674, revolutionary in itself, was also crucial on the long term with a view to bacteriology and reproduction studies, which would only start by the end of the 1700s (Magner 2002: 172). On the other hand, botany and anatomy were essentially descriptive disciplines until well into the nineteenth century, closely connected with art, although the case of anatomy was peculiar in that, ever since the Renaissance, it had an experimental side – that of dissection. Plants, which in the Middle Ages had been studied mainly for what concerned their poisonous or medicinal properties, were now systematically classified by German botanists in an obsessive scholastic tradition to "find a hole for every species" (Pledge 1959: 22) until the eighteenth century, when the increasing trips to the colonies allowed the collection and study of foreign species, and taxonomies elaborated by Linnaeus and

Jussieu became the basis of modern botany (Dampier 1929; Stafleu 1971; Serafini 1993).

In the transit from the 1700s to the 1800s, Cuvier's experimentation in paleontology through his comparison of contemporary animal species with fossils raised the idea of geological catastrophism, which suggested that, periodically, natural disasters such as floods would cause certain species to become extinct and new species to emerge. This vision was strongly opposed by Lamarck, who, along with Erasmus Darwin, opened the doors into the theory of evolution, even though their belief in God the Creator remained unchanged (Serafini 1993; Magner 2002; Ruse 2008). Lamarck also believed that animals could inherit acquired characteristics, a controversial theory that would be dismissed for good only in the early twentieth century. Apart from that, he was the first to use the term *biology* in its modern sense, encompassing "the study of all that pertained to living bodies, their organization, development, special organs, and vital movements" (Magner 2002: 299). But it would not be until the second half of the nineteenth century, with the publication of Charles Darwin's *Origin of the Species* and with Gregor Mendel's discovery of the laws of heredity, that biology started to be regarded not only as the science that studies and classifies living beings, but also as one that unveils certain transcendental truths about the existence and development of life on Earth.

All these developments in life sciences "went hand in hand with developments in philosophy" (Serafini 1993: 140). Biology concerned all living beings, including man, and the study of man raised questions of a more metaphysical nature, such as the immortality of the soul, the existence of God, free will, or causality. In the seventeenth century, Descartes tried to demonstrate rationally that, despite the fact that the laws of nature play a crucial role in human beings, our idea of the existence of God lies in his soul, which is no other than reason and which cannot be acquired by the senses, but, rather, is innate. However, what was also in vogue at that time was the so-called "mechanistic philosophy", developed by d'Holbach and Buffon, according to which man and everything in this world is governed entirely by the laws of physics and chemistry, and, consequently, everything develops according to them only, whether it is the life cycle of a tree or the behaviour of man. In what concerns theology, these physical and mathematical laws that govern the world were essentially regarded as the supreme manifestation of the power of God – a rational explanation for a perfect creation. In the eighteenth century, empiricism replaced rationalism with philosophers such as Locke or Hume (both in the *CC*), who defended the five senses, rather than reason or an immortal soul, as the only true source of human knowledge (Serafini 1993: 140-144). Hume, in particular, fervently denied causality by stating that all causal relationships are mere chimeras and what we actually see is a series of events following one another. By the end of the century, empiricism and rationalism were masterfully combined in Immanuel Kant's *Critique of Pure Reason*, and his philosophical thought would dominate most of the 1800s.

Still, it is documented that "the late eighteenth and early nineteenth centuries were very much transitional periods" (Serafini 1993: 184), where the scientific coexisted with the mystical to some extent. For instance, some biologists, such as Caspar Wolff, believed in the *vis vital* (or vital force governing the human behaviour), and yet, at the same time, conducted studies in embryology. A similar case was that of Spallazani, another physiologist who studied reproductive biology and went as far as to successfully perform artificial insemination in animals, while also holding "ovist" ideas (according to which the sperm is secondary to the egg in reproduction). Insofar as creation is concerned, most philosophers and scientists in general regarded the universe and its laws as an outcome of the "wisdom and providence of a Divine Being" (Serafini 1993: 135), a view that persisted until the end of the nineteenth century, when the modern theory of evolution suggested alternative ways for this world to have originated. And still even some early twentieth-century scientists, such as Einstein, believed in the idea of a "natural order", be it called God or something else, which devised the physical laws that rule our Universe.

In what concerns the study of the Universe, it is Newton's laws of motion and universal gravitation that defined modern astronomy until Einstein's theory of relativity came to light; thus, the task of eighteenth-century astronomers was to demonstrate that the motions of the planets strictly obeyed to Newton's physical laws, and to calculate and predict planetary orbits (Crowe 1994; Dewhirst & Hoskin 1999: 219; Bello 2014: 60). Several important celestial phenomena were identified by eighteenth-century astronomers, such as the motion of the stars, observed by Edmund Halley in 1718; the nutation of the Earth's axis, discovered by James Bradley in 1738; the planet Uranus, spotted by William Herschel in 1781, or his striking detection – two years later – that the sun, being a star, also moved. Telescopes were constantly improved and perfected and new observatories were built. All in all, the unavailability of the most powerful telescopes meant that research in stellar astronomy would not be conducted until the second half of the nineteenth century, when telescopes fitted with prisms permitted to analyse starlight (Crowe 1994: 146-148). On the other hand, it was in the eighteenth century that astronomy finally separated from astrology, when almanacs and ephemerides gave way to research articles, treatises and textbooks on astronomy. By the mid-nineteenth century, astronomers were so numerous in England alone that it was necessary to found an Astronomical Society in London in 1820, and an Astronomical Journal in America in 1849 (Dewhirst & Hoskin 1999: 221), assuring thus the creation of an ample yet professional scientific community, specialised in astronomy.

This professionalisation of astronomy paralleled an increasing professionalisation and institutionalisation of all sciences in the nineteenth century (Bello 2014: 22). Considering that the Royal Society started as a medium of scientific research for mostly learned amateurs (and geniuses), by the mid 1800s it was felt that English science lacked the academic preparation that was available on the continent, where scientific research was conducted mostly in universities. At that time, "Oxford and Cambridge, unrivalled as places of liberal education, were not yet awake to the continental spirit of research" (Dampier 1929: 289). These two eminent institutions were reformed in the second half of the nineteenth century, when, simultaneously, science became more international thanks to increasing travel facilities. Dampier (1929: 290) explains that this was one of the factors that prompted the "segregation of science into sciences" in the second half of the nineteenth century, which also caused philosophy and science to detach from one another. The proliferation of universities and the availability of laboratories encouraged students to focus on the experimental method, leaving no time for expanding their knowledge into other areas: "The growth of knowledge went in so fast that no man could keep track of it all" (Dampier 1929: 290). Advances in biological research went beyond Mendel and Darwin. By the end of the 1800s, processes of cell division such as mitosis and meiosis were explained by Strasburger, Fleming and Beneden, while Boveri and Sutton were the first to notice that chromosomes carry genetic information. Physicists such as Ampère, Faraday, Oersted and Volta experimented with electricity, discovering that it can generate a magnetic field. Later on, electricity and magnetism were unified by Maxwell, who introduced the term ‘electromagnetic force’. As summarised in Serafini (1993: 207),

> [i]t can be no exaggeration to suggest that the Victorian era was one of the most dramatically productive of all previous periods in science, eclipsed perhaps only by the appearance of quantum theory and the theory of relativity in the opening decades in the twentieth century.

Einstein’s theory of special relativity (1905), precisely, may be considered the cutting point between late Modern and contemporary science, in that it displaced Newton’s theory of mechanics which had been regarded as the scientific basis for the previous two hundred years. By that time, the language of science was established, emerging as an international standard. The next section intends to offer a brief outline of the evolution of the English scientific register along the period described in the above paragraphs.

6. Scientific English in the eighteenth and nineteenth centuries

As had been outlined earlier, the eighteenth century was a period of prescriptivism and authoritarianism at all levels, and science and its language were no exception to that. As explained in Camiña-Rioboo (2013: 63), “men of science replaced Latin with English for pragmatic purposes, but they were still anchored in the old-fashioned linguistic model of perfection”. Outside the world of science also, language change was deemed a corruption, as also were any deviations from a standard, urged to be established in order to preserve linguistic “purity”. With that in view, and in order to clean the language from imperfections and vulgarisms, a great deal of grammars and dictionaries were published in the eighteenth century (Hickey 2010: 3; Camiña-Rioboo 2013: 64). Likewise, books in which “hard words”, or words of a Latin or Greek origin, were explained, as they were considered too difficult for English speakers (Hickey 2010: 4), came out. On the other hand, pronunciation was also standardised (Beal 2010), and those coming to the cities (especially, to London) from the country would do their best to polish their speech in order to become assimilated with the ‘native’ population. Even so, in the eighteenth century, education was still a privilege reserved to the higher classes, although efforts were made to spread literacy and etiquette among the less privileged, including women and children (Ticken-Boon van Ostade 2010).

In what strictly concerns the language of science, it both benefitted from the linguistic prescriptivism of that moment and enriched the English language with a learned register, contributing a significant amount of technical terms (Camiña-Rioboo 2013: 66-68). This search for an order and urge to establish a precise scientific terminology was also present in the continent. Linnaeus and Lavoisier established standard nomenclatural systems in botany and chemistry, respectively. Taxonomies are, indeed, a valuable piece of linguistic evidence of the necessity of abstraction that characterises the eighteenth century. Still, scientific English of that time "reveals that true centralization of technical style was lacking" (Montgomery 1996: 102), in that the Royal Society never reached the level of "obedience" to the standards prescribed, contrarily to the case with the Académie des Sciences in France. All in all, this prescriptive attitude continued until the mid-nineteenth century, when the spread of popular education prompted an increasing interest in grammar texts, while the circulation of political pamphlets, newspapers and non-canonical literature urged a reaction from the cultured classes, which sought to "cleanse" the English language from the influence of "low" or "vulgar" expressions. At that time, literacy started growing among the population, while science began to be treated both as a profession and as communication or instruction material for different types of audience – fellow scientists, general readership, students, young ladies.

On the other hand, a "romantic rebellion" took place in the 1800s in an effort to fight the domination of this linguistic purity and revive medieval and Renaissance words and expressions. This entailed a cultivated, literary language not only in poetry fiction, or memoirs, but also in scientific literature, which, in some cases, "seems to have gained an elegance above and beyond anything it had gone before" (Montgomery 1996: 103). This means that, even at that time, the scientific discourse of some authors was still far from fitting into a standard shape consisting of a limited number of linguistic patterns, but showed instead a mastery of literary rhetoric, comparable to that of Sir Walter Scott, Wordsworth or Lord Byron. Nonetheless, in certain scientific fields such as chemistry there appears to be a tendency to use language as a bare tool for the report of experiments, the scientist being hidden behind passivized sentences and nominalisations (Montgomery 1996; Atkinson 1999). These two trends, one towards literary elegance and sophistication and another towards simplicity, would coexist for a while before a definite turn in the second direction took place by the end of the century as the level of abstraction and technicality increased (Bello 2014: 31). On the other hand, the professionalisation of science and the consolidation of scientific communities in every single field of knowledge entailed a simplification of forms: it was no longer necessary to write long introductory speeches, as it had been not long before, when science was still a gentlemanly activity. Researchers were mostly equals in what concerned their dedication and, to a large extent, their expertise in the matters discussed, which allowed for a more rigorous and condensed rhetoric. Likewise, persuasion became a subtler strategy: trustworthiness was perceived through the statement of facts and the availability of scientific data, rather than in the reliability of scientist's account.

Finally, the internationalisation of science could take place thanks to technological improvements that allowed a modernisation of transport and communication systems, which, in turn, favoured a more fluent exchange of research among the scientists around Europe and in America. As explained in Bazerman (1988: 138), "as the character of scientific communication changed from the late seventeenth century to today, publication became essential to research and integrated the working scientists into a communications network". This not only made possible the development of a shared scientific knowledge which, up to a point, would be disseminated across the Western world regardless of borders, but also brought forth the need for certain conventions on an international level, including the adoption of a well-demarcated scientific register, and, later on, of English (and, to some extent, also of French) as an international scientific language. This last factor would also play an important role in the simplification and standardisation of the English scientific discourse.

This chapter aimed to briefly describe the emergence and development of Western science and of English as a scientific register along the late Modern period, comprehended between the seventeenth and twentieth centuries, in an attempt to provide some socio-historical context for the present study. In the next chapter, the corpus material used in this research – eighteenth and nineteenth-century texts on astronomy, philosophy and life sciences – will be described.

Chapter 3

The Corpus

1. Introduction

As has been outlined in Chapter 1, variation studies are usually conducted by analysing relatively large collections of machine-readable texts, or corpora. The latter can be contemporary or historical, providing information on different stages in the evolution of a given language, and often permitting its description through both a synchronic and a diachronic lens. Corpora can also be general, covering a variety of registers in a language, or specialised, containing texts from a particular register. They may also be monolingual or multilingual (McEnery & Hardie 2011). Although particularly useful in the study of linguistic variation and change, corpus linguistics is in fact a methodology that can be applied to any area of linguistics, allowing extensive descriptive works in semantics, syntax and lexicology (Quirk et al. 1988; Biber 1999), as well as morphology (e.g. Baayen & Renouf 1996) and pragmatics (e.g. Aijmer 2008).

The use of observational data in the study of language started as early as in the nineteenth century, while larger samples of recordings or careful transcriptions of utterances began to be collected in the 1920s, often considered as a solid basis of research in the areas of language acquisition and language pedagogy (McEnery &

Wilson 1996: 3-4). In the 1950s, some linguists considered corpora as the only reliable source of linguistic evidence, “the primary explicandum of linguistics” (Leech 1991: 8). This attitude, which defends an entirely empirical approach to the study of language, was heavily criticised by Chomsky (1957, 1962, 1988), who maintained that language study could not be based on natural performance (which would always be poor in one or another aspect), but on competence, which would be based on intuition, being rational, and therefore flawless. Chomsky’s criticism triggered a controversy between the “empiricist” and “rationalist” positions towards the description of language, making corpus linguistics rather unpopular during the late 1950s.

McEnery & Wilson (1996: 5-12) identify several reasons for Chomsky’s negative evaluation of corpora. One reason is that a corpus is finite, while language is infinite. One could try to count and identify all the sentences in a language, but such an attempt, however painstaking, would be doomed to failure. Therefore, a corpus, however large and “real-life”, cannot be the sole basis for the description of a language. Moreover, a corpus is also necessarily skewed in that it will, or will not contain certain linguistic features, depending on whether those linguistic features represent or not the real world and some of its conditionings. Thus, a “true” sentence would be more likely to occur in a corpus, rather than a “false” one; a “polite” sentence would sooner be found, rather than an “impolite” one, and so on (Chomsky 1962: 159, in Leech 1991: 8, and McEnery & Wilson 1996: 10). Finally, Chomsky considers that a corpus is not sufficient (and not really necessary) for determining whether a language construct is grammatical or not. Rather, the intuition of a native speaker of a language should be more than enough to identify an utterance as ungrammatical.

Corpora are nowadays identified with computers. It has to be borne in mind that the first large machine-readable corpus was the *Brown Corpus*, which began to be compiled in the 1960s and was completed by the late 1970s (see Chapter 1). Until the second half of the twentieth century, data processing had to be carried out manually. McEnery & Wilson (1996: 12-13) cite Abercrombie (1965), who criticised the study of language through corpora as a “pseudo-procedure”, implying that it is physically impossible to analyse a several-million-word corpus by hand. This was another reason why corpus linguistics was frowned upon until the last third of the twentieth century, when the compilation of large computerised corpora such as *Brown*, *LOB* or *London-*

Lund were finally finished, and when the emergent corpus-processing software, or concordance programs, made those large corpora available to users. This, according to Leech (1991) was the period when corpus linguistics as we know it today started to develop. Although corpus methodology was used with caution in the 1960s and 1970s due to the disadvantages highlighted by Chomsky, some of his criticisms appear to have actually helped to establish the boundaries of the term *corpus* as it is used today (McEnery & Wilson 1996: 14, 16).

Among the many definitions found in the literature, some appear to be looser, or more "flexible" than others. For instance, early definitions of corpus by Leech include "a source of systematically retrievable data and…a testbed for linguistic hypotheses" (1991: 9), focusing on its serviceability in research, or, otherwise, "a helluva lot of texts, stored in a computer" (1992: 106), referring to its actual shape and content. Kilgarriff & Grefenstette (2003: 334) also emphasise the practical side of corpora, stating that "[a] corpus is a collection of texts when considered as an object of language or literary study" (in Saldanha 2009: 2). From this definition, any collection of texts can be considered a corpus when used in linguistic or literary research. On the other end of the scale, Sinclair (1994: 14) defines corpus as "a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language", while McEnery & Wilson (1996: 24) characterise it as "[a] finite-sized body of machine-readable texts sampled in order to be maximally representative of the language variety under consideration". Biber et al. (1998: 4), in turn, list four main characteristics of a "corpus approach": it is empirical, in that it analyses the actual patterns of language use in natural texts; the collection of natural texts has to be large and principled; computers are used in its analysis; and this analysis combines quantitative and qualitative techniques. All in all, despite the flexibility of the term, there seems to be a general consensus among corpus linguists (Biber 1993; McEnery & Wilson 1996; Martí & Castellón 2000; Tognini-Bonelly 2001; Baker 2002; Bowker & Pearson 2002; McEnery 2003; Taavitsainen 2005) that a corpus, rather than being a mere collection of electronic texts, has to be compiled according to certain criteria. These include size, representativeness, balance and time-span, as well as register selection and research scope.

The following sections describe the corpus used as the data source in our study. Section 2 focuses on the design of the corpus and some of its compilation principles, while Sections 2.1-2.3 present each of the three subcorpora that have been included in the present study. Section 3, on the other hand, deals with ways of processing the corpus in its unannotated and annotated versions by means of two different concordance programs, the *Coruña Corpus Tool* (Parapar & Moskowich 2007) and the *CQPWeb* (Hardie 2012), respectively.

2. The *Coruña Corpus of English Scientific Writing*

As we have seen earlier, specialised diachronic corpora permit to look at the evolution of a particular register along a given period of time. Such is the case of the *Coruña Corpus of English Scientific Writing* (hereafter *Coruña Corpus*), an electronic corpus which is currently being compiled by the members of the Research Group in Multidimensional Studies in English (MuStE)[15] at the University of A Coruña (Spain) and which provided the materials for this piece of research. The *Coruña Corpus* is part of the on-going research project *Coruña Corpus: A Collection of Samples for the Historical Study of English Scientific Writing*, conceived for the diachronic study of variation and change in late Modern scientific English. The corpus covers a period of two hundred years (1700-1900) and consists, to date, of four subcorpora which contain samples from texts on Astronomy, Philosophy, Life Sciences and History (while other subcorpora, dealing with texts on Chemistry, Mathematics, Physics and Linguistics, are currently under compilation). Each subcorpus has a total of twenty texts per century[16], and therefore two texts per decade, while each text sample is ca. 10,000 words long, excluding figures, graphs, tables, formulae and punctuation marks, as well as quotations containing text reproduced literally from other sources, or fragments written in languages other than English. On the other hand, the corpus contains samples of both male and female authors who were educated in different English-speaking regions (England, Scotland, Ireland, the US and Canada) and who used different genres (e.g. treatises, essays, textbooks…) in their writings. However, in order to avoid stylistic idiosyncrasies, only one work per author was selected. The reasons behind the principles followed in the compilation of the *Coruña Corpus*, including representativeness and balance, corpus size and time span, as well as the selection of authors for the different scientific disciplines, are dealt with extensively in Moskowich & Crespo (2007), Moskowich & Parapar (2008), Lareo (2009), and

[15] www.udc.es/grupos/muste

[16] Except for the Astronomy subcorpus (*CETA*), which has 21 texts in each century; see below.

Crespo & Moskowich (2010), as well as in Camiña-Rioboo (2012, 2013), Crespo (2012a, 2012b), Bello (2014) and Moskowich (2016a).

For the present study, we have selected three of the four subcorpora listed above: the *Corpus of English Texts on Astronomy* (*CETA*), the *Corpus of English Philosophy Texts* (*CEPhiT*) and the *Corpus of English Life Sciences Texts* (*CELiST*). Although these subcorpora of the *Coruña Corpus* have been originally encoded in the eXtended Mark-Up Language (XML) format and published without part-of-speech or semantic annotation, recently an annotated version of the three of them has been created (see Section 3). According to the UNESCO (1988) classification of sciences, which is used as a reference in the selection of the scientific disciplines for the *Coruña Corpus*, Philosophy belongs to the category of Social Sciences and Humanities, while Astronomy and Life Sciences fall into the category of Natural Sciences. The latter, however, is not a homogeneous field – as, of course, neither is the former –, and the different scientific disciplines classified under the heading of Natural Sciences present some differences in their use of the scientific method. For instance, Chemistry and Experimental Physics may be considered essentially empirical sciences, while others, such as Mathematics or Theoretical Physics, have a more abstract and speculative nature. Life Sciences, in turn, seem often to be more observational than experimental (e.g. Botany), although in some cases their study must be carried out in a laboratory, as may be the case with Molecular Biology and its related subjects. Although in the previous chapter we have seen that eighteenth- and nineteenth-century natural philosophy was not yet fully developed into what is nowadays considered as science, the choice of using text samples classified as Astronomy, Philosophy and Life Sciences in this study was made in an attempt to find variation among two “hard” disciplines (Astronomy and Life Sciences) and a “soft” one (Philosophy) in a time when each discipline was still evolving into what they are nowadays. On the one hand, this will permit to carry out a diachronic analysis of three scientific subregisters along the eighteenth and nineteenth centuries (e.g. the “Astronomy subregister” vs. the “Philosophy subregister”). On the other hand, given that the texts in each subcorpus are classified under specific genre labels, text samples can likewise be considered as belonging to different subregisters (e.g. the “treatise subregister”, the “essay subregister”, and so on).

The total corpus, which has a total of 1,213,841 words, consists of 122 text samples, each subcorpus being composed of 40 to 42 text samples, evenly distributed across two centuries (18th century and 19th century), as shown in Table 3.1:

Table 3.1

Subcorpus		Number of texts	Number of words
CETA	18th century	21	208,079
(Astronomy)	19th century	21	201,830
(Total *CETA*)		(42)	(409,909)
CEPhiT	18th century	20	200,022
(Philosophy)	19th century	20	201,107
(Total *CEPhiT*)		(40)	(401,129)
CELiST	18th century	20	200,649
(Life Sciences)	19th century	20	202,154
(Total *CELiST*)		(40)	(402,803)
Total corpus		**122**	**1,213,841**

Distribution of text samples and words across three subcorpora (*CETA*, *CEPhiT* and *CELiST*)

The corpus samples are classified according to eight genres, largely based on Görlach's (2004) classification of text-types[17]: Treatise, Essay, Textbook, Letter, Lecture, Article, Dialogue, and Other. The category "Other" has been used in the *Coruña Corpus* with texts that "present miscellaneous features that make them ineligible for any of the previous genres" (Camiña-Rioboo 2013: 181).[18] In this case, the sample classified as "Other" is a dictionary, whereas in other subcorpora not included in this study, such as the *Corpus of Historical English Texts* (*CHET*), this label may encompass different categories, such as travelogue. Moskowich (2011: 182) explains that this genre classification of samples is based on epistemological features and social factors (i.e. the scientific community surrounding the authors, their place and level of education, the epistemological level of their intended audience,

[17] As we saw in Chapter 1, Görlach (2004) considers *text-type* as a formal, rhetorical category, using it as a synonym of *genre*, rather than as a class of texts that are similar in their linguistic features, as defined by Biber (1988, 1989).

[18] Recently, however, the compilers of the *CC* decided not to use the label "Other" for such categories but, rather, to use the name of the genre in question.

etc.), rather than on linguistic features exclusively. Moreover, the author's statement of purpose (normally, in the prefatory material) is often taken into account when considering the genre of the work (Crespo 2012b, 2016). This entails the careful reading of the texts before selecting the samples, which fulfills Biber & Conrad's (2009) requirement that full texts should be used in order to be characterised from a genre perspective.[19]

Tables 3.2 shows the distribution of genres in the corpus:

Table 3.2

Genre	**Number of texts**	**Number of words**
Treatise	61	609,158
Textbook	20	206,307
Essay	14	142,561
Lecture	12	120,373
Article	7	53,857
Letter	5	51,550
Dialogue	2	19,991
Other (dictionary)	1	10,044
Total corpus	**122**	**1,213,841**

Distribution of genres in the corpus

If we want to look at the genres as they appear in the eighteenth- and nineteenth-century parts of the corpus, Table 3.3 and Figure 3.1 (next page) present their distribution as follows:

[19] Nevertheless, it should be borne in mind that this study uses beta versions of the Philosophy (*CEPhiT*) and Life Sciences (*CELiST*) subcorpora, and that the distribution of genres may change in the future (i.e. in their final versions).

Table 3.3

Genre	**18th century**		**19th century**	
	Texts	**Words**	**Texts**	**Words**
Treatise	34	338,138	27	271,020
Textbook	12	124,200	8	82,107
Essay	9	92,231	5	50,330
Lecture	1	9,939	11	110,434
Article	1	4,240	6	49,617
Letter	2	20,051	3	31,499
Dialogue	1	9,907	1	10,084
Other (dictionary)	1	10,044	-	-
Total corpus	**61**	**608,750**	**61**	**605,091**

Distribution of genres across two centuries

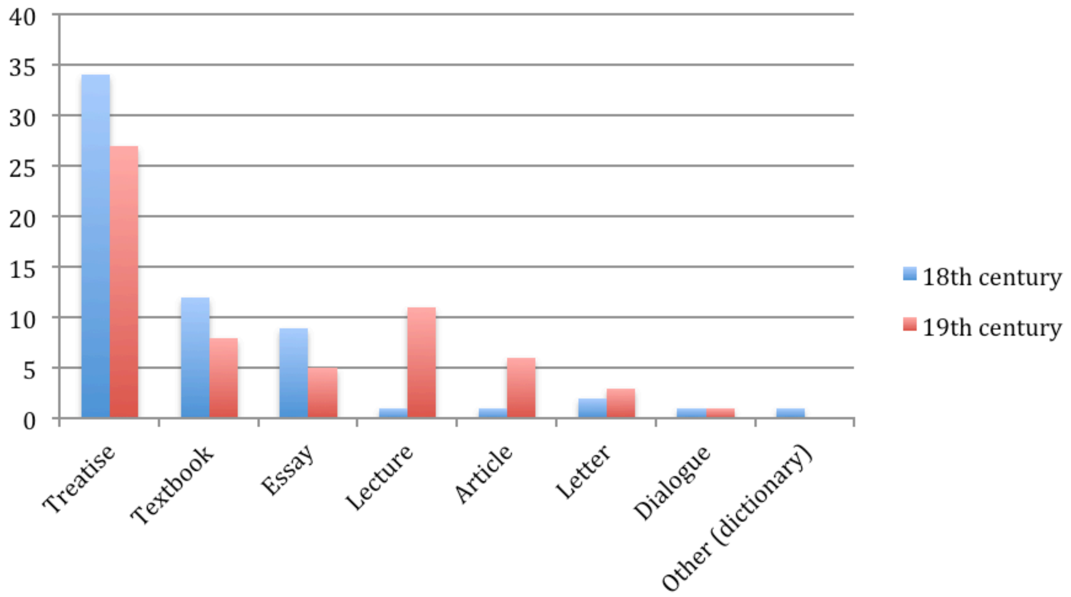


Figure 3.1
Distribution of genres (number of texts) across two centuries

Unlike the distribution of scientific disciplines, the distribution of genres in the *Coruña Corpus* is not uniform. In the next sections we will see how a wider availability of one or another genre in the corpus appears to depend on a particular period (Crespo 2012a: 26-27) and on a particular scientific discipline, suggesting that different genres may be, to some extent, representative of different branches of knowledge (Camiña-Rioboo 2013: 189). As can be seen in Table 3.2, Treatise accounts for half of the texts in our corpus.[20] This genre, first recorded in the late fourteenth century with Chaucer's *A Treatise on the Astrolabe*, has gradually incorporated an empiricist component, changing from "a book of writing which treats of some particular subject" to a book "containing a formal or methodological discussion or exposition of the principles of the subject" (*OED*) (Crespo 2012a: 30). The "scientific" nature of its very definition might justify this preference for Treatises both in the eighteenth and nineteenth century (see Table 3.3). The second most favoured genre is Textbook, which, conversely, is essentially didactic. As shall be seen in Section 2.1, Textbooks are most frequent in eighteenth-century Astronomy, often aimed at the instruction of less learned classes of society, including women. Essay, in turn, comes from the French *éssai,* meaning "first draft" and generally referring to a relatively short composition containing some thoughts or reflections on a subject (*OED*; also Görlach 2004: 88), and appears mainly in Philosophy texts.

On the other hand, Table 3.3 also shows that some genres appear more frequently in the nineteenth century, as is the case with Lectures and Articles. In the case of the former, Lectures were written to be spoken (i.e. to be later read aloud in front of the students; see Gómez-Guinovart & Pérez Guerra 2000). The concentration of science in the Universities and Academies in the nineteenth century might be one of the reasons why this genre started to be used more often. This is also the time when the scientific journal article gradually gains shape as a genre for the publication of research within the scientific community (Bazerman 1988; Moessner 2009). According to Crespo (2016: 30), articles, treatises, letters and essays were normally used to exchange scientific knowledge among peers during the late Modern period, while textbooks, lecture and dialogues were intended for learners, including women. All in all, as we will see in Chapter 6, the epistolary genre may have likewise been used both for scholarly interaction and for didactic purposes. Scientific dialogues, in turn, "introduced a fiction to teach about facts" (Lightman 1997: 192), serving the popularisation of science.

[20] See footnote [4].

Regarding the sex of the author, the distribution of texts written by male and female authors in the corpus is presented in Table 3.4:

Table 3.4

Subcorpus		**Male authors**		**Female authors**	
		Texts	**Words**	**Texts**	**Words**
CETA	18th century	20	197,816	1	10,263
(Astronomy)	19th century	20	191,300	1	10,530
(Total *CETA*)		(40)	(389,116)	(2)	(20,793)
CEPhiT	18th century	17	169,828	3	30,194
(Philosophy)	19th century	20	201,107	-	-
(Total *CEPhiT*)		(37)	(370,935)	(3)	(30,194)
CELiST	18th century	19	190,604	1	10,045
(Life Sciences)	19th century	14	139,441	6	62,713
(Total *CELiST*)		(33)	(330,045)	(7)	(72,758)
Total corpus		**110**	**1,090,096**	**12**	**123,745**

Distribution of male and female authors in the corpus

As we can see, only twelve of the one hundred and twenty-two scientists present in our corpus are women, and most of them belong to the Life Sciences (*CELiST*) subcorpus. This very low percentage (10%) is very likely to be representative of the reality of the time. Indeed, there had been women in Modern Europe who had devoted their life to science, but this was not an easy task to accomplish. With the exception of two Italian ladies[21] who obtained a university degree in the late seventeenth and mid-eighteenth century, academia was essentially all-male institutions and would not admit women until the mid-to-late 1800s. Instead, women scientists would normally work at home, often assisting their husbands, brothers, or fathers (Schiebinger 1989, 2003), who may have even published some of their wives', sisters' or daughters' works under their names (Herrero-López 2007: 75, in Camiña-Rioboo 2013: 184). The nineteenth-century conception of the family as a strictly domestic institution secured the precarious and almost "illegal" position of women scientists, who were usually frowned upon whenever they attempted to participate in the academic world (Abir-Am & Outram 1987). Botany was one of the few exceptions, being somehow

[21] This was the case of Elena Cornaro Piscopia (1646-1684) and Professor Laura Bassi (1711-1778), the latter being even granted a university professorship.

considered a "milder" science (presumably since it involved little more than the observation and description of plants and flowers) and, therefore, more acceptable as a study field for ladies (Slack 1987; Shteir 1987, 2008). This may explain why we have more women in our Life Sciences subcorpus, many of whose texts focus on a variety of botanical species. Conversely, Astronomy (*CETA*) and Philosophy (*CEPhiT*) have only two and three samples from female authors, respectively. This may be due to the fact that ladies who would walk outdoors to contemplate the sky at night were not held in a very high esteem (Herrero-López 2007: 82), as neither must have been those who indulged into existential speculations.

A closer look at each of the three subcorpora is offered in the following sections.

2.1. The Corpus of English Texts on Astronomy (CETA)

Table 3.5 (next pages) lists the authors of the text samples in chronological order (preceded by the ID given to each text in the corpus for this particular study), the date of publication of their work, the title of the part sampled and its extent in words. As has been already described in previous studies (Camiña-Rioboo 2013, Bello 2014), although most of the samples in *CETA* contain around 10,000 words, some texts appear to surpass that limit (with James Hodgson's (1749) and Matthew Stewart's (1761) samples in the eighteenth-century part), while other are visibly smaller (i.e. Alexander Wilson (1773), John Lacy (1779) in the eighteenth century, as well as George Darwin (1880) and Charles Young (1880) in the nineteenth).

As explained in Camiña-Rioboo (2013: 178), the very large samples "are special cases containing many numbers, variables and formulae embedded within sentences, which cannot be deleted without affecting the understanding of the text itself", so that it was necessary to extend the final number of words "until a suitable number of appropriate material analysable under linguistic perspectives could reach the boundaries of 10,000 words". Two smaller samples in each century (articles included *in toto*), in turn, make up for one long text, and in these cases there are three, instead of two, texts that complete the expected number of words in a decade (ca. 20,000). All in all, it must be observed that *CETA* was the first subcorpus of the *Coruña Corpus* to be compiled (see Moskowich & Crespo 2012). *CEPhiT* and *CELiST*, by contrast (see Sections 2.2 and 2.3), have all their samples measuring around 10,000 words.

Table 3.5

Text ID	Author	Year	Title	Extent (words)
astr1	Henry Curson	1702	*The theory of sciences illustrated*	10247
astr2	Robert Morden	1702	*An Introduction to astronomy*	10154
astr3	William Whiston	1715	*Astronomical Lectures*	9939
astr4	John Harris	1719	*Aſtronomical Dialogues Between a Gentleman and a Lady*	9907
astr5	George Gordon	1726	*An introduction to geography, astronomy, and dialling*	10437
astr6	Isaac Watts	1726	*The knowledge of the heavens and the earth made easy*	10407
astr7	Samuel Fuller	1732	*Practical astronomy, in the description and use of both globes, orrery and telescopes*	10232
astr8	Jasper Charlton	1735	*The Ladies Astronomy and Chronology*	10358
astr9	Roger Long	1742	*Astronomy, in five Books*	10474
astr10	James Hodgson	1749	*The theory of Jupiter's satellites*	11106
astr11	John Hill	1754	*Urania*	10044
astr12	James Ferguson	1756	*Astronomy explained upon Isaac Newton's*	10519
astr13	Matthew Stewart	1761	*Tracts, physical and mathematical: containing, an explication of several important points in physical astronomy...*	12180
astr14	George Costard	1767	*The history of astronomy*	10315
astr15	Alexander Wilson	1773	*Observation of the Solar Spots*	4240
astr16	George Adams	1777	*A Treatise describing the construction and explaining the use of celestial and terrestrial globes*	10566
astr17	John Lacy	1779	*The universal system: or mechanical cause of all the appearances and movements of the visible heavens*	5908
astr18	William Nicholson	1782	*An introduction to natural philosophy*	10268
astr19	John Bonnycastle	1786	*An introduction to Astronomy in a Series of Letters*	9975
astr20	Samuel Vince	1790	*A treatise on practical astronomy*	10540
astr21	Margaret Bryan	1797	*A compendious system of astronomy in a course of familiar lectures*	10263

Table 3.5 (continued)

Text ID	Author	Year	*Title*	Extent (words)
astr22	Robert Small	1804	*An Account of the Astronomical Discoveries of Kepler*	10435
astr23	John Ewing	1809	*A Plain Elementary and Practical System of Natural Experimental Philosophy; including Astronomy and Chronology*	9985
astr24	David Brewster	1811	*Ferguson's astronomy explained upon Sir Isaac Newton's Principles*	9824
astr25	William Phillips	1817	*Eight familiar lectures on Astronomy*	10130
astr26	John Gummere	1822	*An Elementary Treatise on Astronomy in Two Parts*	10507
astr27	Thomas Luby	1828	*An Introductory Treatise to Physical Astronomy*	10704
astr28	John Herschel	1833	*A Treatise on Astronomy*	10224
astr29	Landon Garland	1838	*Address on the Utility of Astronomy*	9608
astr30	Denison Olmsted	1841	*Letters on Astronomy, addressed to a Lady*	8742
astr31	Duncan Bradford	1845	*The Wonders of the Heavens*	10268
astr32	William Bartlett	1855	*Elements of natural philosophy. IV Spherical Astronomy*	10858
astr33	William Whewell	1858	*The plurality of worlds*	10079
astr34	Ormsby Mitchel	1860	*Popular astronomy*	10183
astr35	Elias Loomis	1868	*A Treatise on Astronomy*	10323
astr36	William Chauvenet	1871	*A manual of spherical and practical astronomy*	9895
astr37	Joel Steele	1874	*Fourteen weeks in descriptive astronomy*	9979
astr38	George Darwin	1880	*On the Secular Changes in the Elements of the Orbit of a Satellite revolving about a Tidally Distorted Planet*	5181
astr39	Charles Young	1880	*Recent Progress in Solar Astronomy*	6454
astr40	James Croll	1889	*Stellar Evolution and Its Relation to Geological Time*	9390
astr41	Agnes Clerke	1893	*A Popular History of Astronomy during the Nineteenth Century*	10530
astr42	Percival Lowell	1895	*Mars: III Canals*	8531

Authors included in *CETA*

Figure 3.2 below shows the distribution of genres in *CETA*:

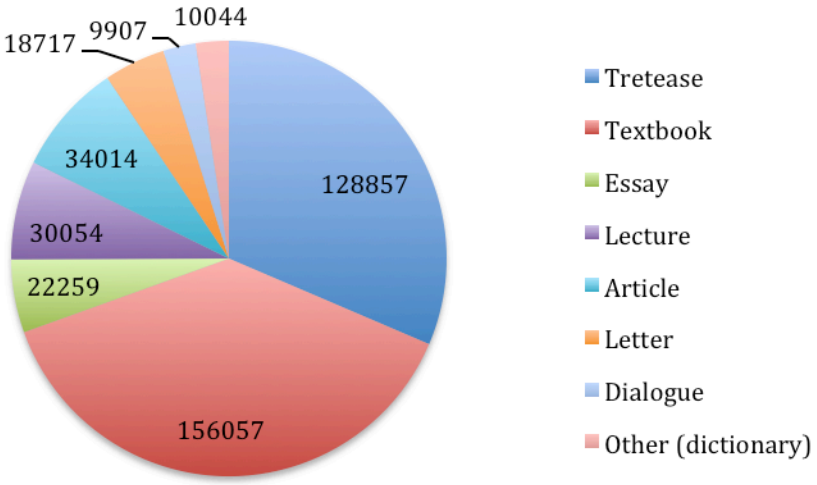


Figure 3.2
Genres in *CETA*

Clearly, Treatise and Textbook appear to be the most common genres in this corpus. As was observed by Bello (2013: 158), *CETA* presents a more or less equal distribution of more formal, or specialised texts (Treatise, Article and Essay) and texts aimed at a learner, less specialised audience (Textbook, Letter, Dialogue and Dictionary), giving thus a balanced representation of different epistemological levels. In fact, Moskowich (2012: 42) refers to Görlach (2004: 1), stating that this distribution of genres "broadly reflects production at the time". She explains how the high number of certain genres is directly connected with the scientific discipline: in this case, the knowledge of Astronomy used to be communicated in accademic settings, which required more formal genres such as treatises and essays. On the other hand, one of the aims of the Modern period was the dissemination of knowledge, which was accomplished by using didactic genres such as textbooks and dialogues which were more accessible to less learned people, as well as letters, which were intended for a presumably less formal way to exchange knowledge. Specialised dictionaries, in turn, were compiled to make a field of knowledge more intelligible to the learner.

A more accurate account of the distribution of genres in *CETA* can be obtained through Figure 3.3, which presents their occurrence in the eighteenth and nineteenth centuries separately:

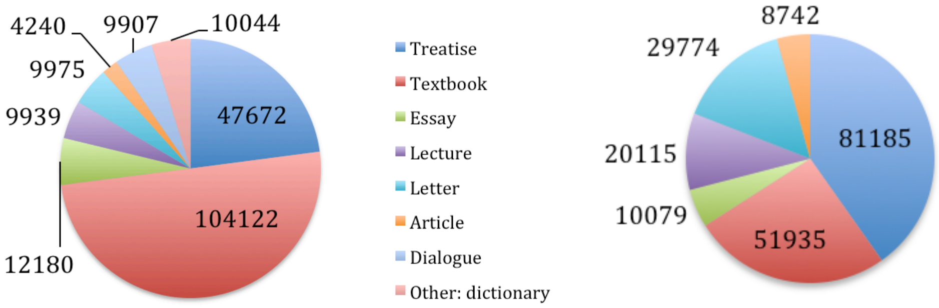


Figure 3.3
Genres in 18th-century (left) and 19th-century (right) *CETA*

Here, the picture is somewhat different. In the eighteenth century, more than half of the samples in *CETA* are textbooks, whereas only a quarter correspond to treatises. Clearly, instruction must have been the primary goal of astronomers at that time. In the nineteenth century, conversely, the Treatise genre occupies more than a third of the samples, while there is also ca. 30% that corresponds to another formal genre, the Article. This seems to point to the fact that Astronomy must have consolidated as a science in the nineteenth century and demanded a more specialised vehicle of transmission between scholars, who, in turn, became more and more rigorous in the genres they used. This might explain why nineteenth-century *CETA* presents less variety of genres than its eighteenth-century counterpart. However, it still contains a notable proportion of instructive genres, but now Lecture appears to have gained some importance, sharing the didactic space with textbooks, and suggesting that the teaching of Astronomy also requires a more formal frame in the nineteenth century.

On the other hand, it is also worth noting that *CETA* has a sample of a work written by a female astronomer in each century. It has been already mentioned that women scientists were not particularly encouraged in the Modern period. Still, teaching at home – and, particularly, young women's education – was a common occupation for ladies, and this was the case of Margaret Bryan, who taught astronomy and natural philosophy to young girls and wrote the *Compendious System of*

Astronomy (1797), a sample of which is included in eighteenth-century *CETA*. The other woman in this corpus is Agnes Mary Clerke, an Irish historian of astronomy and astrophysics whose formation as a scientist would start at her family home in Cork (and, later, Dublin) and would be completed in different libraries in Italy. Although she brought to the light several works, it was *A Popular History of Astronomy in the Nineteenth Century* (1893) that the compilers of *CETA* selected for the nineteenth-century part of the corpus.

The information on authors and their texts is contained in the corpus metadata, accessible with the *Coruña Corpus Tool* (see Section 3). Like every other subcorpus of the *Coruña Corpus*, *CETA* contains a complete set of metadata that provides some interesting details such as the authors' age, level of studies, and place of education, allowing the use of an ample set of extra-linguistic variables in the analysis of the corpus. However, this piece of research only includes the variables time, scientific discipline (as subregister I) and genre (as subregister II). For this reason, all the extra-information previously mentioned will not be included in the description of the corpus as it is not relevant for the present study.

2.2. The Corpus of English Philosophy Texts (CEPhiT)

This was the second subcorpus of the *Coruña Corpus* to be compiled (Moskowich et al. 2016). Although it has been recently published, a beta version of *CEPhiT* has been used here. Like in the previous section, in the following table (3.6) authors are enumerated in chronological order, preceded by the sample ID, and followed by the year of publication, the work title and its length in words. As has been mentioned in Chapter 2, and as we can see in this table, the topics covered by eighteenth- and nineteenth-century Philosophy are extremely varied, ranging from morality and religion to "more mundane subjects such as marriage, feminism and politics" (Camiña-Rioboo 2013: 186). Indeed, the three women philosophers included in *CEPhiT* – Mary Astell, Catharine Macaulay and Mary Wollstonecraft – are all reputed defenders of women's rights, something that was extremely courageous in the patriarchal context of eighteenth-century Europe.

Table 3.6

Text ID	Author	Year	Title	Extent (words)
phil1	Mary Astell	1700	*Some Reflections upon Marriage*	10,079
phil2	George Cheyne	1705	*Philosophical Principles of Natural Religion*	10,060
phil3	John Dunton	1710	*Athenianism*	10,063
phil4	Anthony Collins	1717	*A Philosophical Inquiry Concerning Human Liberty*	9,984
phil5	Robert Greene	1727	*The principles of the philosophy of the expansive and contractive forces*	9,998
phil6	Robert Kirkpatrick	1730	*The Golden Rule of Divine Philosophy*	10,046
phil7	John Balguy	1733	*The Law of Truth*	10,042
phil8	Joseph Butler	1736	*The analogy of religion, natural and revealed, to the constitution and course of nature*	10,050
phil9	George Turnbull	1740	*The Principles of Moral Philosophy*	9,498
phil10	David Hume	1748	*Philosophical Essays Concerning Human Understanding*	10,019
phil11	Henry Bolingbroke	1754	*The Philosophical Works of the late Right Honorable Henry St. John*	9,997
phil12	Francis Hutcheson	1755	*A system of moral philosophy*	9,821
phil13	Thomas Reid	1764	*An Inquiry into the Human Mind, on the Principles of Common Sense*	10,032
phil14	Adam Ferguson	1769	*Institutes of Moral Philosophy*	10,065
phil15	Edmund Burke	1770	*Thoughts on the cause of the present discontents*	10,003
phil16	George Campbell	1776	*The philosophy of rhetoric*	10,008
phil17	Catharine Macaulay	1783	*A treatise on the immutability of moral truth*	10,060
phil18	William Smellie	1790	*The Philosophy of natural history*	9,993
phil19	Mary Wollstonecraft	1792	*Vindication of the Rights of Woman*	10,053
phil20	Alexander Crombie	1793	*An essay on philosophical necessity*	10,026
phil21	Thomas Belsham	1801	*Elements of the Philosophy of The Mind, and of Moral Philosophy*	10,089
phil22	Dugald Stewart	1810	*Philosophical Essays*	10,017
phil23	Richard Kirwan	1811	*Metaphysical Essays*	10,062
phil24	Thomas Brown	1820	*Lectures on the Philosophy of the Human Mind*	10,055

Table 3.6 (continued)

Text ID	Author	Year	Title	Extent (words)
phil25	Sir Richard Phillips	1824	*Two Dialogues between an Oxford Tutor and a Disciple of the Common-Sense Philoſophy*	10,077
phil26	Sir James Mackintosh	1830	*Dissertation on the progress of ethical philosophy, chiefly during the seventeenth and eighteenth centuries*	10,078
phil27	Renn Hampden	1835	*A course of lectures introductory to the study of moral philosophy*	10,019
phil28	Rev. Baden Powell	1838	*The connexion of natural and divine truth: or, the study of the inductive philosophy, considered as subservient to theology*	10,089
phil29	John Mill	1845	*An Examination of Sir William Hamilton's Philosophy*	9,666
phil30	George Combe	1846	*Moral Philosophy*	9,995
phil31	William Lyall	1855	*Intelect, the Emotions, and the Moral Nature*	10,070
phil32	Henry Slack	1860	*The philosophy of progress of human affairs*	9,942
phil33	T. Collyns Simon	1862	*On the Nature and Elements of the External World*	10,065
phil34	Henry Mansel	1866	*The Philosophy of the Conditioned*	10,053
phil35	Thomas Woodward	1874	*A Treatise on the Nature of Man*	10,029
phil36	Arthur Balfour	1879	*A Defence of Philosophic Doubt*	10,048
phil37	Andrew Seth	1885	*Scottish Philosophy*	9,975
phil38	John Mackenzie	1890	*An Introduction to Social Philosophy*	10,028
phil39	James Bonar	1893	*Philosophy and Political Economy*	10,116
phil40	Shadworth Hodgson	1898	*The Metaphysic of Experience*	10,046

Authors included in *CEPhiT*

The nineteenth-century part of the corpus, however, contains no samples written by female authors, suggesting that at that time it must have been even more difficult for women to make their philosophical reflections public. Regarding the genres, Philosophy appears to show quite a different picture than Astronomy (see Figure 3.4, next page):

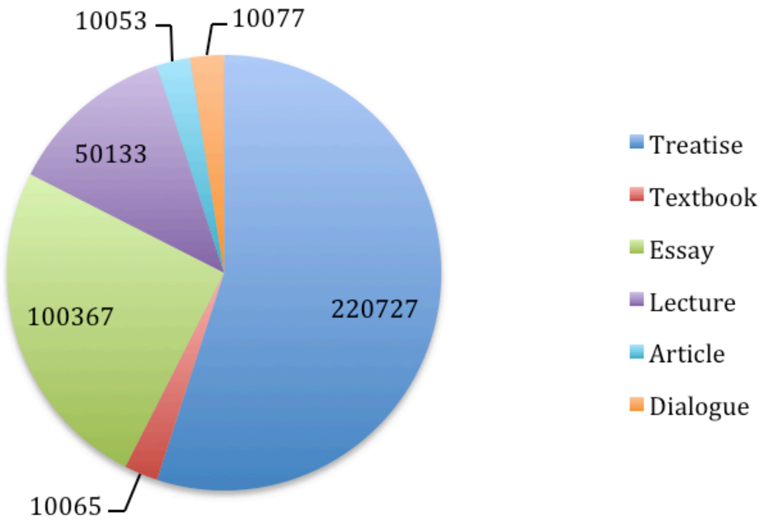


Figure 3.4
Genres in *CEPhiT*

On the one hand, the variety of the genres is, overall, smaller than the one in *CETA*, and the genres Letter and Other (dictionary) are not present in *CEPhiT*. On the other hand, more than half of the text samples in this corpus are treatises, while about a quarter belong to the Essay genre, indicating a clear preference for more formal ways of communication for this branch of knowledge (Moskowich 2011; Crespo 2016). This also seems to be confirmed by the prevalence of lectures (five texts) over textbooks (only one text), suggesting that the didactic end of Philosophy likewise required a more specialised frame of discourse, which was intended to be read in front of an audience, usually in a university. Moreover, the relatively large number of essays suggests that Philosophy admitted a somewhat looser kind of compositions, usually reserved for the author's reflections on a particular topic, which in turn could be more or less abstract in character. For instance, Mary Astell's (1700) essay *Some Reflections Upon Marriage* focus on her own views of certain problems, such as the lack of a proper education among women and their consideration in society as objects for men's entertainment. In contrast, David Hume's *Philosophical Essay Concerning Human Understanding* (1748) deals with the human mind and human nature, by discussing different kinds of philosophy and the formation of ideas and scepticism.

Figure 3.5 shows the distribution of genres in eighteenth- and nineteenth-century *CEPhiT*:

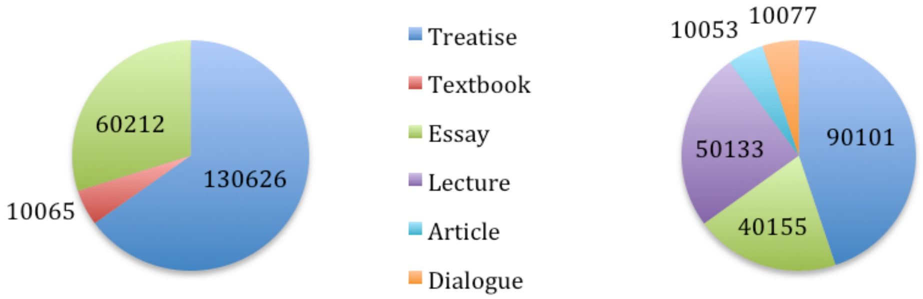


Figure 3.5
Genres in 18th-century (left) and 19th-century (right) *CEPhiT*

Interestingly, despite the variety of topics, eighteenth-century Philosophy does not appear to be particularly diversified in what concerns genres, containing only treatises (largely), essays and one textbook. This, once more, seems to indicate a connection between scientific discipline and genre, suggesting in this case that Philosophy tended to be realised through particular formal genres – Treatise and Essay – in the eighteenth century. In the nineteenth century, however, there is a rise of the Lecture genre, as well as a minor presence of two other, Article and Dialogue (whereas Treatise shrank to less than half of the texts in the subcorpus). While lectures and articles are two different specialised ways to convey the philosophical thought (one through a carefully prepared speech, another through a well-reasoned piece of writing, intended not to be read aloud, but to be published in a prestigious scientific publication), dialogues were still occasionally used in order to instruct ordinary people, normally by presenting a Question-Answer pattern where the latter would be the voice of Reason (see Prince 1996).

2.3. The Corpus of English Life Sciences Texts (CELiST)

This corpus was the third to be compiled and currently exists both in a beta version (which has been used for the present study) and in a definite version, which is soon to be released with minor changes (see Lareo & Esteve-Ramos 2007; Lareo & Moskowich 2009; Lareo 2011b). Once more, a list of the authors in chronological order is given in Table 3.7, accompanied by the sample ID, and also followed by the publication date of the work, its title and its word count. As has been outlined earlier, the compilers of *CELiST* used an inclusive perspective and resorted to a wide range of topics under the label Life Sciences. Thus, some of our texts deal with the animal kingdom (e.g. William Gibson, Thomas Boreman, James Dodd, Thomas Pennant, Alexander Wilson, Edward Jenner, George Dalyell, Alpheus Packard), including horses, herrings, birds, as well as rare and quasi-mythological species. Some texts also study insects (Priscilla Wakefield) and butterflies in particular (Alpheus Packard). Others, on the other hand, focus on flowers, plants, and fungi (Elizabeth Blackwell, William Withering, James Bolton, Maria Jacson, Almira Lincoln, Anne Pratt, Phebe Lankester). Some other texts also deal with different parts of the human body, covering the scientific sub-field of anatomy (James Douglas), while some others like Charles Darwin, or Thomas Huxley, speculate on the origins of the species in the mid-nineteenth century. Finally, by the end of the nineteenth century, some studies, such as Arthur Marshall, focus on embryology. Our Life Sciences corpus, thus, contains studies of zoology, ornithology, entomology, botany and anatomy, as well as research in embryology and on the theory of the evolution.

Table 3.7

Text ID	Author	Year	Title	Extent (words)
life1	James Douglas	1707	*Myographiæ comparatæ specimen: or, a comparative deſcription of all the muscles in a man and in a quadruped*	10,045
life2	Hans Sloane	1707	*The Natural Hiſtory of Jamaica*	10,038
life3	James Keill	1717	*Essays on Several Parts of the Animal Oeconomy*	9,812
life4	William Gibson	1720	*The Farriers new Guide: Anatomy of a Horse*	9,875
life5	Patrick Blair	1723	*Pharmaco-botanologia*	10,089
life6	Thomas Boreman	1730	*A description of three hundred animals*	10,013
life7	Elizabeth Blackwell	1737	*A Curious Herbal*	10,045
life8	John Brickell	1737	*The Natural History of North-Carolina*	10,103
life9	George Edwards	1743	*A Natural History of Birds*	10,028
life10	Griffith Hughes	1750	*The Natural Hiſtory of BARBADOS*	10,044
life11	James Dodd	1752	*An essay towards a natural history of the herring*	10,019
life12	William Borlase	1758	*The Natural History of Cornwall*	9,997
life13	Thomas Pennant	1766	*The British Zoology*	10,037
life14	Edward Bancroft	1769	*An Essay on the Natural History of Guiana, in South America*	10,074
life15	Oliver Goldsmith	1774	*AN HISTORY OF THE EARTH, AND ANIMATED NATURE*	10,103
life16	William Withering	1776	*A botanical arrangement of all the vegetables, naturally growing in Great Britain*	10,091

Table 3.7 (continued)

Text ID	Author	Year	*Title*	Extent (words)
life17	William Speechly	1786	*A Treatise on the Culture of the Pine Apple and the Management of the Hot-house*	10,017
life18	James Bolton	1789	*An History of Fungusses growing about Halifax*	10,052
life19	Edward Donovan	1794	*Instructions for collecting and preserving various subjects of natural history*	10,013
life20	Sir James Smith	1795	*English Botany*	10,048
life21	Maria Jacson	1804	*Botanical Lectures by a Lady*	10,051
life22	Alexander Wilson	1808	*American Ornithology*	10,081
life23	Priscilla Wakefield	1816	*An introduction to the natural history and classification of insects*	9,805
life24	Sir William Lawrence	1819	*Lectures on Physiology, Zoology, and the Natural History of Man*	10,039
life25	Edward Jenner	1824	*Some observations on the migration of birds*	9,775
life26	John Godman	1828	*American Natural History*	10,028
life27	Almira Lincoln	1832	*Familiar Lectures on Botany*	10,028
life28	Sir William Jardine	1835	*THE NATURALIST'S LIBRARY. MAMMALIA*	10,026
life29	Anne Pratt	1840	*Flowers and their associations*	10,023
life30	Sir John Dalyell	1848	*Rare and remarkable animals of Scotland*	10,010
life31	Elizabeth Agassiz	1859	*A FIRST LESSON IN NATURAL HISTORY*	12,959
life32	Charles Darwin	1859	*On the Origin of Species*	10,091
life33	Thomas Huxley	1863	*On the Origin of Species: or, the Causes of the Phenomena of Organic Nature*	10,059
life34	Herbert Spencer	1867	*The principles of Biology*	10,082
life35	Alexander Macalister	1876	*An Introduction to Animal Morphology*	10,083
life36	Phebe Lankester	1879	*Wild Flowers worth Notice*	10,080
life37	Francis Balfour	1880	*A treatise on comparative embryology*	10,080
life38	Sir Francis Galton	1889	*Natural Inheritance*	10,062
life39	Arthur Marshall	1893	*Vertebrate Embryology*	10,044
life40	Alpheus Packard	1898	*A text-book of entomology*	10,016

Authors included in *CELiST*

Regarding genres, *CELiST* presents some similarities with the Philosophy corpus, as shown in Figure 3.6:

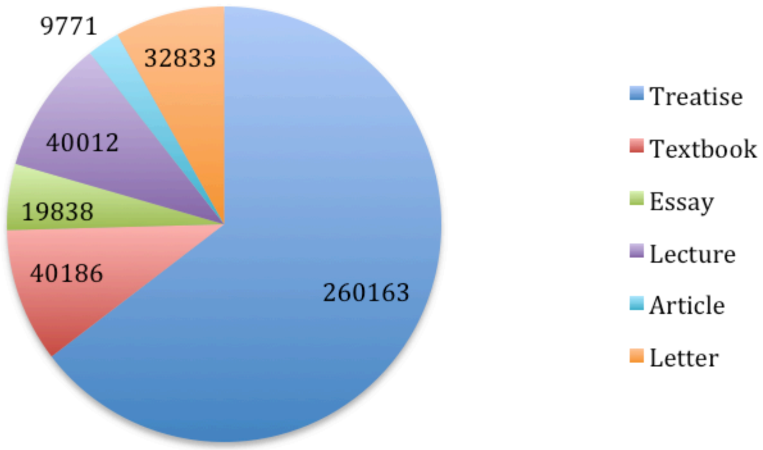


Figure 3.6
Genres in *CELiST*

As it is also the case with *CEPhiT*, Treatise is the genre that, by far, appears to characterise this beta version of *CELiST* the most. Unlike the Philosophy corpus, however, *CELiST* does not contain any dialogues, but, instead, contains a few letters. The didactic genre is also represented, being uniformly distributed between textbooks and lectures. An article and two essays are also included, which again suggests a preference for formal genres, especially in research. However, if we look at Figure 3.7, we shall once again find differences between the two centuries:

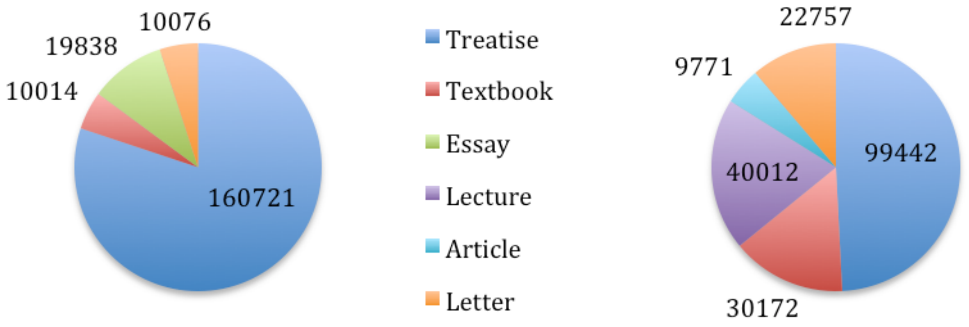


Figure 3.7
Genres in 18th-century (left) and 19th-century (right) *CELiST*

In a more exaggerated way than eighteenth-century Philosophy, Life Sciences in the 1700s is mostly composed of treatises (>75%), suggesting this time again a connection between scientific discipline and genre (even though, as was mentioned earlier, it is probable that the definitive version of *CELiST* will have a different classification and a lesser proportion of treatises). Of the three scientific disciplines contemplated for our study, eighteenth-century Life Sciences appears to have the most formal, specialised frame of discourse, which seems to indicate that the topics listed above were treated with similar rigour and precision (at least, concerning genre conventions). Some diversity is nonetheless present, with two essays, one textbook and one letter. In the nineteenth century, treatises still occupy the main position, but cover now only half of the subcorpus, giving way to the didactic genres (lectures and textbooks), two letters and an article. As was noticed in previous studies (Atkinson 1999; Moskowich & Monaco 2014, 2016), it is likely that letters are used in the nineteenth-century part of *CELiST* as a carefully constructed genre which invites the reader to "observe" something (in this case, nature) from a so-called "personal" (i.e. the writer's) view, often used by women writers. On the other hand, the emergence of the Textbook and the Lecture genres in the nineteenth century seems to suggest that, just like Philosophy, Life Sciences had a more theoretical character in the eighteenth century, whereas the need (or possibility) for using it as a means of instruction may have materialised later in the nineteenth century. The case of Astronomy, by contrast, is completely the opposite in that it was in the eighteenth century that the necessity of teaching seemed to be more apparent, considering the large number of textbooks (see Figures 3.2 and 3.3).

After having looked at the three subcorpora included in this study, we will now proceed to describe the levels of annotation of the *Coruña Corpus*, distinguishing between its semantically and grammatically unannotated and annotated versions.

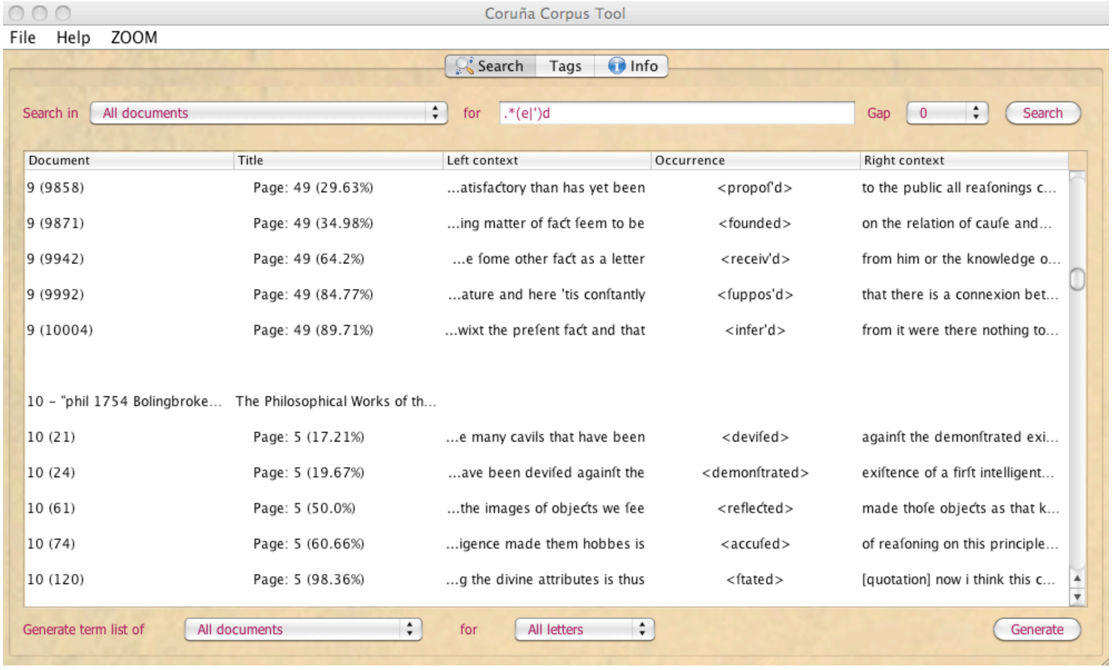
3. Corpus annotation

Leech (1991, 1993) considers automated corpus annotation very important, in that it conveys to the corpus different kinds of interpretative information: either prosodic, morphosyntactic, syntactic, semantic, or pragmatic. He distinguishes between *raw* corpora, which have not been annotated in any way and exist only as plain text, and *annotated* corpora, which contain different kinds of linguistic information. McEnery & Wilson (1996: 34-36) mention several formats of annotation, highlighting the Text

Encoding Initiative (TEI), which is currently often used as a standard for encoding machine-readable texts (see Sperberg-McQueen & Burnard 1994) and adopts the Standard Generalised Markup Language (SGML), which is internationally recognised as a standard. On the other hand, Henry Thompson developed the eXtended (or eXtensive) Mark-up Language (XML) in 1997, which is a subset of SGML but simpler and therefore easier to use. On another level, McEnery & Wilson (1996: 39-45) discuss different types of annotation of corpora, such as textual and extra-textual information, orthographic annotation, linguistic annotation (such as part-of-speech, semantic, or discursive annotation) and phonetic transcription and/or prosodic annotation, which can be used in different cases, depending on the ways in which a corpus will be analysed.

The *Coruña Corpus* is encoded in TEI, and its text samples are annotated in XML. The advantages of XML, as well as the process of annotation, are explained in detail in Camiña & Lareo (2016), and therefore we will not proceed any further to explain this part. However, it is worth noting that, currently, the *Coruña Corpus* has two versions in what concerns part-of-speech and semantic annotation. The original version of the *Coruña Corpus* is not tagged for grammatical nor semantic categories, and it is designed to be processed with the *Coruña Corpus Tool* (hereafter *CCT*), a concordance program created for – although not restricted to – the *Coruña Corpus*, which uses indexes of corpus text files and metadata files to generate word lists and run searches of different complexity. The other recently created version of the *Coruña Corpus* is annotated both for part-of-speech and semantic categories, and can be processed through CQPWeb (see Hardie 2012, 2016), a powerful web-based corpus query processor which stores a large number of corpora. Although the use of the two aforementioned concordance programs will be described in Chapter 4, the following paragraphs will briefly present both the *CCT* and CQPWeb.

The *CCT* was developed by the IRLab in the University of A Coruña, in collaboration with MuStE Research Group (Parapar & Moskowich 2007; Moskowich & Parapar 2008). As the original version of the *Coruña Corpus* contains a number of non-standard characters in order to maintain the corpus samples as close as possible to the original texts, the *CCT* was designed to support those characters whenever it processes a corpus index. This concordance program allows to run searches of simple words, phrases, or sentences, as well as a wide range of wildcard searches, and outputs them in KWIC (Key Word In Context)[22] view, showing the searched element in the centre of the display window and providing a left and right context. Figure 3.8 below is an example of the *CCT* in use:



Search in: All documents for: .*(e|')d Gap: 0

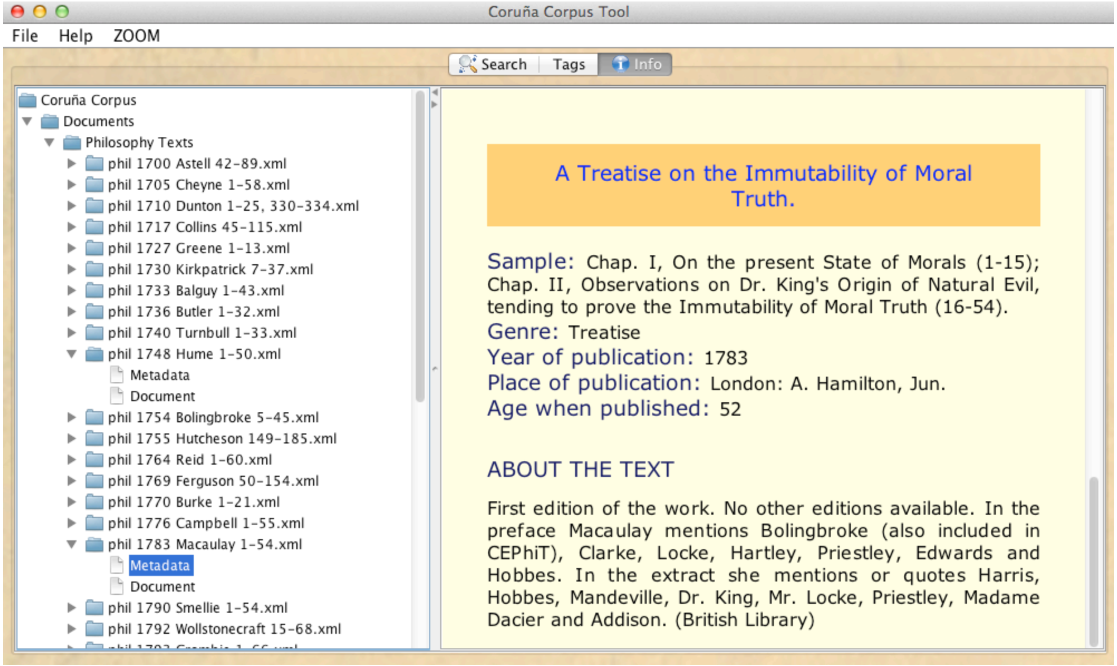
Document	Title	Left context	Occurrence	Right context
9 (9858)	Page: 49 (29.63%)	...atisfactory than has yet been	<propoſ'd>	to the public all reaſonings c...
9 (9871)	Page: 49 (34.98%)	...ing matter of fact ſeem to be	<founded>	on the relation of cauſe and...
9 (9942)	Page: 49 (64.2%)	...e ſome other fact as a letter	<receiv'd>	from him or the knowledge o...
9 (9992)	Page: 49 (84.77%)	...ature and here 'tis conſtantly	<ſuppos'd>	that there is a connexion bet...
9 (10004)	Page: 49 (89.71%)	...wixt the preſent fact and that	<infer'd>	from it were there nothing to...
10 – "phil 1754 Bolingbroke...	The Philosophical Works of th...			
10 (21)	Page: 5 (17.21%)	...e many cavils that have been	<deviſed>	againſt the demonſtrated exi...
10 (24)	Page: 5 (19.67%)	...ave been deviſed againſt the	<demonſtrated>	exiſtence of a firſt intelligent...
10 (61)	Page: 5 (50.0%)	...the images of objects we ſee	<reflected>	made thoſe objects as that k...
10 (74)	Page: 5 (60.66%)	...igence made them hobbes is	<accuſed>	of reaſoning on this principle...
10 (120)	Page: 5 (98.36%)	...g the divine attributes is thus	<ſtated>	[quotation] now i think this c...

Generate term list of: All documents for: All letters

Figure 3.8
Search window of the *CCT*

Apart from the search function, the *CCT* has additional functions such as creating word lists for each file (i.e. text sample) in the corpus, outputting the number of tokens for each type. Also, because every subcorpus of the *Coruña Corpus* is always released with its corresponding metadata, the *CCT* processes metadata indexes at the same time as it processes corpus indexes, allowing to select subsets of the corpus according to different metadata fields, such as "year of publication", "author's place of education", etc. On the other hand, the *CCT* also contains an "Info" section, where each text and its accompanying metadata document can be consulted (see Figures 3.9 and 3.10):

[22] See Luhn (1960).


Coruña Corpus Tool
File
Help
ZOOM
Search
Tags
Info
Coruña Corpus
Documents
Philosophy Texts
phil 1700 Astell 42-89.xml
phil 1705 Cheyne 1-58.xml
phil 1710 Dunton 1-25, 330-334.xml
phil 1717 Collins 45-115.xml
phil 1727 Greene 1-13.xml
phil 1730 Kirkpatrick 7-37.xml
phil 1733 Balguy 1-43.xml
phil 1736 Butler 1-32.xml
phil 1740 Turnbull 1-33.xml
phil 1748 Hume 1-50.xml
Metadata
Document
phil 1754 Bolingbroke 5-45.xml
phil 1755 Hutcheson 149-185.xml
phil 1764 Reid 1-60.xml
phil 1769 Ferguson 50-154.xml
phil 1770 Burke 1-21.xml
phil 1776 Campbell 1-55.xml
phil 1783 Macaulay 1-54.xml
Metadata
Document
phil 1790 Smellie 1-54.xml
phil 1792 Wollstonecraft 15-68.xml

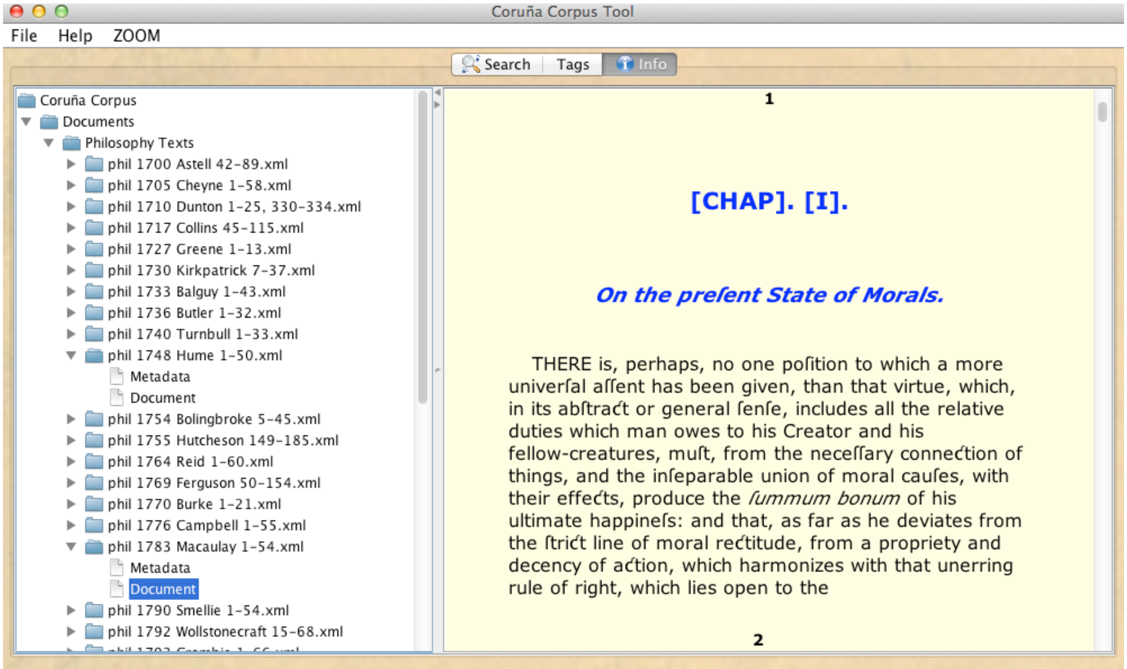
A Treatise on the Immutability of Moral Truth.

Sample: Chap. I, On the present State of Morals (1-15); Chap. II, Observations on Dr. King's Origin of Natural Evil, tending to prove the Immutability of Moral Truth (16-54).
Genre: Treatise
Year of publication: 1783
Place of publication: London: A. Hamilton, Jun.
Age when published: 52

ABOUT THE TEXT

First edition of the work. No other editions available. In the preface Macaulay mentions Bolingbroke (also included in CEPhiT), Clarke, Locke, Hartley, Priestley, Edwards and Hobbes. In the extract she mentions or quotes Harris, Hobbes, Mandeville, Dr. King, Mr. Locke, Priestley, Madame Dacier and Addison. (British Library)

Figure 3.9
Info (metadata) window of the *CCT*


Coruña Corpus Tool
File
Help
ZOOM
Search
Tags
Info
Coruña Corpus
Documents
Philosophy Texts
phil 1700 Astell 42-89.xml
phil 1705 Cheyne 1-58.xml
phil 1710 Dunton 1-25, 330-334.xml
phil 1717 Collins 45-115.xml
phil 1727 Greene 1-13.xml
phil 1730 Kirkpatrick 7-37.xml
phil 1733 Balguy 1-43.xml
phil 1736 Butler 1-32.xml
phil 1740 Turnbull 1-33.xml
phil 1748 Hume 1-50.xml
Metadata
Document
phil 1754 Bolingbroke 5-45.xml
phil 1755 Hutcheson 149-185.xml
phil 1764 Reid 1-60.xml
phil 1769 Ferguson 50-154.xml
phil 1770 Burke 1-21.xml
phil 1776 Campbell 1-55.xml
phil 1783 Macaulay 1-54.xml
Metadata
Document
phil 1790 Smellie 1-54.xml
phil 1792 Wollstonecraft 15-68.xml

1

[CHAP]. [I].

On the preſent State of Morals.

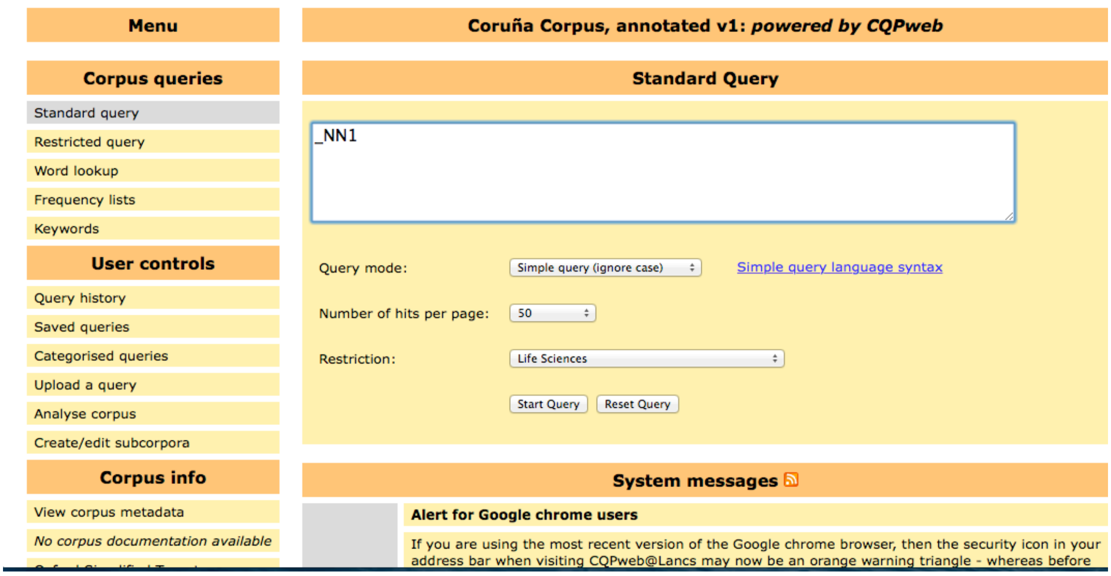
THERE is, perhaps, no one poſition to which a more univerſal aſſent has been given, than that virtue, which, in its abſtract or general ſenſe, includes all the relative duties which man owes to his Creator and his fellow-creatures, muſt, from the neceſſary connection of things, and the inſeparable union of moral cauſes, with their effects, produce the *ſummum bonum* of his ultimate happineſs: and that, as far as he deviates from the ſtrict line of moral rectitude, from a propriety and decency of action, which harmonizes with that unerring rule of right, which lies open to the

2

Figure 3.10
Info (text sample) window of the *CCT*

Extensive research on the *Coruña Corpus* (Moskowich 2011, 2012, 2013; Crespo 2011, 2013; Lareo 2009, 2011; Lareo & Esteve-Ramos 2007; Lareo & Moskowich 2009; Camiña-Rioboo 2010, 2012, 2013; Alonso-Almeida 2012; Bello 2010, 2014; Puente & Monaco 2013, 2016; Puente-Castelo 2014, 2016a, 2016b; Moskowich & Monaco 2014, 2016, among others) has proved the *CCT* to be a very efficient and valuable concordance program, permitting to easily extract information from the corpus either by generating word lists, or by running single-word, multi-word, or wildcard searches. However, because the original version of the *Coruña Corpus* is not annotated for part-of-speech nor for semantic categories, the *CCT* obviously cannot perform part-of-speech searches, nor can it recognise a string of characters as a morphological or syntactic item (e.g. noun, suffix, etc.). Therefore, if the *Coruña Corpus* is part-of-speech and semantically tagged, a different concordance program must be used to process the new version. We thus decided to tag the *Coruña Corpus* with three levels of annotation – part-of-speech tag (CLAWS6 tagset)[23], Lemma and Semantic tag (USAS tagset)[24] – and upload it on CQPWeb[25] (see Hardie 2012, 2016), which not only permits to run single-word, multi-word and wildcard queries, as well as simple part-of-speech-tag and CQP syntax-based queries, but has several additional functions (see Chapter 4).

Figures 3.11 and 3.12 show the standard query window of a part-of-speech search (in this case, a noun) and the output window for this query in CQPweb, respectively:



Coruña Corpus, annotated v1: *powered by CQPweb*

Standard Query

_NN1

Query mode: Simple query (ignore case)

Number of hits per page: 50

Restriction: Life Sciences

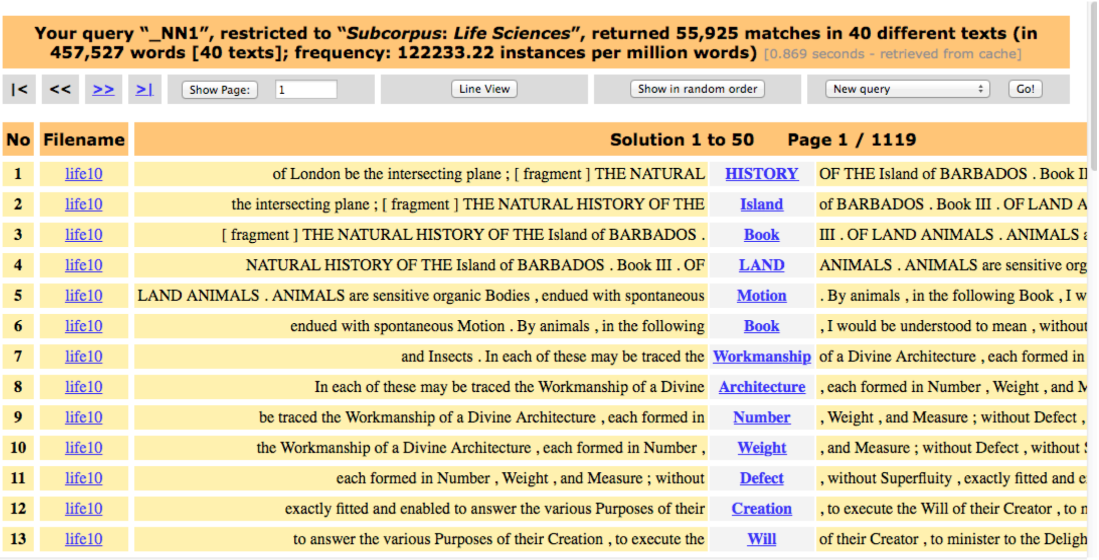
System messages

Alert for Google chrome users

If you are using the most recent version of the Google chrome browser, then the security icon in your address bar when visiting CQPweb@Lancs may now be an orange warning triangle - whereas before

Figure 3.11
Standard query window in CQPWeb

[23] http://ucrel.lancs.ac.uk/claws6tags.html
[24] http://ucrel.lancs.ac.uk/usas/
[25] I am immensely grateful to Dr. Andrew Hardie, who annotated the *Coruña Corpus* for us.



Your query "_NN1", restricted to "*Subcorpus*: *Life Sciences*", returned 55,925 matches in 40 different texts (in 457,527 words [40 texts]; frequency: 122233.22 instances per million words) [0.869 seconds - retrieved from cache]

No	Filename	Solution 1 to 50 Page 1 / 1119		
1	life10	of London be the intersecting plane ; [fragment] THE NATURAL	HISTORY	OF THE Island of BARBADOS . Book I
2	life10	the intersecting plane ; [fragment] THE NATURAL HISTORY OF THE	Island	of BARBADOS . Book III . OF LAND A
3	life10	[fragment] THE NATURAL HISTORY OF THE Island of BARBADOS .	Book	III . OF LAND ANIMALS . ANIMALS a
4	life10	NATURAL HISTORY OF THE Island of BARBADOS . Book III . OF	LAND	ANIMALS . ANIMALS are sensitive org
5	life10	LAND ANIMALS . ANIMALS are sensitive organic Bodies , endued with spontaneous	Motion	. By animals , in the following Book , I w
6	life10	endued with spontaneous Motion . By animals , in the following	Book	, I would be understood to mean , without
7	life10	and Insects . In each of these may be traced the	Workmanship	of a Divine Architecture , each formed in
8	life10	In each of these may be traced the Workmanship of a Divine	Architecture	, each formed in Number , Weight , and M
9	life10	be traced the Workmanship of a Divine Architecture , each formed in	Number	, Weight , and Measure ; without Defect ,
10	life10	the Workmanship of a Divine Architecture , each formed in Number ,	Weight	, and Measure ; without Defect , without S
11	life10	each formed in Number , Weight , and Measure ; without	Defect	, without Superfluity , exactly fitted and e
12	life10	exactly fitted and enabled to answer the various Purposes of their	Creation	, to execute the Will of their Creator , to n
13	life10	to answer the various Purposes of their Creation , to execute the	Will	of their Creator , to minister to the Deligh

Figure 3.12
Query output window in CQPWeb

As we can see, if we look for a part-of-speech-tag in CQPWeb, it returns all the words in the corpus that are annotated for this particular grammatical category. This kind of searches is also possible with lemmas and semantic categories, as will be explained in the Methodology chapter. In order to create this version of the *Coruña Corpus*, original .xml files were converted to .txt files, after which they were annotated for part-of-speech and semantic categories and uploaded on CQPWeb (the trial with the Philosophy subcorpus is extensively explained in Hardie 2016). For this purpose, minimal metadata (such as "genre", "sex of the author", "century", "decade", etc.) were used, so that some metadata searches are also possible with CQPWeb. This allows the *CCT* and CQPWeb to be utilised interchangeably with the *Coruña Corpus*, according to the scope and character of the research in each case.

This research is based on both versions of the *Coruña Corpus*, although in most cases the original version (unannotated for part-of-speech and semantic tags) was used, along with the *CCT*. However, some linguistic categories could not be retrieved with the *CCT* and were searched for with CQPWeb instead. This and the rest of the methodology will be explained in the next chapter.

Chapter 4

Methodology (I): Lexical and grammatical variables

1. Introduction

As we have seen in Chapter 1, register variation can be better attested by taking into account different communicative levels, or different linguistic dimensions. It has been demonstrated that the variation across the different registers in a language (Biber 1988, 1995), or among several subregisters within a register (Biber 2001; Gray 2011), is revealed through the co-occurrence patterns of different linguistic features which are based on some underlying communicative functions, shared by each set of co-occurring features. Using Biber (1988) as its main theoretical and methodological basis (see Chapter 1 Section 3.1), the present study aims to detect and describe the variation among three scientific subregisters, each belonging to a particular scientific discipline (Astronomy, Philosophy and Life Sciences), and each containing smaller subregisters classified on the basis of genre (essays, treatises, textbooks, etc.), along a two-century period (1700-1900). For this purpose, a total of fifty-eight linguistic features have been selected in order to ensure that a sufficient range of communicative functions is included in the analysis.

This chapter describes the first methodological part of our study, which concerns the selection (stage 1) and retrieval (stage 2) of the lexical and grammatical variables that will be later used in the Factor Analysis (see Chapter 5). Section 2 discusses the reasons for the inclusion and exclusion of some of the linguistic features used in Biber's (1988) study, while Section 3 deals with the retrieval of the selected linguistic features from the corpus, offering a description of the concordance programs used for that purpose and giving an account of the query algorithms developed in each case. Finally, Section 4 describes the process of counting the occurrences of those linguistic features in the corpus and the calculation of normalised frequencies.

2. Selection of lexical and grammatical variables

Biber (1988) picked out a total of sixty-seven linguistic features for his analysis on the basis of the communicative functions of certain lexical items and grammatical constructions, specified in previous research, and classed them into sixteen main grammatical and functional subgroups (Biber 1988: 73-75; Biber & Conrad 2001: 17). Initially, our intention was to follow Lee (2000, Chapter 3) in keeping all the features chosen by Biber (1988), as well as his original numeration, both for the subgroups and the linguistic features, so that confusion might be avoided. In the present study, however, only fifty-seven of Biber's (1988) sixty-seven features, plus one additional feature, have been finally selected for the analysis, and only fifteen of the initial sixteen subgroups have been retained (see Section 2.1 for the excluded features). Consequently, in order to avoid creating "ghost" (empty) entries where a feature has been skipped, or new subgroups where an additional feature has been added, we have decided not to retain Biber's (1988) original numeration with a one-to-one correspondence. Instead, we have inserted the additional feature in one of the already existing subgroups, and numbered all the features consecutively based on the order in which Biber (1988) had grouped them.

Table 4.1 lists the fifty-eight linguistic features selected for this study. Most of them follow closely the criteria specified in Biber (1988: 211-245), although one of them has been slightly modified (see discussion below) and marked accordingly with an asterisk (*). The additional linguistic feature, not included in Biber's (1988) study, is marked with two asterisks (**):

Table 4.1

A) Tense and aspect markers	
1. past tense	Any past tense form (e.g. She *was* smart; John *painted* the house)
2. perfect aspect	Any perfect aspect form (e.g. You *haven't* finished yet; *Having* said so, she left.)
3. present tense	Any present tense form (e.g. It usually *rains* in the mornings; The house *stands* alone in the field)
B) Place and time adverbials	
4. place adverbials	*aboard, above, abroad, across, ahead, alongside, around, ashore, astern, away, behind, below, beneath, beside, downhill, downstairs, downstream, east, far, hereabouts, indoors, inland, inshore, inside, locally, near, nearby, north, nowhere, outdoors, outside, overboard, overland, overseas, south, underfoot, underground, underneath, uphill, upstairs, upstream, west* (Quirk et al. 1985: 514ff; Biber 1988: 224)
5. time adverbials	*afterwards, again, earlier, early, eventually, formerly, immediately, initially, instantly, late, lately, later, momentarily, now, nowadays, once, originally, presently, previously, recently, shortly, simultaneously, soon, subsequently, today, tomorrow, tonight, yesterday* (Quirk et al. 1985: 526ff; Biber 1988: 224)
C) Pronouns and pro-verbs	
6. first person pronouns	*I, me, we, us, my, our, myself, ourselves* (plus contracted forms) (Biber 1988: 225)
7. second person pronouns	*you, your, yourself, yourselves* (plus contracted forms) (Biber 1988: 225)
8. third person pronouns (excluding *it*)	*she, he, they, her, him, them, his, their, himself, herself, themselves* (plus contracted forms) (Biber 1988: 225)
9. pronoun *it*	*it*
10. demonstrative pronouns	*this, that, these, those* (e.g. *This* is ridiculous; *Those* are my neighbours)
11. indefinite pronouns	*anybody, anyone, anything, everybody, everyone, everything, nobody, none, nothing, nowhere, somebody, someone, something* (Quirk et al. 1985: 376ff; Biber 1988: 226)
12. Pro-verb *do*	e.g. *the cat did it* (Biber 1988: 226)
D) Questions	
13. *questions	All direct questions (e.g. *What does it mean? Is she coming?*)
E) Nominal forms	
14. nominalisations	All nouns ending in *–tion*, *-ment*, *-ness*, or *–ity*
15. total other nouns	All nouns, except those included in n. 13 and those ending in *–ing*
F) Passives	
16. agentless passives	e.g. Some changes *were considered* necessary
17. *by* passives	e.g. This building *was designed by* a famous architect
G) Stative forms	
18. *be* as main verb	e.g. This *is* Mr. Johnson; The house *is* big
19. existential *there*	e.g. *There is* some butter in the fridge

Table 4.1 (continued)

H) Subordination features	
20. *that* verb complements	e.g. I said *that he went* (Biber 1988: 231)
21. *that* adjective complements	e.g. I'm glad *that you like it* (Biber 1988: 231)
22. **WH clauses in subject position	e.g. *What he told me* was true
23. WH clauses in object position	e.g. I believed *what he told me* (Biber 1988: 231)
24. *to* infinitives	All infinitive forms of verbs preceded by *to*
25. detached past participial clauses with adverbial function	e.g. *Built in a single week,* the house would stand for fifty years (Biber 1988: 233)
26. past participial WHIZ deletion relatives	e.g. *the solution produced by this process* (Biber 1988: 233)
27. present participial WHIZ deletion relatives	e.g. *the event causing this decline is…* (Biber 1988: 233)
28. *that* relativiser in subject function	e.g. *the dog that bit me* (Biber 1988: 124)
29. WH relativiser in subject function	e.g. *the man who likes popcorn* (Biber 1988: 125)
30. WH relativiser in object function	e.g. *the man who Sally likes* (Biber 1988: 125)
31. pied-piping relative clauses	e.g. *the manner in which he was told* (Biber 1988: 125)
32. sentence relatives	e.g. *Bob likes fried mangoes, which is the most disgusting thing I've ever heard of* (Biber 1988: 125)
33. causative adverbial subordinators	*because*
34. concessive adverbial subordinators	*although*, *though*
35. conditional adverbial subordinators	*if*, *unless*
36. other adverbial subordinators	*since, while, whilst, whereupon, whereas, whereby, such that, so that, inasmuch as, forasmuch as, insofar as, insomuch as, as long as, as soon as* (Biber 236)
I) Prepositional phrases, adjectives and adverbs	
37. total prepositional phrases	*against, amid, amidst, among, amongst, at, besides, between, by, despite, during, except, for, from, in, into, minus, notwithstanding, of, off, on, onto, opposite, out, per, plus, pro, re, than, through, throughout, thru, to, toward, towards, upon, versus, via, with, within, without* (Quirk et al. 1985: 665-667; Biber 1988: 236-237)
38. attributive adjectives	e.g. the *big* horse (Biber 1988: 238)
39. predicative adjectives	e.g. the horse is *big* (Biber 1988: 238)
40. total other adverbs	All the adverbial forms except ns. 4-5 and ns. 39-42
J) Lexical classes	
41. conjuncts	*alternatively, altogether, consequently, conversely, eg, e.g., else, furthermore, hence, however, i.e., instead, likewise, moreover, namely, nevertheless, nonetheless, notwithstanding, otherwise, rather, similarly, therefore, thus, viz., in comparison, in contrast, in particular,*

Table 4.1 (continued)

41. conjuncts (continued)	*in addition, in conclusion, in consequence, in sum, in summary, in any event, in any case, in other words, for example, for instance, by contrast, by comparison, as a result, as a consequence, on the contrary, on the other hand, that is* (Biber 1988: 239)
42. downtoners	*almost, barely, hardly, merely, mildly, nearly, only, partially, partly, practically, scarcely, slightly, somewhat* (Quirk et al. 1985: 597-602; Biber 1988: 240)
43. hedges	*at about, something like, more or less, almost, maybe, sort of, kind of* (Biber 1988: 240)
44. amplifiers	*absolutely, altogether, completely, enormously, entirely, extremely, fully, greatly, highly, intensely, perfectly, strongly, thoroughly, totally, utterly, very* (Biber 1988: 240)
45. demonstratives	All demonstrative adjectives (e.g. *this man, that woman, these children, those dogs*)
K) Modals	
46. possibility modals	*can, could, may, might* (plus contracted forms)
47. necessity modals	*must, ought (to), should* (plus contracted forms)
48. predictive modals	*will, would, shall* (plus contracted forms)
L) Specialised verb classes	
49. public verbs	*acknowledge, admit, agree, assert, claim, complain, declare, deny, explain, hint, insist, mention, proclaim, promise, protest, remark, reply, report, say, suggest, swear, write* (Quirk et al. 1985: 1180-1181; Biber 1988: 242)
50. private verbs	*anticipate, assume, believe, conclude, decide, demonstrate, determine, discover, doubt, estimate, fear, feel, find, forget, guess, hear, hope, imagine, imply, indicate, infer, know, learn, mean, notice, prove, realise, recognise, remember, reveal, see, show, suppose, think, understand* (Quirk et al. 1985: 1181-1182; Biber 1988: 242)
51. suasive verbs	*agree, arrange, ask, beg, command, decide, demand, grant, insist, instruct, ordain, pledge, pronounce, propose, recommend, request, stipulate, suggest, urge* (Quirk et al. 1985: 1182-1183; Biber 1988: 242)
52. *seem* and *appear*	e.g. This house *seems* nice; The glass *appears* to be broken
M) Reduced forms and dispreferred structures	
53. split infinitives	e.g. *he wants to convincingly prove that…* (Biber 1988: 244)
54. split auxiliaries	e.g. *they are objectively shown to…* (Biber 1988: 244)
N) Coordination	
55. phrasal coordination	Two nouns, adjectives, adverbs, or verbs, coordinated by *and*
56. clausal coordination	*and* coordinating clauses
O) Negation	
57. synthetic negation	*no*, *neither*, *nor*
58. analytic negation	*not* (plus contracted forms)

List of linguistic features selected for analysis

The communicative functions of individual linguistic features are discussed in detail in Biber (1988: 211-245), but will be also dealt with in Chapter 6. As can be seen in Table 4.1, one of Biber's (1988) original features has been modified for this study (n. 13 questions). In Biber (1988: 227) this feature concerns only direct WH questions (such as *what? why? how?*, etc.), while we have also included the so-called *yes/no* questions. The reason for including all the direct questions in the present study is twofold. On the one hand, direct questions are likely to be infrequent in scientific texts, and it is therefore desirable to increase their number to a maximum in order to be able to include them in the statistical analysis. On the other hand, one of the concordance programs used for the retrieval of the linguistic features from our corpus (CQPWeb) allows searching for punctuation marks, making it thus extremely easy to identify direct questions through question marks. Besides that, feature n. 22 (WH clauses in subject position) has been added to the list as an alternative to n. 23 (WH clauses in object position). In this case, the retrieval of this new feature has been possible thanks to the fact that all the WH-clauses introduced by *what* were searched for with the other concordance program used for this study (the CCT), which carries out semi-automated searches that have to be later manually disambiguated (see Section 3).

As noted in Mohamed (2011: 126), the fact that all these linguistic features are categorised in classes does not mean that each class has a common linguistic function. For instance, phrasal coordination (as in example (4.1) below) and clausal coordination (example (4.2)) have different roles in the discourse: while the former is used for expanding units of ideas, the latter serves as a logical connector (Chafe 1982, 1985; Chafe & Danielewicz 1986; Biber 1988: 245):

(4.1) …after the same Manner may the present *Latitude and Reduction* be found at any other Distance of the Satellite (astr10)
(4.2) …by opening this new Inlet for his Sensations, you also open an Inlet for the Ideas, *and he finds no Difficulty of conceiving these Objects* (phil10)

Similarly, public and private verbs also have opposed functions, the former being related to actions "that can be observed publically" such as complaining, denying, or declaring, whereas the latter refers to mental, or intellectual states, such as thinking, noticing, learning, or understanding (Biber 1988: 242). However, a particular linguistic feature may tend to appear in the company of others from different classes, precisely because all these features share an underlying communicative function.

All in all, despite the fact that Table 4.1 offers quite a complete list of linguistic features which, ideally, embrace a wide range of communicative situations in the English language, one should be aware that some of these features, once retrieved and scrutinised, might drop from the study after running the Factor Analysis (see Chapter 5). On the other hand, some features from Biber's 1988 subset have not been included in this study, either because they were considered unsuitable *a priori*, or because they dropped at the retrieval stage. The reasons for their exclusion are given in what follows.

2.1. Features from Biber (1988) not included in the study

As noted, nine of the sixty-seven features initially included in Biber's (1988) list have been excluded from the present study. Such exclusion took place either *a priori* for the reasons expounded in section 2.1.1 or during the retrieval stage itself as explained in section 2.1.2.

2.1.1. Features discarded *a priori*

The features listed below have been excluded at the selection stage. It must be noted that the main reason for not including all the original sixty-seven features in the present study lies in the necessity to reduce the number of independent variables – in this case, of lexical and grammatical variables – in the dataset used for the Factor Analysis, considered the limited number of texts in our corpus (see detailed explanation of this condition in Chapter 5). However, the procedure of discarding linguistic features should be always carried out with caution, as the exclusion of one feature and its replacement with another will always influence the final results. Therefore, only those linguistic features which were either considered not to have sufficient relevance in the analysis, or particularly difficult to retrieve, were discarded:

- present participial clauses (n. 25 in Biber (1988: 233); e.g. *Stuffing his mouth with cookies, Joe ran out the door*).

Present participial clauses are typically found in narration (Thompson 1983; Beaman 1984; Granger 1997a), which is confirmed through its inclusion in Biber's (1988)

Dimension 2 “Narrative vs. Non-Narrative Concerns”. Considering the difficulty in retrieving this feature (which requires manual disambiguation of all *–ing* forms) and the relatively small weight of Dimension 2 in Biber’s (1988) academic prose, it was decided not to include present participial clauses in the analysis.

- type/token ratio and word length (n. 43 and n. 44 in Biber 1988: 238-239)
Biber (1988) calculated the type/token ratio as the number of different lexical items (types) in a text, as a percentage (i.e. per 100 tokens). Word length, in turn, is equivalent to the mean length of the words in a text, counted in orthographic letters. Both are measures of lexical specificity and have not been included in the present study for technical reasons.

- emphatics (n. 49 in Biber 1988: 241; e.g. *for sure, a lot, such a, real* + ADJ, *so* + ADJ, DO + V, *just, really, most, more*) and discourse particles (n. 50 in Biber 1988: 241; e.g. *well, now, anyhow, anyway, anyways*)
These two features are typical of spontaneous (i.e. live, or real-time) speech elaboration, occurring mostly in conversations and very rarely, or never, in formal written discourse. Furthermore, after consulting the *OED*, it was found that most of these lexical items were not common in writing until the late nineteenth and the early twentieth century.

- contractions (n. 59 in Biber 1988: 243)
Just like emphatics and discourse particles, contractions usually characterise informal spoken discourse. Although they can also be commonly found in eighteenth-century prose, both informal and formal, they seem to indicate stylistic, rather than register-based differences.

- subordinator-*that* deletion (n. 60 in Biber 1988: 244; e.g. I think [*that*] he went to…)
As was the case with pro-verb DO, subordinator-*that* deletion is a feature of ellipsis, typical of conversations, and unlikely to occur in academic writing.

- stranded prepositions (n. 61 in Biber 1988: 244; e.g. *the candidate I was thinking of*)
This is another feature which is normally restricted to conversational registers (Chafe 1985; Johansson & Geisler 1998), which is why it has not been included in the present study.

2.1.2. Features excluded at the retrieval stage

Two linguistic features, initially considered for the analysis, showed problems during or after the retrieval stage:

- gerunds (n. 15 in Biber 1988: 227; e.g. *The reading of books decreased*)

Biber (1988: 227) included in this category all participle forms serving nominal functions. It was initially planned to count gerunds as another type of nominalisations (see Bello 2014), and thus include an extra suffix *–ing* in category n. 13 (see Table 4.1). However, their retrieval turned out to be problematic in that it was rather difficult to distinguish between verbal and nominal *–ing* forms, both automatically and manually (see Section 3). It was decided therefore to exclude gerunds from the analysis to avoid skewed results.

- *that* relativiser in object position (n. 30 in Biber 1988: 234); e.g. *the dog that I saw*)

Although an initially suitable query algorithm was developed with CQPWeb, it was found that it returned too many errors (see a more detailed explanation in Section 3.1.1). This feature was eventually excluded due to the impossibility of developing a query that would increase precision.

The next section deals with the retrieval stage of the linguistic features we have selected for analysis. The concordance programs and query algorithms used for each feature are discussed in it.

3. Retrieval of lexical and grammatical variables

As has been already introduced in Chapter 3 (Section 3), two versions of the *Coruña Corpus* have been used in the present study. One of them has not been annotated for parts of speech and semantic categories and is usually exploited with the CCT (Moskowich & Parapar 2007), a concordance program that permits to run single- and multi-word and wildcard searches, creates frequency lists, and allows establishing criteria for the queries from corpus metadata files. As we have seen on Figure 3.7, the CCT displays a KWIC view of the concordance lines. By clicking on one of the hits another window is opened, showing the query result in the surrounding context (text) and thus allowing a clearer disambiguation when needed. The CCT has been used to retrieve all the linguistic features which did not necessarily require part-of-speech tags for their automated search in the corpus (totalling thirty-five features), which was mostly the case of closed lists of lexical items (e.g. conjuncts, adverbs, or prepositions), or of those features which could be retrieved through a wildcard search (e.g. all the words ending in *–ed* to retrieve past tense or past participle forms).

The other version of the *Coruña Corpus* is the one that has been annotated for part of speech and semantic categories, and is processed through the online corpus query processor CQPWeb (Hardie 2012), which allows running queries of varying complexity, from simple words to part-of-speech tags to elaborated algorithms that capture more or less intricate syntactic structures. The three subcorpora of the *Coruña Corpus* uploaded on CQPWeb are annotated at different levels of annotation with seven different tagsets. However, the three levels of annotation used for this study have been part-of-speech (annotated with the CLAWS6 tagset[26], or C6), lemma, and a simplified version of part-of-speech, or simple POS (annotated with the Oxford Simplified Tagset). The first one (C6) contains a total of 148 part-of-speech tags (e.g. NN1 = singular common noun; NN2 = plural common noun; VV0 = base form of a lexical verb; VVD = past tense of a lexical verb, etc.), which belong to different linguistic categories. The third one (Oxford Simplified Tagset), in turn, contains eleven broad part-of-speech categories (e.g. SUBST, VERB), each of which combines several parts-of-speech tags from the C6 tagset. Finally, the second level of annotation is that of lemmata, which allows to retrieve the different forms of a word. In this study, CQPWeb has been used to retrieve those linguistic features (twenty-three in total) which did require part-of-speech tags for their identification (e.g. nouns, present tense verbs), or for those which were identifiable in function of the syntactic structure in which they were embedded (e.g. attributive vs. predicative adjectives), or which constituted a syntactic structure of their own (e.g. *that* relativiser in subject function).

Following Lee (2000) and Mohamed (2011), in order to develop the queries for CQPWeb we have used the standardised CQP Syntax (see Evert 2003, 2005; Hoffman et al. 2008). The following paragraphs discuss the retrieval of each linguistic feature, giving the algorithms used for those features which were searched for with CQPWeb. At the end of the section, the situation with the feature that dropped from

[26] For a complete list of the CLAWS6 tagset, see http://ucrel.lancs.ac.uk/claws6tags.html.

our study at the stage of retrieval (i.e. *that* relativiser in object position)_is also described.

3.1. Algorithms and queries

A) TENSE AND ASPECT MARKERS

1. past tense

This feature has been searched for with the CCT. The query <.*(e|')d> returned all the strings ending in *–ed* or *'d*, capturing simple past and past participle allomorphs *ed* and '*d* which were later disambiguated manually in order to count only the simple past forms, such as *opened* and *open'd*. CQPWeb was not used for this feature because, although C6 has different POS tags for past tense (V.D) and past participle (V.N), both queries return many tagging errors (which are inevitable when tagging is automatic), capturing either the wrong past form, or else, adjectives ending in *–ed*.

Likewise, individual queries were run for each past form of the irregular verbs listed in Quirk et al. (1985: 115-120), considering all the allomorphs used in the eighteenth and early nineteenth centuries (e.g. *wrote/writ*). In addition, past tense forms of the verbs *be*, *do* and *have* were also counted. Although the latter included both the lexical and the auxiliary verbs *have*, overlapping with perfect aspect forms, these cases were not disambiguated because we consider that tense and aspect are different categories which certainly overlap in English, but one does not cancel the other, and vice versa.

2. perfect aspect

CQPWeb has been used for retrieving perfect aspect forms. Adverbs, negation (*not*, *n't*) and pronouns have been included in the query, allowing to catch combinations such as "has indeed discovered" (astr11)[27], "has not only travelled" (astr37), or "nor had he studied" (life20). The algorithm used for this particular query was:

```
[pos="VH(I|0|Z|D|G)"][pos="RR.*|XX|PN1|PP(H1|HS1|HS2|IS1|IS2|Y)"]
{0,2}[pos="V(B|D|H|V)N"]
```

[27] Text samples will be identified by their short *id*, as specified on tables 3.5-3.7 in Chapter 3.

3. present tense

The query for this feature has also been developed with CQPWeb. Based on Mohamed (2011: 474), it includes all the present forms of the verb *be* plus the finite bare form and third person singular of *do*, *have* and all the lexical verbs. However, bare forms inevitably include imperatives and, therefore, the query has been further restricted to exclude, at least, all cases of the imperative *let* (e.g. *let us consider*), which are especially frequent in the Astronomy subcorpus (more than 400 cases, compared to Philosophy and Life Sciences, which have around 100 cases each). Likewise, the contracted third person singular form *‘s* (as in *it's perfect*) has also been excluded from the query due to the large number of tagging errors where Saxon genitive is captured instead:

```
[pos="VBM|VBR|VBZ|VD0|VDZ|VH0|VHZ|VV0|VVZ"&word!="let"%cd&
word!="\'s"]
```

B) PLACE AND TIME ADVERBIALS

4. place adverbials

This feature was retrieved with the CCT, searching for each adverbial from the closed list given by Biber (1988: 224). Some lexical items, such as *north*, *south*, *east* and *west* had to be disambiguated, as they can be both place adverbs of direction (as in *heading north*) and nouns (as in *in the North of England*).

5. time adverbials

Just as place adverbials, the list of time adverbials (Biber 1988: 224) was searched for in the corpus with the CCT. Likewise, forms such as *early* and *earlier* could be both adverbs (e.g. *he came early)* or adjectives (e.g. *early morning*), and were subsequently manually disambiguated.

C) PRONOUNS AND PRO-VERBS

6. first person pronouns
7. second person pronouns
8. third person pronouns (excluding *it*)

All these features were searched for with the CCT. As noted by Lee (2000: 109), Biber (1988) did not count *mine* and *ours* “for unspecified reasons”, nor did he count *yours*. In order to follow Biber (1988) as closely as possible, we have not included those forms either. To maximise recall, contractions (e.g. *I'll*, *you'd*) were also searched for.

9. pronoun *it*

This feature was also retrieved with the CCT, including contractions (*it's, it'll*) in the query. The contraction *it's* had to be disambiguated because it included not only a contracted form of *it* + *is*, but also an allomorph of the possessive pronoun *its* very common in the late Modern English period (Millward 1989: 230).

10. demonstrative pronouns

CQPWeb was used for the retrieval of this feature, considering the large amount of manual editing needed to separate demonstrative pronouns (as in *this is mine*) and adjectives (as in *this dog is yours*), as well as all the different functions of the form *that*. The following query was used:

```
[word="this|that|these|those"&pos="DD(1|2)"][pos="V.*"|pos="Y.*"]
```

Cases of *that* followed by a comma contained tagging errors where *that* was a conjunction and were therefore manually disambiguated. Other tagging errors, such as those where *that* was a relative, were infrequent.

Biber (1988: 226) also looked for the contraction *that's*. In the C6 tagset in CQPWeb, contractions are tokenised as words and tagged as verb forms *('s = is, 'd=* *had* or *would, 'll = will*), so there was no need to look for contractions specifically as they are already included in the V.* tag in the algorithm, which equals "any verbal form".

11. indefinite pronouns

The closed list used by Biber (1988: 226) was searched for with the CCT. Alternative forms used in the eighteenth century, such as *every body, every one*, *no where* or *every thing*, were included in the searches. Manual disambiguation was used to spot cases where *every body* had a literal sense.

12. pro-verb *do*

This feature was retrieved with CQPWeb. Two different queries have been developed to capture the maximum number of instances of the pro-verb *do*:

```
[word!="to"%cd][pos="VD(0|I|Z|D)"][pos="Y(COM|EX|STP|COL|SCOL|
QUE)"]
```

```
[pos="VD.*"][word="it"]
```

The first one captures cases such as "Let them soak in the Brine double the Time the others *do*, …" (life11), while the second one captures instances of *do* before *it*, as in "they *do it* in so sparing a way, that…" (life25). This also included cases such as "nor does it appear…" (life13), which were eliminated manually.

D) QUESTIONS

13. direct questions

Biber (1988: 227) looked only for direct WH questions, due to technical limitations of the tagging program used. Following Lee (2000: 110), we have considered all the direct questions appearing in the corpus, counting all the question marks (<?>) tokenised with C6 in CQPWeb, using the following query:

```
[word="\?"]
```

E) NOMINAL FORMS

14. nominalisations *(-tion, -ment, -ness, -ity*; Biber 1988: 227)[28]

The CCT was used to retrieve all the endings in *-tion, -ment, -ness, -ity* and their plural forms. Manual disambiguation followed to discard verb forms, such as *cement*(*s*), *torment*(*s*), or *augment*(*s*).

The queries of the CCT also inevitably captured nouns which are not nominalisations, such as *city*, *moment*, or *motion*. Still, those forms were equally counted because neither Biber (1988: 227) nor Lee (2000: 111) disambiguated those cases.

[28] Although the category of nominalisations includes many more forms tan those listed above, such as those ending in *–ism*, *-ship*, *-ance*, *-ence*, *-ery*, *-hood*, etc., it was decided to look only for the closed lists specified in Biber (1988) for this study.

As already mentioned in Section 2.1, gerunds (or nominal forms ending in *–ing*) were initially included in this category, but, following Bello (2014: 197), were later discarded due to the difficulty in distinguishing between nominal and verbal forms.

15. total other nouns (all except n.14)
Nouns were searched for with CQPWeb. This query excludes proper nouns, abbreviated nouns of titles (*Mr., Mrs., M.A.*, etc.), as well as the nominalisations included in category n. 14. Likewise, following Biber (1988: 228), all forms ending in *–ing* (gerunds) were also excluded from the query:

> [class="SUBST"&word!=".+tion(|s)|.+ment(|s)|.+ness(|es)|.+it(y|ies)|.+ing(|s)"
> %cd&pos!="Z.*"&pos!="NP.?"&pos!="NN(A|B)"]

F) PASSIVES
16. agentless passives
17. *by*-passives
Passives were retrieved with the CCT through the same query that was used for past tense in regular verbs. For irregular verbs, the list of past participles from Quirk et al. (1985: 115-120) was searched for, including allomorphs used in the eighteenth and early nineteenth centuries (e.g. *shown/shewn, wrote/writ/written, hid/hidden*).

At a first stage, stative uses, as in "[the muscle] is entirely *covered* with hair" (life 28), were separated from dynamic uses, as in "the lower jaw of man is *distinguished* by the prominence of the chin" (life24), only the latter being included in the analysis. This was done because stative uses indicate state, rather than action deliberately marked with passive voice, and may sometimes be considered borderline cases between a past participle and an adjective.

At a second stage, in order to differentiate between agentless and *by*-passives, and to separate them from categories n. 25 and n. 26 (see below), all these forms were carefully read and manually disambiguated. In the case of *by*-passives, all the instances of *by* were also disambiguated to include only those cases where *by* is followed by an agent, discarding instrumental complements. Thus, examples such as the one used above to illustrate a dynamic use – "the lower jaw of man is distinguished by the prominence of the chin" (life24)" – were all considered agentless passives.

G) STATIVE FORMS

18. existential *there*

CQPWeb was used to retrieve this feature, as the C6 tagset has a specific tag for existential *there:*

```
[pos=“EX”]
```

This tag captures errors such as “There Men are taught, not to glorify God in subservience…” and “There Rewards are propounded to Christians” (phil7), where *there* is a place adverbial. However, if we restricted our query to

```
[pos= “EX”][class!= “SUBST”]
```

we would exclude potential cases where existential *there* + N is valid, such as interrogatives (e.g. *Is there food in the fridge?*). Although there are no such cases in the corpus, we have decided to subtract manually the two invalid cases from the total counts of existential *there* in *CEPhiT.*

On the other hand, the first query also captured a case of “here-and-there”, erroneously tagged as an existential *there* (life10). Although this case can also be subtracted manually, we decided to exclude all potential tagging errors with the following query, specifying that only instances of the word *there* should be returned:

```
[pos="EX"&word="there"%cd]
```

19. *be* as main verb

Copular *be* was also retrieved with CQPWeb, using the following query:

```
[pos!="EX"][pos="VB.*"&word!="'s"][pos="A.*|N.*|P.*|JJ"][pos!=
"V.N|V.G"]
```

The beginning of the query algorithm has been configured to exclude cases of existential *there*, such as “there is a Truth and Falsehood in all Propositions…” (phil10).

H) SUBORDINATION FEATURES

20. THAT verb complements

This feature was searched for with CQPWeb. The query

```
[pos="V.*"&word!="provided"%cd][pos="CST"]
```

was restricted so that instances of the conditional form *provided that* would not be captured.

21. THAT adjective complements

This feature – as in "we may be *sure that*…" (astr12) – was relatively straightforward to search for, with the following query:

```
[pos="JJ.*"][pos="CST"]
```

22. WH clauses in object position

23. WH clauses in subject position

Both forms were retrieved with the CCT, searching for instances of *what* introducing a WH-clause. Biber's (1988: 231) algorithm only allowed to find examples of n. 22, but as we have disambiguated each case manually, instances of n. 23 could also be identified.

24. *to*-infinitives

Infinitives preceded by *to* were retrieved automatically with CQPWeb, the query being:

```
[pos="TO"][pos="V.I"]
```

25. detached past participial clauses with adverbial function

26. past participial WHIZ-deletion relatives

These two features were retrieved through manual disambiguation of regular and irregular past participle forms, just like categories n. 16 and n. 17. In principle, these two forms were easy to distinguish from *by*- and agentless passives at a first stage, in that only the latter are accompanied by a verb, while the former are non-finite forms.

At a second stage, past participle WHIZ-deletion relatives were identified as those post-modifying a noun.

27. present participial WHIZ-deletion relatives
Unlike n. 26, this feature was searched for with an automated query on CQPWeb:

```
[class="SUBST"][pos="VVG"]
```

This query was configured to capture only the past participles of lexical verbs, excluding *be* and *have*, in order to exclude cases such as “the brain of the child in the womb *being* too moist” (phil11), or “Hobbes *having* said that, …” (phil10), which would both be detached present participial clauses.

28. *that* relativiser in subject position (e.g. *the dog that bit me*)
These forms were also retrieved with CQPWeb. The query

```
[class="SUBST"&pos!="NULL"][pos="CST"][pos="R.*"]{0,2}[pos="V.*"]
```

captured instances such as “the sort of thing that really exists” (phil33), “Effects and Appearances that necessarily depend thereupon” (phil2), or “Methods that ever were” (astr3).

29. WH relativiser in subject position
    (e.g. “*the woman who has* only been taught to pleaſe” (phil19))
30. WH relativiser in object position
    (e.g. “*the Deity, whom Berkeley would have* introduced” (phil36))
31. pied-piping relative clauses
    (e.g. “*the orbit in which* it moves” (astr12))
32. sentence relatives
    (e.g. “the fruit will firſt begin to change in the middle, *which is a certain indication of its being ripe*” (life17))

The CCT was used to retrieve these four features. The forms searched for included *who, whom, whose*, and *which*, and had to be subsequently disambiguated by hand.

33. causative adverbial subordinators (*because*)
34. concessive adverbial subordinators (*although, though*)
35. conditional adverbial subordinators (*if, unless*)
These three features were also retrieved with the CCT. Variants used in the eighteenth century, such as *'cause* and *thou*' were also found. As these terms have a straightforward meaning, further disambiguation was not necessary.

36. other adverbial subordinators
Biber's (1988: 236) closed list of what he termed "other adverbial subordinators" was searched for with the CCT, including variants such as *in so far as*, or *for as much as*. Following Lee (2000: 128), the forms *since* and *while* were later manually disambiguated to keep only those cases where they function as logical connectors, as in "Therefore, *since* great accuracy cannot be expected here, multiply…" (astr14), and discarding those cases where they function as time adverbials, such as "And, not 200 years *since*, the great Galileo met with the same fate" (astr19).

I) PREPOSITIONAL PHRASES, ADJECTIVES AND ADVERBS
37. total prepositional phrases
Once more, the list of prepositions adopted from Quirk et al. (1985: 665-667) in Biber (1988: 236-237) was searched for with the CCT. Further disambiguation was not needed.

38. attributive adjectives
The query developed for this feature on CQPWeb was based on the directions given in Lee (2000: 129):

```
[pos="JJ.*"][pos="CC.*"|pos="XX"|pos="RG.*"|pos="JJ.*"]{0,3}[class="SU
BST"]
```

This allowed to retrieve cases such as "whole terrestrial orbit" (astr22), "easy and obvious Manner" (phil10), "beautiful and more engaging Colours" (phil10), "numberless other Parts" (life10), "small continual smothering Fire" (life 10), or "small yellowish kidney-shaped capsule" (life20).

As the concordance programme counts hits, rather than actual instances of adjectives, the resulting combinations had to be revised manually in order to maximise recall.

39. predicative adjectives

A query developed at an earlier stage to retrieve predicative adjectives was adapted from Mohammed (2011: 478) but, like the query for attributive adjectives, also included instances of more than one adjective, degree adverbs and negation:

```
[lemma="seem|become|get|go|grow|prove|come|turn|appear|keep|remain|stay|be"&class="VERB"][pos="XX"|pos="RG.*"]{0,3}[pos="JJ.*"&word!="supposed"%cd][class!="SUBST"]
```

However, this query also retrieved cases such as “Animals are sensitive organic bodies” (life10). Although Lee (2001: 130) seems to classify such structures as instances of predicative adjectives[29], both adjectives in this example clearly premodify a noun and have been counted, therefore, as attributive, under n. 35. The reason is that, in the present query, the restriction [class!="SUBST"] does not work for those cases where there are two adjectives in a row and one of them is optional. It would only work if the query were something like

```
[pos="JJ.*"][pos="JJ.*"][class!="SUBST"]
```

but this would only return the combinations ADJ + ADJ not followed by a noun.

To avoid further complications, we have therefore decided at this point to follow Lee (2001) and retrieve the predicative adjectives by subtracting all the instances of attributive adjectives from the total number of adjectives identified by C6 ([pos="JJ.*"]).

40. total adverbs

To retrieve this feature, all the types tagged as adverbs in the Oxford Simplified Tagset were included in the CQPWeb query, excluding hedges, amplifiers and downtoners (ns. 42-44), place adverbials (n. 1), time adverbials (n. 2), other adverbial

[29] Lee’s (2001: 130) examples were “They are happy, friendly people”, “They are happy, very friendly people” and “They are happy and friendly people”.

subordinators (n. 36) and conjuncts (n. 41), as well as the negation adverbs *no, not, n't* (ns. 57 and 58). This produced a long, if not very complex, query command:

```
[class="ADV"&word!="almost|barely|hardly|merely|mildly|nearly|only|partiall
y|partly|practically|scarcely|slightly|somewhat|maybe|absolutely|altogether|co
mpletely|enormously|entirely|extremely|fully|greatly|highly|intensely|perfectly|
strongly|thoroughly|totally|utterly|very|aboard|above|abroad|across|ahead|along
side|around|ashore|astern|away|behind|below|beneath|beside|downhill|downstai
rs|downstream|east|far|hereabouts|indoors|inland|inshore|inside|locally|near|nea
rby|north|nowhere|outdoors|outside|overboard|overland|overseas|south|underfo
ot|underground|underneath|uphill|upstairs|upstream|west|afterwards|again|earli
er|early|eventually|formerly|immediately|initially|instantly|late|lately|later|mom
entarily|now|nowadays|once|originally|presently|previously|recently|shortly|si
multaneously|soon|subsequently|today|tomorrow|tonight|yesterday|since|while|
whilst|whereupon|whereas|whereby|alternatively|altogether|consequently|conv
ersely|else|furthermore|hence|however|instead|likewise|moreover|namely|never
theless|nonetheless|notwithstanding|otherwise|rather|similarly|therefore|thus"%
cd&pos!="XX"]
```

K) LEXICAL CLASSES

41. conjuncts

As was the case with other place and time adverbials, adverbial subordinators and prepositions, a close list of conjuncts (Quirk et al. 1985: 634-636; Biber 1988: 239) was searched for with the CCT. Manual disambiguation was needed in some cases, such as with the lexical item *however*, where instances such as "*however* insignificant and mean man might be in comparison with…" have been discarded. Likewise, some occurrences of *altogether* were classified under category n. 44 (amplifiers).

42. downtoners

Unlike ns. 43 and 44, downtoners were searched for with CQPWeb, simply because the latter was available at the time of retrieval. The query is based on Mohamed (2011: 132) and includes all the lexical items from Biber's (1988: 240) closed list:

```
[word="almost|barely|hardly|merely|mildly|nearly|only|partially|partly|practically|scarcely|slightly|somewhat"%cd&class="ADV"]
```

43. hedges
44. amplifiers

As hedges and amplifiers were searched for at an earlier stage of this study when CQPWeb was not yet available for the *Coruña Corpus*, they were retrieved with CCT. Once more, each of the terms included in the closed lists offered in Biber (1988: 240) was searched for. Disambiguation was needed for the multi-word term *more or less* (hedge), which appeared sometimes with a literal meaning, as in "we should count a day more or less, according as we went east or west" (astr31).

45. demonstratives

As opposed to demonstrative pronouns (n. 10), demonstrative adjectives precede nouns. The query used for their retrieval, developed on CQPWeb, is based on Mohamed (2011: 137):

```
[word="this|that|those|these"&pos="DD(1|2)"][pos="VV(G|N)"|pos="XX"|pos="RG.*"|pos="JJ.*"]{0,3}[class="SUBST"]
```

This query is extended enough to capture cases such as "this inconceivable self-determining power" (phil20), or "these very small animals" (life10). However, it also returned a few tagging errors where *that* is not a determiner, as in "But may we not hope, *that* Philosophy, if cultivated with care (…) may carry its researches still farther…?" (phil10).

L) MODALS
46. possibility modals
47. necessity modals
48. predictive modals

The three groups of modals were searched for with the CCT. Eighteenth-century variants of *could* (*cou'd*), *should* (*shou'd*), and *would* (*wou'd*), as well as contractions (e.g. *we'll, you'd, shan't, won't*) were included in the searches. The instances retrieved were then manually disambiguated, leaving out cases where *'d* was a contraction of *had*, or when *will* was a noun.

M) SPECIALISED VERB CLASSES

49. public verbs

50. private verbs

51. suasive verbs

These verbs, which belong to three closed lists given in Biber (1988: 242), adopted after Quirk et al. (1985: 1180-1183), were also retrieved with the CCT. The queries included base and third person singular forms, as well as the past simple and the present and past participle forms. As was the case with past tense forms (n. 1) and past participles (ns. 16, 17, 25 and 26), earlier forms of those verbs used in the early and late Modern English periods were also searched for. Manual disambiguation was carried out in some cases, such as when it was necessary to distinguish a nominal form from (*the show*) from a verbal form (*you show*).

52. *seem* and *appear*

CQPWeb was used to retrieve this feature, the following query being used:

```
[lemma="seem|appear"&pos="VV.*"]
```

N) REDUCED FORMS AND DISPREFERRED STRUCTURES

53. split infinitives

The query for split infinitives such as “to rightly understand”, “to exactly correlate” (phil35) was also developed on CQPWeb, and based on Mohamed (2011: 488):

```
[pos="TO"][pos="XX"|pos="R.*"][pos="R.*"]?[pos="V.I"]
```

54. split auxiliaries

The CQPWeb query for split auxiliaries was likewise based on Mohamed (2011: 488), but had to be further restricted to avoid cases such as “then when they *are ready eat* them with this Sauce” (life11). Negation was not included in the query because it would overlap with all the instances of negation in passive structures, modals and perfect aspect:

```
[pos="VB.*"][pos="R.*"][pos="R.*"]?[pos="V.*"&pos!="VV0"]
```

In addition, a second query was developed to include modal auxiliaries and HAVE, which were apparently included in Biber's (1988: 244) searches:

```
[pos="VH.*|VM"][pos="R.*"][pos="R.*"]?[pos="V.*"]
```

O) COORDINATION

Although an ideal way to retrieve instances of phrasal and clausal coordination could have been using the CCT, followed by exhaustive manual disambiguation, thus maximising precision and recall, the relatively large size of the corpus (more than one million words) would have made this task extremely arduous and possibly unproductive. Therefore, it was decided to follow Biber (1988: 245) and Lee (2000: 139) and retrieve both features with precise automated queries.

55. phrasal coordination

Four automated queries were used to retrieve instances of phrasal coordination, including *&* as a variant of *and*:

N *and* N:

```
[class="SUBST"][pos="CC"&(word="and"|word="\&")][class="SUBST"]
```
[30]

ADJ *and* ADJ:

```
[pos="JJ.*"][pos="CC"&(word="and"|word="\&")][pos="JJ.*"]
```

ADV *and* ADV:

```
[pos="R.*"][pos="CC"&(word="and"|word="\&")][pos="R.*"]
```

V *and* V:

```
[pos="V.*"][pos="CC"&(word="and"|word="\&")][pos="V.*"]
```

56. clausal coordination

Unlike Biber (1988) and Lee (2000), we have also included the operators *or* and *but* in this query, which, just like *and*, serve to mark "many different logical relations between clauses" (Biber 1988: 245). As the boundaries between clauses are not easy

[30] Lee (2000: 139) excludes proper nouns from the algorithm, but we see no reason to do that.

to grasp with an automated query, two complementary queries were developed, capturing only those instances of *and*, *or* and *but* which occur after a punctuation mark:

> [pos="Y(EX|STP|COL|SCOL|QUE)"][pos="CC.*"]
>
> [pos="YCOM"][pos="CC.*"][pos="P.*"|pos="R.*"|pos="EX"]

The second query was restricted so that it would avoid capturing instances of phrasal coordination which contain a comma, such as “apples, and oranges”.

P) NEGATION
57. synthetic negation (*no, neither, nor*)
58. analytic negation (*not, n't*)
Both synthetic and analytic negation were retrieved with the CCT. Following Biber (1988: 245), *no* as a response, as in “her sinking and aching heart will answer, no!” (phil30), was excluded.

3.2. Unsuccessful queries
As mentioned at the beginning of this section, the following feature was dropped at the stage of retrieval as a result of an unsuccessful query:

- *that* relativiser in object position (e.g. *the dog that I saw*)

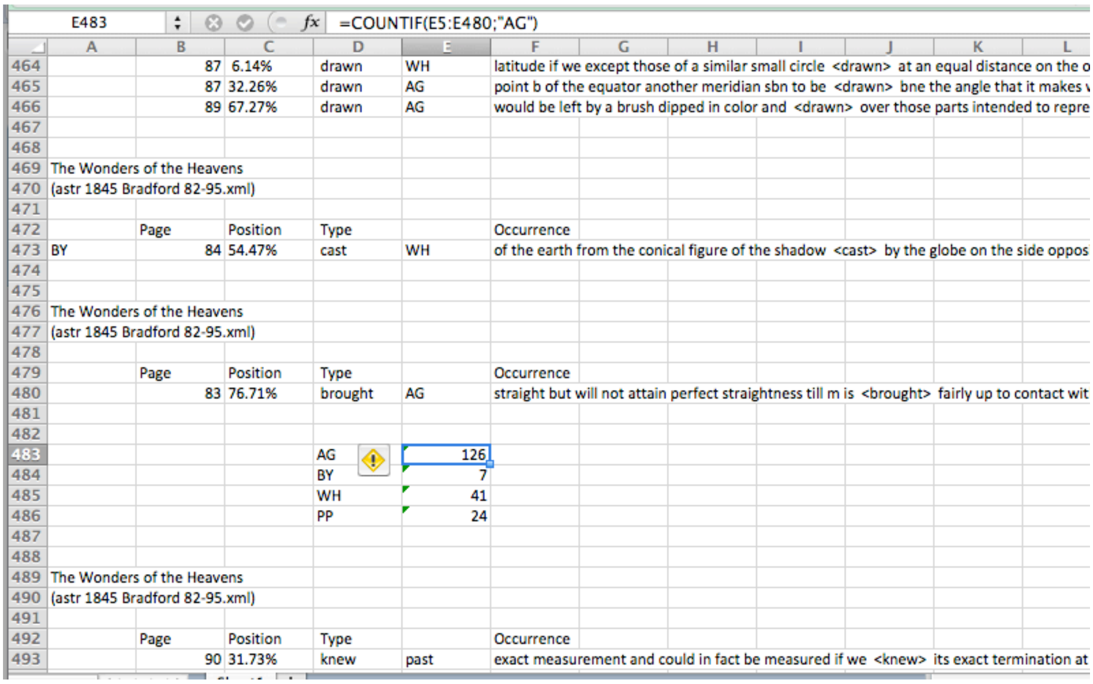
Initially, a CQPWeb query for *that* relativisers functioning as object was developed:

> [class="SUBST"&pos!="NULL"][pos="CST"][pos="A.*|D.*"]{0,1}[class="SUBST"&pos!="NULL"|pos="PN1|PP(H1|HS1|HS2|IS1|IS2|Y)"][pos="V.* "]

This query allowed to retrieve instances such as “the Rank *that* each man holds” (life38). However, it also returned errors such as “the fact *that* it is contrary to common sense” (phil33), which constituted more than 50% of the cases in the Philosophy subcorpus. As it proved impossible to further distinguish in an automated query between *that* as an object of a clause and *that* as a conjunction, it was decided to drop this feature from the present study.

4. Counting the frequencies of the linguistic features

As has been noted earlier, thirty-five of the total fifty-eight features were searched for with CCT, while the twenty-three remaining features were retrieved with CQPWeb. After a feature was searched for in the corpus with CCT, concordance results were stored in spreadsheets created with Microsoft Excel for Mac 2011 (version 14.0.0). At a first stage, a spreadsheet was created for each form searched for (e.g. a particular WH-word, such as *what*, or *which*), and after these were disambiguated, another spreadsheet was created for each text from the corpus, containing all the forms included in a feature (e.g. the different WH-words included in the feature "WH relativiser in object position", such as *who, whom, whose,* or *which*). On the other hand, there were also cases when several features were combined in the same spreadsheet. For instance, all the different passive constructions (agentless passives, *by*-passives, past participle clauses and past participial WHIZ-deletions), as well as past tense, were combined in one single spreadsheet. Thus, a total of 3294 spreadsheets (27 form combinations per 122 texts) were created for each text in the corpus.



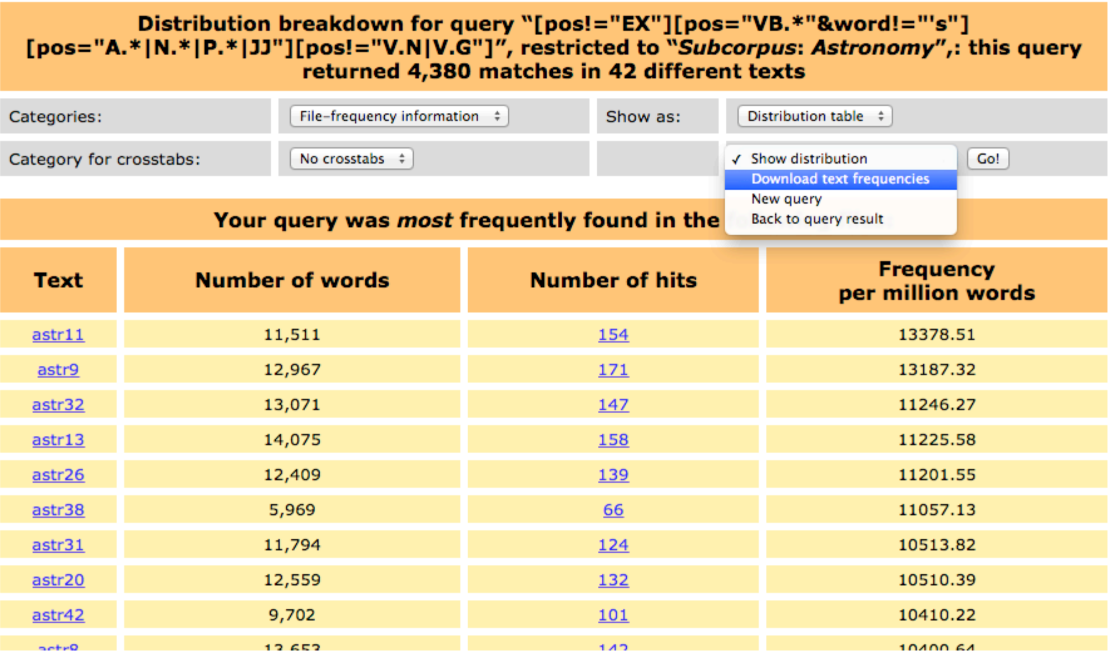
E483 =COUNTIF(E5:E480;"AG")

	A	B	C	D	E	F
464		87	6.14%	drawn	WH	latitude if we except those of a similar small circle <drawn> at an equal distance on the o
465		87	32.26%	drawn	AG	point b of the equator another meridian sbn to be <drawn> bne the angle that it makes
466		89	67.27%	drawn	AG	would be left by a brush dipped in color and <drawn> over those parts intended to repre
467						
468						
469	The Wonders of the Heavens					
470	(astr 1845 Bradford 82-95.xml)					
471						
472		Page	Position	Type		Occurrence
473	BY	84	54.47%	cast	WH	of the earth from the conical figure of the shadow <cast> by the globe on the side oppos
474						
475						
476	The Wonders of the Heavens					
477	(astr 1845 Bradford 82-95.xml)					
478						
479		Page	Position	Type		Occurrence
480		83	76.71%	brought	AG	straight but will not attain perfect straightness till m is <brought> fairly up to contact wit
481						
482						
483				AG	126	
484				BY	7	
485				WH	41	
486				PP	24	
487						
488						
489	The Wonders of the Heavens					
490	(astr 1845 Bradford 82-95.xml)					
491						
492		Page	Position	Type		Occurrence
493		90	31.73%	knew	past	exact measurement and could in fact be measured if we <knew> its exact termination at

Figure 4.1
Excel database for passive constructions and past tense forms in a text from *CETA* (astr31)

The next step was to count the hits of those forms selected after disambiguation in each of these spreadsheets, using the Excel formula COUNTIF or COUNTIFS, depending on each case. This allowed us to obtain raw frequencies of occurrence of each feature for each text. Figure 4.1 offers a screenshot of the spreadsheet that contains the passive structures and past tense forms found in the astronomy text by Bradford (1845), identified as astr31 in the Astronomy subcorpus (*CETA*) (see on the left).

On the other hand, when a feature was retrieved with CQPWeb, the query results were stored in the system. In order to count the raw frequencies of occurrence of a feature, the "Distribution" and the "File frequency information" options were selected from the menu. Downloadable results were then displayed for each text, containing both raw and normalised figures per 1 million words (see Figure 4.2):



Distribution breakdown for query "[pos!="EX"][pos="VB.*"&word!="'s"][pos="A.*|N.*|P.*|JJ"][pos!="V.N|V.G"]", restricted to "*Subcorpus*: *Astronomy*",: this query returned 4,380 matches in 42 different texts

Categories: File-frequency information

Show as: Distribution table

Category for crosstabs: No crosstabs

Your query was *most* frequently found in the

Text	Number of words	Number of hits	Frequency per million words
astr11	11,511	154	13378.51
astr9	12,967	171	13187.32
astr32	13,071	147	11246.27
astr13	14,075	158	11225.58
astr26	12,409	139	11201.55
astr38	5,969	66	11057.13
astr31	11,794	124	10513.82
astr20	12,559	132	10510.39
astr42	9,702	101	10410.22
astr8	13,653	142	10400.64

Figure 4.2
File-frequency information for query results in CQPWeb

However, normalised figures on CQPWeb presented a problem, because CQPWeb tokenises punctuation marks, which means that every punctuation mark in the corpus (including commas, colons, exclamation marks, brackets, etc.) is counted as a word. For this reason, each text in the corpus presents a much larger word count on CQPWeb than it is actually the case. Thus, in order to avoid skewed results, only raw figures given by CQPWeb were recorded.

Once raw counts were obtained for each linguistic feature in each text, normalised frequencies per 1,000 words of texts were calculated. This was carried out by multiplying each raw frequency by 1,000 and dividing it by the total number of words in the text. Normalisation is important when intending to obtain comparable results. Although most samples in our corpus contain approximately 10,000 words, it must be borne in mind that some texts in the Astronomy subcorpus (*CETA*) are only five- or six-thousand-word long (see Chapter 3), which means that raw counts cannot be directly compared. For instance, if past tense verbs occur 49 times in a text which is 5,000 words long and 59 times in another text which is 9,000 words long, normalised figures will show that this feature is more frequent in the former (occurring 9.8 times per 1,000 words) than in the latter (6.5 times per 1,000 words), despite the fact that the latter presents a larger raw number of past tense verbs.

Thus, a dataset containing the normalised frequency counts for each linguistic feature in each text was created in order to be used in the factor analysis. The methodology followed in this statistical procedure is explained in the next chapter.

Chapter 5

Methodology (II): Factor Analysis

1. Introduction

In Chapter 4 we have described the first methodological part of this study, which consisted in the selection of lexical and grammatical variables and their retrieval from the corpus, either automatically through a series of query algorithms, or semi-automatically, with subsequent manual disambiguation. Having retrieved a total of fifty-eight linguistic features, the goal is to reveal underlying communicative patterns through the different combinations of these features in the corpus samples. This chapter deals with statistical methodology, focusing on factor analysis, a multivariate statistical technique for data reduction. Although we have briefly introduced factor analysis in Chapter 1, Section 2 reviews its basics in more detail, focusing on some key issues such as sampling adequacy, communality and uniqueness, and extraction and rotation methods. Section 3 offers a detailed technical description of the application of factor analysis to our data, discussing the problems encountered during the process and considering the different possible resulting combinations, or factor solutions. Finally, the calculation of factor scores for the definitive factor solution is explained in Section 4.

2. Introducing factor analysis

As has already been forwarded in Chapter 1, the purpose of factor analysis is to reduce a large amount of variables into a small number of variables, or factors (Gorsuch 1983; Tabachnick & Fidell 1996; Fabrigar & Wegener 2012). Each factor represents an aspect or area in the original data that can be explained through a small set of variables on the basis of co-occurrence, or 'shared variance' (Biber 1988: 79; Lee 2000: 158). In linguistic studies, the statistical term *variance*[31] refers to "the variability among individual texts for a given lexicogrammatical feature" (Lee 2000: 168). *Shared variance*, therefore, indicates in this case shared variability, or co-occurrence among the linguistic features. Such techniques of data reduction usually help to explain certain complex phenomena in a more simplified way, which is why factor analysis is widely used in psychology and sociology, but nowadays also in linguistics. In multidimensional analyses that follow Biber's (1988) model, factor analysis is used to uncover register variation by analysing the frequencies of co-occurrence of several linguistic features in a large number of texts.

Back to the technical description, factor analysis results from correlations among the different variables. Biber (1988: 79) explains the concept of correlation in the following simplified way. Let us suppose that we have retrieved four linguistic features from an imaginary corpus: first person pronouns, questions, passives and nominalisations, and then have created a dataset with normalised frequency counts for each of those features in each text. A possible correlation matrix resulting from that dataset could be:

Table 5.1

	1st pers. pro.	**questions**	**passives**	**nominalisations**
1st pers. pro.	1.00			
questions	.85	1.00		
passives	-.15	-.21	1.00	
nominalisations	.08	-.17	.90	1.00

Biber's (1988: 79) simplified example of correlation matrix

[31] See Section 2.1 for definitions of common vs. unique variance.

As can be seen on Table 5.1, some correlations are quite large whereas others are not. Biber (1988: 79) explains that "[t]he size of a correlation (whether positive or negative) indicates the extent to which two linguistic features vary together". A large positive correlation indicates that two features tend to co-occur in the texts. Conversely, a large negative correlation means that two features vary in a complementary pattern: that is, when one feature occurs, the other does not, and vice versa. If we square the correlation coefficient (R-squared), we will obtain the percentage of variance shared by the two features, which is a direct indication of how importantly they are related. For instance, if correlations range between 0 and 1, the high correlation of .85 between questions and first person pronouns, when squared, gives an R-squared value of .72, which means that the frequency values for questions and first person pronouns share 72 per cent of the variance. This means that, when first person pronouns occur in a text, questions are also very likely to occur, whereas the texts where first person pronouns do not occur, questions are also unlikely to appear.

Similarly, passive constructions and nominalisations have a positive correlation of .90, which give an R-square of .81, indicating that the frequency values of these two linguistic features share 81 per cent of the variance. However, nominalisations and questions have a negative correlation of –.17, which means that they share only 34 per cent of the variance. In turn, passive constructions and questions have a somewhat higher correlation of –.21, indicating that the amount of shared variance between the two features is 42 per cent. We can see thus that, while passives and nominalisations, and questions and first person pronouns are highly correlated, the rest of the correlations – between passives and first person pronouns, passives and questions, nominalisations and questions, and nominalisations and first person pronouns – are rather low. Biber (1988: 80) identifies two distinct factors from this hypothetical correlation matrix: Factor A, which has first person pronouns and questions, and Factor B, which has nominalisations and passive constructions. These factors are quite independent, or uncorrelated with one another, in that the linguistic features in Factor A do not correlate with those in Factor B.

Even so, it should be noted that, in language studies, perfectly uncorrelated factors are rare, if not impossible, since the different linguistic features are usually related among themselves to some extent, taking into account that most linguistic features have more than just one discursive function. In a factor analysis, factors are defined on the basis of frequently co-occurring linguistic features, so each factor has a group of such features. Normally, a feature co-occurs very frequently with certain features, but may also (though less frequently) appear in the company of another group of features, while never or rarely coinciding with other sets. In function of its frequency of co-occurrence in each case, this feature will have a different weight, or 'loading', on each factor, and it will have more weight (or a larger loading) on a factor where the features with which it frequently co-occurs also have more weight. Thus, the features with the largest loadings are the ones that 'define' a factor. They appear together because they have a common underlying function, and the factor can then be interpreted as a dimension of variation defined by this function. Thus, although our two imaginary factors, A and B, resulting from the correlation matrix on Table 5.1, have only two features defining each of them, each factor could potentially be interpreted as a dimension of variation.

The above example is a very simplified version of a correlation matrix and of how factors are computed. In reality, just as is the case with the application of any other statistical technique, the dataset in question needs to meet certain criteria in order to be regarded as suitable for factor analysis. On the other hand, the term itself – factor analysis – is, in reality, a generic one, because, rather than designating a method, it refers to a whole "family of related procedures" (Lee 2000: 164). The following paragraphs discuss some basic issues that should be considered before conducting a factor analysis, most of which are related to the properties of the data sample analysed. They also briefly discuss the decisions that should be made when choosing the right methods for factor extraction and rotation, as well as when deciding how many factors to extract.

2.1. Size of the dataset

As explained in Tabachnick & Fidell (1996: 640), the correlation coefficients tend to be less reliable when the sample size is small, which is why they recommend that the dataset should contain at least 300 cases for factor analysis, and the same is advised by Norušis (2005). This, however, is one of the many rules of thumb recommended in factor analysis and multivariate statistics textbooks. Gorsuch (1983) and Bryant & Yarnold (1995), for instance, advocate for a five-to-one ratio of participants (in our case, that would be texts) to measured variables (here, linguistic features), while others have even recommended a ten-to-one ratio (Everitt 1975; Nunnally 1978).

Neither of these requirements is met by our data, mainly because of the limitations of the number of samples of our corpus[32] (i.e. only a total of 122 texts are currently available for analysis). However, Kline (1994) suggests that a minimal two-to-one ratio may be enough for an exploratory factor analysis. Fabrigar & Wegener (2012: 26) explain that an adequate sample size depends on certain properties of the data and the model being fit, which means that establishing a fixed ratio of participants to measured variables may either "greatly exaggerate" or "badly underestimate" the required sample size, depending on each case. Thus, when the data present high communalities (.7 and greater; see Section 2.2) and when three-to-five measured variables load on each factor (see Section 3), a dataset of 100 cases may be sufficient (Fabrigar et al. 1999).

2.2. Preliminary tests

Two preliminary statistical tests are normally conducted before running a factor analysis. The first one is the Kaiser-Meyer-Olkin measure of sampling adequacy (KMO MSA; see Kaiser 1970, 1974). As explained in Lee (2000: 177), the MSA is "a measure of the degree to which the variables belong together and are thus appropriate for factor analysis". Kaiser (1974) qualifies MSAs of .90 and greater as 'marvelous', of .80 as 'meritorious', of .70 as 'middling', of .60 as 'mediocre' and of .50 as 'miserable', below .50 being 'unacceptable' (see Norušis 1988: 129; in Lee 2000: 177). MSAs below .60 indicate thus that the data is unsuitable for factor analysis.

The usual output when we run this test with our statistical package is an overall MSA for the whole dataset, followed by a table of individual MSAs for each variable in the dataset. Thus, a low MSA for a particular variable also suggests that this variable might be removed from the dataset and the analysis should be repeated.

The other basic test that is usually performed before running a factor analysis is Bartlett's Test of Sphericity (Snedecor & Cochran 1989), which checks whether the correlations among the data are 0 (that is, that the variables in the dataset are perfectly uncorrelated). In principle, this test requires a normal distribution, which is not the case with our data (see Section 2.3). However, Lee (2000: 176) advocates using this test "as an indicator of inappropriateness", that is, as an initial check whether the data

[32] It should be noted, however, that the actual size of the *Coruña Corpus* (ca. 1,200,000 words) is not much smaller than the total corpus used by Biber in his 1988 study (approx. 1,500,000 words; see Chapter 1), the difference being that Biber's (1988) corpus contains much smaller text samples.

is *un*suitable for factor analysis. According to Norušis (1988: 128, in Lee (2000: 176), if the value of the test for sphericity is large and the significance level is small ($p < .005$), this means that the null hypothesis can be rejected and that, consequently, the data is *not inappropriate* for factor analysis.

Apart from this, after running a factor analysis it is important to take into account the communalities (expressed as h2), or the proportion of variance for a variable that is explained by the common factors, or the factors extracted (Fabrigar & Wegener 2012: 7) and is equal to the sum of the squares of the factor loadings for that variable (Lee 2000: 178). When a variable shows a high communality in a particular factor solution (e.g. > .7), it means that the proportion of variance explained by all the common factors for that variable (or common variance) is large enough and the variable fits well in the model. Any leftover variance which is not explained by the extracted factors is the so-called unique variance, or uniqueness (expressed as u2) – the variance which is 'unique' to that variable only, and is inversely proportional to communality (h2 = 1 – u2). A successful factor analysis, thus, would have variables with high communalities and low uniqueness values, which would mean that a large proportion of the variance in the dataset is explained by the common factors (Grieve 2016: 215). In general, the more factors are extracted, the greater the communalities. If communalities are low, or weak, when a few factors are extracted, more factors need to be extracted so that a greater amount of variance can be accounted for by the final solution.

2.3. Factor extraction methods

As has been mentioned earlier, there are several methods for doing factor analysis. Commonly used extraction methods include principal components analysis (PCA), principal axis factoring (PAF, also called principal factor analysis, or common factor analysis), maximum likelihood, alpha factoring, or image factoring, among others, most of them being included in statistical packages such as SPSS or R. However, there is a primary distinction between PCA and factor analysis in its many variations (often abbreviated as FA in the literature). One distinction between the two methods is that factor analysis yields factors, while PCA outputs components, and, although both terms are often used interchangeably in the literature, they are statistically different (see Tabachnick & Fidell 1996: 635, 637; Lee 2000: 166-168). The main difference between the two methods is in the variance analysed. While PCA analyses all the variance in the data (including unique and error[33] variance), factor analysis analyses only the shared variance. Biber (1988: 82, footnote) remarks that treating unique and error variance as if they were shared variance may produce inflated factor loadings, which is why he considers that solutions produced by factor analysis are better.

Among the several variants of factor analysis, some require a normal distribution of the data (such as maximum likelihood analysis), whereas others do not (such as principal axis factoring, or PAF). Grieve (2016: 216) points out that the differences in the results from using one or another method should be minimal, provided the correlations between variables are high, and that traditionally maximum likelihood is preferred because it is statistically stronger (Everitt & Hothorn 2011; Fabrigar & Wegener 2012). However, because multivariate normality is particularly difficult to achieve in language studies (Lee 2000: 165), and especially if the analysed sample is relatively small, the second method (i.e. PAF) is usually chosen in multidimensional studies following Biber (1988) (see Friginal & Hardy 2014).

2.4. Factor rotation

Rotation is a technique used after the extraction of the factors in order to maximise higher correlations and minimise low ones (Tabachnick & Fidell 1996: 647). This is done with the purpose of transforming a complicated matrix into a simpler one, where the factors are more clearly differentiated from one another and are, thus, easier to interpret (Lee 2000: 188). Although rotation does not change the total percentage of variance accounted for by a factor solution, it "redistributes" this variance and, therefore, changes the percentage of variance accounted for by the individual factors (Norušis 1988: 140). In this study, proportional variance will always be given for rotated solutions.

There are two types of rotation: *orthogonal* rotation and *oblique* rotation. The former is based on the assumption that the factors are completely uncorrelated, or completely independent from one another. The most widely used method for orthogonal rotation is Varimax (Kaiser 1958). Conversely, oblique rotation yields a structure where factors have minor correlations among one another. Oblique rotation can be implemented through a variety of rotation methods, among which Promax

[33] The term *error variance* refers to variance resulting from random or measurement errors (Lee 2000: 166).

(Hendrickson & White 1964) has been frequently used, although relatively recently Oblimin rotation (Jackson 2005) has also become available. Biber (1988: 85 and footnote) advocates for the use of oblique, rather than orthogonal, rotation:

> In the description of textual variation, where the factors represent underlying textual dimensions, there is no reason to assume that factors are completely uncorrelated, and therefore a Promax rotation is recommended […] In fact, oblique solutions might be generally preferable in studies of language acquisition, since it is unlikely that orthogonal, uncorrelated factors actually occur as components of the communication process. That is, from a theoretical perspective, all aspects of language use appear to be interrelated to at least some extent, and thus there is no reason to expect mathematically uncorrelated factors representing those aspects.

Although Biber (1988) recommends using Promax, nowadays Oblimin appears to be the preferred method for oblique rotation (Grieve 2016: 217). In principle, just as it is the case with different methods of extraction, the different methods of rotation will not change a factor solution greatly if the dataset is good (Tabachnick & Fidell 1996; Grieve 2016). As pointed out in Lee (2000: 193), it is generally recommended to try out both types of rotation and, if the results are very similar, choose the simplest (i.e. usually, the orthogonal) solution. Several trials in this study have nevertheless shown that oblique solutions (with Oblimin rotation) offer somewhat clearer structures which are easier to explain (see Section 3).

2.5. How many factors to extract?

This is, possibly, the trickiest part of factor analysis. As quoted in Mohamed (2011: 145), Hair et al. (1998: 103) explain that

> choosing the number of factors to be interpreted is something like focusing a microscope. Too high or too low an adjustment will obscure a structure that is obvious when the adjustment is just right. Therefore, by examining a number of different factor structures derived from several trial solutions, the researcher can compare and contrast to arrive at the best representation of the data.

In a factor analysis, the first factor extracted accounts for the greatest amount of variance, and then the second factor is extracted from the leftover variance and explains the second greatest amount, and so on. Virtually, one can keep extracting factors until all the shared variance is accounted for; however, this would not make much sense because, instead of reducing the data, such an analysis would yield as many factors as there are variables in the dataset (Lee 2000: 180). Grieve (2016: 218-219) highlights three important points to be considered at the moment of factor extraction. First of all, one should extract a sufficient number of factors in order to minimise the number of weak communalities (e.g. < .30 or < .20). If, after an initial extraction, there are many variables that have weak communalities, more factors need to be extracted so that a greater portion of the shared variance in the data matrix is accounted for. Secondly, after rotation, all the extracted factors should load variables strongly (e.g. loadings > .70) and most variables should be loading on one factor. Variables loading weakly and/or most variables loading across many factors are indications of over-factoring (i.e. when too many factors have been extracted). Thirdly, Grieve (2016: 219) insists that "a successful factor analysis should account for a relatively large amount of variance in the values of the variables in the data matrix". This means that the cut should be made in a way that adding more factors would not provide a great amount of additional variance (i.e. would not improve the model greatly).

A widely used method to choose the number of factors and which particularly helps to check this third condition is the so-called scree test, first developed by Cattell (1966). However, like any other graphic illustration of data, a scree plot is better understood through a practical example with real data and will therefore be explained in the next section, which describes each of the steps taken for conducting an exploratory factor analysis on our data.

3. Running a factor analysis

Before determining the final set of lexical and grammatical variables to be used in the final factor analysis, a series of preliminary factor analyses were run with different combinations of linguistic features. The first factor analyses were performed on the initial fifty-eight variable set, using Principal Axis Factoring (PAF), and extracting three, four, five, six, and seven factors. Three different rotation methods (Varimax, Promax and Oblimin) were tested. Although the differences in the results were minimal, Oblimin was the preferred method, both for the arguments offered in favour of oblique rotation by Biber (1988) (see Section 2.4) and also because solutions with Oblimin rotation were easiest to interpret.

The statistical package used for this study was R 3.0.1 for Mac. Following Lee (2000: 279), it should be noted that there are many factor pattern matrices and other tables containing statistical output (e.g. MSA, communalities, proportional and cumulative variance, etc.), resulting from the run of each factor analysis. Although the goal of maximum transparency of results requires that all of them should be published, the large number of 'test' analyses makes this somewhat impractical. It has been therefore decided to include in Appendix IV only the output of those analyses that are directly reported here, the rest being available upon request from the author.

Before running the first factor analyses on fifty-eight linguistic features, the two preliminary statistical tests described earlier (i.e. the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO MSA) and Barlett's Test of Sphericity) were run in order to see whether our dataset was adequate for a factor analysis. In our case, KMO = .68 indicates that the minimum requirement for factor analysis (values above .60) is met (Tabachnick & Fidell 2007; see Gray 2011: 135), although for Kaiser (1970) it is still a 'mediocre' tending to 'middling' value. On the other hand, Bartlett's Test for Sphericity (ChiSquare = 4693.241, df =1653, p < .001) is significant, showing that the null hypothesis that the correlation matrix is an identity matrix (i.e. that all variables are perfectly uncorrelated with one another) can be rejected and meaning, thus, that our dataset is *not inappropriate* for factor analysis (see Section 2.2).

In order to decide on the right number of factors to extract, a scree[34] plot of eigenvalues was used (see Figure 5.1):

[34] The 'scree' on the scree plot refers to what Cattell (1966) also called "factorial littler", or anything that is beyond the point where the slope starts to level off (see Kim & Mueller 1978: 44).

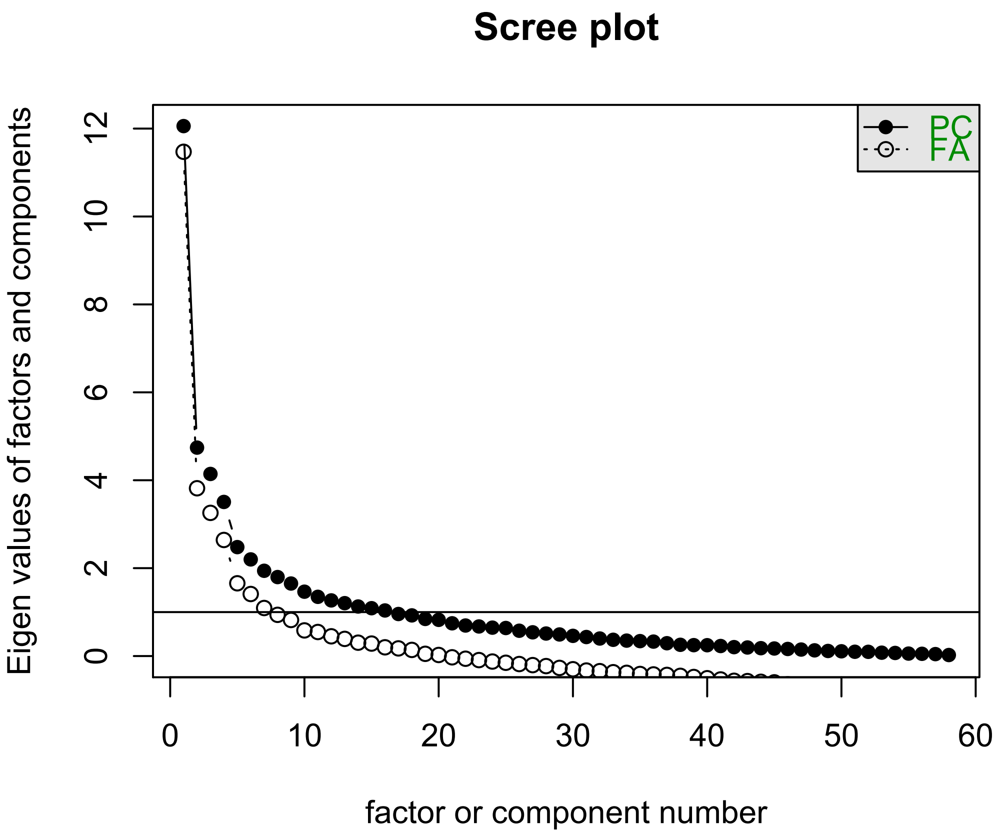


Figure 5.1
Scree plot for a fifty-eight variable correlation matrix

As can be seen on Figure 5.1, the y-axis represents eigenvalues, while the x-axis represents factor (or component) numbers. Without getting into more complex mathematical definitions, eigenvalues are "direct indices of the amount of variance accounted for by each factor" (Biber 1988: 82). A scree plot can be useful both for factor analysis (FA) and principal components analysis (PC), which is why both components (black dots) and factors (white dots) are plotted here. The factor with the largest eigenvalue (Factor 1) accounts for the largest amount of variance; the second factor (after the first fall) accounts for the second largest proportion of variance, and so on (Tabachnick & Fidell 1996: 646). Traditionally, only eigenvalues greater than 1 are selected because factors with very small eigenvalues explain little to no variance and would be therefore meaningless. The "eigenvalue > 1" criterion is sometimes used as an indicator of the 'cutting point' between factors, appearing as a line by default in most statistical packages. In Figure 5.1, the "eigenvalue > 1" line would suggest that six or even seven factors should be extracted. However, this way of selecting the number of factors is rather arbitrary and may not be of great help when interpreting the factors. Generally, it is recommended to cut the number of factors *before* the last major drop on the scree plot, or before the slope on the plot begins to level off (Hatcher 1994, Stevens 2002). According to Grieve (2016: 221), if each factor explains a particular amount of variance, then all the following factors accounting for the same amount of variance should also be extracted. In this particular case, the last major break seems to be the one before the fourth and fifth factors, indicating that a four-factor solution is the best one.

However, if we keep examining Figure 5.1, we can also see that a second (smaller) break occurs between the sixth and seventh factors. Biber (1988: 86) cites Gorsuch (1983) and Farhady (1983), recommending “the more conservative procedure”, which is to select the largest number of factors in order to avoid loss of information as a result of under-factoring, or extracting too few factors. This is why Biber (1988) ended up extracting seven factors in his final analysis, which was criticised by Lee (2000: 286) as over-factoring for a number of reasons, some of which will be discussed in the next sections. If we agreed, then, that the choice to extract the maximum number of factors is the safest, a six-factor solution would be the ideal one.

In any case, there does not seem to be a rule that says where to make the cut, and it is generally agreed that the scree plot is “only a heuristic [sic][35], and somewhat subjective” (Lee 2000: 184). At the end of the day, the final call for selecting the number of factors belongs to the researcher, in the light of his/her own judgment whether the factorial structure makes sense. This is what Lee (2000: 186) calls the principle of interpretability, which entails that “the solution which best explains the data wins”. In fact, interpretation can be critical at the time of choosing among different factor solutions, and the most statistically reliable solution may not be the most convincing one as far as background theory is concerned. In the present study, several different factor solutions have been tried with different datasets, some of which shall be discussed in what follows.

[35] It appears that Lee uses here the adjective *heuristic* as a noun, as an abbreviated use of the term *heuristic technique*.

3.1. Three-factor solution with fifty-eight variables

Table 5.2 shows a rotated factor pattern matrix for a solution with three factors, extracted with the PAF method and rotated with Oblimin:

Table 5.2

	PA1	**PA2**	**PA3**	**h2**	**u2**
PAST	0.24	0.11	0.07	0.0685	0.93
PERF	0.47	0.36	0.30	0.4134	0.59
PRES	0.09	0.08	-0.55	0.3205	0.68
PL_ADV	-0.37	-0.06	-0.21	0.1778	0.82
TIM_ADV	0.19	0.10	0.11	0.0544	0.95
FPERS	0.72	-0.01	-0.02	0.5207	0.48
SPERS	0.17	-0.16	-0.23	0.1139	0.89
TPERS	0.41	-0.05	-0.43	0.3704	0.63
ITPRO	0.47	0.06	-0.42	0.4122	0.59
DEMPRO	0.48	-0.01	-0.13	0.2565	0.74
INDPRO	0.78	0.00	-0.20	0.6544	0.35
PRO_DO	0.37	-0.34	-0.31	0.3746	0.63
QUEST	0.61	-0.06	-0.14	0.4007	0.60
NOM	0.42	0.05	0.55	0.4547	0.55
NOUN	-0.78	-0.01	-0.13	0.6089	0.39
AGPASS	0.02	0.03	0.30	0.0883	0.91
BYPASS	0.14	0.25	-0.12	0.0926	0.91
BE_MAIN	0.32	-0.20	-0.21	0.1991	0.80
EXTHERE	0.51	-0.06	-0.16	0.3007	0.70
THAT_V	0.48	0.06	0.26	0.2864	0.71
THAT_ADJ	0.45	0.13	0.32	0.3034	0.70
WHCL_SUB	0.49	-0.01	0.00	0.2399	0.76
WHCL_OB	0.71	-0.20	-0.15	0.5965	0.40
TO_INF	0.76	-0.04	0.05	0.5848	0.42
PASTPART	-0.33	0.19	0.00	0.1527	0.85
WHIZ	-0.24	-0.06	0.52	0.3409	0.66
PRES_WHIZ	-0.47	-0.13	0.13	0.2485	0.75
THAT_SUB	0.14	-0.21	-0.10	0.0804	0.92
WHREL_SUB	0.21	0.26	-0.01	0.1046	0.90
WHREL_OB	0.57	0.13	0.03	0.3356	0.66
PIP	0.25	0.09	0.26	0.1277	0.87
SREL	0.06	-0.35	-0.24	0.1945	0.81

Table 5.2 (continued)

	PA1	**PA2**	**PA3**	**h2**	**u2**
CAUSADV	-0.05	-0.56	0.02	0.3117	0.69
CONCADV	0.38	0.36	-0.14	0.2882	0.71
CONDADV	0.39	-0.51	0.04	0.4286	0.57
OTHADV	0.17	-0.31	0.06	0.1322	0.87
PREP	-0.15	-0.05	0.67	0.4779	0.52
ATTRADJ	-0.04	0.66	0.32	0.5608	0.44
PREDADJ	-0.01	0.32	-0.20	0.1401	0.86
ADV	0.52	0.13	-0.17	0.3197	0.68
CONJ	0.03	-0.55	0.48	0.5203	0.48
DOWN	0.21	0.44	0.28	0.3147	0.69
HEDG	-0.14	0.64	0.02	0.4368	0.56
AMPL	0.13	0.48	-0.21	0.2849	0.72
DEM	0.52	0.08	-0.08	0.2809	0.72
POSSMOD	0.81	-0.07	0.18	0.6773	0.32
NECMOD	0.61	-0.17	0.19	0.4297	0.57
PREDMOD	0.09	-0.71	0.26	0.5748	0.43
PUBV	0.66	0.01	-0.10	0.4556	0.54
PRIVV	0.72	-0.11	0.12	0.5375	0.46
SUASV	0.53	0.16	0.14	0.3125	0.69
SEEM	-0.01	0.07	0.01	0.0055	0.99
SPLITINF	-0.06	-0.02	0.15	0.0256	0.97
SPLITAUX	0.46	0.38	0.29	0.4193	0.58
PHCOORD	0.11	0.23	-0.34	0.1787	0.82
CLCOORD	0.39	-0.16	0.00	0.1811	0.82
SNEG	0.75	-0.04	-0.17	0.6041	0.40
ANEG	0.88	0.03	-0.07	0.7800	0.22

3-factor solution for 58 features (PAF, Oblimin)

The first column lists the fifty-eight linguistic features (see key in Appendix I), while the second, third and fourth columns present the loadings of each feature on Factors 1, 2 and 3, respectively. (Columns 4 and 5 will be explained a bit later.) As has been explained earlier, a factor loading indicates the strength of the relationship of a particular linguistic feature with a factor, that is, “the extent to which the variation in the frequency of that feature correlates with the overall variation of the factor” (Biber

1988: 87). A cut-off point of .30 has been established for factor loadings, those below that number being disregarded as too low. Important, or, in Biber's (1988: 87) terms, 'salient' loadings (or loadings > .30) are marked in grey. The darkest grey shade marks the largest loading, to indicate that a feature is particularly related to a given factor, whereas the lighter hue marks the 'second largest' loadings (if they are also > .30), which is the case of features loading on more than one factor. As is usually the case, most features load on Factor 1, while only a few load on Factors 2 and 3. This is quite logical because the first factor always accounts for the largest amount of proportional variance (in this case, 20 per cent), whereas each subsequent factor explains a proportion of the additional variance. In total, this three-factor solution explains 33 per cent of the total amount of variance (see proportional and cumulative variance, as well as other details, in the full factorial pattern matrices in Appendix IV).

Factor loadings can be positive or negative. As explained in Biber (1988: 87-88), a positive or negative sign does not mean that a loading is more or less important; rather, positive and negative loadings indicate groups of features that present a complementary distribution in the texts. For instance, here on Factor 1, some features such as indefinite pronouns, questions, first person pronouns, public, private and suasive verbs, analytic and synthetic negation, WH-clauses, *to* infinitives, existential *there*, possibility and necessity modals, adverbs, clausal coordination and copular *be,* all have positive loadings, while nouns, place adverbs, past participle clauses and present participial WHIZ-deletions have negative loadings, which means that these two groups of features occur in a complementary pattern and are unlikely to appear together. In more technical terms, they represent the two ends of a dimension and have therefore opposite discursive functions. In this case, from the features that we have just listed, those forming the first group seem to be related to an involved, possibly spontaneous kind of discourse, while those in the second group appear to indicate a high concentration of information. Initially, the two opposite functions represented in Factor 1 could be interpreted as "involvement vs. informational density", which may remind us of Biber's (1988) Dimension 1 ("Involved vs. Informational Production"), being also similar in Lee (2000) and Gray (2011). However, certain positive features such as analytic and synthetic negation, questions, possibility modals, necessity modals and suasive verbs – all of which have quite large loadings (> .50) – also suggest an argumentative function. Thus, the positive end of Factor 1 might be described as conveying “involved argumentation”.

Factor 2, in turn, groups attributive adjectives, downtoners, hedges, amplifiers and predicative adjectives on one side, and prediction modals, causative and conditional adverbs, conjuncts and sentence relatives on the other. While downtoners and hedges could be considered to form a group of their own, contributing to a ‘cautious’, non-assertive kind of discourse (see, for instance, Hyland 1995, 1998a, 1998b), adjectives may also appear premodified by either downtoners or amplifiers, suggesting subjective description or evaluation. On the other hand, causative and conditional adverbs, as well as conjuncts and sentence relatives appear to mark logical relationships between clauses, while predictive modals of intention or volition may function as indicators of certainty and objectivity.

Finally, Factor 3 has prepositions, nominalisations, past participial WHIZ-deletions, conjuncts and agentless passives with a positive sign, and, on the other end, present tense, third person pronouns, phrasal coordination, as well as the pronout *it* and the pro-verb *do* with a negative sign. Although these two last features load more strongly on Factor 1, their smaller yet still salient loadings on Factor 3 may also be considered important in that, to some extent, they help to understand the underlying construct of this factor, which is why, initially, all the features with salient loadings are taken into account when interpreting the factors.[36] If we go back to Biber’s (1988) study (see Chapter 1 Section 3), the first subset of features may remind us of Dimension 5 which measures abstract or impersonal style, with nominalisations also suggesting a formal, elaborated kind of register (Gray 2011; Bello 2014). Likewise, prepositional phrases are elements of phrasal embedding which contribute to structural elaboration (Gray 2011; Gray & Biber 2012). By contrast, phrasal coordination and the pro-verb *do* are units of integration and syntactic reduction (the pro-verb *do* functioning as a substitute for a finite clause), while third person pronouns and present tense appear to indicate explicit reference to ‘someone who does something now’. This second group of features, therefore, seems to transmit a more personal and presumably more spontaneous type of discourse.

[36] Following Gorsuch (1983), Biber (1988: 93) only counted the features presenting the largest loadings on a factor for the calculation of factor scores, discarding any ‘second largest’ loadings (see discussion in Section 4 on the computation of factor scores).

Although in the light of the above-suggested interpretation a three-factor solution may seem, in principle, acceptable, it is still statistically weak. First of all, as has been noted above, this three-factor solution explains only 30 per cent of the total variance, suggesting that these findings are not particularly reliable. Furthermore, if we go back to Table 1, there are two more columns (four and five) containing the communalities (h2) and uniquenesses (u2) for each feature, respectively. As has been explained in the previous section, communality is inversely proportional to uniqueness (h2 = 1 – u2). When a feature presents a low communality (< .30) in a factor solution, it means that the amount of shared variance explained by it in this particular factor analysis is not enough, because this feature has a high unique variance. In this case, we can see that communalities are pretty low for twenty-four variables, which means that more than a third of the features selected for the analysis – including past tense, time adverbials, predicative adjectives, or amplifiers – are not adequate for this solution. For instance, past tense presents a communality of .17, which means that less that 20 per cent of its shared variance is explained by the three factors extracted, and we can also see that it has no salient loadings on any of those three factors. In principle, all such features should be dropped and the analysis should be rerun.

Another weakness of the present three-factor analysis is a relatively low overall MSA (KMO = 6.8). The MSA is not related to a particular factor solution, but to the dataset in question, which, while it meets the minimum requirements for factor analysis, is still considered to be ‘mediocre’-to-‘middling’ in terms of sampling adequacy. In addition, following Lee (2000: page), individual MSA scores have also been calculated for each feature (see Table5.3, next page):

Table 5.3.

PAST	PERF	PRES	PL_ADV	TIM_ADV	FPERS	SPERS
0.31	0.72	0.33	0.74	0.67	0.74	0.32
TPERS	ITPRO	DEMPRO	INDPRO	PRO_DO	QUEST	NOM
0.60	0.68	0.77	0.89	0.76	0.80	0.73
NOUN	AGPASS	BYPASS	BE_MAIN	EXTHERE	THAT_V	THAT_ADJ
0.80	0.42	0.48	0.59	0.72	0.60	0.79
WHCL_SUB	WHCL_OB	TO_INF	PASTPART	WHIZ	PRES_WHIZ	THAT_SUB
0.72	0.82	0.86	0.66	0.63	0.72	0.36
WHREL_SUB	WHREL_OB	PIP	SREL	CAUSADV	CONCADV	CONDADV
0.52	0.72	0.66	0.64	0.67	0.67	0.75
OTHADV	PREP	ATTRADJ	PREDADJ	ADV	CONJ	DOWN
0.59	0.60	0.59	0.47	0.73	0.45	0.51
HEDG	AMPL	DEM	POSSMOD	NECMOD	PREDMOD	PUBV
0.62	0.60	0.68	0.84	0.70	0.56	0.77
PRIVV	SUASV	SEEM	SPLITINF	SPLITAUX	PHCOORD	CLCOORD
0.71	0.81	0.37	0.25	0.66	0.46	0.69
SNEG	ANEG					
0.84	0.90					

Individual MSAs for 58 linguistic feature

Again, MSAs below .60 indicate that the feature in question – such as past tense, present tense, second person pronouns, agentless and *by* passives, *be* as main verb, *that*-clauses in subject position, WH-relativiser in subject position, other adverbial subordinators, attributive and predicative adjectives, conjuncts, downtoners, predictive modals, *seem* and *appear*, split infinitives and phrasal coordination – does not quite belong in the correlation matrix, especially in those cases where the MSA is very low (such as, for instance, split infinitives, which has a MSA of .25). This once more suggests that some (if not all) of these features should be eliminated and the analysis should be repeated.

However, any deletion must be done with caution because some features with low communalities and mediocre MSA may still load on one of the factors and their potential exclusion may result in loss of information at the time of interpreting a factor. For example, amplifiers present a communality < .30, but still have a salient loading on Factor 2 (.48) and seem to contribute to the underlying function of this factor (possibly, "subjective evaluation or description") along with some of the other features loading on it, such as downtoners, hedges and attributive and predicative adjectives. Likewise, agentless passives have a MSA < .60, but their contribution to Factor 3 seems meaningful in the light of other features with stronger loadings on that factor, such as nominalisations or past participle WHIZ-deletions.

In any case, unlike the overall and individual MSAs, communalities may increase in function of the number of factors extracted, and this entails that some features with unimportant loadings may start showing salient loadings once more factors are extracted. In the following paragraphs a four-factor solution with the initial fifty-eight variables will be discussed.

3.2. Four-factor solution with fifty-eight variables

Table 5.4 shows a factor pattern matrix for a four-factor solution, the factors being extracted once more with the PAF method and rotated with Oblimin:

Table 5.4

	PA1	**PA2**	**PA3**	**PA4**	**h2**	**u2**
PAST	0.20	0.04	-0.05	0.37	0.180	0.82
PERF	0.40	0.24	0.21	0.48	0.512	0.49
PRES	0.18	0.22	-0.38	-0.39	0.380	0.62
PL_ADV	-0.37	-0.09	-0.43	0.32	0.418	0.58
TIM_ADV	0.14	-0.02	-0.15	0.67	0.485	0.51
FPERS	0.71	-0.02	-0.03	0.12	0.528	0.47
SPERS	0.19	-0.15	-0.34	0.15	0.190	0.81
TPERS	0.46	0.02	-0.39	-0.08	0.371	0.63
ITPRO	0.52	0.15	-0.32	-0.17	0.416	0.58
DEMPRO	0.50	0.00	-0.12	0.04	0.259	0.74
INDPRO	0.80	0.02	-0.15	0.01	0.654	0.35
PRO_DO	0.42	-0.28	-0.33	-0.09	0.375	0.62
QUEST	0.63	-0.02	-0.04	-0.15	0.422	0.58
NOM	0.37	0.00	0.74	-0.16	0.698	0.30
NOUN	-0.75	0.03	-0.14	-0.12	0.607	0.39
AGPASS	-0.02	0.00	0.37	-0.09	0.138	0.86
BYPASS	0.16	0.29	0.00	-0.13	0.115	0.88
BE_MAIN	0.39	-0.08	0.01	-0.55	0.459	0.54
EXTHERE	0.53	-0.06	-0.20	0.12	0.328	0.67
THAT_V	0.43	-0.03	0.18	0.30	0.318	0.68
THAT_ADJ	0.38	0.03	0.26	0.29	0.318	0.68
WHCL_SUB	0.50	0.02	0.13	-0.18	0.296	0.70
WHCL_OB	0.73	-0.20	-0.21	0.11	0.627	0.37
TO_INF	0.76	-0.04	0.13	-0.04	0.598	0.40
PASTPART	-0.33	0.22	0.10	-0.17	0.194	0.81
WHIZ	-0.31	-0.15	0.47	0.10	0.340	0.66
PRES_WHIZ	-0.49	-0.18	-0.03	0.21	0.309	0.69
THAT_SUB	0.16	-0.19	-0.14	-0.01	0.084	0.92

Table 5.4 (continued)

	PA1	PA2	PA3	PA4	h2	u2
WHREL_SUB	0.20	0.26	0.06	-0.01	0.108	0.89
WHREL_OB	0.55	0.11	0.06	0.11	0.335	0.66
PIP	0.21	0.06	0.32	-0.02	0.152	0.85
SREL	0.10	-0.31	-0.32	-0.02	0.210	0.79
CAUSADV	-0.03	-0.55	-0.10	-0.01	0.317	0.68
CONCADV	0.39	0.37	-0.08	0.08	0.288	0.71
CONDADV	0.41	-0.49	0.03	-0.14	0.437	0.56
OTHADV	0.16	-0.35	-0.08	0.18	0.184	0.82
PREP	-0.24	-0.17	0.60	0.18	0.476	0.52
ATTRADJ	-0.11	0.59	0.42	0.14	0.568	0.43
PREDADJ	0.02	0.40	-0.02	-0.25	0.212	0.79
ADV	0.53	0.09	-0.30	0.39	0.516	0.48
CONJ	-0.02	-0.62	0.35	0.08	0.510	0.49
DOWN	0.15	0.35	0.25	0.34	0.338	0.66
HEDG	-0.18	0.58	-0.01	0.34	0.512	0.49
AMPL	0.13	0.47	-0.24	0.27	0.372	0.63
DEM	0.52	0.07	-0.09	0.14	0.297	0.70
POSSMOD	0.78	-0.09	0.24	-0.02	0.695	0.31
NECMOD	0.59	-0.19	0.22	-0.02	0.442	0.56
PREDMOD	0.07	-0.77	0.06	0.14	0.612	0.39
PUBV	0.67	0.03	-0.06	0.01	0.454	0.55
PRIVV	0.69	-0.17	0.03	0.27	0.587	0.41
SUASV	0.52	0.17	0.30	-0.14	0.399	0.60
SEEM	-0.03	0.01	-0.14	0.36	0.143	0.86
SPLITINF	-0.08	-0.06	0.11	0.07	0.026	0.97
SPLITAUX	0.39	0.28	0.27	0.34	0.440	0.56
PHCOORD	0.17	0.36	-0.12	-0.40	0.314	0.69
CLCOORD	0.38	-0.18	-0.05	0.11	0.195	0.80
SNEG	0.78	0.01	-0.06	-0.14	0.622	0.38
ANEG	0.89	0.06	0.05	-0.11	0.807	0.19

4-factor solution for 58 features (PAF, Oblimin)

Just like in the three-factor solution, most features keep loading on the first factor, which is the one that explains the greatest amount of shared variance (20 per cent). The second, third and fourth factors explain 7, 6 and 5 per cent, respectively (total amount of shared variance by the four factors = 38 per cent). Most of the features loading on Factor 1 in the three-factor solution also load on this factor in the four-factor solution. This is the case of indefinite pronouns, synthetic and analytic negation, first person pronouns, *to*-infinitives, possibility and necessity modals, existential *there*, private and public verbs, suasive verbs, demonstratives, adverbs, WH-clauses, adverbs and clausal coordination, all of which are positive features. However, some other features, such as copular *be* (or. *be* as main verb) and perfect aspect, load more strongly on other factors, even if they keep loading on Factor 1 also. This is happening because additional factors usually reveal new underlying constructs that cannot be seen in a solution with fewer factors. In this case we can see that both perfect aspect and *be* as main verb are now loading on the different ends of Factor 4. The former is clustering with time adverbs, past tense, and *seem* and *appear*, that are features that did not load on any of the three factors in the previous analysis. On the other hand, copular *be* has formed a group with present tense and phrasal coordination, both of which were loading earlier on Factor 3. We could say thus that Factor 4 has taken a part of the underlying construct in a dimension from a previous factor solution on the negative pole (i.e. some of the negative features from Factor 3 in the three-factor solution, which were interpreted as conveying an informal, 'spontaneous' kind of discourse), while uncovering a new construct on the positive end, formed by features that suggest a narrative dimension (except for *seem* and *appear*, which is tricky to explain in this group as most occurrences actually correspond to present rather than past tense forms[37], but which, on the other hand, does not have much weight on the factor, with a loading of .36). Rather than "elaborated vs. spontaneous", thus, the two poles of Factor 4 suggest a "narrative vs. non-narrative" opposition, similar to Biber's (1988) Dimension 2.

As for the negative features on the present Factor 1, most coincide with the negative features of the first factor in the three-factor solution: nouns, present-participial WHIZ-deletions and past-participle clauses. However, the fourth negative feature – place adverbs – now loads more strongly on Factor 3, along with second person pronouns and sentence relatives, as well as with present tense (even though, as we have just seen, this feature loads stronger on Factor 4), and with third person pronouns and pronoun *it* (both of which have more salient loadings on Factor 1). By contrast, the positive features on Factor 3 – nominalisations, prepositions, past

[37] The "Frequency breakdown" option on CQPWeb allows to see the tokens classified by types in descending order of occurrence.

participial WHIZ-deletions, agentless passives and, to a lesser extent, conjuncts – replicate entirely the positive features on the third factor of the three-factor solution, with the addition of pied-piping constructions (which did not load on any factor previously). The correlation of pied-piping constructions with prepositions could have been expected, in that the former are WH-relative clauses introduced by a preposition, such as

(5.1) Consider whether the year *in which you seek the sun's place* is bissextile (astr16),

while their inclusion in the 'formal', or 'elaborated' group of features is also logical because they function as a marker of structural elaboration. In the present analysis, the two ends of Factor 3 appear, thus, to replicate the same underlying constructs ("elaborated vs. spontaneous") as in the third factor of the three-factor solution.

Finally, just like in the previous analysis, Factor 2 groups downtoners, hedges, attributive adjectives, amplifiers and predicative adjectives (and, now, concessive subordination) against predictive modals, conjuncts, causal subordination, conditional subordination and other adverbial subordinators. Here again, the two underlying functions in binary opposition ("subjective evaluation/description vs. assertive prediction") are very similar to those in the second factor in the three-factor solution. The newly-included concessive subordination, though not very strong, seems to fit perfectly in the first group, in that concessive clauses indicate the possibility of other options, being another marker of non-assertiveness. We have seen thus that the three tentative dimensions identified in the three-factor solution have been largely kept in the four-factor solution, with an additional fourth dimension apparently contrasting narrative and non-narrative discourse. According to the scree plot displayed on Figure 1, a four-factor solution is the one preferred for our fifty-eight variable dataset, although, following Biber's (1988: 82) advice regarding under-factoring, a six-factor solution could also be considered. Conversely, the three-factor solution might be regarded an example of under-factoring in that the three factors extracted do not allow certain features to load and reveal the other underlying construct which emerged with the additional factor.

If we go back to Table 5.4 and look at the communalities, we may observe that sixteen features still have a score < .3, namely, past tense, second person pronouns, demonstrative pronouns, agentless and *by*-passives, past participle clauses, *that* clauses on subject position, WH-relativiser on subject position, pied-piping constructions, concessive subordination, sentence relatives, other adverbial subordinators, predicative adjectives, *seem* and *appear*, split infinitives and clausal coordination. However, only four of these features (*by*-passives, *that*clauses in subject position, WH-relativiser in subject position and split infinitives) do not have salient loadings on any factor. In case the four-factor solution were kept as the final model, it was decided to drop the features with low communalities in two stages and run two comparative factor analyses with two new datasets – one with forty-two variables, deleting *all* the features with low communalities, and another with fifty-four variables, dropping only the four features with no salient loadings. Before that, however, additional factor analyses were run on the initial fifty-eight variable dataset, extracting five, six and seven factors.

The five-factor solution replicated the same structure as in the one with four factors, but the fifth factor did not seem reliable enough to be interpreted as a dimension of variation (see Appendix IV). Therefore, if the five-factor solution were kept for the final analysis, the fifth factor would have been discarded, resulting in another four-factor structure. On the other hand, it was also believed that, based on the scree plot (Figure 5.1), a six-factor solution could provide a better picture than a five-factor solution. In what follows, a six-factor solution will be analysed as a possible alternative to the four-factor solution, as well as a possible example of over-factoring.

3.3. Six-factor solution with fifty-eight variables

Table 5.5 shows a rotated factor pattern matrix, this time for six factors, extracted with the PAF method and rotated with Oblimin:

Table 5.5

	PA1	PA6	PA2	PA3	PA4	PA5	h2	u2
PAST	0.07	-0.03	0.04	-0.06	0.40	0.35	0.284	0.72
PERF	0.26	0.12	-0.25	-0.23	0.41	0.36	0.572	0.43
PRES	0.23	-0.16	-0.12	0.48	-0.46	0.00	0.510	0.49
PL_ADV	-0.06	-0.46	0.10	0.24	0.33	-0.11	0.415	0.58
TIM_ADV	0.19	-0.06	-0.02	0.07	0.67	0.01	0.503	0.50
FPERS	0.60	0.12	0.05	0.03	0.01	0.23	0.584	0.42
SPERS	0.39	-0.33	0.20	0.17	0.06	0.11	0.265	0.73
TPERS	-0.11	0.26	0.25	0.44	0.13	0.36	0.520	0.48
ITPRO	0.27	0.13	-0.01	0.42	-0.16	0.14	0.420	0.58
DEMPRO	0.38	0.14	0.01	0.19	-0.01	0.03	0.279	0.72
INDPRO	0.60	0.25	0.01	0.28	-0.07	0.06	0.710	0.29
PRO_DO	0.17	0.10	0.41	0.28	-0.01	0.14	0.381	0.62
QUEST	0.36	0.29	0.07	0.16	-0.17	0.11	0.429	0.57
NOM	0.06	0.60	-0.11	-0.53	-0.19	0.05	0.689	0.31
NOUN	-0.35	-0.48	-0.05	0.00	-0.13	-0.15	0.611	0.39
AGPASS	-0.22	0.38	-0.08	-0.22	-0.03	-0.13	0.195	0.80
BYPASS	-0.23	0.20	-0.13	0.07	-0.03	0.36	0.213	0.79
BE_MAIN	0.15	0.27	0.16	0.12	-0.54	0.03	0.474	0.53
EXTHERE	0.47	0.02	0.10	0.18	0.05	0.12	0.354	0.65
THAT_V	0.69	0.02	-0.15	-0.15	0.09	-0.14	0.488	0.51
THAT_ADJ	0.52	0.16	-0.22	-0.17	0.12	-0.16	0.413	0.59
WHCL_SUB	0.17	0.47	-0.02	0.09	-0.16	-0.06	0.331	0.67
WHCL_OB	0.62	0.11	0.23	0.19	0.04	0.11	0.661	0.34
TO_INF	0.21	0.55	0.12	0.01	0.02	0.25	0.633	0.37
PASTPART	-0.25	0.01	-0.26	0.02	-0.16	-0.19	0.221	0.78
WHIZ	0.06	-0.05	-0.09	-0.49	-0.03	-0.30	0.381	0.62
PRES_WHIZ	0.08	-0.42	0.01	-0.11	0.10	-0.37	0.399	0.60
THAT_SUB	0.16	0.03	0.19	0.13	0.00	-0.11	0.098	0.90
WHREL_SUB	-0.04	-0.03	-0.10	-0.16	-0.01	0.64	0.402	0.60
WHREL_OB	0.41	-0.03	0.00	-0.18	0.00	0.61	0.614	0.39
PIP	0.24	-0.03	-0.07	-0.42	-0.16	0.37	0.350	0.65
SREL	-0.16	0.11	0.44	0.27	0.15	0.00	0.291	0.71

Table 5.5 (continued)

	PA1	PA6	PA2	PA3	PA4	PA5	h2	u2
CAUSADV	0.15	-0.17	0.52	-0.10	-0.02	-0.11	0.328	0.67
CONCADV	0.06	0.24	-0.27	0.23	0.11	0.20	0.302	0.70
CONDADV	0.07	0.41	0.51	0.01	-0.03	-0.08	0.478	0.52
OTHADV	0.08	0.11	0.33	0.02	0.24	-0.09	0.204	0.80
PREP	-0.14	0.02	0.05	-0.76	0.12	0.14	0.609	0.39
ATTRADJ	-0.01	0.12	-0.73	-0.20	0.01	-0.07	0.597	0.40
PREDADJ	-0.16	0.09	-0.30	0.16	-0.22	0.14	0.209	0.79
ADV	0.28	0.26	-0.07	0.45	0.45	-0.09	0.648	0.35
CONJ	0.19	0.11	0.41	-0.43	0.02	-0.35	0.530	0.47
DOWN	0.27	0.11	-0.51	-0.09	0.20	-0.14	0.411	0.59
HEDG	-0.02	-0.11	-0.66	0.14	0.27	-0.08	0.556	0.44
AMPL	0.10	-0.02	-0.44	0.37	0.25	0.00	0.408	0.59
DEM	0.19	0.11	0.09	0.03	0.17	0.50	0.433	0.57
POSSMOD	0.35	0.65	0.05	-0.01	-0.01	-0.04	0.739	0.26
NECMOD	-0.04	0.75	0.22	-0.02	0.14	0.01	0.627	0.37
PREDMOD	0.08	0.13	0.68	-0.21	0.20	-0.23	0.631	0.37
PUBV	0.63	0.06	-0.01	0.09	-0.13	0.18	0.559	0.44
PRIVV	0.76	0.06	0.10	-0.06	0.10	0.07	0.692	0.31
SUASV	0.24	0.46	-0.20	-0.06	-0.19	0.02	0.412	0.59
SEEM	-0.13	-0.09	0.06	0.02	0.44	0.23	0.241	0.76
SPLITINF	0.02	0.01	-0.03	-0.10	0.05	-0.15	0.037	0.96
SPLITAUX	-0.04	0.57	-0.32	-0.05	0.43	0.06	0.609	0.39
PHCOORD	-0.28	0.29	-0.18	0.33	-0.26	0.18	0.355	0.65
CLCOORD	0.08	0.27	0.23	0.08	0.20	0.08	0.233	0.77
SNEG	0.21	0.57	0.10	0.28	-0.04	0.12	0.661	0.34
ANEG	0.41	0.58	-0.01	0.20	-0.10	0.09	0.826	0.17

6-factor solution for 58 features (PAF, Oblimin)

Unlike in the three- and four-factor solutions, less features load now on the first factor, which accounts for only 11 per cent of the total shared variance (even though it is still the maximum proportional percentage, the total variance explained by the six factors being 45 per cent; see Appendix IV for proportional variance). Some of the positive features have remained, while others have stronger loadings on other factors. Within the first group, positive features include private and public verbs, first person pronouns, indefinite pronouns, WH-clauses on object positions, *that*-clauses as verb complement, *that*-clauses as adjective complement, demonstrative pronouns and questions, while the only remaining negative feature is nouns, but it loads more strongly on Factor 2. Analytic negation, which still loads as a positive feature on Factor 1, also has a more salient loading on the second factor, as is also the case of synthetic negation (which, unlike analytic negation, does not load on Factor 1 anymore). Second person pronouns, by contrast, have a stronger loading on Factor 1 in this six-factor solution, although it also loads on Factor 2. On the other hand, nominalisations and agentless passives have remained on Factor 2, but now appear to form a different construct along with necessity and possibility modals, synthetic and analytic negation, split auxiliaries, suasive verbs, *to* infinitives and WH-clauses in subject position, suggesting a persuasive or argumentative aspect of a specialised type of discourse. We can see here that, in the six-factor solution, the argumentative features are no longer loading on Factor 1, which means that the argumentative function constitutes here an autonomous underlying construct. This group of features is contrasted against nouns, place adverbs, present participial WHIZ-deletions and, to a lesser extent, second person pronouns (that is, the negative features on Factor 2). If nouns and present participial WHIZ-deletions were usually interpreted to convey informational density, their clustering with place adverbs and second person pronouns appears to be harder to interpret, even though the latter do not have much weight on the factor.

Factor 3, however, mirrors the structure of the second factor in the previous analyses, grouping downtoners, hedges, amplifiers, and predicative and attributive adjectives on one side, and predictive modals, sentence relatives, causative and conditional subordination, conjuncts and other adverbial subordinators on the other side, conserving thus the “subjective evaluation vs. assertive prediction” dimension, albeit with reversed polarities. Similarly, in Factor 4, present tense forms a group with third person pronouns, the pronoun *it*, adverbs, and to a lesser extent, amplifiers and phrasal coordination under positive sign, whereas negative features include prepositions, conjuncts, nominalisations, pied-piping constructions and past participial WHIZ-deletions. This structure is identical to that of the third factor (“elaborated vs. spontaneous discourse”) in the three- and four-factor solutions. The same happens with Factor 5, which keeps reproducing the “narrative vs. non-narrative discourse” structure from the fourth factor in the four-factor solution. However, in this case adverbs and split auxiliaries have been added to the ‘narrative’ group, while the only 'non-narrative' features are now *be* as main verb and present tense (which in this analysis loads more strongly on Factor 4).

Notwithstanding, just like Factor 2, Factor 6 offers an altogether new structure to be interpreted. The positive features are WH-relative clauses (with the relativiser either on object and subject positions), demonstratives and *by*-passives, as well as third person pronouns, past tense, pied-piping constructions and perfect aspect, the three latter features having stronger loadings on other factors. On the other end of the scale, in turn, we have present participial WHIZ-deletions, conjuncts and past participial WHIZ-deletions. In principle, the first group of features could be translated into a "human" focus, the label being taken from Gray (2011: 138), in that *by*-passives and third person pronouns highlight the person performing the action, rather than the action itself, while WH-relative clauses are not only introduced by a possibly inanimate *which*, but also by *who*, *whom* and *whose*. However, there is also the possibility that the distinction here is "syntactic elaboration" (considering that the WH-relative clauses have the largest loadings among all the positive features, followed by demonstratives and *by*-passives) and "syntactic compression" (conveyed by present and past participial WHIZ-deletions).

In the light of this six-factor solution we have once more seen how the extraction of additional factors may help to uncover new underlying constructs which may be relatively meaningful for the analysis and contribute to the overall variation picture. However, it is clear that sometimes the interpretation of these new constructs may be rather difficult and should therefore be considered with caution. In this particular case, the interpretation of the 'new' factors – Factor 2 and Factor 6 – seems a bit tricky in that it is not clear to what extent the information they convey is actually new. While Factor 6 appears to constitute an autonomous dimension of variation, Factor 2 seems to be a result of over-factoring in that some of its features make more sense when they load on Factor 1 in the three- and four-factor solutions. In any case, because Factor 2 appears to be somewhat obscure here, we could either adopt only five of the six factors for the final model, or else, bearing in mind that a five-factor solution was not recommended by the scree plot, and did not offer any reliable new constructs, go back to the four-factor solution. The latter option was preferred, based on the initial scree plot (Figure 5.1) and on the interpretation of the factors.

3.4. Towards a final four-factor solution

Although it is probable that Biber (1988) retained all the original variables in his analysis, regardless of the communalities and the MSAs (Lee 2000: 304), another round of factor analyses was run on two modified versions of our initial dataset. At a first stage, it was decided to drop sixteen variables from the initial fifty-eight variable dataset which presented low communalities (i.e. those presenting scores below .3), after which the two preliminary statistical tests were carried out on the new correlation matrix for the forty-two-variable dataset, with a resulting improved KMO MSA = .79 (now ‘middling’) and Chi Square = 3365.793, df = 861, p < .005 (which, again, indicates that the correlation matrix is not an identity matrix). The scree plot obtained from the correlation matrix suggested a five-factor solution (see Appendix IV). A factor analysis was carried out, three, four and five factors being extracted. Despite the fact that the five-factor solution was statistically considerably stronger and explained 51 per cent of the total variance, some of the factors were particularly difficult to interpret. The reason for this was, presumably, the absence of twelve features which, despite presenting low communalities, yielded salient loadings on some factors and had thus contributed to their interpretation in previous analyses (see discussion earlier in Section 3.2). This five-factor solution was therefore rejected and will not be discussed in this section, although it can be found in Appendix IV (the rest being available upon request from the author).

After that, twelve of the sixteen features that had previously been dropped were restored, with a resulting fifty-four variable dataset. This time, only those features which failed to load on any of the factors in the previous four-factor solution (i.e. *by*-passives, *that* clauses on subject position, WH-relativiser on subject position and split infinitives) were excluded from the analysis. Once more, the KMO MSA and Bartlett’s Sphericity tests were run on the correlation matrix, resulting in a somewhat worsened overall MSA of .71 (although still ‘middling’ and therefore better than with the fifty-eight variable dataset), and a Chi Square = 4381.731, df = 1431, p = < .005 (which means that the results of the FA are statistically significant). Figure 5.2 shows a scree plot for the fifty-four-variable dataset, which suggests once more a four-factor solution:

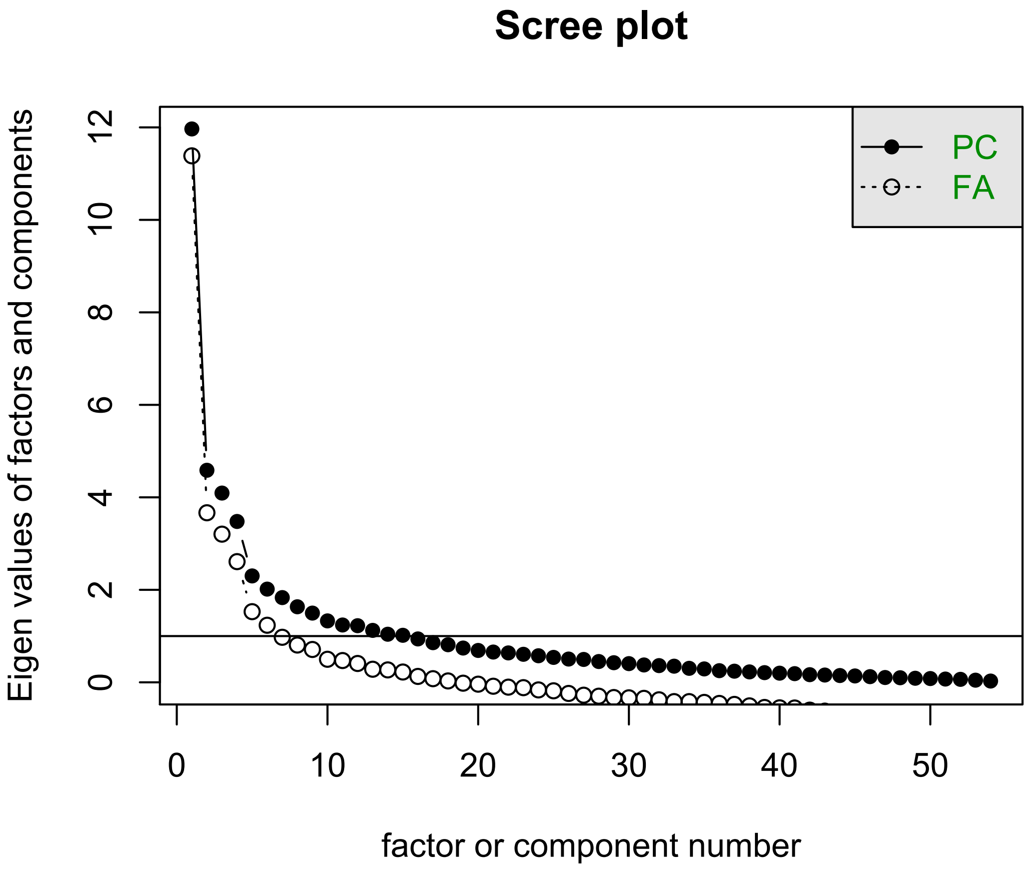


Figure 5.2
Scree plot for a 54-variable correlation matrix

A factor analysis was then run, using, just like in the previous cases, the PAF extraction method and Oblimin rotation, and four factors were extracted from the fifty-four-variable dataset (see Table 5.6, next page):

Table 5.6

	PA1	PA2	PA3	PA4	h2	u2
PAST	0.20	-0.02	-0.06	0.38	0.19	0.81
PERF	0.40	-0.24	0.20	0.48	0.51	0.49
PRES	0.17	-0.21	-0.39	-0.41	0.39	0.61
PL_ADV	-0.37	0.09	-0.43	0.32	0.42	0.58
TIM_ADV	0.15	-0.01	-0.15	0.66	0.47	0.53
FPERS	0.72	0.02	-0.03	0.11	0.53	0.47
SPERS	0.19	0.15	-0.33	0.14	0.18	0.82
TPERS	0.45	0.04	-0.40	-0.06	0.37	0.63
ITPRO	0.52	-0.12	-0.33	-0.18	0.41	0.59
DEMPRO	0.50	-0.01	-0.12	0.02	0.26	0.74
INDPRO	0.80	-0.03	-0.16	-0.01	0.66	0.34
PRO_DO	0.41	0.31	-0.32	-0.08	0.38	0.62
QUEST	0.63	0.02	-0.05	-0.16	0.43	0.57
NOM	0.38	-0.02	0.73	-0.16	0.69	0.31
NOUN	-0.75	-0.04	-0.13	-0.13	0.61	0.39
AGPASS	-0.01	-0.02	0.36	-0.09	0.14	0.86
BE_MAIN	0.39	0.10	0.01	-0.55	0.47	0.53
EXTHERE	0.52	0.07	-0.20	0.11	0.33	0.67
THAT_V	0.44	-0.01	0.18	0.28	0.31	0.69
THAT_ADJ	0.40	-0.08	0.26	0.27	0.32	0.68
WHCL_SUB	0.50	-0.03	0.12	-0.18	0.30	0.70
WHCL_OB	0.73	0.19	-0.20	0.11	0.62	0.38
TO_INF	0.76	0.04	0.12	-0.04	0.59	0.41
PASTPART	-0.32	-0.24	0.09	-0.19	0.20	0.80
WHIZ	-0.30	0.10	0.48	0.10	0.34	0.66
PRES_WHIZ	-0.49	0.14	-0.01	0.20	0.29	0.71
WHREL_OB	0.55	-0.08	0.04	0.11	0.32	0.68
PIP	0.21	-0.04	0.31	-0.01	0.14	0.86
SREL	0.09	0.35	-0.30	0.00	0.22	0.78
CAUSADV	-0.05	0.56	-0.06	0.02	0.32	0.68
CONCADV	0.38	-0.35	-0.10	0.07	0.28	0.72
CONDADV	0.40	0.50	0.05	-0.12	0.44	0.56
OTHADV	0.16	0.35	-0.06	0.20	0.18	0.82
PREP	-0.24	0.16	0.60	0.20	0.49	0.51
ATTRADJ	-0.09	-0.65	0.39	0.09	0.60	0.40
PREDADJ	0.02	-0.38	-0.04	-0.27	0.21	0.79
ADV	0.53	-0.11	-0.30	0.38	0.52	0.48
CONJ	-0.02	0.59	0.39	0.11	0.50	0.50

Table 5.6 (continued)

	PA1	**PA2**	**PA3**	**PA4**	**h2**	**u2**
DOWN	0.16	-0.38	0.23	0.32	0.35	0.65
HEDG	-0.17	-0.61	-0.04	0.32	0.53	0.47
AMPL	0.14	-0.48	-0.26	0.25	0.38	0.62
DEM	0.51	-0.03	-0.10	0.15	0.29	0.71
POSSMOD	0.79	0.07	0.25	-0.02	0.70	0.30
NECMOD	0.59	0.18	0.23	-0.02	0.44	0.56
PREDMOD	0.07	0.74	0.10	0.17	0.58	0.42
PUBV	0.68	-0.02	-0.07	0.00	0.46	0.54
PRIVV	0.70	0.15	0.03	0.26	0.58	0.42
SUASV	0.53	-0.19	0.29	-0.16	0.42	0.58
SEEM	-0.03	0.02	-0.15	0.37	0.16	0.84
SPLITAUX	0.40	-0.29	0.25	0.33	0.44	0.56
PHCOORD	0.17	-0.33	-0.15	-0.41	0.31	0.69
CLCOORD	0.38	0.19	-0.05	0.12	0.20	0.80
SNEG	0.78	0.01	-0.07	-0.13	0.62	0.38
ANEG	0.89	-0.05	0.04	-0.11	0.81	0.19

4-factor solution for 54 features (PAF, Oblimin)

If we compare Tables 5.6 and 5.4, we will see that the present analysis reproduces almost exactly the four dimensions identified in the previous four-factor solution. Most features load on Factor 1, which accounts for 21 per cent of the shared variance, the total amount of variance explained by the four factors being 41 per cent (see Appendix IV). In order to offer a clearer picture for the interpretation of the factors, Table 5.7 (below and next page) summarises the factorial structure:

Table 5.7

FACTOR 1		**FACTOR 2**		**FACTOR 3**	
Analytic negation	.89	Predictive modals	.74	Nominalisations	.73
Indefinite pronouns	.80	Conjuncts	.59	Prepositions	.60
Possibility modals	.79	Causative subordination	.56	Past participial WHIZ-deletions	.48
Synthetic negation	.78	Conditional subordination	.50	(Conjuncts)	.39
To infinitives	.76	Other adverbial subordinators	.35	(Attributive adjectives)	.39
WH-clauses on object position	.73	Sentence relatives	.35	Agentless passives	.36

Table 5.7 (continued)

FACTOR 1 (continued)		**FACTOR 2 (continued)**		**FACTOR 3 (cont.)**	
First person pronouns	.72	(Pro-verb do)	.31	Pied-piping constructions	.31
Private verbs	.70	--------------------		--------------------	
Public verbs	.68	Attributive adjectives	-.65	Place adverbs	-.43
Questions	.63	Hedges	-.61	(Third person pronouns)	-.40
Necessity modals	.59	Amplifiers	-.48	(Present tense)	-.39
WH-relativiser on object position	.55	Downtoners	-.38	(Pronoun *it*)	-.33
Adverbs	.53	Predicative adjectives	-.38	Second person pronouns	-.33
Suasive verbs	.53	(Concessive subordination)	-.35	(Pro-verb *do*)	-.32
Existential *there*	.52	(Phrasal coordination)	-.33	(Sentence relatives)	-.30
Pronoun *it*	.52			(Adverbs)	-.30
Demonstratives	.51				
Demonstrative pron.	.50	**FACTOR 4**			
WH-clauses on subject position	.50	Time adverbs	.66		
Third person pronouns	.45	Perfect aspect	.48		
that-clauses as verb complements	.44	Past tense	.38		
Pro-verb *do*	.41	Adverbs	.38		
(Perfect aspect)	.40	Seem/appear	.37		
that-clauses as adjective complements	.40	(Split auxiliaries)	.33		
Split auxiliaries	.40	(Place adverbs)	.32		
(Conditional subordination)	.40	(Downtoners)	.32		
(*be* as main verb)	.39	(Hedges)	.32		
(Clausal coordination)	.38	--------------------			
		be as main verb	-.55		
(Nominalisations)	.38	Present tense	-.41		
--------------------		Phrasal coordination	-.41		
Nouns	-.75				
Present participial WHIZ-deletions	-.49				
Place adverbs	-.37				
Past participle clauses	-.32				
(Past participial WHIZ-deletions)	-.30				

Summary of the factorial structure for the final four-factor solution with 54 variables

The features in parentheses are those loading more strongly on another factor, while those which have been greyed out indicate that their weight on the factor is not very important (< .40). Indeed, it should be borne in mind that not every feature loading on a factor is equally important for the interpretation of that factor. As noted in Biber (1988: 87), the loadings of all the features are not equally large and, therefore, not all the features are equally representative of the dimension underlying a factor. This means that, even if those features with less important loadings contribute to the overall justification of a dimension, it is nevertheless the features with the largest loadings that define the factor and should thus be given special attention. For instance, among the positive features loading on Factor 4, time adverbs and perfect aspect are the ones which have the largest loadings (.66 and .48, respectively) and are therefore the defining features of this factor, while past tense (.38), general adverbs (.38) and *seem*/*appear* (.37) have all relatively small loadings and therefore should not be used as the primary reference when interpreting the factor, although some of them – such as past tense – certainly fit in the whole underlying structure, justifying its apparent narrative function. Likewise, we might say that Factor 2 is defined by predictive modals, conjuncts, causative subordination and conditional subordination, whereas other adverbial subordinators and sentence relatives have only a minor contribution in the underlying dimension (which, notwithstanding, also makes sense in that both of them are elements that organise, or structure, the discourse).

As for the features in parentheses, which have stronger loadings on another factor, their importance is, in principle, also secondary, not only because their loadings are also relatively weak (< .40 in most cases, except for perfect aspect and conditional subordination on Factor 1, and third person pronouns on Factor 3), but also because a 'second largest' loading indicates that the relationship of the feature with this factor is not as strong as its relationship with the factor where it has the largest loading. Even so, as has been noted earlier, these features may also contribute to the underlying dimension of the factor and, therefore, to its interpretation. Thus, for instance, present tense loads more strongly on Factor 2 (–.41) because it is in binary opposition with perfect aspect and past tense (i.e. the narrative dimension), but it also loads on Factor 3 (–.39) along with other features that presumably indicate immediacy or spontaneity, such as place adverbs, or second and third person pronouns. This can be illustrated through example (5.2) below, where second and third person pronouns are in *italics*, while present tense verbs are underlined:

(5.2) if *you* plunge *them* into well rectified spirit of wine (…) *they* soon expire and retain *their* golden appearance (life19)

As noted in Biber (1988: 92) and elsewhere (Tabachnick & Fidell 1996; Fabrigar & Wegener 2012), the interpretation of the factors is always tentative until confirmed by further research. An in-depth analysis of the dimensions of variation produced by this four-factor solution is offered in Chapter 6. The next step in a factor analysis is the calculation of factor scores, which are estimates of the values of each factor for each of the observed variables (Grieve 2016: 217). In multidimensional analyses of register variation, factor scores are computed for each text, and then for each register, in order to analyse the similarities and differences among the different registers with regard to each factor, or dimension of variation. The process is explained in what follows.

4. Calculating factor scores

Strictly speaking, factor scores are estimated, or approximated, rather than calculated, in that they are always indeterminate since, for certain mathematical reasons, it is not possible to measure them directly (Steiger & Schöneman 1978; see Grice 2001: 432). Hence, several methods have been developed to estimate factor scores. Two commonly used procedures are the regression method (Thomson 1951) and the Bartlett method (Bartlett 1937). Just like factor analysis itself, each of them is based on complex mathematical equations that are not going to be dwelled upon here. However, it should be remarked that both use standardised, rather than normalised, frequencies of variables (or z-scores; see explanation below). While the regression method is considered to be the simplest to interpret and is normally used after Maximum Likelihood analysis, the Bartlett method is preferred if the factors have been extracted with Principal Axis Factoring (PAF) (Grieve 2016: 218). Even so, Tabachnick & Fidell (1996: 678) explain that, in the Bartlett method, factor scores have the same mean and standard deviation as in the regression approach, which should yield similar results.

Another traditional way of estimating factor scores is a more straightforward one, which consists of summing the frequencies of variables that load highly on each factor. This "quick and dirty" approach (Tabachnick & Fidell 1996: 678) is the one used by Biber in all his multidimensional studies (1988, 1995, 2001, etc.) and in many of those following his method (e.g. Conrad 1996; Lee 2000; Gray 2011). In this procedure, the frequencies for all the variables with salient loadings on a factor are simply added together, while those with negative salient loadings are subtracted. At this point, the normalised frequencies (in our case, per 1,000 words of text) that have been obtained at a previous stage need to be standardised, that is, transferred to a scale where the mean frequency for a feature in the sample is 0.0 and its standard deviation is 1.0. The formula used for this procedure is the following:

$$z = (x - \mu) / \sigma$$ [38]

The result of this formula is a so-called standard score, or z-score, with either a positive or a negative value. Thus, for instance, a positive z-score of 1.3 for a particular linguistic feature (e.g. past tense) in a text indicates that the frequency of this feature is 1.3 standard deviations higher in this text, relative to the overall sample mean for this feature (which is always 0.0). If this happened, it would mean that that particular text is 'marked' for the presence of that linguistic feature (in this case, past tense). Likewise, a negative z-score for past tense in another text would indicate that this text contains fewer past tense forms than the mean of all the texts in the corpus. As explained in Biber (1988: 94), standardised, rather than normalised frequencies are used in order to avoid cases where very frequently occurring features (e.g. with a normalised frequency of 140.36 per 1,000 words) would have a much larger weight on a factor than those features with smaller frequencies (e.g. 0.9 per 1,000 words). This would create factor scores which are not proportionate for all the features, thus overestimating the influence of some features on a factor and underestimating the importance of others.

In this study, we have estimated factor scores by trying both the regression and the Bartlett methods, as well as the straightforward method (which, for convenience, will be termed here 'the Biber method'). Both the regression and the Bartlett methods are available in the R statistical package, which allowed factor scores to be calculated automatically for each text. After that, mean, or average, factor scores were calculated for each scientific discipline in each century, by adding together the factor scores for all the texts in that scientific discipline in a given period

[38] Normalised frequency value minus mean value in the corpus, divided between the standard deviation.

and dividing the resulting figure by the number of texts. As was expected, differences between the two types of scores were minor (see Figures 5.3 and 5.4 below):

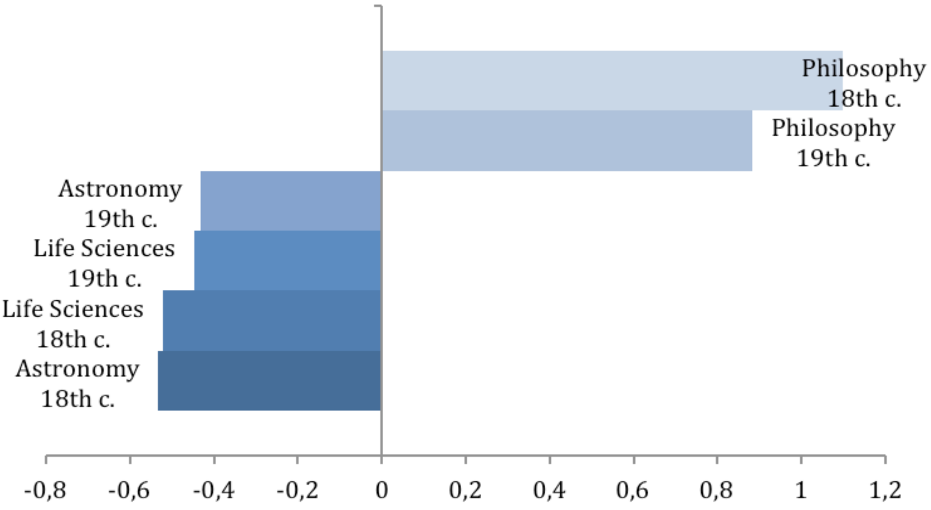


Figure 5.3
Mean scores of Dimension 1 for three scientific disciplines and two centuries (Regression method)

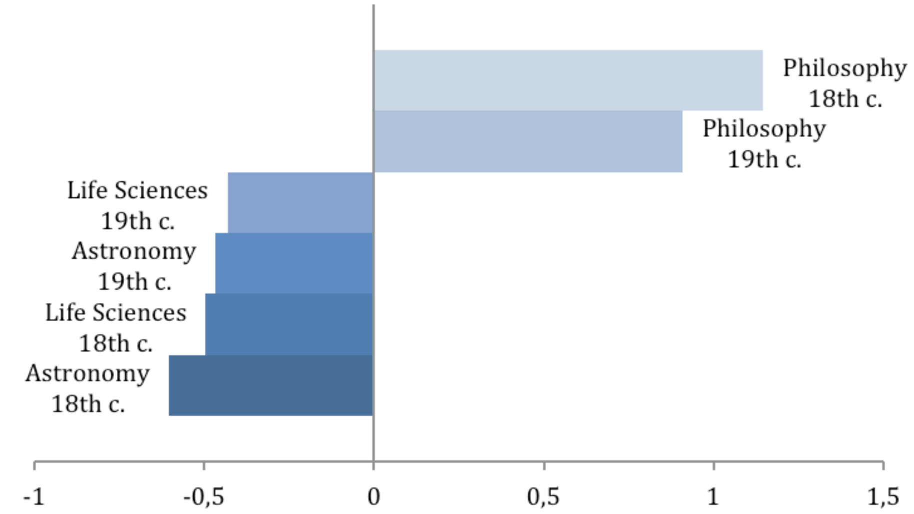


Figure 5.4
Mean scores of Dimension 1 for three scientific disciplines and two centuries (Bartlett's method)

Figures 5.3 and 5.4 plot the factor scores for Dimension 1 (i.e. Factor 1), tentatively labelled as "involved argumentation vs. informational density", for three scientific disciplines (Life Sciences, Philosophy and Astronomy) in the eighteenth and nineteenth centuries, using the regression method and the Bartlett method, respectively. The x-axis represents the dimension of variation, while the bars to the left (negative scores) or to the right (positive scores) show the dimension scores for each discipline in each century, in ascending order. As we can see on both figures, two scientific disciplines – Life Sciences and Astronomy – have negative scores in both centuries, which means that they are characterised by a high frequency of negative features (such as nouns, present participle WHIZ-deletions and past participle clauses) and a rather low frequency of positive features from Factor 1 (e.g. first person pronouns, questions, private and public verbs, etc.; see Table 5.7 in the previous section). Conversely, both eighteenth- and nineteenth-century Philosophy has rather high positive scores (with values between .8 and 1.0 in the nineteenth century and 1.1 and 1.3 in the eighteenth), which indicates that the positive features from Factor 1 listed on Table 5.7 are frequent in philosophy texts, while the negative ones are relatively infrequent.

If we were to interpret these results in the light of the tentative label suggested for Dimension 1, we might say Philosophy appears to be a scientific discipline marked by an involved and argumentative style both in the eighteenth and the nineteenth centuries, although slightly less so in the latter period. By contrast, Astronomy and Life Sciences appear as rather informational disciplines, with a concentrated nominal style which increases as we move forward in time. The minor differences between the values obtained with the regression and the Bartlett methods can be appreciated better in this group, in that Bartlett scores are always more 'informational' for Astronomy than for Life Sciences, whereas regression scores yield a higher negative value for nineteenth-century Life Sciences (– .45) with respect to nineteenth-century Astronomy (– .43). On the other hand, regression scores show eighteenth-century Astronomy and Life Sciences standing closer to each other, while Bartlett scores appear to have moved Astronomy farther to the 'informational' end of the scale.

In order to calculate factor scores using the Biber method, z-scores have been first computed automatically (with R) for each linguistic feature in each text, after which those of the salient features in each factor were summed. At a first stage, in order to follow Biber (1988) as closely as possible, only the features that had their largest loading on a factor were considered for that factor, features with 'second largest' loadings being discarded. In other words, features were never 'recycled' (i.e. used for more than one factor) because, statistically, they have a closer relationship with the factor on which they load more strongly. However, later studies by Biber (2001, 2006) and other multidimensional studies, such as Lee (2000) or Gray (2011) do appear to 'recycle' the features loading on several factors. Thus, another trial version consisted in calculating factor scores using the z-scores of *all* the features loading on a factor, regardless of the fact whether those features loaded stronger on another factor. Figures 5.5. and 5.6 plot factor scores for Dimension 1 for the same three scientific disciplines, following the model discussed earlier, without and with ‘recycled’ features, respectively:

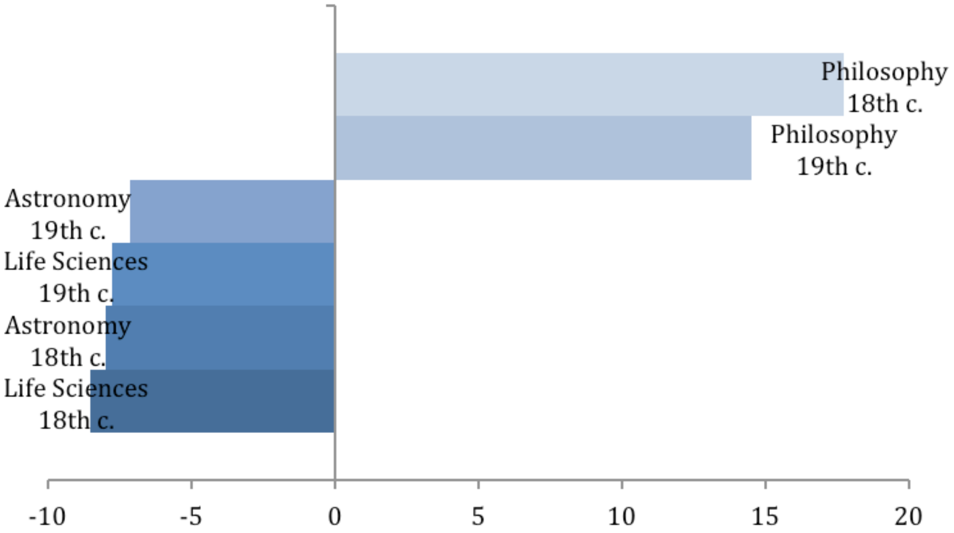


Figure 5.5
Mean scores of Dimension 1 for three scientific disciplines and two centuries (Biber method; no ‘recycling’ of features)

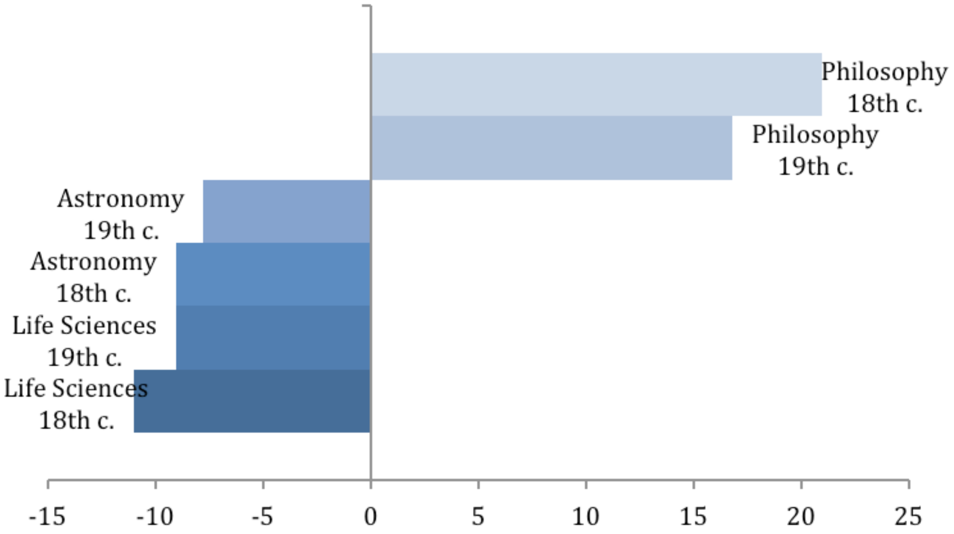


Figure 5.6

Mean scores of Dimension 1 for three scientific disciplines and two centuries (Biber method; ‘recycled’ features)

Just like in Figures 5.3 and 5.4, the x-axis represents the dimension, but on Figures 5.5 and 5.6 its range is considerably bigger (–10 to 20 without 'recycled' features, –15 to 25 with 'recycled' features). The reason lies in the fact that in the regression and the Bartlett methods z-scores were weighted, in a way that the final factor scores were proportional to the importance of each feature to each factor, thus features with larger loadings having larger weighted z-scores, and those with less important loadings having proportionally smaller z-scores. In the Biber method, however, the z-scores for the linguistic features are not weighted, that is, all the z-scores are treated equally, regardless of their factor loadings (of course, as long as they are > .30). Potentially, this may somewhat distort the picture shown in the factorial structure because, clearly, not all the features loading on a factor are equally representative of the structure underlying that factor. This could be a problem both in the case when features are 'recycled' and when they are not. In the first case, supposing a factor has a number of features with relatively small loadings, most of which are also 'second largest', then if all these features are included in the factor score, they may add a disproportionate weight to that score. On the other hand, when several features with 'second largest' are *not* counted for the computation of a factor score, but other features with 'primary' (or largest), yet not very important loadings *are* counted, the picture is once more distorted, this time because several potentially meaningful features have been taken out of the sum, while the least important ones have been given unnecessary attention.

Here, if we compare Figures 5.5 (with no 'recycled' features) and 5.6 (with 'recycled' features), we can see, first of all, that Philosophy has very high positive scores (15 to 17 for the eighteenth century, 18.5 to 22 in the nineteenth) as a result of directly summing all the z-scores for the positive features in Factor 1. This is usually the case in most factor analyses when factor scores are obtained directly because, as has been noted earlier, most features load on the first factor. In the case of the 'negative' group (Astronomy and Life Sciences), fewer features are counted in each case (because Factor 1 has only five negative features, four if we do not 'recycle' past participle WHIZ deletions), which results both in a lower negative factor score. However, it also results in a different arrangement of the two disciplines in each century, depending on whether this 'recycled' feature is counted or not. Thus, when past participle WHIZ-deletions are not included in the factor score, eighteenth-century Astronomy is more 'informational' than nineteenth-century Life Sciences, whereas if we include this 'recycled' feature, we get the opposite picture and, actually, Life Sciences becomes, overall, more 'informational' than Astronomy. On the other hand, the fact that Life Sciences has larger negative scores than Astronomy with the Biber method, contrarily to what happened when we calculated factor scores with the Bartlett method, seems to indicate, precisely, that features with smaller loadings (such as past participle clauses or place adverbials) were particularly frequent in Life Sciences and, due to their large z-scores, have a larger weight on the factor scores.

All in all, however, Figures 5.3-5.6 show that, despite all the (relatively) minor differences discussed above, the main picture of Dimension 1 does not change, which indicates that the underlying structure of Factor 1 is quite strong. This was not the case with statistically weaker factors, such as Factors 3 and 4, which have fewer features and most of them loading less strongly, and where, consequently, part of the picture may change considerably, depending on the method used (i.e. if we compare Bartlett or regression scores with Biber scores)[39]. Although it is true that the Biber method has been used in most multidimensional analyses, weighted scores appear to be, overall, more reliable. Following the advice offered in Grieve (2016), we have thus decided to use Bartlett scores, based on the fact that we have also use PAF as a factor extraction method. The next chapter offers a discussion of each of the four dimensions of variation that have resulted from the factor analysis here reported, as well as an analysis of the variation of three scientific disciplines and eight registers across the eighteenth and nineteenth centuries.

[39] Factor scores calculated with the Bartlett method have been included in Appendix III. Those calculated with the regression method and with the Biber method, both with and without 'recycled' features, are available on request.

Chapter 6

A Multidimensional Analysis of Late Modern English scientific texts

1. Introduction

In the two previous chapters we have described the methodology followed in this study, which consisted in the selection and retrieval of fifty-eight linguistic features from the corpus samples and in the processing of their frequencies through a multivariate statistical technique, namely, factor analysis. This permitted to transform the initial large amount of linguistic features into a much smaller subset of latent constructs, or factors. For this particular research we have identified four such constructs which, in their totality, explain 41 per cent of the total variation in the corpus. After a series of trial stages, which included the dropping of variables with low communalities, a four-factor solution with fifty-four variables was agreed on as our final model. These four factors can now be interpreted as underlying dimensions of variation, each of them responsible for a particular discursive function.

Although a preliminary interpretation of the four-factor model was offered in Chapter 5 (Section 3.2), in this chapter we will focus on each dimension of variation in detail. Section 2 examines the functions of the groups of features constituting each factor through examples from the texts and offers a definitive label for each dimension of variation. After that, Section 3 discusses the four dimensions in function of the relations among three scientific disciplines (Astronomy, Philosophy and Life Sciences) and eight subregisters (treatise, essay, textbook, lecture, letter, article, dialogue and dictionary) with respect to each of these dimensions.

2. Dimensions of variation in late Modern scientific English

Following Biber (1988: 101), Table 5.7 from Chapter 5 is repeated here for convenience as Table 6.1, summarising the factorial structure:

Table 6.1

FACTOR 1		**FACTOR 2**		**FACTOR 3**	
Analytic negation	.89	Predictive modals	.74	Nominalisations	.73
Indefinite pronouns	.80	Conjuncts	.59	Prepositions	.60
Possibility modals	.79	Causative subordination	.56	Past participial WHIZ-deletions	.48
Synthetic negation	.78	Conditional subordination	.50	(Conjuncts)	.39
To infinitives	.76	Other adverbial subordinators	.35	(Attributive adjectives)	.39
WH-clauses on object position	.73	Sentence relatives	.35	Agentless passives	.36
First person pronouns	.72	(Pro-verb do)	.31	Pied-piping constructions	.31
Private verbs	.70	--------------------		--------------------	
Public verbs	.68	Attributive adjectives	-.65	Place adverbs	-.43
Questions	.63	Hedges	-.61	(Third person pronouns)	-.40
Necessity modals	.59	Amplifiers	-.48	(Present tense)	-.39
WH-relativiser on object position	.55	Downtoners	-.38	(Pronoun *it*)	-.33
Adverbs	.53	Predicative adjectives	-.38	Second person pronouns	-.33
Suasive verbs	.53	(Concessive subordination)	-.35	(Pro-verb *do*)	-.32
Existential *there*	.52	(Phrasal coordination)	-.33	(Sentence relatives)	-.30
Pronoun *it*	.52			(Adverbs)	-.30
Demonstratives	.51				
Demonstrative pron.	.50				

Table 6.1 (continued)

FACTOR 1 (continued)		**FACTOR 4**	
WH-clauses on subject position	.50	Time adverbs	.66
Third person pronouns	.45	Perfect aspect	.48
that-clauses as verb complements	.44	Past tense	.38
Pro-verb *do*	.41	Adverbs	.38
(Perfect aspect)	.40	Seem/appear	.37
that-clauses as adjective complements	.40	(Split auxiliaries)	.33
Split auxiliaries	.40	(Place adverbs)	.32
(Conditional subordination)	.40	(Downtoners)	.32
(*be* as main verb)	.39	(Hedges)	.32
(Clausal coordination)	.38	--------------------	
		be as main verb	-.55
(Nominalisations)	.38	Present tense	-.41
--------------------		Phrasal coordination	-.41
Nouns	-.75		
Present participial WHIZ-deletions	-.49		
Place adverbs	-.37		
Past participle clauses	-.32		
(Past participial WHIZ-deletions)	-.30		

Summary of the factorial structure for the final four-factor solution with 54 variables

This table presents the linguistic features loading on each factor in descending order of importance. As has been explained earlier, the features with the largest loadings are the ones that bear the closest relationship with, or define, the factor, whereas those with smaller loadings have a looser relationship with the factor, even though they usually contribute to the overall interpretation of the underlying structure. The features loading on each factor tend to co-occur in the texts from the corpus. Features with positive and negative weights occur in the texts in a complementary distribution, meaning that when the former are frequent in a text, the latter are not, and vice versa. This is due to the fact that the communicative functions of the positive and negative subsets of features oppose each other (e.g. "narrative vs. non-narrative", "personal vs. impersonal", etc.).

As we can see on Table 6.1, some features have loaded on more than one factor, albeit with different weights. In the previous chapter we had agreed that features with 'second largest' loadings (that is, those which load primarily on another factor) would be taken into account in the present study because they may be important for a correct interpretation of the dimension. A discussion of each group of features is offered in the sections that follow.

2.1. Interpretation of Factor 1

As was noted earlier, in order to interpret a factor it is important to focus, first of all, on those features which present the largest loadings on that factor. In the case of the positive features on Factor 1, eight of them have a loading of .70 and larger (analytic negation, indefinite pronouns, possibility modals, synthetic negation, *to* infinitives, WH-clauses in object position, first person pronouns and private verbs), while twelve more features show a loading equal to or larger than .45 (which includes public verbs, questions, necessity modals, WH-relativiser in object position, adverbs, suasive verbs, existential *there*, pronoun *it*, demonstratives, demonstrative pronouns, WH-clauses in subject position and third person pronouns). The first group of features is the one that has the strongest relationship with Factor 1. Some of these features, such as possibility modals, synthetic and analytic negation, first person pronouns and private verbs appear to indicate an internal debate in the first person, through which the author often expresses his/her own views on a subject from an involved, rather than detached, perspective. Possibility modals, in particular, express the estimation that certain events or situations *can* or *may* or *might* take place. Private verbs usually express emotions or thoughts, while *to*-infinitives are often used as adjective or verb complements (as in *happy to see you*; *hope to go*), where the adjective or verb expresses a personal attitude or stance (Biber 1988: 111).

In the following example (6.1), first person pronouns are in **bold**, possibility modals are in *italics*, private verbs are underlined, analytic negation is highlighted in grey, the adjective expressing stance is double underlined, and the infinitival clause complementing the adjective is wave underlined:

> (6.1) Upon this principle, **I** think, **we** *might* freely have rejected* any theories, hitherto entertained (…) Sometimes, however, it *may* not be

> improper, to throw out hints and conjectures, when **we** *can* attain to nothing better… (astr15)

The verb form *rejected* is marked with an asterisk (*), because, although it was not included in any of Biber's (1988) closed lists of verbs, it can be interpreted here as a private verb which expresses attitude. Likewise, the expression *to throw out hints and conjectures* indicates, once more, a mental action. On the other hand, direct questions and public verbs appear to convey an interactive style, while necessity modals and suasive verbs seem, once more, to indicate involved argumentation, in which a particular idea must be defended[40]:

> (6.2) {If the will *can* not set ***itself*** in motion,} ***it*** must be moved by something else. **I** agree with Mr. Locke in thinking, that the only stimulus to action is some pressing uneasiness [which the mind feels], prompting ***it*** to change ***its*** present state. Says he, This uneasiness then **I** consider to be the immediate cause of volition, and absolutely essential to every act of the will. But WHENCE DOES THIS UNEASINESS ARISE? (phil20)

In example (6.2) above, the necessity modal *must* (dashed underlined) marks the insistence on the hypothesis that "something else necessarily moves the will", while the suasive verb *agree* (thick underlined) indicates that the author is complying with Mr. Locke's opinion (which he expresses as a thought, or an act of thinking by means of the corresponding mental verb, here underlined). Another feature which contributes to the argumentative character of this extract is the conditional clause (between curled brackets {}) which introduces the hypothetical state of the art at the beginning of this extract. Although conditional subordination is a 'recycled' feature from Factor 2 (that is, it loads more strongly on that factor), it also loads on the first factor, contributing to the persuasive character of the dimension.[41] Likewise, the public verb *says* (double underlined) introduces a direct quotation[42], presumably by Mr. Locke, whose idea is introduced by a mental verb (*consider*) which, in turn, is complemented by an infinitival clause (*to be the immediate cause of volition…*). Finally, the direct

[40] Suasive verbs and necessity modals appear in Biber's (1988) Dimension 4, labelled "Overt expression of persuasion".

[41] Conditional subordination is another feature loading on Biber's (1988) Dimension 4.

[42] This must be a paraphrase, rather than a quotation, since all direct quotations have been eliminated from the *Coruña Corpus* to avoid including text belonging to a different author.

rhetorical question (in SMALL CAPS) reflects an interactive style, through which the author seems to be directly addressing either the reader, or else, his own thoughts.

On the other hand, the WH-relative clause (with the head *which* in object position, in square brackets []) are explicit identifications of referents in the text (Biber 1988: 110), just as the pronoun *it* in its different forms (in ***bold italics***). Likewise, the demonstrative *this* (dot-dot-dash underlined) which premodifies *uneasiness* refers to the *pressing uneasiness* mentioned immediately before, whereas the third person pronoun *he* refers to *Mr Locke*. These deictic elements contribute to a more fluid exposition of the idea developed in this extract, while the direct personal references (*I*, *Mr Locke*, *he*, the *mind*), all of which perform an action, whether mental (think, feel) or verbal (say), related to this idea, point once more to an involved style.

It might seem surprising that private, public and suasive verbs, which, according to Quirk et al. (1985) belong to different semantic subclasses and which actually load on three different factors in Biber's (1988) seven-factor model[43], all load on the first factor in this study. This, in principle, is something that may happen when too few factors have been extracted, which results in too many features loading on a factor and, consequently, in two or more different structures collapsing together. However, this does not seem to be the case with the present analysis because all three types of verbs kept loading together on Factor 1 in the five-factor solution as well (see Appendix IV), even though in Chapter 5 (Section 3.3) we had seen that suasive verbs had eventually moved to Factor 2 in the six-factor solution. It might be argued, in fact, that private and public verbs, despite having labels with opposed meanings, are pragmatically very similar in that, in a more direct or covert manner, most of them convey personal stance. For instance, private verbs *assume, conclude* and *doubt* express mental actions linked to a particular attitude, but the two latter verbs could also be used as public verbs, or verbs of saying. Likewise, public verbs *agree*, *admit* or *acknowledge*, which, presumably, imply an open communication of a thought, might be also used as private (mental) verbs, depending on the context. On the other hand, some of the suasive verbs are also included in the public and private verbs groups, such as *agree* or *decide* (see list of features in Chapter 4, and Biber (1988: 242)).

[43] In Biber's (1988) multidimensional analysis, private verbs load on Factor 1, public verbs load on Factor 2 and suasive verbs load on Factor 4.

Example (6.3), which belongs to an 18th-century Philosophy text by David Hume, illustrates the involved and persuasive character conveyed by the positive features loading on Factor 1. First person pronouns are in **bold**, third person pronouns are thick underlined, possibility modals are in *italics*, private and public verbs are underlined, analytic and synthetic negation is highlighted in dark and light grey, respectively, and necessity modals are dashed underlined. Also, infinitives introducing clauses complementing verbs and adjectives are waved underlined, the WH-relative clause is in square brackets [], conditional clauses are in curled brackets {}, while the direct question is in SMALL CAPS.

> (6.3) By bringing Ideas into so clear a Light, **we** *may* reasonably hope to remove all Dispute, that *may* arise, concerning their Nature and Reality. It is probable, that no more was meant by those, [who denied innate Ideas,] than that all **our** Ideas were Copies of **our** Impressions; though it must be confessed, that the Terms they employed were not chosen with such Caution, nor so exactly **defined** as to prevent all Mistakes about their Doctrine.
>
> For WHAT IS MEANT BY INNATE? {If innate be equivalent to natural,} then all the Perceptions and Ideas of the Mind must be allowed to be innate or natural, in whatever Sense **we** take the latter Word, whether in Opposition to what is uncommon, artificial, or miraculous. {If by innate be meant, contemporary to **our** Birth,} the Dispute seems to be frivolous; nor is it worth while to enquire in what time Thinking begins, whether before, at, or after **our** Birth (phil10)

At the other end of Factor 1 we have the negative features. Here, nouns and present participial WHIZ-deletions have the largest loadings (both > .45), and appear to be, therefore, the ones that bear the closest relationship with Factor 1. The three other negative features are place adverbs, past participle clauses and past participial WHIZ-deletions. Most of these features also load negatively on Biber's (1988) Dimension 1, except for past participle clauses, which load on Dimension 5 (although the latter also includes some features which have not been counted for this study, such as type-token ratio and word length). Biber (1988: 104) describes nouns as "the primary bearers of referential meaning in a text", a high quantity of nouns indicating thus "a great density of information". Both present and past participial WHIZ-deletions, in turn, are used to post-modify nouns, "further elaborating the nominal content" (Biber 1988: 105), while past participle clauses function as elements of syntactic compression (Granger 1995a, 1995b). As for place adverbials, Biber (1988: 105) considers their co-occurrence with the other features surprising, but thinks that they may function as elements of "text internal deixis in highly informational texts (e.g. *it is shown here*; *it was shown above*)".

Example (6.4) below has been extracted from a nineteenth-century Life Sciences text, while examples (6.5), (6.6) and (6.7) belong to one from the eighteenth-century subcorpus. Nouns are **in bold**, present-participial WHIZ-deletions are underlined, past-participial whiz-deletions are between square brackets [], place adverbials are in *italics*, and past participle clauses are between curled brackets {}:

> (6.4) The **pancreas** secretes a **fluid** containing **water**, **phymatin**, **salts** of **soda**, &c., **tyrosin**, **leucin**, **guanin** in **traces** (life35)
>
> (6.5) …hence the **angle** [formed between this **line** and the horizontal one [*above* described]] is most open (life24)
>
> (6.6) When the facial **angles** of the **anthropo-morphus simi**, {as *above* stated,} are compared to those of some **Negroes**… (life24)
>
> (6.7) The **absence** of the **rete mirabile**, and of all analogous **provision*** for moderating the **influx** of the **blood** into the **brain**, accords, with the other **circumstances** [enumerated *above*], in showing that **man** is entirely unfit for the **attitude** on all **fours** (life24)

As we can see, all four extracts show a dense informational style. Example (6.4) contains many nouns, most of which function as the direct object of the present participle WHIZ-deletion post-modifying the noun *fluid*. Example (6.5), though does not present a particularly large amount of nouns, has two past participial WHIZ-deletions, one of which is embedded in the other. Example (6.6), in turn, contains a past participle clause (*as above stated*), which, being a non-finite passive construction, contributes to a compressed syntax and a somewhat detached style. Finally, example (6.7) presents a high frequency of nouns, one of which (*provision*) is marked with an asterisk (*) because, although it has been counted a noun in this study, it could also be considered a nominalisation. What seems surprising, nevertheless, is that nominalisations appear to load (even if weakly) among the positive features on Factor 1, being thus in binary opposition with nouns. This, however, may be somehow related to the abstraction of the matters dealt with in the discipline of Philosophy (which, as shall be seen in Section 3.1, presents the highest involved scores), as opposed to the material world objects observed in Astronomy and Life Sciences. On the other hand, in what respects co-occurrence, if we look at examples from the texts, we will see that nominalisations are more frequent in the company of past participle WHIZ-deletions, as in

> (6.8) The **observations** [detailed at length in the preceding chapter,] will aid the **explanations** [given *here*] (life30)

which appears to justify their loading together on Factor 3 (see Section 2.3), whereas nouns often appear with present participial WHIZ-deletions, forming combinations such as “the Stars forming this Constellation” (astr11), “the body describing the ellipse” (astr13), “a white Ring encircling its Neck” (life10).

It appears therefore, from what we have seen so far, that the positive group of features loading on Factor 1 conveys, as a whole, personal stance and a verbal style, whereas the one composed by negative features transmits a detached, densely informational style. This dimension of variation is very similar to Dimension 1 in Biber (1988[44], 2001) and Gray (2011), all of which appear to deal with the general communicative style of the texts, ranging from involved to informational. Gray (2011: 139) considers it remarkable that this dimension appears in multidimensional analyses of a type of written texts, the primary purpose of which is, in fact, informational. All in all, the involved side of Factor 1 (now Dimension 1) in the present study also appears to have a persuasive aspect to it, and we have thus considered that it should be called “Involved/persuasive vs. informational style”. This means that some texts in the *Coruña Corpus* tend towards an involved, more personal kind of discourse,

[44] In Biber (1988), Dimension 1 is called “Involved vs. informational production” because in that particular study Biber took into account the circumstances in which his texts were produced (i.e. spoken vs. written texts). In fact, the features indicating involvement occurred most frequently in spoken texts. In the present study, all the texts analysed are similar in their production circumstances, all of them being written scientific prose, and, thus, the term *style* has been considered more appropriate.

whereas others convey an unconcerned and more or less densely informational rhetoric. A classification of texts by subregisters (disciplines and genres, distributed across the eighteenth and nineteenth centuries) with respect to Dimension 1 will be offered later in Section 3.1.

2.2. Interpretation of Factor 2

Among the positive features loading on Factor 2, we have five with weights of .50 and larger (predictive modals, conjuncts, causative subordination, and conditional subordination), and three more features which, although they have smaller loadings, may help to interpret the whole underlying structure (adverbial subordinators, sentence relatives and pro-verb *do*). Predictive modals are "direct pronouncements that certain events *will* occur" (Biber 1988: 111) and thus, as their name suggests, are used to predict a certain state of affairs (see Coates & Leech 1980; Coates 1983). Its past form (*would*), in turn, expresses the possibility that something *would* occur if a condition were fulfilled. Both the present and the past forms of predictive modals are therefore often used in the apodosis of conditionals (see Comrie 1986; Athanasiadou & Dirven 1997; Ferguson 2001; Gabrielatos 2010; Puente-Castelo, forthcoming), which makes the co-occurrence of these features consistent. (In fact, both predictive modals and conditionals loaded on Biber's (1988) Dimension 4 "Overt expression of persuasion".) On the other hand, conjuncts (e.g. *hence*, *thus*, *therefore*), causative subordinators (*because*, *'cause*), conditionals (*if*, *unless*), and other adverbial subordinators (*since*, *while*, *whereas*) usually function as logical connectors between clauses.

Finally, pro-verb *do* is an element of text-internal ellipsis, or what Biber (1988: 106) calls a "reduced surface form", in that it substitutes a more complete verb phrase or a clause, whereas sentence relatives express the speaker's attitude to an idea which has just been expressed (Biber's (1988: 108) example being *He went to the store today, which I think is ridiculous*). The co-occurrence of these two features on this factor might seem rather unexpected since they both loaded among the positive features on Biber's (1988) Dimension 1 and have been demonstrated to be more typical of an involved rather than of an informational type of discourse. However, examples from the corpus have shown that sentence relatives may also convey a judgment of a particular situation from a professional, rather than an emotional, perspective:

> (6.9) The objects had better not be above 3 or 4 miles off, because the observer might, if they were very distant, have to change his place very considerably, **which would be inconvenient** (astr20)

On the other hand, reduced surface forms, such as the pro-verb *do*, can function as elements of structure compression in order to allow the transmission of information in the most efficient way (i.e. by using fewer words). In example (6.10) below, predictive modals are in **bold**, conjuncts are underlined, causative subordinators are in *italics*, conditional clauses are in curled brackets {}, and the pro-verb *do* is thick underlined:

> (6.10) *Because* the Moon describes an eccentrical orbit about the Earth at E, and the action of the Sun upon her sometimes increases her tendency towards the Earth, and sometimes diminishes it, i.e. makes her gravity towards the Earth increase, or decrease, too fast; {if while the Moon ascends from her lower apid A, her gravity towards the Earth decreases too fast, instead of describing the semi-ellipsis ABC, and coming to the higher apid at C, as she **would** otherwise do,} she **will** run out in the curve BFD, and come to the higher apid at F. But the curve ABFD is more eccentric than the curve ABCD. Therefore when the gravity of the Moon towards the Earth decreases too fast, the eccentricity of her orbit **will** increase. On the other hand, {if the Moon is going from her higher apid C to her lower A, and her gravity towards the Earth increases too fast,} instead of describing the same ellipsis CDA, and so coming to the lower apid at A, she **will** approach nearer to the Earth (astr14)

Meanwhile, the negative features loading on Factor 2 include attributive adjectives, hedges, and amplifiers (all > .45), as well as downtoners, predicative adjectives, concessive subordination and phrasal coordination. As we had remarked in Chapter 5, while hedges, downtoners and amplifiers are often used in scientific writing to either mitigate or boost categorical claims and thus make the discourse less ‘impersonal’ (Hyland 1995, 1998; Crompton 1997), attributive and predicative adjectives suggest an evaluative or descriptive language. Concessive subordination (introduced by *although*, *though*, *thou*’) serves to counteract a claim by another one, sometimes being also used as a mitigating strategy. Phrasal coordination, on the other hand, is used for unit expansion (Chafe 1982, 1985; Chafe & Danielewicz 1986), and may be used as a concluding link in an enumeration.

In example (6.11), attributive adjectives are underlined, predicative adjectives are in *italics*, hedges, downtoners and amplifiers are in **bold**, concessive subordinators are thick underlined and phrasal coordination is in square brackets []:

> (6.11) The wound is *venomous*, and **extremely** *painful*, though not *fatal*, as I once observed in a Negro wench, who was stung by one of these animals in the right side, a little below the short ribs. The wound was **almost** *imperceptible*, and without any apparent tumefaction; but the wench, whom I saw within a few minutes after the accident happened, which was in November 1763, complained of being **excessive** *cold*, though the weather was **very** *hot*, and had a violent shivering like the paroxysm of an ague, with a quick, weak, tremulous, and sometimes intermitting pulse, sometimes [yawning and stretching], and frequently gasping for breath (life14)

In the example above, the concessive subordinator *though* is used twice, first to mitigate the description of the wound produced, presumably, by a reptile or insect (*venomous, and extremely painful, though not fatal*), and, later, to contrast two opposing ideas (i.e. that the wench was *excessive cold* while *the weather was very hot*). Notice that in the first instance the initial affirmation is stressed by an amplifier (*extremely painful*) and the immediately following mitigation only denies the possibility of a further misfortune (*not fatal*), while in the second instance both adjectives are premodified by an amplifier (*excessive* (a lModE variant of the adverb *excessively*), *very*), which implies that the two ideas being contrasted are equally strong. In total, this extract contains seven attributive and five predicative adjectives, four of which are preceded by either an amplifier (*extremely*, *excessive*, *very*) or a downtoner (*almost*), showing a clearly descriptive and evaluative character.

Thus, in the light of the positive and negative features loading on Factor 2, and of examples (6.10) and (6.11) (among others to be found in the corpus, which will be given in Section 3.2) we may say that this Dimension groups, on one side, scientific texts which transmit deductive and logical reasoning and, on the other side, those which offer a detailed description of different things or processes. This seems to suggest that, just like Dimension 1 marks two different communicational styles, Dimension 2 indicates two different types of scientific focus. We have therefore decided to label this dimension as "Argumentative vs. descriptive focus".

2.3. Interpretation of Factor 3

The positive features on the third factor include nominalisations, prepositions, past participial WHIZ-deletions (all presenting loadings > .45), as well as attributive adjectives, conjuncts, agentless passives, and pied-piping constructions. In Chapter 5 we have noted that this group of features replicates to some extent Biber's (1988) Dimension 5 "abstract vs. non-abstract style", which is here reinforced with the addition of nominalisations. As explained in Bello (2014: 117),

> Nominalizations are a result of objective thought. Unlike finite clauses, which are near the speaker/listener's perspective because they require chronological sequencing, tense and overt agency expression, nominalizations allow the presentation of abstract ideas and the expression of reason and causality (Downing, 1997, 2000; Eggins, 1994).

On the other hand, nominalisations are very frequent in formal, technical and professional discourse. In the case of the scientific register, which was emerging largely through the creation of new technical terms, nominalisations occurred as the result of transforming verbs, adjectives or adverbs into nouns in order to refer to scientific procedures, natural processes or states (Halliday 1985b, 1988; Downing 1997; Bello 2014) and became frequent during the time span covered by this study.

Attributive adjectives, prepositions and past participial WHIZ-deletions, in turn, serve to pre- and postmodify nouns and to thus further elaborate on this nominal style, which results in complex noun phrases such as (6.12) and (6.13). Prepositional phrases, on the other hand, also function as time, place or manner adverbials (see example 6.14). Likewise, pied-piping constructions constitute an element of structural elaboration, consisting in a relative clause introduced by a preposition, such as the one in (6.14). In the following examples, nominalisations are in **bold**, while an asterisk (*) marks those nominal forms of verbs ending in *–ency*, and *–ion*, which are not included in Biber's (1988) list of nominalisations. Also, attributive adjectives are in *italics*, prepositions are underlined, and past participial WHIZ-deletions and pied-piping constructions are in squared [] and curled {} brackets, respectively:

> (6.12) the *Synodical* **Revolution** of the *Second* Satellite" (astr10)

> (6.13) these pursuits of a tendency* [opposed to all *religious* impressions*] (phil28)

> (6.14) an *absolute* being, {in whose nature these **conditions** and **relations**, in some manner unknown to us, disappear in a *simple* and *indivisible* **unity**} (phil34)

The somewhat longer example (6.15) offered below, extracted from the nineteenth-century part of the Astronomy subcorpus, appears to show quite a heavy presence of all these features, revealing a dense nominal style which characterises an elaborated register dealing with highly technical matters:

> (6.15) Its **conformation** reveals itself indirectly through **irregularities** in the **distribution** of light and **darkness**. The forms of its **elevations** and depressions* can be inferred only from the shapes of the *black*, *unmitigated* shadows [cast by them]. But these shapes are in a state of *perpetual* and *bewildering* **fluctuation**, partly through changes in the angle of **illumination**, partly through changes in our point of view, [caused by what are called the moon's "**librations**"] (astr41)

By contrast, the negative features loading on Factor 3 – place adverbs and third person pronouns (with values of .40 and higher), as well as present tense, pronoun *it*, second person pronouns, pro-verb do (and also, to a smaller extent, sentence relatives and adverbs) – seem to portray a very different type of discourse. As we had noted earlier, third person pronouns, place adverbs and present tense appear to make reference to either habitual or immediate events or actions that someone does or suffers, while second person pronouns indicate interaction with an explicit or implicit reader or listener of the discourse. On the other hand, the pro-verb *do* and the pronoun *it* are elements of syntactic reduction that substitute a longer finite clause or a noun phrase, respectively, thus making the discourse more compact.

In the following examples (6.16) and (6.17), third person pronouns are in *italics*, second person pronouns are thick underlined, the pronoun *it* is wave underlined, present tense verbs are in **bold**, and the pro-verb *do* is additionally underlined:

> (6.16) On the other Hand, when the Herrings **are** to be cured Red, as soon as *they* are caught, *they* **wash**, **cut**, and **lay** *them* in Brine as for pickled Herring; but let *them* soak in the Brine double the Time the others **do**, that is to say, twenty four Hours. As *they* **are** to take *their* whole Salt here, which the pickled ones **do** not, *they* taking half *theirs* in the Barrel. When the Herrings are taken out of the Brine, *they* **spit** *them*, that is, **string** *them* by the Head on little wooden Spits; and thus **hang** *them* in a Kind of Chimney made for that Purpose (life11)

> (6.17) But this is only when it **pleases** *them* to spread out *their* little bodies, and flaunt all *their* pretty fringes; and, as you will see, when I **tell** you a little more about it, *they* **can** shut *themselves* up, and look as ugly and dull as *they* **please**. In this you **see**, *they* **differ** very much from a flower, which **cannot** fold up its leaves and put them away when it **likes**. It is true that some flowers **close** at night, and **open** in the day, but it is not because *they* **want** to **do** so, but because the state of the atmosphere **causes** *them* to shut and open (life31)

The first of these examples (6.16) corresponds to an extract from a recipe for pickling herrings, and appears to be of a descriptive rather than directive character in that it contains a detailed account of what certain people usually do when they prepare the fish (*they wash, cut and lay them in Brine*; *they spit them*; *string them*, etc.). The second example (6.17), on the other hand, seems to be part of a letter, in which the author offers a description of flowers, directly addressing the reader (*as you will see*, *when I tell you…*) from time to time. The object of the description, in turn, appears to be animate and functions as the subject of different verbs which indicate attitude (e.g. *when it pleases them to spread out their little bodies*), ability (e.g. *they can shut* *themselves up*), or volition (*it is not because they want to do so*). In both cases, the discourse appears to transmit immediacy and straightforwardness, with an addition of a certain involvement through personal references, external to the text.

In the light of the above examples we can see that form seems here to go hand in hand with content. The simplicity and straightforwardness of the language in examples (6.16) and (6.17) is used to describe fairly concrete, everyday things, such as herrings and flowers. On the other hand, the complexity and abstractness conveyed by the positive features of Factor 3 appears to refer to equally complex and abstract matters, such as certain physical or metaphysical properties. This dimension of variation appears thus to deal with complexity, or elaboration, often not only linguistic, but also of the subject matter dealt with (and, as such, of the whole discourse), and has been labelled as “elaborate vs. non-elaborate discourse”.

2.4. Interpretation of Factor 4

Finally, the positive features loading on Factor 4 are time adverbials and perfect aspect (> .45), as well as past tense, general adverbs, seem/appear, split auxiliaries, place adverbials, downtoners, and hedges (the last four loading primarily on other factors). Time adverbials are sometimes used as text-internal deictics, but more commonly refer to times outside the text (Biber 1988: 110). Their grouping with place adverbials and general adverbs reminds us of a similar cluster in Biber’s (1988) Dimension 3, indicating situation-dependent reference. On the other hand, time adverbials, perfect aspect and past tense also clearly indicate a narrative type of discourse, replicating to some extent Biber’s (1988) Dimension 2, “Narrative vs. non-narrative concerns”. In the following examples (6.18) and (6.19), time adverbials are in *italics*, past tense and perfect aspect verbs are in **bold**, and other time and place references are underlined:

> (6.18) Thus along the sun 's path it **became** possible to select a number of stars over which the sun **passed**, and which **would** by their position mark his route in the heavens. To aid in this investigation, as well as for some other purposes, the ancients **erected** a vertical staff on a level plane, and *then* **noted** where the shadow of the top of the staff **fell** at noon each day throughout the year. This instrument **was** called a gnomon, and its use

> **revealed** many important facts in the solar motion, and **detected** others *hitherto* overlooked (astr34)

> (6.19) The powers of song in some individuals of the Wood Thrush **have** *often* **surprised** and **delighted** me. Of these I remember one, many years ago, whose notes I could *instantly* recognize on entering the woods, and with whom I **had been** as it were **acquainted** from his first arrival. The top of a large white oak that **overhung** part of the glen, **was** *usually* the favourite pinnacle from whence he **poured** the sweetest melody; to which I **had** *frequently* **listened** till night **began** to gather in the woods and the fire-flies to sparkle among the branches.
>
> But alas! in the pathetic language of the poet. A few days afterwards, passing along the edge of the rocks, I **found** fragments of the wings and broken feathers of a Wood Thrush killed by the Hawk, which I **contemplated** with unfeigned regret, and not without a determination to retaliate on the first of these murderers I could meet with (life22)

Both examples present a narration in the past tense. Extract (6.18) describes an astronomical experiment, carried out in the ancient times, whereas extract (6.19) reports a series of observations related to the behaviour of a bird. Past tense and perfect aspect indicate a succession of events in the past, while time adverbials help to connect those events (*the ancients erected / and then noted*; *he poured the sweetest melody / to which I had frequently listened*). Likewise, other temporal and place references help to organise this narration in time and space: *A few days afterwards* [time reference], *passing along the edge of the rocks* [place reference], *I found...*

Nevertheless, a less clear component among the positive features of Factor 4 is *seem/appear*, which is usually found in present tense in the texts. However, the following example (6.20) reveals that in some cases *seem* and *appear* are found in a narrative context, with other verbs and also modal auxiliaries in the past tense (which are underlined, whereas *seem* and *appear* are in **bold**), even though the latter were not included in the past tense query (see Chapter 4):

> (6.20) The movements of the planets were to the ancients extremely complex. Venus, for instance, was sometimes seen as “evening star” in

> the west, and then again as “morning star” in the east. Sometimes she **seemed** to be moving in the same direction as the sun, then going apparently behind the sun, **appeared** to pass on again in a course directly opposite. At one time she would recede from the sun more and more slowly and coyly, until she would **appear** to be entirely stationary; then she would retrace her steps, and **seem** to meet the sun (astr37)

By contrast, the negative features loading on Factor 4 are *be* as main verb, present tense, and phrasal coordination, all of which have values > .40. As explained in Biber (1988: 106), “*be* as main verb is typically used to modify a noun with a predicative expression, instead of integrating the information into the noun phrase itself” (as in *the house is big*, as opposed to *the big house*). Present tense, in turn, usually appears in non-narrative discourse and is often used in descriptions or for referring to states of the art and either immediate or habitual or typical events or behaviours:

> (6.21) But a Philosopher, who **proposes** only to represent the common Sense of Mankind in more beautiful and more engaging Colours, if by Accident he **commits** a Mistake, **goes** no farther; but renewing his Appeal to common Sense, and the natural Sentiments of the Mind, **returns** into the right Path, and **secures** himself from any dangerous Illusions (phil10)

Finally, phrasal coordination is also often used in descriptions, functioning as a mechanism of unit expansion. In examples (6.22) and (6.23), verbs in present tense are in **bold**, *be* as main verb is additionally **underlined**, and phrasal coordination is in *italics*:

> (6.22) The Panther **is** of the species or kind of Cats, **is** near as large as the Tiger, and much of the same shape, the Skin **is** of a reddish or whitish Colour, finely mottled with small round black Spots, and the Hair **is** *short and mossy*. It **is** said, all four-footed Beasts **are** wonderfully *delighted and enticed* by the smell of the Panther, but that their frightful Countenances soon **scars** them away, wherefore they **hide** their Heads 'till they **come** within reach of their Prey, which they **leap** upon and quickly **devour** (life8)

> (6.23) THIS Fly **is** somewhat *larger and blacker* than a Cock-Roch; and **derives** its Name from the tinkling Noise it **makes**. The Head and Back **are** *hard and shining*; the former divided from the latter by a broad close Joint. As it *bends* its Head backward, the upper Joint **falls** as a regular Spring into the Socket of the lower; and, when it **bows** its Head forward, it **opens** with a sharp tinkling Note, as the Spring of the outward Case of a Watch, when pressed (life10)

Although this dimension, represented by Factor 4, seems in principle more complex (and probably less clear) than Biber's (1988) Dimension 2 in that some of the positive features listed above (e.g. *seem/appear*) are not very easy to fit in the model, most of its features indicate a dichotomous relationship between presence and absence of narrativity, which is why we have labeled it "Narrative vs. non-narrative discourse".

Our model, therefore, consists in the following four dimensions of variation: Dimension 1, which deals with communicative style, ranging from involved and persuasive to informational; Dimension 2, which accounts for the scientific focus of the texts, which in turn can be either argumentative or descriptive; Dimension 3, which measures the degree of complexity, or elaboration, of the discourse; and Dimension 4, which marks narrativity and classifies texts as narrative or non-narrative (and more or less narrative, depending on each case). Table 6.2 below summarises this information:

Table 6.2

Dimension	**Deals with**	**Label**
Dimension 1	Communicative style	"Involved/persuasive vs. informational style"
Dimension 2	Scientific focus	"Argumentative vs. descriptive focus"
Dimension 3	Discourse elaboration	"Elaborate vs. non-elaborate discourse"
Dimension 4	Narrativity	"Narrative vs. non-narrative discourse"

Four dimensions of variation for a sample of the *Coruña Corpus* (including *CETA*, *CEPhiT*, and *CELiST*)

Each of these four dimensions of variation will be now used to describe the relations among the three scientific disciplines and eight genres present in our corpus, each of them being analysed as subregisters.

3. Textual relations among scientific disciplines and genres

The descriptive statistics (that is, mean dimension score, minimum and maximum scores, range and standard deviation) for each subregister (i.e. discipline and genre) in the two centuries with respect to each dimension of variation have been included in Appendix II. The range is the difference between the minimum and maximum scores, whereas the standard deviation measures the spread of the distribution – that is, how widely the scores of the texts in a subregister are scattered, or how tightly they are grouped, around the mean score. Large standard deviations indicate high variability within a subregister, while small ones show that most texts have a dimension score close to the mean.

At this point a note on the distribution of the genres in the corpus must be made. Table 3.3. is repeated here for convenience as Table 6.3, showing the distribution of genres in the eighteenth and nineteenth centuries:

Table 6.3

Genre	**18th century**		**19th century**	
	Texts	**Words**	**Texts**	**Words**
Treatise	34	338,138	27	271,020
Textbook	12	124,200	8	82,107
Essay	9	92,231	5	50,330
Lecture	1	9,939	11	110,434
Article	1	4,240	6	49,617
Letter	2	20,051	3	31,499
Dialogue	1	9,907	1	10,084
Other (dictionary)	1	10,044	-	-
Total corpus	**61**	**608,750**	**61**	**605,091**

Distribution of genres across two centuries

As we have seen earlier in Chapter 3, while the *Coruña Corpus* has a balanced representation of the scientific disciplines (with twenty texts per century in Philosophy and Life Sciences, and twenty-one in the case of Astronomy), the eight genres present in the corpus are not distributed in an equitable way. As has been noted in Puente-Castelo (forthcoming), this unequal distribution of genres appears to be a result of choosing representativeness over balance at the time of compiling the different subcorpora of the *Coruña Corpus* and seems fairly justified if, on the one hand, we assume that such a distribution "broadly reflects production at the time" (Görlach 2004: 1; Moskowich 2012: 42) and, on the other hand, if we consider the difficulty to achieve a balanced representation of the genres while meeting, at the same time, the strict compilation principles of the corpus (Camiña-Rioboo 2012: 97). However, the lack of an equitable distribution of the genres makes it impracticable to use certain statistical tests for comparing means, such as ANOVA, that are normally used in multidimensional studies. As a consequence, despite the fact that this study is based on a multivariate statistical technique and may therefore be considered statistically reliable to the extent to which (i.e. 41 per cent) the resulting model explains the variation present in the corpus, the results discussed in the following sections will be regarded as tentative until they can be statistically proved in further research.

3.1. Relations along Dimension 1 "Involved/persuasive vs. informational style"

Figure 6.1 (earlier Figure 5.4 in Chapter 5) plots the mean dimension scores for three scientific disciplines – Astronomy, Philosophy and Life Sciences – in the eighteenth and nineteenth centuries along Dimension 1, "Involved/persuasive vs. informational style".

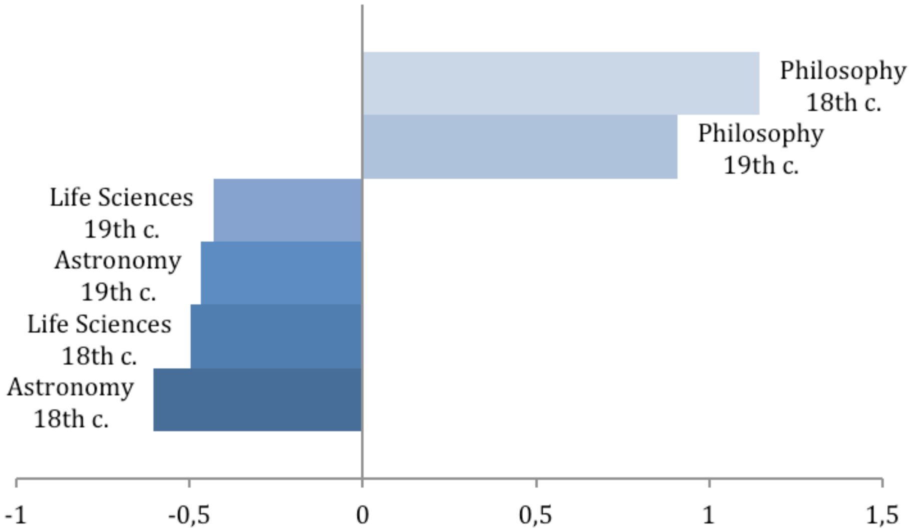


Figure 6.1
Relations among three scientific disciplines across two centuries along Dimension 1 "Involved/persuasive vs. informational style" (descending order)

The x-axis represents the dimension, with positive scores on the right and negative scores on the left. Thus, disciplines characterised by an involved, argumentative discourse appear on the right side (with positive values), whereas those tending towards an informational style appear on the left (with sub-zero values). Zero, in turn, represents the overall corpus mean score for Dimension 1. This means that the closer the mean score of a text, or register, is to zero, the less marked that text or register is with respect to Dimension 1 (i.e. neither involved or informational, having a more or less balanced distribution of positive and negative features).

As can be seen on Figure 6.1, and as it had already been forwarded in the previous chapter, both eighteenth- and nineteenth-century Philosophy show a clearly involved style, in contrast with Astronomy and Life Sciences, which appear to be rather informational disciplines in both centuries. This involved vs. informational distinction appears to mark a first clear difference between the humanities and the natural sciences. If we consider the scientific disciplines in question, the distinction may seem quite logical. Philosophy is a discipline based on the development of thoughts and often controversial ideas, which are usually conveyed through a debate, entailing an argumentative – and therefore persuasive – kind of discourse. Moreover, the expression of intimate thoughts often means that the text is written in the first person, which indicates a clear involvement of the author in the matters discussed. This picture coincides with that of Gray's (2011: 140) MD analysis of present-day English academic writing, where Theoretical Philosophy appears likewise isolated with a very high involvement score. By contrast, Astronomy and Life Sciences are disciplines based on observation (and experiment, to some extent), rather than reasoning, and their focus is to convey a great deal of information with minimal personal involvement.

This distinction may be observed in examples (6.2) and (6.7a), belonging to eighteenth-century Philosophy and Life Sciences, respectively, already been discussed in the previous section and are repeated here for convenience. (Example (6.7a) is an extended version of example (6.7).) As above, positive and negative features for Dimension 1 will be always marked in the following ways, unless specified otherwise:

- Positive: First person pronouns are in **bold**, third person pronouns are thick underlined, possibility modals are in *italics*, private and public verbs are underlined, the pronoun it is in ***bold italics***, indefinite pronouns are double underlined, analytic and synthetic negation is highlighted in dark and light grey, respectively, and necessity modals are dashed underlined, and existential *there* is dot-dash underlined. Likewise, infinitives introducing clauses complementing verbs and adjectives are waved underlined, WH-relative clauses are in square brackets [], conditional clauses are in curled brackets {}, and direct questions are in SMALL CAPS.

- Negative: Nouns are **in bold**, present-participial WHIZ-deletions are underlined, past-participial whiz-deletions are between square brackets [], place adverbials are in *italics*, and past participle clauses are between curled brackets {}; those nouns counted as nominalisations, though formally not included in Dimension 1, fulfil the same grammatical function as other nouns, and are therefore marked with an asterisk (*):

> (6.2) {If the will *can* not set ***itself*** in motion,} ***it*** must be moved by something else. **I** agree with Mr. Locke in thinking, that the only stimulus to action is some pressing uneasiness [which the mind feels], prompting ***it*** to change ***its*** present state. Says he, This uneasiness then **I** consider to be the immediate cause of volition, and absolutely essential to every act of the will. But WHENCE DOES THIS UNEASINESS ARISE? (phil20)

> (6.7a) The **absence** of the **rete mirabile**, and of all analogous provision* for moderating the **influx** of the **blood** into the **brain**, accords, with the other **circumstances** [enumerated above], in showing that **man** is entirely unfit for the **attitude** on all **fours**. In most **animals**, the great occipital **foramen** is placed at the back of the **head**; the **jaws** are considerably elongated; the **occiput** forms no projection* beyond this **opening**, the **plane** of which is vertical, or at least very slightly inclined. (life24)

Likewise, examples (6.24) and (6.25) below show that both eighteenth- and nineteenth-century Astronomy also used a detached kind of writing, characterised by strong informational density:

> (6.24) Again, in the **descent** of this **Particle** downwards from **E**, being still in the action* of the **sun**, its **descent** will be hastened; and therefore, instead of going to **c**, as it would otherwise have done, it will cut the **Plane** of the **Ecliptic** in some other **point** nearer to **B** than **c**, as, e.g. at **e**. But wherever this **Ring** cuts the **Plane** of the **Ecliptic** is the next **Node**, {as was observed before with regard to the **Moon**} (astr14)

> (6.25) The question* is not yet settled, and no collation* of **data [**obtained from small portions* of the **earth's surface]**, and covering short **periods** of **time**, can ever settle it. It requires observations* from all **parts** of the **world** and covering several **sun-spot periods** to form the foundation* of any safe conclusion*. It is to be noted that the distinguished French **astronomer**, Faye, has recently called in question* even the connection* between the **sun-spot period** and the magnetic **state** of the **earth**, a relation* which has been considered as perfectly demonstrated for the last thirty **years** (astr39)

On the other hand, it appears from Figure 6.1 that all three disciplines tend towards a less marked style with respect to Dimension 1 in the nineteenth century, Philosophy becoming somewhat less involved, and Astronomy and Life Sciences becoming less informational. Although in the previous chapter we have seen that minor differences sometimes depend on the method chosen for estimating factor scores and should, therefore, be interpreted with caution, this apparent general tendency away from the extremes seems to suggest that the involved vs. informational difference decreases progressively as the three scientific registers evolve towards their twentieth-century standards.

Mean dimension scores have also been calculated for each of the eight genres contained in our corpus. The relations among the different genres with respect to Dimension 1 are shown on Figure 6.2:

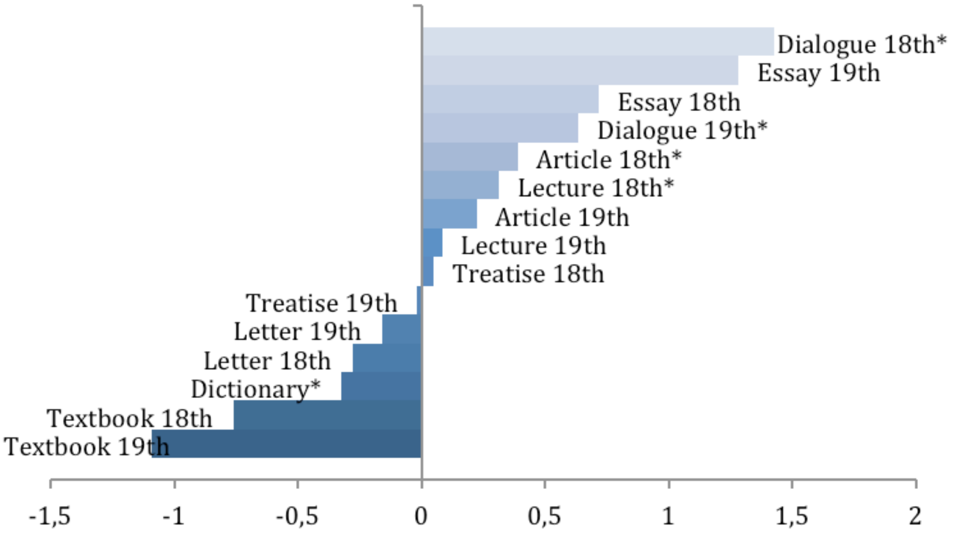


Figure 6.2
Relations among eight genres across two centuries along Dimension 1 "Involved/persuasive vs. informational style" (descending order)

As we can see, eighteenth-century Dialogue is at the the top of the involved/persuasive list of genres, with a score of 1.43. As we have noted at the beginning of this section, cases such as this one will be treated with discretion because it is not certain to what extent a single text can be representative of the genre to which it belongs. Still, if we consider that the dialogic form entails direct reported speech and a continuous interaction between two conversants, it may not seem surprising that both dialogues included in our corpus are on the involved side of Dimension 1. However, what is still more remarkable, is that, while the dialogue with the lower score (0.63) belongs to the nineteenth-century Philosophy subcorpus, it is the one dealing with astronomical matters in the eighteenth-century the one that appears to contain the highest proportion of involvement features. This can be seen in example (6.26) below, which contains an extract from *Astronomical Dialogues between a Gentleman and a Lady* by John Harris (1719):

> (6.26) But **I** have a great Mind to learn, from **my** Friend, something of the Nature and Use of them; for they appear to be made and finished up with that Curiosity and Care, that sure some very useful Knowledge is to be learnt from them, and IS ***IT*** NOT BARBAROUS IN YOU MEN TO CONFINE *IT* ALL TO YOUR SELVES?

> MADAM, said **I**, you will give **me** a new Rise to value any thing that **I** understand; if **I** *can* render ***it*** acceptable to you.
>
> WELL then, Sir, said she, all Compliments apart, both to your self and **me**, pray let **us** go to **our** Business, the Tea won't be ready this Hour, and there is a little too much Dew for **us** to take a Walk in the Garden. Let **me** understand then, first the Difference between these two Globes, and why one hath the Cities, Countries, and Places of the Earth drawn on ***it***, like a Map; and the other Circles and Stars, and these odd uncouth Figures of Beasts, Birds and Fishes: PRAY WHY DO **THEY** TURN ROUND? WHAT DOTH THIS BRASS HOOP SIGNIFY IN WHICH THEY HANG? For **I** perceive that ***it*** also hath Numbers engraved upon ***it***: And WHAT DOTH THIS BROAD WOODEN THING SERVE FOR, THAT HATH THE DAYS OF THE MONTH AND OTHER LETTERS, AS WELL AS FIGURES, PASTED UPON ***IT***?
>
> **I** am glad, said **I**, Madam, by the warm Manner of your Enquiry, to find that you are in earnest; and **I** have often wished that the same Curiosity and Love of Knowledge would inspire more of the fair Sex… (astr4)

The next most involved/persuasive subregister appears to be Essay. This does not seem too surprising, considering that ten of the fourteen essays in our corpus belong to the Philosophy subset. The following example (6.27), an extract from a philosophical essay by Henry Bolingbroke (1754), reflects the characteristics of both discipline and genre, which seem to have an ideal form-content relationship, considering that the essay is regarded as an open form established for the deliberation over certain thoughts or reflections on a subject (see definitions in Chapter 3):

> (6.27) **I** mean, that {if **our** senses were able to discover to us the inmost constitutions, and the real essences of outward objects}, such senses would render **us** unfit to live and act in the system to which we belong.{If the system was not made for **us**, [who pretend on very weak grounds, **I** think, to be the final cause of it,]} **we** at least were made for the system, and for the part **we** bear among terrestrial animals. Other creatures *there may be*, and, **I** believe readily, *there are*, [who have finer senses than men,

> as well as superior intelligence to apply and improve the ideas they receive by sensation] (phil11)

However, it is also remarkable that, paradoxically, essays become more involved in the nineteenth century while the overall mean dimension score for Philosophy texts decreases. Although we only have five essays written in the 1800s (four in Philosophy and one in Astronomy), it appears that the Essay subregister tends to reinforce, rather than lose, its personal and argumentative style with time. The following example (6.28), taken from Arthur J. Balfour's *A Defence of Philosophic Doubt* (1879), illustrates this tendency:

> (6.28) As **we** have seen, the ultimate beliefs which *may* or rather must be accepted with confidence are, according to him, of two kinds: the beliefs **we** have respecting **our** own actual mental states, and the beliefs, {if any,} [which are part of the original furniture of the mind]. He frequently asserts that **we** hold both these kinds of belief on the authority of consciousness. ARE **WE** THEN TO ATTRIBUTE TO HIM THE THEORY WHICH **I** HAVE ATTRIBUTED TO SIR WILLIAM HAMILTON THE THEORY, I MEAN, THAT CONSCIOUSNESS IS AN INTERNAL WITNESS WHICH MUST BE DISTINGUISHED LIKE OTHER WITNESSES FROM THE STATEMENTS TO WHICH IT CERTIFIES? (phil36)

Article, Lecture and Treatise, nevertheless, appear to evolve in the opposite direction, becoming less involved/persuasive in the nineteenth century, but still staying on the right (or involved) side. This seems to suggest that the most formally scientific genres shifted towards the informational end of the scale gradually and slowly, with Article and Lecture being still far in the nineteen hundreds from the conventionally impersonal equivalents of these genres established as a standard in the twentieth century. Still, just as is the case with dialogues, even if we consider the only eighteenth-century article and lecture included in our corpus as a possible reference, they should not be necessarily regarded as representative of the genre. Treatise, staying in the middle of the axis, appears not to be markedly involved nor informational, having, overall, around fifty per cent of each component in the mean dimension score. Even so, some treatises, such as the one represented in example (6.29), have a larger amount of involved/persuasive features, whereas others (see example (6.30)), appear to have a more informational nature, which seems to be owing to the scientific disciplines to which they belong (Philosophy and Astronomy, respectively):

> (6.29) To complain of the age **we** live in, to murmur at the present possessors of power, to lament the past, to conceive extravagant hopes of the future, are the common dispositions of the greatest part of mankind; indeed the necessary effects of the ignorance and levity of the vulgar. Such complaints and humours have existed in all times; yet as all times have not been alike, true political sagacity manifests ***itself***, in distinguishing that complaint, [which only characterizes the general infirmity of human nature,] from those [which are symptoms of the particular distemperature of **our** own air and season.] Nobody, **I** believe, will consider ***it*** merely as the language of spleen or disappointment, if **I** say, that there is something particularly alarming in the present conjuncture (phil15)

> (6.30) **Azimuth** is the **Distance** betwixt the North **Point** of the **Horizon**, and that **Point** where a Vertical **Circle** passing through the **Body** of the **Sun** or **Star**, cuts the **Horizon**, or the **Distance** betwixt the Prime **Vertical**, and the **Vertical** the **Sun** or **Star** is upon. **Altitude** of the **Sun** or **Star**, is an **Arch** of an Azimuth **Circle** [comprehended between the **Horizon** and the **Parallel** of **Altitude** the **Sun** or **Star** is upon] (astr5)

What appears to be surprising, however, is that both eighteenth- and nineteenth-century letters appear on the informational side of Dimension 1. Indeed, one would expect the Letter genre to contain more features common in personal interaction, and yet dimension scores show the contrary picture. Again, this is probably due to the fact that most letters included in our corpus belong to the natural sciences (i.e. Astronomy and Life Sciences), which explains the highly informational character of some of them (see example (6.31)). Moreover, it must be taken into account that the epistolary genre in science was little more than one of the several conventional forms of transmission of scientific knowledge used in the late Modern period (see Atkinson 1999: 81-84) and, as such, contained a rather formal kind of register:

> (6.31) The learned **societies** [established in various **centres** of civilization*] have more especially directed their attention* to the advancement* of physical **astronomy**, and have stimulated the **spirit** of **enquiry** by a **succession** of **prizes**, [offered for the solutions* of **problems** <u>arising out of the difficulties* which were progressively developed by the advancement* of astronomical **knowledge**</u>.] Among these questions*, the determination* of the **return** of **comets**, and the **disturbances** which they experience in their **course**, by the action* of the **planets** near which they happen to pass, hold a prominent **place** (astr30)

Finally, Dictionary and Textbook appear to be the most informational subregisters. Although John Hill's *Urania* (1754) is the only dictionary present in our corpus, one of the definitions of dictionary applicable to the present case – "a book of information or reference on any subject in which the entries are arranged alphabetically; an alphabetical encyclopedia" (*OED*) – suggests that dictionaries are informational by nature. Textbook, in turn, is a genre used for didactic purposes and, as such, is characterised by a register conceived for a maximally efficient transmission of information (see example (6.32)):

> (6.32) Observe *here*, that as any **Place**, **Town** or **City** on **Earth** is found and determined by the **Parallel** of its **Latitude** <u>crossing its **Line** of **Longitude**</u>; so the proper **Place** of the **Sun** or **Star** in the **Heavens** is found and determined by the **Point** where its **Parallel** of Declination* crosses its **Meridian** or **Line** of Right **Ascension**; which indeed are but the self same **things** on both the **Globes**… (astr6)

Thus, we have seen that Dialogue and Essay are the most involved and persuasive genres, followed by Lecture and Article, while Letter, Dictionary and Textbook are clearly informational. Treatises, in turn, can be both types, which up to a point may depend on the scientific discipline they belong to. Except for essays and letters, which become more involved in the nineteenth century, all the other genres appear to move in the informational direction with time. This seems to suggest that, while the general tendency is the progressive impersonalisation (as a way of standardisation), of scientific discourse, Essay and Letter follow the contrary path (which may be one of the reasons why they will not become prototypical scientific genres). The gradual replacement of the first person and persuasive features by an impersonal, passivised and nominal style in the scientific texts analysed in this study appears to go hand in hand with Atkinson's (1999: 78-80) findings that scientific discourse shifts from author-centered to object-centered towards the end of the nineteenth century. This suggests that Atkinson's (1999) characterisation of the article genre with respect to the "involved/informational" dichotomy can be extended to other genres of scientific writing of that time. As we had seen in Chapter 2, this was the period when the need to persuade the reader of the trustworthiness of the experiment decreased as the experiment itself, rather than the scientist who carried it out, became the focus, and it is therefore in natural sciences (which are based on observation and may involve experiments) where this shift in style was more patent. Philosophy, on the contrary, being a discipline dialectical by nature, preserved to a much larger extent its involved character along the nineteenth century, and, according to Gray's (2011: 140) findings, still shows involvement in the present day.

Textual relations with respect to Dimension 2 will be analysed in what follows.

3.2. Relations along Dimension 2 "Argumentative vs. descriptive focus"

The mean Dimension 2 scores for Astronomy, Philosophy and Life Sciences in the eighteenth and nineteenth centuries are plotted on Figure 6.3 (on the right). The right extreme of the x-axis indicates a high proportion of positive features, interpreted as conveying argumentation or logical reasoning, while the left end indicates a clustering of negative features, characterising descriptive texts.

Figure 6.3 shows that Astronomy (especially, its eighteenth-century part) contains a high proportion of features expressing logical relationships, whereas Life Sciences appears to be a descriptive discipline, both in the 1700s and 1800s. This may seem curious, considering that both disciplines may, in principle, be characterised as observational sciences.

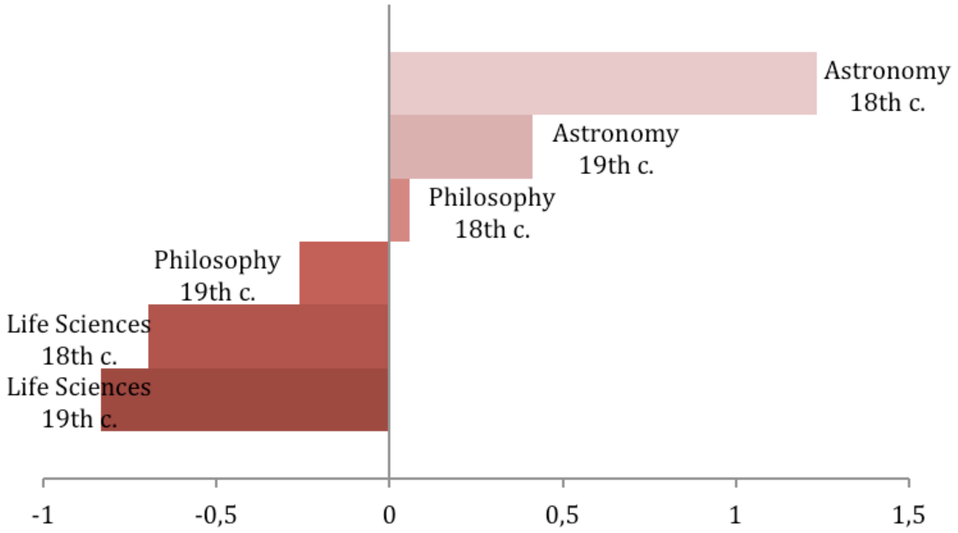


Figure 6.3
Relations among three scientific disciplines across two centuries along Dimension 2 "Argumentative vs. descriptive focus" (descending order)

Examples (6.33) and (6.34), containing excerpts from eighteenth-century Astronomy and Life Sciences texts, respectively, illustrate their difference with respect to Dimension 2, in that the observations of the former are primarily based on mathematics. In the first example, containing mainly positive features, conjuncts are in **bold**, predictive modals are in *italics*, causative adverbs are underlined, and other adverbial subordinators are thick underlined; conditional subordination, when present, will be marked with curled brackets {}. In the second extract, which focuses on negative features, attributive adjectives are in **bold**, predicative adjectives are in *italics*, and hedges, downtoners and amplifiers are all underlined. All the subsequent examples of positive and negative features for Dimension 2 will be marked, accordingly, in the same way, unless specified otherwise.

> (6.33) Again, because the square of Aye is to the square of TA as TA to TG; **therefore** Aye *will* be to TA in the subduplicate ratio of TA to TG; **therefore** the rectangle contained by Aye, Tx, *will* be to the rectangle contained by TA, Tx, in the subduplicate ratio of TA to TG: and because the rectangle contained by TA, Tx, is to the square of TA as Tx to TA; that is, in the subduplicate ratio of TP to TA, (because TP, Tx, TA, may be considered as proportionals); **therefore** the rectangle contained by Aye, Tx, *will* be to the rectangle contained by TA, Tx , in the subduplicate

> ratio of TA to TG: and because the rectangle contained by TA, Tx, is to the square of TA as Tx to TA; that is, in the subduplicate ratio of TP to TA, (because TP, Tx, TA, may be considered as proportionals); **therefore** the rectangle contained by AYE, Tx, *will* be to the square of TA in the subduplicate ratio of TP to TG; that is, in the subduplicate ratio of the cube of TP to the parallelopiped whose base is the square of TP, and altitude TG… (astr13)

> (6.34) JOHN COOK'S HORSE, or HAG'S HORSE. THOUGH this hath all its Limbs in Perfection; yet it is so **shapeless** an Animal, that, without a **narrow** Inspection, it can hardly be distinguished at first from a **dry half-rotten** Piece of Straw of about Three Inches long. Its Legs, which are Four in Number, are very near as *fine* as those of a **large** Spider. It seems to be every way very *inoffensive*; and it is generally to be found upon Shrubs and Bushes. A great many Negroes have a Notion, that, if they kill one of these, they will be very *unlucky* in breaking all **Earthen** Wares they handle: Of this they are so strongly persuaded, that I have seen a Negro Wench suffer a Whipping, rather than, when commanded to do it, kill one of them. The whole Body and Legs are speckled alternately with a **russet** Brown, and a **dull** White; but not *discernible* at any **great** Distance (life10)

If we look at the diachronic evolution of Astronomy, Philosophy and Life Sciences along Dimension 2, all three disciplines appear to have shifted in the negative direction in the nineteenth century. The progressive abandonment of the expression of relationships of conditionality and causality, albeit at different speeds in the three scientific disciplines, appears to partially coincide with the findings of Puente-Castelo (forthcoming), who suggests that the scientific discourse in the 1700s was still very much influenced by the logical frames of reasoning characteristic of Scholasticism. Although the distribution of positive and negative features in nineteenth-century Astronomy and both eighteenth- and nineteenth-century Philosophy is somewhat more balanced, closer to the average, Philosophy presents a positive score in the 1700s a negative one in the 1800s. This does not mean, though, that all nineteenth-century Philosophy texts are predominantly descriptive. The following extract from the philosophical lecture by Andrew Seth (1885), reproduced in examples (6.35) and (6.46) below, shows that both components may be present, even if neither appears to be very abundant:

> (6.35) Negative features:
>
> Berkeley would have been *ready* to admit that it must at least be referred to me as mine that this relation, therefore, at the lowest is *necessary* to render it *knowable*. But the unknowableness of sense-atoms or **mere** data, except as somehow related to one another, had not forced itself upon him at the date of his **epoch-making** works. He says unhesitatingly in the "Principies" Section 89 that He is thinking, of course, of the **undoubted** truth that we may first consider an object by itself, as we say, and then add to this survey a consideration of its relations to **other** things to its environment, for example, or the past of which it is the outcome. But is the thing, as originally considered, absolutely without relations? On the contrary, it is simply *impossible* to consider anything in **sheer** isolation from its **temporal** and **spatial** environment… (phil37)

> (6.36) Positive features:
>
> Berkeley *would* have been ready to admit that it must at least be referred to me as mine that this relation, **therefore**, at the lowest is necessary to render it knowable. But the unknowableness of sense-atoms or mere data, except as somehow related to one another, had not forced itself upon him at the date of his epoch-making works. He says unhesitatingly in the "Principies" Section 89. that He is thinking, of course, of the undoubted truth that we may first consider an object by itself, as we say, and then add to this survey a consideration of its relations to other things to its environment, **for example**, or the past of which it is the outcome. But is the thing, as originally considered, absolutely without relations? **On the contrary**, it is simply impossible to consider anything in sheer isolation from its temporal and spatial environment… (phil37)

The distribution of positive and negative Dimension 2 scores by genres shows a more detailed picture (see Figure 6.4 on the next page):

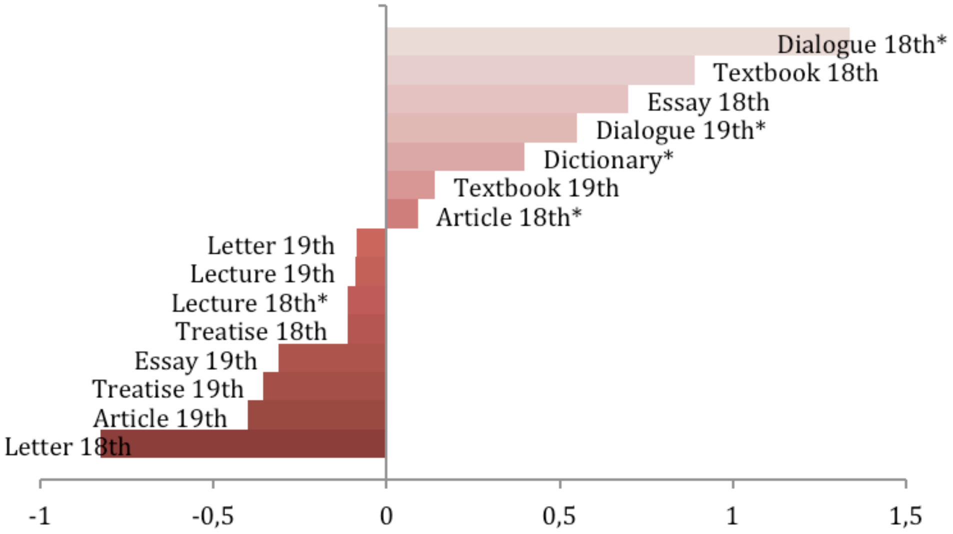


Figure 6.4
Relations among eight genres across two centuries along Dimension 2 "Argumentative vs. descriptive focus" (descending order)

Here, subtler differences (which could not be spotted by looking at disciplines only) may be appreciated. Once more, the eighteenth-century dialogue included in the Astronomy subcorpus has the highest dimension score. In this case, the explanation for this relatively large proportion of positive features, associated with logical reasoning, seems to lie in the scientific discipline, rather than in the genre, although the function of the latter in the scientific literature of the time may also play an important role. In the following extract (6.37), the positive features are marked as above, and the conditional subordinate clause is between squared brackets {}:

> (6.37) VERY many and substantial ones, Madam, said I, and you *will* be fully convinced by them, when they occur to your Reading hereafter, {if you proceed on in that Way you are now going}: But, **however**, the Sun shining so bright into this Room, *will* furnish me now with one Argument to make that Notion plain to you. You see, Madam, when I hold any solid Body in this Light of the Sun, its Shadow *will* be nearly like the Shape and Form of that of the Body; when I hold this Book in the Light, its Shadow *will* be square at the Sides, as the Book is; but when I hold this Orange in the same Light, the Shadow, you see, hath a round Edge; and **therefore** <u>since</u> in the Eclipses of the Moon, the Shadow of the Earth, which you know, Madam, occasions the Moon 's being covered with

> Darkness, appearing always exactly round or circular, we justly conclude that the Figure of the Earth is round or spherical too, or **else** the Termination or Out-Line of its Shadow could never be always in a Circular Form (astr4)

As we can see in the above excerpt, the text contains mainly features of involvement and interaction from Dimension 1, which characterise the dialogical register, but also conjuncts and adverbial subordinators, which organise the ideas in the text. If we bear in mind that dialogues in science served a didactic purpose, and that the subject matter of this one is Astronomy, it may not be surprising that logical markers and predictive modals are used to express a rationally connected sequence of events. Textbook, which is another genre adopted for educational purposes, also has a remarkably high score on Dimension 2 in the eighteenth century. Example (6.38) is an excerpt from George Costard's *History of Astronomy* (1767), while example (6.39) belongs to Adam Ferguson's *Institutes of Moral Philosophy* (1769):

> (6.38) **However**, as in the case of the waters, above explained, the whole effect must be ascribed to the joint actions of both Luminaries, which, **therefore**, *will* be greatest when they conspire together, and least when the action of one checks, or, in part, counterbalances the action of the other. The transverse axis of the spheroid of the Atmophere *will* be longest when it passes through the centers of the Sun and Moon at, or near, the Equator; and **therefore** the greatest forms *will* be about the times of the two Equinoxes (astr14)

> (6.39) And in this sense every law must be strictly observed; <u>because</u> it is law only <u>so far as</u> it is observed. Gravitation is a law only <u>because</u> all bodies actually gravitate. But in this sense, too, the intellectual system hath its laws; for there are facts relating to the operations of mind which are fixed and invariable. In this sense, **therefore**, the laws of the intellectual system are equally well observed with those of the material. The term law, **however**, has a farther signification, and means some rule of choice, or expression of what is good (phil14)

Eighteenth-century essays also present a very large proportion of positive features; however, in the nineteenth century, essays appear on the other end of the scale, with a negative dimension score. Examples (6.40) and (6.41) below, extracts from eighteenth- and nineteenth-century essays, respectively, illustrate this difference:

> (6.40) **Because** CE is less than CD, the ratio of BC to CE *will* be greater than the ratio of BC to CD; therefore, by composition, the ratio of BE to EC *will* be greater than the ratio of BD to DC; therefore the rectangle contained by BE, DC, *will* be greater than the rectangle contained by CE, BD; that is, the rectangle ABE *will* be greater than the rectangle contained by CE, BD: but the rectangle AFE is equal to the rectangle contained by CE, BD; therefore the rectangle ABE is greater than the rectangle AFE; therefore HAVE is greater than OF; and therefore BD is greater than IF. Again, **because** the rectangle AFE is equal to the rectangle contained by BD, CE, BD *will* be to IF as OF to CE; and **because** BD is greater than IF, therefore OF is greater than CE (astr13)

> (6.41) The pursuit of such a "**high** priori road" has in **modern** times fallen somewhat into discredit, especially with regard to questions which have a distinctly **practical** bearing. And indeed it is evident that the application of **philosophical** principles to such questions must be expected to be among the **latest** results of **philosophic** study, and that it will be *dangerous* to attempt to apply them before we have succeeded in making our **first** principles thoroughly *clear* and *certain*. This we can hardly hope to see immediately accomplished; and consequently it would seem that our method of investigation must for the present be somewhat more *tentative* in its character. Yet it seems equally *evident* that, until we can secure such a **systematic** method of study as has now been indicated, there can not be, in any **proper** sense of the word, a **Social** Philosophy (phil38)

Although example (6.40) belongs to the Astronomy subcorpus and contains a great deal of mathematical reasoning throughout, the following excerpt (6.42) from an eighteenth-century Philosophy essay by Henry Bolingbroke (1754) in which the existence of God is debated through the undisputable arms of logic, likewise shows a presence of positive features. Notice that, in this extract, causative conjunctions other than *because* (i.e. *for*) have been highlighted, despite the fact that they have not been counted as such in the searches:

> (6.42) I RETURN to the subject immediately before me, and I say, that, since there must have been something from eternity, **because** there is something now, the eternal Being must be an intelligent Being, **because** there is intelligence now; **for** no man *will* venture to assert that non-entity can produce entity, or non-intelligence, intelligence: and such a Being must exist necessarily, whether things have been always as they are, or whether they have been made in time; **because** it is no more possible to conceive an infinite, than a finite, progression of effects without a cause. Thus the existence of a God is demonstrated; and cavil against demonstration is impertinent (phil11).

The eighteenth-century Dictionary, in turn, also contains a high frequency of elements of logical reasoning and argumentation which help to organise its densely informational discourse:

> (6.43) ALIQUOT part. A part of any number, or of any quantity, which, being repeated a certain number of times, *will* produce the whole quantity. **Thus**, in numbers, three is an aliquot part of twelve, because being four times repeated it produces twelve; and, in measure, a line of a foot long is an aliquot part of a yard, because three times repeated it makes the whole yard. **On the contrary**, five being ever so many, or ever so few times repeated, *will* not make twelve, and **therefore** five is not an aliquot part of twelve, but an aliquant part (astr11)

On the other hand, treatises, lectures and letters, along with nineteenth-century articles, appear on the descriptive side of Dimension 2, showing few features of discourse structuring. This may be explained, partly, by the fact that almost half of the total number of treatises, (26 of a total of 61) and most letters (3 of 5) belong to the discipline of Life Sciences, which is – as we it had been shown earlier – of a predominantly descriptive nature. This may be seen in examples (6.44), (6.45) and (6.46) below, which illustrate the three subregisters in the above-mentioned order:

> (6.44) From such **special** adaptations, the similarity of the larvae or **active** embryos of **allied** animals is sometimes much obscured; and cases could be given of the larvae of two species, or of two groups of species, differing quite as much, or even more, from each other than do their **adult** parents. In most cases, however, the larvae, though active, still obey more or less closely the law of **common embryonic** resemblance. Cirripedes afford a **good** instance of this: even the **illustrious** Cuvier did not perceive that a barnacle was, as it certainly is, a crustacean; but a glance at the larva shows this to be the case in an **unmistakeable** manner. So again the two **main** divisions of cirripedes, the pedunculated and sessile, which differ widely in **external** appearance, have larvae in all their several stages barely distinguishable (life32)

> (6.45) The **prickly** pear exhibits a **thick** and **expanded** stem, which is formed of leaves imperfectly developed. The stamens and pistils through excess of nourishment, swell out, and become petals; all **double** flowers are formed in this manner. The poppy in its **natural** state has many stamens, and but four petals; but you often see **double** poppies, with scarcely the vestige of a stamen left; the same change may be observed in the rose, which naturally has but five petals and many stamens and pistils, but in a very **full**, **double** rose, scarcely any appearance of either stamen or pistil is to be seen (life27)

> (6.46) Their usefulness is also very *important* in preserving a **due** proportion among plants, in consuming what is *dead* or *decayed*, and in yielding a **large** supply of food to other animals; birds and fishes especially, of which they are the **constant** prey. To those who love to indulge their taste with the view of the most **luxurious** and **elegant** objects, this branch of natural history will afford the most **unlimited** gratification, from the **infinite** variety of form and colour, excelled in

> richness and beauty by no part of nature, not even by the **gay** tribes of our **favourite** flowers (life23)

Even so, as we have seen earlier, some texts, including those with a larger proportion of negative features, also have positive ones, although the latter are overshadowed by the former. This can be appreciated in the extract from a nineteenth-century Life Sciences lecture by Thomas Huxley (1863), exemplified through excerpts (6.47) and (6.48). The former focuses on negative features only, whereas in the latter it is the positive ones that are highlighted:

> (6.47) Negative features:
>
> I do not see anything very *wonderful* in the fact, if it took all that trouble to get it from a **wild** state, that it should go back into its **original** state as soon as you remove the conditions which produced the variation to the **domesticated** form. There is an **important** fact, however, forcibly brought forward by Mr. Darwin, which has been noticed in connection with the breeding of **domesticated** pigeons; and it is, that however *different* these breeds of pigeons may be from each other, and we have already noticed the **great** differences in these breeds, that if, among any of those variations, you chance to have a **blue** pigeon turn up, it will be sure to have the **black** bars across the wings, which are *characteristic* of the **original wild** stock, the Rock Pigeon (life33)

> (6.48) Positive features:
>
> I do not see anything very wonderful in the fact, {if it took all that trouble to get it from a wild state,} that it *should* go back into its original state as soon as you remove the conditions which produced the variation to the domesticated form. There is an important fact, **however**, forcibly brought forward by Mr. Darwin, which has been noticed in connection with the breeding of domesticated pigeons; and it is, that however different these breeds of pigeons may be from each other, and we have already noticed the great differences in these breeds, that {if, among any of those variations, you chance to have a blue pigeon turn up,} it *will* be sure to

> have the black bars across the wings, which are characteristic of the original wild stock, the Rock Pigeon (life33)

Eighteenth-century letters, on the other hand, are at the very end of the negative side of Dimension 2. In principle, this agrees with the fact that Life Sciences was already very descriptive in the 1700s (see extract (6.49) below, taken from a letter on birds by Edward Bancroft, published in 1769):

> (6.49) It is somewhat larger than a **common** House Sparrow, and has a **conical**, **straight**, **sharp** bill, of a **light** carnation colour. Its feathers are a **confused** assemblage of all the most **lively** and **beautiful** colours in nature: among these, yellow, scarlet, green, and a **blackish** purple, or indigo colour, have the greatest share: besides these, there are white, black, and blue. All these colours are mixed with such **beautiful** disorder, that it is impossible to convey an idea of their disposition (life14)

However, the following example (6.50) shows that, in this particular case, a letter on Astronomy could also be of a descriptive nature at that time:

> (6.50) And since this **obscure** part is always bounded by a **circular** line, the earth itself, for that reason, must certainly be *spherical*. Because it is *evident*, that none but a **spherical** body can, in all situations, cast a **circular** shadow. Nor are the **little** unevenneses on the earth 's surface, arising from hills and valleys, any **material** objection to its being considered as a **round** body; since the **highest** mountains we are *acquainted* with, bear a less proportion to the whole bulk of the earth than the **small** rings on the coat of an **orange** bear to that fruit; or a grain of sand, to an **artificial** globe of nine inches diameter (astr19)

All in all, once more, if we shift the focus on the positive features in the latter example, we shall see that this excerpt from an astronomical letter by John Bonnycastle (1786) also contains some adverbial subordinators and other operators expressing a relationship of logical causality (see example (6.51) below). Here again, adverbial expressions other than *because* have been highlighted, even if they have not been included in the original query:

> (6.51) And since this obscure part is always bounded by a circular line, the earth itself, for that reason, must certainly be spherical. Because it is evident, that none but a spherical body can, in all situations, cast a circular shadow. Nor are the little unevennesses on the earth 's surface, arising from hills and valleys, any material objection to its being considered as a round body; since the highest mountains we are acquainted with, bear a less proportion to the whole bulk of the earth than the small rings on the coat of an orange bear to that fruit; or a grain of sand, to an artificial globe of nine inches diameter (astr19)

Finally, nineteenth-century articles also contain a relatively large proportion of negative features. This, contrarily to what was the case with treatises and lectures, does not seem to depend on scientific discipline (with only one of the six articles belonging to the Life Sciences subcorpus) but appears to be a characteristic of the subregister; see examples (6.52) and (6.53) below, corresponding to nineteenth-century Philosophy and Astronomy, respectively:

> (6.52) As a **religious** and **moral** being, man is *conscious* of a relation of a **personal** character, *distinct* from any suggested by the phenomena of the **material** world, a relation to a **supreme Personal** Being, the object of his **religious** worship, and the source and judge of his **moral** obligations and conduct. To adopt the name of God in an **abstract** speculation merely as a **conventional** denomination for the **highest** link in the chain of thought, and to believe in Him for the **practical** purposes of worship and obedience, are two very **different** things; and for the latter, though not for the former, the conception of God as a Person is *indispensable*. Were man a being of **pure** intellect, the problem of the Unconditioned would be divested of its **chief** difficulty; but he is also a being of **religious** and **moral** faculties, and these also have a claim to be satisfied by any **valid** solution of the problem (phil34)

> (6.53) Shall he have made the sun to rule by day, the moon by night; shall he have drawn out the “hosts of heaven”, and regulated their **rapid,** yet **calm** and **harmonious** motions, by laws the most **beautiful** and **simple**, and evidently the **mere** extension of those which are in **daily** operation around us; and we be not allowed to investigate these things, because they do not directly place shillings and pence in our pockets? This were not only to extinguish a source of the **highest** pleasure, but to bury some of the **richest** talents *committed* to our care; and to yield up some of the most **ennobling** impulses of our nature to motives of the most **sordid** selfishness (astr27)

We have therefore seen that, with respect to the dimension of language just analysed, the three scientific disciplines appear to have different scientific focuses, with Astronomy and Life Sciences standing at two different extremes: while the former relies mainly on mathematical reasoning, logical argumentation and drawing of conclusions, the latter may be characterised as having a chiefly descriptive focus. Philosophy, in turn, combines both, shifting towards the descriptive side with time. Likewise, Astronomy also decreases in the argumentative component in the nineteenth century. On the other hand, some subregisters, such as Dialogue, Dictionary, or Textbook appear to be predominantly argumentative, while others such as Letter, Lecture or Treatise are more of a descriptive character. This seems to go hand in hand with discipline, in that the first group is mostly present in Astronomy, whereas the second group appears more often in Life Sciences. Similarly, the fact that Essay is markedly argumentative in the eighteenth century but becomes descriptive in the nineteenth appears to coincide with the movement in the ‘descriptive’ direction of the Philosophy subcorpus, which contains most samples of this genre.

In the following section, textual relations with respect to Dimension 3 will be analysed.

3.3. Relations along Dimension 3 “Elaborate vs. non-elaborate discourse”

Figure 6.5 plots the mean scores for Astronomy, Philosophy and Life Sciences in the eighteenth and nineteenth centuries on Dimension 3. As was also the case with Dimensions 1 and 2, positive scores are on the right side, indicating a high frequency of positive features, characteristic of a complex, elaborate kind of discourse, whereas negative values are on the left side, indicating the absence (or quasi-absence) of positive features and the presence of negative features, which characterises non-elaborate discourse.

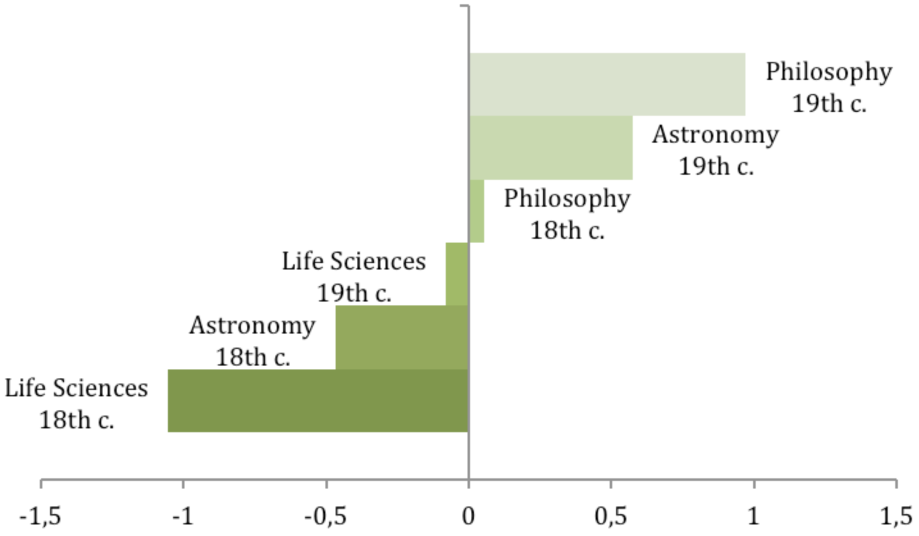


Figure 6.5
Relations among three scientific disciplines across two centuries along Dimension 3 "Elaborate vs. non-elaborate discourse" (descending order)

As shown in Figure 6.5, all three disciplines become more elaborate with time. Life Sciences, however, is predominantly non-elaborate, as neither is eighteenth-century Astronomy, whereas nineteenth-century Astronomy and both eighteenth- and nineteenth-century Philosophy are elaborate. As we have observed earlier in Section 2.3, the low presence of elaborate features in Life Sciences might be related to the subject matters of the discipline, which are, essentially, the different living organisms, whether vegetal or animal, and their characteristics, behaviour, and internal composition. In the eighteenth century, this discipline was predominantly based on the classification and cataloguing of species, which required a fairly simple and straightforward language. In the nineteenth century, however, Life Sciences also gained abstraction and technicality, as shall be seen later in this section. The situation is very different with Philosophy, where the matters dealt with are of an abstract nature and are therefore conveyed through an abstract, densely nominalised and passivised language. All in all, the gap, or difference between the scores in the two centuries appears to be quite large for each discipline, suggesting a sensible diachronic change in the elaborate direction. Moreover, it has been shown in Gray (2011: 117-118) that present-day Philosophy is the scientific register that shows the highest content of structural complexity, which seems to justify the high Dimension 3 score of Philosophy in our study.

In Astronomy this escalation of elaborate features is also present. Examples (6.54) and (6.55) below illustrate this change, through excerpts from eighteenth- and nineteenth- century samples of *CETA*, respectively. In example (6.55), positive features are marked in the following way: nominalisations are in **bold**, prepositions are underlined, past participial WHIZ-deletions are between square brackets [], pied-piping constructions are between curled brackets {}, and agentless passives are thick underlined. Additionally, all the nominalisations not included in Biber's (1988) closed list (see Table 4.1 in Chapter 4) are marked with an asterisk (*). Conversely, in example (6.54), which highlights negative features, present tense verbs are in **bold**, place adverbials are underlined, third and second person pronouns as well as the pronoun *it* are in *italics*, and general adverbs are dashed underlined. The above-mentioned patterns will be followed in all the subsequent examples of positive and negative features for Dimension 3, unless specified otherwise.

> (6.54) …as the Sun, in *his* diurnal Motion, always **moves** parallel to the Equinoctial, *he* **must** be longer above the Horizon than below. By moving the Sun 1 Degree every Day, according to his annual Course, in a quarter of a Year, or about 91 Days from the 10th of March, viz., the 10 of June, *he* will be at the beginning of Cancer, (*his* greatest Declination, then our Days are at the longest) the Sun being in the first Degree of Cancer, *his* Place; bring it to the Meridian, and fix the Index as before, by turning the Sun westward, according to *his* diurnal Motion, we then **see** *he* **sets** about a quarter of an Hour after 8 in the Evening… (astr8)

> (6.55) **Operations** for determining the figure of the earth have been carried out during the present century on an *unprecedented* scale. The Russo-Scandinavian arc, {of which the **measurement** was completed under the **direction** of the elder Struve in 1855,} reached from Hammerfest to Ismailia on the Danube, a length of 2520'. But little inferior to it was the Indian arc, [begun by Lambton in the first years of

> the century,] [continued by Everest,] [revised and extended by Walker.] The *general* upshot is to show that the polar compression* of the earth is somewhat greater than had been supposed. The *admitted* **fraction** until lately was 1/300 that is to say, the **thickness** of the *protuberant equatorial* ring was taken to be 1/300 of the *equatorial* radius. (astr41)

This increase of elaborate features in the Astronomy subcorpus across time agrees with the constant increase of nominalisations in *CETA* registered in Bello (2014), which, in fact, includes a much wider range of nominalisation suffixes than Biber's (1988) list. On the other hand, while both eighteenth- and nineteenth-century Philosophy texts appear to contain a fairly large proportion of elaborate features, the latter seems to present a denser clustering of nominalisations and prepositional phrases, as may be observed in examples (6.56) and (6.57), respectively:

> (6.56) AMONG the many cavils, that have been devised against the *demonstrated* existence* of a first, *intelligent*, *self-existent* cause of all things, this has been one, That things [known] must be anterior to knowledge*; and that we may as well assert that the images of objects we see reflected made those objects, as that knowledge*, or intelligence* made them. HOBBES is accused of reasoning on this principle in his Leviathan, and his book De civ, by the author of the *Intellectual* system of the universe; and his **argument**, in the place where he mentions the **notions** that reason dictates to us concerning the *divine* attributes, is thus stated. Now I think this charge a little too hastily brought, and a little too heavily laid. So will any man who reads the context (phil11)

> (6.57) What we call comparing and weighing reasons are processes* of **consciousness** dependent on these neuro-cerebral processes*, and are the evidence of these latter being engaged in the **adjustment** and **settlement** of their *original* conflict, [evidenced by the **opposition** of the alternatives.] The **associations** which each *conflicting neuro-cerebral* process* calls up are evidence* of its spreading* to other parts of the brain, and being either reinforced or weakened by the *neuro-cerebral* processes* which it sets up in those other parts. In this way the **action** of

> the whole, or of a comparatively *large* part, of the brain is brought to bear upon the comparatively *small* part implicated in the *original neuro-cerebral* processes* sustaining the *alternative* contents* of **consciousness** (phil40)

And, as has already been demonstrated earlier with respect to other dimensions, the presence of positive features does not always entail the total absence of negative ones. Example (6.58) below corresponds to the same excerpt from Henry Bolingbroke's essay (1754), reproduced in example (6.56), but this time with the negative features highlighted:

> (6.58) AMONG the many cavils, that **have** been devised against the demonstrated existence of a first, intelligent, self-existent cause of all things, this has been one, That things known **must** be anterior to knowledge; and that we **may** as well assert that the images of objects we **see** reflected made those objects, as that knowledge, or intelligence made *them*. HOBBES **is** accused of reasoning on this principle in *his* Leviathan, and **his** book De civ, by the author of the Intellectual system of the universe; and *his* argument, in the place where *he* **mentions** the notions that reason **dictates** to us concerning the divine attributes, **is** thus stated. Now I **think** this charge a little too hastily brought, and a little too heavily laid. So will any man who **reads** the context. (phil11)

Finally, Life Sciences presents little complexity both in the eighteenth and nineteenth centuries, as shown in examples (6.59) and (6.60), respectively:

> (6.59) Pliny **reports** that the young Ones **are** carried off in the following manner in India, viz. The Hunters **lie** in wait to espy when the Tigress is abroad, that *they* may have an opportunity to carry off the whole Litter of Whelps at once, upon very swift Horses prepared for that End. But when the Tigress **returns** and **finds** her young ones gone, *she* **pursues** most swiftly those that carried *them* away, by the Scent. But as soon as *they* **perceive** the Tigress approaching near them, *they* **let** fall one of the Cubs, which *she* **takes** in her Mouth, and **runs** back to her Den with *it*, and

> immediately **pursues** again in quest of the rest of *her* Whelps, thus *she* **runs** to and from *her* Den, until such time as the Hunters **have** an Opportunity to embark and get off with part of the young Ones (life8)

> (6.60) Scarce a winter **passes** but innumerable thousands of *them* **are** seen in the lower parts of the whole Atlantic states, from New Hampshire to Carolina, particularly in the neighbourhood of our towns; and from the circumstance of *their* leaving, during that season, the country to the north-west of the great range of the Alleghany, from Maryland northward, it would appear, that *they* not only **migrate** from north to south, but from west to east, to avoid the deep snows that generally **prevail** on these high regions for at least four months in the year. The Robin **builds** *his* nest, often on an apple tree, **plasters** *it* in the inside with mud, and **lays** five eggs of a beautiful sea green (life22)

If we bear in mind that, with respect to Dimension 2, Life Sciences has been characterised as a predominantly descriptive discipline, the large number of present tense verbs and place adverbials seems to be logical in that the former refer to immediate or habitual states and actions, characteristic of certain species (in this case, animal), whereas the former are used to designate places where they dwell or directions in which they move. Likewise, third person pronouns appear to be necessary deictic elements used to refer to those species, which, in turn, appear as direct agents of the actions designated by the verbs. All in all, it may also be proved that, just like it was relatively easy to find negative features in eighteenth-century Philosophy, it is also possible to find positive features in nineteenth-century Life Sciences, as may be observed in example (6.61):

> (6.61) The works of Linneus being now translated, botany has a language peculiar to itself; that language is, perhaps, somewhat less difficult to learn than any other language; and should tenfold the difficulty be found in the **acquirement** of it, the time might be esteemed well spent. The term **fructification** is defined by Linnaeus to be a *temporary* part of vegetables dedicated to **germination**; that is, all the parts of the blossom, which are intended for the **production** and **preservation** of the seed, and which,

> having brought that to **perfection**, wither and fall off. All these parts, however, are not essential to the **production** of *perfect* seed, as will be seen hereafter; nor are all these parts present in every flower (life21)

Nevertheless, the following example (6.62) illustrates that negative features are also present, to some extent, in the excerpt that we have just analysed, even though in example (6.61) they appear overshadowed by positive ones:

> (6.61) The works of Linnaeus being now translated, botany **has** a language peculiar to *itself*; that language **is**, perhaps, somewhat less difficult to learn than any other language; and should tenfold the difficulty be found in the acquirement of *it*, the time might be esteemed well spent. The term fructification **is** defined by Linnaeus to be a temporary part of vegetables dedicated to germination; that is, all the parts of the blossom, which **are** intended for the production and preservation of the seed, and which, having brought that to perfection, **wither** and **fall off**. All these parts, however, **are** not essential to the production of perfect seed, as will be seen hereafter; nor **are** all these parts present in every flower (life21)

For the analysis now of our corpus with respect to Dimension 3 by genre, Figure 6.6 offers the following picture:

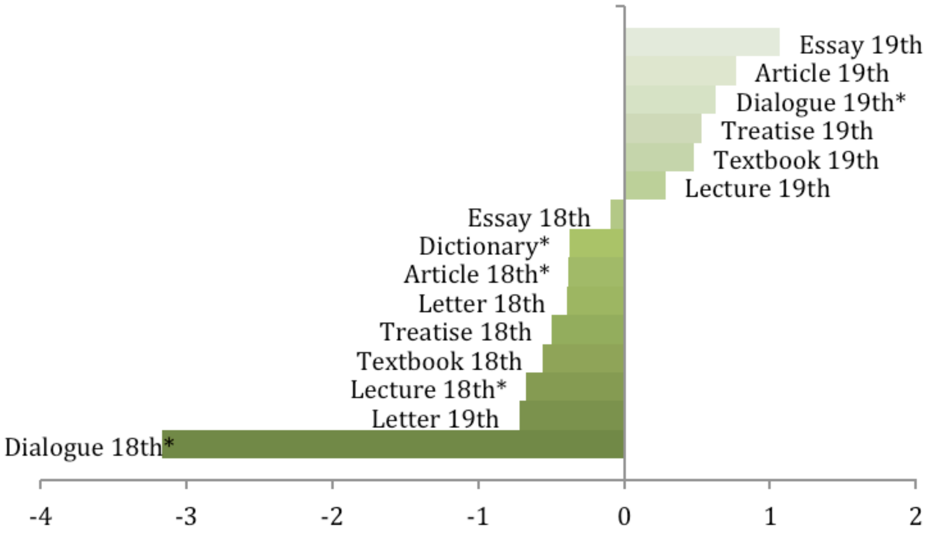


Figure 6.6
Relations among eight genres across two centuries along Dimension 3 “Elaborate vs. non-elaborate discourse” (descending order)

On the above picture we can see that, with the exception of letters, all the subregisters based on genre classification also shift in the direction of discourse complexity and elaboration with time. To be more exact, all the eighteenth-century genres appear to be relatively non-elaborate, while all the nineteenth-century genres, except letters, become elaborate. This seems to coincide with Atkinson's (1999: 126-129) demonstration that non-epistolary research articles became more elaborate and, at the same time, more impersonal along the eighteenth and nineteenth centuries, whereas epistolary articles, conversely, became less abstract across time.[45] Likewise, this steady increase in nominalisations in all three disciplines appears to confirm the statement made in Bello (2014: 322) on nominalisations gradually consolidating as a marker of the scientific register in English.

The subregister presenting the largest proportion of elaborate features in the nineteenth century is that of Essays, as can be seen in the following fragment (example 6.63) from Arthur James Balfour's *A Defence of Philosophic Doubt* (1879):

> (6.63) It is perfectly accurate to talk of a *permanent* **possibility** of **sensation** in the same sense; as equivalent, that is, to a set of *permanen*t causes of **sensation** {by which, when they are properly supplemented by causes which are not permanent but only occasional, a **sensation** will actually be produced.} But though Science may be consistent with a belief* in a world [composed of such **possibilities**,] the teaching* of Idealism* certainly is not. Again, the permanence* [attributed to the **possibilities** of **sensation**] might be a permanence* not of the **conditions** {by which **sensations** are produced} but of the laws which regulate their **production** (phil 36)

Eighteenth-century essays, conversely, appear to have a moderate amount of both negative and positive features, as has been shown through examples (6.57) and (6.58) above, respectively, when analysing eighteenth-century Philosophy. However, while the overall mean dimension score for the Philosophy subregister in the 1700s is positive (0.05), that of the Essay subregister is a negative one (-0.08). Likewise,

[45] See also Biber (1988) on Dimension 5, "Abstract vs. non-abstract style", on which Academic Prose has a large "abstract" score.

eighteenth-century treatises also appear to be non-elaborate (see example (6.64)), although some of them, such as the one exemplified in excerpts (6.65) and (6.66), may contain both non-elaborate and elaborate features. Among the latter, detached past participle clauses, although not included in Factor 3, are marked as passive constructions with a double forward slash // because, all in all, this feature fulfills a similar function to that of agentless passives and past participial WHIZ-deletions:

> (6.64) LOW CLAVARIA. TAB. CXII. FIG. II. THIS **arises** singly or in clusters, from a very small root, which **is** furnished with numerous downy fibres. The branches **are** small near the base, increasing in thickness upwards, and, in *their* ascent, **are** divided and subdivided into numerous branches, all of which **are** lopped off at top, with a broad termination, which **is** often decorated with small rising points round the margin. Sometimes the margin **is** dentated or crenated, it **is** most commonly of a yellow or golden colour, but sometimes **varies** to white or purple. **Grows** in barren pastures, about Halifax (life18)
>
> (6.65) Negative features:
> *Its* Eyes, which **are** Two, **are** small, shining, and hemispherical, situated near the upper Part of the Head, for the Convenience of seeing before, as well as behind. Below these stand Two Horns, or Feelers, of about an Inch long. The Back **is** black and shining, joined by a strong Ligament to the Abdomen, which **is** made up of Six Annuli, or Sections. The Rapidity of *its* Flight **depends** upon Four glossy Wings (life10)
>
> (6.66) Positive features:
> Its Eyes, which are Two, are small, shining, and hemispherical, //situated near the *upper* Part of the Head,// for the Convenience* of seeing before, as well as behind. Below these stand Two Horns, or Feelers, of about an Inch long. The Back is black and shining, //joined by a strong **Ligament** to the Abdomen,// which is made up of Six Annuli, or **Sections**. The **Rapidity** of its Flight depends upon Four *glossy* Wings (life10)

Nineteenth-century Treatise, conversely, is a register characterised by a high level of structural elaboration, with long noun phrases, postmodified by past participial WHIZ-deletions (some of which contain other WHIZ-deletions embedded; see example (6.67)), or by a sequence of embedded prepositional phrases (as in example (6.68)):

> (6.67) The mundus intelligibilis or thought-world, on the other hand, which we build out of the phenomena of *objective* thought, taking it apart from these systems of symbolism*, logical and mathematical, and containing the *unfilled* blanks which we have spoken of, is thus a world of *provisional* images, **conceptions**, and hypotheses, [framed on the lines of sense-**presentations** and their forms of time and space, and awaiting, in some cases the **verification** [afforded by sense-**presentations**,] in others the *concrete* filling up* of its *abstract* skeletons by either *presented* or *represented* details, in others the fiat of *decisive* **volitions** (phil40)

> (6.68) In the *present* day we might doubt the **correctness** of the division* of the **faculties** of the soul and their **functions**, and we should certainly doubt whether on the strength of the analogy it could be maintained that the **separation** of classes should be a *fixed* fate even for the individual 's life-time (phil39)

Similarly, nineteenth-century Article, Textbook and Lecture all have positive scores, showing a densely nominalised and, in most cases, passivised discourse (see examples (6.69), (6.70) and (6.71), respectively):

> (6.69) The only *remaining* work for *future* astronomers, is to determine with the extreme of accuracy* the consequences* of its rules, by the *profoundest* **combinations** of mathematics; and the magnitude* of its data by the *minutest* **scrupulousness** of **observation**. And in this *last* respect, but little may be hoped for, unless instruments* can be constructed and adjusted with a **nicety** which seems almost incompatible with the **productions** of the most *consummate* skill. All the phenomena of this science depend upon a *single* law, which may be deduced from the

> *simplest* among them and by the *rudest* **observation**; and which has been put repeatedly to the *severest* trial, by a series of discoveries* unparalleled in number and delicacy*: such as the precession* of the equinoxes; the **nutation** of the earth 's axis; the **aberration** of light; the **oscillations** both of the ocean and the atmosphere and those **variations** in the elements of the *planetary* motions* and orbits, [termed secular,] requiring in some cases the lapse of ages for their **development** (astr29)

> (6.70) A body [subjected to the **action** of a *central* force, whose **intensity** varies as the square of the distance* inversely,] must describe one or other of the *conic* **sections**, depending upon the **relation** between its **velocity** and the **intensity** of the *central* force. The orbits that are known to belong to the *solar* system are ellipses. 21. Those primaries which move in *elliptical* orbits of *small* **eccentricities** are called PLANETS. Those primaries having orbits of *great* **eccentricities** are called COMETS. Comets are also distinguished from planets in having a degree of **density** so low as to give some the appearance* more of a vapour than of a *solid* body (astr32)

> (6.71) The essence of *Scottish* philosophy, as it appears in Reid, is accordingly a **vindication** of **perception**, as **perception**, in **contradistinction** to the *vague sensational* idealism*, which had ended in the **disintegration** of knowledge*. **Sensation** is the **condition** of **Perception**; but so far from the two terms being interchangeable, **sensation**, as **sensation**, does not enter into **perception** at all. It is significant that the two points {on which Reid takes his stand} should be (1) the **reassertion** of the *essential* difference* between the *primary* and the *secondary* **qualities**, or, in other words, the **proclamation** of the *impassable* gulf between extension*, as a percept, and any feeling* or series of feelings*; and (2) the **assertion** that the unit of knowledge* is an act of **judgment** (phil37)

Interestingly, it may be observed that these examples, and, especially, example (6.71), illustrate that this linguistic complexity also seems to reflect a particular elegance and literariness, characteristic of the scientific prose of the nineteenth century, suggesting that the latter, apparently, could be *both* elaborate *and* abstract, the two terms not necessarily standing in opposition as was suggested earlier in Chapter 2. Conversely, the eighteenth-century dialogue (example 6.72 below), belonging to the Astronomy subcorpus, has a very high negative score (-3.16). This may be due to its relatively informal, *ad hoc* style, typical of a conversation, characterised by a high amount of second and third person pronouns (while first person pronouns, though also abundant, are not counted in Dimension 3), as well as by present tense verbs. The latter may be either referring to immediate states of mind (e.g. *I hope*) or stating speech acts (e.g. *I thank you*), or used for the explanation of the matter that is being discussed:

> (6.72) I **THANK** *you*, Sir, said *she*, for that Information; I **shall**, I **hope**, be able to understand a little of Books of this Kind, by Degrees; But, pray, **have** *you* any thing more to show me, relating to these Circles?
>
> MADAM, said I, it'll be proper for *you* to know, that as our Astronomers **make** six greater, so *they* **make** also four lesser Circles of the Sphere; two of which *they* **call** the Tropicks, and the other two the Polar Circles. The Meaning of the Word Tropicks **is**, **returns** back again; for indeed neither the Sun seemingly, nor the Earth really, **goes** any further in its Annual Course, to the North or Southward of the Equinoctial than 23 Degrees… ; but after *it* **hath** gone so far, **returns** again toward *it*… (astr4)

On the other hand, letters, which also have negative scores in both centuries, show here an evolution which is the opposite to that followed by the rest of the genres. Examples (6.73) and (6.74) below are excerpts from an eighteenth-century letter on Astronomy and a nineteenth-century letter on Life Sciences, respectively. Although Life Sciences becomes somewhat more elaborate with time, letters go in the contrary direction, furthering themselves away from the gradually consolidating scientific register, and eventually disappearing from the scientific literature:

> (6.73) The planet next above the Earth's orbit, **is** Mars. *His* distance from the Sun **is** computed to be about 144 millions of miles; and by travelling at the rate of fifty-five thousand miles an hour, *he* **goes** round the Sun in a

> little less than two of our years. *His* diameter **is** 4200 miles; and *his* diurnal rotation upon *his* axis **is** performed in about twenty-four hours and thirty-nine minutes. This planet sometimes **appears** gibbous, but never horned; which plainly **shows**, that *his* orbit **includes** that of the earth, and that *he* **shines** not by *his* own native light (astr19)

> (6.74) When the prey **is** caught in this way, the tentacles **close** upon *it* and **pass** *it* into the mouth; but in order that *you* **may** understand this, I **must** tell *you* something about the mouth, and about the inside of our little Sea-Anemone. If we **look** down upon *him* from above, we **shall** see in the centre of the fringes a hole, and that hole **is** the mouth which **opens** into a kind of sac that **hangs** down below *it*, inside the animal, and **is** *its* stomach, into which all the food **passes** and where it **is** digested. If we could make a cut across our little friend, so as to get a glimpse of *his* internal arrangement, we should see this sac which **makes** a cavity in the middle of the body… (life31)

Finally, the next section of this chapter analyses the relations among the different registers with respect to the last dimension in our four-dimensional model, which deals with the presence or absence of narrativity in the texts.

3.4. *Relations along Dimension 4 “Narrative vs. non-narrative discourse”*

The mean scores of the three scientific disciplines across two centuries for Dimension 4 are plotted on Figure 6.7 on the right. Here once more, registers with positive values are on the right side of the x-axis, showing that they contain a relatively large amount of narrative features, while those with negative values are on the left side, indicating a low to very low proportion of them.

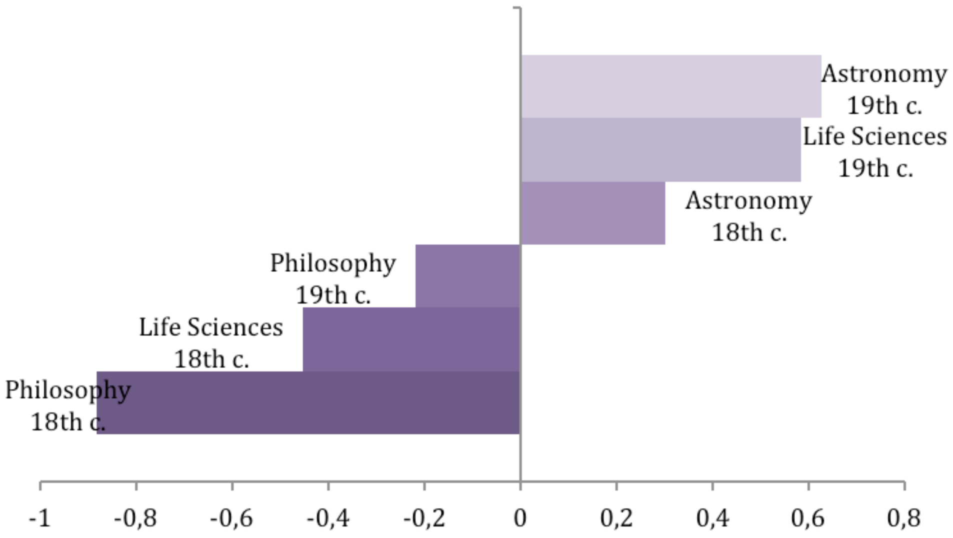


Figure 6.7
Relations among three scientific disciplines across two centuries along Dimension 4 "Narrative vs. non-narrative discourse" (descending order)

Figure 6.7 shows that all three disciplines become more narrative with time. Philosophy is overall non-narrative, although it presents a relatively lower negative score in the nineteenth century, whereas Astronomy is narrative in both centuries, with a higher positive score in the 1800s. Life Sciences, in turn, appears to be non-narrative in the first period (see example (6.75)) and to become narrative in the second (as shown through example (6.76)). In the first example, and in all the examples marking negative features unless indicated otherwise, present-tense verbs are in **bold**, *be* as main verb is underlined, and instances of phrasal coordination are in *italics*. In the second example, as well as in all the subsequent examples highlighting positive features, past tense and perfect aspect verbs are in **bold**, while time adverbials are in *italics* and other time references are underlined:

(6.75) These branches **terminate** in fingered divisions, which **are** of a pale ochre colour. At the height where these branches **terminate**, the stem **emits** a single compressed horn, an inch long, the upper part ochre coloured. About three inches above this, the stem **begins** to spread out, and **forms** itself into a large open expansion, rudely resembling a patha of the Calla Indica, but **is** broader at both extremities; the margins **are** rolled back, and gently *waved and laciniated*; on the upper side, near the middle of the expansion, **is** a projecting line, which **seems** to consist of a wrinkled membrane, *hard and dry*; on the under side, all round, the margin **is** of a pale ochre colour, and full of angular pores, which **have** no proper tubes; all the rest of the under side **is** beset with tubes, placed in an oblique direction, their mouths *torn and laciniated*, but their lower parts cylindrical (life18)

(6.76) In rambling through the woods one day, I **happened** to shoot one of these birds, and **wounded** him slightly in the wing. Finding him in full feather, and seemingly but little hurt, I **took** him home, and **put** him into a large cage, made of willows, intending to keep him in my own room, that we **might** become better acquainted. As soon as he found himself enclosed on all sides, he **lost** no time in idle fluttering, but throwing himself against the bars of the cage, **began** *instantly* to demolish the willows, battering them with great vehemence, and uttering a loud piteous kind of cackling, similar to that of a hen when she is alarmed, and takes to wing. Poor baron Trenck never **laboured** with more eager diligence at the walls of his prison than this son of the forest in his exertions for liberty; and he **exercised** his powerful bill with such force, digging into the sticks, seizing and shaking them so from side to side, that he *soon* **opened** for himself a passage; and though I repeatedly **repaired** the breach, and **barricaded** every opening in the best manner I **could**, yet on my return into the room I always **found** him at large, climbing up the chairs, or running about the floor, where from the dexterity of his motions, moving backwards, forwards, and sideways with the same facility, it **became** difficult to get hold of him *again* (life22)

Although both extracts may be characterised as descriptive, the first one focuses on giving the visual details of the branches of a tree and contains therefore, mainly, verbs in the present tense and copular *be*, often followed by adjectives, whereas the second one gives a full account of an experiment carried out by the author – which consisted in catching a bird and putting it in a cage, and observing then how it behaved under such conditions – and thus presents a large proportion of past tense verbs and time adverbials. Likewise, the following fragment (example (6.77)) from a nineteenth-century Astronomy letter by Denison Olmsted (1841) is also full of time references (which are here underlined, including multi-word expressions), telling a story about several astronomers who had devoted their lives to the discovery of comets in the past:

> (6.77) On the night of Christmas-day, 1758, George Palitzch, of Politz, near Dresden, “a peasant,” says Sir John Herschel, “by station, an astronomer by nature,” first **saw** the comet. An astronomer of Leipzic **found** it *soon* after; but, with the mean jealousy of a miser, he **concealed** his treasure, while his contemporaries throughout Europe **were** vainly directing their anxious search after it to other quarters of the heavens. At this time, Delisle, a French astronomer, and his assistant, Messier, who, from his unweared assiduity in the pursuit of comets, **was** called the Comet-Hunter, **had been** constantly engaged, for eighteen months, in watching for the return of Halley's comet. Messier **passed** his life in search of comets. It is related of him, that when he was in expectation of discovering a comet, his wife **was** taken ill and **died**. While attending on her, being withdrawn from his observatory, another astronomer **anticipated** him in the discovery. Messier **was** in despair. A friend, visiting him, **began** to offer some consolation for the recent affliction he **had suffered**. Messier, thinking only of the comet, **exclaimed**, “I had discovered twelve: alas, that I should be robbed of the thirteenth by Montagne!” and his eyes **filled** with tears. Then, remembering that it **was** necessary to mourn for his wife, whose remains **were** still in the house, he **exclaimed**, “Ah! this poor woman!” (…), and *again* **wept** for his comet (astr30)

By contrast, the following excerpt (6.78) from an eighteenth-century essay by Mary Astell (1700), which belongs to the Philosophy subcorpus, deals with a general – if controversial – matter in a detached present tense, describing with exquisite irony the duties of what at that time was considered a good wife:

> (6.78) It **is** a Woman's Happiness to hear, admire and praise them, especially if a little Ill-nature **keeps** them at any time from bestowing due applauses on each other. And if she **aspires** no further, she **is** thought to be in her proper Sphere of Action, she **is** *as wise and as good* as can be expected from her. She then who **Marries ought** to lay it down for an indisputable Maxim, that her Husband **must** govern *absolutely and entirely*, and that she **has** nothing else to do but to *Please and Obey*. She **must** not attempt to divide his Authority, or so much as dispute it, to struggle with her Yoke will only make it gall the more, but **must** believe him *Wise and Good* and in all respects the best, at least he **must** be so to her. She who **can't** do this **is** in no way fit to be a Wife… (phil1)

Likewise, the following example (6.79) from a nineteenth-century Philosophy lecture by George Combe (1846) shows that some topics, such as morality, are universal and are not therefore restricted by time:

> (6.79) Our first inquiry **is** into the basis of morals regarded as a science; that **is**, into the natural foundations of moral obligation. There **are** two questions very similar in terms, but widely different in substance which we **must** carefully distinguish. The one **is**, What actions **are** virtuous? and the other, What **constitutes** them virtuous? The answer to the first question, fortunately, **is** not difficult (phil30)

However, the excerpt below (6.80), taken from the same text, shows that this lecture also contains narrative passages, where the author brings episodes he is familiar with to illustrate his ideas and opinions:

> (6.80) To illustrate the possibility of discriminating natural dispositions and talents by means of observations on the head, I may be permitted to allude to the following cases.
>
> On the 28th October 1835, I **visited** the jail at Newcastle, along with Dr. George Fife (who is not a phrenologist) and nine other gentlemen, and the procedure adopted **was** this: I **examined** the head of an individual criminal, and, before any account of him whatever **was** given, **wrote** down my own remarks… (phil30)

As we have seen, thus, Philosophy appears to have a relatively non-narrative discourse in that it deals with matters of a general, universal nature, even though sometimes – especially in the nineteenth century – authors resort to narration of particular happenings from which general conclusions can be drawn. Astronomy and Life Sciences, by contrast, evolve towards a markedly narrative standard in the 1800s, which suggests a lasting importance of experimental reports in these two disciplines at a time when, as we shall see in what follows, the experimental article (or experimental essay, as it was commonly called at that time (Gotti 2001), loses its narrative component.

Turning now to Dimension 4 from the point of view of genres, the picture will be somewhat different (see Figure 6.8):

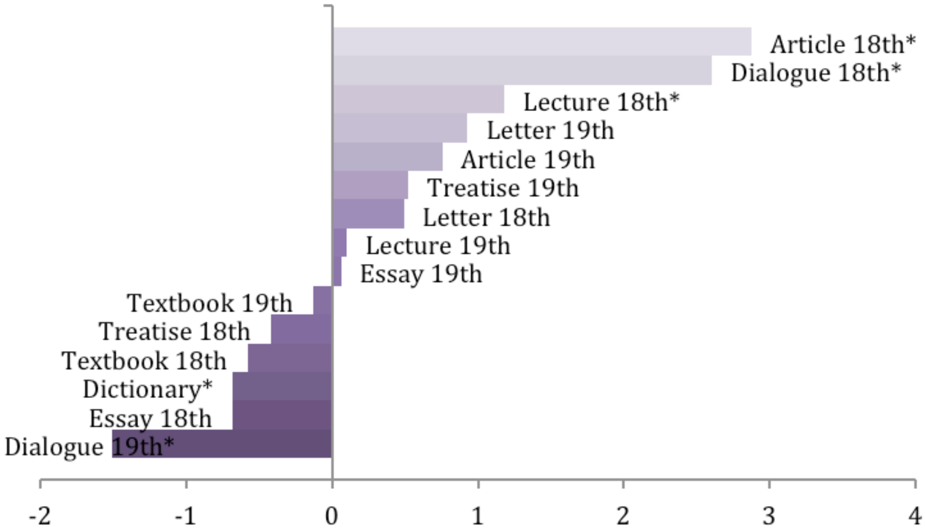


Figure 6.8
Relations among eight genres across two centuries along Dimension 4 “Narrative vs. non-narrative discourse” (descending order)

Indeed, Figure 6.8 shows a picture that contradicts the classification by discipline. Unlike the distribution of genres with respect to Dimension 3, here not all genres become more narrative with time. Rather, while Essay, Textbook, Treatise and Letter appear to gain narrative features, Article, Dialogue and Lecture seem to move in the contrary direction. This, however, could be due to the fact that both eighteenth- and nineteenth-century Dialogue, as well as the eighteenth-century Article and Lecture are constituted by only one sample and may not, therefore, be necessarily representative of the genre at the time. Thus, as already noted earlier, these cases will be treated with caution and will not be used as a basis for generalisations. However, the fact that these three genres lose narrative features over time appears to coincide with Atkinson's (1999: 144) findings in his analysis of the *Transactions* with respect to Biber's (1988) Dimension 2 "Narrative vs. non-narrative concerns", which show that scientific research articles become less narrative along the eighteenth and nineteenth century.

All in all, it might be interesting to look at examples from the texts in order to prove whether these examples justify the dimension scores. For instance, the eighteenth-century Article, relating an astronomical discovery, presents the largest positive score (2.9) and is illustrated in example (6.81) below:

> (6.81) At last, on December 11th, I *again* **discovered** it, on the opposite side of the disc, it **having** by that time **advanced** a little way from the eastern limb, being distant from it 1'30". And *now* I **could** only perceive three sides of the umbra, namely, the upper and under sides, and that towards the limb, which **was** the side that *formerly* **had vanished**. The side towards the center of the disc **was** not as yet visible; but I **concluded**, upon the same grounds as *formerly*, that it **was** hid from my sight, by its averted position only, and that, after the spot **had advanced** a little further, it **would** make its appearance. Accordingly, the next day, being December 12th at ten o'clock, it **came** into view, and I **saw** it distinctly, though narrower than the other sides. After this, my observations **were** interrupted, by unfavourable weather, till the 17th, when the spot **had passed** the center of the disc, the umbra *now* appearing to surround the nucleus equally (astr15)

This is an example of an experimental report, typical of eighteenth-century scientific prose (Bazerman 1988; Atkinson 1999). The details of the observation are given day by day, the narration containing quite a few time references (i.e. time adverbials such as *again*, *now*, or *formerly*, which indicate that the narrator had already made that discovery before and contrasts the past (in the past) with the present (also in the past); or multi-word expressions denoting dates (e.g. *on December 11th* or *till the 17th*), times (e.g. *at ten o'clock*), or other text-internal deictics, such as *by that time*, or [*a*]*fter this*, referring to different moments in the narration. This kind of time references is also present in the eighteenth-century Dialogue, where a short narration of the encounter of the author with his interlocutor, at the beginning and at the end of example (6.82), is followed by direct reported speech, marked by verbs of saying in the past tense:

> (6.82) IT is *now* about seven Years ago, since I **presented** the most Engaging Lady M..... with Mr. Fontenelle's Book of the Plurality of Worlds: And I remember well what she **said** a few Days after.
>
> I **have looked** over your Book, Sir, **said** she, as my way is, first cursorily, and I intend to give it a very careful second Reading; but I perceive by it, you **have cut** out much more Trouble for your self, than perhaps you **imagined**: For I find there are many things *previously* necessary to the understanding it, which you must oblige me with explaining; but, **continued** she, a Conversation of that kind with me, I doubt, will be too dull and tedious since I am not blessed with any of those shining Qualifications, with which Mr. Fontenelle **hath complimented** M. la Marquiee…; I should indeed, **said** she, except those two, which I suppose, in Complaisance to our Sex, he makes the Foundation of Philosophy, viz. Ignorance and Inquisitiveness for those I'm sure, I have in Perfection, as you **have** long **experienced**.
>
> I need not mention the Return I **made**, nor how prettily she **changed** the Discourse to something more general, when she **found** I **was** going to say just things of her… (astr4)

Conversely, no such narration is present in the nineteenth-century Dialogue (example (6.83)) from the Philosophy subcorpus, where the interventions of the two interlocutors are directly reported and, therefore, verbs of saying in the past tense are absent:

> (6.83) (COMMON SENSE). We **have** the same law of varied force without it; and no occasion for a vacuum, because the planets **swim** in the medium of space like ships in a current. Besides, void space, while gas **is** elastic, **is** a solecism, and the phenomena of meteors, comets, and planets, **prove** that the supporters of *light and flame* **are** to be found every where.
>
> OXONIAN. I **confess** I always **have** had difficulties in conceiving of void space while such a body as an elastic atmosphere existed in it; and whether gas **exist** universally and independently, which **seems** to be your notion, or whether the atmospheres of planets **expand**, it **seems** reasonable to conclude that space **must** be occupied; and if so, then the resistance would cause gravitation to overcome the projectile force, and the planets would be carried to the sun in spiral lines, which we **know** **is** not the fact. But how **does** your gaseous lever operate?
>
> COMMON SENSE. By *action and re-action*, proportioned to the quantities of *matter and momenta* of the bodies; and with reference to particular bodies, inversely as the square of the distances (phil25)

Indeed, the fact that the nineteenth-century Dialogue has very few narrative features does not constitute a valuable proof that all scientific dialogues were not narrative at that time. Rather, as we have just seen in examples (6.82) and (6.83), this difference between the two dialogues can be appreciated through the way the subject is tackled in each text: while the eighteenth-century dialogue on astronomical matters contains accounts of diverse experiments and the indirect speech of the dialogue itself is conveyed in the past tense, the nineteenth-century philosophical debate is expressed through directly reported speech which, in turn, mostly contains axioms expressed through present tense.

In the eighteenth-century Lecture, in turn, present perfect is used to refer to the actions that have been happening for a considerable period of time and continue to happen at the moment, or did not change until very recently:

> (6.84) Because, therefore, from the beginning of the new Astronomy unto the present Age, the Fixed Stars **have been** reckoned to have no annual Parallax; (for as for the Diurnal, no one in his right Wits *ever* **dreamed** that they **were** subject to that) it is no wonder, if the most Sagacious Astronomers **have determined**, that both their Distances and Magnitudes **were** altogether unknown. Nor indeed **have** the most excellent Observers, who **set** themselves to it in good earnest, **been** able, till very *lately*, by observing the Fixed Stars most accurately at divers Times of the Year, to obtain even the least apparent Difference of Place (astr3)

On the other hand, as we had observed before with the excerpt from Astronomy (see example (6.77)), nineteenth-century letters often contain narrations of past discoveries. The following excerpt (6.85), now from the Life Sciences subcorpus, is mostly narrated in the past tense, while past perfect is used to refer to a time that was previous to the time of the discovery of a flower species:

> (6.85) *Sometimes* in breaking up or blasting rocks, there **have been** found upon them impressions that **looked** as if some large but graceful flowers, not unlike a widely opened tulip or lily, only of great size, **had been** roughly drawn there. At first, the persons who **found** these strange old flowers, as they **seemed**, buried in the rocks, **could** not understand how they **came** to be there, or what they **were**, but from their appearance they **were** called "stone lilies". But when they **were** more closely examined, and carefully studied by naturalists, who **were** familiar with animal structures, it **was** found that what **looked** like a flower-cup **was** a kind of Star-Fish, growing upon a tall stalk, which must **have been** attached to the ground when the creature **was** alive. And so they **were** no longer considered as flowers of old times that **had been** hidden away in the rocks, and they **lost** their pretty name of "stone lilies", and are *now* called Crinoids, the first animals of this kind that *ever* **lived** (life31)

Likewise, nineteenth-century Treatise also appears to be a fairly narrative genre, as may be observed through example (6.86):

> (6.86) Though the imperfections of the Ptolemaic system **were** not *immediately* perceived, especially during the confusion which **attended** the decline and destruction of the Roman empire, their effects **did** not fail, in process of time, to become fully evident. In the ninth century, on the revival of science in the east, under the encouragement of the caliphs, surnamed Abassides, Ptolemy's astronomical tables **were** found to deviate so widely from the actual situations of the celestial bodies, as to be no longer useful in calculations: and it **became** necessary for the Saracen astronomers at Bagdat to form tables entirely new. The Saracen's **carried** their astronomical knowledge with them into Spain; and, in the thirteenth century *again*, the new tables **were** found unfit to represent the celestial motions; and, to supply their place, the tables, called Alphonsine, **were** constructed, by the direction of Alphonso the 10th, king of Castile. The errors even of the Alphonsine tables **became**, in the fifteenth century, equally sensible with the former (astr22)

By contrast, eighteenth-century Treatise is predominantly non-narrative (see example (6.87)), focusing, just like eighteenth-century Textbook (example (6.88)) and Essay (example (6.89)), on descriptions, either of the parts of a plant, or of a terrestrial globe, or of abstract concepts such as vice and virtue:

> (6.87) For an explanation of these see the glossary; or look at a rose, and the green covering that *encloses and supports* the blossom **is** called the CUP. Pl. 3. fig . I. The Cup of a Polyanthus **is** represented in pl. 3 fig. 10. Linnaeus **says** the Empalement **is** formed by the outer bark of the plant. The BLOSSOM **is** that beautifully coloured part of a flower, which **commands** the attention of everybody. If it **is** *entire and undivided*, as in the Polyanthus, or Auricle; it **is** said to be a blossom of one Petal; but if it **is** composed of several parts, it **is** accordingly said to be a blossom of one, two, three, &c. or many parts or Petals. Thus the Blossom of the Tulip **is** formed of six Petals; and the Garden Roses **bear** Blossoms composed of many Petals. The Blossom **is** supposed to be an expansion of the inner bark of the plant. The CHIVES **are** slender thread-like substances, generally placed within the Blossom, and surrounding the Pointals (life16)

(6.88) Each globe **hath** a brass wire circle, TWY, placed at the limits of the crepusculum, or twilight, which, together with the globe, **is** set in a wooden frame: the upper part BC **is** covered with a broad paper circle, whose plane **divides** the globe into two hemispheres, and the whole **is** supported by a neat *pillar and claw*, with a magnetic needle in a compass box at M. On our new terrestrial globe, the division of the face of the earth into *land and water*, **is** accurately laid down from the *latest and best* astronomical, geographical, and nautical discoveries. There **are** also many additional circles, as well as the rhomb-lines, for the greater *ease and convenience* in solving all the necessary *geographical and nautical* problems. On the surface of our new celestial globe, all the southern constellations, lately observed at the Cape of Good-Hope by M. de la Caille, and all the stars in Mr. Flamted's British catalogue, **are** accurately laid down, and marked with *Greek and Roman* letters of reference, in imitation of Bayer (astr16)

(6.89) Virtue, of all Objects, **is** the most *valuable and lovely*; and accordingly this Species of Philosophers **paint** her in the most amiable Colours, borrowing all Helps from *Poetry and Eloquence*, and treating their Subject in an *easy and obvious* Manner, such as **is** best fitted to please the Imagination, and engage the Affections. They **select** the most striking *Observations and Instances* from common Life; **place** opposite Characters in a proper Contrast; and alluring us into the Paths of Virtue, by the Views of Glory and of Happiness, **direct** our Steps into these Paths, by the soundest Precepts and most illustrious Examples. They **make** us feel the Difference betwixt *Vice and Virtue*; they **excite** and **regulate** our Sentiments; and so they **can** but bend our Hearts to the Love of Probity and true Honour, they **think**, that they **have** fully attained the End of all their Labours.

THE other Species of Philosophers **treat** Man rather as a reasonable than an active Being, and **endeavour** to form his Understanding more than cultivate his Manners (phil10)

Finally, nineteenth-century Textbook, which, much like nineteenth-century Essay and Lecture on the positive side, has a negative score close to zero (-0.13), appears to contain both non-narrative and narrative features, and yet, being overall unmarked for Dimension 4, contains the two of them in a rather low proportion; see examples (6.90) and (6.91), respectively:

> (6.90) Negative features:
>
> From some minute changes in the situations of some of the fixed stars, called the Proper motions of those stars, Dr. Herschell has inferred that the centre of gravity, and consequently the whole system, of the *sun and planets*, <u>**is**</u> in motion towards the constellation Hercules. But the investigations of *Dusejour and Burckhardt* **have** shown that the observations hitherto made, **are** not sufficient to prove the existence of any such motion.
>
> KEPLER'S LAWS. Kepler's laws, with regard to the motions of the planets, **have** been thus far considered as rigorously true. It **may** now be proper to inform the student that the mutual actions of the heavenly bodies on each other, **cause** slight deviations from those laws, as they **are** stated in the preceding part of the work (astr26)

> (6.91) Positive features:
>
> From some minute changes in the situations of some of the fixed stars, called the Proper motions of those stars, Dr. Herschell **has inferred** that the centre of gravity, and consequently the whole system, of the sun and planets, is in motion towards the constellation Hercules. But the investigations of Dusejour and Burckhardt **have shown** that the observations hitherto made, are not sufficient to prove the existence of any such motion.
>
> KEPLER'S LAWS. Kepler's laws, with regard to the motions of the planets, **have been** <u>thus far</u> considered as rigorously true. It may *now* be proper to inform the student that the mutual actions of the heavenly bodies on each other, cause slight deviations from those laws, as they are stated in the preceding part of the work (astr26)

We have seen, thus, that while all three scientific disciplines tend to become more narrative with time, some genres such as Essay, Textbook, Treatise and Letter also become more narrative, whereas others, such as Article, Dialogue or Lecture, are already at the top of the narrative scale in the eighteenth century and, consequently, move in the non-narrative direction. Although we have agreed to avoid drawing hasty conclusions about genres represented by one single sample, the fact that these three genres become less narrative in the nineteenth century appears to back Atkinson's (1999: 131, 144) findings that scientific discourse becomes less narrative over time. On the other hand, the overall movement of all disciplines and some genres towards, rather than away from, the narrative side actually contradicts Atkinson's (1999) results, in the light of which, our sample is moving *away from* the non-narrative standard that would consolidate in the twentieth century. This contradiction might be attributed to the fact that Atkinson's (1999) corpus only contained research articles, whereas our sample comprises a variety of genres. This seems to indicate that the behaviour of a subregister should not necessarily be interpreted as common to the larger register to which this subregister belongs. In this case, the genre Article in our corpus does appear to behave according to the pattern identified in Atkinson (1999), becoming less narrative over time. However, this is not the case with some other genres, which nonetheless are equally labelled as *scientific writing*, as we have seen earlier in Chapter 3.

One of the reasons for this tendency may be, once more, the dependence of the genres on the scientific disciplines to which they belong. If we go back to Chapter 3, we will see that ten of the twenty-seven nineteenth-century treatises belong to the Life Sciences subcorpus, while eight belong to Astronomy, both of which are highly narrative in the 1800s. The same happens with nineteenth-century Letters, two of the total three dealing with "natural history" (as appears in the titles of both; see Table 3.7 and Appendix III) and one with Astronomy. Another tentative explanation might be that, as we have seen in Chapters 2 and 3, some of the genres in our corpus, including letters and some of the treatises, were likely used for the popularisation of knowledge in the nineteenth century, for which narration may have been thought as more accessible for the masses. Notice that the original purpose of these genres was not didactic *per se* (as would be Lecture or Textbook), but, rather, they were intended for the exchange of scientific knowledge among colleagues, just as was the case with articles and essays (Crespo 2016: 30). Still, an educational aim his can be appreciated if we look at some of the titles of those works. For instance, the *Letters on Astronomy, Addressed to a Lady* (1841) by William Olmsted were intended, as the title indicates, for women, which, at that time, were mostly learners, except for a few who could in fact be considered as colleagues (see Chapter 3 Section 2). Likewise, the two Life Sciences letters – Priscilla Wakefield's *Introduction to the Natural History and Classification of Insects* (1816) and Elizabeth Aggasiz's *First Lesson in Natural History* (1859) – were also intended for instruction, and most likely also to women, which was the usual audience of female scientists. This may also be the case with some treatises, such as Anne Pratt's *Flowers and Their Associations* (1840), as well as of Agnes Mary Clerke's *Popular History on Astronomy During the Nineteenth Century* (1893), both of which appear to be addressed to a wider audience than the restricted academic world (the latter, however, ensuring Clerke's reputation as a competent scientist in the astronomical community (Brück 1991, 2004).

The present analysis of register variation at four different dimensions has shown us that, while certain sociolinguistic trends can be detected in the patterns revealed by our model, sometimes it is difficult to determine whether a particular linguistic preference is owed to a particular subregister category (that is, scientific discipline or genre). As it appears from what we have seen in this chapter, while we can envisage certain coherence and some well-defined tendencies in the scientific disciplines, this is not the case with the genre categories, which seem to have rather fuzzy boundaries, reflecting the lack of a technical standard in the English scientific writing which characterises the late Modern period (see Chapter 2). In the following pages, the findings presented so far will be reviewed and summarised with the attempt to draw some conclusions from the present research and to critically analyse the validity of this study. Likewise, questions for further research will be suggested.

Conclusions

In this study we have looked at register variation and change in a sample of eighteenth- and nineteenth-century English scientific texts from the *Coruña Corpus of English Scientific Writing*. Our sample comprises three different subcorpora, each containing texts of a particular scientific discipline: *CETA* (the *Corpus of English Texts on Astronomy*), *CEPhiT* (the *Corpus of English Philosophy Texts*) and *CELiST* (the *Corpus of English Life Sciences Texts*), the latter being a beta version awaiting publication. According to the UNESCO classification of sciences (1988), one of these scientific disciplines – namely, Philosophy – belongs to the humanities, whereas the other two – Astronomy and Life Sciences – are considered natural sciences. This initial partition was the starting point where we expected to find register variation. Apart from that, as stated in the Introduction, this study had two main goals: 1) to identify variation and change across the three aforementioned scientific disciplines, and 2) to spot variation and change across the eight genres that shape the different texts in the corpus: Treatise, Textbook, Essay, Lecture, Article, Letter, Dialogue and Dictionary.

After agreeing on using Biber and Conrad's (2009) definition for the term *register* – namely, a situational variety that can be described in function of its pervasive linguistic features – and reviewing several approaches to trace register variation and change, we have decided to carry out Biber's (1988) Multidimensional

Analysis in order to identify variation and change on several dimensions of variation (Chapter 1). For this purpose, first of all, we have characterised the register under study – the late Modern scientific register – in its socio-cultural context, providing an insight into Western science along the broad period stretching from the Scientific Revolution to the early twentieth century (Chapter 2). Subsequently, we have described the *Coruña Corpus* and the sample analysed, focusing on its compilation principles (such as time-span, representativeness vs. balance, and size), and have then separately examined each of the three subcorpora included in the sample in terms of genres (Chapter 3). After that, we have proceeded with the methodological part of our Multidimensional Analysis, which contained two main steps: 1) selection, retrieval and counting of linguistic features, and 2) the running of a factor analysis.

The first step, thus, consisted in the selection of fifty-eight lexical and grammatical features on the basis of previous research, and in their retrieval from each text in the corpus, for which we re-used and/or developed a series of query algorithms (Chapter 4). The frequencies of the features retrieved, normalised to 1,000 words, have then undergone a factor analysis (Chapter 5), which, after a series of preliminary tests and trials, including the dropping of some features that presented weak factor loadings, yielded a four-factor solution for a fifty-four variable dataset which explained 41 per cent of the total variation in the corpus. Each factor could be interpreted as an underlying dimension of variation, in that the lexical and grammatical features loading on a factor reflect their frequent co-occurrence in a text, and therefore convey a particular discursive function.

The second part of the factor analysis consisted in calculating factor scores for each text, and, eventually, for each subregister (based on discipline or genre), which could then be placed, according to its score, on each of the four dimensions of variation. Thus, based on the features that presented salient loadings on each factor, four dimensions of variation have been identified and labelled for the texts under survey: Dimension 1 "Involved/persuasive vs. informational style", Dimension 2 "Argumentative vs. descriptive focus", Dimension 3 "Elaborate vs. non-elaborate discourse", and Dimension 4 "Narrative vs. non-narrative discourse" (first part of Chapter 6). Each subregister has been then described with respect to each of these dimensions of variation, analysing a) its place on a dimension with regard to other subregisters, and b) the diachronic change of each subregister on a dimension, whenever this subregister was present both in the eighteenth and nineteenth century samples in our corpus (second part of Chapter 6). Having done all this, the following paragraphs summarise our findings and suggest some tentative conclusions drawn from these.

First of all, regarding the first goal of our study, which was to spot **variation and change in English across scientific disciplines**, we can say that both variation and change have been found. In what concerns the former, we have seen that the major difference between the humanities and the natural sciences can be observed on Dimension 1 "Involved/persuasive vs. informational style", which shows Philosophy as a highly involved/persuasive discipline, while Astronomy and Life Sciences appear to be clearly informational. As for change, it can be observed that, as time goes by, the three disciplines move moderately towards the average, suggesting a moderate tendency towards a standard that is less marked with respect to communicative style. Despite that, we have also observed that the dichotomous relationship between Philosophy and other scientific disciplines with respect to Dimension 1 mirrors the situation in the present-day English scientific register (as shown in Gray (2011)), suggesting that Philosophy will keep involvement as its characteristic discursive feature over time. This confirms that Philosophy has always been and continues to be a discipline of a dialectic nature where a variety of transcendental matters are usually dealt with through a debate. Some of our examples have shown how the author often conveys "personal stance and evaluation" (Gray 2011: 143) and/or quotes other authors, sometimes through a somewhat confrontational tone, in order to put across the contradictory character of certain ideas.

Dimension 2 "Argumentative vs. descriptive focus", in turn, isolates Life Sciences to some extent as a fundamentally descriptive discipline, with Philosophy having a more or less balanced distribution of positive and negative features, and Astronomy ranging between highly to moderately argumentative in the eighteenth and nineteenth centuries, respectively. This dimension highlights a major difference between Life Sciences and Astronomy, despite the fact that both may be described as observational disciplines. As we have seen in Chapter 2, Life Sciences was, for a long period, based on the cataloguing of diverse animal and vegetal species, as well as in the detailed description of the organisation and functioning of their internal organs. In Astronomy, by contrast, the observations of the sky are performed through the lens of mathematics and physics, which makes it possible not only to describe the positions and explain the behaviour of the celestial bodies, but also to predict their movements based on precise calculations. All in all, all three disciplines appear to become more descriptive with time, showing a gradual loss of logic markers, which is particularly apparent in the astronomical and philosophical discourse. This progressive disappearance of the chains of causality suggests that the English scientific discourse in the eighteenth century still followed the Scholastic trends of logical reasoning, which would be gradually abandoned along the 1800s.

With respect to Dimension 3 “Elaborate vs. non-elaborate discourse”, we have spotted a similar movement across the centuries in the direction of elaborate discourse, with Life Sciences shifting from markedly to moderately non-elaborate, and Philosophy on the other end of the scale, starting as very moderately elaborate (close to the average) and becoming highly elaborate with time. Astronomy, in turn, appears as non-elaborate in the eighteenth century and as elaborate in the nineteenth, with a gap of 1.1 (see Appendix II). As we had observed earlier, a key ‘elaborate’ feature of Dimension 3 is the nominalisation, the “grammatical metaphor” (Halliday 1985b, 1988) often used in the English scientific discourse to refer to actions and natural processes (e.g. *illumination*, *fluctuation*), and states (*darkness*), their use of which has already been demonstrated to increase along the eighteenth and nineteenth century in Astronomy (Bello 2014). The moderately nominalised discourse in Philosophy already in the eighteenth century appears to be justified by the subject matters of this discipline, which are often of a highly abstract character (e.g. morality and justice; equality of the sexes; the immortality of the soul; causality, etc.; notice that the subjects themselves are nominalisations). The development of English philosophical discourse is shown to go hand in hand with its increase in abstraction and elaboration, becoming highly elaborate in the 1800s. Life Sciences, by contrast, deals with concrete animate and inanimate objects of the material world, which invites to use a simpler kind of language when describing their characteristics and explaining their behaviour, although we have also observed that the technicality of this field likewise increased relatively in the nineteenth century. On the other hand, considering that other important features of discourse elaboration are the prepositional phrases, past participial WHIZ-deletions, or pied-piping constructions (which sometimes appear embedded, creating long, complex sentences), the high dimension scores of Philosophy also coincide with Gray’s (2011: 117-118) findings, according to which present-day Philosophy is characterised by a dense structural elaboration.

Finally, in what regards Dimension 4 “Narrative vs. non-narrative discourse”, Astronomy stands out as the most narrative discipline, unlike Philosophy, which switches from markedly to moderately non-narrative. Life Sciences, in turn, starts with a low score on this dimension (-0.5; see Appendix II) but becomes more narrative in the nineteenth century. Thus, all three disciplines change in the narrative direction. Nineteenth-century Astronomy and Life Sciences have very similar positive scores, suggesting an approach to a highly narrative standard in the natural sciences, as has been shown through text excerpts in the previous chapter. This appears to reflect that the scientist’s accounts of experiments and observations continued to be important in these two disciplines (and, by extension, in diverse genres; see below), despite the fact that the experimental article moved in the contrary direction along the late Modern period, as was demonstrated in Atkinson (1999), as well as in our study (see below on variation among genres). Philosophy, in turn, stays relatively non-narrative throughout both centuries, which has been justified through text examples by its dealing with subject matters of a general or universal nature, free from time constraints. Still, we have also observed that this discipline likewise becomes more narrative in the nineteenth century, by alternating general statements with particular instances (often from past experiences) where they may be applied.

Our analysis of **variation and change across genres** – which was the second goal of our study – has shown that, despite the ‘bigger picture’ described above, variation and change *within* disciplines is also present (which can also be appreciated if we look at the standard deviations for each dimension and discipline in Appendix II). Starting with variation with respect to Dimension 1, we have seen that some genres, such as Textbook, Dictionary, or Letter, are relatively informational, whereas others such as Dialogue or Essay, appear to be involved and/or persuasive. As we had observed earlier, Dictionary is a genre informational by nature, in that it shapes knowledge with definitions. Likewise, Textbook is used for instruction and the authors using this genre are likely procuring to convey information in the most efficient manner. Dialogue, by contrast, is an interactional genre, where matters are debated or discussed through a conversation between two participants which exchange opinions and experiences from a personal perspective, often resorting to persuasive strategies in order to convince the other party. Essays, in turn, have been characterised as an open genre where the authors express their view of a particular subject, which justifies the inclusion of personal stance. Treatise, in turn, presents mostly moderate scores, containing samples from both sides. This may be accounted for by the fact that treatises abound in the three disciplines and deal with a variety of subjects. In what concerns letters, it was initially expected that they would present involvement features in that they are directly addressed to a correspondent and presumably entail personal interaction. However, their relatively high informational scores may indeed be due both to their dealing with natural sciences and to the didactic character of some of them, as we have recently seen in Chapter 6.

On the other hand, we have also seen that all the genres except Essay and Letter (and except for Dictionary, of which we only have a nineteenth-century sample) become more informational with time, suggesting a general tendency of the scientific discourse towards a more impersonal and informationally dense standard, from which both essays and the epistolary genre would gradually drop. This gradual replacement of involvement features with informational ones also coincides with Atkinson's (1999: 78-80) findings in his diachronic multidimensional analysis of research articles from the *Philosophical Transactions*, which show a shift author- to object-centred discourse. Our findings, in turn, suggest that this characteristic can be extended to other English scientific genres in the late Modern period. As we have seen in Chapter 2, it was during that time that the trustworthiness of the scientist was gradually replaced with the importance of the object of research and the experiment it underwent, which entailed a shift of focus from the scientist to the experiment. In our corpus this growing impersonalisation of the scientific discourse appears to be primarily reflected in the natural sciences, while Philosophy, as we have seen earlier, will maintain a more personal and involved focus through time.

In what regards scientific focus (Dimension 2), it has been shown that some registers which stand on different ends on Dimension 1, such as Dialogue (involved) and Dictionary and Textbook (informational), appear together on Dimension 2 as markedly argumentative. In the case of the latter, their high content of logical connectors appears to be justified by the fact that both genres convey a type of discourse that presents facts and data in a tightly packed manner and needs a high proportion of elements of discourse cohesion. The former, in turn, was chosen in order to convey a debate on an astronomical and a philosophical nature (Harris (1719) and Phillips (1824), respectively) by means of a dialectical battle of a didactic character, carried out through rationalising and logical reasoning. We may agree thus that all three registers have informative purposes. Furthermore, most textbooks deal with astronomical matters, which, as has earlier been observed, are discussed largely through mathematical reasoning. Lecture, on the other hand, albeit also of an instructive character, appears to be a relatively descriptive genre, as are also Letter and Treatise. This phenomenon has been found to be related to the scientific discipline, in that most letters and treatises correspond to Life Sciences, which is a descriptive discipline. Essays, in turn, are predominantly argumentative in the eighteenth century and descriptive in the nineteenth. This may likewise be explained by the fact that the majority of essays, both eighteenth- and nineteenth-century, are found in Philosophy, which, just like the Essay genre, shifts from argumentative to descriptive with time.

In what concerns discourse complexity, the distribution of the genres along Dimension 3 has shown that, overall, all the eighteenth-century genres except letters are non elaborate, becoming elaborate in the nineteenth century, just as is the case with the scientific disciplines. As we have pointed out earlier, this picture once more supports Atkinson's (1999: 126-129) data, which showed that non-epistolary research articles become more elaborate and impersonal along the eighteenth and nineteenth centuries, while epistolary articles go in the contrary direction. Here again, our findings suggest that this characteristic of late Modern English research articles may be extended to other scientific genres in that period. Likewise, the high scores of most of the nineteenth century sample on Dimension 3 appear to back Bello's (2014: 322) claim that nominalisations gradually consolidate as a marker of the English scientific register.

On the other hand, unlike on Dimensions 1 and 2, where essays appear to go "against the flow" (or the apparent general tendency of late Modern English scientific writing), on Dimension 3 Essay not only likewise becomes more elaborate with time, but, in fact, has the highest positive score in the nineteenth century, followed by Article, Dialogue, Treatise, Textbook and Lecture. This may be once more justified by the fact that most essays deal with Philosophy, a discipline whose discourse has been characterised as highly elaborate both in the late Modern period (as shown in the present study) and in the present day (as showin in Gray 2011). By contrast, the least elaborate genre is the eighteenth century dialogue, which, unlike its nineteenth-century counterpart, conveys immediacy and little formality, which seems to fit with its high involvement score on Dimension 1. Furthermore, the oral origins of Dialogue and Lecture appear to be reflected to some extent on this dimension, in that both present negative scores in the eighteenth century and positive scores in the nineteenth, the score being always higher in the case of the Dialogue, suggesting that both abandoned the immediacy characteristic of orality with time. Letters, on the contrary, which were already moderately non-elaborate in the 1700s, appear to lose structural complexity in the nineteenth century, possibly justifying, once more, their gradual disappearance from academic discourse.

Finally, as far as the dichotomy narrative vs. non-narrative is concerned, Dimension 4 scores for genre-based subregisters show that, contrary to what happens on Dimension 3, here the majority of genres do not mirror the disposition of the scientific disciplines to become more narrative with time. Rather, as we have seen earlier, three eighteenth-century genres – namely, Dialogue, Article and Lecture – stand out as highly narrative from the rest of eighteenth-century genres. Although we have agreed that these genres, being represented by one sample, should be treated with caution, they might be nonetheless considered, to some extent, indicative of the importance of the experimental accounts in the eighteenth-century astronomical discourse. On the other hand, we have also noted that the loss of narrative features in these three genres once more coincides with Atkinson's (1999: 144) findings in his study of the *Transactions*, this time with respect to Biber's (1988) Dimension 2 "Narrative vs. non-narrative concerns", which reveal that scientific research articles become less narrative along the eighteenth and nineteenth century.

However, we have also seen that, by removing the abovementioned genres from the general picture, the rest – namely, essays, treatises, textbooks and letters – do appear to reflect the progressive shift to the narrative side common to the scientific disciplines, with letters and treatises, as well as nineteenth-century articles, coinciding with the high narrative scores of Life Sciences and Astronomy in the nineteenth century. Some tentative explanations for this phenomenon have been suggested, one being that, contrarily to what happens with research articles, experimental accounts may have maintained their importance in other English scientific genres in the late Modern period, something which has been pointed out to account for the movement in the narrative direction in the three scientific disciplines. A second possible reason could, again, lie in the scientific disciplines to which the concerned genres belong, considering that all the nineteenth-century letters and the largest proportion of nineteenth-century treatises belongs to the natural sciences, which appear to be markedly narrative at that time. Finally, another possibility has been suggested to be the didactic character of those letters and of some of the treatises, particularly those written by women scientists, who, despite having used a genre considered as strictly professional in its original purpose (Crespo 2016), may have intended their writings for the instruction of the masses, including other women, and resorted to narration as a more accessible style.

Having reviewed variation and change in our corpus both at the scientific discipline and genre levels, as revealed through our four-dimensional model, we have seen that sometimes it is tricky to assign a particular linguistic or communicative tendency to the influence of a particular scientific discipline or genre. As we have seen in the above paragraphs, the discipline and genre variables seem interdependent in some cases, particularly in those where the discipline appears to condition certain aspects of the genre, as well as where the diachronic change of most genres coincides with that of the scientific disciplines, as can be best appreciated on Dimensions 2 and 3, respectively. In the case of Dimension 2, it may be suggested that the characteristics of some genres were dependent on the discipline, as can be seen, for instance, in the example of Treatise and Letter, which stay descriptive because they belong to Life Sciences, a discipline which in the late Modern period was focused on the sorting, classification and cataloguing of different species, based on their characteristics and natural environment. By contrast, Textbook has been characterised as highly argumentative in the 1700s, and appears to be the preferred genre for many eighteenth-century authors writing on Astronomy, an exact science conveyed through a discourse built of on estimations and predictions based on calculations. However, although Astronomy becomes more elaborate in the nineteenth century, textbooks move in the contrary direction, which is likely to be caused by the fact that a comparatively larger proportion of nineteenth-century textbooks are found in the Life Sciences subcorpus than in the Astronomy one.

As for the diachronic change on Dimension 3, it has been suggested that the movement towards discourse elaboration may have in fact been a general trend of English scientific writing, rather than a disposition of individual genres. This could be supported by the fact that this tendency did not spread to the epistolary genre, which would later disappear from scientific literature. The fact that essays would likewise become less popular as a scientific genre but, notwithstanding, gained discourse complexity, seems to stem from the fact that the Essay genre is tightly related to the discipline of Philosophy, being one of those "expository [genres]" which "allow for the presence of authorial views, generating discussion" (Crespo 2016: 32). On the other hand, we have likewise seen that even those essays which do not deal with philosophical matters have involvement (Dimension 1) as their distinctive characteristic, as also do our two dialogues, one of which belongs to Astronomy, a discipline characterised on the whole as informational on Dimension 1. This shows that certain genres such as Essay and Dialogue appear to be independent from the scientific discipline they deal with in some (but not all) of their discursive aspects. Finally, the analysis of discipline and genre-based variation with respect to Dimension 4 has shown that, while the scientific disciplines become more narrative over time, some of their genres, on the contrary, lose narrative features, suggesting once more the importance of variation at the different subregister levels and the difficulty in determining whether this variation is due to one or another level.

Attempting to put our findings in a tighter relation with late Modern English scientific writing as a whole, as well as with the socio-historical context of this study as outlined in Chapter 2, we have drawn the following conclusions:

First of all, as has already been observed, the progressive shift from an author-centred to an object-centred scientific discourse which can be appreciated in the general movement away from personal involvement and towards a more impersonal/informational communicative style on Dimension 1, shared by all the genres except for essays and letters, appears to reflect the gradual decrease in importance of the figure of the scientists in scientific literature, which becomes replaced by either the object of the study or the experiment that is being related. The need to persuade the reader of the trustworthiness of the experiment through rhetoric (i.e. the persuasive features of Dimension 1) became smaller as the experiment, or else precise mathematical data, became a proof of trustworthiness by themselves.

This also seems to be supported by the general increase in nominalisations that can be appreciated on Dimension 3, which causes that, during the late Modern period, noun phrases "gradually take over" verb phrases, representing "nominalized reifications of scientific activity" (Atkinson 1999: 143; see also Halliday 1988; Halliday and Martin 1993). In other words, nominalisations – together with passive structures and absence of personal pronouns – contribute to the abstractness of the scientific discourse. On the other hand, according to Bello (2014: 325), nominalisations were increasingly used as scientific discourse markers as an indication that the authors belonged to the discourse community. Thus, the progressive impersonalisation and elaboration of the texts in our corpus – especially, those belonging to the natural sciences – appear likewise to reflect the importance of the scientific community at that time, whether it was the Royal Society (especially in the eighteenth century), or the growing university circles in the 1800s. This proliferation of nominalisations is likewise due to the increasing specialisation of the scientific lexicon (Camiña-Rioboo 2013: 66-68), which added new technical terms in the different sciences during that period.

Secondly, it must also be admitted that, despite what has been observed in the above paragraphs, persuasion is still present in most of the genres in the eighteenth-century part of our corpus, suggesting that it was still important in the writing of science of that time. Furthermore, despite the apparently sharp division between involved/persuasive and informational scientific discourse on Dimension 1, we have seen that certain samples which belong to an informational discipline, such as Astronomy, had very high involved scores (such as, for instance, the eighteenth-century Dialogue and Article). In fact, if we look at Appendix II (or else to Figure 6.2 back in Chapter 6), we will see that even the nineteenth-century Article has an overall positive score, characterising it as a moderately involved/persuasive subregister. This appears to reflect the fuzzy boundaries between the natural sciences and the humanities during the period studied, justifying the rather heterogeneous term Natural Philosophy which was still in use at that time.

Notwithstanding, in what concerns the nature of individual scientific disciplines, it does seem that some dimensions in our model are more suitable than others for their description. For instance, Philosophy might be considered as most accurately represented by Dimension 1 as an involved/persuasive discipline in that it usually deals with topics subject to controversy, whether of a moral, political or religious kind, usually conveyed not merely through reasoning, but through the defence of a cause and the attack and criticism of those ideas that opposed this cause. The very titles of some philosophical works in our corpus, such as Mary Wollstonecraft's *Vindication of the Rights of Women* (1792), or Arthur Balfour's *A Defence of Philosophical Doubt* (1879), seem to do justice to their high involvement scores (1.1 and 2.3, respectively; see Appendix III). Life Sciences, in turn, appears to be better characterised by Dimension 2, which succeeds to highlight its descriptive character in the late Modern period, as opposed to Astronomy, a science based on mathematical and physical properties, conveyed at that time through logical inferences and the expression of relationships of cause and effect (especially in the eighteenth century). Dimension 3, in turn, reveals the closeness between nineteenth-century Philosophy and Astronomy in that, contrarily to Life Sciences, both become rather elaborate, something that appears to indicate that these two disciplines, despite belonging to different scientific fields and despite their difference with respect to Dimension 1, acquire a highly technical discourse with time, reflecting the general growth of a learned register for the different sciences in the 1800s. In the case of Philosophy, this growing technicality and clausal complexity does not appear to be incompatible with involvement features (even if the average for these decreases slightly in the nineteenth century). Finally, in what concerns Dimension 4, we have seen that both Astronomy and Life Sciences suggest a narrative standard in the nineteenth century, which might be justified by the continual importance of experimental accounts in the natural sciences, and also by the apparent fact that narration was resorted to in certain genres which were used with a didactic or popularising purpose.

Having attempted to answer our two initial research questions, we have seen that, first of all, this study has confirmed previous findings about the English scientific register, demonstrating once more its gradual increase in informational density, technicality, abstraction and structural complexity along the eighteenth and nineteenth century. On the other hand, it has also shown that this does not happen equally in the three disciplines and in all the genres that have been included in our sample, having revealed patterns of internal variation which are not always straightforward to interpret. Rather, as we had observed earlier, our model appears to reflect the lack of a technical standard in the English scientific discourse during the late Modern period, the necessity of which was so eagerly defended at that time. In this respect, it needs however to be remembered that we have looked at variation in only three scientific disciplines. This means that, in the light of Biber and Gray's (2014) article on the importance of variation at the subregister level, the above-suggested extension of the general characteristics of our sample to the whole of the English scientific register should be taken as a hypothesis until confirmed by further studies. On the other hand, we have likewise observed that, while some of the variation patterns we have spotted appear to be justifiable by certain aspects common to the subregister to which they belong (whether analysed as a discipline or as a genre), other patterns have proved less explicable at this stage and seem to need further investigation.

This said, and on the basis of certain difficulties that we have found during the different stages of this research, we would like to suggest certain improvements for a further study. Thus, for instance, one of the challenges that we have encountered is the unbalanced distribution of genres in our sample, which, however, is fully justified by representativeness – that is, in that the distribution it presents attempts to reflect the production at the time (see Chapter 3). While this lack of balance at the genre level by no means invalidates the study, it does not allow us to describe genres which are represented by very few text samples, as it would open the possibility to false generalisations. However, an advantage of our corpus sample is that it does present a balanced distribution of scientific disciplines, which has proved very helpful in the interpretation of our model at the discipline level. Secondly, during the factor analysis stage, it would be ideal to have a dataset of texts that would be proportionally larger to the number of variables retrieved from the corpus (i.e. linguistic features). Although our dataset meets all the preliminary requirements of sampling adequacy, it is likely that a larger dataset would further improve them, which would in turn result in the dropping of less variables – or, rather, in the necessity to drop less variables due to low individual MSAs or low communalities. In the present study, we justify the decision we took during the factor analysis stage to "save" the maximum number of linguistic features in spite of their low communalities with the premise that a statistically weaker but more interpretable model is better for an exploratory variation study than a statistically stronger one that might be hardly interpretable on a linguistic level.

All in all, although our model only explains 41 per cent of the total variation in the corpus and does not, therefore, aim to be regarded as decisive for its description, it is our hope that the patterns revealed through its four dimensions of variation might be useful for a better understanding of some characteristics of late Modern scientific English. In this light, this analysis intends to be a trial study that might be used as a reference for further research, in the hope that it might also serve for the improvement of some of the technical aspects mentioned above. In view of further research, we also hope that a confirmatory factor analysis of an enlarged version of this corpus, individually or in combination with other historical or contemporary corpora, will yield a statistically stronger model that would, in turn, reveal clearer patterns of variation – and, possibly, unveil new ones – at the different sublevels of the late Modern English scientific register.

References

Abercrombie, David (1965). *Studies in phonetics and linguistics*. London: Oxford University Press.

Abir-Am, Pnina & Outram, Dorinda (1987). Introduction: In P. Abir-Am & D. Outram (Eds.), *Uneasy careers and intimate lives: Women in science (1789-1979)* (pp. 1-16). New Brunswick, NJ: Rutgers University Press.

Aijmer, Karin (1986). Why is *actually* so frequent in spoken English? In G. Tottie & I. Bäcklund (Eds.), *English in speech and writing: A symposium* (pp. 119-129). Stockholm: Almqvist and Wiksell.

Aijmer, Karin (2008). At the interface between grammar and discourse: A corpus-based study of pragmatic markers. In Romero-Trillo, J. (Ed.), *Pragmatics and corpus linguistics: A mutualistic entente* (pp. 11-36). Berlin: Mouton de Gruyter.

Allen, Bryce, Jian Qin & F. W. Lancaster (1994). Persuasive communities: a longitudinal analysis of references in the *Philosophical Transactions of the Royal Society*, 1665-1990. *Social Studies of Science* 24(2), 279-310.

Alonso-Almeida, Francisco (2012). An Analysis of Hedging in Eighteenth Century English Astronomy Texts. In Moskowich, I. & Crespo, B. (eds.), *Astronomy 'playne and simple': The Writing of Science between 1700 and 1900* (199-220). Amsterdam; Philadelphia: John Benjamins.

Alonso-Almeida, Francisco & Inés Lareo (2016). The status of *seem* in the nineteenth century. In I. Moskowich, G. Camiña, I. Lareo & B. Crespo (Eds.), *The Conditioned and the Unconditioned': Late Modern English Texts on Philosophy* (pp. 145-165). Amsterdam: John Benjamins.

Andersen, Henning (2006). Synchrony, diachrony, and evolution. In O. N. Thomsen (Ed.), *Competing models of linguistic change: evolution and beyond*. Amsterdam: John Benjamins.

Anthony, Lawrence (2009). AntConc: A freeware concordance program for Windows, Macintosh OS X, and Linux. Available at: www.antlab.sci.waseda.ac.jp/software.html.

Asención-Delaney, Yuly (2014). A Multi-Dimensional analysis of advanced written L2 Spanish. T. In Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-dimensional analysis, 25 years on: a tribute to Douglas Biber* (pp. 35-79). Amsterdam: John Benjamins.

Asención-Delaney, Yuly & Collentine, Joe (2011). A multidimensional analysis of a Written L2 Spanish Corpus. *Applied Linguistics* 32, 299-322.

Aston, Guy & Burnard, Lou (1998). *The BNC handbook*. Edinburgh: Edinburgh University Press.

Atkinson, Dwight (1992). The evolution of medical research writing from 1735 to 1985: the case of the Edinburgh Medical Journal. *Applied Linguistics* 13(4), 337-374.

Atkinson, Dwight (1996). The Philosophical Transactions of the Royal Society of London, 1675-1975: A sociohistorical discourse analysis. *Language in Society* 25, 333-371.

Atkinson, Dwight (1999). *Scientific discourse in sociohistorical context: The Philosophical Transactions of the Royal Society of London, 1675-1975*. Mahwah, NJ: Erlbaum.

Baayen, Harald & Renouf, Antoinette (1996). Chronicling The Times: Productive lexical innovations in an English newspaper. *Language* 72, 69-96.

Bacon, Francis (1605 [1875]). Of the Dignity and Advancement of Learning. In J. Spedding, R. L. Ellis & D. D. Heath (Eds.), *The Works of Francis Bacon*, Vol. IV (pp. 273-498). Longmans, Cumpers, and Co.

Bacon, Francis (1620 [1875]). The New Organon. In J. Spedding, R. L. Ellis & D. D. Heath (Eds.), *The Works of Francis Bacon*, Vol. IV (pp. 39-248). Longmans, Cumpers, and Co.

Bailey, Richard (1996). *Nineteenth-century English*. Ann Arbor: University of Michigan Press.

Bailey, Richard (2003). The ideology of English in the long nineteenth century. In M. Dossena & C. Jones, *Insights into Late Modern English* (pp. 21-44). Bern: Peter Lang.

Baker, Mona (2002). Corpus-based studies within the larger context of translation studies. *Genesis: Revista Científica do ISAI* 2. Available at: https://www.research.manchester.ac.uk/portal/en/publications/corpusbased-studies-within-the-larger-context-of-translation-studies(9eb8126a-f38b-4352-a98d-d64a841ba37b).html.

Banks, David (2002). Systemic Functional Linguistics as a model for text analysis. *ASp, la revue du GERAS* 35/36, 23-35.

Banks, David (Ed.) (2004). *Text and texture, systemic functional viewpoints on the nature and structure of text*. Paris: L'Harmattan.

Banks, David (2005). *Introduction à la linguistique systémique fonctionnelle de l'anglais*. Paris: L'Harmattan.

Banks, David (2008). *The Development of Scientific Writing*. *Linguistic Features and Historical Context*. London/Oakville: Equinox.

Barber, Charles (1976). *Early Modern English*. Edinburgh: Edinburgh University Press.

Barber, Charles (2004). *The English Language: A Historical Introduction* (5th ed.) Cambridge: Cambridge University Press.

Bartlett, Maurice S. (1937). The statistical conception of mental factors. *British Journal of Psychology* 28, 97-104.

Baugh, Albert & Cable, Thomas (2002). *A history of the English language* (5th ed.) London: Routledge.

Bazerman, Charles (1984). Modern evolution of the experimental report in Physics: Spectroscopic articles in Physical Review, 1893-1980. *Social Studies of Science* 14, 163-196.

Bazerman, Charles (1988). *Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science*. Madison: The University of Wisconsin Press.

Bazerman, Charles (1993). From cultural criticism to disciplinary participation: Living with powerful words. In A. Herrington and C. Moran (Eds.), *Writing, teaching and learning in the disciplines* (pp. 61-68). New York: Modern Language Association.

Bazerman, Charles (1994). *Constructing experience*. Carbondale: Southern Illinois University Press.

Beal, Joan (2004). *English in Modern Times*. London: Arnold.

Beal, Joan (2010). Prescriptivism and the suppression of variation. In R. Hickey (Ed.), *Eighteenth-century English: Ideology and change* (pp. 21-37). Cambridge: Cambridge University Press.

Beal, Joan (2012). Late Modern English in its historical context. In I. Moskowich & B. Crespo, *Astronomy 'playne and simple': the writing of science between 1700 and 1900*. Amsterdam: John Benjamins.

Bello Viruega, Iria M. (2010). Nominalizations in astronomical texts in the eighteenth century. In I. Moskowich, B. Crespo, I. Lareo & P. Lojo (Eds.), *Language windowing though corpora. Visualización del lenguaje a través de corpus* (pp. 75-89). A Coruña: Universidade da Coruña.

Bello Viruega, Iria M. (2014). On How *the Motion of the Stars* Changed the Language of Science: A Corpus-based Study of Deverbal Nominalizations in Astronomy Texts from 1700 to 1900. Unpublished PhD Dissertation. University of A Coruña.

Bello Viruega, Iria M. (2016). Reflections on our astronomical undertaking: nominalizations and possessive structures in the Coruña Corpus. In F. Alonso-Almeida, L. Cruz García & V. González Ruiz (Eds.), *Corpus-based Studies on Language Varieties. Linguistic Insights* (pp. 217-232). Bern: Peter Lang.

Berkenkotter, Carol & Huckin, Thomas N. (1995). *Genre Knowledge in Disciplinary Communication: Cognition/Culture/Power*. New Jersey: Lawrence Erlbaum Associates, Publishers.

Berber Sardinha, Tony (2014). 25 years later: comparing Internet and pre-Internet registers. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-dimensional* *analysis, 25 years on: a tribute to Douglas Biber* (pp. 81-107). Amsterdam: John Benjamins.

Berber Sardinha, Tony, Kaufmann, Carlos & Acunzo, Cristina M. (2014). Dimensions of register variation in Brazilian Portuguese. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-dimensional analysis, 25 years on: a tribute to Douglas Biber* (pp. 35-79). Amsterdam: John Benjamins.

Berns, Jan & Van Marle, Jaap (Eds.) (2002). *Present-day dialectology: problems and findings*. Berlin; New York: Mouton de Gruyter.

Bértoli-Dutra, Patricia (2014). Multi-Dimensional analysis of pop songs. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-dimensional analysis, 25 years on: a tribute to Douglas Biber* (pp. 149-175). Amsterdam: John Benjamins.

Besnier, Nico (1988). The linguistic relationships of spoken and written Nukulaelae registers. *Language* 64, 707-736.

Besnier, Nico (1994). Involvement in linguistic practice: an ethnographic appraisal. *Journal of Pragmatics* 22, 279-299.

Bhatia, Vijay (1993). *Analysing genre: Language use in professional settings*. Applied Linguistics and Language Study Series. London: Longman.

Bhatia, Vijay (1996). Methodological issues in genre analysis. *Hermes, Journal of Linguistics* 16, 39-59.

Bhatia, Vijay (2002). Applied genre analysis: a multi-perspective model. *Ibérica* 4, 3-19.

Bhatia, Vijay (2004). *Worlds of written discourse. A genre-based view*. London: Continuum International.

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

Biber, Douglas (1989). A typology of English texts. *Linguistics* 27, 3-43.

Biber, Douglas (1993). Representativeness in corpus design. *Literary and Linguistic Computing* 19, 219-241.

Biber, Douglas (1994). An analytical framework on register studies. In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 31-57). New York: Oxford University Press.

Biber, Douglas (1995). *Dimensions of register variation: a cross linguistic comparison*. Cambridge: Cambridge University Press.

Biber, Douglas (1999). A register perspective on grammar and discourse: Variability in the form and use of English complement clauses. *Discourse Studies* 1(2), 131-150.

Biber, Douglas (2001). Dimensions of variation among eighteenth-century speech-based and written registers. In S. Conrad & D. Biber (Eds.), *Variation in English: Multidimensional studies* (pp. 200-214). Harlow: Longman.

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 & R. Jones (Eds.), *Advances in discourse studies* (pp. 100-114). London: Routledge.

Biber, Douglas (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14(3), 275-311.

Biber, Douglas (2014). Multi-Dimensional Analysis: A personal history. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-dimensional analysis, 25 years on: a tribute to Douglas Biber* (xxix-xxxviii). Amsterdam: John Benjamins.

Biber, Douglas & Burges, Jená (2000). Historical change in the language use of women and men: Gender differences in dramatic dialogue. *Journal of English Linguistics* 28(1): 21-37.

Biber, Douglas & Burges, Jená (2001). Historical shifts in the language of women and men: Gender differences in dramatic dialogue. In S. Conrad & D. Biber (Eds.), *Variation in English: Multidimensional studies* (pp. 157-170). Harlow: Longman.

Biber, Douglas & Susan Conrad (2001). Introduction: Multi-dimensional analysis and the study of register variation. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-Dimensional Studies* (pp. 3-12). Harlow: Longman.

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

Biber, Douglas & Finegan, Edward (1986). An initial typology of English text types. In J. Aarts & W. Meijs (Eds.), *Corpus linguistics. Vol. II: New studies in the analysis and exploitation of computer corpora* (pp. 19-46). Amsterdam: Rodopi.

Biber, Douglas & Finegan, Edward (1989). Drift and the evolution of English style: A history of three genres. *Language* 65(3), 487-517.

Biber, Douglas & Finegan, Edward (1992). The linguistic evolution of five written and speech-based English genres from the seventeenth to the twentieth centuries. In M. Rissanen, O. Ihalainen, T. Nevalainen & I. Taavitsainen (Eds.), *History of Englishes: New methods and interpretations in historical linguistics* (pp. 688-704). Berlin: Mouton.

Biber, Douglas & Finegan, Edward (1994). Introduction: Situating register in sociolinguistics. In D. Biber and E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 3-12). New York: Oxford University Press.

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

Biber, Douglas & Finegan, Edward (2001a). Diachronic relations among speech-based and written registers in English. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-Dimensional Studies* (pp. 66-83). Harlow: Longman.

Biber, Douglas & Finegan Edward (2001b). Intra-textual variation within medical research articles. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-Dimensional Studies* (pp. 108-123). Harlow: Longman.

Biber, Douglas & Gray, Bethany (2013a). Being specific about historical change: the influence of sub-register. *Journal of English Linguistics* 41, 104-134.

Biber, Douglas & Gray, Bethany (2013b). Identifying multi-dimensional patterns of variation across registers. In M. Krug & J. Schlüter (Eds.), *Research methods in language variation and change* (pp. 402-420). Cambridge: Cambridge University Press.

Biber, Douglas & Mohamed Hared (1992). Dimensions of register variation in Somali. *Language Variation and Change* 4, 41-75.

Biber, Douglas & Mohamed Hared (1994). Linguistic correlates of the transition to literacy in Somali: Language adaptation in six press registers. In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 182-216). Oxford: Oxford University Press.

Biber, Douglas, Edward Finegan & Dwight Atkinson (1994). ARCHER and its challenges: Compiling and exploring A Representative Corpus of Historical English Registers. In U. Fries, P. Schneider & G. Tottie (Eds.), *Creating and using English language corpora. Papers from the 14th International Conference on English Language Research on Computerized Corpora, Zurich 1993* (pp. 1-13). Amsterdam: Rodopi.

Biber, Douglas, Susan Conrad & Randi Reppen (1998). *Corpus linguistics: Investigating language structure and use.* Cambridge: Cambridge University Press.

Biber, Douglas, Susan Conrad & Viviana Cortes (2004). If you look at… Lexical bundles in university lectures and textbooks. *Applied Linguistics* 25, 371-405.

Biber, Douglas, Mark Davies, James K. Jones & Nicole Tracy-Ventura (2006). Spoken and written register variation in Spanish: A Multi-Dimensional analysis. *Corpora* 1, 7-38.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan (1999). *Longman Grammar of Spoken and Written English*. London: Pearson Education.

Bitzer, Lloyd (1968). The rhetorical situation. *Philosophy and Rhetoric* 1. 1-14.

Bourcier, Georges (1981). *An introduction to the history of the English language*. Translated and adapted by C. Clark. Cheltenham: Thornes.

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

Boyle, Robert (1661 [1772]). Proemial Essay. In T. Birch (Ed.), *The works of Robert Boyle*, Vol I. (pp. 192-204). London: J. & F. Rivington, 1772, rpt. Hildesheim: Georg Olms, 1965.

Bright, William (1976). *Variation and change in language: essays*. Stanford, CA: Stanford University Press.

Brown, Gillian & Yule, George (1983). *Discourse analysis*. Cambridge: Cambridge University Press.

Brown, Penelope & Colin Fraser (1979). Speech as a marker of situation. In K. Schrerer & H. Giles (Eds.), *Social markers in speech* (pp. 33-62). Cambridge: Cambridge University Press.

Brück, Mary T. (1991). Companions in astronomy: Margaret Lindsay Huggins and Agnes Mary Clerke. *Irish Astronomical Journal* 20(2), 70-77.

Brück, Mary T. (2004). *Agnes Mary Clerke and the Rise of Astrophysics*. Cambridge: Cambridge University Press.

Bruthiaux, Paul (1994). 'Me Tarzan, you Jane': Linguistic simplification in 'personal ads' register. In D. Biber & E. Finegan (Eds.), *Sociolinguistic Perspectives on Register* (pp. 136-154). New York: Oxford University Press.

Bruthiaux, Paul (1996). *The discourse of classified advertising.* New York: Oxford University Press.

Bryant, Walter W. (2014) [1907]. *A History of Astronomy*. London; New York. Routledge.

Bryant, Fred B. & Yarnold, Paul R. (1995). Principal-components analysis and exploratory and confirmatory factor analysis. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 99-136). Washington, DC: American Psychological Association.

Bunton, David (2002). Generic moves in PhD thesis introductions. In J. Flowerdew (Ed.). *Academic Discourse* (pp. 57-75). London: Longman.

Burke, Peter (2000). *A social history of knowledge: from Gutenberg to Diderot*. Cambridge: Polity.

Burnley, David (2014). *The History of the English Language: A Sourcebook*. London: Routledge.

Cabezón, José I. (1998). Introduction. In J. I. Cabezón (Ed.), *Scholasticism: cross-cultural and comparative perspectives* (pp. 1-17). Albany: State University of New York Press.

Camiña-Rioboo, Gonzalo (2010). New nouns for new ideas. In M. LL. Gea-Valor, I. García, and M. J. Esteve (Eds.), *Linguistic and Translation Studies in scientific Comunications*, Vol. 86 (pp. 157-176). Bern: Peter Lang.

Camiña-Rioboo, Gonzalo (2012). Accounting for observations of the heavens in the 18th century: New nouns to explain old phenomena. In I. Moskowich & B. Crespo (Eds.), '*Astronomy 'playne and simple': The Writing of Science between 1700 and 1900* (pp. 93-121). Amsterdam: John Benjamins.

Camiña-Rioboo, Gonzalo (2013). *Noun Formation in the Scientific Register of Late Modern English: A Corpus-based Approach*. Unpublished PhD dissertation. University of A Coruña.

Camiña-Rioboo, Gonzalo & Inés Lareo (2016). Editorial policy in the *Corpus of English Philosophy Texts*. In I. Moskowich, G. Camiña, I. Lareo & B. Crespo

(Eds.), *The Conditioned and the Unconditioned': Late Modern English Texts on Philosophy* (pp. 45-60). Amsterdam: John Benjamins.

Carkin, Susan (2001). *Pedagogic language in introductory classes: A multi-dimensional analysis of textbooks and lectures in Biology and Macroeconomics*. Unpublished PhD dissertation. Northern Arizona University.

Carroll, John B. (1960). Vectors of prose style. In T. A. Sebeok (Ed.), *Style in language* (pp. 283-292). Cambridge, Mass.: MIT Press.

Cattell, Raymond B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research* 1, 245-276.

Chafe, Wallace (1982). Integration and involvement in speaking, writing, and oral literature. In D. Tannen (Ed.), *Spoken and written language: Exploring orality and literacy* (pp. 35-54). Norwood, NJ: Ablex.

Chafe, Wallace & Jane Danielewicz (1987). Properties of Spoken and Written Language. In R. Horowitz & S. J. Samuels (Eds.)*, Comprehending Oral and Written Language* (pp. 83-113). San Diego: Academic Press.

Chambers, Jack & Trudgill, Peter (1998). *Dialectology*. Cambridge: Cambridge University Press.

Channel, Joanna (1981). Applying semantic theory to vocabulary teaching. *English Language Teaching Journal* 35(12), 115-122.

Chomsky, Noam (1957). *Syntactic structures*. The Hague; Paris: Mouton.

Chomsky, Noam (1962). A transformational approach to syntax. In A. Hill (Ed.), *Proceedings of the Third Texas Conference on Problems of Linguistic Analysis in English, May 9-12, 1958* (pp. 124-148). Austin: University of Texas Press.

Chomsky, Noam (1988). *Generative grammar: Its basis, development and prospects*. Kyoto: University of Foreign Studies.

Condi de Souza, Renata (2014). Dimensions of variation in *TIME* magazine. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-dimensional analysis, 25 years on: a tribute to Douglas Biber* (pp. 177-193). Amsterdam: John Benjamins.

Conrad, Susan (1996). Investigating academic texts with corpus-based techniques: An example from biology. *Linguistics and Education* 8, 299-326.

Conrad, Susan (2001). Variation among disciplinary texts: A comparison of textbooks and journal articles in biology and history. In S. Conrad. & D. Biber (Eds.), *Variation in English: Multi-Dimensional Studies* (pp. 94-107). Harlow: Longman.

Conrad, Susan (2002). Corpus linguistic approaches for discourse analysis. *Annual Review of Applied Linguistics* 22, 75-95.

Conrad, Susan & Biber, Douglas (2001). Multi-dimensional methodology and the dimensions of register variation in English. In S. Conrad & D. Biber (Eds.), *Variation in English: multi-dimensional studies* (pp. 13-42). Harlow: Longman.

Cormack, Leslie & Ede, Andrew (2012). *A History of Science in Society: From Philosophy to Utility*. Second Edition.

Couture, Barbara (1986). Epistemologies and Methodologies: Research in Written Language Function. In B. Couture (Ed.), *Functional Approaches to Writing: Research Perspectives* (pp. 1-10). Norwood, N.J.: Ablex Publishers.

Crespo, Begoña (2011). Persuasion markers and ideology in eighteenth century philosophy texts. *Revista de Lenguas para Fines Específicos* 17, 199-228.

Crespo, Begoña (2012a). Astronomy as scientific knowledge in Modern England. In I. Moskowich & B. Crespo (Eds.), *Astronomy 'playne and simple': The Writing of Science between 1700 and 1900* (pp. 15-34). Amsterdam: John Benjamins.

Crespo, Begoña (2012b). Astronomical discourse in 18th century texts: a new-born model in the transmission of science. In I. Moskowich & B. Crespo (Eds.), *Astronomy 'playne and simple': The Writing of Science between 1700 and 1900* (pp. 57-78). Amsterdam: John Benjamins.

Crespo, Begoña (2014). Female Authorial Voice: Discursive Practices in Prefaces to Scientific Works. In M. Gotti & D. S. Giannoni (Eds.), *Corpus Analysis for Descriptive and Pedagogic Purposes: English Specialised Discourse* (pp. 189-202). Bern: Peter Lang.

Crespo, Begoña (2016). Genre categorisation in *CEPhiT*. In I. Moskowich, G. Camiña, I. Lareo & B. Crespo (Eds.), *The Conditioned and the Unconditioned': Late Modern English Texts on Philosophy* (pp. 25-44). Amsterdam: John Benjamins.

Crespo, Begoña & Moskowich, Isabel (2010). CETA in the context of the Coruña Corpus. *Literary and Linguistic Computing* 25(2), 153-164.

Croft, William & Poole, Keith (2008). Inferring universals from grammatical variation: Multidimensional scaling for typological analysis. *Theoretical Linguistics* 34(1), 1-37.

Crombie, Alistair (1953). *Robert Grosseteste and the origins of Modern science*. Oxford: Clarendon Press.

Crombie, Alistair (1959). *Augustine to Galileo*. London: Heinemann.

Crombie, Alistair (1969). The significance of medieval discussions of scientific method for the Scientific Revolution. In M. Clagett (Ed.), *Critical problems in the history of science* (pp. 79-102). Madison, WI: University of Wisconsin Press.

Crombie, Alistair (1990). *Science, optics, and music in Medieval and Early Modern thought*. London: A&C Black.

Crompton, Peter (1997). Hedging in academic writing: some theoretical problems. *English for Specific Purposes* 16(4), 271-287.

Crowe, Michael J. (1994). *Modern theories of the universe: From Herschel to Hubble*. New York: Dover.

Crystal, David & Davy, Derek (1969). *Investigating English Style*. Harlow: Longman.

Csomay, Eniko (2000). Academic lectures: An interface of an oral and literate continuum. *Novelty* 7, 30-46.

Csomay, Eniko (2002). Variation in academic lectures. In R. Reppen (Ed.), *Using corpora to explore linguistic variation* (pp. 203-224). Philadelphia: John Benjamins.

Culpeper, Jonathan & Kytö, Merja (2010). *Early Modern English Dialogues: Spoken interaction as writing*. Cambridge: Cambridge University Press.

Dampier, William (1948 [1929]). *A History of Science and Its Relations with Philosophy and Religion* (3rd ed.). Cambridge: Cambridge University Press.

Davies, Mark (2007). *TIME Magazine Corpus: 100 million words, 1920s-2000s*. Available online at http://corpus.byu.edu/time/.

Davies, Mark. (2008-). *The Corpus of Contemporary American English*: *400+ million words, 1990-present*. Available online at http://www.americancorpus.org/.

Davies, Mark (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics* 14(2), 159-190.

Davies, Mark (2010-). *The Corpus of Historical American English: 400 million words, 1810-2009*. Available online at http://corpus.byu.edu/coha/.

Dear, Peter (1995). *Discipline and experience: the mathematical way in the Scientific Revolution*. Chicago: University of Chicago Press.

Denison, David (1998). Syntax. In S. Romaine (Ed.), *The Cambridge History of the English Language, Volume IV: 1776-1997* (pp. 92-329). Cambridge: Cambridge University Press.

Deuber, Dagmar (2014). *English in the Caribbean: variation, style and standards in Jamaica and Trinidad*. Studies in English Language. Cambridge: Cambridge University Press.

Dewhirst, David & Michael Hoskin (1999). The message of starlight: the rise of astrophysics. In M. Hoskin (Ed.), *The Cambridge Concise History of Astronomy* (pp. 219-305). Cambridge: Cambridge University Press.

Downing, Angela (1997). Encapsulating discourse topics. *Estudios Ingleses de la Universidad Complutense* 5, 147-168.

Dreyer, John L. E. (1953). *A History of Astronomy from Thales to Kepler*. New York: Dover Publications.

Dudley-Evans, Tony (1986). Genre analysis: an investigation of the introduction and discussion sections of M. Sc. dissertations. In M. Coulthard (Ed.) *Talking about text* (pp. 128-145). Birmingham: English Language Research. University of Birmingham, UK.

Ellis, Jeffrey & Ure, Jean (1969). Language varieties: Register. In A. R. Meetham (Ed.), *Encyclopaedia of Linguistics, Information and Control* (pp. 251-259). Oxford: Pergamon.

Erwin-Trypp, Susan (1972). On sociolinguistic rules: alternation and co-occurrence. In J. Gumperz and D. Hymes (Eds.), *Directions in sociolinguistics* (pp. 213-250). New York: Holt, Rinehart and Winston.

Esteve, María José, Inés Lareo & Gonzalo Camiña-Rioboo (2010). A study of nouns and their provenance in the astronomy section of the Coruña Corpus of Scientific English Writing. Some preliminary considerations. In Cifuentes Honrubia, J. L., A. Gómez González-Jover, A. Lillo Buades, J. Mateo Martínez & F. Yus Ramos (Eds.), *Los caminos de la lengua. Estudios en homenaje a Enrique Alcaraz Varó* (pp. 537-546). Alicante: Universidad de Alicante.

Everitt, Brian (1975). Multivariate analysis: The need for data, and other problems. *British Journal of Psychiatry* 126, 2S7-240.

Everitt, Brian & Hothorn, Torsten (2011). *An introduction to applied multivariate analysis with R*. New York: Springer.

Fabrigar, Leandre & Wegener, Duane (2012). *Exploratory factor analysis*. Oxford: Oxford University Press.

Fabrigar, Leandre, Duane Wegener, Robert MacCallum & Erin Strahan (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods 4,* 272-299.

Farhady, Hossein (1983). On the plausibility of the unitary language proficiency factor. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 11-28). Rowley, Mass.: Newbury House.

Ferguson, Charles A. (1983). Sports announcer talk: Syntactic aspects of register variation. *Language in Society* 12(2), 153-172.

Forbes, Robert J. & Dijksterhuis, Eduard J. (1963). *A history of science and technology*: *the eighteenth and nineteenth centuries*. London: Pelican.

Francis, W. Nelson & Kučera, Henry (1964). *Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*. Providence, Rhode Island: Department of Linguistics, Brown University.

Fried, Mirjam (2010). Introduction: from instances of change to explanations of change. In M. Fried, J. Ostman & J. Verschueren, *Variation and change: pragmatic perspectives*. Amsterdam: John Benjamins.

Friginal, Eric (2009). *The language of outsourced call centers: a corpus-based study of cross-cultural interaction* (1-16). Amsterdam; Philadelphia: John Benjamins.

Friginal, Eric & Hardy, Jack A. (2014). *Corpus-based sociolinguistics: A guide for students*. London: Routledge.

Gerzymisch-Arbogast, Heidrun (1993). Contrastive scientific and technical register as a translation problem. In S. E. Wright & L. D. Wright (Eds.), *Scientific and Technical Translation* (pp. 21-52). Amsterdam; Philadelphia: John Benjamins.

Giltrow, Janet (2010). Genre as difference: The sociality of linguistic variation. In H. Dorgeloh & A. Wanner (Eds.), *Syntactic variation and genre* (pp. 29-52). Berlin: Mouton de Gruyter.

Gómez Guinovart, Javier & Pérez Guerra, Javier (2000). A multidimensional corpus-based analysis of english spoken and written-to-be-spoken discourse. *Cuadernos de Filología Inglesa* 9(1), 39-70.

González-Álvarez, Dolores & Pérez Guerra, Javier (1998). Texting the written evidence: On register analysis in late Middle English and early Modern English. *Text* 18(3), 321-348.

Görlach, Manfred (1991). *Introduction to Early Modern English*. Cambridge: Cambridge University Press.

Görlach, Manfred (1994). *The Linguistic History of English*. London: Macmillan.

Görlach, Manfred (2001). *Eighteenth-century English*. Heidelberg: Universitätsverlag.

Görlach, Manfred (2004). *Text types and the history of English*. Berlin: Mouton de Gruyter.

Gorsuch, Richard L. (1983). *Factor analysis*. Hillsdale, NJ: Erlbaum.

Gotti, Maurizio (1996). *Robert Boyle and the science of language*. Milano: Guerini Scientifica.

Gotti, Maurizio (2001). The experimental essay in Early Modern English. *European Journal of English Studies* 5(2), 221-239.

Gotti, Maurizio (2003). *Specialised discourse: linguistic features and changing conventions*. Bern: Peter Lang.

Gotti, Maurizio (2008). *Investigating specialised discourse*. Bern: Peter Lang.

Gray, Bethany (2011). *Exploring academic writing through corpus linguistics: When discipline tells only part of the story*. Unpublished PhD dissertation. Northern Arizona University, Flagstaff, AZ.

Gray, Bethany & Biber, Douglas (2012). The emergence and evolution of the pattern N + PREP + V-ing in historical scientific texts. In I. Moskowich & B. Crespo (Eds.), *Astronomy 'playne and simple': The writing of science between 1700 and 1900* (pp. 181-198). Amsterdam: John Benjamins.

Gregory, Michael (1967). Aspects of varieties differentiation. *Journal of Linguistics* 3, 177-197.

Gregory, Michael & Carroll, Suzanne (1978). *Language and situation: Language varieties and their social contexts.* London: Routledge and Kegan Paul.

Grice, James W. (2001). Computing and evaluating factor scores. *Psychological Methods* 6(4), 430-450.

Grieve, Jack (2014). A Multi-Dimensional analysis of regional variation in American English. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-dimensional analysis, 25 years on: a tribute to Douglas Biber* (pp. 3-33). Amsterdam: John Benjamins.

Grieve, Jack (2016). *Regional variation in written American English*. Cambridge: Cambridge University Press.

Grieve, Jack, Douglas Biber, Eric Friginal & Tatiana Nekrasova (2011). Variation among blogs: A Multi-Dimensional analysis. In A. Mehler, S. Sharoff & M. Santini (Eds.), *Genres on the web: Computational models and empirical studies* (pp. 303-322). London: Springer.

Groom, Nicholas (2010). Closed-class keywords and corpus-driven discourse analysis. In M. Bondi & M. Scott (Eds.), *Keyness in texts* (pp. 59-78). Amsterdam: John Benjamins.

Gumperz, Jonh J. (1982a). *Discourse strategies*. Cambridge: Cambridge University Press.

Gumperz, John J. (1982b). *Language and social identity*. Cambridge: Cambridge University Press.

Gumperz, John J. & Hymes, Dell (eds.) (1972). *Directions in sociolinguistics*. New York: Holt, Rinehart, and Winston.

Hair, J. F., Anderson, R. E., Tatham, R. L., and Black, W. C. (1998). *Multivariate data analysis* (5th ed.). Upper Saddle River: Prentice Hall.

Hall, Alfred R. (1954). *The Scientific Revolution 1500-1800: The formation of the modern scientific attitude*. London; New York: Longmans, Green and co.

Halliday, Michael A. K. (1978) *Language as social semiotic: the social interpretation of language and meaning*. London: Edward Arnold.

Halliday, Michael A. K. (1985) *Spoken and written language*. Geelong, Victoria: Deakin University Press.

Halliday, Michael A. K. (1988). On the language of physical science. In M. Ghadessy (Ed.), *Registers of written English: situational factors and linguistic features* (172-168). London: Pinter.

Halliday, Michael A. K (1989). Some grammatical problems in scientific English. *Australian Review of Applied Linguistics: Genre and Systemic Functional Studies*, Series 5(6), 13-37.

Halliday, Michael A. K. (1990). The construction of knowledge and value in the grammar of scientific discourse: Charles Darwin's The Origin of the Species. In C. De Stasio, M. Gotti & R. Bonadei (Eds.), *La rappresentazione verbale e iconica* (pp. 57-80). Milano: Guerini.

Halliday, Michael A. K. & Hasan, Ruqaiya (1980). Text and context: aspects of language in a social-semiotic perspective. *Sophia Linguistica: working papers in linguistics* 6, 4-91.

Halliday, Michael A. K. & Hasan, Ruqaiya (1989). *Language, context and text: aspects of language in a social-semiotic perspective*. Oxford: Oxford University Press.

Halliday, Michael A. K., McIntosh, Angus & Strevens, Paul D. (1964). *The linguistic sciences and language teaching*. London: Longmans Green and co ltd.

Hämäläinen, Riitta-Liisa (2008). Text type distinctions and variation in English software Engineering. Unpublished MA dissertation. University of Jyväskylä.

Hardie, Andrew (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17(3), 380-409.

Hardie, Andrew (2016). Infrastructure for analysis of the *CEPhiT* corpus: Implementation and applications of corpus annotation and indexing. In I. Moskowich, G. Camiña, I. Lareo & B. Crespo (Eds.), *The Conditioned and the Unconditioned': Late Modern English Texts on Philosophy* (pp. 61-76). Amsterdam: John Benjamins.

Hasan, Ruqaiya (1973). Code, register, and social dialect. In B. Bernstein (Ed.), *Class, codes and control* (pp. 253-292). Berlin & New York: Mouton de Gruyter.

Hatcher, Larry (1994). *A step-by-step approach to using the SAS System for factor analysis and structural equation modeling*. Cary, NC: SAS Institute Inc.

Hatim, Basil & Mason, Ian (1990). *Discourse and the translator*. New York: Longman.

Heath, Shirley B. & Juliet Langman (1994). Shared thinking and the register of coaching. In D. Biber & E. Finegan (Eds.), *Sociolinguistic Perspectives on Register* (pp. 82-105). Oxford: Oxford University Press.

Helt, Marie E. (2001). A multi-dimensional comparison if British and American spoken English. In S. Conrad and D. Biber (Eds.), *Variation in English: Multi-Dimensional Studies* (pp. 171-184). Harlow: Longman.

Hendrickson, Alan E. & White, Paul O. (1964). Promax: A quick method for rotation to oblique simple structure. *British Journal of Mathematical and Statistical Psychology* 17(1), 65-70.

Henry, John (2002). *The Scientific Revolution and the origins of Modern science*. New York: Palgrave.

Herrero-López, Concepción (2007). Las mujeres en la investigación científica. *Criterios* 8, 75-96.

Hickey, Raymond (2010). Attitudes and concerns in eighteenth-century English. In R. Hickey (Ed.), *Eighteenth-century English: Ideology and change* (pp. 1-20). Cambridge: Cambridge University Press.

Hoenigswald, Henry M. (1960). *Language change and linguistic reconstruction*. Chicago: University of Chicago Press.

Hoffmann, Sebastian, Stefan Evert, Nicholas Smith, David. Y. W. Lee & Ylva Berglund Prytz (2008). *Corpus Lnguistics with BNCWeb — a Practical Guide*. Frankfurt am Main: Peter Lang.

Hundt, Marianne (1997). Has British English been catching up with American English over the past thirty years? In M. Ljung (Ed.), *Corpus-Based Studies in English: Papers from the Seventeenth International Conference on English-Language Research Based on Computerized Corpora (ICAME 17)* (pp. 135-151). Amsterdam: Rodopi.

Hundt, Marianne, Andrea Sand & Paul Skandera (1999a). *Manual of Information to accompany The Freiburg – Brown Corpus of American English ('Frown')*. Freiburg: Department of English. Albert-Ludwigs-Universität Freiburg.

Hundt, Marianne, Andrea Sand & Rainer Siemund (1999b). *Manual of Information to accompany The Freiburg – LOB Corpus of British English ('FLOB')*. Freiburg: Department of English. Albert-Ludwigs-Universität Freiburg.

Hunston, Susan (1993). Evaluation and ideology in scientific discourse. In M. Ghadessy (Ed.), *Register analysis: Theory and practice* (pp. 57-73). London: Pinter.

Hunter, Michael C. (1989). *Establishing the new science: The experience of the early Royal Society*. New York: Boydell & Brewer.

Hyland, Ken (1995). The author in the text: hedging scientific writing. *Hong Kong Papers in Linguistics and Language Teaching* 18, 33-42.

Hyland, Ken (1996). Writing without conviction? Hedging in science research articles. *Applied Linguistics* 17(4), 433-454.

Hyland, Ken (1998a). Boosting, hedging and the negotiation of academic knowledge. *Text* 18, 349-382.

Hyland, Ken (1998b). *Hedging in scientific research articles*. Amsterdam; Philadelphia: John Benjamins.

Hyland, Ken & Marina Bondi (Eds.) (2006). *Academic discourse across disciplines*. Bern: Peter Lang.

Hymes, Dell (1962). The ethnography of speaking. In T. Gladwin & and W. Sturtevant (Eds.), *Anthropology and human behaviour* (pp. 13-53). Washington, DC: Anthropology Society of Washington.

Hymes, Dell (1974). *Foundations in sociolinguistics*. Philadelphia: University of Pennsylvania Press.

Hymes, Dell (1976). Models of the interaction of language and social setting. *Journal of Social Issues* 23(2), 8-38.

Hymes, Dell (1984). Sociolinguistics: Stability and Consolidation. *International Journal of the Sociology of Language* 45, 39-45.

Irvine, Judith (1979). Formality and informality in communicative events. In J. Baugh & J. Sherzer (Eds.), *Language in use: Readings in sociolinguistics* (pp. 211-228). Englewood Cliffs, NJ: Prince-Hall.

Jackson, J. Edward (2005). Oblimin rotation. *Encyclopedia of Biostatistics* 6. Wiley Online Library.

Jang, S.-C. (1998). *Dimensions of spoken and written Taiwanese: A corpus-based register study*. Unpublished PhD dissertation. University of Hawaii.

Johansson, Christine & Christian Geisler (1996). Pied piping in spoken English. In A. Renouf (Ed.), *Explorations in corpus linguistics* (pp. 67-82). Amsterdam: Rodopi.

Johansson, Stig (Ed.) (1982). *Computer corpora in English language research*. Bergen: Norwegian Computing Centre for the Humanities.

Johansson, Stig, Geoffrey Leech & Helen Goodluck (1978). *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers*. Department of English, University of Oslo.

Johnstone, Barbara (2002). *Discourse analysis*. Oxford: Blackwell.

Joos, Martin (1961). *The five clocks*. New York: Harcourt.

Kaiser, Henry F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika* 23(3), 187-200.

Kaiser, Henry F. (1970). A second-generation Little Jiffy. *Psychometrika* 35, 401-415.

Kaiser, Henry F. (1974). An index of factorial simplicity. *Psychometrika* 39, 31-36.

Kennedy, Graeme (2003). Amplifier collocations in the British National Corpus: Implications for English language teaching. *Tesol Quarterly* 37(3), 467-487.

Kiesling, Scott F. (2011). *Language variation and change*. Edinburgh: Edinburgh University Press.

Kilgarriff, Adam & Grefenstette, Gregory (2003). Introduction to the special issue on Web as a Corpus. *Computational Linguistics* 29(3), 1-15.

Kim, Y. J. & Biber, Douglas (1994). A corpus-based analysis of language variation in Korean. In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 157-181). Oxford: Oxford University Press.

Kim, Jae-On & Mueller, Charles W. (1978). *Introduction to factor analysis: What it is and how to do it*. Beverly Hills, CA: Sage.

Kinneavy, J. (1971). *A theory of discourse*. Toronto: Prentice-Hall, Inc.

Kline, Paul (1994). *An easy guide to factor analysis*. New York, NY: Routledge.

Kodytek, V. (2008). *On the replicability of the Biber model: The case of Czech*. Unpublished manuscript.

Kortmann, Bernd (Ed.) (2004). *Dialectology meets typology: dialect grammar from a cross-linguistic perspective*. Berlin; New York: Mouton de Gruyter.

Kress, Gunther (1987). Genre in a social theory of language: A reply to John Dixon. In I. Reid (Ed.), *The place of genre in learning: Current debates* (pp. 35-45). Geelong, Australia: Deakin University Press.

Kress, Gunther (1993). Genre as Social Process. In B. Cope & M. Kalantzis (Eds.), *The Powers of Literacy: A Genre Approach to Teaching Writing* (pp. 22-37). Pittsburgh: University of Pittsburgh Press.

Kuhn, Thomas (1970). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

Kytö, Merja & Rissanen, Matti (1992). A Language in Transition: The Helsinki Corpus of English Texts. *ICAME Journal* 16, 7-27.

Labov, William (1963). The social motivation of a sound change. *Word* 19, 273-303.

Labov, William (1965). On the mechanism of linguistic change. *Georgetown Monographs on Language and Linguistics* 18, 91-114.

Labov, William (1966). *The social stratification of English in New York City*. Washington DC: Center for Applied Linguistics.

Labov, William (1972a). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, William (1972b). Some features of the English of Black Americans. In R. W. Bailey & J. L. Robinson (Eds.), *Varieties of Present-Day English* (pp. 236-255). New York: MacMillan.

Lange, Claudia (2012). *The syntax of spoken Indian English*. VEAW G45, Amsterdam: John Benjamins.

Lareo, Inés (2006). *Colocaciones verbo-nombre en Inglés Moderno Tardío. Algunos aspectos de su comportamiento y evolución*. PhD. Dissertation. A Coruña: Universidade da Coruña.

Lareo, Inés (2008). Analysing a type of collocation. Make complex predicate in nineteenth-century science and fiction". *Revista Canaria de Estudios Ingleses* 57, 165-180.

Lareo, Inés (2009). El Coruña Corpus. Proceso de compilación y utilidades del Corpus of English Texts on Astronomy (CETA). Resultados preliminares sobre el uso de los predicados complejos en CETA. In P. Cantos Gómez & A. Sánchez Pérez (Eds.), *A Survey on Corpus-based Research* (pp. 267-280). Murcia: Asociación Española de Lingüística de Corpus.

Lareo, Inés (2011a). Uso de predicados complejos en los escritos de Astronomía del siglo XIX en lengua inglesa. Explotación del Corpus of English Texts on Astronomy. *Revista de Lenguas para Fines Específicos* 17, 229-252.

Lareo, Inés (2011b). Colocacións con *make*, *take* e *do* + nome nun corpus do século XIX de textos ingleses científicos e literarios escritos por mulleres. *Cadernos de Fraseoloxía Galega* 13, 191-214.

Lareo, Inés (2012). A corpus-driven approach to explore the use of complex predicates in 18th century English scientific writings. In I. Moskowich & B. Crespo, *Astronomy 'playne and simple': The Writing of Science between 1700 and 1900*. Amsterdam: John Benjamins.

Lareo, Inés & Esteve-Ramos, María José (2007). 18th century scientific writing. A study of make complex predicates in the *Coruña Corpus*. *ICAME* 32, 69–96.

Lareo, Inés & Moskowich, Isabel (2009). Make plus adjective in eighteenth-century science and fiction: Some issues made public. *English Studies* 90(3), 345-362.

Lareo, Inés & Montoya-Reyes, Ana (2007). Scientific writing: following Robert Boyle's principles in experimental essays – 1704 and 1998. *Revista Alicantina de Estudios Ingleses* 20, 119-137.

Lass, Roger (Ed.) (1999). *The Cambridge History of the English Language, Volume III: 1476-1776*. Volume three. Cambridge: Cambridge University Press.

Lee, David Y. W. (2000). *Modelling Variation in Spoken and Written Language: The Multi-Dimensional Approach Revisited*. Unpublished PhD dissertation. Lancaster University.

Lee, David Y. W. (2008). Corpora and discourse analysis: New ways of doing old things. In V. Bhatia, J. Flowerdew & R. H. Jones (Eds.), *Advances in Discourse Studies* (pp. 86-99). London: Routledge.

Leech, Geoffrey (1991). The state of the art in corpus linguistics. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics: Studies in honour of Jan Svartvik* (pp. 8-29). London: Longman.

Leech, Geoffrey (1992). Corpora and theories of linguistic performance. In J. Svartvik (Ed.), *Directions in corpus linguistics* (pp. 105-122). Berlin: Mouton de Gruyter.

Leech Geoffrey, Paul Rayson & Andrew Wilson (2001). *Word frequencies in written and spoken English*. London: Longman.

Lewin, Beverly (2005). Hedging: An Exploratory Study of Authors' and Readers' Identification of 'Toning Down' in Scientific Texts. *Journal of English for Academic Purposes*, *4*, 163- 178.

Lightman, Bernard (1997). The voices of nature: Popularizing Victorian science. In B. Lightman (Ed.), *Victorian Science in Context* (pp. 187-211). Chicago/London: University of Chicago Press.

Longacre, Robert (1976). *An anatomy of speech notions*. Lisse: Peter de Ridder Press.

Love, Alison (2002). Introductory concepts and 'cutting edge' theories: Can the genre of the textbook accommodate both? In J. Flowerdew (Ed.), *Academic Discourse* (pp. 76-92). Harlow: Longman.

Lovejoy, James (1995). Prepositions in British and American English – a computer-aided corpus study. *Arbeiten aus Anglistik und Amerikanistik* 20, 55-74.

Luhn, Hans P. (1960). Key word-in-context index for technical literature (kwic index). *Information Science and Technology* 11(4), 288-295.

Magner, Lois N. (2002). *A history of the Life Sciences, revised and expanded.* New York; Basel: Marcel Dekker.

Mair, Christian (1995). Changing patterns of complementation, and concomitant grammaticalisation, of the verb help in present-day British English. In B. Aarts & C. F. Meyer (Eds.), *The Verb in Contemporary English: Theory and Description* (pp. 258-272). Cambridge: CUP.

Mair, Christian (1997). Parallel corpora: a real-time approach to language change in progress. In M. Ljung (Ed.), *Corpus-Based Studies in English: Papers from the Seventeenth International Conference on English-Language Research Based on Computerized Corpora (ICAME 17)* (pp. 195-209). Amsterdam: Rodopi.

Mair, Christian (2002). Short term diachronic shifts in part-of-speech frequencies: A comparison of the tagged LOB and F-LOB corpora. *International Journal of Corpus Linguistics* 7(2), 245-264.

Marckworth, Mary L. & William J. Baker (1974). A discriminant function analysis of co-variation of a number of syntactic devices in five prose genres. *American Journal of Computational Linguistics*, *Microfiche 11.*

Martí, M. Antònia & Castellón, Irene (2000): *Lingüística computacional*. Barcelona: Universitat de Barcelona.

Martin, James R. (1985). *Factual writing: exploring and challenging social reality*. Geelong: Deakin University Press.

Martin, James R. (1997). Analysing genre: functional parameters. In F. Christie & J. R. Martin (Eds.), *Genre and institutions: social processes in the workplace and school* (pp. 3-39). London; New York: Continuum.

Martin, James R. (2001). Language, register and genre. In A. Burns & C. Coffin (Eds.), *Analysing English in a global context* (pp. 149-166). London: Routledge.

Martin, James R., Frances Christie & Joan Rothery (1987). Social processes in education: A reply to Sawyer and Watson (and others). In I. Reid (Ed.), *The* *place of genre in learning: Current debates* (pp. 35-45). Geelong, Deakin University Press, Australia.

Matthews, David (2000). *The Invention of Middle English: An Anthology of Primary Sources*. University Park, PA: The Pennsylvania State University Press.

Matthiessen, Christian M. (1993). Register in the round: diversity in a unified theory of register analysis. In M. Ghadessy (Ed.), *Register analysis: theory and practice* (pp. 221-292). London; New York: Pinter Publishers.

McCormmach, Russel (2004). *Speculative truth: Henry Cavendish, natural philosophy, and the rise of modern theoretical science*. Oxford: Oxford University Press.

McEnery, Tony (2003). Corpus linguistics. In R. Mitkov (Ed.), *The Oxford handbook of computational linguistics* (pp. 448-463). Oxford: Oxford University Press.

McEnery, Tony & Hardie, Andrew (2011). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

McEnery, Tony & Wilson, Andrew (1996). *Corpus linguistics: An introduction*. Edinburgh: Edinburgh University Press.

Mele-Marrero, Margarita & Martín-Díaz, M. Auxiliadora (2001). Formación y desarrollo del inglés moderno. In I. De la Cruz Cabanillas & F. J. Martín Arista (Eds.), *Lingüística histórica inglesa* (pp. 573-596). Barcelona: Ariel.

Miller, Carolyn R. (1984). Genre as social action. *Quarterly Journal of Speech* 70, 151-167.

Millward, Celia M. (1989). *A Biography of the English Language*. New York: Holt, Rinehart and Winston.

Milroy, James (1980). *Language and social networks*. Oxford: Blackwell.

Milroy, James (1992). *Linguistic variation and change: On the historical sociolinguistics of English*. (Language in Society 19). Oxford: Basil Blackwell.

Milroy, James (1997). External motivations for linguistic change. *Multilingua* 16, 311-323.

Moessner, Lilo (2009). The influence of the Royal Society on 17th-century scientific writing. *ICAME Journal* 33, 65-87.

Mohamed, Ghada (2011). *Text classification in the BNC using corpus and statistical methods*. Unpublished PhD dissertation. Lancaster University.

Montgomery, Scott (1996). *The scientific voice*. New York; London: The Guilford Press.

Moon, Rosamund (1998). *Fixed expressions and idioms in English: a corpus-based approach*. Oxford: Clarendon Press.

Moskowich, Isabel (1995). *Los escandinavos en Inglaterra y el cambio léxico en ingles medieval*. A Coruña: Universidade da Coruña.

Moskowich, Isabel (2001). Morfología flexiva del inglés moderno. In I. De la Cruz Cabanillas & F. J. Martín Arista (Eds.), *Lingüística histórica inglesa* (pp. 624-654). Barcelona: Ariel.

Moskowich, Isabel (2011). "The Golden Rule of Divine Philosophy" exemplified in the Coruña Corpus of English Scientific Writing. *Lenguas para fines específicos* 17, 167-198.

Moskowich, Isabel (2012). *Language contact and vocabulary enrichment: Scandinavian elements in Middle English*. Studies in English Medieval Language and Literature, 34. Bern: Peter Lang.

Moskowich, Isabel (2013). Eighteenth-century female authors: Women and science in the *Coruña Corpus of English Scientific Writing*. *Australian Journal of Linguistics* 33(4), 467-487.

Moskowich, Isabel (2016a). Philosophers and scientists from the Modern Age: Compiling the *Corpus of English Philosophy Texts*. In I. Moskowich, G. Camiña, I. Lareo & B. Crespo (Eds.), *The Conditioned and the Unconditioned': Late Modern English Texts on Philosophy* (pp. 1-23). Amsterdam: John Benjamins.

Moskowich, Isabel (2016b). When sex talks. Evidence from the Coruña Corpus of English Scientific Writing. In F. Alonso-Almeida, L. Cruz García, and V. González Ruiz (Eds.). Corpus-based Studies on Language Varieties (pp. 233-248). Bern: Peter Lang.

Moskowich, Isabel & Crespo, Begoña (2007). Presenting the Coruña Corpus: A Collection of Samples for the Historical Study of English Scientific Writing. In J. Pérez Guerra. D. González-Álvarez, J. L. Bueno-Alonso & E. Rama-Martínez (Eds.), *Of Varying Language and Opposing Creed: New Insights into Late Modern English* (pp. 341-357). Bern: Peter Lang.

Moskowich, Isabel & Crespo, Begoña (2012) (Eds.). *Astronomy 'playne and simple': The Writing of Science between 1700 and 1900*. Amsterdam: John Benjamins.

Moskowich, Isabel & Monaco, Leida Maria (2014). Abstraction as a Means of Expressing Reality: Women Writing Science in Late Modern English. In M. Gotti & D. S. Giannoni (Eds.), *Corpus Analysis for Descriptive and Pedagogical Purposes* (pp. 203-224). Bern: Peter Lang.

Moskowich, Isabel & Parapar, Javier (2008). Writing science, compiling science: The Coruña Corpus of English Scientific Writing. In M. J. Lorenzo Modia (Ed.), *Proceedings from the 31st AEDEAN Conference* (pp. 531-544). A Coruña: Universidade da Coruña.

Moskowich, Isabel, Gonzalo Camiña, Inés Lareo & Begoña Crespo (Eds.) (2016). *The Conditioned and the Unconditioned': Late Modern English Texts on Philosophy*. Amsterdam: John Benjamins.

Myers, Greg (1989). The pragmatics of politeness in scientific articles. *Applied Linguistics* 10(1), 1-35.

Nelson, Gerald (2012a). *International Corpus of* English Markup manual for spoken texts (1-18).

Nelson, Gerald (2012b). *International Corpus of English* Markup manual for written texts (1-14).

Neumann, Stella (2013). *Contrastive register variation: A quantitative approach to the comparison of English and German*. Berlin; Boston: Walter de Gruyter.

Nevalainen, Terttu & Raumolin-Brunberg, Helena (1994). Sociolinguistics and language history: The Helsinki Corpus of Early English Correspondence. *Hermes, Journal of Linguistics* 12, 135-143.

Nevalainen, Terttu & Raumolin-Brunberg, Helena (1995). Constraints in politeness: the pragmatics of address formulae in early English correspondence. In A. Jucker (Ed.), *Historical pragmatics: Pragmatic developments in the history of English* (pp. 541-601). Amsterdam: John Benjamins.

Norušis, Marija J. (1988). *SPSS advanced statistics user's guide*. Chicago: SPSS Inc.

Norušis, Marija J. (2005). *SPSS 14.0 advanced statistical procedures companion*. Upper Saddle River, NJ: Prentice Hall.

Nunnally, Jum C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw Hill.

Ochs, Elinor (1979). Planned and unplanned discourse. In T. Givón (Ed.), *Discourse and syntax* (pp. 51-80). New York: Academic Press.

Oster, Sandra (1981). The use of tenses in reporting past literature. In L. Selinker, E. Tarone & V. Hanzeli (Eds.), *English for academic and technical purposes: Studies in honor of Louis Trimble* (pp. 76-90). Rowley, MA: Newburg House.

Paltridge, Brian (1996). Genre, text type, and the language learning classroom. *ELT Journal* 50(3), 237-243.

Paltridge, Brian (1997). *Genre, frames and writing in research settings*. Amsterdam: John Benjamins.

Pahta, Päivi (2001). Creating a new genre: Contextual dimensions in the production and transmission of early scientific writing. *European Journal of English Studies*, *5*(2), 205-220.

Pahta, Päivi (2003). On structures of code-switching in medical texts from medieval England. *Neuphilologische Mitteilungen* 104, 197-210.

Pahta, Päivi (2004). Code-switching in medieval medical writing. In I. Taavitsainen, & P. Pahta (Eds.), *Medical and Scientific Writing in late Medieval English*. Cambridge: Cambridge University Press.

Park, Katharine & Lorraine Daston (2006). Introduction: The Age of the New. In R. Porter, K. Park & L. Daston (Eds.), *The Cambridge History of Science: Volume 3, Early Modern Science* (pp. 1-17). Cambridge: Cambridge University Press.

Parapar, Javier & Moskowich, Isabel (2007). The Coruña Corpus Tool. *Revista del Procesamiento del Lenguaje Natural* 39, 289-290.

Parodi, Giovanni (2007). Variation across registers in Spanish. In G. Parodi (Ed.), *Working with Spanish corpora* (pp. 11-53). London: Continuum.

Pledge, Humphrey T. (1959). *Science since 1500: A short history of mathematics, physics, chemistry, and biology*. New York: Harper.

Porter, Roy, Katharine Park & Lorraine Daston (Eds.) (2006), *The Cambridge History of Science: Volume 3, Early Modern Science*. Cambridge: Cambridge University Press.

Prince, Michael (1996). *Philosophical dialogue in the British Enlightenment: theology, aesthetics, and the novel*. Cambridge: Cambridge University Press.

Puente-Castelo, Luis (2016a). Conditional constructions and their uses in eighteenth-century philosophy and life sciences texts. In F. Alonso-Almeida, I. Ortega Barrera, E. Quintana Toledo & M. Sánchez Cuervo (Eds.), *Input a Word,* *Analyse the World: Selected Approaches to Corpus Linguistics* (pp. 241-255). Newcastle upon Tyne: Cambridge Scholars Publishing.

Puente-Castelo, Luis (2916b). Explaining the use of *If...then...* structures in *CEPhiT*. In I. Moskowich, G. Camiña, I. Lareo & B. Crespo (Eds.) *'The Conditioned and the Unconditioned': Late Modern English Texts on Philosophy* (pp. 167-181). Amsterdam: John Benjamins.

Puente-Castelo, Luis (Forthcoming). *On conditionality: A corpus-based study of conditional structures in Late Modern English scientific texts*. PhD dissertation. University of A Coruña.

Purvis, T. M. (2008). *A linguistic and discursive analysis of register variation in Dagbani*. Unpublished PhD dissertation. Northern Arizona University, Flagstaff, AZ.

Quirk, Randolph, Sydney Greenbaum, Geoffrey Leech & Jan Svartvik (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Reppen, Randi (1994). *Variation in elementary student writing*. Unpublished PhD Dissertation. Indiana University, Bloomington, IN.

Reppen, Randi (2001). Register variation in student and adult speech. In S. Conrad and D. Biber (Eds.), *Variation in English: Multi-Dimensional Studies* (pp. 187-199). Harlow: Longman.

Romaine, Suzanne (2000). *Language in society: an introduction to sociolinguistics*. Oxford: Oxford University Press.

Roy, Louis (1998). Medieval Latin scholasticism: some comparative features. In J. I. Cabezón (Ed.), *Scholasticism: cross-cultural and comparative perspectives* (19-34). Albany: State University of New York Press.

Ruse, Michael (2008). *The evolution wars: A guide to the debates*. Millerton, NY: Grey House Publishing.

Salager-Meyer, Françoise (1994). Hedges and textual communicative function in medical English written discourse. *English for Specific Purposes* 13(2), 149-170.

Saldanha, Gabriela (2009). Principles of corpus linguistics and their application to translation studies research. *Tradumática* 7. Available at: http://www.raco.cat/index.php/Tradumatica/article/view/154828/206722.

Samraj, Betty (2002). Disciplinary variation in abstracts: the case of wildlife behavior and conservation biology. In J. Flowerdew (Ed.), *Academic Discourse* (pp. 40-56). Harlow: Longman.

Sánchez-Barreiro, Estefanía (2010a). Los elementos de prolongación copulativos en textos científicos ingleses del siglo XVIII. In I. Moskowich B. Crespo, I. Lareo & P. Lojo (Eds.), *Language windowing through corpora. Visualización del lenguaje a través de corpus* (pp. 801-815). A Coruña: Universidade da Coruña.

Sánchez-Barreiro, Estefanía (2010b). Adjunctive and Disjunctive Lists in the Modern English Scientific Discourse. In M.LL. Gea-Valor, I. García, & M. J. Esteve (Eds), *Linguistic and Translation Studies in Scientific Comunications*. Vol. 86 (177-193). Bern: Peter Lang.

Sánchez-Barreiro, Estefanía (2010). Los elementos de prolongación de listas enumerativas en textos científicos ingleses del siglo XVIII. Unpublished MA dissertation, University of A Coruña.

Sánchez-Barreiro, Estefanía (Forthcoming). Los extenders como recurso estilístico en el discurso científico inglés del siglo XVIII. PhD dissertation. University of A Coruña.

Schiebinger, Londa (1989). The history and philosophy of women in science: A review essay. *Signs* 12(2), 305-332.

Schiebinger, Londa (2003). The philosopher's beard: Women and gender in science. In R. Porter (Ed.), *The Cambridge History of Science: Volume 4, Eighteenth-Century Science* (pp. 184-210). Cambridge: Cambridge University Press.

Schiffrin, Deborah, Deborah Tannen & Hamilton, Heidi E. (Eds.) (2001). *Handbook of discourse analysis*. Oxford: Blackwell.

Scott, Mike (1999). *WordSmith Tools*. Oxford, UK: Oxford University Press.

Serafini, Anthony (1993). *The epic history of biology*. Cambridge, MA: Perseus Publishing.

Shapin, Steven (1988). The house of experiment in seventeenth-century England. *Isis* 79(3), 373-404.

Shapin, Steven (1996). *The Scientific Revolution*. Chicago, IL: University of Chicago Press.

Shteir, Ann B. (1987). Botany in the breakfast room: Women and early nineteenth-century British plant study. In P. Abir-Am & D. Outram (Eds.), *Uneasy* *careers and intimate lives: Women in science (1789-1979)* (pp. 31-44). New Brunswick, NJ: Rutgers University Press.

Shteir, Ann B. (2008). Elegant recreations? Configuring science-writing for women. In B. Lightman (Ed.), *Victorian science in context* (pp. 236-255). Chicago: University of Chicago Press.

Sinclair, John (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, John (1994). Corpus typology: A framework for classification. EAGLES document (1-18), now published as Sinclair, John (1995). In G. Melchers & B. Warren (Eds.), *Studies in Anglistics*. Stockholm: Almqvist and Wiksell, International.

Slack, Nancy (1987). Nineteenth-century American women botanists: Wives, widows, and work. In P. Abir-Am & D. Outram (Eds.), *Uneasy careers and intimate lives: Women in science (1789-1979)* (pp. 77-103). New Brunswick, NJ: Rutgers University Press.

Smith, Jeremy (1996). *An historical study of English*. London: Routledge.

Snedecor, George W. & Cochran, William G. (1989) *Statistical methods* (8th edition). Ames: Iowa State University Press.

Sollaci, Luciana B. & Pereira, Maurizio G. (2004). The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *Journal of Medical Library Association* 92(3), 364-371.

Sperberg-McQueen, C. M. & Burnard, Lou (1994) (Eds.). *Guidelines for electronic text encoding and interchange*. Chicago, Oxford: Text Encoding Initiative.

Sprat, Thomas (1667). *The History of the Royal Society of London for the Improving of Natural Knowledge*. London: Martyn & Allestry.

Stafleu, Frans A. (1971). *Linnaeus and the Linnaeans: the spreading of their ideas in systematic botany, 1735-1789*. Utrecht: International Association for Plant Taxonomy.

Steiger, James H., & Schonemann, Peter H. (1978). A history of factor indeterminacy. In S. Shye (Ed.), *Theory construction and data analysis* (pp. 136-178). Chicago: University of Chicago Press.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Hillsdale, NS: Erlbaum.

Strenström, Anna-Britta (1986). What does *really* really do? Strategies in speech and writing. In G. Tottie & I. Bäcklund (Eds.), *English in speech and writing: A symposium* (pp. 149-163). Stockholm: Almqvist and Wiksell.

Stubbs, Michael (1983). *Discourse analysis: the sociolinguistic analysis of natural language*. Oxford: Blackwell.

Svartvik, Jan & Quirk, Randolph (Eds). 1980. *A Corpus of English Conversation*. Lund: CWK Gleerup.

Swales, John (1974). Notes on the function of attributive -en particles in scientific discourse. *Paper in English for Special Purposes* 1. Khartoum: University of Khartoum.

Swales, John (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Swales, John (2004). *Research genres: Exploration and applications*. Cambridge: Cambridge University Press.

Swales, John & Najjar, Hazem (1987). The writing of research article introductions. *Written Communication* 2(4), 175-191.

Sweet, Henry (1873-1874). The History of English Sounds. *Transactions of the Philological Society*, 461-623.

Taavitsainen, Irma (2001). Changing conventions of writing: the dynamics of genres, text types, and text traditions. *European Journal of English Studies* 5(2), 139-150.

Taavitsainen, Irma (2005). On corpus linguistics: Computers and the history of English. In I. Moskowich & B. Crespo (Eds.), *Re-interpretations of English (II)* (pp. 325-345). A Coruña: Universidade da Coruña.

Taavitsainen, Irma & Päivi Pahta (Eds.) (2004). *Medical and scientific writing in Late Medieval English*. Cambridge: Cambridge University Press.

Taavitsainen, Irma & Päivi Pahta (2010). *Early modern English medical writing: Corpus description and studies*. Amsterdam: John Benjamins.

Taavitsainen, Irma, Päivi Pahta, Turo Hiltunen, Martti Mäkinen, Ville Marttila, Maura Ratia, Carla Suhr & Jukka Tyrkkö (Eds.) (2010). *Early Modern English Medical Texts*. CD-ROM. Amsterdam: John Benjamins.

Tabachnick, Barbara & Fidell, Linda (1996). *Using multivariate statistics*. New York: HarperCollins.

Tabachnick, Barbara & Fidell, Linda (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson.

Takahashi, K. (2006). *Typology of registers in the British National Corpus: Multi-feature and multi-dimensional analyses*. Unpublished PhD dissertation. University of Lancaster.

Thompson, Sandra A. (1983). Grammar and discourse: The English detached participial clause. In F. Klein-Andreu (Ed.), *Discourse perspectives on syntax* (pp. 43-65). New York: Academic Press.

Thompson, Sandra A. & Robert E. Longacre (1985). Adverbial clauses. In T. Shoppen (Ed.), *Language typology and syntactic description* Vol. 2 (pp. 171-233). Cambridge: Cambridge University Press.

Thomson, Godfrey H. (1951). *The factorial analysis of human ability*. London: University of London Press.

Tieken-Boon van Ostade, Ingrid (2010). Eighteenth-century women and their norms of correctness. In R. Hickey (Ed.), *Eighteenth-century English: Ideology and change* (pp. 59-72). Cambridge: Cambridge University Press.

Tognini-Bonelli, Elena (2001). *Corpus linguistics at work*. Amsterdam: John Benjamins.

Tribble, Christopher (1999). *Writing difficult texts*. Unpublished PhD dissertation. Lancaster University.

Trudgill, Peter (1972). Sex, covert prestige and linguistic change in the urban British English of Norwich. *Language in Society* 1, 179-195.

Trudgill, Peter (1978). *Sociolinguistic patterns in British English*. London: Edward Arnold.

Trudgill, Peter (2000). *Sociolinguistics: an introduction to language and society*. London: Penguin.

Trudgill, Peter (2002). *Sociolinguistic variation and change*. Edinburgh: Edinburgh University Press.

Ulijn, Jan (1989). The scientific and technical register and its cross-linguistic constants and variants. *Acta Universitatis Wratislaviensis* 1130, 183-231.

UNESCO (1988). Proposed International Standard Nomenclature for Fields of Science and Technology UNESCO/NS/ROU/257. Paris: United Nations Educational Scientific and Cultural Organization.

Ure, Jean (1982). Introduction: approaches to the study of register range. *International Journal of the Sociology of Language* 35, 5-23.

Valle, Ellen (1996). A scientific community and its texts: A historical discourse study. In B. Gunnarsso, P. Linell & B. Nordberg (Eds.), *The construction of professional discourse* (pp. 76-98). London: Longman.

Veirano Pinto, Marcia (2014). Dimensions of variation in North American movies. In T. Berber Sardinha, & M. Veirano Pinto (Eds.), *Multi-dimensional analysis, 25 years on: a tribute to Douglas Biber* (pp. 109-147). Amsterdam: John Benjamins.

Weinreich, Uriel, William Labov & Marvin Herzog (1968). Empirical foundations for a theory of language change. In W. Lehmann & Y. Malkiel (Eds.), *Directions for historical linguistics* (pp. 95-195). Austin: University of Texas Press.

Werlich, Egon (1982). *A text grammar of English*. Heidelberg: Quelle und Meyer.

White, M. (1994). *Language in job interviews: Differences relating to success and socioeconomic variables*. Unpublished PhD dissertation. Northern Arizona University, Flagstaff, AZ.

Wilkins, John (1668). *An Essay towards a Real Character, and a Philosophical Language*. London: Samuel Gellibrand and John Martin.

Wyld, Henry C. (1920). *A History of Modern Colloquial English*. Oxford: Basil Blackwell.

Xiao, Richard (2009). Multidimensional analysis and the study of World Englishes. *Word Englishes,* 28(4), 421-450.

Xiao, Zhonghua & Tony McEnery (2005). Two approaches to genre analysis: Three genres in modern American English. *Journal of English Linguistics* 33(1), 62-82.

Zerbe, Michael J. (2007). *Composition and the rhetoric of science: Engaging the dominant discourse*. Carbondale: Southern Illinois University Press.

Appendices

Appendix I

Key to linguistic features included in the analysis

Linguistic feature	Short name
1. Past tense	PAST
2. Perfect aspect	PERF
3. Present tense	PRES
4. Place adverbials	PL_ADV
5. Time adverbials	TIM_ADV
6. First person pronouns	FPERS
7. Second person pronouns	SPERS
8. Third person pronouns (excluding *it*)	TPERS
9. Pronoun *it*	ITPRO
10. Demonstrative pronouns	DEMPRO
11. Indefinite pronouns	INDPRO
12. Pro-verb *do*	PRO_DO
13. Questions	QUEST
14. Nominalisations	NOM
15. Total other nouns	NOUN
16. Agentless passives	AGPASS
17. *by* passives	BYPASS
18. *be* as main verb	BE_MAIN
19. Existential *there*	EXTHERE
20. *that* verb complements	THAT_V
21. *that* adjective complements	THAT_ADJ
22. WH clauses in subject position	WHCL_SUB
23. WH clauses in object position	WHCL_OB
24. *to* infinitives	TO_INF
25. Detached past participial clauses with adverbial function	PASTPART
26. Past participial WHIZ deletion relatives	WHIZ
27. Present participial WHIZ deletion relatives	PRES_WHIZ
28. *that* relativiser in subject function	THAT_SUB
29. WH relativiser in subject function	WHREL_SUB

Linguistic feature	Short name
30. WH relativiser in object function	WHREL_OB
31. Pied-piping relative clauses	PIP
32. Sentence relatives	SREL
33. Causative adverbial subordinators	CAUSADV
34. Concessive adverbial subordinators	CONCADV
35. Conditional adverbial subordinators	CONDATV
36. Other adverbial subordinators	OTHADV
37. Total prepositional phrases	PREP
38. Attributive adjectives	ATTRADJ
39. Predicative adjectives	PREDADJ
40. Total other adverbs	ADV
41. Conjuncts	CONJ
42. Downtoners	DOWN
43. Hedges	HEDG
44. Amplifiers	AMPL
45. Demonstratives	DEM
46. Possibility modals	POSSMOD
47. Necessity modals	NECMOD
48. Predictive modals	PREDMOD
49. Public verbs	PUBV
50. Private verbs	PRIVV
51. Suasive verbs	SUASV
52. *seem* and *appear*	SEEM
53. Split infinitives	SPLITINF
54. Split auxiliaries	SPLITAUX
55. Phrasal coordination	PHCOORD
56. Clausal coordination	CLCOORD
57. Synthetic negation	SNEG
58. Analytic negation	ANEG

Linguistic features and their short names (continued)

Appendix II

Descriptive statistics for the dimension scores of all subregisters

Dimension	Mean	Minimum value	Maximum value	Range	Standard deviation
Astronomy (18th century); N=21					
D1	–0.6	–2.4	1.4	3.8	0.8
D2	1.2	–0.2	4.7	4.9	1.1
D3	–0.5	–3.2	1.3	4.5	1.0
D4	0.3	–1.0	2.9	3.9	1.0
Astronomy (19th century); N=21					
D1	–0.5	–1.5	1.1	2.6	0.7
D2	0.4	–0.7	1.9	2.6	0.7
D3	0.6	–1.0	1.6	2.6	0.8
D4	0.6	–1.0	2.3	3.3	0.9
Philosophy (18th century); N=20					
D1	1.1	–0.5	2.5	3.0	0.8
D2	0.1	–1.6	1.3	2.9	0.8
D3	0.1	–2.0	1.7	3.7	1.1
D4	–0.9	–3.4	0.8	4.2	1.0
Philosophy (19th century); N=20					
D1	0.9	–0.5	3.4	3.9	0.9
D2	–0.3	–1.2	0.5	1.7	0.5
D3	1.0	–1.1	2.0	3.1	0.8
D4	–0.2	–1.5	1.0	2.5	0.7
Life Sciences (18th century); N=20					
D1	–0.5	–1.4	0.1	1.5	0.4
D2	–0.7	–2.5	1.4	3.9	1.0
D3	–1.1	–2.9	0.1	3.0	0.8
D4	–0.5	–4.3	0.9	5.2	1.1
Life Sciences (19th century); N=20					
D1	–0.4	–1.3	1.1	2.4	0.6
D2	–0.8	–2.1	0.4	2.5	0.7
D3	–0.1	–2.1	1.0	3.1	0.8
D4	0.6	–1.7	2.6	4.3	1.0
Treatise (18th century); N=34					
D1	0.0	-1.4	2.1	3.5	1.0
D2	-0.1	-2.5	2.6	5.1	1.2
D3	-0.5	-2.8	1.6	4.4	1.0
D4	-0.4	-4.3	0.9	5.2	1.0
Treatise (19th century); N=27					
D1	0.0	-1.2	3.4	4.6	1.0
D2	-0.4	-2.1	1.9	4.0	0.9
D3	0.5	-1.1	2.0	3.1	0.9
D4	0.5	-1.0	2.6	3.6	0.9
Textbook (18th century); N=12					
D1	-0.8	-1.4	0.0	1.4	0.5
D2	0.9	-1.3	2.2	3.5	1.0
D3	-0.6	-2.7	1.7	4.4	1.1
D4	-0.6	-3.4	1.1	4.5	1.1
Textbook (19th century); N=8					
D1	-1.1	-1.5	-0.4	1.1	0.4
D2	0.1	-1.7	1.4	3.1	1.1
D3	0.5	-0.8	1.2	2.0	0.6
D4	-0.1	-1.7	1.3	3.0	1.0

Dimension	Mean	Minimum value	Maximum value	Range	Standard deviation
Essay (18th century); N=9					
D1	0.7	-2.4	2.5	4.9	1.5
D2	0.7	-0.7	4.7	5.4	1.6
D3	-0.1	-1.6	1.3	2.9	1.0
D4	-0.7	-1.7	0.6	2.3	0.9
Essay (19th century); N=5					
D1	1.3	0.8	2.3	3.1	0.6
D2	-0.3	-1.0	0.5	1.5	0.6
D3	1.1	0.2	1.7	1.9	0.6
D4	0.1	-1.3	0.8	2.1	0.9
Lecture (18th century); N=1					
D1	0.3	0.3	0.3	-	-
D2	–0.1	–0.1	–0.1	-	-
D3	–0.7	–0.7	–0.7	-	-
D4	1.2	1.2	1.2	-	-
Lecture (19th century); N=11					
D1	0.1	-1.0	1.6	2.6	0.9
D2	-0.1	-0.8	1.8	2.6	0.7
D3	0.3	-1.3	1.8	3.1	0.9
D4	0.1	-0.9	1.5	2.4	0.8
Article (18th century); N=1					
D1	0.4	0.4	0.4	-	-
D2	0.1	0.1	0.1	-	-
D3	–0.4	–0.4	–0.4	-	-
D4	2.9	2.9	2.9	-	-
Article (19th century); N=6					
D1	0.2	-0.4	0.9	1.3	0.5
D2	-0.4	-0.7	0.0	0.7	0.3
D3	0.8	-0.5	1.6	2.1	0.9
D4	0.8	-1.2	2.3	3.5	1.2
Letter (18th century); N=2					
D1	–0.3	–0.5	0.0	0.5	0.3
D2	–0.8	–1.4	–0.2	1.2	0.8
D3	–0.4	–0.6	–0.1	0.5	0.4
D4	0.5	–0.1	1.0	1.1	0.8
Letter (19th century); N=3					
D1	-0.2	-0.5	0.4	0.9	0.5
D2	-0.1	-0.6	0.4	1.0	0.5
D3	-0.7	-2.1	0.5	2.6	1.3
D4	0.9	-0.7	1.9	2.6	1.4
Dialogue (18th century); N=1					
D1	1.4	1.4	1.4	-	-
D2	1.3	1.3	1.3	-	-
D3	–3.2	–3.2	–3.2	-	-
D4	2.6	2.6	2.6	-	-
Dialogue (19th century); N=1					
D1	0.6	0.6	0.6	-	-
D2	0.5	0.5	0.5	-	-
D3	0.6	0.6	0.6	-	-
D4	–1.5	–1.5	–1.5	-	-
Dictionary (18th century); N=1					
D1	–0.3	–0.3	–0.3	-	-
D2	0.4	0.4	0.4	-	-
D3	–0.4	–0.4	–0.4	-	-
D4	–0.7	–0.7	–0.7	-	-

Descriptive statistics (continued)

Appendix III

Factor scores* per text

Life Sciences

Text ID	TEXT	GENRE	FACTOR1	FACTOR2	FACTOR3	FACTOR4
life1	1707 Douglas	Treatise	-1,383308122	-0,81217221	-1,040200204	-0,204847139
life2	1707 Sloane	Treatise	-0,239920913	-1,274011613	-1,465916654	0,131589093
life3	1717 Keill	Essay	0,060323734	1,417248887	0,011299423	0,440384874
life4	1720 Gibson	Treatise	-0,642073691	-0,531364639	-0,83361555	-0,828673524
life5	1726 Blair	Treatise	-0,074521086	-1,00573823	-1,094213464	-0,091133482
life6	1730 Boreman	Textbook	-0,231488417	-1,302117707	-2,68068435	-1,477098312
life7	1737 Blackwell	Treatise	-1,319158064	-1,491572934	-0,943867122	-4,276033639
life8	1737 Brickell	Treatise	0,145873799	0,016883588	-2,844328085	-0,522776014
life9	1743 Edwards	Treatise	-0,683456078	-1,457075178	-1,324162575	-1,132841942
life10	1750 Hughes	Treatise	-0,443203443	-1,694593073	-1,285212386	0,091204428
life11	1752 Dodd	Essay	-0,411280985	-0,037033039	-1,563415091	-0,443877977
life12	1758 Borlase	Treatise	-0,641851521	-0,471431528	-0,913732152	-0,371916961
life13	1766 Pennant	Treatise	-0,592219293	-1,060798879	-0,397325498	-0,254691034
life14	1769 Bancroft	Letter	-0,51205786	-1,407978435	-0,646817558	-0,058420722
life15	1774 Goldsmith	Treatise	-0,275237368	-0,548924767	-1,149804503	0,139514456
life16	1776 Withering	Treatise	-0,191460539	0,112284881	-0,60006993	-0,361315873
life17	1786 Speechly	Treatise	-0,033494721	0,962521026	0,004561485	0,910247749
life18	1789 Bolton	Treatise	-1,267598167	-1,648183067	-1,159787783	-0,872833675
life19	1794 Donovan	Treatise	-0,496051783	0,840692944	0,141331884	0,098502877
life20	1795 Smith	Treatise	-0,69731284	-2,52271168	-1,311535185	0,003165707
life21	1804 Jacson	Lecture	-0,462971885	-0,154245101	-0,277253291	-0,771203445
life22	1808 Wilson	Treatise	-0,582744006	-1,431098767	-0,885431579	1,339605033
life23	1816 Wakefield	Letter	-0,494015164	-0,577488379	-0,601751676	-0,698191853
life24	1819 Lawrence	Lecture	-0,654749323	-0,799791154	0,337980169	0,012726811
life25	1824 Jenner	Article	0,253119822	-0,619179591	-0,163192435	1,52622229
life26	1828 Godman	Treatise	-1,018300732	-1,289995137	-0,578182542	0,827781684
life27	1832 Lincoln	Lecture	-0,827010541	-0,379664851	0,018247129	-0,027441579
life28	1835 Jardine	Treatise	-0,859537176	-1,418087538	0,048466354	0,546159383
life29	1840 Pratt	Treatise	-0,575825734	-1,160575502	-0,577791378	0,428951015
life30	1848 Dalyell	Treatise	-0,327452789	-1,616346257	0,443154981	1,546125808
life31	1859 Agassiz	Letter	0,356296118	0,393279293	-2,070425742	1,585940173
life32	1859 Darwin	Treatise	0,454757064	-0,790633622	1,034979218	2,628443654
life33	1863 Huxley	Lecture	1,092009183	0,124874793	-1,262970783	1,462286497
life34	1867 Spencer	Treatise	-0,102826715	-0,480812686	0,355686155	0,188005676
life35	1876 Macalister	Textbook	-1,411538672	-1,02334055	0,963784152	-1,682871654
life36	1879 Lankester	Treatise	-0,515028655	-2,144458368	-0,239972367	0,077787158
life37	1880 Balfour	Treatise	-0,700051596	-1,055426255	0,413818579	0,800804796
life38	1889 Galton	Treatise	0,197439312	0,060664624	1,039314805	0,896580408
life39	1893 Marshall	Textbook	-1,340605064	-1,73450535	0,212243597	1,092268772
life40	1898 Packard	Textbook	-1,121481425	-0,552441053	0,15693262	-0,144435368

* Bartlett method

Philosophy

Text ID	TEXT	GENRE	FACTOR1	FACTOR2	FACTOR3	FACTOR4
phil1	1700 Astell	Essay	2,505417456	0,248341575	-1,636038013	-1,571833527
phil2	1705 Cheyne	Treatise	1,643380283	0,42349976	-0,374858602	0,841348745
phil3	1710 Dunton	Treatise	1,297195687	0,347503111	-1,483241186	-0,093525106
phil4	1717 Collins	Treatise	1,732199092	1,280730047	-0,133500637	-1,95461659
phil5	1727 Greene	Treatise	0,443628121	0,781906955	-0,784491249	-0,945341391
phil6	1730 Kirkpatrick	Treatise	2,107387225	1,333362225	-1,98609996	0,340172259
phil7	1733 Balguy	Essay	2,268041306	-0,006676149	0,248502735	-1,611876974
phil8	1736 Butler	Treatise	1,821996503	1,005696382	-0,219872383	0,190341837
phil9	1740 Turnbull	Treatise	1,350305398	0,059779964	0,899380797	-1,746120076
phil10	1748 Hume	Essay	0,913421174	-0,44235576	0,283771764	-1,068568819
phil11	1754 Bolingbroke	Essay	1,627439901	0,823765803	-0,706449972	-0,146776962
phil12	1755 Hutcheson	Treatise	0,290664572	-0,49398743	1,565509034	-0,556340005
phil13	1764 Reid	Treatise	1,938394929	0,034546709	-0,749275934	-0,637888333
phil14	1769 Ferguson	Textbook	-0,003218162	-0,399726918	1,699442477	-3,447755642
phil15	1770 Burke	Treatise	0,518184675	-0,373273058	0,656168923	-0,056944007
phil16	1776 Cambell	Essay	0,597270091	-0,723999496	0,715846219	-0,623442258
phil17	1783 Macaulay	Treatise	-0,05998984	-1,590986962	1,461325273	-1,255007867
phil18	1790 Smellie	Treatise	-0,498997735	-1,214609713	0,077931305	-1,283023734
phil19	1792 Wollstonecraft	Treatise	1,107084184	-0,309348478	0,971691756	-0,271977472
phil20	1793 Crombie	Essay	1,335041048	0,334802701	0,562605698	-1,734101815
phil21	1801 Belsham	Lecture	-0,474757757	-0,472221268	1,789016745	-0,803645586
phil22	1810 Stewart	Essay	0,805042708	-1,041216688	1,226217055	0,792666787
phil23	1811 Kirwan	Essay	1,043649145	-0,665956716	1,684287246	-1,318636066
phil24	1820 Brown	Lecture	1,574604388	-0,055727594	0,103827104	1,020825572
phil25	1824 Phillips	Dialogue	0,634335439	0,548274445	0,631834825	-1,504817211
phil26	1830 Mackintosh	Treatise	0,010241849	-0,716364193	1,329710484	0,15413944
phil27	1835 Hampden	Lecture	0,605054141	-0,077065096	1,093893713	0,197431152
phil28	1838 Powell	Treatise	0,084131711	-1,232030775	1,867969407	0,687600324
phil29	1845 Mill	Treatise	1,850256036	-0,068076018	0,413262445	-0,17749583
phil30	1846 Combe	Lecture	0,537898865	-0,343891603	1,333488523	-0,880263383
phil31	1855 Lyall	Treatise	1,938721781	-0,372769663	-0,234517436	-0,974782616
phil32	1860 Slack	Treatise	-0,029455695	-0,048620898	1,97451199	0,203453761
phil33	1862 Simon	Treatise	3,366596063	0,052921867	-1,062638948	0,319286215
phil34	1866 Mansel	Article	0,906044426	-0,716396916	1,498568662	-1,152264087
phil35	1874 Woodward	Treatise	0,221597116	0,335856048	0,883656301	-0,583413424
phil36	1879 Balfour	Essay	2,288630252	0,481350111	0,737246037	-0,122409642
phil37	1885 Seth	Lecture	0,953405222	-0,502769637	0,925468775	-0,082380552
phil38	1890 Mackenzie	Essay	1,209458931	0,020988357	1,517421414	0,75693625
phil39	1893 Bonar	Treatise	0,567175458	-0,461028342	0,099453749	-0,33588363
phil40	1898 Hodgson	Treatise	0,087800908	0,163575632	1,61591523	-0,576892234

Factor scores (continued)

Astronomy

Text ID	TEXT	GENRE	FACTOR1	FACTOR2	FACTOR3	FACTOR4
astr1	1702 Curson	Textbook	-0,759681625	1,398129113	-0,991660479	-0,674066454
astr2	1702 Morden	Textbook	-1,133237905	0,594432458	-0,362387311	-0,565172201
astr3	1715 Whiston	Lecture	0,314578071	-0,113341904	-0,674804095	1,183402958
astr4	1719 Harris	Dialogue	1,426760284	1,337639473	-3,165577365	2,602763982
astr5	1726 Gordon	Treatise	-1,050949638	1,191811781	-0,772773229	0,487979844
astr6	1726 Watts	Textbook	-0,969528474	0,848130879	-1,449663468	-0,277117145
astr7	1732 Fuller	Textbook	-1,179240851	0,764227553	-0,916546255	-0,090944971
astr8	1735 Charlton	Textbook	-0,866635536	1,261275018	-1,654272376	0,174714601
astr9	1742 Long	Textbook	-0,811472679	2,097127056	-0,756708847	-0,446416827
astr10	1749 Hodgson	Textbook	-1,393674492	2,200247296	0,478673598	-0,455231171
astr11	1754 Hill	Dictionary	-0,326489183	0,398800604	-0,377791428	-0,67982586
astr12	1756 Ferguson	Treatise	-0,957799351	1,628697839	-0,942133739	0,591572137
astr13	1761 Stewart	Essay	-2,43178032	4,6659471	1,286113659	0,64076648
astr14	1767 Costard	Textbook	-0,471269945	1,990023871	-0,286547401	1,053062698
astr15	1773 Wilson	Article	0,389212253	0,090177822	-0,37863629	2,877033213
astr16	1777 Adams	Textbook	-1,252395066	0,640222482	-0,270665351	-0,980433005
astr17	1779 Lacy	Treatise	-0,350376512	1,214018456	-0,5611074	-0,15527627
astr18	1782 Nicholson	Treatise	-0,427280497	0,826306072	0,867320212	0,059862483
astr19	1786 Bonnycastle	Letter	-0,04635293	-0,247140326	-0,143948141	1,04209056
astr20	1790 Vince	Treatise	-0,359433827	2,570403823	0,726835001	-0,335496591
astr21	1797 Bryan	Textbook	-0,048702519	0,56527301	0,503229849	0,28164679
astr22	1804 Small	Treatise	-0,495043613	0,23978795	1,372728789	0,23716931
astr23	1809 Ewing	Lecture	-0,975914139	1,758967985	-0,625041472	0,228444014
astr24	1811 Brewster	Treatise	-0,960982435	0,857335666	-0,952514705	1,321177326
astr25	1817 Phillips	Lecture	-0,433615338	-0,074222629	-0,298518052	0,764665871
astr26	1822 Gummere	Textbook	-0,997113116	1,305184289	1,160959623	-0,76054975
astr27	1828 Luby	Treatise	-0,906966868	1,949107031	0,889043096	-0,516591254
astr28	1833 Herschel	Treatise	-0,074303749	0,041184961	0,796069256	0,340098142
astr29	1838 Garland	Article	-0,085098569	-0,030022891	0,957233599	0,15334678
astr30	1841 Olmsted	Letter	-0,332935694	-0,079854074	0,530790368	1,898139076
astr31	1845 Bradford	Textbook	-0,354972548	0,629213495	0,331300985	-0,044222115
astr32	1855 Bartlett	Textbook	-1,531784941	0,760245139	0,822476274	-1,02891286
astr33	1858 Whewell	Essay	1,069063229	-0,358684097	0,189411761	0,194206333
astr34	1860 Mitchel	Treatise	-0,627710957	0,539161727	0,740415581	1,317485886
astr35	1868 Loomis	Textbook	-1,31000558	1,447130902	1,001488631	0,284212376
astr36	1871 Chauvenet	Treatise	-1,195389512	0,574493137	1,179000802	-0,138309382
astr37	1874 Steele	Textbook	-0,64348136	0,280736612	-0,808017534	1,276104846
astr38	1880 Darwin	Article	-0,363129686	-0,183049062	1,57841216	0,202111758
astr39	1880 Young	Article	0,06662775	-0,723717872	1,22926394	2,275183488
astr40	1889 Croll	Treatise	0,49695809	0,257150661	1,246958627	1,761353562
astr41	1893 Clerke	Treatise	-0,749460286	-0,370151453	1,233930265	1,833467403
astr42	1895 Lowell	Article	0,565748504	-0,124862177	-0,499423177	1,526632288

Factor scores (continued)

Appendix IV

Statistical output

3-factor solution for 58 features

```
> fa(r=corMat, nfactors=3, rotate="oblimin", fm="pa")
Factor Analysis using method =  pa
Call: fa(r = corMat, nfactors = 3, rotate = "oblimin", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
            PA1   PA2   PA3     h2   u2 com
PAST       0.24  0.11  0.07 0.0685 0.93 1.6
PERF       0.47  0.36  0.30 0.4134 0.59 2.6
PRES       0.09  0.08 -0.55 0.3205 0.68 1.1
PL_ADV    -0.37 -0.06 -0.21 0.1778 0.82 1.6
TIM_ADV    0.19  0.10  0.11 0.0544 0.95 2.2
FPERS      0.72 -0.01 -0.02 0.5207 0.48 1.0
SPERS      0.17 -0.16 -0.23 0.1139 0.89 2.7
TPERS      0.41 -0.05 -0.43 0.3704 0.63 2.0
ITPRO      0.47  0.06 -0.42 0.4122 0.59 2.0
DEMPRO     0.48 -0.01 -0.13 0.2565 0.74 1.1
INDPRO     0.78  0.00 -0.20 0.6544 0.35 1.1
PRO_DO     0.37 -0.34 -0.31 0.3746 0.63 2.9
QUEST      0.61 -0.06 -0.14 0.4007 0.60 1.1
NOM        0.42  0.05  0.55 0.4547 0.55 1.9
NOUN      -0.78 -0.01 -0.13 0.6089 0.39 1.1
AGPASS     0.02  0.03  0.30 0.0883 0.91 1.0
BYPASS     0.14  0.25 -0.12 0.0926 0.91 2.1
BE_MAIN    0.32 -0.20 -0.21 0.1991 0.80 2.4
EXTHERE    0.51 -0.06 -0.16 0.3007 0.70 1.2
THAT_V     0.48  0.06  0.26 0.2864 0.71 1.6
THAT_ADJ   0.45  0.13  0.32 0.3034 0.70 2.0
WHCL_SUB   0.49 -0.01  0.00 0.2399 0.76 1.0
WHCL_OB    0.71 -0.20 -0.15 0.5965 0.40 1.3
TO_INF     0.76 -0.04  0.05 0.5848 0.42 1.0
PASTPART  -0.33  0.19  0.00 0.1527 0.85 1.6
WHIZ      -0.24 -0.06  0.52 0.3409 0.66 1.4
PRES_WHIZ -0.47 -0.13  0.13 0.2485 0.75 1.3
THAT_SUB   0.14 -0.21 -0.10 0.0804 0.92 2.2
WHREL_SUB  0.21  0.26 -0.01 0.1046 0.90 1.9
WHREL_OB   0.57  0.13  0.03 0.3356 0.66 1.1
PIP        0.25  0.09  0.26 0.1277 0.87 2.3
SREL       0.06 -0.35 -0.24 0.1945 0.81 1.8
CAUSADV   -0.05 -0.56  0.02 0.3117 0.69 1.0
CONCADV    0.38  0.36 -0.14 0.2882 0.71 2.3
CONDADV    0.39 -0.51  0.04 0.4286 0.57 1.9
OTHADV     0.17 -0.31  0.06 0.1322 0.87 1.7
PREP      -0.15 -0.05  0.67 0.4779 0.52 1.1
ATTRADJ   -0.04  0.66  0.32 0.5608 0.44 1.4
PREDADJ   -0.01  0.32 -0.20 0.1401 0.86 1.7
ADV        0.52  0.13 -0.17 0.3197 0.68 1.3
CONJ       0.03 -0.55  0.48 0.5203 0.48 2.0
DOWN       0.21  0.44  0.28 0.3147 0.69 2.2
HEDG      -0.14  0.64  0.02 0.4368 0.56 1.1
AMPL       0.13  0.48 -0.21 0.2849 0.72 1.6
DEM        0.52  0.08 -0.08 0.2809 0.72 1.1
POSSMOD    0.81 -0.07  0.18 0.6773 0.32 1.1
NECMOD     0.61 -0.17  0.19 0.4297 0.57 1.3
PREDMOD    0.09 -0.71  0.26 0.5748 0.43 1.3
PUBV       0.66  0.01 -0.10 0.4556 0.54 1.0
PRIVV      0.72 -0.11  0.12 0.5375 0.46 1.1
SUASV      0.53  0.16  0.14 0.3125 0.69 1.3
```

```
SEEM       -0.01  0.07  0.01 0.0055 0.99 1.1
SPLITINF   -0.06 -0.02  0.15 0.0256 0.97 1.4
SPLITAUX    0.46  0.38  0.29 0.4193 0.58 2.7
PHCOORD     0.11  0.23 -0.34 0.1787 0.82 2.0
CLCOORD     0.39 -0.16  0.00 0.1811 0.82 1.3
SNEG        0.75 -0.04 -0.17 0.6041 0.40 1.1
ANEG        0.88  0.03 -0.07 0.7800 0.22 1.0

                        PA1  PA2  PA3
SS loadings           11.45 4.11 3.59
Proportion Var         0.20 0.07 0.06
Cumulative Var         0.20 0.27 0.33
Proportion Explained   0.60 0.21 0.19
Cumulative Proportion  0.60 0.81 1.00

 With factor correlations of
      PA1   PA2   PA3
PA1  1.00 -0.04 -0.05
PA2 -0.04  1.00  0.03
PA3 -0.05  0.03  1.00

Mean item complexity =  1.6
Test of the hypothesis that 3 factors are sufficient.

The degrees of freedom for the null model are  1653  and the
objective function was  46.54
The degrees of freedom for the model are 1482  and the objective
function was  27.97

The root mean square of the residuals (RMSR) is  0.09
The df corrected root mean square of the residuals is  0.09

Fit based upon off diagonal values = 0.86
Measures of factor score adequacy
                                                 PA1  PA2  PA3
Correlation of scores with factors              0.98 0.94 0.93
Multiple R square of scores with factors        0.97 0.89 0.86
Minimum correlation of possible factor scores  0.94 0.78 0.73
```

4-factor solution for 58 features

```
> fa(r=corMat, nfactors=4, rotate="oblimin", fm="pa")
Factor Analysis using method =  pa
Call: fa(r = corMat, nfactors = 4, rotate = "oblimin", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
             PA1   PA2   PA3   PA4    h2   u2 com
PAST        0.20  0.04 -0.05  0.37 0.180 0.82 1.6
PERF        0.40  0.24  0.21  0.48 0.512 0.49 2.9
PRES        0.18  0.22 -0.38 -0.39 0.380 0.62 3.0
PL_ADV     -0.37 -0.09 -0.43  0.32 0.418 0.58 2.9
TIM_ADV     0.14 -0.02 -0.15  0.67 0.485 0.51 1.2
FPERS       0.71 -0.02 -0.03  0.12 0.528 0.47 1.1
SPERS       0.19 -0.15 -0.34  0.15 0.190 0.81 2.5
TPERS       0.46  0.02 -0.39 -0.08 0.371 0.63 2.0
ITPRO       0.52  0.15 -0.32 -0.17 0.416 0.58 2.1
DEMPRO      0.50  0.00 -0.12  0.04 0.259 0.74 1.1
INDPRO      0.80  0.02 -0.15  0.01 0.654 0.35 1.1
PRO_DO      0.42 -0.28 -0.33 -0.09 0.375 0.62 2.8
QUEST       0.63 -0.02 -0.04 -0.15 0.422 0.58 1.1
NOM         0.37  0.00  0.74 -0.16 0.698 0.30 1.6
NOUN       -0.75  0.03 -0.14 -0.12 0.607 0.39 1.1
AGPASS     -0.02  0.00  0.37 -0.09 0.138 0.86 1.1
BYPASS      0.16  0.29  0.00 -0.13 0.115 0.88 2.0
BE_MAIN     0.39 -0.08  0.01 -0.55 0.459 0.54 1.9
EXTHERE     0.53 -0.06 -0.20  0.12 0.328 0.67 1.4
THAT_V      0.43 -0.03  0.18  0.30 0.318 0.68 2.2
THAT_ADJ    0.38  0.03  0.26  0.29 0.318 0.68 2.7
WHCL_SUB    0.50  0.02  0.13 -0.18 0.296 0.70 1.4
WHCL_OB     0.73 -0.20 -0.21  0.11 0.627 0.37 1.4
TO_INF      0.76 -0.04  0.13 -0.04 0.598 0.40 1.1
PASTPART   -0.33  0.22  0.10 -0.17 0.194 0.81 2.6
WHIZ       -0.31 -0.15  0.47  0.10 0.340 0.66 2.1
PRES_WHIZ  -0.49 -0.18 -0.03  0.21 0.309 0.69 1.7
THAT_SUB    0.16 -0.19 -0.14 -0.01 0.084 0.92 2.8
WHREL_SUB   0.20  0.26  0.06 -0.01 0.108 0.89 2.0
WHREL_OB    0.55  0.11  0.06  0.11 0.335 0.66 1.2
PIP         0.21  0.06  0.32 -0.02 0.152 0.85 1.8
SREL        0.10 -0.31 -0.32 -0.02 0.210 0.79 2.2
CAUSADV    -0.03 -0.55 -0.10 -0.01 0.317 0.68 1.1
CONCADV     0.39  0.37 -0.08  0.08 0.288 0.71 2.2
CONDADV     0.41 -0.49  0.03 -0.14 0.437 0.56 2.1
OTHADV      0.16 -0.35 -0.08  0.18 0.184 0.82 2.1
PREP       -0.24 -0.17  0.60  0.18 0.476 0.52 1.7
ATTRADJ    -0.11  0.59  0.42  0.14 0.568 0.43 2.0
PREDADJ     0.02  0.40 -0.02 -0.25 0.212 0.79 1.7
ADV         0.53  0.09 -0.30  0.39 0.516 0.48 2.6
CONJ       -0.02 -0.62  0.35  0.08 0.510 0.49 1.6
DOWN        0.15  0.35  0.25  0.34 0.338 0.66 3.2
HEDG       -0.18  0.58 -0.01  0.34 0.512 0.49 1.8
AMPL        0.13  0.47 -0.24  0.27 0.372 0.63 2.4
DEM         0.52  0.07 -0.09  0.14 0.297 0.70 1.3
POSSMOD     0.78 -0.09  0.24 -0.02 0.695 0.31 1.2
NECMOD      0.59 -0.19  0.22 -0.02 0.442 0.56 1.5
PREDMOD     0.07 -0.77  0.06  0.14 0.612 0.39 1.1
PUBV        0.67  0.03 -0.06  0.01 0.454 0.55 1.0
PRIVV       0.69 -0.17  0.03  0.27 0.587 0.41 1.4
SUASV       0.52  0.17  0.30 -0.14 0.399 0.60 2.0
SEEM       -0.03  0.01 -0.14  0.36 0.143 0.86 1.3
SPLITINF   -0.08 -0.06  0.11  0.07 0.026 0.97 3.3
SPLITAUX    0.39  0.28  0.27  0.34 0.440 0.56 3.6
PHCOORD     0.17  0.36 -0.12 -0.40 0.314 0.69 2.5
```

```
CLCOORD    0.38 -0.18 -0.05  0.11 0.195 0.80 1.6
SNEG       0.78  0.01 -0.06 -0.14 0.622 0.38 1.1
ANEG       0.89  0.06  0.05 -0.11 0.807 0.19 1.0

                         PA1  PA2  PA3  PA4
SS loadings            11.53 4.05 3.52 3.10
Proportion Var          0.20 0.07 0.06 0.05
Cumulative Var          0.20 0.27 0.33 0.38
Proportion Explained    0.52 0.18 0.16 0.14
Cumulative Proportion   0.52 0.70 0.86 1.00

 With factor correlations of
      PA1   PA2  PA3  PA4
PA1  1.00 -0.02 0.02 0.01
PA2 -0.02  1.00 0.00 0.05
PA3  0.02  0.00 1.00 0.05
PA4  0.01  0.05 0.05 1.00

Mean item complexity =  1.9
Test of the hypothesis that 4 factors are sufficient.

The degrees of freedom for the null model are  1653  and the
objective function was  46.54
The degrees of freedom for the model are 1427  and the objective
function was  25.2

The root mean square of the residuals (RMSR) is  0.07
The df corrected root mean square of the residuals is  0.08

Fit based upon off diagonal values = 0.9
Measures of factor score adequacy
                                                PA1  PA2  PA3  PA4
Correlation of scores with factors             0.99 0.94 0.94 0.93
Multiple R square of scores with factors       0.97 0.89 0.89 0.86
Minimum correlation of possible factor scores  0.94 0.78 0.77 0.71
```

5-factor solution for 58 features

```
> fa(r=corMat, nfactors=5, rotate="oblimin", fm="pa")
Factor Analysis using method =  pa
Call: fa(r = corMat, nfactors = 5, rotate = "oblimin", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
            PA1   PA2   PA3   PA5   PA4    h2   u2 com
PAST       0.05  0.00 -0.03  0.22  0.48 0.271 0.73 1.4
PERF       0.27 -0.25  0.18  0.09  0.55 0.548 0.45 2.3
PRES       0.32 -0.20 -0.31  0.08 -0.56 0.511 0.49 2.6
PL_ADV    -0.21  0.06 -0.55 -0.08  0.21 0.418 0.58 1.7
TIM_ADV    0.19 -0.02 -0.28 -0.05  0.62 0.486 0.51 1.6
FPERS      0.70  0.02  0.02  0.04  0.08 0.539 0.46 1.0
SPERS      0.31  0.14 -0.36 -0.03  0.05 0.207 0.79 2.3
TPERS      0.19  0.15 -0.16  0.61  0.11 0.518 0.48 1.5
ITPRO      0.50 -0.08 -0.18  0.24 -0.21 0.424 0.58 2.2
DEMPRO     0.54 -0.01 -0.08  0.02 -0.04 0.282 0.72 1.1
INDPRO     0.85 -0.02 -0.08  0.04 -0.09 0.712 0.29 1.1
PRO_DO     0.33  0.36 -0.18  0.25 -0.04 0.382 0.62 3.3
QUEST      0.58  0.05  0.08  0.13 -0.15 0.428 0.57 1.3
NOM        0.18 -0.02  0.79 -0.06 -0.02 0.700 0.30 1.1
NOUN      -0.61 -0.05 -0.23 -0.14 -0.18 0.609 0.39 1.6
AGPASS    -0.11 -0.01  0.38 -0.04 -0.01 0.144 0.86 1.2
BYPASS    -0.10 -0.18  0.16  0.42  0.05 0.215 0.79 1.9
BE_MAIN    0.33  0.14  0.18  0.12 -0.52 0.473 0.53 2.3
EXTHERE    0.57  0.06 -0.16  0.04  0.04 0.343 0.66 1.2
THAT_V     0.65 -0.10  0.03 -0.41  0.11 0.480 0.52 1.8
THAT_ADJ   0.54 -0.15  0.13 -0.35  0.14 0.415 0.58 2.2
WHCL_SUB   0.43  0.00  0.22  0.07 -0.16 0.300 0.70 1.9
WHCL_OB    0.77  0.20 -0.15  0.03  0.03 0.651 0.35 1.2
TO_INF     0.52  0.11  0.31  0.30  0.10 0.634 0.37 2.5
PASTPART  -0.27 -0.24  0.06 -0.08 -0.20 0.201 0.80 3.2
WHIZ      -0.15  0.01  0.28 -0.49  0.02 0.384 0.62 1.8
PRES_WHIZ -0.19  0.06 -0.26 -0.48  0.02 0.400 0.60 2.0
THAT_SUB   0.23  0.19 -0.13 -0.05 -0.07 0.095 0.90 3.0
WHREL_SUB -0.05 -0.17  0.20  0.38  0.18 0.218 0.78 2.5
WHREL_OB   0.39 -0.06  0.16  0.24  0.20 0.357 0.64 2.7
PIP        0.11 -0.06  0.34  0.00  0.05 0.154 0.85 1.3
SREL       0.00  0.40 -0.21  0.24  0.07 0.255 0.75 2.3
CAUSADV    0.04  0.53 -0.12 -0.15 -0.04 0.317 0.68 1.3
CONCADV    0.27 -0.31  0.01  0.27  0.11 0.295 0.71 3.2
CONDADV    0.29  0.53  0.15  0.09 -0.05 0.447 0.55 1.8
OTHADV     0.15  0.35 -0.08 -0.02  0.20 0.187 0.81 2.1
PREP      -0.37  0.13  0.53 -0.16  0.33 0.545 0.46 2.9
ATTRADJ   -0.04 -0.68  0.28 -0.20  0.06 0.593 0.41 1.6
PREDADJ   -0.06 -0.34  0.07  0.22 -0.21 0.211 0.79 2.7
ADV        0.56 -0.10 -0.31  0.10  0.31 0.514 0.49 2.3
CONJ       0.09  0.52  0.23 -0.44  0.03 0.532 0.47 2.5
DOWN       0.29 -0.46  0.09 -0.27  0.20 0.398 0.60 2.9
HEDG      -0.05 -0.65 -0.17 -0.10  0.21 0.531 0.47 1.4
AMPL       0.21 -0.48 -0.29  0.07  0.16 0.380 0.62 2.4
DEM        0.30  0.01  0.05  0.37  0.28 0.375 0.63 2.9
POSSMOD    0.68  0.09  0.34  0.04  0.02 0.695 0.30 1.5
NECMOD     0.34  0.25  0.37  0.23  0.15 0.502 0.50 3.8
PREDMOD    0.08  0.73  0.04 -0.19  0.18 0.614 0.39 1.3
PUBV       0.71 -0.04 -0.01  0.00 -0.07 0.499 0.50 1.0
PRIVV      0.79  0.10 -0.01 -0.18  0.16 0.648 0.35 1.2
SUASV      0.45 -0.17  0.37  0.01 -0.13 0.409 0.59 2.4
SEEM      -0.15  0.03 -0.14  0.20  0.47 0.245 0.75 1.8
SPLITINF  -0.01  0.01  0.04 -0.16  0.03 0.032 0.97 1.2
SPLITAUX   0.23 -0.28  0.28  0.14  0.43 0.480 0.52 3.4
PHCOORD   -0.03 -0.23  0.09  0.44 -0.27 0.340 0.66 2.4
```

```
CLCOORD    0.26  0.22  0.03  0.18  0.20 0.222 0.78 3.7
SNEG       0.60  0.07  0.13  0.31 -0.05 0.625 0.37 1.7
ANEG       0.78 -0.02  0.20  0.19 -0.09 0.809 0.19 1.3

                       PA1  PA2  PA3  PA5  PA4
SS loadings           9.99 4.10 3.57 3.49 3.05
Proportion Var        0.17 0.07 0.06 0.06 0.05
Cumulative Var        0.17 0.24 0.30 0.36 0.42
Proportion Explained  0.41 0.17 0.15 0.14 0.13
Cumulative Proportion 0.41 0.58 0.73 0.87 1.00

 With factor correlations of
     PA1   PA2   PA3   PA5   PA4
PA1 1.00  0.07  0.17  0.30  0.12
PA2 0.07  1.00  0.02  0.00 -0.05
PA3 0.17  0.02  1.00 -0.03  0.07
PA5 0.30  0.00 -0.03  1.00 -0.12
PA4 0.12 -0.05  0.07 -0.12  1.00

Mean item complexity =  2
Test of the hypothesis that 5 factors are sufficient.

The degrees of freedom for the null model are  1653  and the
objective function was  46.54
The degrees of freedom for the model are 1373  and the objective
function was  23.48

The root mean square of the residuals (RMSR) is  0.06
The df corrected root mean square of the residuals is  0.07

Fit based upon off diagonal values = 0.92
Measures of factor score adequacy
                                               PA1  PA2  PA3  PA5
PA4
Correlation of scores with factors            0.98 0.95 0.95 0.92
0.93
Multiple R square of scores with factors      0.96 0.89 0.90 0.85
0.86
Minimum correlation of possible factor scores  0.93 0.79 0.80 0.71
0.72
```

6-factor solution for 58 features

```
> fa(r=corMat, nfactors=6, rotate="oblimin", fm="pa")
Factor Analysis using method =  pa
Call: fa(r = corMat, nfactors = 6, rotate = "oblimin", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
             PA1   PA6   PA2   PA3   PA4   PA5    h2   u2 com
PAST        0.07 -0.03  0.04 -0.06  0.40  0.35 0.284 0.72 2.1
PERF        0.26  0.12 -0.25 -0.23  0.41  0.36 0.572 0.43 4.3
PRES        0.23 -0.16 -0.12  0.48 -0.46  0.00 0.510 0.49 2.8
PL_ADV     -0.06 -0.46  0.10  0.24  0.33 -0.11 0.415 0.58 2.7
TIM_ADV     0.19 -0.06 -0.02  0.07  0.67  0.01 0.503 0.50 1.2
FPERS       0.60  0.12  0.05  0.03  0.01  0.23 0.584 0.42 1.4
SPERS       0.39 -0.33  0.20  0.17  0.06  0.11 0.265 0.73 3.2
TPERS      -0.11  0.26  0.25  0.44  0.13  0.36 0.520 0.48 3.6
ITPRO       0.27  0.13 -0.01  0.42 -0.16  0.14 0.420 0.58 2.6
DEMPRO      0.38  0.14  0.01  0.19 -0.01  0.03 0.279 0.72 1.8
INDPRO      0.60  0.25  0.01  0.28 -0.07  0.06 0.710 0.29 1.9
PRO_DO      0.17  0.10  0.41  0.28 -0.01  0.14 0.381 0.62 2.6
QUEST       0.36  0.29  0.07  0.16 -0.17  0.11 0.429 0.57 3.2
NOM         0.06  0.60 -0.11 -0.53 -0.19  0.05 0.689 0.31 2.3
NOUN       -0.35 -0.48 -0.05  0.00 -0.13 -0.15 0.611 0.39 2.2
AGPASS     -0.22  0.38 -0.08 -0.22 -0.03 -0.13 0.195 0.80 2.7
BYPASS     -0.23  0.20 -0.13  0.07 -0.03  0.36 0.213 0.79 2.8
BE_MAIN     0.15  0.27  0.16  0.12 -0.54  0.03 0.474 0.53 2.0
EXTHERE     0.47  0.02  0.10  0.18  0.05  0.12 0.354 0.65 1.6
THAT_V      0.69  0.02 -0.15 -0.15  0.09 -0.14 0.488 0.51 1.3
THAT_ADJ    0.52  0.16 -0.22 -0.17  0.12 -0.16 0.413 0.59 2.2
WHCL_SUB    0.17  0.47 -0.02  0.09 -0.16 -0.06 0.331 0.67 1.7
WHCL_OB     0.62  0.11  0.23  0.19  0.04  0.11 0.661 0.34 1.7
TO_INF      0.21  0.55  0.12  0.01  0.02  0.25 0.633 0.37 1.9
PASTPART   -0.25  0.01 -0.26  0.02 -0.16 -0.19 0.221 0.78 3.5
WHIZ        0.06 -0.05 -0.09 -0.49 -0.03 -0.30 0.381 0.62 1.8
PRES_WHIZ   0.08 -0.42  0.01 -0.11  0.10 -0.37 0.399 0.60 2.3
THAT_SUB    0.16  0.03  0.19  0.13  0.00 -0.11 0.098 0.90 3.6
WHREL_SUB  -0.04 -0.03 -0.10 -0.16 -0.01  0.64 0.402 0.60 1.2
WHREL_OB    0.41 -0.03  0.00 -0.18  0.00  0.61 0.614 0.39 2.0
PIP         0.24 -0.03 -0.07 -0.42 -0.16  0.37 0.350 0.65 3.0
SREL       -0.16  0.11  0.44  0.27  0.15  0.00 0.291 0.71 2.5
CAUSADV     0.15 -0.17  0.52 -0.10 -0.02 -0.11 0.328 0.67 1.6
CONCADV     0.06  0.24 -0.27  0.23  0.11  0.20 0.302 0.70 4.3
CONDADV     0.07  0.41  0.51  0.01 -0.03 -0.08 0.478 0.52 2.0
OTHADV      0.08  0.11  0.33  0.02  0.24 -0.09 0.204 0.80 2.4
PREP       -0.14  0.02  0.05 -0.76  0.12  0.14 0.609 0.39 1.2
ATTRADJ    -0.01  0.12 -0.73 -0.20  0.01 -0.07 0.597 0.40 1.2
PREDADJ    -0.16  0.09 -0.30  0.16 -0.22  0.14 0.209 0.79 3.9
ADV         0.28  0.26 -0.07  0.45  0.45 -0.09 0.648 0.35 3.4
CONJ        0.19  0.11  0.41 -0.43  0.02 -0.35 0.530 0.47 3.5
DOWN        0.27  0.11 -0.51 -0.09  0.20 -0.14 0.411 0.59 2.3
HEDG       -0.02 -0.11 -0.66  0.14  0.27 -0.08 0.556 0.44 1.6
AMPL        0.10 -0.02 -0.44  0.37  0.25  0.00 0.408 0.59 2.7
DEM         0.19  0.11  0.09  0.03  0.17  0.50 0.433 0.57 1.7
POSSMOD     0.35  0.65  0.05 -0.01 -0.01 -0.04 0.739 0.26 1.6
NECMOD     -0.04  0.75  0.22 -0.02  0.14  0.01 0.627 0.37 1.3
PREDMOD     0.08  0.13  0.68 -0.21  0.20 -0.23 0.631 0.37 1.8
PUBV        0.63  0.06 -0.01  0.09 -0.13  0.18 0.559 0.44 1.3
PRIVV       0.76  0.06  0.10 -0.06  0.10  0.07 0.692 0.31 1.1
SUASV       0.24  0.46 -0.20 -0.06 -0.19  0.02 0.412 0.59 2.4
SEEM       -0.13 -0.09  0.06  0.02  0.44  0.23 0.241 0.76 1.8
SPLITINF    0.02  0.01 -0.03 -0.10  0.05 -0.15 0.037 0.96 2.1
SPLITAUX   -0.04  0.57 -0.32 -0.05  0.43  0.06 0.609 0.39 2.6
PHCOORD    -0.28  0.29 -0.18  0.33 -0.26  0.18 0.355 0.65 5.1
```

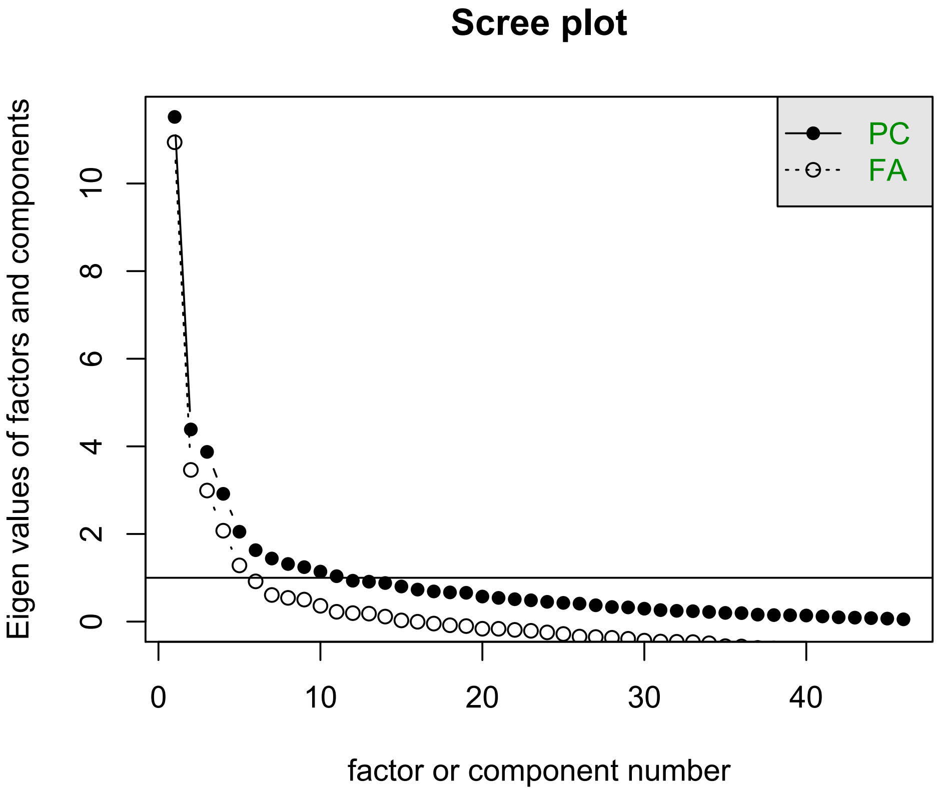
```
CLCOORD    0.08  0.27  0.23  0.08  0.20  0.08 0.233 0.77 3.4
SNEG       0.21  0.57  0.10  0.28 -0.04  0.12 0.661 0.34 2.0
ANEG       0.41  0.58 -0.01  0.20 -0.10  0.09 0.826 0.17 2.2

                       PA1  PA6  PA2  PA3  PA4  PA5
SS loadings           6.21 5.89 4.13 3.74 2.86 3.19
Proportion Var        0.11 0.10 0.07 0.06 0.05 0.06
Cumulative Var        0.11 0.21 0.28 0.34 0.39 0.45
Proportion Explained  0.24 0.23 0.16 0.14 0.11 0.12
Cumulative Proportion 0.24 0.46 0.62 0.77 0.88 1.00

 With factor correlations of
     PA1  PA6   PA2   PA3   PA4   PA5
PA1 1.00 0.43  0.09  0.12  0.09  0.23
PA6 0.43 1.00  0.05  0.04  0.02  0.25
PA2 0.09 0.05  1.00  0.05 -0.05 -0.05
PA3 0.12 0.04  0.05  1.00 -0.03  0.16
PA4 0.09 0.02 -0.05 -0.03  1.00 -0.02
PA5 0.23 0.25 -0.05  0.16 -0.02  1.00

Mean item complexity =  2.3
Test of the hypothesis that 6 factors are sufficient.

The degrees of freedom for the null model are  1653  and the
objective function was  46.54
The degrees of freedom for the model are 1320  and the objective
function was  21.77

The root mean square of the residuals (RMSR) is  0.06
The df corrected root mean square of the residuals is  0.06

Fit based upon off diagonal values = 0.94
Measures of factor score adequacy
                                                 PA1  PA6  PA2  PA3
PA4
Correlation of scores with factors              0.96 0.97 0.95 0.94
0.93
Multiple R square of scores with factors        0.92 0.95 0.90 0.89
0.86
Minimum correlation of possible factor scores  0.85 0.89 0.80 0.78
0.72
                                                 PA5
Correlation of scores with factors              0.92
Multiple R square of scores with factors        0.84
Minimum correlation of possible factor scores  0.69
```

5-factor solution for 46 features and scree plot (*)

```
> fa(r=corMat, nfactors=5, rotate="oblimin", fm="pa")
Factor Analysis using method =  pa
Call: fa(r = corMat, nfactors = 5, rotate = "oblimin", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
            PA1   PA2   PA3   PA5   PA4   h2   u2 com
PERF       0.26 -0.18  0.21  0.01  0.50 0.48 0.52 2.2
PRES       0.34 -0.24 -0.33  0.13 -0.43 0.45 0.55 3.7
PL_ADV    -0.17  0.08 -0.53 -0.09  0.18 0.37 0.63 1.6
TIM_ADV    0.13  0.04 -0.22 -0.02  0.62 0.43 0.57 1.4
FPERS      0.67  0.03  0.07  0.05  0.08 0.53 0.47 1.1
TPERS      0.15  0.15 -0.14  0.61  0.09 0.49 0.51 1.4
ITPRO      0.56 -0.10 -0.24  0.22 -0.19 0.47 0.53 2.1
DEMPRO     0.51  0.02 -0.10  0.07  0.04 0.29 0.71 1.1
INDPRO     0.85 -0.03 -0.07  0.06 -0.07 0.72 0.28 1.0
PRO_DO     0.33  0.37 -0.19  0.26 -0.04 0.40 0.60 3.4
QUEST      0.54  0.03  0.15  0.15 -0.16 0.44 0.56 1.5
NOM        0.12 -0.03  0.82 -0.06 -0.06 0.72 0.28 1.1
NOUN      -0.55 -0.08 -0.24 -0.16 -0.21 0.60 0.40 2.0
BE_MAIN    0.39  0.06  0.16  0.11 -0.61 0.55 0.45 2.0
EXTHERE    0.59  0.07 -0.18  0.02  0.05 0.36 0.64 1.2
THAT_V     0.68 -0.07  0.05 -0.42  0.12 0.50 0.50 1.8
THAT_ADJ   0.55 -0.11  0.13 -0.35  0.17 0.42 0.58 2.1
WHCL_SUB   0.42 -0.03  0.23  0.09 -0.15 0.31 0.69 2.0
WHCL_OB    0.73  0.21 -0.12  0.07  0.05 0.63 0.37 1.3
TO_INF     0.43  0.13  0.35  0.34  0.12 0.65 0.35 3.2
WHIZ      -0.14  0.03  0.27 -0.49  0.01 0.37 0.63 1.7
PRES_WHIZ -0.14  0.04 -0.24 -0.49 -0.03 0.38 0.62 1.6
WHREL_OB   0.44 -0.04  0.13  0.11  0.11 0.31 0.69 1.5
SREL      -0.04  0.40 -0.22  0.28  0.06 0.27 0.73 2.5
CAUSADV    0.06  0.49 -0.09 -0.21 -0.13 0.30 0.70 1.6
CONCADV    0.24 -0.30  0.03  0.27  0.16 0.29 0.71 3.5
CONDADV    0.20  0.57  0.15  0.17  0.01 0.49 0.51 1.6
PREP      -0.33  0.13  0.48 -0.29  0.19 0.48 0.52 3.0
ATTRADJ   -0.10 -0.67  0.35 -0.12  0.15 0.63 0.37 1.8
PREDADJ   -0.08 -0.37  0.10  0.24 -0.17 0.22 0.78 2.5
ADV        0.45 -0.04 -0.25  0.23  0.44 0.56 0.44 3.1
CONJ       0.09  0.50  0.23 -0.46 -0.03 0.51 0.49 2.5
DOWN       0.29 -0.41  0.09 -0.28  0.28 0.41 0.59 3.6
HEDG      -0.05 -0.60 -0.15 -0.08  0.32 0.54 0.46 1.7
AMPL       0.18 -0.42 -0.30  0.14  0.33 0.42 0.58 3.5
DEM        0.30  0.03  0.04  0.29  0.22 0.30 0.70 2.9
POSSMOD    0.59  0.12  0.35  0.13  0.10 0.69 0.31 1.9
NECMOD     0.17  0.30  0.44  0.38  0.24 0.62 0.38 3.7
PREDMOD   -0.01  0.78  0.05 -0.11  0.18 0.64 0.36 1.2
PUBV       0.73 -0.04  0.02 -0.02 -0.10 0.52 0.48 1.0
PRIVV      0.81  0.13  0.02 -0.21  0.11 0.67 0.33 1.2
SUASV      0.38 -0.18  0.42  0.09 -0.08 0.42 0.58 2.5
SPLITAUX   0.12 -0.20  0.29  0.21  0.55 0.55 0.45 2.3
PHCOORD   -0.10 -0.28  0.16  0.51 -0.24 0.40 0.60 2.4
SNEG       0.54  0.05  0.19  0.36 -0.05 0.64 0.36 2.1
ANEG       0.73 -0.01  0.23  0.22 -0.04 0.81 0.19 1.4

                      PA1  PA2  PA3  PA5  PA4
SS loadings           9.0 3.68 3.35 3.48 2.73
Proportion Var        0.2 0.08 0.07 0.08 0.06
Cumulative Var        0.2 0.28 0.35 0.42 0.48
Proportion Explained  0.4 0.17 0.15 0.16 0.12
Cumulative Proportion 0.4 0.57 0.72 0.88 1.00
```

```
 With factor correlations of
      PA1   PA2  PA3   PA5   PA4
PA1 1.00  0.10 0.22  0.34  0.15
PA2 0.10  1.00 0.05  0.01 -0.07
PA3 0.22  0.05 1.00  0.00  0.10
PA5 0.34  0.01 0.00  1.00 -0.09
PA4 0.15 -0.07 0.10 -0.09  1.00

Mean item complexity =  2.1
Test of the hypothesis that 5 factors are sufficient.

The degrees of freedom for the null model are  1035  and the
objective function was  34.64
The degrees of freedom for the model are 815  and the objective
function was  13.21

The root mean square of the residuals (RMSR) is  0.05
The df corrected root mean square of the residuals is  0.06

Fit based upon off diagonal values = 0.96
Measures of factor score adequacy
                                                 PA1  PA2  PA3  PA5
PA4
Correlation of scores with factors              0.98 0.94 0.94 0.93
0.92
Multiple R square of scores with factors        0.96 0.89 0.89 0.86
0.84
Minimum correlation of possible factor scores  0.91 0.78 0.78 0.72
0.69
```

(*)



4-factor solution for 54 features

```
> fa(r=corMat, nfactors=4, rotate="oblimin", fm="pa")
Factor Analysis using method =  pa
Call: fa(r = corMat, nfactors = 4, rotate = "oblimin", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
            PA1   PA2   PA3   PA4   h2   u2 com
PAST       0.20 -0.02 -0.06  0.38 0.19 0.81 1.6
PERF       0.40 -0.24  0.20  0.48 0.51 0.49 2.9
PRES       0.17 -0.21 -0.39 -0.41 0.39 0.61 2.8
PL_ADV    -0.37  0.09 -0.43  0.32 0.42 0.58 2.9
TIM_ADV    0.15 -0.01 -0.15  0.66 0.47 0.53 1.2
FPERS      0.72  0.02 -0.03  0.11 0.53 0.47 1.1
SPERS      0.19  0.15 -0.33  0.14 0.18 0.82 2.5
TPERS      0.45  0.04 -0.40 -0.06 0.37 0.63 2.0
ITPRO      0.52 -0.12 -0.33 -0.18 0.41 0.59 2.1
DEMPRO     0.50 -0.01 -0.12  0.02 0.26 0.74 1.1
INDPRO     0.80 -0.03 -0.16 -0.01 0.66 0.34 1.1
PRO_DO     0.41  0.31 -0.32 -0.08 0.38 0.62 2.9
QUEST      0.63  0.02 -0.05 -0.16 0.43 0.57 1.1
NOM        0.38 -0.02  0.73 -0.16 0.69 0.31 1.6
NOUN      -0.75 -0.04 -0.13 -0.13 0.61 0.39 1.1
AGPASS    -0.01 -0.02  0.36 -0.09 0.14 0.86 1.1
BE_MAIN    0.39  0.10  0.01 -0.55 0.47 0.53 1.9
EXTHERE    0.52  0.07 -0.20  0.11 0.33 0.67 1.4
THAT_V     0.44 -0.01  0.18  0.28 0.31 0.69 2.1
THAT_ADJ   0.40 -0.08  0.26  0.27 0.32 0.68 2.7
WHCL_SUB   0.50 -0.03  0.12 -0.18 0.30 0.70 1.4
WHCL_OB    0.73  0.19 -0.20  0.11 0.62 0.38 1.3
TO_INF     0.76  0.04  0.12 -0.04 0.59 0.41 1.1
PASTPART  -0.32 -0.24  0.09 -0.19 0.20 0.80 2.7
WHIZ      -0.30  0.10  0.48  0.10 0.34 0.66 1.9
PRES_WHIZ -0.49  0.14 -0.01  0.20 0.29 0.71 1.5
WHREL_OB   0.55 -0.08  0.04  0.11 0.32 0.68 1.1
PIP        0.21 -0.04  0.31 -0.01 0.14 0.86 1.8
SREL       0.09  0.35 -0.30  0.00 0.22 0.78 2.1
CAUSADV   -0.05  0.56 -0.06  0.02 0.32 0.68 1.0
CONCADV    0.38 -0.35 -0.10  0.07 0.28 0.72 2.2
CONDADV    0.40  0.50  0.05 -0.12 0.44 0.56 2.1
OTHADV     0.16  0.35 -0.06  0.20 0.18 0.82 2.1
PREP      -0.24  0.16  0.60  0.20 0.49 0.51 1.7
ATTRADJ   -0.09 -0.65  0.39  0.09 0.60 0.40 1.7
PREDADJ    0.02 -0.38 -0.04 -0.27 0.21 0.79 1.8
ADV        0.53 -0.11 -0.30  0.38 0.52 0.48 2.6
CONJ      -0.02  0.59  0.39  0.11 0.50 0.50 1.8
DOWN       0.16 -0.38  0.23  0.32 0.35 0.65 3.0
HEDG      -0.17 -0.61 -0.04  0.32 0.53 0.47 1.7
AMPL       0.14 -0.48 -0.26  0.25 0.38 0.62 2.3
DEM        0.51 -0.03 -0.10  0.15 0.29 0.71 1.3
POSSMOD    0.79  0.07  0.25 -0.02 0.70 0.30 1.2
NECMOD     0.59  0.18  0.23 -0.02 0.44 0.56 1.5
PREDMOD    0.07  0.74  0.10  0.17 0.58 0.42 1.2
PUBV       0.68 -0.02 -0.07  0.00 0.46 0.54 1.0
PRIVV      0.70  0.15  0.03  0.26 0.58 0.42 1.4
SUASV      0.53 -0.19  0.29 -0.16 0.42 0.58 2.1
SEEM      -0.03  0.02 -0.15  0.37 0.16 0.84 1.3
SPLITAUX   0.40 -0.29  0.25  0.33 0.44 0.56 3.6
PHCOORD    0.17 -0.33 -0.15 -0.41 0.31 0.69 2.6
CLCOORD    0.38  0.19 -0.05  0.12 0.20 0.80 1.7
SNEG       0.78  0.01 -0.07 -0.13 0.62 0.38 1.1
ANEG       0.89 -0.05  0.04 -0.11 0.81 0.19 1.0
```

```
                        PA1  PA2  PA3  PA4
SS loadings           11.43 3.92 3.48 3.06
Proportion Var         0.21 0.07 0.06 0.06
Cumulative Var         0.21 0.28 0.35 0.41
Proportion Explained   0.52 0.18 0.16 0.14
Cumulative Proportion  0.52 0.70 0.86 1.00

 With factor correlations of
     PA1   PA2  PA3   PA4
PA1 1.00  0.03 0.01  0.01
PA2 0.03  1.00 0.00 -0.05
PA3 0.01  0.00 1.00  0.05
PA4 0.01 -0.05 0.05  1.00

Mean item complexity =  1.8
Test of the hypothesis that 4 factors are sufficient.

The degrees of freedom for the null model are  1431  and the
objective function was  42.89
The degrees of freedom for the model are 1221  and the objective
function was  21.81

The root mean square of the residuals (RMSR) is  0.07
The df corrected root mean square of the residuals is  0.08

Fit based upon off diagonal values = 0.92
Measures of factor score adequacy
                                                PA1  PA2  PA3  PA4
Correlation of scores with factors             0.98 0.94 0.94 0.92
Multiple R square of scores with factors       0.97 0.89 0.88 0.85
Minimum correlation of possible factor scores  0.94 0.78 0.77 0.70
```

RESUMEN DE LA TESIS DOCTORAL

Un análisis multidimensional de textos científicos en inglés moderno tardío del Coruña Corpus

Leida Maria Monaco

En este estudio se analizan la variación y el cambio lingüísticos en una muestra de 122 textos científicos escritos en lengua inglesa, publicados a lo largo de los siglos dieciocho y diecinueve, y pertenecientes al *Coruña Corpus of English Scientific Writing* (en adelante *Coruña Corpus*). Dicha muestra contiene tres subcorpus, cada uno de los cuales corresponde a una disciplina científica: *CETA*, o *Corpus of English Texts on Astronomy* (astronomía); *CEPhiT*, o *Corpus of English Philosophy Texts* (filosofía), y *CELiST*, o *Corpus of English Life Sciences Texts* (ciencias de la vida, incluyendo biología, zoología y anatomía). De acuerdo con la clasificación de las ciencias y los campos del saber de la UNESCO (1988), una de las mencionadas disciplinas, filosofía, pertenece al campo de las humanidades, mientras que las otras dos, astronomía y ciencias de la vida, son consideradas como ciencias naturales. Esta división es el primer punto donde se espera encontrar variación lingüística. Por otra parte, esta tesis tiene dos objetivos principales: 1) identificar variación y cambio entre las tres disciplinas científicas mencionadas, y 2) detectar variación y cambio entre un total de ocho géneros textuales que dan forma a los distintos textos en el corpus: Treatise (tratado), Textbook (libro de texto), Essay (ensayo), Lecture (conferencia), Article (artículo), Letter (carta), Dialogue (diálogo) y Dictionary (diccionario).

Este estudio adopta la definición de Biber & Conrad (2009) para el término *register* (registro), según la cual éste se considera una variedad situacional que puede describirse en función de los rasgos lingüísticos que lo caracterizan. Después de analizar varios enfoques para detectar variación y cambio lingüísticos, se ha decidido utilizar el método del análisis multidimensional (*Multidimensional Analysis*) de Biber (1988) para poder captar variación y cambio en varias dimensiones comunicativas, tal y como aparece recogido en el primer capítulo de esta tesis. El análisis multidimensional se utilizó por primera vez en un estudio de la variación entre diversos registros ingleses orales (conferencias, conversaciones telefónicas, emisiones de radio, etc.) y escritos (tales como artículos científicos, artículos de prensa, novelas de ficción o documentos legales). Dicho estudio, gran parte del cual consistía en un análisis factorial (Gorsuch 1983; Tabachnik & Fidell 1996) de sesenta y tres categorías léxicas y gramaticales, sacadas de cientos de textos escritos o de transcripciones de textos orales, analiza la co-ocurrencia de esas categorías lingüísticas en cada uno de los textos y establece, en base a dicha co-ocurrencia, cinco "dimensiones de variación" (por ejemplo, dimensión "emotiva vs. informativa"; dimensión "persuasiva"; dimensión "narrativa", etc.) de la lengua inglesa contemporánea, a lo largo de las cuales se sitúan los diversos registros, presentando así variación los unos con respecto a los otros. Tomando el estudio de Biber (1988) como modelo, esta tesis consiste también en un análisis multidimensional, analizando las dimensiones de variación lingüística entre los distintos subregistros de la filosofía, astronomía y ciencias de la vida en lengua inglesa, tal y como era entre 1700 y 1900.

Para ello, primeramente se ha querido caracterizar el registro objeto de estudio (es decir, el registro científico del inglés moderno tardío, o *late Modern English scientific register*) dentro de su contexto socio-cultural, ofreciendo una visión general de la ciencia occidental en el amplio período que se extiende entre la Revolución Científica y las primeras décadas del siglo veinte (capítulo 2). Seguidamente, se ha procedido a describir el *Coruña Corpus* y la muestra analizada, prestando especial atención a los principios de recopilación del corpus (tales como el período de tiempo que abarca, el contraste entre la representatividad y el equilibrio de la muestra, así como su tamaño en palabras), y a examinar cada uno de los tres subcorpus en cuanto a su composición en géneros textuales (capítulo 3). Después de estos dos trámites iniciales, se ha proseguido con la parte metodológica del presente análisis multidimensional. Dicha parte se ha completado en dos pasos, consistiendo el primero en seleccionar, recuperar y contar los diferentes rasgos lingüísticos que aparecen en cada uno de los textos de la muestra (capítulo 4), y el segundo, en aplicar un análisis factorial (capítulo 5).

Así pues, el capítulo 4 recoge la selección un total de cincuenta y ocho rasgos léxicos y gramaticales, basándose dicha selección en estudios previos, así como la recuperación de los datos seleccionados del corpus, para lo cual se han reutilizado y/o desarrollado una serie de algoritmos de búsqueda. Parte de las búsquedas se llevaron a cabo con la *Coruña Corpus Tool*, un programa de concordancia desarrollado para realizar búsquedas en el *Coruña Corpus* y que permite detectar variantes morfológicas, mientras que la mayoría de las búsquedas fueron procesadas con *CQPWeb*, una plataforma online que permite búsquedas por categorías gramaticales, posibilitando la detección y posterior recuento de estructuras sintácticas complejas. Una vez realizado el recuento, las frecuencias de aparición de cada rasgo en los diversos textos se han normalizado a 1000 palabras de texto para poder así ser directamente comparables. Estas frecuencias normalizadas fueron seguidamente sometidas a un análisis factorial, descrito detalladamente en el capítulo 5, el cual, luego de una serie de pruebas preliminares y ensayos (incluyendo la eliminación de ciertos rasgos que presentaban saturaciones o cargas factoriales débiles), produjo una solución de cuatro factores en un conjunto de cincuenta y cuatro variables que explica un 41 por ciento de la variación total del corpus. Cada uno de los factores se ha interpretado como una dimensión de variación, ya que los rasgos léxicos y/o gramaticales que saturan (es decir, presentan saturaciones o cargas fuertes) en un factor reflejan una coocurrencia frecuente de los mismos en un texto, donde, en su conjunto, cumplen una función comunicativa o discursiva concreta.

La segunda parte del análisis factorial consistió en calcular puntuaciones factoriales para cada texto, y, seguidamente, para cada subregistro (en base a cada disciplina científica y género textual), para así poder distribuir los mismos, de acuerdo con su puntuación, en cada una de las cuatro dimensiones de variación. Dichas dimensiones se han identificado y caracterizado de la siguiente manera: Dimension 1 "Involved/persuasive vs. informational style" (estilo implicado/persuasivo vs. informativo); Dimensión 2 "Argumentative vs. descriptive focus" (enfoque argumentativo vs. descriptivo); Dimensión 3 "Elaborate vs. non-elaborate discourse" (discurso elaborado vs. no elaborado), y Dimensión 4 "Narrative vs. non-narrative discourse" (discurso narrativo vs. no narrativo), tal y como aparece detallado en la primera parte del capítulo 6 (análisis de datos). La segunda parte de dicho capítulo analiza la distribución de los diversos subregistros a lo largo de cada dimensión, tal y como se detalla a continuación.

Antes de nada, en lo que se refiere al primer objetivo de este estudio, que consiste en detectar variación y cambio lingüísticos entre las tres disciplinas científicas, podemos afirmar que se ha encontrado tanto variación como cambio. Respecto a la primera, se ha observado que la mayor diferencia entre las humanidades y las ciencias naturales se nota en la Dimensión 1 (estilo implicado/persuasivo vs. informativo), donde puede verse que la filosofía tiende a caracterizarse por un discurso más bien personalmente involucrado y persuasivo, mientras que la astronomía y las ciencias de la vida presentan un discurso más neutro y, por lo general, estrictamente informativo. En cuanto al cambio, puede observarse que, a medida que el tiempo avanza, las tres disciplinas se mueven hacia la media, sugiriendo una tendencia moderada a un discurso estándar menos marcado en lo referente al estilo comunicativo. A pesar de ello, también se ha observado que esta relación dicotómica entre la filosofía y las otras dos disciplinas con respecto a la Dimensión 1 refleja la situación del discurso científico inglés actual, tal y como se muestra en Gray (2011), sugiriendo así que la filosofía tenderá a conservar el carácter implicado/persuasivo como una característica discursiva propia con el paso del tiempo. Esto parece ir acorde con la naturaleza dialéctica de la filosofía, disciplina que trata una amplia variedad de temas trascendentales, a menudo a través del debate.

Por otro lado, la Dimensión 2 (enfoque argumentativo vs. descriptivo) aparta las ciencias de la vida hasta cierto punto como una disciplina caracterizada por un discurso fundamentalmente descriptivo. La filosofía, en cambio, presenta una distribución más o menos equilibrada de rasgos de ambos lados, mientras que el discurso astronómico oscila entre alta y moderadamente argumentativo en los siglos dieciocho y diecinueve, respectivamente. Esta dimensión destaca una diferencia considerable entre las ciencias de la vida y la astronomía, a pesar de que ambas puedan describirse como disciplinas observacionales. Tal y como se ha mostrado en el capítulo 2, las ciencias de la vida se basaron durante mucho tiempo en la catalogación de diversas especies animales y vegetales, así como en la descripción detallada de la organización y funcionamiento de sus órganos internos. En la astronomía, en cambio, el cielo se observa a través de las lentes de la matemática y la física, lo que posibilita tanto describir la posición de los cuerpos celestiales como predecir sus movimientos en base a cálculos precisos. Con todo, las tres disciplinas muestran una ligera tendencia hacia el polo “descriptivo” de la dimensión con el tiempo, especialmente a través de una gradual pérdida de los marcadores lógicos del discurso, algo aparente en el caso del registro astronómico y filosófico. Esta desaparición, o, por lo menos, disminución progresiva de las cadenas de causalidad parece sugerir que el discurso científico inglés del siglo dieciocho aún seguía el patrón escolástico de razonamiento lógico, algo que se irá abandonando paulatinamente a lo largo del siglo siguiente.

Con respecto a la Dimensión 3 (discurso elaborado vs. no-elaborado), se ha detectado un movimiento general similar, a lo largo de los dos siglos, hacia un discurso más elaborado, encabezado por la filosofía, que comienza con rasgos moderadamente elaborados en el siglo dieciocho, acentuándose éstos en el diecinueve. La astronomía, por su parte, pasa de ser un registro medianamente “no elaborado” a moderadamente “elaborado” con el tiempo, mientras que las ciencias de la vida presentan un discurso no elaborado en ambos siglos, con puntuaciones más próximas a la media en el diecinueve. Cabe mencionar que un rasgo clave del discurso elaborado es la nominalización, descrita por Halliday (1985, 1988) como la “metáfora gramatical” que tan a menudo se utiliza en el discurso científico inglés para hacer referencia a acciones y procesos naturales (por ejemplo, *illumination* (iluminación), *fluctuation* (fluctuación), etc.), o estados físicos (por ejemplo, *darkness* (oscuridad)), cuyo uso se demostró haber aumentado a lo largo de los siglos dieciocho y diecinueve en la astronomía (Bello 2014).

Por su parte, la notable presencia de nominalizaciones en el discurso filosófico inglés, ya en el siglo dieciocho, parece justificarse por las materias tratadas por la disciplina, soliendo ser éstas de un carácter altamente abstracto, tales como la moral y la justicia, la igualdad entre los sexos, la inmortalidad del alma, la causalidad, etc. (Nótese que las materias mencionadas son, gramaticalmente, nominalizaciones.) Así, el desarrollo del discurso filosófico inglés aparece íntimamente ligado a su aumento en abstracción y elaboración, llegando a ser altamente elaborado en el siglo diecinueve. Las ciencias de la vida, por el contrario, tratan de objetos concretos del mundo material, sean éstos animados o no, lo que invita un uso más sencillo del lenguaje a la hora de describir sus características y explicar su comportamiento, habiéndose aún así observado un relativo aumento de los tecnicismos propios de la disciplina hacia el final del período estudiado. Por otro lado, teniendo en cuenta que otra parte fundamental de un discurso elaborado son las frases preposicionales, los modificadores postnominales, o las así llamadas *pied-piping constructions* (cláusulas relativas encabezadas por una preposición), creando todos ellos oraciones subordinadas complejas con una gran cantidad de elementos incrustados, se ha observado que las elevadas puntuaciones del discurso filosófico en la Dimensión 2 coinciden con los hallazgos de Gray (2011: 117-118), según los cuales el registro filosófico inglés contemporáneo se caracteriza por una gran densidad estructural.

Finalmente, en lo que se refiere a la Dimensión 4 (discurso narrativo vs. no narrativo), la astronomía se sitúa en la cima del eje como la disciplina más narrativa, al contrario que la filosofía, que pasa a ser moderadamente a marcadamente no narrativa. El discurso de las ciencias de la vida, por su parte, tiene una puntuación negativa en el siglo dieciocho que pasa a positiva en el diecinueve. A pesar de las diferencias en cuanto a su posición, también aquí las tres disciplinas van encaminadas en una dirección, la narrativa. El caso de las ciencias de la vida y la astronomía del siglo diecinueve, que muestran puntuaciones positivas muy similares, parece sugerir un movimiento hacia un estilo narrativo estándar, lo que refleja a su vez tanto una continuada importancia del papel de los informes de experimentos y observaciones en la literatura de estas disciplinas, como una creciente popularización de la ciencia para que esta sea más asequible a un público menos restringido, sobre todo en ciertos géneros (véase más abajo). La filosofía, a su vez, mantiene un discurso poco narrativo a lo largo de los dos siglos, algo que podría justificarse con el hecho de que las materias que trata son de carácter universal y atemporal, tendiendo a usarse para ellas el presente. Así y todo, se ha observado que el discurso filosófico inglés también presenta rasgos narrativos en algunas muestras del siglo diecinueve, correspondiendo éstos a relatos de experiencias personales que sirven para ilustrar afirmaciones generales.

En cuanto al segundo objetivo de este estudio, que consistía en averiguar si había variación y cambio entre géneros textuales, esta tesis ha demostrado que éstos también tienen lugar. Empezando por la Dimensión 1, se ha observado que algunos géneros, tales como Textbook, Dictionary o Letter, son relativamente informativos, mientras que otros como Dialogue o Essay presentan un estilo implicado/persuasivo. Esta diferencia podría justificarse analizando el objeto comunicativo de estos géneros. Por ejemplo, tanto el diccionario como el libro de texto son géneros cuyo objetivo es informar al lector, sea estrictamente a través de definiciones, como sucede en el caso del primero, o mediante explicaciones más detalladas de conceptos o fenómenos en el caso del segundo. El diálogo, en cambio, es un género que transmite debates sobre diversas materias a través de una supuesta conversación entre dos participantes que intercambian opiniones y experiencias desde una perspectiva más personal, haciendo uso para ello de estrategias persuasivas para convencer al interlocutor. El ensayo, por su parte, se caracteriza como un género abierto, permitiendo que los autores expresen su visión particular sobre un tema, generalmente a través de un lenguaje que refleja una postura personal. En cuanto al género epistolar, se esperaba que presentase rasgos de implicación o persuasión, puesto que las cartas, en principio, también implican interacción personal. Sin embargo, la abundancia de rasgos informativos en la mayoría de éstas parece deberse a su naturaleza didáctica, así como a la disciplina científica a la que pertenecen (ciencias de la vida). El género Treatise presenta puntuaciones más moderadas, agrupadas alrededor de la media, lo que puede deberse a que los tratados abundan en las tres disciplinas y tratan una amplia variedad de temas.

En lo que se refiere al cambio entre géneros con respecto a la Dimensión 1, se ha observado que todos los géneros con la excepción de Essay y Letter (que desaparecerán progresivamente del discurso académico) tienden hacia un estándar relativamente impersonal y densamente informativo y al abandono de un discurso que, en el siglo dieciocho, tenía un carácter más implicado y personal. Este hallazgo coincide con las observaciones de Atkinson (1999) sobre la progresiva sustitución del enfoque en el sujeto (es decir, el científico autor de las observaciones o experimentos) por un enfoque en el objeto tratado (el experimento en sí) en los artículos científicos durante el período del inglés moderno tardío. Esta coincidencia, por su parte, sugiere que el fenómeno mencionado se extiende a otros géneros utilizados para transmitir la ciencia en este período, si bien se nota más en la astronomía y las ciencias de la vida, manteniendo la filosofía un enfoque más personal que las otras dos disciplinas a lo largo del tiempo.

En cuanto a la Dimensión 2, que mide el enfoque científico, algunos de los géneros que aparecen en polos opuestos en la Dimensión 1, tales como Dialogue, Dictionary y Textbook, se unen aquí como principalmente argumentativos todos ellos. Mientras que el diálogo y el libro de texto precisan de conectores discursivos para ordenar un discurso que presenta la información de una forma extremadamente densa, el primero transmite una batalla dialéctica sobre temas astronómicos y filosóficos que se expresa principalmente a través de razonamientos matemáticos y lógicos, respectivamente. El género Lecture, en cambio, si bien también tiene un carácter instructivo, presenta rasgos descriptivos, al igual que Letter y Treatise. Esto parece deberse a que los tres géneros pertenecen en su mayoría a las ciencias de la vida, caracterizadas como altamente descriptivas en el análisis por disciplinas. Los ensayos, por el contrario, no se mantienen constantes, presentando rasgos argumentativos en el siglo dieciocho y descriptivos en el diecinueve. Puesto que la mayoría de los ensayos se han encontrado en la disciplina filosófica, no parece sorprendente que reflejen la tendencia de esta última con respecto a esta dimensión.

En lo que concierne a la complejidad del discurso, reflejada en la Dimensión 3, la distribución de los géneros muestra que todos ellos se caracterizan por un discurso poco elaborado en el siglo dieciocho y relativamente elaborado en el diecinueve. Estos datos también apoyan los hallazgos de Atkinson (1999: 126-129) sobre la progresiva elaboración del discurso en los artículos científicos ingleses escritos a lo largo del período inglés moderno tardío, sugiriendo una vez más que este fenómeno tiene un carácter más general, pudiendo atribuirse a otros géneros textuales de este período, además del artículo. Asimismo, las elevadas puntuaciones de la mayor parte de la muestra del siglo diecinueve justifican la afirmación de Bello (2014: 322), según la cual las nominalizaciones se han consolidado gradualmente como un marcador propio del registro científico inglés. Finalmente, en cuanto a la dicotomía narrativo vs. no narrativo (Dimensión 4), ciertos géneros, tales como Article, muestran un comportamiento distinto al de las disciplinas científicas, yendo en la dirección contraria hacia un estándar no narrativo, coincidiendo este fenómeno también con las observaciones de Atkinson (1999: 144). La mayoría de los géneros, sin embargo, tales como Essay, Treatise, Textbook y Letter, muestran una tendencia distinta, lo que sugiere que la literatura científica didáctica aumenta sus rasgos narrativos a la vez que crece su accesibilidad hacia un publico menos restringido o especializado, especialmente en el caso de los tratados escritos por mujeres (Crespo 2016). Otra explicación para este fenómeno radica en que la mayoría de los tratados se encuentran en la muestra que pertenece al corpus de las ciencias de la vida, una disciplina que presenta una gran abundancia de rasgos narrativos en el siglo diecinueve.

Tal y como se resume en el apartado de Conclusiones, en este estudio se ha pretendido caracterizar el inglés de tres disciplinas científicas en los siglos dieciocho y diecinueve con respecto a cuatro dimensiones del lenguaje, siguiendo el modelo de Biber (1988). El análisis ha confirmado hallazgos previos sobre el inglés científico moderno tardío, demostrando una vez más su aumento en densidad de información, nivel técnico, abstracción y complejidad estructural a lo largo de los siglos dieciocho y diecinueve. Por otro lado, también se ha demostrado que esto no sucede de manera uniforme en las tres disciplinas y en los ocho géneros incluidos en la muestra analizada, revelando patrones de variación interna que reflejan una falta de estándar técnico cuya necesidad se defendía con tanto ímpetu por aquel entonces. Así y todo, esta tesis pretende ser una hipótesis sobre el estado del discurso científico inglés moderno tardío que puede utilizarse de base para un análisis factorial confirmatorio sobre una muestra que contenga más disciplinas científicas. El objetivo para un estudio futuro será comprobar si los patrones de variación y cambio aquí descritos se mantienen, así como detectar patrones nuevos que podrían salir a la luz.