# LEXICAL RICHNESS IN MODERN WOMEN WRITERS: EVIDENCE FROM THE *CORPUS OF HISTORY ENGLISH TEXTS**

Isabel Moskowich
MuStE Group
Universidade da Coruña

Abstract

This paper addresses the issue of lexical density and its popularity after the arrival of corpus linguistics and its methodology. In fact, this is now one of the most frequently used descriptive tools in the analysis of register and genre. Researchers have often trusted lexical density as it is quantifiable and measurable by applying a formula and this has made its use very popular both for scrutinising grammatical and lexical forms and their frequencies. Lexical richness is a related concept although it does not refer exactly to the same, This paper aims to examine lexical richness, understood as the degree of variety of terms used in texts written by women during the eighteenth and nineteenth centuries. To this end, I will analyse samples drawn from the *Corpus of English History Texts* (*CHET*) to see whether the communicative format (genre) of the sample has any influence on vocabulary in a discipline with discursive patterns that were not probably as standardised as those of other fields of knowledge.

Keywords: Lexical richness, late modern English, scientific writing, Coruña Corpus, women authors.

Resumen

Este artículo aborda el tema de la densidad léxica y su popularidad desde la llegada de la lingüística de corpus como metodología. De hecho, es una de las herramientas descriptivas que más se usan en el análisis de registros y géneros. Los investigadores han recurrido a ella a menudo, ya que es cuantificable y medible por medio de una fórmula y esto la ha hecho que se use muy a menudo tanto para el análisis de palabras gramaticales como léxicas. La idea de riqueza léxica es un concepto relacionado con el de densidad léxica aunque no son exactamente lo mismo. Este artículo se centrará en la riqueza léxica, entendida como el grado de variedad de términos usados en unos textos concretos, los escritos por mujeres en los siglos XVIII y XIX. Con tal fin, analizaré muestras extraídas del *Corpus of English History Texts* (*CHET*) para comprobar si los formatos comunicativos (géneros) influyen sobre el vocabulario en una disciplina cuyos patrones discursivos no estaban tan desarrollados como en otros campos del saber.

Palabras clave: riqueza léxica, inglés modern tardío, escritura científica, Coruña Corpus, autoras científicas.

## INTRODUCTION

The development of corpus linguistics has meant that lexical density is now one of the most frequently used descriptive tools in the analysis of register and genre. As a quantifiable parameter to which a particular formula can be applied, it can be employed to assess either functional or lexical words. A closely related concept is that of lexical richness, and although the two terms have been often used interchangeably, they are not necessarily the same. This paper aims to examine lexical richness, understood as the degree of variety of terms used in particular texts, and to do so in texts written by women during the eighteenth and nineteenth centuries. To this end, I will analyse samples drawn from the *Corpus of English History Texts* (*CHET*) to see whether genre has any influence on vocabulary in a discipline that might be considered less standardised than many others. In what follows, section 1 will briefly provide the context in which the texts under scrutiny were produced, as well as describing the social circumstances of women and the broader situation of science writing at the time. Section 2 will explore the idea of lexical richness and will present the initial hypothesis of the study, plus the methodology. A description of the corpus follows in section 3, after which the analysis and results are presented. Finally, section 5 will offer some concluding remarks.

## 1. SOME BACKGROUND ON SCIENCE IN LATE MODERN ENGLISH

Although today's scientific discourse is said to be highly conventionalised, this has not always been so. The structure *IMRD* (Introduction, Method, Results and Discussion) of research articles, which is found in experimental studies of all kinds, is often accompanied by particular linguistic features, including lexical ones, that convey specific communicative purposes (Swales; Biber and Finegan). Also, as Monaco (forthcoming) notes, "Another example [of conventionalised style] could be 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."

I have argued elsewhere (Moskowich, "Morfología flexiva" 625) that it is this tendency towards standardisation that characterises scientific writing at the beginning of the late Modern English period. It is at this time that authors in general start to prefer the use of the standard variety of the language, independently of their

* The research report herehas been funded by the Spanish Ministerio de Economía y Competitividad (MINECO), grant number FFI2013-42215-P. This grant is hereby gratefully acknowledged.

geographical origin, thus ignoring their own dialectal idiosyncrasies, and it is also the time when academics, grammarians and lexicographers try to "fix" and "correct" the English language. Yet this tendency in language in general may not have affected scientific writing to the same extent, since English had been adopted as a valid language for the transmission of knowledge only one century earlier, alongside the gradual abandonment of Latin for this purpose. Moreover, Crombie (95) notes "linguistic inertia" as "evidence of continuity with earlier forms of thought, whatever changes the requirements of successful scientific practice may have brought about." And this seems to have occurred in spite of Bacon's 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), he claimed 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 74). Contrary to Crombie's idea of linguistic inertia, others have claimed that the members of the Royal Society adopted Bacon "as their linguistic messiah" (Montgomery 75). Perhaps the truth is somewhere between these two points of view, and the unadorned style advocated by Bacon was to be seen more in syntactic structures than in the choice of particular lexical items, where the constraints of particular scientific disciplines were often determinant.

During the seventeenth to nineteenth centuries, fields of knowledge were conceived of very differently from the way they are today. In fact, Newton was not considered to be astronomer, mathematician or physicist (Monaco, forthcoming) but rather a *natural philosopher*, that is, someone who "had a breadth of comprehension, perceived analogies and other irregularities, derived rules that explain phenomena, and predicted the future", and who also combined "accuracy of observations", "precision of judgment", and "speculative curiosity" (McCormmach 17). Other disciplines, falling outside the experimental realm, were not considered science at all. History (or historiography) is not generally regarded a science at all and was certainly not regarded as such in the eighteenth and nineteenth centuries, when science was essentially associated with the observation and explanation of the natural world.

In the period under study here, the eighteenth and nineteenth centuries, women were excluded from this official world of "real" science. For this reason, the *Coruña Corpus of English Scientific Writing* contains eight female authors in the *Corpus of History English Texts* (*CHET*) whereas only two in *CETA* (*Corpus of English Texts on Astronomy*) (Moskowich *et al.*), reflecting the scientific, social and hence linguistic reality in Modern times.

Women were not readily accepted in the world of knowledge until very recently. Even Margaret Cavendish, frequently mentioned in the literature as an active member of seventeenth-century scientific circles together with her husband, was never received as a full member of any academic or scientific society. It was perhaps often the case that scientific circles were no more than the result of a fashionable habit of the high classes, with women also affected by this. Indeed, according to Crespo (103), there were two main reasons for women's participation in science: the fact that they and their families belonged to high social strata; that science at the time was a

non-institutionalised activity, often considered a hobby or part-time occupation and thus not something in itself serious or important. The women I have included for this study all enjoyed the kind of very favourable personal circumstances conducive to becoming history writers. Crespo (105) claims that "the fathers of our female writers normally occupied important social posts, being bankers, landowners, members of parliament or merchants interested to a certain extent in intellectual matters." This is exactly the case of one of the eighteenth-century authors under study here, Sarah Scott, whose background accurately matches this description. The case of Elizabeth Justice is different, but also one of educational privilege; her father was one of those people concerned with ladies' education and sent her to a boarding school, as well as providing her with a private tutor. Women in the nineteenth century saw some social changes regarding their position, but access to education was still for a small minority. Both Scott and Justice wrote texts of an instructional character, Scott's work being addressed to children, whereas Justice's travelogue included observations on everyday life, and Crespo (104) has argued that the latter presents significant differences from comparable works by contemporary male authors.

The first of the authors whose work will be considered from the nineteenth century, Mercy Otis Warren (1728-1814), did not receive any formal education until she attended a preparatory school with her brothers in Massachusetts. This was a pattern seen in other early American women writers from comfortable families, where male relatives encouraged studying and writing within the confines of the home (*ODNB*). By contrast, Lady Maria Callcott (1785-1842) attended school from very early and was an enthusiastic student, studying Latin, French and Italian, among other disciplines. Lucy Aikin (1781-1864) and Alice Cooke (1867-1940) are the only writers in the corpus recognised as historians by the *ODNB*. The latter also read extensively in French, Italian and Latin. Elizabeth Sewell (0815-1906), in turn, was first sent to a school when she was four and received formal education until she was fifteen, when she returned home to help with the education of her two younger sisters. This was a turning point in her life since it marked the beginning of her interest in the education of middle-class girls. We do not have any information about the early life and education of Martha Freer. As regards Alice Cooke, not only did she receive formal education in schools for girls but also attended Victoria University of Manchester.

Even with formal training and university degrees, women were generally excluded from official seats of knowledge and worked at the lower end of the scientific scale. They were not admitted to any of the institutions and societies founded throughout the seventieth and eighteenth centuries (Solsona i Pairó 86-87). Exclusion from institutionalised science served to reinforce these women's desire to spread knowledge to those who were similarly pushed into the educational background, and may have had an effect on the language they used to convey knowledge to their audience.

According to Crespo (105), most female authors share certain notable characteristics in their writings: the new empirical and observational approaches to science which provoked the use of expressions such as "I observ'd" (Justice xiv), and "I flatter myself, that my imparting to general curiosity what in my researches

I have been able to discover concerning it" (Scott xiii); positive references to the female sex (Margart Bryan, author included in *CETA*); allusions to the vices, virtues and religious morals of the time, in prefaces such as that in Justice's (1739) work. It is perhaps because women's claims to equality, including intellectual equality, were still questioned that many of them felt they could play a role in society by writing for the weakest (children and other women), this in an attempt to make their audience conscious in a subtle way of the real role women play in human development. It seems that such an orientation must have been accompanied by particular ways of using language, and in this sense their vocabulary may reflect the kind of intellectual richness which they were not otherwise allowed to show.

## 2. LEXICAL RICHNESS AND METHODOLOGY

While late Modern English is a period of lexical innovation often fostered by social and technological developments, my concern here is not with such innovation but with lexical variety or richness. My initial hypothesis is that texts contain fewer different types as they become more specialised, since some terms are always preferred as being appropriate to refer to particular extra-linguistic entities in specific domains. That is, lexical richness decreases over the course of time as vocabulary becomes more discipline-related. Although I agree with Smitterberg and Kytö (129) that genre is an indispensable parameter in historical linguistics, its role as a limiting factor will not be explored in the present study since my sample is a small one, certainly not sufficient to produce any definitive results in this respect. Besides, genre may not be a constraint if we consider that genres have not always been well delimited, as shown for the genre "letter" by Kytö and Romaine (213), where they found that substantial differences could be observed within the genre as a result of revisions made by authors: "professional letters can be highly informational and are often written and revised with care. Thus, they often show a greater degree of lexical variety and informational density compared with personal letters." We also know that the vocabulary of English increased from the eighteenth century onwards, especially with the introduction of new terms adopted from other languages arising from contact through colonisation and commerce. Such an increase did not stop in the following hundred years (Görlach 1999, 93), and indeed, "many of the new words in the *OED* from the Late Modern English period reflect recent advances in technology" (Tieken-Boon van Ostade 65). It is therefore my intention to look at whether specialised texts reflect the growth in vocabulary in general or, rather, if there is some kind of stagnation due to the adoption of history-related terminology imposed by genre. This will be done by assessing texts in search of the true extent of their lexical variety, density or richness.

Some discussion seems in order here, since, as mentioned above, lexical richness and lexical density are often used as interchangeable terms. The latter is very clearly defined in computational linguistics as the estimated measure of content per units, either functional or lexical, and is calculated according to a formula the results of which are used in discourse analysis as a parameter varying across registers

and genres. In general terms, the calculation establishes the proportion of lexical (nouns, adjectives, verbs, adverbs) tokens in relation to the total number of tokens in the text. However, while this measure provides a good portrait of informational density (that is, the proportion of lexical and functional words in the text), it does not really express the idea of variety as such, that is, the many different terms that can be found in a particular sample. For this, Peirce's 1906 type-token distinction, in which a descriptive class is distinguished from the elements that instantiate that class, is fundamental. In terms of linguistic variety, types (the class or lexeme) and not only tokens[1] (each use of a form) must be taken into account. Content analysis, in Krippendorff's (24-31) concept of the term, may also be said to be at stake in dealing with the analysis of vocabulary in its context, if we assume that the words mentioned most often (that is, with the greatest number of tokens) are those reflecting important concerns in the texts. There are also other ways to measure lexical richness, such as Brunet's Index (W), Honore's Statistic (R) and Type-Token Ratio (TTR). The latter has been seen to be unsatisfying in some studies on child language (Richards 1987) and has been said not to be directly related to what one would consider a rich language[2], yet it may be useful when dealing with samples of a similar size as the ones under study here, and hence it will be used for this study. However, TTR analysis will be complemented with typical frequencies using normalised figures, as well as with a detailed account of each text, a more typical approach in microscopic analyses.

## 3. CORPUS MATERIAL

The data for the present analysis has been taken from the *Corpus of History English Texts* (*CHET*), a beta version of one of the sub-corpus of the *Coruña Corpus of English Scientific Writing* (*CC*, henceforth). The samples me*et al*l the criteria already employed for the other subcorpora, that is, they are text samples of approximately 10,000words published between 1700 and 1900, all written directly in English so as to avoid any interference from other languages or mistakes derived from translation, and only one sample per author has been collected to avoid the abundance of any particular idiosyncratic linguistic features. Each sample is accompanied, as is the case with other subcorpora in the *CC*, by a file with extralinguistic information, relating to both the author and the text itself, including genre, date of publication and bibliographical information (for the text), the place where the author acquired her (scientific) linguistic habits,her age when the work was published,and her occupation (Moskowich 2012, 36).

[1] Ignoring Quine's (180) claims, *token* and *occurrence* will be used to refer to the same entity in this paper.

[2] Mike Scott, the developer of Word Smith Tools, claims that Shakespeare's TTR is certainly not high (http://www.lexically.net/downloads/version5/HTML/index.html?type_token_ratio_proc.htm).

*CHET* contains, at the moment of writing this paper,[3] 404, 511 words, split equally between texts published in the eighteenth and the nineteenth centuries (see Table 1 below).

| TABLE 1. WORDS PER CENTURY IN CHET | |
|---|---|
| Words in the 18th c. | 201,951 |
| Words in the 19th c. | 202,560 |
| total words in CHET | 404,511 |

Of the forty authors whose writings have been compiled in *CHET*, only eight are women. The total number of words of the texts under examination, then, is 81,578, representing 20.17% of the total. This seems to be in accordance with the social and historical context of women during the late Modern English period as described in section 2, especially so if we consider that only two of these eight women published their work during the eighteenth century, thus representing one quarter of all the material for this study.

Table 2 below provides details of the eight works under survey here, including the genre to which the texts have been ascribed in the compilation process.

| TABLE 2. WOMEN AUTHORS IN CHET | | | | |
|---|---|---|---|---|
| Date | Words | Author | Work Title | Genre/TT |
| 1739 | 10,005 | Justice, Elizabeth | Voyage to Ruſſia: describing the Laws, Manners, and Cuſtoms, of that great Empire, as govern'd, at this preſent, by that excellent Princeſs, the Czarina. Shewing the Beauty of her Palace, the Grandeur of her Courtiers, the Forms of Building at Petersburgh, and other Places: with ſeveral entertaining Adventures, that happened in the Paſſage by Sea, and Land. York: printed by Thomas Gent | Travelogue |
| 1762 | 10,114 | Scott, Sarah | The History of Mecklenburgh, from the Firſt Settlement of the Vandals in that Country, to the Present Time; including a Period of about Three Thouſand Years. London: printed for J. Newbery | Treatise |
| 1805 | 10,214 | Warren, Mercy Otis | History of the rise, progress and termination of the American revolution. Interspersed with Biographical, Political and Moral Obſervations. In three volumes. Vol. I. Boston: printed by Manning and Loring, for E. Larkin | Treatise |
| 1828 | 10,332 | Callcott, Maria | A Short history of Spain. In two volumes. Vol. II. London: John Murray | Treatise |
| 1833 | 10,013 | Aikin, Lucy | Memoirs of the Court of King Charles the First. In two volumes. Vol. I. London: printed for Longman, Rees, Orme, Brown, Green, and Longman | Treatise |

[3] The material is referred to as a beta version of *CHET* because some texts for the nineteenth century are still under revision, which might eventually alter the word counts for some samples slightly.

| | | | | |
|---|---|---|---|---|
| 1857 | 10,037 | Sewell, Elizabeth Missing | A first history of Greece. New York: printed by D. Appleton and company | Textbook |
| 1860 | 10,102 | Freer, Martha Walker | History of the reign of Henry IV. King of France and Navarre. London: printed by Hurst and Blackett, publishers | Treatise |
| 1893 | 10,761 | Cooke, Alice M. | The Settlement of the Cistercians in England. The English Historical Review, vol. 8, No. 32. (625-648). Oxford: Oxford UP | Article |

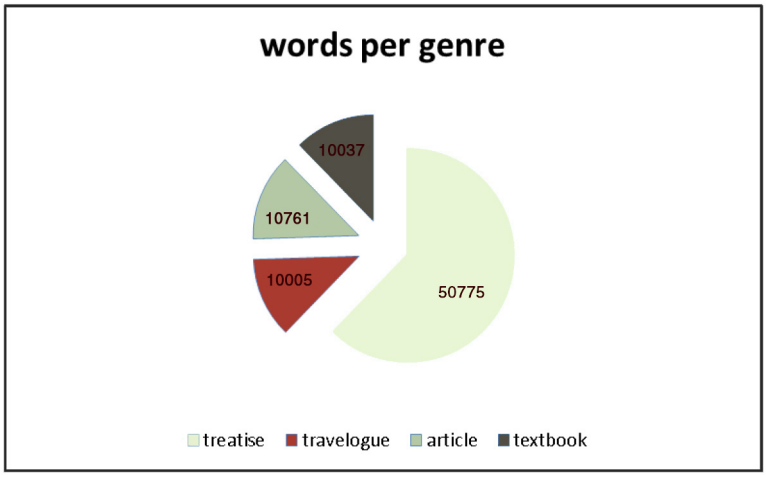
As can be seen, there is not much variety regarding genre or choice of text type[4] on the part of authors. It is not only a question of the popularity of particular genres at any point in time, but also that such choices may have been provoked by factors such as text reception. In other words, the fact that we have one travelogue and one article versus five treatises and a textbook may be in direct relation to the key role of women writers in the transmission of instruction (usually found in treatises) rather than in the advancement in knowledge (typically represented by articles) during the two centuries. Graph 1 below illustrates this distribution.

Curiously, the genres to be found at both endpoints of the period under survey are precisely those most representative of their time: travelogue was greatly in vogue as an identifiable written form in the eighteenth century but the term itself had almost collapsed and nobody uses it any more, whereas in the nineteenth century the article was blooming as the means of scientific communication *par excellence*, and treatises, which tended to offer fuller surveys of particular fields of knowledge, were also very numerous.

An analysis of these samples from the point of view of their lexical diversity will be presented in the following section, using frequency lists and type-token ratios (TTR) as tests of diversity in their vocabulary. Since this ratio is highly text-length dependent, in the sense that the longer a text is, the lower the TTR will automatically be (Arnaud, 1984), the samples contained in the *CC* have the advantage of all being of a similar size and are therefore easily comparable. However, frequencies will be normalised to 10,000 words to ensure optimal comparability.

After generating individual frequency lists with the Coruña Corpus Tool, the lists were exported to Microsoft Excel 2010. All different forms, that is, both content words and function words, have been considered for this survey. Some variability within forms (counted as different types) may occasionally arise not as a result of an author's decisions but by those of the editor or printer, for example in spelling alternations such as the use of <s> and <ſ> in *sun* and *ſun*. However, other differences, such as the use of abbreviations (*'em* for *them*, for instance), may indicate the author's wish to provide variety and thus make her writing a little more vivid, to avoid repetition, orto achieve a better prose rhythm in the case of texts intended to be read aloud in schools. Table 3 below shows the number of different forms

[4] The concepts of genre and text type will be used interchangeably here, in that the differences between them are not relevant for the purpose of this research.




Graph 1. Genres in women's history writing.

initially identified (Types raw) and those left after removing figures and editorial marks by compilers (Types cleaned) for each text.

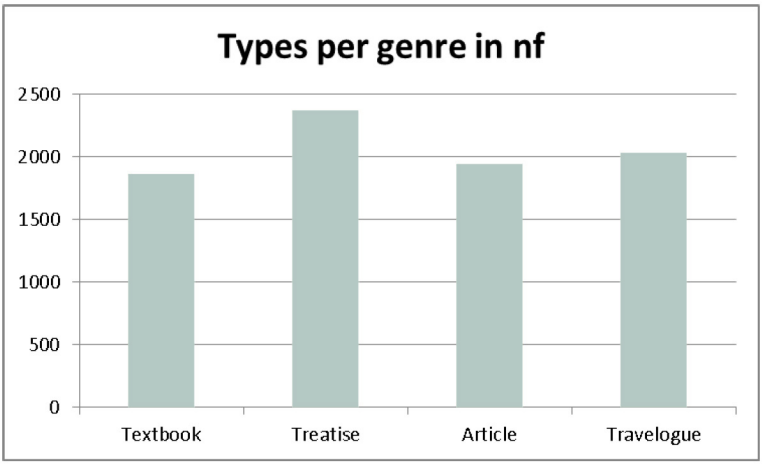
TABLE 3. TYPES FOR CONSIDERATION IN SAMPLES

| Author | Genre/TT | Words | Types_raw | Types_cleaned |
|---|---|---|---|---|
| Sewell, Elizabeth M. | Textbook | 10,037 | 1,870 | 1,861 |
| Scott, Sarah | Treatise | 10,114 | 1,996 | 1,981 |
| Justice, Elizabeth | Travelogue | 10,005 | 2,045 | 2,028 |
| Cooke, Alice M. | Article | 10,761 | 2,409 | 2,088 |
| Callcott, Maria /lady | Treatise | 10,332 | 2,390 | 2,358 |
| Warren, Mercy Otis | Treatise | 10,214 | 2,480 | 2,463 |
| Aikin, Lucy | Treatise | 10,013 | 2,659 | 2,603 |
| Freer, Martha Walker | Treatise | 10,102 | 2,668 | 2,610 |

The following section will provide an analysis of these data.

## 4. DATA ANALYSIS

Since the size of the corpus under analysis is small, the study requires a microscopic orientation. To this end, each text sample has been dealt with individually, following the methodology described in section 4 above, followed by some detail analysis.




Graph 2. Distribution of types per genre, in normalised frequencies.

The first thing to be observed (see Table 3 above) is that the number of types within each sample ranges from 1,861 (in 10,037 words) in Sewell's text to 2,610 (in 10,102 words) in the sample from Martha Freer. The difference in the number of types in these two samples does not seem to be caused by sample size, since all samples in the *CC* contain around 10,000 words. Time as a variable can also be disregarded, as it does not seem to have any direct influence on my data; only three years separate the publication of the two works containing the highest and lowest number of types (published in 1857 and 1860, respectively). So, where does the difference lie? Genre seems to be a determining factor, since the work containing the least varied vocabulary is a textbook and the one with the greatest use of different terms is a treatise. In fact, Table 3 seems to reveal a gradiance here which establishes some kind of correlation between genre and lexical diversity, thus: textbook, travelogue, article, treatise. This sequence is disrupted only by the sample from the text written by Sarah Scott in 1762 (*The History of Mecklenburgh*). The fact that this work had two editions in its very year of publication suggests it was very popular. Although a treatise, Scott's book is certainly close to its readers in style and content, offering anecdotes from the everyday lives of the historical characters she portrays, as the author herself recognises as one of her aims in the preface: "They all required, with the moſt ardent curioſity, for anecdoteſ concerning the houſe of Mecklenburh" (Scott, *The History of Mecklenburgh* ix)

In this sense, the work may be said to be an atypical type of treatise. So, with the exception of Scott's work, we can see that genre, more than any other variable, seems to be a plausible cause of the difference in lexical richness to be found in the samples under survey. Graph 2 provides a general overview of the distribution of vocabulary as used by female authors across genres:

Although sample size is very similar in all cases, types have been normalised (per 10,000 words) in order to confirm our first impressions of the data. In fact, we observe that the number of different terms contained in each of the four genres shows that these genres can be arranged in decreasing order, with treatises first (2,366.22 nf), followed by the travelogue by Justice (2,026.98 nf), article (1,940.34 nf) and textbook, with has the least rich vocabulary (1,860.11 nf). However, it is the type-token ratio that will yield the clearest picture of the vocabulary in the different genres.

The calculation of the TTR as set out in Table 4 confirms the idea that genres can be somehow sorted in decreasing order of lexical richness, and at first sight this seems to be in relation to the level of technicality of texts. Indeed, it might be said that genres are chosen precisely according to such technical level.

TABLE 4. TTR PER GENRE

| Genre | TTR |
|---|---|
| Textbook | 18.6011757 |
| Treatise | 23.6632201 |
| Article | 19.4034012 |
| Travelogue | 20.2698651 |

So, we have seen that normalisation of frequencies provides the same initial results as the TTR. However, given that there are relatively few women represented in the *CHET*, a more detailed analysis may be needed to be sure that none of the text extracts here is skewing the findings due to the linguistic idiosyncrasies of particular authors. Such idiosyncrasies might explain the fact that although treatises are usually highly technical texts addressed to members of the same epistemic community as the author, one particular treatise in our data is the second least lexically rich sample.

Görlach (2004) defines treatise as a "Discussion of a topic including some methodological issues", a definition that coincides broadly with those provided by the *OED* and other dictionaries. In fact, the *Merriam-Webster Dictionary* defines the term as "a systematic exposition or argument in writing including a methodical discussion of the facts and principles involved and conclusions reached <a *treatise* on higher education>." Such a definition, and its implications for the richness of the vocabulary employed, cannot be applied to the treatise by Sarah Scott, unless we simply consider this text an "Account" or a "tale (as proposed by the *Merriam-Webster Dictionary* as an obsolete meaning of the term), a description which seems like a far more adequate one of *The History of Mecklenburgh, from the Firſt Settlement of the Vandals in that Country, to the Present Time*.

It is surprising that the difference in the number of types for the genres article (in principle a genre used to address specialised readership within the scientific community) and travelogue (a piece intended for a wider audience) is very small: article has a TTR of 19.40% and travelogue one of 20.26%. Following this trend, textbook is the category expected to show least variety. Textbooks are conceived of as a means of instruction, and therefore they must repeat ideas and words so that

readers assimilate them. At the same time, we will see that in our sample the book is clearly addressed to children, which may also imply that vocabulary has been deliberately restricted.

For the third level of analysis, I intend to consider individual texts. In each of them I have scrutinised the use of *hapax legomena* and of those types with four tokens or fewer, in that these are indexical of lexical variety. I have also looked into the most abundant forms, both functional and content words, since these will provide important information as to the communicative intentions of the texts. Finally, I have tried to establish some kind of relationship between genre and vocabulary, looking for possible constraints imposed on the latter by the former. Table 4 below offers a general overview of the lexical structure of the samples. In it we can see the genre or text type to which the work belongs, the number of types contained in the sample (those labelled "cleaned" in Table 5, that is, excluding numbers and editorial marks by the corpus compilers) and the *hapax legomena* in each text extract.

TABLE 5. USE OF *HAPAX LEGOMENA* IN TEXTS BY WOMEN FROM *CHET*

| Words | Author | Genre/TT | Types cleaned | Hapax Legomena |
|---|---|---|---|---|
| 10,037 | Sewell, Elizabeth M. | Textbook | 1,861 | 975 |
| 10,114 | Scott, Sarah | Treatise | 1,981 | 1,087 |
| 10,005 | Justice, Elizabeth | Travelogue | 2,028 | 1,206 |
| 10,761 | Cooke, Alice M. | Article | 2,088 | 1,203 |
| 10,332 | Callcott, Maria /lady | Treatise | 2,358 | 1,370 |
| 10,214 | Warren, Mercy Otis | Treatise | 2,463 | 1,451 |
| 10,013 | Aikin, Lucy | Treatise | 2,603 | 1,659 |
| 10,102 | Freer, Martha Walker | Treatise | 2,610 | 1,622 |

The first thing to note here is that in all the texts more than half the words are *hapax legomena*, that is, they are used only once. This ranges from 52.4% in Sewell's textbook to almost 64% in Aikin's treatise. This itself serves as an answer to one of my research questions, namely, the degree of dependence between genre and lexical richness, assuming that this is twofold. On the one hand, more technical sorts of texts (and the more so as textual categories become more standardised) should exhibit greater lexical variety. On the other hand, this same standardisation would probably cause a preference for certain terms over others, since we know that domains or disciplines tend to have a typical vocabulary associated with them (Coxhead and Nation 5-7). In the case of the eight texts under study here, we can see that the assumption of technicality seems to be true, in that the textbook tends to use fewer types and repeat them more often, whereas the treatise does exactly the opposite. However, the intermediate numbers indicate that this is not absolutely true, since Scott's travelogue (with 59.5%) contains more *hapax legomena* than the article by Alice Cooke (57.6%) and also more than the three treatises by Sarah Scott, Maria Callcott and Mercy Warren (with 54.1%, 58.1% and 58.9%, respectively).

According to these TTRs, neither discipline nor genre can be identified as the cause for such irregular distribution of *hapax legomena*.

The specific kind of vocabulary found in a text may also be illuminating, and thus we will turn to this now, going from those samples containing the greater number of different types to those with the smaller. In 1857, Sewell wrote a textbook on the history of Greece, and mentioned in the preface that her intention was to provide children with an easy and understandable history of Greece. The genre, then, may have demanded a relatively greater than average use of repetition, and thus less variety. It seems that this might explain why her sample is the least rich lexically. In fact, the thirty three most frequently repeated words which she uses are function words such as *the*, *to*, *of* and *and*. The frequency list generated by the Coruña Corpus Tool shows that the most frequent content words are *Athens* (39 tokens), *Cyrus* (38 tokens), *Athenians* (34 tokens) and *Alcibiades* (33 tokens), all proper names. The text may thus reflect not only the constraints of subject matter, but also of the readership and pedagogical tendencies of a time where history was conceived of as a succession of deeds carried out by particular individuals rather than a conjunction of social or economic forces and their interpretation. In fact, although we have seen above that Sewell's *hapax legomena* are the lowest, these include very "exotic" items such as *Aeolis* as well as very common ones such as *weak*, *repeat*, *buy*, *leave* and *stone*. Also, there are few types ranging from 2 to 4 tokens, which seems to indicate that her vocabulary is also less varied than that of the other authors considered.

Sarah Scott's treatise, contrary to what we have just seen, does not have many repeated function words. In fact, there are only 18 types for functional words such as *the*, *for*, *a* and *in*, and immediately after these the frequency list shows the first content words, *king* (72 occurrences), *Mecklenburgh* (65), *Albert* (55) and *Duke* (50). Again, all these highly frequent content words make reference to particular characters or places, in accordance with the subject matter and following the trends of the discipline at that time in Europe.

The article by Alice Cooke came out in 1893, a moment at which this particular genre was blooming as an essential part of scientific dialogue. The frequency list here shows that the 20 most frequent terms are function words and the most frequent content word is the noun *order* (56 tokens). At first sight, this may suggest that Cooke is different from the other authors we have seen inasmuch as she is not making constant reference to particular concrete entities or people but rather to an abstract entity, in this case perhaps notions of social order. However, the second most frequent word in Cooke's vocabulary is *Cistersian* (51 tokens), hence we see that *order* is simply the collective noun for groups of religious people. It is worth mentioning that the third most frequent content word, with 43 occurrences, is the French term *citeaux*, no doubt in connection with the subject matter of the work. As shown in Table 5 above, this sample has the lowest *hapax legomena* count (1,203) after the textbook. If we assume that the level of technicality in a text is somehow related to the amount of single occurrences of terms, in that this increases lexical richness, how can we account for such a low *hapax legomena* count in this text, when we consider articles to be a genre typical of nineteenth-century scientific communication? Maybe one of the arguments mentioned earlier should be reconsidered

at this point: that specialised genres and specialised disciplines tend to create their own dominion-specific vocabulary, thus avoiding the use of the kind of linguistic resources that would naturally lead to increased variety.[5] The hypothesis of a relation betweenthe number of *hapax legomena* and specialised genre on the one hand, and subject matter constraints on the other, is reinforced by the analysis of the *hapax legomena* in Cooke's text, where we discover that many such items are place names or very common terms such as *useful* or *speak* which occur only once in the whole sample. This seems to confirm that she tends to repeat more technical terms because there are no synonyms for them. Besides this, there is an abundance of terms used only twice (335 types), three times (165 types) and four times (86 types), including in the latter group words such as *then* and *themselves*. This analysis seems to show that the author makes use of synonyms for non-specific words where no important shade of meaning will be lost, yet repeats those terms which are domain-specific, since the new trends of scientific and academic writing in the nineteenth century demands the use of a specialised lexicon associated with specialised, technical texts.

Among the *hapax legomena* employed by Callcott in her treatise, we find some terms that are not so unusual or specialised. Such is the case of *stirred*, *accusers* or *survey*. Very striking is the case of *across*, with one single use in all the 10,000-word sample. Callcott's richness of vocabulary is to be also seen in the fact that many terms (430 in all) are used only twice in the whole extract. Some of them are loans (*vizir*) and their exoticism may explain their infrequent use, whereas others are common, familiar English words such as *used*. Their scarcity may therefore be accounted for by the desire on the part of the writer to avoid repetition, as was the case with Cooke, except when this is strictly necessary, that is, with technical terms when there is no synonym available. As for the content word Callcott uses most frequently, this is *Ferdinand* (following 25 types on the frequency list representing functional categories). *Ferdinand* appears 53 times and is followed by *King* with 46 tokens, and is a clear illustration of the text's subject matter and arguments.

In Warren's treatise, the most frequent content word is *general* (with 46 tokens) following the 23 more frequent (functional) types such as *the* and *of*. *General* is followed by *British* (27 tokens), *army* (25), *Rome* (24) and *inhabitants* (also 24). Once more we see subject matter determining lexicon choice, but on this occasion the most frequent terms denote collective rather than individual referents. The work itself, *History of the rise, progress and termination of the American revolution. Interspersed with Biographical, Political and Moral Obfervations*, deals with a topic where collective entities are as important as individual heroes, and this may determine vocabulary choice. As a treatise, it contains a high proportion of *hapax legomena* (451) and also of other types not frequently used. Thus, there are many types with

[5] Let us consider our own practice when writing papers on linguistics. We find that words that have synonyms in general language use tend to lose such synonymy under the demands of precise scientific expression. Thus, in the current paper I cannot easily resort to the use of the various synonyms for *type* in non-specialised English, such as *class*, *sort* or *kind* when expressing *type* in its discrete linguistic sense.

only two tokens (4,444), with 3 tokens (190) and with 4 tokens (91). Among the *hapax legomena* used by this North American writer, we can find very common words such as *according* and *commerce*, as well as others such as *allure* and *exemption*.

My analysis of Elizabeth Justice's 1739 travelogue *Voyage to Russia* reveals that she does not use many *hapax legomena*, and hence her writing is not as rich as those of others from a lexical point of view. In fact, many of the terms we find only once are spelling variants of other forms (*'tis*, *'twould*) and in fact might have been produced not by the author herself but by the editor or the printer. On the other hand, there are certain types with many repetitions (function words such as *the* or *of*). This extract is particularly illuminating in that the frequency list shows that the eighth most frequent type is not a function word[6], but the personal pronoun *I*, with 147 tokens. This is evidence, once more, of the power of genre and subject matter on vocabulary choice. *Voyage to Russia* is a travelogue, and in this sense is almost a diary, a narration where the author is present at all times. In turn, subject matter may also have some influence on genre selection andtone, and consequently on vocabulary. Following the pronoun *I*, *great* (with 42 tokens) and *place* (with 33 tokens) are the two most frequent content words, both of which have a relatively vague meaning. As for the terms that are not often repeated throughout the sample, we should mention that 296 types have only 2 tokens, 165 types have only 3 tokens and 94 types are used only on four occasions. It is striking that, this being a travelogue, one of the less frequent types in the text is the word *road*, used only once in the whole sample.

The treatise by Martha Freer, published in 1860, contains 1,622 terms that are used only once, a very high number of *hapax legomena*. These include extremely common words one would expect to find frequently in a history book, such as *year,* and loanwords from French (*étudier*). Also abundant are types with less than four occurrences. Thus, the sample contains 435 different types with only two tokens, 189 with 3 and 95 terms that are used only four times. These figures point at a desire on the part of the author to avoid repetition. At the other end of the frequency list here, among the types with most tokens, the ten first words are function words, as expected. Immediately after these, *King* is the first content word with the highest level of use, appearing 105 times, and we also find, though with far less use, *Henry* (56 tokens), *Duke* (55) and *Majesty* (with 35 tokens), a vocabulary clearly focused on particular individuals in clear opposition to the sort of lexicon denoting collective referents noted in Warren's work. However, for a study of the contrastive discursive patterns denoting social or political tendencies in American and European history books, a larger corpus would be needed.

Finally, the richest text, in terms of containing the highest number of types, is *Memoirs of the Court of King Charles the First*, by Lucy Aikin. The first twenty-

[6] It is generally accepted that the first ten lines (therefore, terms) of a frequency list generated by a concordancer can be disregarded for the analysis of content, in that they will normally be function words.

three most frequent forms are all functional forms, and, not surprisingly, these are followed by the content word *King*, thus indicating that subject matter plays an important role in vocabulary choices here. However, we must not forget that two other determining factors are to be found in the work of this author. On the one hand, together with Cooke, she is one of the only two women in my material recognised as a historian by the *ODNB*, and as such she may be addressing a specialised readership and therefore resorting to a more varied vocabulary, the shades of meaning she believes such readership can grasp. On the other hand, the sample of her writing here, despite its title, is a treatise, and genre has also been seen to be a determining factor. So, these two elements, and not just subject matter, may also have some kind of influence on her choice of words.

## 5. CONCLUSIONS

According to the analysis of the *Oxford English Corpus* carried out jointly by *Oxford Online* and the*Oxford English Dictionary*, *the* is the most common word in the English language, and the same is true of my material. All the texts analysed, not surprisingly, are characterised by having functional words as the most frequently found tokens, *the* being the first in all cases. However, there are some differences among them. When counting content words in decreasing order of frequency we can see that they appear with different distributions, and that these may be directly related to subjectmatter as well as to genre.

As Tse and Hyland (177) claim, the use of community discourses helps speakers become members of social groups, defining them in relation to others. They go on to argue that institutional contexts privilege certain ways of making meanings, and it is in this sense that I have used the words "specialised vocabulary" and "domain-specific vocabulary." In a way, using a particular set of words helps create some kind of professional identity. But the women whose works I have studied here, although discipline-insiders, operated outside institutionalised circles due to the circumstances in which they lived in the late Modern English period. Tse and Hyland also claim that discipline is an important source of variation (179) and this is why I have chosen to study one single discipline and to explore what happens within it. Language choice seems to be heavily influenced by discipline and subject matter much more than by gender, and thus I have chosen not to compare gender-related language but language, vocabulary in particular, and to do so within the discipline of history at its very beginnings as an independent field of knowledge, after the emergence of Empiricism, once science had stopped being a *totum revolutum* and when different branches with their own rules began to appear.

My analysis seems to indicate that the language of these female authors is influenced not only by the genres they are using, but that these are chosen precisely because of the writers' intended readerships. Following the tendency noted by Jücker and Kopaczyk (2), I believe that the use of language by authors must not be considered in isolation but rather through considering it within its community and social-cultural context. In this sense, the authors here may have been constrained

quite severely when writing by the thought of whom they were addressing. In other words, the choice of vocabulary is conditioned by subjectmatter and genre and the latter, in turn, may be influenced by the potential or expected readership, since language use depends on the social and historical context of speakers constituting a particular epistemic community.

Recibido: 15-12-2015
Aceptado: 14-2-2016

## WORKS CITED

Arnaud, Pierre J.L. "The Lexical Richness of L2 Written Productions and the Validity of Vocabulary Tests." *Practice and problems in language testing: papers from the International Symposium on Language Testing.* Eds. Terry Culhane, Christine Klein-Bradley and Douglas K. Stevenson. Colchester: University of Essex, 1984. 14-28. Print.

Biber, Douglas and Finegan, Edward. "Intra-textual variation within medical research articles." *Variation in English: Multi-dimensional Studies*. Eds. Susan Conrad, and Douglas Biber. London: Routlege. 2001. 108-137. Print.

Coxhead, Averil and Nation, Paul. "The specialised vocabulary of English for academic purposes." *Perspectives on English for academic purposes.* Eds. John Flowerdew, and Matthew Peacock. Cambridge: Cambridge UP. 2001. 252. Print.

Crespo, Begoña. "Women writing science in the eighteenth century: a preliminary approach to their language in use." *Anglica, An International Journal of English Studies* 24.2 (2015): 103-128. Print.

Crombie, Alistair C. *Augustine to Galileo: The History of Science A.D. 400-1650*. London: Penguin. 1969. Print.

Görlach, Manfred. *English in Nineteenth-Century England. An Introduction*. Cambridge: Cambridge UP. 1999. Print.

—— *Text Types and the History of English*. Berlin: Mouton de Gruter. 2004. Print.

Jücker, Andreas H. and Kopaczyk, Joanna. "Communities of practice as a locus of language change." *Communities of Practice in the History of English*. Eds. Joanna Kopaczyk and Andreas H. Jücker. Amsterdam: Benjamins. 2013. 1-16.

Krippendorff, Klaus. *Content Analysis: An Introduction to its Methodology*.Beverly Hills, CA: Sage.1980. Print.

Kytö, Merja and Romaine, Suzanne. "Adjective Comparison in Nineteenth-century English." *Nineteenth-century English: Stability and Change*. Eds. Merja Kytö, Mats Rydén, and Erik Smitterberg. Cambridge: Cambridge UP. 2006. 194-214. Print.

McCormmach, Russell. *Speculative Truth : Henry Cavendish, Natural Philosophy, and the Rise of Theoretical Science*.Oxford: Oxford UP. 2004. Print.

Merriam-Webster Dictionary. http://www.merriam-webster.com/. Retrieved 4 October 2015.

Monaco, Leida Maria. A Multidimensional Analysis of Late Modern English Scientific Texts from the Coruña Corpus. Universidade da Coruña. Unp. PhD.

Montgomery, Martin. *An Introduction to Language and Society*. London: Routledge. 1995. Print.

Moskowich, Isabel. "Morfología flexiva del inglés moderno." *Lingüística histórica inglesa*. Eds. Isabel, and Javier Martín Arista. Barcelona: Ariel. 2001. 624-654. Print.

—— "CETA as a tool for the study of modern astronomy in English." *Astronomy 'playne and simple'. The writing of science between 1700 and 1900*. Ed. Isabel Moskowich & Begoña Crespo. Amsterdam: John Benjamins. 2012. 35-56.

—— *et al*., comps. *Corpus of English Texts on Astronomy*. Amsterdam: John Benjamins.2012. Print.

ODNB. *Oxford Dictionary of National Biography*. http://www.oxforddnb.com/.

"The Oxford English Corpus: Facts about language." Web. http://www.oxforddictionaries.com/words/the-oec-facts-about-the-language. Retrieved 17 november 2015.

Oxford English Dictionary. Web. http://www.oed.com/.

Peirce, Charles Sanders. "Prolegomena to an apology for pragmaticism." *The Monist* 16.4 (1906): 492-546. Print.

Quine, Wullard Van Orman. *Word and object*. Cambridge, MA: The MIT Press.1960. Print.

Richards, Brian. "Type/Token Ratios: what do they really tell us?" *Child Language* 14 (1987): 201-209. Print.

Smitterberg, Erik and Kytö, Merja. "English Genres in Diachronic Corpus Linguistics." *From Clerks to Corpora: essays on the English Language Yesterday and Today*. Eds. Philip Shaw, Britt Erman, Gunnel Melchers and Peter Sundkvist. Stockholm: Stockholm UP. 2015. 117-133. Print. DOI: http://dx.doi.org/10.16993/bab.g License: CC-BY.

Solsona i Pairó, Nuria. *Mujeres científicas de todos los tiempos*. Madrid: Talasa Ediciones.1997. Print.

Swales, John. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge UP. 1990. Print.

Tieken-Boon van Ostade, Ingrid. *Introduction to Late Modern English*. Edinburgh: Edinburgh UP. 2009. Print.

Tse, Polly and Hyland, Ken. "Gender and discipline: Metadiscourse variation in academic book reviews." *Academic discourse across Disciplines*. Ed. Ken Hyland & Marina Bondi. Bern: Peter Lang. 2006. 177-202. Print.