EPiC Series in Language and Linguistics
Volume 1, 2016, Pages 351–360
CILC2016. 8th International Conference on Corpus Linguistics

"it is proper ſubſerviently, to inquire into the nature of experimental chemiſtry": Difficulties in reconciling discipline-based characteristics and compilation criteria in the selection of samples for *CECheT*.

Luis Miguel Puente Castelo[1*] and Leida Maria Monaco[1]
[1] Universidade da Coruña.
luis.pcastelo@udc.es, leidamaria.monaco@udc.es

Abstract

This paper discusses the compilation of the beta version of *CECheT*, the subcorpus devoted to Chemistry in the *Coruña Corpus of English Scientific Writing*, and reflects on the difficulties faced during this process and how they were overcome. The historical context of science will be examined, particularly that of chemistry, and how this affects the process of compilation.

Attention will also be paid to the compilation criteria used in the whole *Coruña Corpus*, including those regarding the appropriateness of authors and text samples (Moskowich, 2012), and how these criteria have been applied to the compilation of *CECheT* in order to make it representative of the practices of the discipline at the time.

Finally, the corpus will be described briefly, looking at a series of parameters: the topics of the texts, the size of samples, their chronological distribution, as well as the geographical origin and sex of the authors represented.

1 Introduction

The *Corpus of English Chemistry Texts* (henceforth, *CECheT*), part of the *Coruña Corpus of English Scientific Writing*, has been being compiled by the *Research Group on Multidimensional Corpus-Based Studies in English* (*MUSTE*) at the Universidade da Coruña, beginning in 2014. The aim of this paper is to discuss the process of compilation and selection of samples in *CECheT*, which is now in its first beta version. On the one hand, it will focus on the difficulties faced during the process of selecting a set

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

of samples representative of the language used in chemistry writing in the period, and on how these difficulties were overcome. At the same time, it will also present a first parameter-based description of the subcorpus.

The paper will be divided into seven sections. After the introduction, Section 2 will set out the general design of the *Coruña Corpus*, of which *CECheT* is a part, and its general compilation criteria will be explained in Section 3. Following this, Section 4 will review the situation of Chemistry during the eighteenth and nineteenth centuries, and Section 5 will analyse the difficulties in reconciling the general criteria presented in Sections 2 and 3 and discipline-specific aspects outlined in Section 4. Finally, Section 6 will examine *CECheT* through looking at a series of parameters, namely the distribution of samples over time, their topics, and the sex and origin of their authors, in order to provide a first description of the corpus. A short conclusion will be provided in Section 7.

2 The *Coruña Corpus*

The *Coruña Corpus* is a "purpose-built electronic corpus conceived of as a resource for the study of scientific writing in English" (Moskowich, 2012, p. 35), currently under compilation by the *Research Group on Multidimensional Corpus-Based Studies in English* (*MUSTE*) at the Universidade da Coruña. This corpus allows research at all linguistic levels except phonology. In what follows, the *Coruña Corpus* is analysed in detail, focusing on three main aspects: its general structure, its size, and the time span it covers.

2.1 General structure

The *Coruña Corpus* contains samples of scientific texts from the Late Modern English period (here understood as comprising the eighteenth and nineteenth centuries), and consists of several twin subcorpora, each dealing with a particular field of knowledge but all designed and compiled under the same principles.

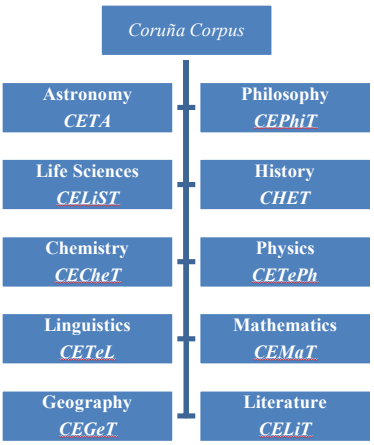

Figure 1: Subcorpora of the *Coruña Corpus*

As can be seen in Figure 1 above, the *Coruña Corpus* has been initially designed as a ten-subcorpora project. Of these ten subcorpora, the subcorpus on astronomy, *CETA*, was published in 2012 with John Benjamins (Moskowich & Crespo, 2012), along with a book containing a collection of pilot studies, while the subcorpus on philosophy, *CEPhiT*, has recently come out, also together with an accompanying

book (Moskowich, Camiña-Rioboo, Lareo, & Crespo, 2016), in the same format as *CETA*. Work is underway, at different stages of development, in the subcorpora on life sciences, history and chemistry; meanwhile, the remaining subcorpora (dealing with physics, mathematics, linguistics, geography, and literature) are still in initial stages of development.

2.2 Size

As already explained, all subcorpora present the same structure. Samples in the *Coruña Corpus* are approximately 10,000 words long, and two samples are selected per decade in each subcorpus, thus totalling *c*.20,000 words per decade and discipline (and, consequently, *c*.200,000 words per century and discipline, and *c*.400,000 words per subcorpus). This would make the whole *Coruña Corpus*, once finished, approximately 4,000,000 words long, which it can argued makes it sufficiently large and varied to consider the corpus representative of general scientific writing of the period.

The 10,000-word sample size is not arbitrary, but rather is the result of a very conscious selection. During the process of the general design of the *Coruña Corpus*, the compilers took into consideration Biber's (1993) position, in which he argued that, in order to study variation in scientific register, 1,000-word samples should be sufficient. However, the compilers of the *Coruña Corpus* decided, *contra* Biber, to select 10,000-word samples instead, as they considered that the scientific register was less standardised during the period under study than it is today, and consequently 1,000-word samples would not provide a good representation of the register. Moreover, given the relative scarcity of valid texts, the use of 1,000-word samples would make it almost impossible to achieve a sufficiently large corpus. On the other hand, samples have been selected in such a way that they cover all sections of texts (introductions, methods, results, discussions, conclusions…), thus avoiding the danger of selecting "arbitrarily cut-out chunks" (instead of full texts) as noted by Claridge *et al*. (1999).

2.3 Timespan

The *Coruña Corpus* covers the eighteenth and the nineteenth centuries, coinciding with a crucial period in the development of science and scientific writing. The period is delimited by two important scientific breakthroughs, effectively acting as chronological bookends. The first of these – the start of the eighteenth century – is the moment at which the long process of change in science, having begun in the late sixteenth and early seventeenth centuries, culminates in the final demise of Scholasticism and its substitution by new scientific paradigms. This coincides with the popularisation of Newton's ideas on gravity, which produced a major breakthrough in physics, providing the basis for much of the scientific research in the eighteenth and nineteenth centuries. On the other hand, the turn of the twentieth century also coincides with a major scientific advance: Einstein's 1905 paper on the Special Theory of Relativity, which opened up a new era of research.

At the linguistic level, this period corresponds approximately to the period referred to as Late Modern English. This is a period in which the English language experiments comparatively little change in phonetics, morphology, or syntax, but witnesses instead the appearance of the foundations for the development of a definite scientific register, with the emergence of a specific terminology and a distinctive genre, the research article. The end of the period coincides with a new trend towards the final configuration and consolidation of the scientific register as we know it today. For instance, at the 1897 International Congress of Mathematics Thomas Huxley argued that a new scientific style was needed, thus foreshadowing the emergence of the contemporary scientific register.

3 General compilation criteria

The main aim of the *Coruña Corpus* compilers has been to try to collect sets of samples which reflect the use of the language in each discipline (and in science as a whole) during the period as faithfully as possible, thus making the corpus representative (Biber, 1993) (McEnery & Wilson, 1996) (Biber, Conrad, & Reppen, 1998) (McEnery, Xiao, & Tono, 2006) of what was considered scientific writing during the period. This representativeness manifests itself in two different sets of criteria during the compilation of the corpus. The first of these comprises a series of standards concerning the eligibility of particular samples as examples of scientific writing, that is, requisites for a text to be considered for inclusion. The second set concerns the need for the set of samples to fulfil a series of parametric rules so that they are sufficiently varied, including examples of the different types of scientific writing during the period under study, and so that they constitute, as a set, a balanced representation of the register during the period.

3.1 Criteria concerning the eligibility of particular texts

The main criteria regarding the eligibility of samples concern their suitability as valid examples of scientific writing during the period, as well as more practical questions, such as the degree to which their computerisation is possible. In what follows, four of these criteria are described.

The first criterion is that only scientific texts which have been written, edited, and published can be selected. The exclusion of oral texts is obvious, as it is evident that it is impossible to obtain oral data from most of the period (although published transcriptions of lectures have been included). This exclusion extends also to unpublished material or unedited manuscripts, among others. In addition, texts written in verse are discarded, as their inherent constraints imply a distorted use of language.

The second criterion is that only texts written by native speakers are included in the corpus. This is done in order to avoid both the possible influence of a writer's native language(s) on their English style, and the inclusion of possible distorted uses of English.

Thirdly, only texts written originally in English are eligible. This excludes translations, even if authors were native speakers of English and translated their work themselves. The aim here is to avoid interferences from the source language in the translation (at the time, mainly Latin).

Finally, first editions are preferred. This is because the time reference used in the *Coruña Corpus* is the year of publication of the sample. By preferring first editions, compilers ensure that the language in the sample represents the language of a period near the time reference selected, and avoid distorting the results by including samples representing an earlier variety of language. When first editions are not available, compilers select subsequent editions published within a thirty-year timespan starting from the publication of the first edition. This duration follows Kytö *et al.*'s (2000, p. 92) assumption that language change can be observed after thirty years, thus implying that samples within thirty years from their first publication would present only minor changes, yet not sufficiently important to be considered as unequivocal examples of language change.

3.2 Criteria concerning the balance of the set of samples

Apart from complying with the above criteria, during the compilation process each sample is also analysed in relation to all the other samples in the corpus, so that, when all samples are considered as a whole, the set of samples is balanced, representative of scientific writing, and includes examples of all the different types of scientific writing used during the period. In order to do so, only one work per author has been allowed in the whole *Coruña Corpus*, thus avoiding jeopardising representativeness by the over-representation of idiosyncratic varieties. Moreover, the selection of samples has been carefully parameterised, each sample being classified according to the parameters of the discipline and period of the text, its genre, the sex of the author and their origin.

However, the most important step in order to ensure this representativeness in any subcorpus of the *Coruña Corpus* is to study the history of the discipline in the period and to take into account the particular features which are characteristic of the discipline during the process of compilation, as described in the following section.

4 Chemistry during the eighteenth and nineteenth centuries

Chemistry developed relatively late as a discipline. In fact, there was no chemistry as such during the scholastic period. Instead, there were two prior disciplines, more or less independent of each other, which would slowly evolve into chemistry as it is understood nowadays. The first of these antecedents was "materia medica" or "medical miscellanea", a conglomerate of scientific and quasi-scientific practices related to the process of healing, and which included traces of what are now medicine, pharmacy and chemistry. The second was alchemy, a discipline which combined experimental science, magic, and philosophy, with the objective of achieving the transformation of matter; nevertheless, its processes influenced the development of experimental methods in the new scientific practices that would appear from the seventeenth century onwards.

The Enlightenment and the emergence of New Science entailed the beginning of the process of specialization in scientific practice, which continues today. Chemistry, however, became an individualised science later than most other disciplines; the first academic chair would not be established until 1727 and its first journals (Crell's *Chemisches Journal* and Lavoisier's *Annales de Chimie*) even later than that, during the final quarter of the eighteenth century (1778 and 1789 respectively). These late dates are perhaps surprising, since new scientific practice was characterized by a keen interest in the real world which extended to the composition of the matter.

4.1 Topics of chemical research during the period

During the first years of the eighteenth century, chemistry was still very much mixed with medicine and pharmacy. There are three notable types of writing on chemistry in this period: coursebooks, which reflect the interest of New Science in the dissemination of knowledge, and which would continue to appear throughout the period; *pharmacopoeiae* or collections of the description of pharmaceutical compounds and their applications; and writings examining the composition of medical remedies, particularly spa waters and curative salts. This would continue until 1787, when Lavoisier published his *Méthode de Nomenclature Chimique*, a publication which would revolutionize research in chemistry during the following century, giving rise to the typical methodological and procedural particularities of chemistry, and influencing the production of authors as profoundly as Newton's discovery of gravity had influenced physics in the previous century.

It is during this period that two debates stemming from Lavoisier's work began to take shape. On the one hand, there were important discussions about chemical nomenclature and its spelling, with different models which would continue to compete until the final part of the nineteenth century. On the other hand, and also influenced by the publication in 1805 of Dalton's proposals on the weight and combination of atoms, there were a number of debates about the character of the atom and the calculation of atomic weights. The controversy over this matter would last for more than fifty years, serving as a basis for the design of the periodic table, presented by Mendeleev in 1869.

Following the publication of the periodic table, and during the final decades of the period, several new elements were discovered which had been supposed to exist but had hitherto not been found, including caesium and rubidium. This coincided with a further major breakthrough, the discovery of radioactivity and its related properties, by Pierre and Marie Curie, which would greatly influence research during the twentieth century with the rise of nuclear chemistry and physics. At the same time, new interest developed in the uses of chemistry in everyday life, such as in agriculture and cleaning,

giving way to a new sort of research, specifically oriented to the exploration of the commercial applications of chemical compounds.

5 Difficulties faced during the process of compilation and application of the criteria to solve them

As explained above, the process of sample selection has been made in such a way so as to assure that the set of samples is representative of disciplinary practices at the time. In the case of *CECheT*, this selection process has been particularly difficult for the start of the time period. The three main reasons for this are now discussed.

First, as already mentioned, during the first decades of the eighteenth century chemistry was still not a definitively individualised discipline. This means not only that there is a general scarcity of samples during the period, but also, given that the limits between medicine, chemistry and pharmacy were not clear-cut at that time, it is difficult to decide whether a sample can indeed be classified as belonging to the field of chemistry. Bearing in mind the general criterion of the *Coruña Corpus*, under which all categories are classified according to the category to which they were considered to belong during the period in question, and not according to how they are viewed today, the way chosen to overcome these difficulties was to examine the opinion of the authors themselves, as expressed in the abstracts and titles of their works, as to the field and subject matter of their research. After this, a new criterion was adopted: samples would be regarded as part of the discipline of chemistry if they had a focus on the composition (rather than on the effects or applications) of the elements analysed, be they water, chemical elements, or remedies. This, obviously, led to the inclusion in *CECheT* of works which would not be considered part of the discipline (or, indeed, scientific) nowadays.

Second, the process of vernacularisation was slower in chemistry than in other disciplines, and thus works in Latin were common (if not the majority) during the first part of the eighteenth century. This, again, has led to a restriction in the number of possible samples. Moreover, it has also been necessary to discard a high number of English translations from Latin. These were particularly difficult to identify, as they were often not advertised as translations, in particular when the authors themselves translated their own texts.

A third and final problem in selecting samples from the start of the eighteenth century is the general lack of information about the authors of that period. The principles of compilation of the *Coruña Corpus* state that it is preferable to select authors "about whom we could find basic biographical information and hence whose linguistic habits we could infer" (Moskowich, 2012, p. 48), and this has led to the elimination of an important number of valid samples where no author information was available.

Regarding the criterion of the balance of the set of samples, it should be noted that in some cases balance has been sacrificed for the sake of representativeness. This has affected the parameters of text genre and sex of the authors. The presence of samples representing the different genres throughout the period is uneven: articles, for instance, are more prevalent in the nineteenth than in the eighteenth century, in that they only started to be used at the beginning of the period and during the nineteenth century underwent a notable evolution, developing into their present prominent position. Textbooks, on the other hand, enjoyed a greater presence in the eighteenth century, highlighting their key role in the dissemination of knowledge at the time[†]. Special care has been taken to include in *CECheT* one of what is perhaps the most characteristic genres of chemistry in English during the first part of the eighteenth century, that is, the very short articles published in the *Philosophical Transactions*. In this case, and even though it does not reach the 10,000-word limit for samples, one of these has been included *in toto*, ensuring that the total number of words for the decade approximates the 20,000-word mark.

[†] For more information on the process of generic classification in chemistry, see Moskowich & Crespo (this volume).

The balance of the samples regarding the parameter of sex of the author has been, perhaps, more complicated. The majority of samples of the *Coruña Corpus* are written by men, as was the case with science in general during the period under study, since women faced serious difficulties in gaining access to scientific knowledge and had to overcome very significant obstacles in becoming part of the community of scientists, both socially and professionally. Thus, despite the fact that every subcorpus of the *Coruña Corpus* includes texts written by women, the selection of such texts for *CECheT* has been particularly difficult in that the biographical information for most female chemists was inexistent, a situation which was found more often than with male chemists or with the women scientists in other subcorpora. Moreover, it was apparently also more common for women chemists to publish under pseudonym or anonymously, which makes the identification of works by female writers in the field of chemistry in that period even more problematical.

6 Description of CECheT

In this final section, *CECheT* will be described in terms of a series of parameters: the distribution of samples over time, the topics of the samples, and the sex of the authors and their origins.

CECheT includes forty-one samples, which are distributed as shown in Figure 2. All samples contain around 10,000 words, except two samples in the 1740s. These texts – which, as noted above, are shorter and have thus been included *in toto* – are "Reflexions Concerning the Virtues of Tar Water" (1744) by Humphrey Jackson and "A Discourse concerning the Usefulness of Thermometers in Chemical Experiments" (1746) by Cromwell Mortimer, the latter of which is a short article published in the *Philosophical Transactions*.

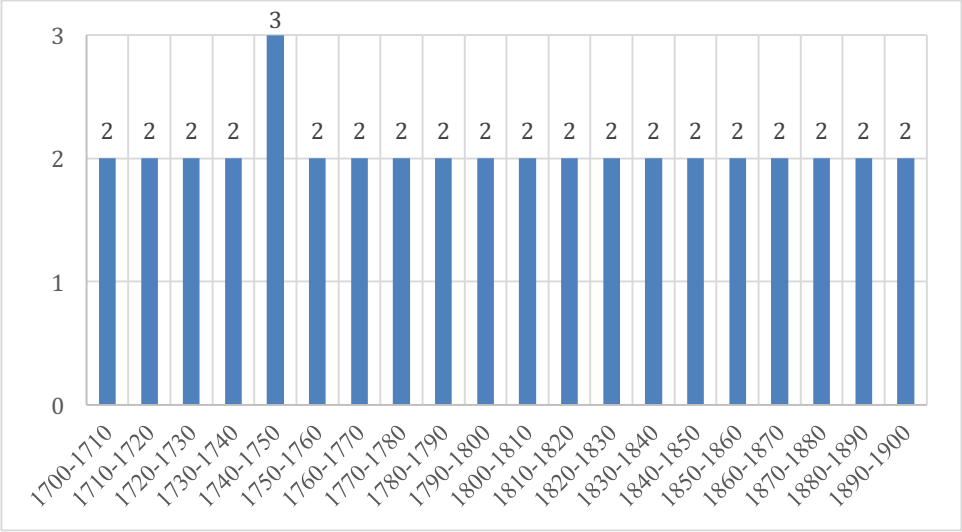

Figure 2: Distribution of samples per decade

Regarding the topics of the samples, and as shown in Figure 3 below, *CECheT* appears to be divided into two main periods. During the first period, there are a significant number of works (represented in different shades of purple) which analyse the composition of diverse elements, especially medicines, spa waters and salts, as well as a number of compendia of experiments (in orange) and works detailing the uses of diverse instruments, such as microscopes or thermometers (in grey). From the final quarter of the eighteenth century, concurrently with Lavoisier's ideas, new topics appear. These include

theoretical works (in red), including debates on the nature of the atom and chemical terminology, practical applications of chemistry in topics such as cooking and agriculture (in green), and popularisation works (in yellow). Coursebooks (in blue), reflecting their importance, appear throughout the whole period.

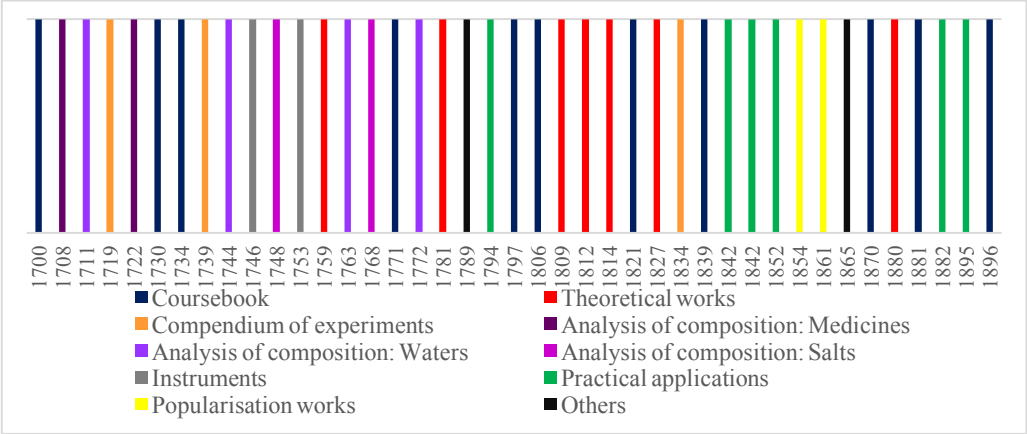

Figure 3: Thematic of the samples in *CECheT*

Regarding the sex of the authors, only three of the forty-one samples are written by women, this distribution shown in Figure 4 below. This represents a proportion of 7%, which the team of compilers agree to be more or less representative of the discipline in the period.

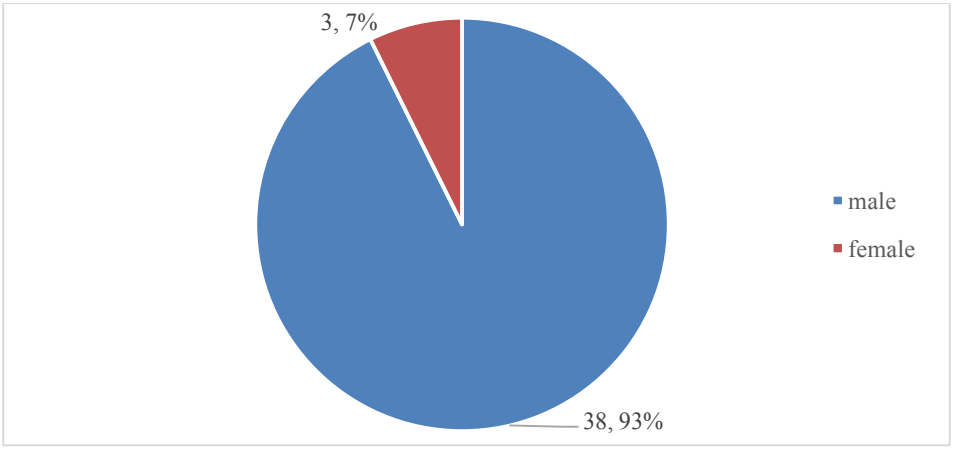

Figure 4: Samples per sex of the author in *CECheT*

Finally, regarding the geographical origin of the authors, as shown in Figure 5 below, most of the samples (46%) were written by English chemists. Scotland is the second most frequent origin (20%), whereas a further 17% – i.e. the seven samples marked as "others" – includes authors about whom there is no information or who were educated in more than one place, making it impossible to ascertain where they acquired their linguistic habits. These are followed by Irish and North American authors (7% each), as well as by one text belonging to a Welsh chemist.

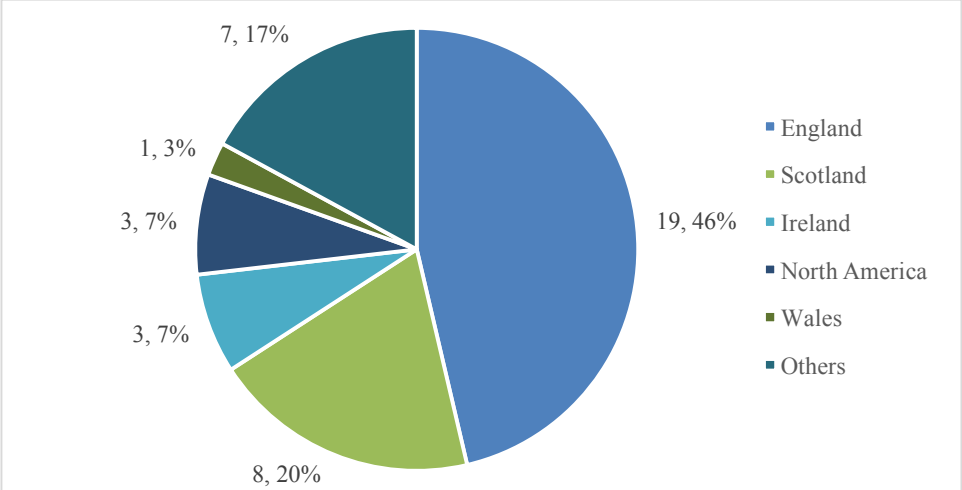

Figure 5: Samples per origin of the author in *CECheT*

7 Conclusion

This paper has presented both the main characteristics of *CECheT* and the problems faced during its compilation. Once this process is concluded, *CECheT* will be a valuable resource for research into early writing on chemistry in English, and, as part of the *Coruña Corpus*, will also contribute to the representation of scientific writing in the eighteenth and nineteenth century as a whole. *CECheT* is scheduled to be completed in December 2016.

Referencias

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing, 8/4*, 243-257.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating language structure and use.* Cambridge: Cambridge University Press.

Claridge, C., Schmied, J., & Siemund, R. (1999). The Lampeter Corpus of Early Modern English Tracts. In K. Hofland, A. Lindebjerg, & J. Thunestvedt, *ICAME Collection of English Language Corpora (CD-ROM)*. Bergen: The HIT Centre, University of Bergen.

Kytö, M., Rudanko, J., & Smitterberg, E. (2000). Building a Bridge between the Present and the Past: A Corupus of 19-century Englis. *ICAME, 24*, 85-97.

McEnery, T., & Wilson, A. (1996). *Corpus Linguistics.* Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based Language Studies: An Advanced Resource Book.* London: Routledge.

Moskowich, I. (2012). CETA as a tool for the study of modern astronomy in English. In I. Moskowich, & B. Crespo, *Astronomy 'playne and symple'. The writing of science between 1700 and 1900.* (pp. 35-56). Amsterdam/Philadelphia: John Benjamins.

Moskowich, I., & Crespo, B. (2012). *Astronomy 'playne and simple'. The writing of science between 1700 and 1900.* Amsterdam/Philadelphia: John Benjamins.

Moskowich, I., Camiña-Rioboo, G., Lareo, I., & Crespo, B. (2016). *'The Conditioned and the Unconditioned' Late Modern English Texts on Philosophy*. Amsterdam/Philadelphia: John Benjamins.