Exploring multimodality in historical texts: the value of paratextual features in the making of the Coruña Corpus[1] (CC)

Multimodalumas istoriniuose tekstuose: paratekstinių bruožų vertė kuriant Coruña Corpus (CC)

MARÍA JOSÉ ESTEVE-RAMOS, Universitat Jaume I, Spain

ISABEL MOSKOWICH, Universidade da Coruña, Spain

Abstract

In this article, we explain how the project of the Coruña Corpus of English Scientific Writing (CC) included non-linguistic elements in the corpus design from the very inception of the project, which allow for the analysis of paratextual devices such as layout, punctuation, decoration, and/or other visual language, this in order to explore how a text interacts with its context and hence how such aspects of a work are also important and meaningful in themselves. These elements are key to an understanding of how texts were perceived by their contemporary audiences, and can only be fully appreciated by considering their meaning to the reader. The considerable time and effort involved in the compilation of these minute details in the CC reflects the belief that paratextual features in the Late Modern English period can be considered vectors of additional meaning, and hence need to be included when designing and studying the linguistic material of the period. Multimodality, a current trend in various forms of linguistic analysis, provides the appropriate and necessary framework for the way we work with historical corpora and has the potential to open up new and enriching avenues of research.

KEYWORDS: multimodality, corpus linguistics, historical linguistics, scientific English, Late Modern English.

Introduction

Like other human activities, language is a socially-bound phenomenon, and as such is determined at least partly by society itself. As with other areas of human activity, this may involve more than what is immediately apparent, and the notion that there exists an intimate relationship between language and society has been fundamental in the compilation of the Coruña

[1] See www.udc.es/grupos/muste

Corpus of English Scientific Writing (henceforth CC). The project, officially launched in 2005, currently occupies a prominent position in the field of historical corpus linguistics, especially in the context of scholarship in Spain. From the outset, one of the most important decisions that needed to be taken was that of what criteria would be adopted towards accurately reflecting the interrelation between the language of the texts and the intended users, that is, between the scientific register and its writers and readers. Whereas the majority of similar projects at the time simply involved texts transcribed in .txt format, the CC has stood out from its inception in that it uses xml markup, and also seeks to incorporate practically all available elements from the compiled texts: tables, figures, mathematical symbols, abbreviations and all types of signs and visual representations, including errors and all punctuation apparatus. It is common to understand corpus studies as a matter of analysing purely linguistic elements, but it is our aim in this paper to describe the full range of possibilities that the CC offers in terms of exploring other equally important aspects of the scientific register, in that such elements are key to situating a text in its precise context, as well as understanding the dialogue between text and reader.

The main objective here, then, is to underline the value of the criteria used in the compilation of the CC, which allow for the analysis of paratextual devices such as layout, punctuation, decoration, and/or other visual language, in order to explore how a text interacts with its context and hence how such aspects of a work are also important and meaningful in themselves. These elements are key to an understanding of how texts were perceived by their contemporary audiences, and can only be fully appreciated by considering their meaning to the reader, which reflects the printing conventions of the time, and also shows how such conventions may have changed in order to adapt to new ideas and trends in scientific discourse.

In this article, the possibilities of the CC for those wishing to combine techniques of corpus linguistics with other, closer approaches to textual analysis, are outlined. To that end, the section entitled *Multimodality: the next turn in corpus linguistics* offers a review of multimodal elements of texts over time. The following section of the article addresses the criteria and principles governing both the compilation and encoding of the CC and how these are of use for a variety of alternative kinds of textual studies. Following this, a number of proposals for future research will be offered in the last section, along with some concluding remarks.

Multimodality: The Next Turn in Corpus Linguistics

In this section we provide a brief outline of several notions and ideas within the umbrella term *multimodality*, and explain why such a concept is pertinent to the design of a historical corpus. The Coruña Corpus of English Scientific Writing project was pioneering in the sense that it integrated such non-linguistic material into the design of the corpus, as will be seen below. We will also show the importance of recognising that older texts cannot be understood without their non-linguistic elements; indeed, even in terms of the linguistic material contained in a text, typographical information may operate as a significant feature of interaction and thus carry additional meaning, providing a further layer of communicative significance between the author/printer and the potential reader. This information needs to be incorporated into the corpus design, and together with the metadata file (with information related to the social and personal aspects of texts and authors) is essential to validate the full meaning of the selected material, towards being able to answer a wide range of research questions, as is the case with any other original/primary sources.

Multimodality as a Trend

Multimodality is currently of considerable interest in a variety of fields of linguistic research. Thus, we need to understand the notion in detail and how it can be implemented into the design of diachronic corpora. A basic idea here is that multimodality encompasses what we understand as semiotics, embracing elements such as image and gesture, in addition to language itself. On these lines, the following definition by Hiipala & Bateman (2022, p. 405) based on Wildfeuer et al.'s (2020) is apposite:

> Multimodality research is an emerging discipline that examines how communication builds on appropriate combinations of "modes" of expression, such as natural language, illustrations, drawings, photography, gestures, layout, and many more.

In the Late Modern English period, which the CC covers, these elements are typically illustrations, graphical material, and rhetorical and typographical elements, among others. In this respect, it is essential that these be considered as part of the message:

> What may be the most important question for Kress and also for Iedema (2003) may be more to do with what we can accomplish as communicators through different kinds of communicative resources, such as images, colour, materials, and gesture, than whether these can be described and accounted for in the same manner as Language (Machin, 2013, p. 349).

Multimodality studies arose in linguistics largely as a result of the publication of the work of Kress and Van Leeuwen, especially *Multimodal Discourse* (2001). This work was interdisciplinary in nature, emphasising the idea that texts can express meaning not only through the textual content itself, but also by means of colour, images, material objects and even the architecture of a text. Such an appreciation of the ways in which a text is conceived is fundamental today, in terms of what has come to be known as multimodal studies. In another important contribution, Iedema (2003, p. 29) refers to the same concept as an evolution or "mutation" of discourse analysis, which is considered to be "an offshot" of sociological linguistics. In his treatment of the subject, which moves from discourse analysis to the concept of resemiotisation (which we will introduce at the end of this section), Iedema (2003) explains the different steps that led to the current approach, which itself stems from a primary concern for studying language use beyond the sentence and with a focus on interaction, and evolves into a discipline that focuses on the idea of cohesion and text grammar. Furthermore, the work of Halliday (1978) introduced the idea that language is, above all else, a mode of social action:

> But most importantly, it was the systemic-relational principle underpinning Halliday's depiction of language which ultimately proved to be the most fruitful connection: semiosis not analysed in terms of discrete building blocks or structures, but in terms of socially meaningful tensions and oppositions which could be instantiated in one or more (structural) ways. (Iedema, 2003, p. 31)

As this author continues explaining, this idea was reflected convincingly in the first issue of the *Journal of Social Semiotics* in 1990, and paved the way for a great deal of research that conceptualised language as something that transcended the idea of language-as-text. Following this, we jump to 2001, when Kress and Van Leuwen coined the term *multimodality*, as noted above. The impact of these ideas is reflected in a broader and more dynamic vision of language, now extended to embrace visual and multimedia material. Today, cybergenres are a clear fusion of the visual, audible and written, similarly to how the printing press fostered its own kind of multimodality centuries ago.

As mentioned above, one especially interesting concept arising from the trend towards multimodality is expressed by Iedema (2003, p. 41): resemiotisation. This is defined as "[...] how meaning making shifts from context to context, from practice to practice, or from one stage of a practice to the next" and in this respect "[...] is crucially interested in how materiality ('expression') serves to realize the social, cultural and historical structures, investments and circumstances of our time." The concept also provides a fascinating field of research for diachronic linguistics, where the design of our corpus, including all the visual and paratextual features therein, can be compared with the same genres from different periods in order to see if the meaning of these elements has been altered or modified over time. This is why we aim to contextualise the whole communicative reality of texts and their readers by not excluding their original paratextual material. It is, above all, a process of providing historically accurate evidence for the potential researcher.

Multimodality in Corpus Studies: A Methodological Proposal

Multimodal analysis has become an important, indeed a fundamental aspect of research in a wide range of disciplines. In this section, we will discuss how the notion has been conceived and applied to the design and use of corpora, especially those containing materials from other historical periods of the English language. As we will see in the next section, the inclusion of paratextual elements in the design of a corpus is technically demanding, and not all corpora do so. As Hiipala and Bateman (2022, p. 405) observe: "Although the field of

multimodality is increasingly oriented towards empirical analysis, compiling multimodal corpora to support such analyses is still highly labour-intensive." This is especially the case when we think of historical corpora, where non-conventional material, typographical elements and genre structures are found. On this latter question, Stöckl (2017, p. 267) adresses the issue of genre clusters and their intertextual relations, with the aim of obtaining "an illustration profile of individual genres." Hiipala and Bateman (2022, p. 406) also refer to the difficulties faced by corpora designers and compilers, noting that "whereas it is common for a range of automatic processing techniques to be applied to linguistic corpora, the possibilities for multimodal data remain limited."

Of course, this idea applies to corpora that compile contemporary material. When we turn to historical or diachronic corpora, limitations still exist, but there are tools -as explained in the following section- that allow for the inclusion of the paratextual elements of older texts. What is important is the interpretation of such elements, and in this sense the inclusion of explanatory notes about the context and/or the biography of the author becomes a powerful tool to understand the potential function of these features. In the CC, information about the age, life, education and training of authors, as well as details about the publication and circulation of their texts, are provided. Taatvisainen and Suhr (2012), in a discussion of historical corpora, refer to this as a desirable aspect of corpus design:

> Sociolinguistic parameters are helpful, as they give more precise information about the authors and their audiences by specifying e.g. the author's age, education and gender, the regional origin of the writing, circulation and publishing facts, to name some relevant parameters. In an ideal case this information is encoded in a corpus and readily available to the researcher.

In the CC, this information is encoded in metadata files that are available to users. The files contain all the details relating to the life and work of each sample's author, and can be read as .xml files or used as parameters for sample selections and searches in the software provided with the corpus. Indeed, this metadata is one of the most remarkable aspects of the corpus, together with the inclusion of paratextual elements, a fact which we will underline in the present paper. It is clear that multimodality needs to be a fundamental part of the design in every compilation, and that all technologies -more so in our digital world- have become necessary allies in order to incorporate such elements. However, it is not just a matter of incorporating them so that they are visible and available; we also need to design tools that allow users to resolve their research questions easily and effectively:

> What is required at this stage in the development of multimodal studies as a field is the sort of empiricism of extensive text analysis such as revolutionized the study of language during the 1960s and 1970s. To do this will inevitably require that analysts of multimodality learn to use (particularly contemporary digital) multimodal resources and techniques in order to appropriately deal with the natures of such media [...]. (O'Halloran & Smith, 2012, p. 11)

In what follows we will briefly discuss multimodality as an element of added value in the design of the Coruña Corpus.

Multimodality in Historical Texts: Adding Value to the Coruña Corpus Design

In their introduction on one of the most innovative works in this area, Varila et al. (2017, p. 1) begin thus: "When reading a text, our understanding of its meaning is influenced by the visual form and material features of the page on which it appears." Multimodality has been a new trend in the study of contemporary English, but until very recently it has been neglected in work on historical texts, even more so in studies using historical corpora and materials. However, interest here has been growing in recent years, with visual and paratextual elements in early texts now being studied.

Our own interest centres on the validity of corpus linguistics as a tool, not just to study the discursive and linguistic elements of texts, but also to observe the other components of each sample from the primary source, serving as they do to help establish an intentional communicative bond with the audience. In order to do so, compilers and editors need to be aware that multimodality can also be a fundamental part of the way we understand, study and analyse language. Back in 2003, when the first plans for the CC were being made, it was clear that the

criteria for compilation had to include all these elements alongside the textual transcription, because corpora – that is, the technology itself – need to be able to provide researchers with resources that allow access to a direct bridge between the primary material and the reader/user. According to Varila (2017, p. 7–8), multimodal analysis:

> […] approaches representation, communication and interaction as something more than language; writing is thus only one of the communicational and semiotical meaning-making tools explored alongside image, gesture, speech and other modes […] although multimodal studies typically focus on present-day material, they offer a useful perspective for studying early written communication as well.

A variety of proposals have been made regarding approaches to the multimodal study of earlier texts. Varila et al. (2017, p. 5) suggest that, when addressing the multimodal aspect of early texts, consideration should be made, on the one hand, of what they call the "higher-level elements", which include features of layout and mise-en-page, and, on the other, about lower-level elements, such as the use of typographical variants and the script. These two levels are normally connected, but a focus on one or the other may lead to different emphases on the communicative purpose of a text. This concept of high-level elements also relates to the earlier concept of *ordinatio* by Parkes (1991), and has been used extensively in the study of manuscripts and early print. In this sense, we might consider elements such as the dimensions of the page, the amount of text (vs blank space), the use of colours, titles and rubrics, types of initials and their dimensions and shapes, punctuation, tables of contents, and illustrations. We tend to think of corpora as being text-only resources, with the compiled material stripped of its cultural and paratextual meaning, yet such a notion is necessarily challenged if we consider the multimodal dimension of language; this applies both to the present and to the past. Thus, page layout and all other multimodal elements merit close attention:

> Mak notes that while the page has long remained a basic unit of "graphic communication of ideas" despite the changing methods of text production, its role in the transmission of knowledge has not been adequately appreciated. She maintains that the "page" is an expressive space for text, space and image, it is a cultural artefact; it is a technological device. But it is also all of these at once. (Varila et al., 2017, p. 10)

All these elements are reflected in the transcription of the CC material. They are important because, from the point of view of the printers/authors, their use adds an extra layer to the communicative act, reinforcing or modifying the linguistic message itself. Furthermore, printers at the time gave specific instructions on these aspects, in that they were considered a fundamental part of the text. Thus, Varila (2017, p. 12) cites a manual of "print letters" produced by Joseph Moxon: "He advises one designing a title for an inscription to "consider the words of *emphasis*, make those words to vary from the letter your Discourse is in, as either *Roman*, *Italik*, or English [black letter], according as the words may properly require".

By choosing a specific script and/or typography the printer was effectively intervening in the process of composition. We also need to bear in mind that specific letters and typographical elements were sometimes modified, or were even newly introduced at the page-setting stage, and it is not always clear whether these were at the hand of the printer or the author. Traditional abbreviations were no longer used widely, and the use of the *ct ligature* and other similar spellings became part of publishing conventions, all this at a time when prescription began to be imposed:

> Above all, it is significant that the English spelling system that emerged from the seventeenth century is not a collection of random choices from the ungoverned mass of alternatives that were available at the beginning of the century but rather a highly ordered system taking into account phonology, morphology, and etymology and providing rules for spelling the new words that were flooding the English lexicon. (Brengelman, 1980, p. 334)

One of the main reasons for assessing the use of spelling in these texts is that authors at this time would have had a degree of control over their texts, although printers would also have intervened, in that "it was not until well into the eighteenth century that authors expect printers to "spell, point, and digest their copy" and to correct far-fetched spellings." (Brengelman, 1980, p. 342).

The design of the CC allows the presence of these spellings to be seen, and therefore affords us greater insight into the text, that is, a more multimodal vision of what the material represents. The fundamental reason why the inclusion of the non-linguistic elements is a key issue here is because multimodality is the lens through which we understand communication. It has even more significance in understanding texts from other periods, where such paratextual features were actors of meaning, especially in a non-digital era.

The Making of the Coruña Corpus and the Importance of Defining Criteria: Examples and Relevance

One of the most time-consuming tasks when compiling a corpus is its design. In the case of the CC the planning and designing process took several years and included various tests and pilot studies, with beta versions of the first corpora. Although unconsciously to begin with, the initial steps of our methodology coincided with the approach outlined by Kennedy (1998, p. 70–85):

1 Corpus design
2 Planning a storage system and keeping records
3 Obtaining permissions
4 Text capture
5 Markup

The CC covers the years 1700–1900, a period in which science and scientific communication were flourishing in the Western world. In the belief that corpora must represent real languages or registers, the design of the corpus aimed to cover different scientific disciplines independently, in separate subcorpora, but maintaining the same structure for all of these, so that they could be used in the same way, with comparability ensured. The initial idea meant that comparisons could be made between the soft and hard sciences (Hyland, 2005), taking the UNESCO *Classification of the Fields of Science and Technology* (1988) as a starting point for the construction of our own taxonomy.

The subcorpora of the CC all share the same structure, with each one containing two samples of approximately 10 000 words per decade. Thus, each discipline-specific subcorpus is c. 400 000 words in size. Another principle of compilation is that of only representing the language of the authors. Hence, all selected texts were written directly in English, and all language not pertaining to the author in question is excluded from the repertoire of tokens detected by the search engine[2]. However, as compilers we also want to provide users with information which is present in the works, but which transcends the texts themselves. To this end, we have attempted to represent more content than that contained in the sample text itself. This, in turn, implies that the format in which our files are saved prior to indexing in each corpus cannot be plain text. As early as 2007 we decided to use XML as a means of incorporating details that could be relevant for users who sought to go beyond the text, that is, elements that we believe provide additional information that might be of interest to users. The CC is a specialised corpus in that it is "delimited by a specific register, discourse domain or subject-matter" (De Beaugrande, 2001, p. 11). So, it also includes those paratextual elements which may be characteristic of such specialised registers and which, of course, can form part of the reading experience for the intended audience in various ways. All these paratextual elements have been retained in transcribing and editing the samples, because they form part of the way in which the author communicates, and thus their nature and place in the texts is relevant in terms of the scientific knowledge conveyed. In this way, the text goes beyond the text.

Driscoll (2007) distinguishes two possible levels at which we can approach the question of preparing electronic manuscripts, depending on whether we focus on the structure of the work (chapters, sections, paragraphs) or the structure of the document (layout, arrangement of the text on the page, etc.). In encoding the CC we have tried to do both of these, and hence we have indicated where paratextual elements occur in the samples, even in those cases where we do not in fact reproduce them as part of the sample itself.

That said, it is clear that of Kennedy's (1988) five steps, those involving text capture and encoding, are of the greatest import for us, in that we want to provide the most faithful machine-readable version possible of each sample. Therefore, extra textual material and layout details were both carefully considered, with figures, tables, page numbers, quotations and symbols all included in the corpus in one way or another. In what follows we will describe how these elements were dealt with and the various solutions adopted to that end.

[2] For a complete account of the structure and details of the Coruña Corpus, see Crespo and Moskowich (2020) and Moskowich (2021).

Quotations

Quotations are clearly not instances of the language of the author whose linguistic features are being compiled in a corpus. However, the very fact that a writer decides to mention other writers may be relevant for our understanding of scholarly traditions. In the CC, we have marked up such quotations so that they are excluded from the index used by the Coruña Corpus Tool (CCT), the software accompanying the corpus, to carry out searches. The location of quotations is indicated with the editorial mark [quotation]. However, since not all quotations are the same, we have dealt with them in different ways.

In those cases where removing the quotation would affect the logical discourse of the text, we include them in the typed text but mark them with the tag <del> so that the quoted words are not considered by the CCT when creating the index of samples. Such words can be seen in a window of the CCT, and can provide relevant context for understanding a text sample. Fig. 1 below illustrates the way in which these elements are shown in the Display Window of the CCT:
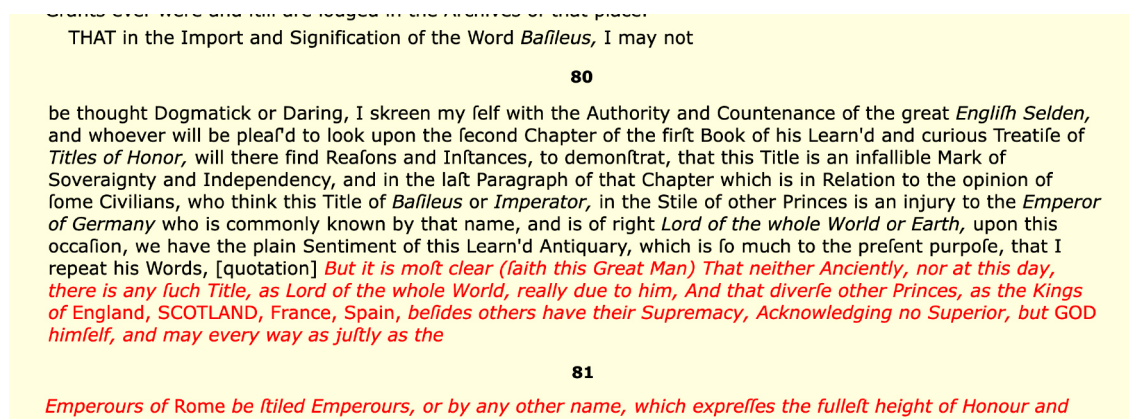
Fig. 1 In red, words by other authors (not searchable by the CCT)

When the deletion of quotations does not affect the comprehension of the text, they are not reproduced in the sample at all. This is also the case for quotations in languages whose alphabets we do not represent. Here, we not only indicate the place where the quotation occurs in the original, but also provide information on the language used by means of other editorial marks ([quotgreek] [quotrussian]) as shown in Fig. 2, where the quotation occurs in a note and inside an extract that has been deleted:

Fig. 2 Marking of quotations in languages not reproduced

Figs. **3** and **4** below illustrate the way in which quotations in other languages are identified when they are short. These also indicate the language used:

Arms of the Family, Two Coats quarterly, 1ſt *Azure,* a Bend betwixt Six croſs Croſlets *Or,* for the Name of *Marr;* 2d *Argent* a Pale *Sable;* the Paternal Coat of *Erskine* ſupported by Two Griffons. The Creſt is a Hand holding a crooked Sword; the Motto, [quotfrench] Je penſe plus.

In the Year 1638, *John* Earl of *Marr* alienated the Barony of *Erskine* to Sir *John Hamilton* of *Orbieſtoun,* one of the Senators of the College of Juſtice, Lineally deſcended to *Gavin,* a younger Son of *James* Firſt Lord *Hamilton; William Hamilton* of *Orbieſtoun,* his Grandchild, ſold theſe Lands, *[an].* 1703. to *Walter* Lord *Blantyre.*

Fig. 3 Marking of quotations in languages other than English

77

His Armorial Bearing is *Gules,* a Lion Rampant, *Argent,* within a Border ingraled, of the 2d, and for Creſt a Lion's Paw, with this Motto, [quotlat] Cogit in hoſtem.

Fig. 4 Marking of quotations in languages other than English

The CC only represents scientific prose. For this reason, material such as poetry is omitted, which occurs with some frequency in our samples. The screenshot in **Fig. 5** illustrates a sample taken from *The Natural History Of Barbados*, published in 1750 by Griffith Hughes. It is compiled in the Corpus of English Life Sciences Texts (CELiST) and shows that whereas a poem is included by the author, we do not reproduce it in the sample, although we do indicate where it occurs, because it may have some relevance for certain interpretations of this epistemic tradition:

Moſt of theſe Birds of Paſſage never fail to appear here between the Nineteenth and the Twenty-ſeventh of *Auguſt,* eſpecially if it be then wet Weather; but if about that time it is very dry, the greateſt Part of them are ſeen to fly very high, and to keep their direct Courſe towards the Eaſt; and, as *Milton* expreſſes it,

[quotation] [poem]

Fig. 5 Representation of poems in the CC

Some other kinds of material are omitted when encoding the samples, such as those labelled fragments, which will be dealt with in the following subsection.

Fragments

As noted above, the CC samples of c. 10 000 words are taken from longer texts. Where a complete work is shorter than this, it is included in its entirety. This number of words per author is substantially greater than what has been suggested by some scholars as a working minimum for the study of variation: Biber (1993), for example, recommends restricting the size of samples to 1000 words for Present-day English. Also, since we do not want to resort to the same parts of texts all the time, in order to avoid the repetition of rhetorical patterns and the linguistic devices that these may involve, we have sometimes started a sample in *media res* or set the cut-off point before the end of a particular section or chapter.

There are several circumstances in which parts of a text are not reproduced, and where we indicate that a fragment has been deleted and that there was more text by the author there. In such cases, we use the tag <del> and the editorial mark [fragment] as shown in **Example 1** below:

Ex 1   ~~[fragment]~~

In this way users of the corpus can see where we have removed a fragment of the work, and at the same time the use of the tag <del> avoids the CCT indexing our editorial mark in its repertoire, hence it is not considered as part of the author's language. On some occasions, the mark [fragment] is also used when we omit a passage

that does not seem to deal with a scientific topic *stricto sensu*, which is often the case in eighteenth-century works. Nevertheless, the position of such fragments is always visible (see **Fig. 6**), either when they are too long or not easily reproduced because they are in non-Latin script.

Fig. 6 Use of [fragment] to indicate that a sample does not start at the beginning of a section

When the deleted fragments are not in English, we also indicate the language in which those words were written. The use of editorial marks such as [fragment], [fragmentgreek] or [poem] are one of the characteristics of the CC that make it different from other historical (or non-historical) corpora, and is useful for a more comprehensive approximation to late Modern English scientific writing in that paralinguistic material offers potentially relevant information here.

Figures and Tables

Scientific writing in the period following the demise of Scholasticism tended to resort to the illustration of what was being communicated by means of figures (of objects, instruments, animals or plants) or tables containing data. Such elements are not considered to be prose (whose analysis is the primary interest of the CC) and are therefore omitted when codifying the samples. However, the fact that they were included in a work by its author cannot be overlooked, as this may also provide us with information about the tendencies of particular authors or tendencies more widely in the epistemic community. Thus, while figures in VML are not represented, we do indicate the places where these occur in the original book or article. The same procedure as for other kinds of omitted material is followed, using the tag <del> so that the software never includes the editorial mark in the word-count for the sample or considers it part of the language of the author. The location of such graphs, figures or drawings and engravings either in the middle of a paragraph or outside it is thereby indicated.

The transformation of a text sample into searchable corpus material is carried out so that compilers have preserved if not the particular objects themselves, then at least the type of objects they are and the place in the

text where they occur. Hence, in **Fig. 7** we can see how we deal with images, rather than simply ignoring them. This sample, from Curson's *The theory of sciences illustrated; or, the grounds and principles of the seven liberal arts: grammar, logick, rhetorick, musick, arithmetick, geometry, astronomy* (1702), contains a figure in the upper left-hand part of the page in the original, which we identify in our computerised corpus (on the right):
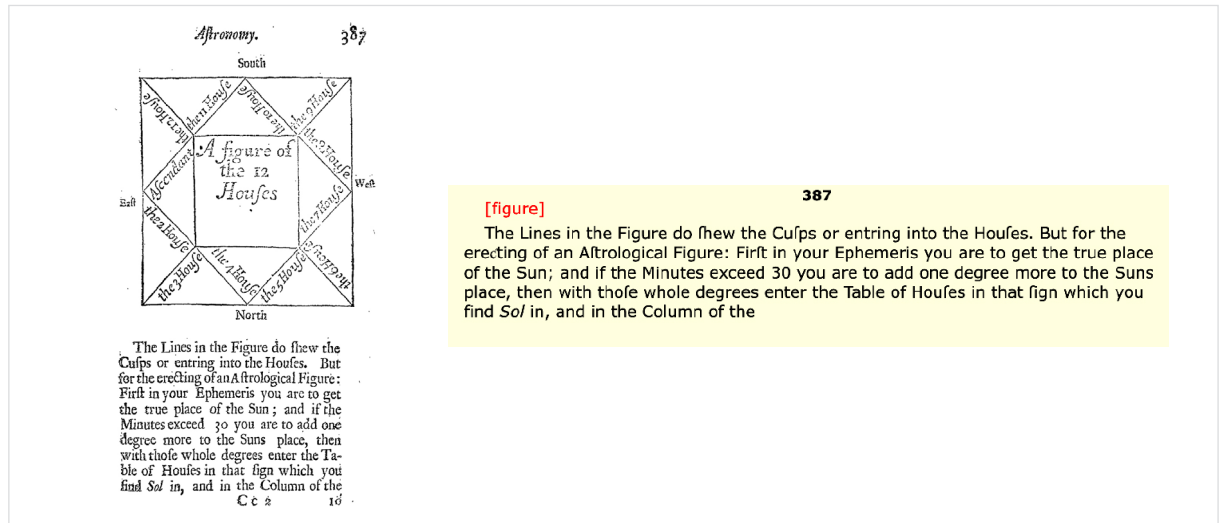

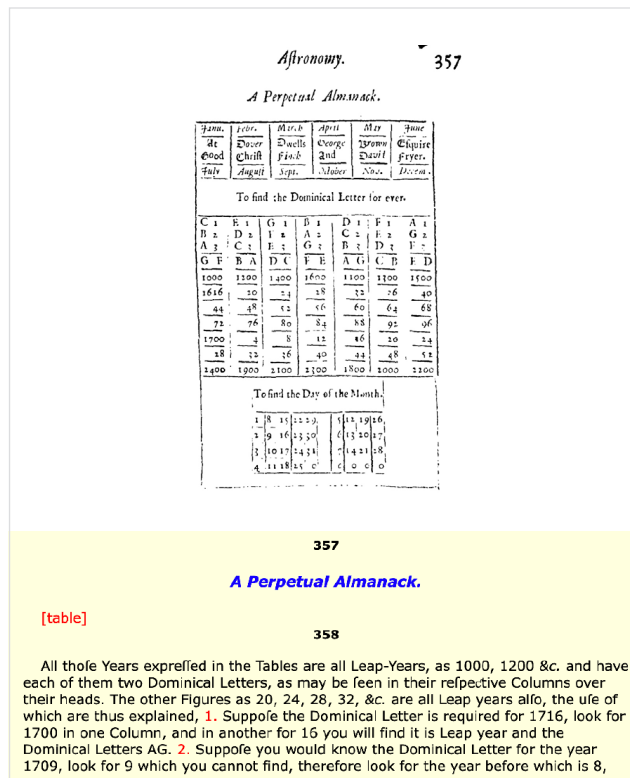**Fig. 7** Image of printed figure and its representation in the CCT

Fig. 8 Page 357 of the book by Curson (1702) and how it is rendered in the CC

The same protocol has been applied in the case of tables and diagrams of all kinds. The two images on the left (**Fig. 8**) exemplify this. The image on the left is of page 357 of the same book by Curson (1702), which is included in the Corpus of English Texts on Astronomy (CETA). At the bottom there is a screenshot of how users can see the same information in the Display Window of the CCT: the image shows the content of the preceding page and the page number (357). Immediately following this we reproduce the title given to the table by the author, and we also indicate that the table occupies the whole page (there is nothing else on it), plus the fact that the words within the table itself will not appear as part of the terms used by the author, since it is tagged <del>, which makes clear that it is part of the compilation process and the content of the table will not be reflected anywhere in the repository of author's words.

Retaining this information allows users to have a general idea of how to approach the study of multimodal techniques in the samples of the various subcorpora, that is, in the different scientific disciplines covered by the CC.

Spelling Variations

Corpora compiled following strict principles differ from others in several key ways. While it is true that they cannot compete in terms of size, they can do so in terms of rigorousness. Examples of such careful compilation and encoding can be found with certain historical corpora, and it is in this sense that the CC can be considered. For instance, the preservation of the spelling variations found in the compiled texts has been sought, and to this end a variety of symbols and characters have been encoded, as can be seen in the Display Window of the CCT (see underlined forms in Fig. 9 below). We have also made sure the software can find all of them.
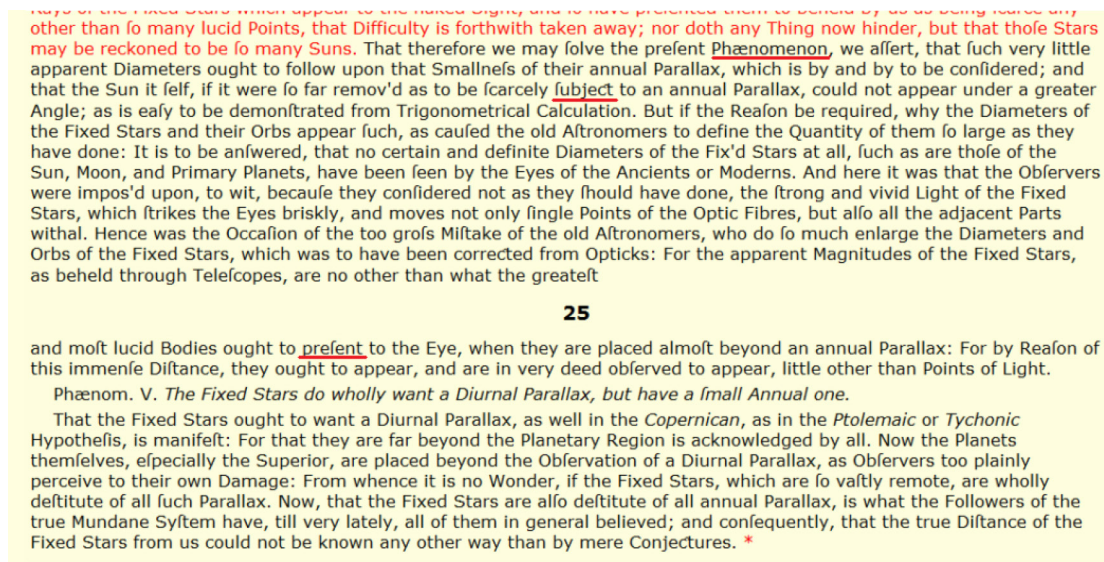

Fig. 9 Examples of outdated spellings

As corpus linguists we also wanted our software to retrieve very accurate data. This involves the CCT being able to recognise different spelling variants, even if only one such variant is searched. Fig. 10 illustrates a search for the term *economical*; note how the application also provides variant spellings of the original search term:

Fig. 10 Spelling variants in the search window of the CCT

Abbreviations

Late Modern authors of scientific writing often used abbreviations. Many of these occur frequently, whereas others are characteristic of particular domains. In all cases, we have used the tag <abbr> to mark these, as in Example 2:

Ex 2   <p>As the moment of commencing the expedition approached, Marquette was engaged in his labors at the Mission of <abbr>[St].</abbr> Francis Xavier, and was then joined by Joliet from Quebec</p>

Together with abbreviations, we have preserved full stops and any other punctuation marks as they appear in the original.

Notes

There are two kinds of notes in the samples of the CC: margin notes and footnotes. We follow TEI guidelines, which require that all notes are placed immediately after their reference, which implies that at first glance they are not visually distinguishable from one another. However, we indicate the type of note we are dealing with:

Ex 3   <p>[margin note] the text [endnote]</p>

Margin notes are not always attributable to the writer and in fact are often notes included by the editor or even the printer to indicate what a particular extract is about. In these cases, when a margin note does not seem to be by the author, we do not reproduce it at all.

Footnotes often contain comments but may also contain references. Bibliographical references are included together with other material that can serve to help reconstruct the author's epistemic universe. This is what we see in the sample represented in Fig. 11, taken from *The plan of a dictionary of the English language; addressed to the Right Honourable Philip Dormer, Earl of Chesterfield; One of His Majesty's Principal Secretaries of State*, by Samuel Johnson (1747, p. 6):
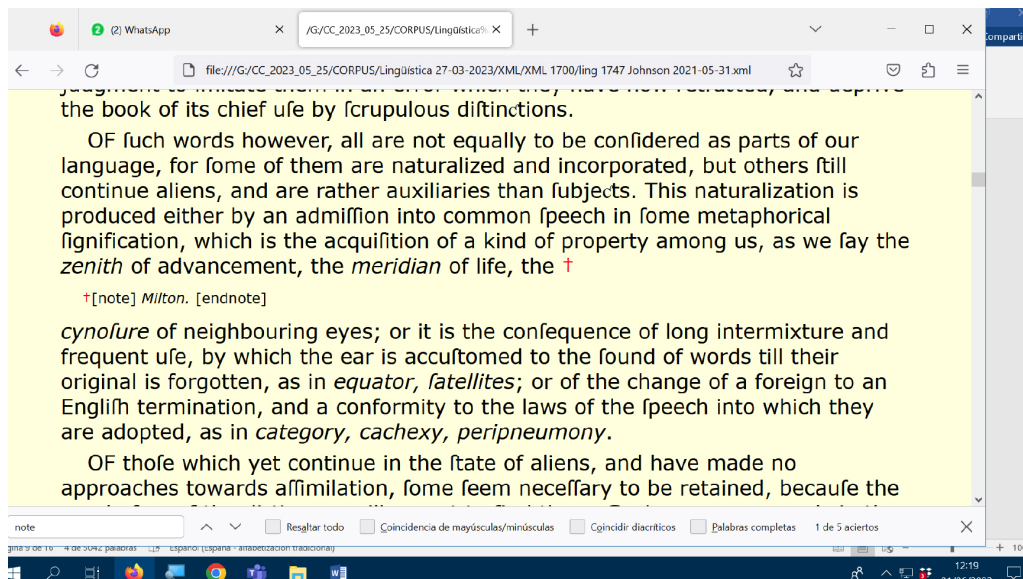

Fig. 11 Representation of end notes in the Corpus of English Texts on Languages (CETeL)

Hence this bibliographical information remains visible; it is of potential use, for example, in studies about epistemic networks.

Representation of Non-English Text

When encoding the Corpus of English Texts on Languages (CETeL) we encountered some difficulties that were specific of this particular subcorpus. These related to the presence of foreign words, which were often the subject matter of the sample itself. In such cases, and although these words were not part of the linguistic habits of authors, they were not deleted; rather, all such elements were maintained as visible, searchable and analysable.

As we have already noted, in those cases where the author uses foreign languages that we cannot directly represent with our Latin-based alphabet, we resort to editorial marks such as [fragmentgreek]. Thus, readers know that there is an extract in a foreign language that is not reproduced, in that the corpus includes English text only. Despite this, the fact that we indicate the presence of these elements is useful, since it provides information on the author's background. **Fig. 12** below illustrates how such forms are dealt with for the text sample from Henry Curson's *The theory of sciences illustrated; or, the grounds and principles of the seven liberal arts: grammar, logick, rhetorick, musick, arithmetick, geometry, astronomy*, which was published in London in 1702 and is included in the Corpus of English Texts on Astronomy (CETA):
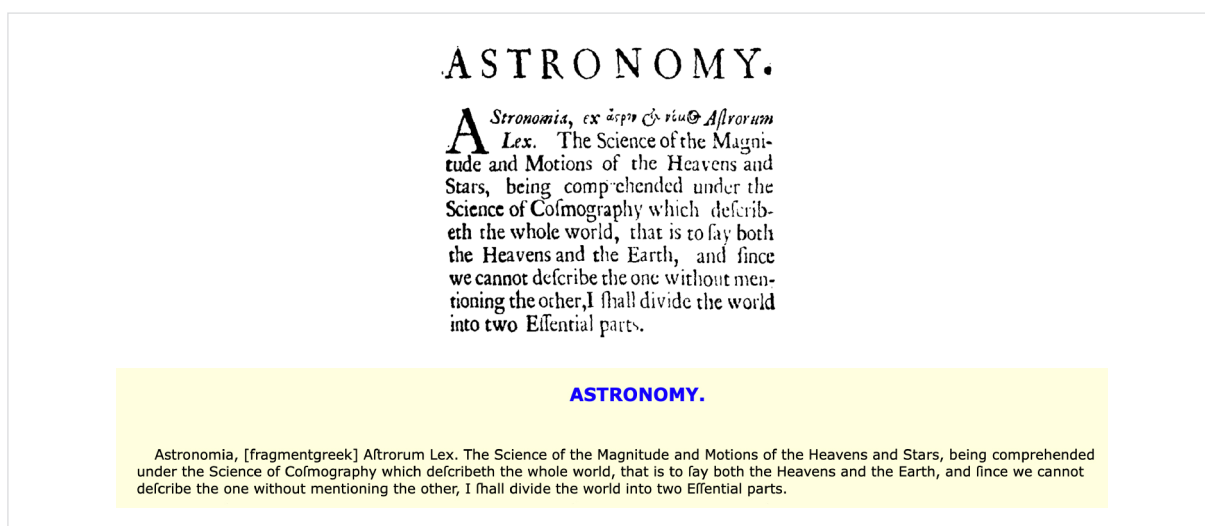
Fig. 12 Original image of page 337 from the work by Curson (1702). And solution given by the compilers to non-Roman alphabet terms

Unconventional Signs and Symbols

Our samples do not contain only text, and this implies that we need to represent things that cannot be found on a conventional keyboard. Signs of the zodiac, symbols indicating paragraph or section references, and all other symbols in a text that do not have a syntactic function, are tagged as elements that will not be considered by the CCT, that is, as non-analysable:

Ex 4   ~~symbol~~

In order to represent symbols and special characters we have resorted to the Unicode standard[3] as it is a less limited character encoding system. As already noted, many of these special symbols appear in Astronomy texts, and are tagged with <del> as non-analysable. However, other symbols are both shown and indexed for searches, as in the case of ct ligature and long s, both of which are very abundant in eighteenth-century samples, since authors and printers still uses them widely at that time.

[3] http://unicode.org

Punctuation

We have maintained punctuation marks as they appear in the original texts, except in those cases where they are themselves the subject of a discussion, something which occurs in texts dealing with languages and linguistics (CETeL). The tag <seg> is used to represent these marks when they have the value of a word. This decision does not affect the reading of the sample and allows for a better analysis, as can be seen in the example below, which illustrates the encoding of a fragment from Sweet (1892, p. 14):

Ex 5   <div3>

<head>PHONETICS.</head>

<p>~~41.~~ Phonetics is the science of speech-sounds.</p>

<p>~~42.~~ As the ordinary spelling does not always show the real pronunciation, it is necessary to use a <emph>phonetic</emph> spelling, which, to prevent confusion, we enclose in <seg type="punctuation">(</seg> <seg type="punctuation">)</seg>.</p>

Formulas

In the disciplines compiled thus far we have found a variety of formulas: mathematical, physical, chemical, etc. Complex formulas have not been represented, although their place in the original text has been indicated. In fact, formulas, as with certain other elements, have been only retained and reproduced (where possible) in cases where they perform a syntactic function, that is, where a passage would not be fully comprehensible without them. The same criterion has been applied to alphanumeric strings.

Text Divisions

Particular layouts and text-nesting might in themselves constitute relevant aspects of specific epistemic models, and for this reason we have retained these different levels of embedding. Tags such as <div2> or <div3> are used to indicate the level of embedding of fragments of text such as section and subsection. Meanwhile, <head>text</head> are used to mark chapter titles within first-level divisions, as in the following example:

Ex 6   Original text:

CHAP. VIII

THE COMETS

Tagged text:

<div1>

<head><abbr>[CHAP].</abbr> 0VIII

THE COMETS</head>

texto

</div1>

As with the other strategies described above, the aim here is to preserve the integrity of non-verbal elements of a text.

Things that Have Been Edited

As editors, we have modified certain other things. Thus, we normally correct typos so that the CCT can find all instances of a particular type. To ensure that a specific form does indeed constitute a typo, the *OED* is consulted to confirm that the form did not exist with that particular spelling at the time the text was written. Information regarding typos and misprints is then introduced in the metadata file accompanying each sample. Where errata are found in a work, we also include this information in the metadata.

Another decision we had to take relates to truncated words. To ensure that the software recognises all forms correctly, truncated words are reproduced in a non-truncated form, thus limiting the use of hyphens to compound words. This in turn implies a limitation in terms of the study of certain aspects of late Modern English spelling, but the technical issues arising from the retention of truncated items made the decision to avoid them inevitable.

Conclusions

In this final section we will provide some suggestions for how the CC can be used from a multimodal perspective. In order to do so, we need to identify potentially discursive and rhetorical patterns in a specific subcorpus, and to interpret these in the Late Modern English context of the corresponding discipline. Such aspects will serve to reveal the fluidity of the dialogue between the text and the reader.

Within the various elements present in the design of the corpus, as described in the second section of the article, we find tables and figures, archaic typographical elements, quotations and notes, among others. A further element, that of the use of capitalisation (especially in this period) and/or italics can also be meaningful in an analysis. Yet more features of interest are the use of signs and abbreviations, in that they potentially relate to the level of technicality of a text, an important question when considering the kind of material compiled in the CC.

Some preliminary searches of the Corpus of History English Texts (CHET) – and more specifically in the sample from 1805 – reveal interesting evidence for future research questions. In this text, the typography imitates old script. The author is American, and perhaps in this text there is some emphasis on the epic and archaic tone of the subject, which at the same time seems to lend the writing greater authenticity. Another example appears in 1810, where a sample about the history of Spain uses many kinds of archaic letters and scripts, probably with the same purpose. Britton's sample (1814) shows an interesting example of this symbiosis between image and text.
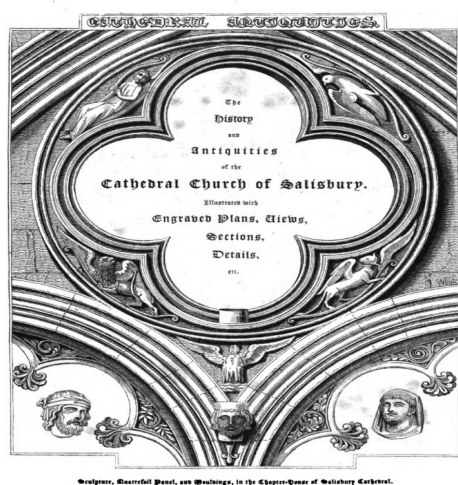

Fig. 13 Original image of Britton, 1814. *The History and Antiquities of the Cathedral Church of Salisbury.*

Another interesting line of research involves a comparison of the different disciplines represented in the CC. For example, the subcorpus on Astronomy (CETA) presents no instances of the ct ligature or the long s after 1790. However, in the subcorpus of Philosophy (CEPhil) we find examples of these archaic graphemes until 1811. Meanwhile, in the subcorpus of Life Sciences (CELiSt) we find them until 1804, as illustrated in Fig. 14.

The reasons for this may be related to the level of technicality and perhaps also to the characteristics of the audience. Readers and their social strata need to be taken into account when analysing this aspect of a text, because there may be an intention within the text not only to help the reader navigate the text, but also to persuade them that the text needs to be interpreted in a certain way. So, we can say that persuasion is part of multimodality. In fact, Varila et al. (2017, p. 15) refer to the Aristotelian concept of ethos as the first and most important mode of persuasion, thus implying that persuasion is based on the character of
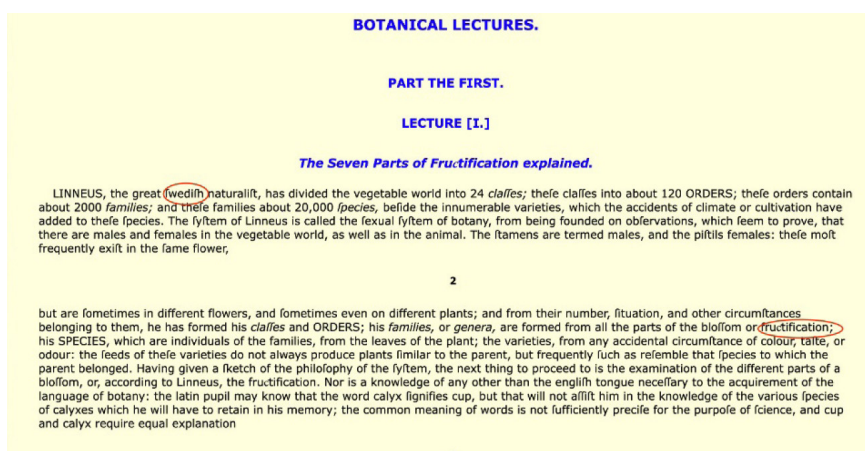
Fig. 14 Representation of the late use of archaic spelling in CELiST

the speaker. By way of example, Varila et al. mention a book showing "the arms of a well-known university or a famous printer's device on its cover" (p. 15) probably to warn the reader about the efficiency of the producers, and therefore also implying that the contents of the book will be of high quality. These same authors claim that while ethos focuses on the speakers or writers, the second mode of persuasion, pathos, focuses on the emotions of the hearer or reader and assert that "images are perhaps the most obvious visual means of achieving this." (p. 15)

The considerable time and effort involved in the compilation of these minute details in the CC reflects the belief that paratextual features in the Late Modern English period can be considered vectors of additional meaning, and hence they need to be included when designing and studying the linguistic material of the period. Multimodality, a current trend in various forms of linguistic analysis, provides the appropriate and necessary framework for the way we work with historical corpora and has the potential to open up new and enriching avenues of research.

Acknowledgement

Conflict of Interest

The authors declare no conflict of interest regarding the publication of this article.

References

1 Biber, D. (1993). Representativeness in corpus design. Literary and Linguistic Computing, 8(4), pp. 243-257. https://doi.org/10.1093/llc/8.4.243

2 Brengelman, F. H. (1980). Orthoepists, printers, and the rationalization of English spelling. The Journal of English and Germanic Philology, 79(3), pp. 332-354.

3 Crespo, B., & I. Moskowich. (2020). Astronomy, philosophy, life sciences and history texts: Setting the scene for the study of modern scientific writing. English Studies, 101(6), pp. 665-684. https://doi.org/10.1080/0013838X.2020.1798635

4 De Beaugrande, R. (2001). Large corpora, small corpora and the learning of language. In Ghadessy, M., A. Henry, & R. L. Roseberry (Eds.). Small corpus studies and ELT. Amsterdam/ Philadelphia: John Benjamins, pp. 3-28. https://doi.org/10.1075/scl.5.04bea

5 Driscoll, M. J. (2007). Levels of transcription. In L. Burnard, K. O'Brien O'Keeffe, & J. Unsworth (eds.), Electronic textual editing. Retieved April 2023 from https://tei-c.org/Vault/ETE/Preview/driscoll.html.

6 Halliday, M. A. K. (1978). Language as Social Semiotic: The Social Interpretation of

Language and Meaning. London: Edward Arnold.

7   Hiipala, T., & J. A. Bateman. (2022). Semiotically-grounded distant viewing of diagrams: insights from two multimodal corpora. Digital Scholarship in the Humanities, 37(2), 405-425. https://doi.org/10.1093/llc/fqab063

8   Hyland, K. (2005). Stance and engagement: a model of interaction in academic discourse. Discourse Studies, 7(2), 173-192. https://doi.org/10.1177/1461445605050365

9   Iedema, R. (2003). Multimodality, resemiotization: extending the analysis of discourse as multi-semiotic practice. Visual Communication, 2(1), pp. 29-57. https://doi.org/10.1177/1470357203002001751

10 Kennedy, G.D. (1998). An Introduction to Corpus Linguistics. New York: Longman.

11 Kress, G., & Van Leeuwen, T. (2001). Multimodal Discourse: The Modes and Media of Contemporary Communication. London: Arnold Publishers.

12 Machin, D. (2013). What is multimodal critical discourse studies. Critical Discourse Studies, 10(4), 347-355. https://doi.org/10.1080/17405904.2013.813770

13 Moskowich, I. (2021). The making of the corpus of English life sciences texts (CELiST), a bunch of disciplines. In Moskowich, I., I. Lareo, & G. Camiña Rioboó (Eds.), All Families and Genera: Exploring the Corpus of English Life Sciences Texts. Amsterdam/Philadelphia: John Benjamins, pp. 2-19. https://doi.org/10.1075/z.237.01mos

14 O'Halloran, K. L., & Smith, B. A. (2012). Multimodal text analysis. The encyclopedia of applied Linguistics, Wiley online library, pp. 1-5. https://doi.org/10.1002/9781405198431.wbeal0817

15 Parkes, M. B. (1991). Scribes, Scripts and Readers. Studies in the Communication, Presentation and Dissemination of Medieval Texts. London: The Hambledon Press.

16 Stöckl, H. (2017). Multimodality in a diachronic light: Tracking changes in text-image relations within the genre profile of the MIT Technology Review. Discourse, Context & Media, 20, 262-275. https://doi.org/10.1016/j.dcm.2017.07.001

17 Taatvisainen, I., & Suhr, C. (2012). Developing historical corpus pragmatics towards multimodality. Studies in variation, contacts and change in English 11: Developing corpus methodology for historical pragmatics. Retrieved June 2023 from https://varieng.helsinki.fi/series/volumes/11/taavitsainen_suhr.html

18 UNESCO (1988). Proposed international standard nomenclature for fields of science and technology. UNESCO/ROU257 rev. 1. Paris.

19 Varila M. L., Salmi, H., Mäkilähde, A., Skaffari, J., & Peikola, M. (2017). Disciplinary decoding: Towards understanding the language of visual and material features. In Peikola, M.,A. Mäkilähde, H. Salmi, M. L.Varila, & J. Skaffari (Eds.), Verbal and visual communication in early English texts. Turnhout: Brepols, pp. 1-20. https://doi.org/10.1484/M.USML-EB.5.114128

Santrauka

María José Esteve-Ramos, Isabel Moskowich

Multimodalumo tyrinėjimas istoriniuose tekstuose: paratekstinių bruožų vertė kuriant Coruña Corpus (CC)

Šiame straipsnyje paaiškinama, kaip vėlyvosios anglų mokslo kalbos tekstyno Coruña Corpus (CC) projektas nuo pat jo sudarymo pradžios į tekstyno dizainą įtraukė nekalbinius elementus, leidžiančius analizuoti tokias paratekstines priemones, kaip maketas, skyrybos ženklai, puošyba ir (arba) kita vaizdinė raiška, siekiant ištirti, kaip tekstas sąveikauja su jo kontekstu ir kaip tokie kūrinio aspektai yra svarbūs ir reikšmingi patys savaime. Šios priemonės yra labai svarbios norint suprasti, kaip tekstus suvokė jų tuometinė auditorija, ir juos galima visiškai įvertinti tik įvertinus jų reikšmę skaitytojui. Nemažai laiko ir pastangų, pareikalavusių rengiant šias smulkmenas CC, atspindi įsitikinimą, kad vėlyvosios moderniosios anglų kalbos paratekstinės ypatybės gali būti laikomos papildomos reikšmės vektoriais, todėl juos reikia įtraukti kuriant ir studijuojant to laikotarpio kalbinę medžiagą. Multimodalumas, dabartinė įvairių kalbinės analizės formų tendencija suteikia tinkamą ir būtiną pagrindą darbui su istoriniais tekstynais ir gali atverti naujų ir praturtinančių tyrimų galimybių.

About the Authors

MARÍA JOSÉ ESTEVE-RAMOS

Phd, Senior Lecturer, Universitat Jaume I, Spain

Research interests
Historical linguistics, scientific English, manuscript studies, corpus studies

Address
Dpt d'Estudis Anglesos (FCHS). Universitat Jaume I. 12071 Castelló de la Plana, Spain

E-mail
resteve@ang.uji.es

Orcid ID
0000-0003-4935-2413

ISABEL MOSKOWICH

Phd, Professor, Universidade da Coruña, Spain

Research interests
Corpus studies, historical linguistics, scientific English

Address
Facultade de Filoloxia, Campus da Zapateira, Universidade da Coruña, 15071 A Coruña, Spain

E-mail
isabel.moskowich-spiegel. fandino@udc.es

Orcid ID
0000-0003-4380-2487