Materialidades da Literatura

Vol. 4.2 (2016)
ISSN 2182-8830

'Estudos Literários Digitais 2'
Manuel Portela e
António Rito Silva (orgs.)

Representativeness in Corpora of Literary Texts: Introducing the C18P Project

IRIS GEMEINBÖCK
University of Vienna

Abstract
Currently there are very few specialised corpora of literary texts that are tailored to the needs of literary critics who are interested in corpus stylistic analyses of prose fiction. Many existing corpora including literary texts were compiled for linguistic research interests and are often unsuitable for corpus stylistic purposes. The paper addresses three of the main problems: the absence of labelling of the texts for literary genre, the use of extracts, and the prevalence of linguistic periodisation schemes. C18P is a corpus of prose fiction designed specifically to address these issues. It traces the early development of the novel from 1700 up until the Victorian era. It can, for instance, be used for an analysis of the characteristic linguistic features of individual literary genres and forms. The following paper introduces the design of the corpus as well as some of its potential uses. **Keywords:** corpus analysis; corpus stylistics; corpus building; eighteenth century; prose fiction; representativeness.

Resumo
Existem atualmente poucos *corpora* específicos de textos literários que estejam concebidos para servir as necessidades de críticos literários interessados na análise estilística de *corpus* de prosa ficcional. Muitos dos *corpora* de textos literários existentes foram compilados para efeitos de investigação linguística e, em muitos casos, não se adaptam aos objetivos da análise estilística de *corpus*. Este artigo aborda três dos principais problemas: a ausência de uma classificação dos textos por género literário, o uso de excertos e a prevalência de esquemas de periodização linguística. O CP18 é um *corpus* de prosa ficcional concebido para resolver aquelas limitações. Traça o desenvolvimento inicial do romance desde 1700 até à época Vitoriana. Pode ser utilizado, por exemplo, para análise das características linguísticas específicas de determinados géneros e formas literárias. Este artigo apresenta o desenho do *corpus* bem como alguns dos seus possíveis usos. **Palavras-chave:** análise de *corpus*; estilística de *corpus*; construção de *corpus*; século XVIII; prosa ficcional; representatividade.

1. Introduction

Corpus stylistics has proved to be an interesting approach to literary texts and a large number of studies have been done in the past decade (see inter alia Fischer-Starcke, 2010; Mahlberg, 2007; O'Halloran, 2007). Corpus analyses have for example gainfully been used to analyse the characteristics of Dickens's style, typical features in the speech of characters in Shakespeare plays, and the mood as well as the role of ambigui-

MATLIT 4.2 (2016): 29-48. ISSN 2182-8830
http://dx.doi.org/10.14195/2182-8830_4-2_2

ty in Conrad's *Heart of Darkness* (Mahlberg, 2007; Culpeper, 2009; Stubbs, 2005).[1]

So far, corpus stylistic studies of British literature have by and large either focused on the work of a particular author, such as Dickens or Austen (Mahlberg, 2007; Fischer-Starcke 2010), or on the stylistics of an individual text, such as *Eveline* or *The Heart of Darkness* (O'Halloran, 2007; Stubbs, 2005), and not on the analysis of literary genres. This might be a reason why despite the popularity of literary stylistics as an approach, there are relatively few ready-built downloadable corpora available that have been designed specifically for the study of different literary kinds and their stylistic properties. Conversely, the lack of specialised corpora of British literature might be a reason why literary genre thus far has not been a focus in literary stylistics.[2]

One of the main issues with existing corpora which will be discussed below is the lack of labelling for literary genre, which—obviously enough—is a prerequisite for being able to analyse genre in prose fiction. One example of a corpus of present day English that does have a genre classification system in place is the COCA (Davies, 2008). For the existing eighteenth and nineteenth century corpora literary genres are not usually used as categories.

However the case may be, corpus stylistic studies of literary genres are scarce and currently, there are few corpora that could be used for such analyses without substantial changes. There are certainly a number of corpora—some of them quite large—which contain a selection of literary texts, but to be useful for corpus stylistic analyses of literary genre, they need to be heavily adapted or entirely rebuilt for reasons which will be outlined and become clear in this discussion. This is especially true when the object of interest are historical texts.

A case in point is the literature of the eighteenth century, a period which is particularly interesting from a literary point of view since during that era the novel as we know it today started taking shape as a genre or 'super-genre'

---

[1] It should probably be pointed out that whenever this paper refers to 'corpus studies', what is meant are analyses in literary corpus stylistics, and not the closely related field of stylometry (also called computational stylistics). Although there may be some overlap between the two approaches, broadly speaking, *stylometry* tends to focus more on *describing and distinguishing* kinds of texts and literary styles through quantitative parameters—authorship attribution is a typical application—and showing *clusters or networks* of texts within a genre or other group of texts, whereas in literary corpus stylistics the quantitative data is used to enrich *interpretations* of texts. For studies in stylometry see inter alia Hoover (2007), or Peng and Hengartner (2011).

[2] Among the small number of literary corpora that are currently freely available for download in packaged form is for instance the *Corpus of English Novels* (De Smet)—covering the English novel from the late nineteenth century to the beginning of the twentieth century. I could find only one corpus stylistic study by Dillon (2007) that includes the description of two literary genres, namely romantic fiction and erotic fiction, and another one by Gerbig (2008) that includes travel fiction beside non-fictional texts on travelling.

as the case may be (see below). Yet there is no ready assembled specialised corpus of literary texts available for this period.

In this article I will introduce the C18P project, describe how it seeks to fill this gap, as well as suggest some potential applications for the corpus. Before going into the particulars of this corpus, however, I will address the question of why existing corpora proved unsuitable for a corpus analysis of literary genre. The need for a corpus representing the state of pre-Victorian prose fiction from the eighteenth and early nineteenth centuries originally arose in the course of my PhD project, which analyses the characteristics of early Gothic fiction integrating literary and corpus stylistic methods. To perform my analysis a reference corpus of general fiction from around the same period was needed but the corpora that are available proved to require substantial reworking.

One of my central points will be that most corpora including British literary texts were designed by linguists for linguistic research. This makes certain changes necessary when the corpora in question are adapted for questions of a more literary bend, or it might even be necessary to build a corpus 'from scratch' to suit the requirements of researchers in literary stylistics and literary studies more broadly. In the following sections I want to explore how the notion of what is perceived as representative changes, depending on the field of study that the corpus is built in and the purpose it is conceived for, and more specifically how corpora designed by and for historical linguistics differ from corpora needed to do literary stylistics. Then the design and corpus make up of C18P will be introduced, and the final section will briefly discuss the results of a pilot study on Gothic fiction (see also Gemeinböck, 2015) and suggest further potential uses of C18P.

2. Representativeness, context, and purpose

> The guiding principles that relate corpus and text are concepts that are not strictly definable, but rely heavily on the good sense and clear thinking of the people involved, and feedback from a consensus of users. However unsteady is the notion of representativeness, it is an unavoidable one in corpus design, and others such as sample and balance need to be faced as well. (Sinclair, 2005: 5)

Representativeness is a central concept in corpus linguistics and in particular with regard to corpus building, given that corpora are models of a particular variety of language or the state of languages at a particular stage in history more generally; they model the characteristics of that variety or, in other words, they *represent* the variety. However, representation is not a simple matter of 'mirroring' as closely as possible an objectively existing variety of language, but what is perceived as representative—and therefore a good

model—will depend on the field of research that the corpus is designed for as well as which kinds of research questions the corpus is designed to answer.

Many currently existing 'reference corpora', such as the *British National Corpus*, the *Corpus of Contemporary American English*, the *International Corpus of English* and many more have been assembled by linguists, designed to fulfil criteria that are useful and meaningful to linguists. These corpora along with research papers on corpora and corpus building (see for example Biber, 1993; Wynne 2005) have influenced what is by many researchers now considered good practice and the state of the art in corpus building.[3]

Most of the criteria that have been suggested to ensure that a corpus is as representative as possible or to make the design of a corpus as transparent as possible are of course also applicable when assembling a corpus of literary texts. Sinclair, for example, suggests six steps to ensure representativeness, which can be roughly summarised as: 1. selecting a well-defined variety or varieties of language to create a framework of corpus components; 2. drawing up an inventory of text types for each component; 3. deciding which text types are particularly central and which more marginal; 4. estimating a target size for each component and for the corpus overall; 5. monitoring how closely the corpus approximates the initial plans as it comes together; 6. documenting the process for future users (Sinclair, 2005: 12). It is probably safe to say that these steps are relevant to any kind of corpus building endeavour. However, some common practices interfere with what would be perceived as 'representative' from a literary studies' perspective, or at least my perspective as someone from the field of literary studies (see below).

Before addressing those points in which the expectations of linguists and literary critics probably diverge, I want to briefly point out in how far what is perceived as a representative selection is also influenced by the kind of variety to be represented, which in turn has a bearing on *the number of texts* expected. Broadly speaking, there are corpora that aim to model the state of a more extensive kind of language, such as British English in the 1990s, and corpora that focus on more specialised languages, such as the language of fiction during the same period. In the former case, the corpus will comprise many different components of spoken and written registers, domains, or genres, each of which will only need a small number of texts to be representative within the context of this large corpus. For example, the category of fiction of the ICE-GB contains twenty texts. The small number of texts included per category means that it will not make sense to introduce any fine-grained distinctions within a category, so there will be no sub-genres in the category of fiction. Therefore, the characteristics that can be extracted will be fairly general features of fiction. The individual parts of such corpora are also

[3] It should, however, also be pointed out that there are of course many controversies about what is good practice and appropriate in which use case. In addition, what is considered good practice changes considerably over time (see inter alia Kilgarriff, Atkins and Rundell 2007).

often too small to be used as stand-alone corpora by themselves. That is to say, the twenty texts of fiction in the ICE-GB are not sufficient in and of themselves to constitute a corpus representing the state of fiction in the 1990s. However, as a component of a larger corpus, twenty texts can be sufficient to represent a genre or other type of text.

By contrast, more specialised corpora that aim to represent a more restricted variety of language, such as a corpus of 1990s fiction, tend to be larger than the corresponding components of reference corpora but are often smaller than the overall reference corpus because of their narrower scope. Some specialised corpora are structured into sub-genres so that the corpus has different components that can be compared with each other. This enables the researcher to analyse more fine-grained characteristics of the language of, for example, different kinds of 1990s prose fiction and to contrast different kinds of fiction. Therefore, the overall number of texts needs to be quite large, but again, as with general purpose corpora, there may be some components that do not have as many samples because they are part of a larger whole. It is also worth noting that some specialised corpora do not have any internal subdivision, especially if the variety they try to represent is already very narrow and they are designed to be compared to one of the existing larger reference corpora.

How large a corpus should be to be representative thus depends on the kind and scope of the variety to be modelled and the kind of research interests it is designed for, and hence corpora can vary quite dramatically in terms of their size. However, it should have become apparent that unless the corpus in question is very large, it is not generally speaking a good idea to separate an individual component from the rest and use it as a stand-alone corpus. So to compile a specialised corpus it is very often not enough to simply use part of an already existing corpus by itself and this is one of the reasons why it was necessary to build the *Corpus of Eighteenth-Century Prose Fiction*, which will henceforth be referred to as 'C18P'.

There simply were no specialised corpora of narrative fiction from that period and while the *Corpus of Late Modern English Texts* does contain a comparatively large number of texts of narrative fiction, there are not enough to serve as a corpus in and of itself. Furthermore, there is no systematic subdivision into components for the fiction part of the corpus, a point which will be discussed in more detail in the following section together with other issues that make existing corpora of fictional texts less than ideally representative from a literary studies perspective.

3. A literary perspective on corpus building

Apart from the insufficient number of literary texts they contain, there are essentially three main methodological issues with existing corpora including

fictional texts and their design that run counter to what many literary critics would probably expect. All of the following propositions for steps to ensure that a corpus will be perceived as representative by users from the field of literary studies might seem obvious and somewhat trivial—hence commonsensical and not worth being dwelt upon. However, there are to my knowledge currently *no* corpora of literary texts that fulfil these criteria, and certainly none for eighteenth-century prose fiction. The following section will outline the proposed adjustments and in how far available corpora fail to meet them.

As already explained, potential expectations of future corpus users should be taken into account when building corpora and undertaking analyses of user needs has become common practice (see also Králík and Sulc, 2005; Santos and Frankenberg-Garcia, 2007). While carrying out the laborious process of interviewing prospective users is beyond the scope of the C18P project, the theoretical considerations underpinning the corpus design should aim at fulfilling its future users' requirements. Since C18P was conceived explicitly as a corpus that should serve the interests and needs of researchers doing corpus stylistics of literary texts and literary critics, it tries to take into consideration which corpus components might be expected by these groups. My project exploits basic shared knowledge and shared concepts from the field of literary studies—such as literary periods and genres—to construct a framework to make the corpus suitable for and more readily accessible to studies in literary stylistics. Some theoretical observations on the design of the corpus will be made in this section, while the next section will focus on their practical implementation in C18P.

The first issue concerns the division of the texts in the corpus into manageable groups that can be compared. According to Lee, most corpus studies—and presumably also most corpus building projects—rely on genre as one of the categories that texts are divided into, or on linked concepts such as register, text type, domain, style or sublanguage (Lee 2001: Introduction).[4] Obviously many of the available categories are from the field of linguistics and not widely used in literary studies, with the exception of 'genre'. However differently this concept has been treated in linguistics and literature, literary genre is also a notion used in everyday language and very rich in meanings that can be exploited to find a sort of tentative common ground between literary and linguistic approaches.[5] Lee's very 'generous' definition of genre, for instance, can also be brought into congruence with the literary use of the term: genre is a category based on criteria external to the text (as opposed to

[4] Lee however also remarks that these categories are rarely used in a systematic fashion in corpus building and that a systematic distinction between domains, genres and subgenres is often absent so that genres and subgenres are frequently jumbled together on the same level of categorisation (Lee, 2001: Genres in Corpora).

[5] For a view on problematic issues surrounding the use of 'genre' in corpus linguistics versus literary studies see Mauranen (1998).

categories formed on the basis of the properties of the language the texts use) and used to speak about texts as members of a culturally shared grouping of texts. The notion of genre is thus used to understand texts as cultural artefacts with a shared purpose and with a shared textual structure. In Lee's view the term genre also brings into focus the ideological and social purposes of texts (Lee, 2001: Genre, Register, and Style), a perspective that very much resonates with current approaches to genre in literary studies, which have variously seen genres as social institutions and shared modes of interpreting textual worlds.[6]

On a more practical note, Lee further suggests various advantages of using genre as a category in corpora. Firstly, it makes navigating a corpus much easier for users and facilitates genre-based analyses, that is, analyses that look into how language varies due to social and situational constraints or other genre constraints. In addition, using a category as widely understood as genre—despite the vagueness and controversies attached to the concept—makes it easier to quickly ascertain the composition of a corpus and whether the selection of the corpus builder is deemed suitable for the researcher's purposes (Lee, 2001: Introduction). This can be very valuable when trying to judge whether a corpus can be used for a project or if and how it can be adapted. Corpora that use less intuitive categories or very broad categories are difficult to evaluate in terms of their representativeness and usefulness in a particular research context.

Lee also remarks that genre seems to be what is called a basic-level category in prototype theory, making it an intuitively accessible category and therefore very powerful (Lee, 2001: Genres as Basic-Level Categories in a Prototype Approach). However, what is perceived as 'basic-level' will of course vary from context to context. To wit, Lee uses 'the novel' as an example for a basic level concept and thus as its own genre (ibid). To a certain extent this is an understandable decision, since in a larger corpus novels as prototypical prose fiction will be contrasted with other kinds of literary texts, such as plays and poetry. When taking a rather broad perspective and using form as the criterion for categorising literary texts, the novel can be seen as a basic level category. However, when switching perspectives to a field specialising in the study of literary texts and novels in particular, 'the novel' loses its status as basic level category and becomes what Lee calls a 'super-genre', a level above the basic level. Today novels make up a substantial part of the production of prose fiction, making novels too large and heterogeneous a group to be a basic level category. Synchronically as well as diachronically there are clearly distinctive novelistic genres all with their own cultural meanings and prototypical features. Especially for a corpus of literary texts or prose fiction more concretely, it seems therefore more appropriate to rely on

[6] For more on literary approaches to genre see, for example, Bawarshi and Reif, 2010; Beebee, 1994; Frow, 2001.

novelistic genres and other literary prose genres, such as the sentimental novel or the travelogue, which form the basic-level within literary studies.

So far, there has been no systematic attempt to apply literary genre labels to corpora of eighteenth-century texts, which might be due to the purposes these corpora were designed for, or to the considerable effort required to label all the texts for genre in a systematic way. The *Corpus of Late Modern British and American English Prose* (COLMOBAENG), for instance, only uses the label 'fiction' for what appears to be a selection of various kinds of novels and other short prose fiction. The *Corpus of Late Modern English Texts* (CLMET) uses 'Narrative fiction' as genre and somewhat confusingly 'FICT', which presumably stands for 'fiction', as sub-genre with only three texts also having the sub-genre labels 'FICT/TRAVEL' or 'FICT/TREAT'. In addition there are free-form text notes for a few of the texts that state either the form, such as "epistolary novel" or "short stories", or the literary genre, such as "science fiction", but these are sporadic and cannot be used to divide the texts into coherent groups to be compared. Furthermore, as mentioned above, a lack of internal structure within a large group of texts makes it hard to judge the contents of those groups and to assess if the corpus in question is suitable for answering a given research question. Thus, to make existing corpora accessible for studies of literary genre and form, they have to be substantially reworked.

The second issue with existing corpora concerns the use of extracts from texts, which is common practice in corpus building with a focus on linguistic research. Many corpora consist entirely of extracts of a certain length, such as 2000-word segments, which is for example the case with the *Century of Prose Fiction Corpus*. This is a measure to balance the corpus and ensure that no individual writer's style dominates. It also facilitates making comparisons between corpus components since they are all exactly the same size so that a normalisation of word frequencies is not necessary. Other corpora, such as the first version of the CLMET, use extracts where the complete text would exceed the given word limit per author (200,000 words in the case of CLMET).

In any case, there are many researchers in corpus linguistics who caution against using extracts, since as Sinclair points out, it is not safe to assume that an extract from a text is representative of the complete text (Sinclair, 2005: 11). For prose fiction it may be assumed that the results when using extracts from the beginnings of texts would be very different than, for instance, when using extracts from the endings. Any difficulties arising from disparities in the length of texts are secondary to the losses of information about entire sections of text that occur when using only relatively short extracts. Instead, it is better to produce a corpus that is large enough so that the effects of particularly long texts can be evened out by virtue of the overall size of the corpus (Sinclair, 2005:11).

Since the aim of C18P is to represent prose fiction and not to focus on the stylistic choices made at the beginnings or other specific parts of the texts, complete texts have been used, with the exception of cases where only certain volumes of a text were available and the text was judged important enough to be included despite its incompleteness. To ensure that the corpus would still represent genre styles and the style of certain periods rather than the individual styles of a few dominant writers, a word limit of approximately 250,000 words per author was decided on, and texts falling within that limit were preferred over ones violating that limit if there was a choice of several texts by the same author.[7]

Finally, and perhaps most trivially, in terms of the periodisation scheme corpora use, existing corpora tend to use linguistic periods, such as 'Late Modern English', as their basic framework. As sub-categories they might either simply use decades or the duration of generations, i.e. 70 year segments—as is the case with the CLMET. From a literary studies perspective it is of course more desirable to use literary periods, such as Postmodernism, and other eras that are seen as possessing a distinctive literary style, like the Restoration, as the basis for the selection of a target period and for further subdivision within the corpus. In addition, in the case of eighteenth-century prose fiction, there is a link between the literary periods and the number of publications (see below), so that a division into literary periods helps reflect this development.

The C18P corpus tries to address the three issues described above so that the corpus can be easily used by researchers interested in the characteristics of literary styles at the dawning of the age of prose fiction.

4. Issues with using existing corpora in my project

Having discussed issues with existing corpora from a literary studies perspective in general, the next section will turn to some of the implications using one of the existing corpora would have had in the concrete case of my project, namely the analysis of early Gothic fiction as a literary genre. Firstly, in existing corpora the number of words and the number of prose fiction texts for the target period from 1700 to 1830 of my project is not sufficient. Two of the corpora covering the eighteenth century, COLMOBAENG and the *Century of Prose Fiction Corpus*, consist entirely of extracts from texts so that

[7] The original version of CLMET by contrast has a word limit of 200,000 words per author. The lower word limit might reflect the fact that the CLMET also contains a substantial number of non-literary texts, such as pamphlets and letters, which might not be as long as many novels and other prose fiction. However, for the purpose of including novels in their entirety the word limit had to be raised somewhat (version 3.0 of CLMET has entirely abandoned the original word limit, which introduces other problems, see section 5).

their number of words is not sufficiently large in comparison to that of my group of 24 complete texts of Gothic fiction. To compare a target corpus to a reference corpus, the number of words of the reference corpus should ideally be larger than that of the target corpus, since the purpose of the reference corpus is to represent a broader variety of language than that of the test group. However, COLMOBAENG for instance has an overall number of only 372,000 words to represent the period from 1700 to 1799 and the *Century of Prose Fiction Corpus* has 500,000, while my Gothic corpus has over 2 million words. In addition, there is no information on how the extracts in those corpora were chosen, so that to judge what kind of data the corpora contain, each extract would have to be examined with regard to its place in the original text in order to identify patterns that might have been applied in the selection process. This also makes expanding these corpora difficult, since they can only be supplemented by further appropriately selected extracts, so as not to introduce any imbalances.[8]

Even the largest existing open-access corpus (CLMET 3.0), which is superior in the number of words to the other two corpora, has only 45 texts, of which 8 are Gothic fiction, leaving merely 37 texts to serve as a reference corpus. This number of texts is not large enough to adequately represent the variety of genres and writers during that period in a specialised stand-alone corpus of prose fiction texts covering the literary production of more than a century. This becomes all the more apparent when compared to the number of prose fiction texts produced in Britain, which experienced a steep rise at the end of the eighteenth century. Raven estimates that 1,421 works of prose fiction were published between 1770 and 1799 alone (2000: 26), so that in light of the range of authors and genres during the era in question, a wider selection of texts is needed. In addition, the fewer texts, the more danger there is of a bias of one kind or another. So to safely draw conclusions in how far Gothic fiction is different from other contemporary prose fiction genres, a larger number of texts to which to compare the Gothic group was imperative.

Concerning the labelling of literary genres, to the best of my knowledge none of the existing corpora covering the eighteenth century provide any systematic genre labelling. When working with prose fiction texts and thinking of the ways in which they can be divided into meaningful groups, literary genre certainly must be one of the most obvious categories to use. The lack of any such sub-division makes working with a corpus—although not impossible—very difficult in practice. Being faced with the undefined mass of texts in CLMET 3.0 and COLMOBAENG at the outset of my project, for instance, it was impossible to gauge from the list of titles and authors whether the most important genres of the eighteenth century, such as the sentimental

[8] For reasons why it is more desirable to use complete texts in any case, see section 3 above.

novel, Gothic fiction, or adventure and travel fiction, were represented in sufficient numbers and if so in which proportions. This means that even when working with existing corpora, the first step in using them for an analysis of literary genre is to roughly subdivide them into genre groups to be able to judge if the corpus in question is suitable for a project or whether it needs to be extended in any way. While COLMOBAENG does not use any literary genre labels at all, in CLMET 3.0 the only literary genre label denoting manageable groups that was applied systematically in the table describing its data was 'children's book'—albeit not in the genre field, but only as an additional note to the actual genre label 'narrative fiction'. This, as already pointed out, is of course a consequence of the fact that CLMET 3.0 is not designed for fine-grained analyses of literary genre and therefore only classifies texts as 'narrative fiction', while literary genre is treated rather as an afterthought or sporadic additional comment.

After a preliminary assessment of CLMET's distribution of texts between genres, it became evident that most groups, such as travel and adventure fiction or historical fiction, had to be supplemented heavily and that some genres were almost entirely missing, especially political satire from the beginning of the eighteenth century. Furthermore, some authors like Richardson are decidedly over-represented at 1,206,567 words, making up approximately a fifth of the total word count for prose fiction from 1700 to 1830 and thus running the risk of biasing the corpus towards Richardson's style, whereas some important authors like Defoe are conspicuous by their absence. From a history of literature perspective, therefore, CLMET 3.0 does not represent the period satisfactorily, which is not surprising given the project's focus on linguistic rather than literary analyses.

To summarise, the eighteenth century is a period for which no specialised open-access corpora of literary texts exist, making corpora originally conceived for linguistic research, such as CLMET 3.0, the only resources available. In order for these corpora to be useful for literary corpus analyses, however, they have to be extended and adapted to such an extent that the result is in effect a new corpus. This was the point of departure for the C18P project, which seeks to provide a specialised corpus of prose fiction texts of sufficient size to be used as a stand-alone corpus in projects focusing on the interpretation of literary texts and genres from a corpus perspective. The next sections will outline the design rationale for C18P and discuss the make-up of its components.

5. Corpus design of C18P

The *Corpus of Eighteenth-Century Prose Fiction* (C18P) aims to represent the variety of genres and authors in the production of prose fiction during the early stages in the development of the novel, i.e. post-Restoration up to the

Victorian era. It includes both seminal works and authors which are frequently mentioned in reference books on the period as well as some less well-known texts and writers for 'balance'.[9] As already mentioned, the corpus was originally conceived as a reference corpus for my PhD project, which looks into the stylistic characteristics of early Gothic fiction from a corpus stylistic perspective.[10] To extract the characteristic features of Gothic fiction, the reference corpus has to contain a broad range of prose fiction from around the same period and from the era before. C18P is such a corpus that contains a wide range of prose fiction texts from 1700 to 1830.

Concerning its structure, the corpus is divided into three periods as a temporal framework. In addition, the texts are labelled for literary genre, form, and the gender of the author, dividing the corpus into further groups which can be compared and contrasted.

The periodisation scheme has been adopted from the *Norton Anthology of British Literature* (Greenblatt and Abrams, 2006 Vol. 1: 2070) and is also mirrored in Raven's account of the era (Raven, 2000: 27-35). The first period covers the years from 1700 to 1745 and its focus rests on short prose fiction, with a large amount of political and satirical fiction by writers like Addison and Arbuthnot, as well as early examples of adventure novels and travel fiction by Defoe, Swift, and Chetwood (Greenblatt and Abrams, 2006 Vol. 1: 2075-2077). The second period stretches from 1746 to 1785 and initiates what the *Norton Anthology* describes as the "age of prose", with a heavy focus on sentimental novels and some picaresques, fictional biographies and Oriental tales as well as the first examples of Gothic fiction (Greenblatt and Abrams, 2006 Vol. 1: 2077-2080). Famous writers from this period include, for instance, Burney, Fielding, Goldsmith, Lennox, Pratt, Richardson, Smollett, and Sterne. The last section is the Romantic period and covers the years from 1786 up until the beginning of the Victorian era in 1830 (Greenblatt and Abrams, 2006 Vol. 2: 1-22). Here the emphasis is on Gothic fiction and some historical fiction. As should already have become obvious, there is a link between genres and periods, with each period focusing on particular genres, but this connection is not exclusive so that there are, for instance, satirical texts in all three periods and sentimental novels in the last two periods.

The framework of the corpus is thus based on genre categories and a periodisation scheme from a literary studies background that will be meaningful to literary critics familiar with eighteenth-century literature. Being built on the shared knowledge of potential users from the field of literary studies, the

[9] See inter alia Greenblatt and Abrams, 2006; Raven, 1987 & 2000; Richetti, 1996; Watson, 1971.

[10] The corpus will of course be made available including the sub-corpus of Gothic fiction, which forms a substantial part of the corpus component from 1786-1830, when early Gothic fiction was at its pinnacle.

corpus should be well-suited to aid enquiries into the style of the prose fiction of the pre-Victorian era.

6. Corpus make-up of C18P

Overall, C18P currently consists of 143 texts and approximately 9.7 million words. As Table 6.1 and Figure 1 show, there is a noticeable rise in the number of texts per subperiod. This is by design and has two reasons, one of them being practical: there are more digitised texts available for the later periods and if all components had to be of equal size—for instance so that they are easier to compare—then the lowest number of texts per period would be the limit for all periods. This would mean a great loss of data for the later periods, for which more texts are available. Secondly, to represent the development of prose fiction writing and publishing it is important to take into account current knowledge on publication data from that period.

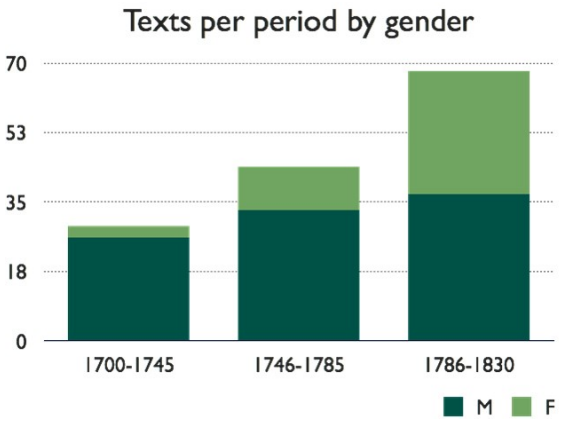

Figure 1.

According to data by Raven (2000: 26-27), although the number of novels published per year fluctuates, overall there is a considerable rise in the number of texts published from the 1780s onwards, with an especially sharp increase in the late 1790s—the dawning of the "age of prose" (Greenblatt and Abrams, 2006 Vol. 1: 2077). C18P reflects that increase in publications in general, and it also aims to represent the increase in the number of texts published by female authors. Raven states that publications by female writers rose quite dramatically at the end of the eighteenth century to equal or even surpass the number of prose fiction texts published by male authors (Raven, 2000: 48). As evident from the data in Figure 1, C18P is still somewhat biased towards texts by male writers due to a greater availability of texts, but with 31 texts by female writers in the Romantic period and 37 by male writers, the count is close to even for the era when the number of female authored texts soared.

Period	Authors	Texts	Words
1700-1745	12	30	1,001,417
1746-1785	32	44	2,984,203
1786-1830	45	69	5,729,918
Total	89	143	9,715,538

Table 6.1. Texts and word count by period.

The majority of the texts was retrieved from the *University of Oxford Text Archive*, and a lesser number from other databases, such as *Project Gutenberg* and *Project Gutenberg Australia*. Concerning the editions used, there is usually only one version of the text available in the *University of Oxford Text Archive* and information on the edition is documented in the texts' metadata provided on the OTA-website. On Project Gutenberg, there are sometimes several digitised transcriptions of texts, which may be based on different editions. Therefore, the original *Project Gutenberg*-text ID is given in the description file of C18P, so that researchers interested in which edition was used in the corpus can find the relevant information. The orthography of the texts has been left untouched, with the exception of an automatic substitution of the modern short *s* for any instances of long s-letters.

As already stated above, another classification marker in C18P, apart from the three periods, is literary genre. Regarding prominence in the corpus, the most important genres are: satire, Gothic fiction, political fiction (including Jacobin novels), sentimental novels, fictional biography, historical fiction, travelogues and adventure fiction, didactic and moral fiction, picaresques, scandal fiction and *roman à cléf*, novels of manner, and oriental fiction (see Table 6.2). It should be noted that the large number of satirical and political texts is mainly due to the fact that most of those texts are short fiction, rather than full-length novels. In addition it is worth pointing out that one text may have several genre labels, since literary texts rarely confine themselves to only one genre, but in most cases participate in several genres (compare Frow, 2006: 45).

Genre/Form	Number of texts
Satire	31
Gothic	24
Political, Jacobin	21

Genre/Form	Number of texts
Sentimental novel	19
Fictional biography	16
Historical	12
Travel, adventure	11
Didactic, moral fiction	9
Picaresque	6
Scandal, *Roman à cléf*	5
Novel of manners	4
Oriental tale	4
Short fiction	48
Epistolary novel	21
Other novels	74

Table 6.2. Genres and forms in C18P.

One problem regarding literary genre, or any labelling for genre based on criteria external to the text—that is to say criteria rooted in presumably shared cultural knowledge—is that ultimately the decision which genres to attribute to a text and which genre labels to use in the first place is a somewhat subjective choice. For the labelling of texts in C18P a number of reference books were consulted,[11] but even so the labelling can of course never be 'objective' and always remains a matter of individual judgement. However, as Lee remarks, even this subjective categorisation into genres is preferable to no such grouping (Lee, 2001: The BNC Bibliographical Index) and subsequent users may adapt the labelling to suit their sensibilities.

Probably less contentious than genre is the labelling for literary form, with the categories of *short fiction*, *epistolary novels*, and the rather broad category of *other novels* being used. This classification of the texts by form together with the other category labels of sub-periods and genre, as well as the fact that the texts are tagged for the gender of their author, enable researchers to perform contrastive analyses between two or more groups (see below). The final section of this paper will discuss some of these potential research applications for C18P.

[11] See Burwick, 2012; Day and Lynch, 2015; Greenblatt and Abrams, 2006; Punter, 2012; Punter and Byron, 2004; Richetti, 1996; Watson, 1971.

7. Potential uses

At 10 million words, C18P is a medium sized corpus that—like the CLMET (De Smet, 2005: 78)—falls somewhere in between small but highly annotated corpora, such as ICE-GB or the Helsinki corpus and modern mega-corpora like the *Corpus of Contemporary American English*. Regarding the research questions the corpus can be used to answer, this means that many of the corpus components are large enough for quantitative research, such as an analysis of keywords, collocations, or characteristic syntactic patterns. For instance, the differences between periods, forms, as well as the difference in the styles of male and female writers can all be explored with quantitative methods. Similarly, many of the genre groups are large enough to warrant quantitative analyses. Although there is no fixed rule about how large a corpus component has to be to be usefully analysed with quantitative-statistical methods, judging from existing research, samples of around twenty to thirty texts seem to be suitable, provided of course that the group is reasonably homogeneous.[12]

To give a concrete example of such a genre-based analysis using C18P, as mentioned above, the aim of my PhD project is to analyse Gothic fiction using a corpus stylistic approach. So far, a pilot study of the keywords of Gothic fiction, with the Gothic genre being compared to the rest of C18P using the Mann-Whitney U-statistic, has yielded promising results. The top ten keywords include evocative items such as *midnight*, *fled*, *hastened*, and *trembled*. Further investigation showed that several groups of semantically related items are strongly represented, such as words referring to motion, strong emotions, auditory as well as visual perception, and parts of the body (see Table 7.1). These groups can all be related to central themes in Gothic fiction, like pursuit and escape, extreme psychological states, and highly stylised gestures (see also Gemeinböck 2015), and can form the point of departure for further in-depth analysis and interpretation of the genre.

Group	**Example keywords**
Motion	*hastened*, *rushed*, *darted*, *fled*, *escape*, *approached*, *followed*
Emotions	*surprise*, *anxious*, *terror*, *horror*, *impatience*, *dreaded*
Perception	*marked*, *perceived*, *watched*, *regarded*, *listened*, *sounds*
Parts of the body	*bosom*, *brow*, *eyes(s)*, *arms*, *lips*, *ear*

Table 7.1. Selection of keyword groups.

[12] See for instance the work by Bednarek 2012, Fischer-Starcke 2010.

Other similar corpus stylistic studies that could be undertaken using C18P include an analysis of sentimental novels, satirical fiction, or fictional biographies. Furthermore, if desired the other smaller genre groups can be extended to approximate the appropriate sample size of around twenty to thirty texts to match the larger existing groups and then any characteristic patterns of interest, such as keywords, n-grams, or syntactic patterns can be statistically analysed.

As De Smet states about CLMET, even the smaller components of the corpus that cannot be used for quantitative analysis are still useful for finding quotations and references to particular words or phrases of interest and can thus be practical for use in qualitative studies with the corpus serving as a kind of quotation database (De Smet, 2005: 80).

Overall, C18P is a valuable addition to the range of existing corpora, providing the means to perform quantitative analyses in the field of eighteenth-century literary studies, a highly interesting period in the development of prose fiction and the novel in particular. It uses concepts from literary studies as basic building blocks, making corpus stylistics more accessible to literary critics. The fact that the sample texts are labelled for literary genre and form makes it easy to understand and its contents more transparent than is the case with many existing corpora. In addition the corpus should be easily extendible and at 9.7 million words it is also quite a substantial corpus of literary texts that can serve as a starting point for any quantitative exploration of eighteenth-century prose fiction.

The corpus will be available shortly via GitHub.[13] Until the release version of the corpus is ready, preliminary versions can be obtained from the author directly (see contact details).

Acknowledgements

I wish to thank everyone who attended my presentation of this paper during the *Digital Literary Studies* conference 2015 at the University of Coimbra and gave me feedback or made suggestions for improving my work, either during the session or over coffee. I am particularly grateful to Francesca Frontini, Leif-Jöran Olsson, and Jan Rybicki for their help. I also would like to thank the three reviewers and the editors who took the time to read my paper and make numerous suggestions for improvement.

[13] The corpus will be available from the following url: https://github.com/antiquary/c18p.

References

BAWARSHI, Anis S. and Mary Jo Reif (2010). *Genre: An Introduction to History, Theory, Research, and Pedagogy*. West Lafayette: Parlor Press.

BIBER, Douglas (1993). "Representativeness in Corpus Design." *Literary and Linguistic Computing* 8.4: 243–257.

BURWICK, Frederick, ed. (2012). *The Encyclopedia of Romantic Literature*. Chichester: John Wiley.

CULPEPER, Jonathan (2009). "Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet."*International Journal of Corpus Linguistics* 14.1: 29–59.

DAVIES, Mark (2004-). *BYU-BNC*. (Based on the British National Corpus from Oxford University Press). 30 Jun. 2015. http://corpus.byu.edu/bnc/.

———— (2008-). *The Corpus of Contemporary American English: 450 million words, 1990-present*. 30 Jun. 2015. http://corpus.byu.edu/coca/.

DAY, Gary and Jack Lynch, eds. (2015). *The Encyclopedia of British Literature 1660 - 1789*. Chichester: John Wiley.

DE SMET, Hendrik (2005). "A corpus of Late Modern English text." *ICAME Journal* 29: 69–82.

———— (n.d.). *The Corpus of English Novels (CEN)*. 15 Mar. 2015. https://perswww.kuleuven.be/~u0044428/cen.htm.

DE SMET, Hendrik, Hans-Jürgen Diller, and Jukka Tyrkkö (2013). "The Corpus of Late Modern English Texts, version 3.0." 29 Jan. 2015. https://perswww.kuleuven.be/~u0044428/.

FANEGO, Teresa (2012). "COLMOBAENG: A Corpus of Late Modern British and American English Prose." *Creation and use of historical English corpora in Spain*. Ed. Nila Vázquez. Newcastle upon Tyne: Cambridge Scholars Publishing: 101–117.

FISCHER-STARCKE, Bettina (2010). *Corpus Linguistics in Literary Analysis: Jane Austen and Her Contemporaries*. London: Continuum.

FROW, John (2005). *Genre*. Oxon: Routledge.

GEMEINBÖCK, Iris (2015). "Containing chaos: compiling a corpus of eighteenth century prose fiction." *On-line Proceedings of the Annual Conference of the Poetics and Linguistics Association (PALA)*. 29 Jan. 2016. http://www.pala.ac.uk/uploads/2/5/1/0/25105678/gemeinboeck_iris.pdf.

GREENBLATT, Stephen and M.H. Abrams, eds. (2006). *Norton Anthology of English Literature*. New York: Norton.

HOOVER, David L. (2007). "Corpus Stylistics, Stylometry, and the Styles of Henry James."*Style* 2.41: 174–203.

"ICE-GB Corpus Design" (28 May 2015). *The International Corpus of English – Britain*. University College London. 30 Jun. 2015. http://www.ucl.ac.uk/english-usage/projects/ice-gb/design.htm.

KILGARRIFF, Adam, Sue Atkins and Michael Rundell (2007). "BNC Design Model Past its Sell-By." *Corpus Linguistics Conference*, Birmingham, UK, 2007. 8 Dec. 2015. http://www.kilgarriff.co.uk/Publications/2007-KilgAtkinsRundell-CL-Sellby.pdf.

KRÁLÍK, Jan and Michal Sulc (2005). "The Representativeness of Czech corpora." *International Journal of Corpus Linguistics* 10.3: 357-366.

LEE, David Y. W. (2001). "Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path Through the BNC Jungle." *Language, Learning & Technology* 5.3. 24 Mar. 2015. http://llt.msu.edu/vol5num3/lee/.

MAHLBERG, Michaela (2007). "Clusters, key clusters and local textual functions in Dickens." *Corpora* 2.1: 1–31.

MAURANEN, Anna. (1998). "Another look at genre: corpus linguistics vs. genre analysis." *Studia Anglica Posnaniensia: International Review of English Studies*: 303.

MILIC, Louis T. (1995). "The Century of Prose Corpus: A half-million word historical database." *Computers and the Humanities* 29: 327–337.

O'HALLORAN, Kieran (2007). "The subconscious in James Joyce's 'Eveline': a corpus stylistic analysis that chews on the 'Fish hook'." *Language and Literature* 16.3: 227–244.

PENG, Roger and Nicolas Hengartner (2011). "Quantitative Analysis of Literary Styles." *Department of Statistics Papers*, UCLA. 25 Oct. 2011. 8 Dec. 2015. http://escholarship.org/uc/item/883831vz.

Project Gutenberg. 29 Jan. 2015. https://www.gutenberg.org.

PUNTER, David (2012). *A New Companion to the Gothic*. Oxford: Blackwell.

PUNTER, David and Glennis Byron (2004). *The Gothic*. Malden: Blackwell Publishing.

RAVEN, James (1987). *British Fiction 1750–1770: A Chronological Check-List of Prose Fiction Printed in Britain and Ireland*. Newark: University of Delaware Press.

———— (2000). "Historical Introduction: The Novel Comes of Age." *The English Novel 1770–1829: A Bibliographical Survey of Prose Fiction Published in the British Isles: Volume I*. Eds. Peter Garside, James Raven, and Rainer Schöwerling. Oxford: Oxford UP. 15-121.

RICHETTI, John, ed. (1996). *The Cambridge Companion to the Eighteenth-Century Novel*. Cambridge: Cambridge UP.

SINCLAIR, John (2005). "Chapter 1: Corpus and Text—Basic Principles." *Developing Linguistic Corpora: a Guide to Good Practice*. Ed. Martin Wynne: 4–24. 29 Jan. 2015.

STUBBS, Michael (2005). "Conrad in the computer: examples of quantitative stylistic methods." *Language and Literature* 14.1: 5–24.

University of Oxford Text Archive. University of Oxford. 4 May 2015. https://ota.ox.ac.uk/.

WATSON, George, ed. (1971). *The New Cambridge Bibliography of English Literature: 1660–1800*. Cambridge: Cambridge UP.

WYNNE, Martin (2005). *Developing Linguistic Corpora: a Guide to Good Practice*. 29 Jan. 2015. http://www.ahds.ac.uk/creating/guides/linguistic-corpora/.