Containing chaos: compiling a corpus of eighteenth century prose fiction

Iris Gemeinböck

1 Introduction

The *Corpus of Eighteenth Century Prose Fiction* or C18P is a corpus that aims to represent the range of prose fiction that was produced in Britain in the period from 1700 to 1830. The corpus focuses on a period when the novel as a prose fiction form was emerging and female writers became increasingly more productive, fundamentally changing the course of literary production. C18P is divided into three subperiods and in its current version it consists of 146 texts from 18 literary genres by 89 authors, amounting to a total of approximately 9.7 million words. Thus C18P is a valuable resource for researchers interested in this prolific literary period.

My own PhD project may serve as an example to illustrate which kinds of questions the corpus is designed to answer. In my PhD project I am analysing early Gothic fiction from the eighteenth and early nineteenth centuries as a literary genre using corpus linguistic methods. One of the steps in the process of analysis is to extract keywords and to then categorise them by semantic field to gain a better understanding of what the Gothic genre is about from a quantitative point of view. To determine keywords a reference corpus is needed. When performing keyword tests on literary texts, it is not unusual to use general reference corpora, such as the BNC, or a sampler of a reference corpus for this purpose.[1]

However interesting the results achieved by using general purpose corpora can be, there is an important reason not to use them as reference corpora in the specific case of analysing a literary genre: a literary genre does not constitute itself as distinctive against non-literary genres like postcards or student essays, but relative to other literary genres, ideally from the same period. Therefore its linguistic resources should be compared to those of other literary genres. There are of course a lot of differences between non-literary genres, such as essays, pamphlets and letters, and literary genres, such as Gothic fiction, which could be described in great detail, but the result would perhaps not be a very useful portrait of Gothic fiction as a *literary* genre, or in any case it would be a very different undertaking from what I envisioned for my PhD.[2] In order

[1] Sometimes the reference corpus even covers a different period than the test corpus (see for instance O'Halloran, 2007: 233, Gerbig, 2008). Stubbs (2005: 10-11) uses only 'imaginative fiction' in his reference corpus, but it is from a different period than the text he is analysing.

[2] Of course I do not think that there are never any good reasons to use a corpus of non-literary texts when analysing a literary genre. Using non-literary texts, however, would clearly bring other features and characteristics to the fore than those which set the genre in question apart from other genres in the overall landscape of literary texts from a given period.

to describe a literary genre as such and to discuss its characteristics, it seems more productive to compare it to other literary genres and thus understand it against the background of literary writing (not 'general' writing, which is the case when literary texts or genres are compared against general reference corpora). This is also the approach that Dillon chooses, when analysing the language of romance fiction versus the language of erotic fiction; he compares both these genres against the fiction section of the BNC (2007: 170).[3]

One of the reasons why some researchers still sometimes choose to use general purpose reference corpora – when there is no particular reason to prefer them over literary reference corpora – might be that there are not many specialised corpora of literary texts available. Even with resources like *Project Gutenberg*, the *University of Oxford Text Archive*, or *Eighteenth Century Collections Online*, assembling a corpus can be quite laborious, especially since databases often make it difficult to efficiently retrieve groups of texts according to parameters like year of publication, genre and so on. A case in point is *Project Gutenberg*, where the metadata of texts does not contain the original year of publication for texts, which makes compiling a corpus covering a specific period by far more cumbersome. Additionally, the further back in time one looks, the more difficult it is to obtain a sufficient number of texts, let alone to conveniently retrieve them from a single database.

The section above briefly summarises my motivation for building a specialised corpus of eighteenth and early nineteenth century prose fiction. In the following section I want to outline my reasons against using any of the few existing corpora covering this period.

2 Existing Corpora

Currently, there are relatively few free corpora of eighteenth century British texts.[4] There are some databases that offer quite a substantial amount of texts, which can serve as 'raw materials' for corpora, but these databases are not corpora in and of themselves. Databases and collections offering eighteenth and early nineteenth century texts include the *University of Oxford Text Archive*, the *Eighteenth Century Collections Online*, or *Project Gutenberg*. As already indicated, these are not corpora because they are compiled to different standards than linguistic corpora. For example, databases do not usually have a word limit per author, which is needed to balance the corpus and ensure that it is not biased towards the language of a small number of individual authors. Furthermore, the texts of many corpora are subdivided into several categories that are designed to be compared to each other, which is not the case for databases.

So, despite (relatively) easy access to quite a number of source texts, there are comparatively few pre-assembled corpora available. The largest open-access corpus containing eighteenth century British English texts at the time of writing is the *Corpus of Late Modern English Texts*, an extended version of which has recently be made available, comprising over 34 million words. The texts of this

[3]There are currently very few examples of attempts to study literary genre from a corpus stylistics perspective. One is by Dillon (2007) (mentioned above) and there is another by Gerbig (2008), which however does not focus exclusively on fiction, but rather on the theme of travel across various genres, including travel fiction.

[4]Among the existing corpora that are *not* freely available are, for instance, the *Century of Prose Corpus* and the *Corpus of Late Modern British and American English Prose*.

corpus cover a range of genres such as narrative fiction, drama, letters, and treatises. It covers the entire Late Modern English period, so that only the first sub-period (1710-1780) and part of the second sub-period (1780-1850) are of interest for my purposes. In the CLMET version that was available at the time C18P was assembled (CLMET 3.0), the relevant CLMET sub-sections contained 45 texts of which 8, however, are Gothic fiction, leaving 37 texts of non-Gothic fiction.[5]

The first issue therefore was that the sub-corpus of fiction that could be extracted from the CLMET was too small, with only a few texts more than the test group. This could have been remedied by adding more texts and indeed there is a substantial overlap in the selection of texts of C18P and CLMET. The relevant parts of CLMET could, however, still not be used as they were for two main reasons: firstly, CLMET frequently uses extracts from texts to increase the diversity of texts without exceeding the word limit of 200,000 words per author set by the corpus builders.[6] However, for an interpretative analysis of literary texts using short extracts is not desirable, since it would be hard to select an extract to adequately represent the overall meaning of the work – after all different parts of the text fulfil quite different functions. Therefore, for C18P complete texts were used, with only a few exceptions where only parts of the text were available.

The second issue with CLMET – and the two other corpora mentioned in the footnote – was the lack of labelling for literary genre. Without any further subdivision or categorisation in the class of prose fiction, it is impossible to judge the contents of the corpus. Broad labels, such as 'narrative fiction' or 'imaginative fiction' used in CLMET and the *Corpus of Late Modern American and British Prose* respectively, obscure the internal structure of those categories and the criteria for selecting sample texts; they reveal nothing about what to expect from the corpus regarding the contents of those categories. While such broad categories still might be useful if the aim is to contrast narrative prose texts with drama or non-fictional categories, when working exclusively with prose fiction, these labels are not precise enough to be of use. Both to efficiently evaluate if the corpus in question is suitable for a specific research question and to allow for more fine grained analyses, literary prose fiction has to be further subdivided into more manageable, meaningful groups.

Literary genre suggests itself as basic building block for a corpus of literary texts and as Lee points out, 'genre' in general is a very useful category for corpus builders, because it is a notion that is used in everyday life and therefore can be grasped intuitively (Lee, 2001). The fact that genre is a shared and widely understood concept facilitates working with the corpus. A disadvantage is of course that the attribution of genres to texts is always somewhat subjective, even when it is supported by reference books, such as literary encyclopaedias and bibliographies. In addition, using literary genre as a category may have profound implications for the structure and usage of the corpus (this issue is discussed below).

[5]Even the most recent version of CLMET contains fewer prose fiction texts than C18P and the other issues (see below) remain.

[6]The same problem applies to the *Century of Prose Corpus* and the *Corpus of Late Modern British and American English Prose* – they both consist entirely of extracts.

3 Basic components of C18P

Before going into detail about the complexity or complications – in short the chaos – that using literary genre as category adds to the structure of a corpus, I want lay out the basic design and the components of C18P. To reiterate, C18P covers the period from 1700 to 1830 and tries to represent the production of prose fiction in Britain during that era. The period from the eighteenth century up until the Victorian age was a key stage in the early development of what is nowadays called 'the novel', and therefore of great interest in literary studies. From its tentative beginnings in adventure and travel prose fiction, the novel had turned into a recognisable form by the end of the eighteenth century. Within the period covered by C18P, three distinct subperiods are commonly distinguished (see for example Greenblatt and Abrams, 2006). This periodisation is mirrored in the structure of C18P (see Table 1). The first sub-period focuses on satirical writing and early adventure or travel fiction, with writers such as Defoe and Chetwood. The second sub-period starts the age of the sentimental novel and features writers like Richardson or Sterne. Finally, the last period is the Romantic era with a focus on Gothic fiction and historical novels. Important writers from this era include Austen, Edgeworth and Radcliffe.[7]

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 1: Subperiods in C18P

In the *Norton Anthology of British Literature* the decades between 1746 and 1785 are referred to as 'the age of prose' since the output of prose fiction began to increase dramatically during that period (Greenblatt and Abrams, 2006: 2077-2080). Not only did the overall number of works of prose fiction rise, but especially the number of texts by female writers, which adds extra interest to the era as a turning point in literary history.[8] C18P tries to reflect both these trends, so that the number of texts rises from each period to the next and the number of texts by female writers also rises to a point where it is almost equal to the number of texts by male writers in the last period (see Table 2). Historically this is not quite accurate, since according to Raven the number of texts by published female writers per year often surpassed that of male writers, depending on the year, in the last decades of the eighteenth century (Raven, 2000: 48). However, due to availability of texts by female writers the ratio of male to female writing in C18P does not quite reach those proportions.

[7]It may seem counterproductive to include Gothic fiction in a corpus that was originally conceived to serve as a reference corpus to study Gothic fiction. However, since the corpus will be available to other researchers and Gothic fiction was the most published genre of the Romantic period, Gothic texts are included in the corpus. For the purposes of my study, I separate the Gothic fiction from the rest of the corpus, which is then used as a reference corpus (see Section 5). The number of texts is such that there is still enough material to fulfil this purpose more than adequately.

[8]For publication data from the period between 1770 and 1799 see Raven (2000).

Period	Female	Male	Unknown
1700-1745	3	26	1
1746-1785	11	33	0
1786-1830	31	37	1
Total	45	96	2

Table 2: Number of texts by gender of author

To keep the three periods as homogeneous as possible internally, while maximising differences between the periods, texts were selected in such a way that each writer's texts fall within only one sub-period. For instance, Defoe is part of the period between 1700 and 1745, as is Chetwood. Richardson's work is restricted to texts published in the second sub-period from 1746-1785, because that is when he produced most of his novels. This unfortunately also means that *Pamela*, one of his most important texts, which was published in 1740, is not in the corpus. Interestingly and fortunately however, there were few cases in which the restriction of each writer to the period when he or she published most meant having to leave out any key texts. Another restriction to keep the corpus balanced is a word limit of around 300,000 words per author. Initially, the word count per writer was to be 200,000, like for the CLMET, but this turned out too restrictive for an era when a lot of texts were published in multiple volumes and with word counts exceeding 200,000 words by far.

4 Containing chaos: genre labelling in C18P

> **contain** |kənˈteɪn|
> verb [with obj.]
>
> I. To have in it, to hold; to comprise, enclose.: *the corpus contains chaos.*
>
> II. To hold together; to keep under control, restrain, restrict, confine.: *the corpus contains chaos.*
>
> –*Oxford English Dictionary* 'contain'
> (2015, example sentences my own)

While the three major subperiods in the corpus mirror the historical developments in publication of prose fiction texts, as mentioned before, another important complementary principle of organisation in the corpus is the categorisation of texts into literary genres. Given that genre labelling is a common way of organising a corpus into coherent groups, using literary genres for a corpus of prose fiction seems a logical step. Interestingly, there are currently very few corpora that systematically use literary genre labels to group their fiction section. One of these exceptions is the *British National Corpus*, where genre labels were introduced by Lee (2001).

Lee remarks that using genre as an organising principle has several advantages: Lee refers to genre as a 'basic level category', making it intuitively accessible to

corpus users (2001). It thus makes the corpus easier to navigate and facilitates focused search queries as well as comparisons between groups of texts. Finally, as mentioned above, genre labelling also makes the contents of the corpus more transparent. For both corpus users and the corpus builder it becomes easy to judge how balanced the corpus is and whether all the genres and sub-genres that are, for instance, expected in a corpus of eighteenth century prose fiction are actually represented. Thus, gaps can be remedied more easily and the corpus can be adapted to the individual requirements of a project more readily.

Genre labelling is therefore a highly desirable part of corpus meta-data. On the other hand, it adds a few complications and a measure of chaos to the corpus. 'Literary genre' is an elusive concept with a myriad of definitions and uses. Genre is variously understood as a property of the text itself or a quality readers or critics retroactively attribute to texts in an effort to make meaning and to categorise texts into coherent groups that are part of an overall system of kinds of texts; 'genre' has been compared to biological species undergoing an evolution, members of a family, social institutions, and speech acts (see for instance Bawarshi and Reiff, 2010, Fishelov, 1993).

In the present case, while of course all of these possibilities co-exist at the same time, genre is understood as both a category to make meaning of texts by readers and at the same time a shared resource tapped into by writers composing texts. The former is relevant when corpus builders and critics try to bring some order into and to stabilise – at least momentarily – the diversity of texts produced at a given time in history and to divide them into meaningful categories, for instance, to add some perspective on the connection between genre and socio-historical developments. To a certain extent then 'genre' is in this approach is a 'post-mortem', a retroactive attribution of properties to integrate a text into a more or less coherent system of kinds of texts existing during a given period. On the other hand, corpus linguistics assumes a set of concrete, shared resources among language users and writers of a certain period, an ongoing development of common linguistic means to express certain meanings in particular social situations, which can be extracted from texts. The role of corpus stylistics, in this project, as in many projects on non-literary genres, is thus to better understand the connection between an assumed genre from a reader's or critic's perspective and the shared resources found in the texts that signal 'genrehood'. (However, this intention still bypasses the difficult question where to locate genre, but rather dislocates and postpones it.)

Genre	Texts
Satire	31
Gothic	24
Political, Jacobin	21
Sentimental novel	19
Fictional biography, memoir	16
Historical	12
Travel, Adventure	11
Didactic, Moral	9
Picaresque	6

Table 3: Selected genre labels from C18P

On a more practical note regarding the concrete task of building a corpus based on these lofty ideas, there is unfortunately no generally accepted exhaustive list of genres or genre names that can be referred to when assigning literary genre labels to texts. While there are some fairly common labels that are used across a wide range of reference texts and criticism, such as *Gothic novel* or *sentimental novel*, there many alternative terms that might (or might not) have slightly different connotations. A case in point are the terms *sentimental novel* and *novel of sensibility* or *travel fiction* and *travelogue*, which seem to be used synonymously in many cases but might potentially have slightly different uses.

Apart from the fact that there are some closely related genre terms, the list of genre terms is also open-ended. To uphold a semblance of objectivity, various reference books, such as the encyclopaedias and bibliographies by Burwick (2012), Day and Lynch (2015), and Watson (1971) respectively were consulted for the labelling of texts in the corpus. Still, the ultimate decision of which labels to attribute remains somewhat subjective. However, Lee (2001) states that even with a subjective bias, it is still more desirable to have some sort of genre labelling in place than having none at all. Tables 3 and 4 show the most important genre labels, forms, and the number of texts per category in C18P.

Form	Texts
Short fiction	48
Epistolary novel	21
Other novels	74

Table 4: Prose fiction forms in C18P

The most complicated complication, however, concerns the fact that most literary texts participate in several genres. While sometimes it might be possible to agree on a 'main' genre, this is very often not possible and very much depends on how a reader reads the text in question. I therefore decided to allow several genre labels per text. On the one hand, this might reflect the perceived properties of the text more accurately provided that the labels are fitting, but on the other hand it fundamentally changes the working of the internal structure of the corpus. Most existing corpora use one genre category per text, resulting in a simple relationship between texts and group membership. To use a visual metaphor, each text is part of one single container and the contents of individual containers can easily be compared to one another. Furthermore this relationship is stable. In C18P however, one text may be contained in several containers at once; that is to say a text can be a member of several genres, such as travel fiction, sentimental, and Gothic fiction at once. As a consequence, there is no neat structure with clearly separate permanent categories anymore.

This prompts the question whether the resulting categories are not too heterogeneous – if in other words they can yield useful results or if the distinctions are too 'blurry'. In addition, there is the issue of how to use the corpus without causing any conflicts between the groups that are to be compared. The following section then offers insight into which kinds of research questions the corpus might be used for and into the kinds of results that can be expected.

5 Pilot study: keywords of early Gothic fiction

As stated above, C18P is primarily designed for the analysis of genres and forms of prose fiction, and thus the usual way of proceeding would be to single out an individual genre or form that is of interest and to use rest of the corpus as reference corpus. This avoids the conflicts mentioned above that could arise from overlaps between categories due to multiple genre labels per text.

Another concern, apart from the question of how to best use a corpus with such unstable, overlapping categories was the issue of whether such a corpus can deliver useful, meaningful results, given that the overlaps between groups might erode any distinct boundaries between kinds of texts. To ensure that C18P produces useful, interpretable results, a pilot study was undertaken. Such pilot studies cannot, of course, guarantee that all categories will yield useful results, but they do give some measure of assurance.[9]

The group selected for this preliminary analysis was Gothic fiction because after all the purpose of building the corpus in the first place was to create a reference corpus to analyse Gothic fiction. Therefore, all the texts in this category were separated from the rest of the corpus into a test group; then the Mann-Whitney U test was used to determine keywords.[10]

Before commenting on the results of this pilot study, I would like to briefly discuss the composition of the test group. Given that Gothic fiction is the focus of my PhD project it might be suspected that the selection of texts and the composition of the group might have been treated differently than that of other subcategories in the corpus and that this might influence the results obtainable. However, the make-up of the Gothic group is much the same as for other genre categories in the corpus. Exactly the same process of consulting reference books to identify potential texts for inclusion was used as for other categories. Furthermore, the contents of the group is governed by the availability of public domain digital, machine-readable versions of texts, as is the case for all the other genre categories in the corpus. And finally, the texts that make up the Gothic category are just as diverse as those of other genre groups. Most texts in the Gothic group have multiple genre labels: *Caliph Vathek*, for instance, is both 'Gothic fiction' and an 'Oriental tale', and *Caleb Williams* is an example of political fiction – 'Jacobean fiction' to be precise – as well as a Gothic novel.

Despite C18Ps shortcomings regarding the number of texts available and the internal heterogeneity of the groups, the results of the pilot study are very promising. It seems that there are still sufficient and sufficiently distinct genre markers that can be identified, even for fairly complex kinds of texts studied under suboptimal circumstances. In the case of Gothic fiction, even at first glance a lot of typically 'Gothic' items could be identified on the list of keywords: *midnight*,

[9] Due to the fact that this project is undertaken by a one researcher-team no further testing could be done so far. I would, however, be very grateful for more user feedback to be able to improve the corpus (contact details below).

[10] The Mann-Whitney U test was used firstly to avoid biases towards the vocabulary of long texts and to eliminate words that occur frequently in individual texts, such as names of characters and places. Secondly, the distribution of keywords across the groups are taken into account to a certain extent using the Mann-Whitney U test. Word frequencies are replaced with 'ranks', from highest number of occurrence to lowest, which are then summed up for the two groups (test group and reference group), and strong keywords therefore will generally also have good coverage among the texts of the test group, which is highly desirable when trying to identify the shared linguistic resources of a genre.

victim, and *fate* all rank among the top twenty keywords. In addition, further analysis revealed that the keywords extracted cover several key occupations of Gothic fiction, such as the body, the theme of flight and pursuit, or extreme emotions.

The group of words – very often verbs – describing different kinds of motion is probably one of the most noticeable and prominent groups of related words. It contains items such as *hastened*, *rushed*, *darted*, *fled*, *escape*, *approach*, *approached*, *arrival*, *followed*, *advanced*, *entered*, *wandered*, or *steps*. This prevalent interest in motion, as indicated above, can be linked to the theme of flight and pursuit that is one of the core topics in Gothic fiction: a character is pursued by some evil being, be it a supernatural being or a human villain. In many cases the victim is a damsel in distress who is beset by a dark villain, like Isabella in *The Castle of Otranto* or Emily in *The Mysteries of Udolpho*. Often heroes and heroines of Gothic fiction are also involved in a pattern of advance and retreat, like Doctor Frankenstein in Shelley's eponymous novel. These patterns are reflected in the prominence of keywords associated with 'motion'.

Another conspicuous and well represented theme in the Gothic keywords is perception. There is a large number of items that can be linked to perception in general, such as *marked* or *perceived*, and more specifically to visual or auditory perception with words like *watched*, *glance*, *concealed*, *appeared*, *beheld*, *visions*, *regarded*, *listened*, *voice*, *sighed*, *silence*, or *sound(s)*. These items show the preoccupation of Gothic fiction with the characters' notoriously unreliable perception and their constant, anxious interpretation of noises and visual input, which is a typical device for creating suspense and an uncanny atmosphere for the reader in Gothic fiction.

Of course words describing feelings, and especially extreme psychological states, also feature prominently in the list of keywords. There is a large range of emotions represented, with items such as *surprise*, *anxious*, *anxiously*, *terror*, *horror*, *impatience*, *dreaded*, *bewildered*, *terrified*, *agitation*, *rage*, *gloom*, or *alarmed* all ranking high in the list of keywords. This is hardly surprising given that various critics see the depiction of extreme emotional states as one of the key characteristics of Gothic fiction (Clery, 2004: 13, Howells, 1995: 27). The Gothic thrives on theatrically stylised displays of emotion and dwells on characters' psychological torments at great length, making a preoccupation with extreme psychological states one of the most distinctive features of Gothic writing.

Other notable groups include 'parts of the body' with items such as *bosom*, *brow*, *eyes(s)*, *arms*, *lips*, and *ear* or 'housing/buildings and furniture' with, for example, *apartment*, *dwelling*, *walls*, *roof*, *couch* and of course words that describe the actants/agents/characters in the narrative like, for instance, *victim*, *attendant(s)*, *domestics*, or *stranger*. All these groups were formed on the basis of the first 150 keyword items. In addition, there are some minor groups already discernible – such as items related to 'stopping' – that could be extended, for instance, by searching further down the list of keywords, but for the purpose of establishing the usefulness of the results delivered by the Gothic group and C18P as a reference corpus this more elaborate process was eschewed, but it will of course be part of my PhD project.

Overall then it seems that despite its unconventional structure, C18P delivers meaningful results when analysing literary genres. While no in-depth analyses of the keywords have been undertaken yet, the preliminary results of the pilot

study indicate that C18P can add interesting findings to extend and elaborate on existing insights from qualitative analyses of literary genres such as Gothic fiction.

6 Conclusion

To conclude, C18P is a corpus of literary texts from the period between 1700 and 1830, an era when the modern novel came into being. C18P tries to cover as much of the range of genres and authors that made up the landscape of British prose fiction in the eighteenth century as possible. It roughly mirrors the rise in the number of texts produced over the course of the three subperiods it is divided into and it also takes into account the increase in the number of texts by female writers. Thus C18P tries to reflect the production of prose fiction of the era as best as possible with the texts currently available in the public domain.

C18P is primarily designed to be used for genre analyses, selecting one genre to be examined and using the rest of corpus as reference corpus. However, since the texts are not only labelled for literary genres but also for form, it should also be suitable for analysing literary forms, such as for instance the epistolary novel or short fiction, although it must be granted that the subdividsion into literary forms is not very fine-grained and might have to be improved by researchers with an interest in questions of form and structure. Apart from literary genre and form, differences in the styles of female versus male writing, or changes between the three subperiods of the corpus are further potential fields of inquiry.

One of the drawbacks of the corpus is that some genre categories, such as the adventure novel, the oriental tale, or the historical novel, are still underrepresented. This is largely due to the limited availability of texts in the public domain and a lack of genre labelling in those texts that are accessible, which makes finding potential texts to include a difficult and laborious process. However, since the corpus will be available under a Creative Commons license it can be supplemented and changed according to individual researchers' research focus and their access to texts (not all of which might be publicly accessible). As already mentioned, the corpus has a fairly flexible structure and thus extending it should not offset the balance of texts as long as the overall sample size does not exceed roughly 20 to 30 texts per genre, which is currently the size of the largest genre subcategories.

Overall, C18P tries to fill a gap in the landscape of specialised corpora, where there are thus far comparatively few examples of corpora composed solely of literary texts and even fewer corpora in which literary texts are labelled for genre other than in the broadest terms. It is a fairly large corpus of around 10 million words that should be of use to any researcher interested in the style and linguistic properties of eighteenth century prose fiction, as well as the development of the novel as a literary kind.

Mag. Iris Gemeinböck
PhD candidate
University of Vienna
a0202543@unet.univie.ac.at
https://github.com/antiquary

References

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

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

'contain, v'. (September 2015) *OED Online*. Oxford University Press. [Online] Available from `http://www.oed.com/view/Entry/40041` [Accessed: 9th November 2015].

Clery, E. J. (2004) *Women's Gothic: From Clara Reeve to Mary Shelley*. 2nd Ed. Hordon: Northcote House.

Davies, Mark. (2004) *BYU-BNC*. (Based on the British National Corpus from Oxford University Press). [Online] Available from `http://corpus.byu.edu/bnc/` [Accessed: 30th June 2015.]

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.

De Smet, Hendrik, Hans-Jürgen Diller, and Jukka Tyrkkö. (2013) 'The Corpus of Late Modern English Texts, version 3.0'. Corpus Page. [Online] Available from `https://perswww.kuleuven.be/~u0044428/` [Accessed: 29th January 2015].

Dillon, George L. (2007) 'The genres speak: Using large corpora to profile generic registers'. *Journal of Literary Semantics* 36: 159-187.

Eighteenth Century Collections Online – ECCO-TCP. [Online] Available from `http://quod.lib.umich.edu/e/ecco/` [Accessed: 29th January 2015.]

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

Fishelov, David. (1993) *Metaphors of Genre: The Role of Analogies in Genre Theory*. University Park: Pennsylvania State UP.

Gerbig, Andrea. (2008) 'Travelogues in time and space: A diachronic and intercultural genre study', in Andrea Gerbig and Oliver Mason (eds) *Language, People, Numbers: Corpus Linguistics and Society*, pp. 157-175. Amsterdam: Rodopi.

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

Howells, Coral Ann. (1995) *Love, Mystery and Misery: Feeling in Gothic Fiction*. London: Athlone Press.

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. [Online] Available from `http://llt.msu.edu/vol5num3/lee/` [Accessed: 23rd March 2015]

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.

Project Gutenberg. [Online] Available from `https://www.gutenberg.org/` [Accessed: 29th January 2015]

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

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. [Online] Available from `https://ota.ox.ac.uk/` [Accessed: 4th May 2015]

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