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.

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 ambigu-
ty in Conrad’s *Heart of Darkness* (Mahlberg, 2007; Culpeper, 2009; Stubbs, 2005).

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.

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.