

#CORPUS

Concerning the term #CORPUS, the Oxford English Dictionary already sketches a conceptual extension towards analysis, a means for comparison. The well-established humanistic meaning of a “body or complete collection of writings or the like” or “the whole body of literature on any subject” dating back to the 18th c, was extended in mid-20th c by the understanding that it might represent a “material upon which a linguistic analysis is based” and by the emergence of corpus linguistics as a “branch of linguistics concerned with analysis of corpora as a means of studying language”. *

From this, it is only a small step to the application of digital approaches to fields such as digital linguistics or digital literary studies although the challenge of defining a perfect or even but suitable #CORPUS remains. The projects of mixed methods take it for granted and discuss the concept of corpus in two further directions. On the one hand, they extend it unselfconsciously from text and language towards music, visual media and even material objects. On the other hand, the terminus #CORPUS is at the centre of diverse methodical considerations. Some projects focus on source material from humanities itself and its usability for digital analysis (Handwriting; ArchiMediaL; Rhythmicalizer; Digital Plato), some on the intrinsic characteristics of a specific corpus (QuaDrama) or the specific transformation from analogous to digital corpus with the resulting promises of knowledge (ReadingAtScale).

Most projects address challenges at the intersection between humanities and computer science as a crucial phenomenon: “Material in corpora is often manually or automatically enriched (...) for conducting large-scale experiments and following specific research questions” (QuaDrama), needs translation into machine-readable formats (BachBeatles) or additional means of sourcing to create an adequate data base (ArchiMediaL; see #HETEROGENEITY)

* “Corpus, n.”. in: *Oxford English Dictionary (OED)*, first published 1893; most recently modified version published online March 2022 with draft additions from 1997 and 2013, <https://www.oed.com/> [accessed: 20.05.2022].

Title: Doing DH in a mixed-methods scenario. Experiences in the QuaDramaA Project

Team: Quantitative Drama Analytics (Nils Reiter, Marcus Willand, Benjamin Kraut-ter, Janis Pagel)

Corpus: German Drama (1730–1930)

Field of Study: Literary Studies

Institution: University of Stuttgart

Methods: Annotation, machine learning, quantitative analysis

Tools: CorefAnnotator, DramaAnalysis, DramaNLP, RStudio

Technology: Apache UIMA, NLP, XML