Historical Models and Serial Sources

Michael Piotrowski – Université de Lausanne

Serial sources such as records, registers, and inventories are the “classic” sources for quantitative history. Unstructured, narrative texts such as newspaper articles or reports were out of reach for historical analyses, both for practical reasons—availability, time needed for manual processing—and for methodological reasons: manual coding of texts is notoriously difficult and hampered by low inter-coder reliability. The recent availability of large amounts of digitized sources allows for the application of natural language processing, which has the potential to overcome these problems. However, the automatic evaluation of large amounts of texts—and historical texts in particular—for historical research also brings new challenges. First of all, it requires a source criticism that goes beyond the individual source and also considers the corpus as a whole. It is a well-known problem in corpus linguistics to determine the “balancedness” of a corpus, but when analyzing the content of texts rather than “just” the language, determining the “meaningfulness” of a corpus is even more important. Second, automatic analyses require operationalizable descriptions of the information you are looking for. Third, automatically produced results require interpretation, in particular, when—as in history—the ultimate research question is qualitative, not quantitative. This, finally, poses the question, whether the insights gained could inform formal, i.e., machine-processable, models, which could serve as foundation and stepping stones for further research.