Using matrix algebra to analyze a massively parallel Bible corpus

Thomas Mayer & Michael Cysouw

The aim of this talk is basically twofold. First, we present the current status of our massively parallel Bible corpus and the current efforts to improve its coverage and meta information. The basic idea of the corpus is to provide a large-scale multilingual corpus that can be easily accessed by researchers. This also includes a web-interface with search functions for non-quantitative analyses. We also discuss the problem of copyright issues and how to allow for linguistically meaningful studies on the basis of the texts without violating copyright restrictions. One of the solutions to this problem leads to the second part of the talk in which the parallel structure of the texts is transformed into large (but sparse) matrices.

Second, we discuss the usefulness of matrix representations when dealing with parallel texts. There are three main reasons why matrix algebra can be helpful in analyzing parallel texts. (i) Many calculations that are necessary for comparing parallel texts can be performed much faster with matrix manipulations; (ii) matrices give a concise representation of the data types, which makes it easier to talk about different types and facilitates storing them in a pipeline of computational methods; (iii) methods from matrix algebra will hint at decompositions or calculations that are useful for the analysis of parallel texts. We illustrate this with a few examples from our own research.