Reg. No: 2016/23/B/HS2/01323; Principal Investigator: prof. dr hab. Adam Tomasz Pawłowski

Methods and tools of corpus linguistics in the research of bibliography of Polish publications from the period 1997-2017.

Histories of states, societies and nations are not written exclusively by professional historians. Similarly, the knowledge about the surrounding world can be found not only in scientific monographs. Valuable knowledge can be also generated from text sources prepared for practical purposes: an efficient use of libraries is a purpose of this kind. This is precisely why we made a bibliography the object of investigation in our proposed project. We do not mean partial bibliographies (even if they encompass thousands or dozens of thousands of records). We will work on the records representing the totality of books published in Poland during the period (1997–2017). Data sets of this size can be certainly regarded as particularly rich knowledge resources. They allow discovering long cultural trends (for instance a curve representing frequency of terms *terrorism* or *computer* in titles of books published during a sufficiently long period of time; one can also check, if titles become shorter in the world running faster and faster). One can discover relationships between titles (automatic clustering of titles in respective domains or disciplines). It is also possible to discover in titles language features which in many cases allow to determine gender of the author.

Can a small team of researches cope with such tasks? Processing of more than five hundred thousand bibliographical records (according to tests carried out so far this is the expected size of the database generated from the resources of the Polish National Library) is unfeasible using traditional (manual) methods. However, a combination of expertise in automatic text processing tools, experience in corpus linguistics and information science will let us carry out the project successfully. We intend to apply Natural Language Processing tools an corpus linguistic methods to the analysis of bibliographical “micro texts”. We intend to verify, if titles allow to determine authors sex (gender?), we would like to create a map of toponyms (i.e. geographical proper names), we will cluster titles automatically and evaluate efficiency of classification method. We will also generate histograms of cultural trends in bibliographical data.

What encouraged us to undertake this topic was certainly the lack of such research in Poland. However, great possibilities of digital humanities in the domain of text processing were a also powerful practical argument.