Large-Scale Text Analysis and Methodological Foundations of Computational Stylistics

In social sciences, earth sciences, and particularly in the humanities, we witness nowadays a phenomenon usually referred to as the “digital turn”. This is connected with the advent of computer-based and computer-assisted technologies, the ever-growing computing power, and – more importantly – the exponential increase of the amount of information that has been electronically produced, analyzed, managed, viewed, stored, and made available to the public in the last decades.

A vast majority of the data stored on innumerable servers across the world are obviously text-centric. As the most notable example, the Google Books database now contains some 30 million books, hundreds of hypertext editions gathered from universities, libraries, and private initiatives (their quality varying from excellent to unacceptable), thousands of webpages, millions of blog posts, billions of tweets, and so forth. The Internet itself is still a highly textual medium and a potential multilingual corpus at the same time. The “digital turn” means access to previously unheard-of amounts of data; at the same time, however, this presents non-trivial challenges. To name but a few, these include information search and retrieval, data analysis, classification, plagiarism detection, surveillance, forensic textual studies, and many others. With the ever-increasing corpus of electronic data available, scholars can nowadays engage with texts on an unprecedented scale. Innovative research questions and methodologies emerge; attempts to answer the former and make use of the latter begins to provide new data-driven answers to age-old questions in the humanities. In the Digital Humanities, much attention is currently invested in the development of computational methodologies to assist in this process.

The present project, even if focused solely on computational stylistics of literary texts, provides a significant contribution to this great challenge: one of its main aims is to develop, evaluate and apply an innovative methodology of comparing texts in order to identify hidden patterns and similarities invisible to the naked eye. This will allow us to automatically distinguish texts written by men and women, to find, say, first-person narration novels (or any other genre) out of a mass of textual data, to retrieve chronological trends among texts in a corpus, and so forth. Simultaneously, we want to make our method resistant to large amounts of noise – a crucial yet underestimated factor when one deals with “web-scraped” textual data. The technique will be thoroughly tested for scalability, that is its applicability to corpora of any size (it is assumed to process thousands of full-sized novels at once). The methodology developed should be suitable for different languages, and easily applicable into cross-language studies; we are going to study different aspects of style in Polish, English, Latin or Greek literatures. Last but not least, we want to propose a novel way of visualizing relations between dozens of texts at once (which is already a non-trivial task for a few hundred novels).

The project is divided in five discrete yet closely connected subprojects. Subproject A is planned to explore theoretical problems of stylometric tests, including but not limited to reliability issues and visualization techniques; subproject B is aimed at testing the established methodology on large corpora of literary texts in different languages, genres, themes and authorial genders in the original and in translation; subproject C, conducted by a PhD student, will examine grammatical features as indicators of stylistic variation; subproject D, planned as another PhD thesis, will discuss style differentiation in one particular literary tradition, say, Latin (depending on the profile of the PhD student); last but not least, subproject E will explore sequential measures and their applicability to stylometry.