Stylistic Differentiation between Literary Texts: Linguistic Perspective

One of the reasons of the increasing popularity Digital Humanities is its intrinsic interdisciplinary nature. A good example here it stylometry, or applying computational techniques in order to address questions of stylistic relations between (literary) texts. A successful application of stylometric methodology is conditioned by (1) research questions introduced by literary studies, (2) statistical inference involving formal language of mathematics, as well as (3) theoretical framework heavily inspired by contemporary linguistics. The present paper will explore the linguistic perspective.

In classical approaches to stylometry – which includes various analyses varying from solving authorship attribution cases, to large-scale investigations of literary corpora – frequencies of the most frequent words (MFWs) are claimed to outperform other types of style-markers. Importantly, the "words" are defined here as space-delimited strings of letters, while "most frequent" mean mostly if not exclusively function words, such as particles, articles, conjunctions or prepositions.

What it basically means, is that a stylometric test – be it authorship attribution or a distant-reading analysis of literature using quantitative methods – can be applied to any web-scraped plain text file with a high probability of achieving acceptable results. However, such an approach is inevitably difficult to digest for a linguist, who might ask: why word forms, and not other units of language? Why not measure base forms (i.e. lemmas), or morphemes? And above all: why would we restrict ourselves to lexis, while neglecting syntax?

On theoretical grounds, grammar should always constrain the authorial freedom of choice to a significantly greater degree than it constrains the (usually very individual) lexical repertoire. If an author wishes to describe a given entity with an adjective, there exist numerous words to choose from: e.g. the entity's size may be big, large, great, considerable etc. However, if we take into account grammatical categories, the entity will inevitably be represented by a sequence [Adjective] + [Noun]. Moreover, these limitations are much more rigid on the syntactic level than on the lexical level.

The presentation will explore different realms of linguistic features in order to test to which extent they exhibit authorial idiosyncrasies. Disappointing as it might seem, the results clearly suggest that most of the authorial individual voice can be traced using very shallow textual features, such as the aforementioned word forms or – even worse – mere sequences of letters.