Data-driven text analysis for digital humanities: some thoughts on how and why

Andrew Salway
Language and Language Technology Group, Uni Research, Bergen, Norway

Over the last two decades significant advances in computational power and concomitant techniques for data analysis and visualization have led to a shift in research practice across the natural sciences, with the establishment of data-driven research paradigms. Now the humanities are witnessing a nascent data-driven turn of their own, stimulated by rapidly growing digital archives of cultural resources which include large volumes of text. This paper will reflect on how and why data-driven text analysis methods could and should be used as part of digital humanities research methods, by making particular reference to literary studies and by considering the relevant state-of-the-art technology for natural language processing – including our own work on new technologies.

With the availability of text corpora comprising many millions or billions of words, and the computational power to run inductive algorithms over them, data-driven methods become a feasible and necessary companion to hypothesis-driven approaches. In broad terms, inductive text analysis techniques highlight interesting linguistic patterning in a corpus – for a researcher to then interpret – thus alleviating the problem of “seek and ye shall find” by minimising prior assumptions about the content of the material. Inductive techniques are suited to the exploratory phases of research during which they can provide researchers with a manageable overview of the content of a corpus in order to stimulate new hypotheses. They may also guide the development of coding schemes, i.e. systematic mappings between countable linguistic forms (e.g. words, phrases, grammatical structures) and concepts related to the phenomena of interest to researchers (e.g. topic, style, framing) which are essential if further automated techniques are to be used to test hypotheses on a large scale.

Approaches to literary and narrative inquiry have traditionally been grounded in the phenomenology of reading (Herman 2005: 125-6) whereby analysts consult their readerly intuitions about a text. In turn, these intuitions find expression in terms and categories deriving from a history of similar engagements by a larger community of expert readers (cf. Toolan 2001: 22). Drawing on these resources, scholars classify literary devices, analyse how they are used in different texts and theorise about how the devices enable readers to understand and engage with narratives. By contrast, the notion of ‘distant reading’ suggests that keeping a distance from the intricacies of individual works provides new insights by allowing interconnections to become apparent through the analysis of larger collections of works (Moretti 2013).

An interesting approach that blended distant and close reading was taken by Mahlberg in her corpus stylistics work on a Dickens corpus of 4.5 million words (Mahlberg 2007). A computer system identified unusually frequent word clusters compared with a reference corpus of other 19th-century novels. The clusters were assumed to reflect local textual functions and they were manually filtered, organised and interpreted as such, and then used as the basis of stylistic analysis; a similar approach was presented by Herman and Salway (2007), and Salway and Herman (2011). A key point about data-driven approaches is that once an interesting phenomenon has been identified (e.g. a local textual function), its linguistic realisation is already known (e.g. a word cluster that was induced from the text). This means that further data about the distribution of the phenomenon in the corpus, i.e. for testing hypotheses, can be generated automatically without the need for prohibitively expensive manual coding.

The potential utility of large text corpora depends upon the extent to which interesting phenomena manifest in linguistic patterning that is sufficiently regular and frequent that it can
be automatically induced. In part there is a fundamental limit since context and sometimes co-text are not accessible to inductive text analysis techniques. There is also a limit due to the state-of-the-art in natural language processing. Whilst insights can be gained using relatively simple techniques (e.g. frequency and keyword lists, n-grams, concordances and collocation data), there seems to be a need for techniques that can induce richer linguistic patterning from corpora. To that end we have adapted a grammar induction algorithm (Solan et al. 2005) and demonstrated how it can be used to induce salient information structures from a variety of corpora (Salway and Touileb 2014; Salway, Touileb and Tønnereim 2014; Touileb and Salway 2014). The induced structures extend and enrich the overview of corpus content provided by extant techniques. In the current paper we will show and discuss preliminary results from applying our method to a literary corpus, and compare results with Mahlberg (2007).

References