Newton open-sourced: Performing cross-corporal Latent Semantic Analysis on EEBO-TCP and the Corpus Newtonicum

Cornelis J. Schilt
Newton Project, University of Sussex, United Kingdom; c.schilt@sussex.ac.uk

When Isaac Newton died in 1727 he left behind a huge amount of draft manuscripts, many of them related to the posthumously published Chronology of Ancient Kingdoms Amended. Throughout these drafts we witness a composer at work, skilfully mastering a huge number of source texts by ancient historians and poets, many of them in Greek or Latin originals, and others in their Latin or English translations. Homer, Ovid, Flavius Josephus, Herodotus, Eusebius, they are all there, both with literal quotations and in paraphrase. Newton was very aware of variations in original Greek and Latin translations and deliberately selected particular versions to quote from. It is not always clear however which sources he uses, or whether a particular line of text is his, a paraphrase or even a direct quote. Also, it is remarkable that on a number of occasions Newton apparently misquotes particular sources; that is, if we compare his citations to well-known contemporary versions, we witness discrepancies. There are a number of possible answers to this. It might be that Newton didn't always quote directly from a particular source that was readily available to him, but was sometimes remembering from mind, which would inevitable lead to less accuracy. Alternatively, he might have used other sources than the obvious ones, including maybe even his own translations. And then there's the third option: Newton is deliberately adjusting and modelling quotes to fit a particular argument. We know that Newton sometimes adapted a similar strategy in his scientific data analysis, so it would be particularly interesting to see whether and how this returns in his scholarly writings.

Regarding his sources we do not know exactly which versions Newton had at his disposal, since many of the works he refers to were not in his library. Similarly, we do not exactly know which contemporary historians and writers influenced Newton in writing his chronology. For example, he regularly refers to John Marsham's works for supportive evidence of his claims, but mentions Joseph Scaliger only once. According to researchers of Newton's library his copy of Scaliger's Opus Novum shows annotations and dog-earing. Are there any hidden Scaliger references in his texts? And what about other historians and chronologers of his days: in what way is Newton indebted to them? What can engagement with intelligently marked up texts tell us about the working practices of one early modern scholar? What can this tell us about Newton’s memory and creativity, and more generally, his attitude to textual sources? More reflexively, how might we in the future use TCP-IP texts to their full extent?

Many of the works that Newton might have read and/or directly used are part of the Early English Book Online Text Creation Partnership. The public domain release of the EEBO-TCP texts enables us for the first time to do a large-scale comparative analysis involving both Newton's chronological draft texts and any works that he might have used in composing these. Most of the Corpus Newtonicum has been transcribed using TEI-P5 XML and is available for research. With the availability of the huge corpus of EEBO-TCP texts, we will be able to gain a better understanding of Newton's use of sources and quotations by performing cross-corporal Latent Semantic Analysis and topic modelling. By using the rich Newton materials as an example, I will demonstrate how sophisticated encoded texts enable us to gain a much deeper understanding of the connections between historical texts and the particular idiosyncrasies of their authors. More specifically, I will show how the public availability of the EEBO-TCP texts challenges us to ask questions that were previously unthinkable, and to answer questions that were previously unanswerable due to matters of sheer scale and complexity.

On the other hand, I will also address the technical and editorial implications that Latent Semantic Analysis and other forms of comparative analysis have for the EEBO-TCP corpus. For instance, it is worrying that a large part of the corpus is still in TEI-P3, with the Oxford Text Archive in P5. It is for all intents and purposes pivotal that all texts comply to the same standard at all times. There will be always be time windows when conversions are taking place, but these should be done simultaneously to guarantee comparative data integrity. Another key issue to do with comparative analysis has to do with data quality and a rigorous implication of standards. A total disambiguity in the coding of ligatures and abbreviations is an absolute necessity when comparing large amounts of texts; but also the ability to choose – as an editor or user – a particular format or option in textual representation. It is essential that texts are faithfully transcribed and not narrowed down by their editors. For instance, at the moment handwritten additions to printed texts are not transcribed " due to the numerous complications and uncertainties associated with handwriting in texts", notable illegibility, sometimes due to the quality of the images used by the TCP editors, or because of faded ink. Other issues involve "the lack of standardization in handwriting and the various shorthands that existed in the 16th century" and difficulties in determining the relationship of the handwritten addition to the text. However, as the same post from which these quotes are taken painfully shows, much might be lost by this systematically discarding of handwritten additions: "the discovery of a potentially unique version of a poem, found in one of our texts." This is of course a rather special find, but it illustrates that it should not be for transcribers to discard potential useful information that might appear unrelated. Especially marginal annotations can provide us with a rich source of information about how a particular text is used.

The above are just some examples involving data quality issues, of which there are many more. The public availability of the EEBO-TCP texts is a major development in digital history and a potential source of new key insights in history in general. With the right quality management we will be able to address a whole range of new and exciting questions. Newton's draft papers, the Corpus Newtonicum, provide such a challenging set of questions, for which comparative joint analysis with the EEBO-TCP corpus will undoubtedly provide us a unique set of answers.

Quotations above taken from (http://www.textcreationpartnership.org/2013/07/05/an-unexpected-discovery/#more-1820; consulted 23-05-2014)