Wrangling large-scale data for specialised corpora

Andrew Hardie
Lancaster University
a.hardie@lancaster.ac.uk

With vast amounts of text now readily accessible via the web, a “specialised corpus” need not be a “small corpus”. However, the immensity of web resources presents challenges. Automatically-spidered data comes with none of the structure that characterises carefully-constructed corpora; when the research goal is to approach language of a very specific type, a “flat” corpus of this kind will typically not be satisfactory.

The ESRC-funded project “Metaphor in End-of-Life Care” (MELC) aims to examine metaphoricity in language associated with terminal illness – not only of patients but also of carers and medics. We mass-downloaded message-boards amounting to XXX words; but we then faced a number of problems: (1) structuring the data in a way that reflects the conceptual divisions of the original message-board; (2) allowing analysts routes of access into this dataset; (3) labelling the different classes of participant.

By using a bespoke spidering program, rather than an off-the-shelf mass-download tool, we made a single message-board thread correspond to a single corpus text. Within each thread/text, the mark-up identifies different posts, as well as the user responsible for each. A relational database was created in which all threads, users and posts are represented and cross-linked. A web-interface to this database allowed us to annotate user types – identifying users as patients, carers etc. by examining how they identify themselves in their first posts. We can then extract specified “slices” of the corpus for detailed analysis.

These techniques illustrate how a very large web-derived corpus can be made tractable as a resource for detailed analysis such as the investigation of metaphor, whilst respecting the conceptual structure of the original online resource.