A gentle introduction to textometry

Àâòîð Serge Heiden
04.10.2009 ã.
Ïîñëåäíåå îáíîâëåíèå 28.10.2009 ã.

Ìàòåðèàëû ê ëåêöèè (ïðåçåíòàöèÿ)

Textometry is a methodology designed for humanities researchers to work on digitized texts corpora with computers and statistics. After having digitized and encoded the texts in the computer and organized them in a coherent set called ‘corpus’, textometrical tools help to analyse the corpus with search engines and frequency based statistical tools.

Search engines look in the texts for qualified elements like lexical items (words/compound words) or structural elements (chapter/sentence…) and can be tailored to catch variations through pattern matching. For example, one can search for ‘a word beginning with “anti-“ some words before another one at the end of a sentence’. If the elements have been linguistically annotated (for example with lemma or part of speech), the search engine can also use that information to express more constraints on the pattern to look for.

Statistical tools can work on the number of occurrences of all or specific lexical elements in a particular structural element with respect to other structural elements or the whole corpus. For example, if the texts of the corpus have an ‘author’ property specified, one can extract the most specific words used by a given author with respect to the others, or if the texts have a ‘date’ property specified, one can extract the most specific words of a given period of time. Specificity is measured by a statistical score, with a definite statistical meaning. Statistical tools can also work on the linear sequence of words in texts. For example, one can analyse how often pair of words occur together within sentences or paragraphs, compute the specificity score of those encounters and build the network of all the specific pairs of words in the vocabulary of a text.The Textometry research project (http://textometrie.ens-lsh.fr) is developping a new software platform which implements that methodology and which will be demonstrated.