METHOD FOR DETERMINING THE REPRESENTATIVITY OF A CORPUS

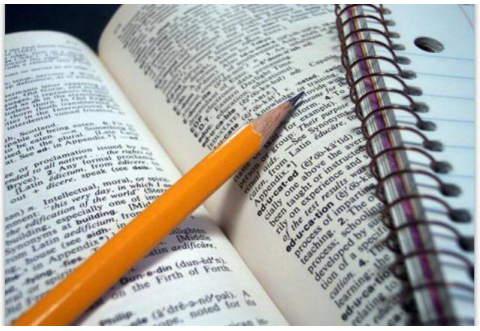
Description:

The question of representativeness remains today one of the most controversial aspects of corpus linguistics. In the case of specialized corpus, which tend to have a much smaller size than the so-called “general” or “reference” corpus, the question of representativeness is really key, indeed, it is one of its defining characteristics. In practice, the quantification of the minimum size that a specialized corpus must have has not yet been objectively determined. And it is that there is no consensus on what is the minimum number of documents or words that a certain corpus must have in order for it to be considered valid and representative of the population to be represented. Thus, the present invention is an efficient solution to determine a posteriori the minimum size of a corpus or textual collection, regardless of the language or textual type of said collection, establishing, therefore, the minimum representativeness threshold through an algorithm (N-Cor) analysis of the lexical density as a function of the incremental increase of the corpus. Starting from this premise, a proposal for a computer implementation has been arrived at, which has resulted in an application developed in Java, and which we have called ReCor. Said system has the following main classes: a) Words (algorithm of computation, reading and writing to file); b) Gui (user interface); and c) Graphic Window (adapter for graphic representation).

Keywords:
Corpus, Linguistics, Language Processing, Software

Sectors:
ICT, Others

Areas:
Software / Procedures, Education



Advantages:

Among the advantages of the present invention are: • It is independent of the language or textual type of the collection of documents analyzed. • Establishes the minimum threshold of representativeness of a corpus. • Includes input data, output data, graphical representation and output files. • It is easy to use, with a friendly interface.

Uses and Applications:

The present invention is used as a computer-implemented data processing method, particularly linguistic data and information, to determine the representativeness of a corpus.

Patent Number: ES2320511B1
Applicants: Universidad De Málaga
Inventors: Gloria Corpas Pastor, Miriam Seghiri Dominguez, Romano Maggi
Filing Date: 05/12/2006
Protection Level: National (Spain)
Processing Status: Spanish patent