'Constructing comparable and parallel corpora for terminology extraction – work in progress'

Belinda Maia

Faculdade de Letras da Universidade do Porto
Via Panorâmica s/n,
4150-564 Porto,
Portugal.
bmaia@mail.telepac.pt

This paper will focus on the theory and practice of creating bi- or multilingual parallel and comparable corpora for research into specialized language. Such corpora pose specific problems because – apart from the usual problems of copyright and accessibility of texts - the texts need to be evaluated on the basis of domain, register and style. One also needs to consider the relative advantages and disadvantages of parallel and comparable corpora and the type of research for which they are required.

Parallel corpora are understood to be originals texts aligned with their translations. Aligned texts can be made into the translation memories that assist translators with their work, and they can be used by researchers to analyze a wide variety of linguistic phenomena related to translation, particularly at the lexical, syntactic and semantic levels. Comparable corpora are original texts in two or more language that are similar in subject matter, or domain, register and style. The theoretical side of compiling such corpora is complex, because of the differing opinions on the comparable aspects of texts in different languages.

There is considerable interest in corpora of this kind for areas of research that range from terminology management, through information retrieval to knowledge engineering. We shall begin by suggesting ways in which texts can be selected, and how the cooperation of domain specialists can be maximized. We shall analyze the possibilities and problems of using such corpora to extract terminology and examine the methodology being developed from this for information retrieval. We shall also look at other ways in which such corpora can be used for research, and special attention will be paid to the distinction between general and specific language. The practical aspects of such corpora building will be discussed, with special reference to our experience of working with corpora and terminology with projects and dissertations at postgraduate level.

The work undertaken requires cooperation between the linguist and experts in information technology and the specialized domains, and we shall explain why such interdisciplinary work is essential to progress in a field which at present suffers from fragmentation and lack of communication. The paper will describe the efforts of the Porto branch of the Linguateca (see: http://www.linguateca.pt) to develop a faculty site for corpora as well as bi-lingual specialized corpora of comparable and parallel texts in Portuguese and English.