XVIth European Symposium on Language for Special Purposes
27-31 August 2007
University of Hamburg

Bilingual English-Chinese Parallel Texts and Homothematic Corpora

Benjamin K. Tsou
Language Information Sciences Research Centre
City University of Hong Kong

Abstract

The successful introduction of balanced corpora in the last century, such as the now popular COBUILD, BNC, for example, has heralded an important milestone in computational linguistics and NLP, including developments in the domain of LSP. An important guiding principle underlying balanced corpora is the purposeful avoidance of duplication of content.

The subsequent appearance of parallel corpora has notionally led to an opposite guiding principle of HOMOTHEMATIC corpus construction, where EXACT content duplication is now anticipated in at least the two languages involved. This has led to important findings and fruitful applications, relevant to lexicological and lexicographical work, terminology bank building and bilingual dictionary production as well as to Information Retrieval, and to Example Based Machine Translation, and Translation Memory.

On the basis of a parallel Chinese-English legal corpus at the City University of Hong Kong, I shall discuss some issues relating to textual equivalence and content duplication, and their contributions to cross-lingual comparisons. I shall also review the use of Latin terms found in the English texts and their renditions and impact with reference to Chinese.

Comparisons will be made with another homothematic corpus, LIVAC, a synchronous corpus of Chinese based on the cultivation and dynamic maintenance of media material across 6 major Chinese speech communities.

/1