Daniel Gallego Hernández (Universidad de Alicante) and Ramesh Krishnamurthy (Aston University)

COMENEGO (Corpus Multilingüe de Economía y Negocios): design, creation and applications

This paper will describe the initial stages of the COMENEGO project, which is initially creating comparable corpora of Business texts in Spanish and French. The language of business remains a vital field in global activities, and the globalised market requires frequent cross-border, cross-linguistic, and cross-cultural interaction. The need for rapid and accurate translations places increasing demands on the business community, on translation practitioners, and those who train the translators. High-frequency activities also tend to react, innovate, and adapt more quickly to changes in their environment and practices, and therefore there is a constant need to renew outdated resources. However, this pressure often leads to ad hoc collections. The highly competitive nature of commerce also poses problems of data availability, accessibility, and cost. Translated texts may contain traces of source language interference and other non-native-speaker features, hence especially for advanced and specialised translation purposes, comparable corpora may be more suitable. For Spanish and French, these factors and features are evident in previous corpora. For example, MLCC(*1) is reasonable in size (10m words for each language) but is now outdated (early 1990s), and lacks textual variety (single sources – Expansion for Spanish and Le Monde for French). IULA(*2) has variety, but is rather small (1m words of Spanish), and contains many translated texts. CLUVI(*3) is also smaller (0.4m and 1.8m Spanish), and contains translated texts, single-source subcorpora (Consumer Eroski) and some outdated texts (1998 onwards). Vicente(*4) is diachronic and contrastive (1995 and 2006), contains only press articles (French: Le Monde, Les Echos; Spanish: El Pais, Expansion), and is not publicly available.

COMENEGO focuses on up-to-date texts, textual variety and balance, and the necessary compromise between more idealised corpus design and practical factors, and will constitute a valuable resource for researchers, translators, and translator trainers and trainees. The first cycle of data collection has acquired 7m words of Spanish and 9m words of French. Copyright permissions are being obtained, and pilot analyses are revealing imbalances and deficiencies which are being addressed. This paper will discuss not only the process of corpus design and creation, but also the applications of comparable corpora in translation pedagogy.

References

MLCC (*1) - MLCC Multilingual and Parallel Corpora
http://www.elda.org/catalogue/en/text/W0023.html

IULA (*2) - Corpus técnico del IULA
http://bwananet.iula.upf.edu/indexes.htm

CLUVI (*3) - Corpus Lingüístico de la Universidad de Vigo
http://sli.uvigo.es/CLUVI/index.html

Vicente (*4) - VICENTE, Christian (en prensa): «Lingüística de corpus y traducción especializada: aplicaciones a la traducción francés-español de la economía», XXV CILPR: Congrès international de linguistique et de philologie romanes, Innsbruck, 2007.