Translation-oriented annotation of a multimedia parallel corpus of subtitles

Patricia Sotelo-Dios (University of Vigo, Spain)

This poster presents an ongoing research project that involves annotating a multimedia parallel corpus of subtitles with translation-related information. The Veiga corpus of English-Galician film subtitling currently covers 35 audiovisual products (ca. 450,000 tokens) and can be accessed freely at http://sli.uvigo.gal/CLUVI/vmm_en.html. The corpus search application allows for complex as well as parallel searches. It shows the bilingual equivalences of the searched terms in context and enables users to stream the video clips where the bilingual pair appears, thus giving them access to the (co-)text in its original, multi-semiotic form. The Veiga already contains two levels of annotation: on the one hand, the omission, addition and reordering of translation units; and, on the other hand, the in-cue and out-cue times and line breaks in the subtitles. All of the above mentioned aspects are tagged according to the XML CLUVI specification for parallel corpora (Guinovart & Sacau, 2004), and a new set of tags is now being defined to annotate certain elements that are particularly relevant to the practice of subtitling. The aim of the project is to enrich the corpus with translation-related data so that users can easily search for potentially problematic issues and observe the specific techniques used by the translator to render them in the subtitles.

A first pilot experiment is now being carried out with one of the films. In particular, the items currently being considered for annotation are the following: linguistic variation (dialect, slang, taboo language), idioms, culture-specific references, named entities, and orality features such as paralinguistic elements, discourse markers, interjections, vocatives, false starts or repetitions. Although a number of studies have already empirically investigated the translation of some of these issues in a variety of parallel subtitles corpora, with a special focus on the transfer of humour, cultural references, taboo language and linguistic variation, none of these corpora have yet been tagged with this type of information and made available to the public. In this respect, the Veiga corpus, with this additional layer of metadata, intends to make a small contribution to the annotation of parallel corpora for descriptive translation studies by providing a freely searchable database that could be used for translation teaching and research.

References

Gómez Guinovart, X., & Sacau Fontenla, E. (2004). Parallel corpora for the Galician language: building and processing of the CLUVI (Linguistic Corpus of the University of Vigo). In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, R. Silva, F. Carvalho, ... S. Barros (Eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)* (pp. 1179-1182). Paris, France: European Language Resources Association.