El corpus paral·lel del Diari Oficial de la Generalitat de Catalunya: compilació, anàlisi i exemples d'ús

Antoni Oliver (Barcelona)

Summary: In this paper the process of compilation of the parallel corpus from the Official Diary of the Catalan Government (DOGC) is presented. It describes the downloading process, the tools and processes for the treatment and linguistic analysis. The final result is a big parallel corpus that is freely available in several formats and with several annotation levels. This corpus is a very valuable resource for different applications. As example, three possible fields of application are described: as a translation memory to be used in a Computer-Assisted Translation tool; for terminology extraction and query and for training statistical machine translation systems.
Keywords: Parallel corpus, translation memory, terminology extraction, statistical machine translation, Natural Language Processing

Zeitschrift für Katalanistik 30 (2017), 269–291
ISSN 0932-2221 · eISSN 2199-7276