Irene Renau/Rogelio Nazar

TOWARDS A MULTILINGUAL DICTIONARY OF DISCOURSE MARKERS

Automatic extraction of units from parallel corpus

Abstract This paper presents a multilingual dictionary project of discourse markers. During its first stage, consisting of collecting the list of headwords, we used a parallel corpus to automatically extract units from texts written in Spanish, Catalan, English, French and German. We also applied a method to create a taxonomy structure for automatically organising the markers in clusters. As a result, we obtain an extensive, corpus-driven list of headwords. We present a prototype of the microstructure of the dictionary in the form of a standard XML database and describe the procedure to automatically fill in most of its fields (e. g., the type of DM, the equivalents in other languages, etc.), before human intervention.

Keywords Computational lexicography; corpus-driven lexicography; discourse markers; multilingual lexicography

Contact Information

Irene Renau
Pontificia Universidad Católica de Valparaíso, Chile
irene.renau@gmail.com

Rogelio Nazar
Pontificia Universidad Católica de Valparaíso, Chile
rogelio.nazar@pucv.cl

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress. Mannheim: IDS-Verlag.