Especificar COMUNICACIÓN

Título: Anotación semántica y discursiva de textos bilingües para aplicaciones lingüísticas y computacionales

Ponente: Julia LAVID-LÓPEZ, Juan Rafael ZAMORANO y Lara MORATÓN
Autor o autores: Julia Lavid-López, Juan Rafael Zamorano, Lara Moratón, Marta Carretero, Jorge Arús
Correo de contacto: lavid@filol.ucm.es

Línea temática: Entorno digital e investigación

Grupo de investigación UCM: FUNCAP: Lingüística Funcional (inglés-español) y sus aplicaciones 930175
Director del Grupo UCM: M. Julia Lavid López
Otros grupos de investigación:
Proyecto/s de investigación:

- proyecto MULTINOT, financiado por el Ministerio de Economía y Competitividad
- TEXTLINK: Structuring Discourse in Multilingual Europe, financiado por la Comisión Europea

Resumen: 300 palabras

Las bases de datos textuales o córpora en formato digital y anotadas con información lingüística compleja (e.g. aspectos semánticos, pragmáticos y discursivos) son fundamentales para el entrenamiento y validación de algoritmos en el campo del Procesamiento del Lenguaje Natural (PLN), y son esenciales como 'goldstandard' para comprobar el funcionamiento de las tecnologías lingüísticas. En el ámbito de la Lingüística de Corpus, además, la anotación de textos añade un valor fundamental a los corpora textuales, enriqueciéndolos con información lingüística que puede ser reutilizada para diferentes aplicaciones. Además, la anotación de corpus tiene un gran potencial como método de investigación de fenómenos lingüísticos, ya que permite explorar dichos fenómenos de forma empírica, pudiendo validar hipótesis y teorías que antes no estaban comprobadas (véase Hovy y Lavid 2010, Lavid 2012, Lavid et al. 2016a, 2016b).

En esta presentación se describen los principales avances que se están realizando por varios investigadores del grupo FUNCAP en el campo de la anotación, tanto automática como manual, de textos bilingües (inglés-español) en el marco del proyecto MULTINOT (véase Lavid et al. 2015), así como la reciente colaboración que se ha establecido a nivel internacional en el marco de TEXTLINK, una iniciativa europea para la creación de corpus anotados con mecanismos discursivos (Discourse Relational Devices). Con respecto a MULTINOT, se presenta el corpus bilingüe desarrollado y se perfilan los últimos avances en la anotación de fenómenos como la tematización, los marcadores discursivos y la epistemicidad. Con respecto a TEXTLINK, se perfila la colaboración en el ámbito de la anotación de marcadores discursivos en inglés y español, utilizando los corpus bilingües compilados para el proyecto MULTINOT.

REFERENCIAS

Hovy, E.H. and J.M. Lavid. (2010): Towards a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics.International Journal of Translation Studies. Vol. 22, No. 1, Jan-June 2010.

Lavid, J., Carretero, M. and JR Zamorano (2016a): A linguistically-motivated annotation model of modality in English and Spanish: Insights from MULTINOT. Linguistic Issues in Language Technology volume 14, issue (4) August 2016. Special Issue on Modes of Modality in NLP. CSLI Publications, Stanford University.

Lavid J., Carretero, M. and JR Zamorano (2016b): Contrastive Annotation of Epistemicity in the Multinot Project: Preliminary Steps. In Harry Bunt (ed.). Proceedings of the ISA-12, Twelfth Joint ACL - ISO Workshop on Interoperable Semantic Annotation, held in conjunction with Language Resources and Evaluation Conference 2016. 81-88.

Lavid, J., Arús, J. DeClerck, B and V. Hoste (2015): Creation of a high-quality and register-diversified (English-Spanish) parallel corpus for linguistic and computational investigations. In Pedro A. Fuertes-Olivera et. al. (eds.). Special Issue of Procedia-Social and Behavioral Sciences. Current Work in Corpus Linguistics: Working with Traditionally- conceived Corpora and Beyond. Selected Papers from the 7th International Conference on Corpus Linguistics (CILC2015). Volume 198, pages 1-556 (24 July 2015).