Procesamiento del lenguaje Natural, nº 41 (2008), pp. 311-312 recibido 7-05-2008; aceptado 16-06-2008

CLARIN: Common Language Resources and Technology Infrastructure

Núria Bel, Montserrat Marimon
Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra
Pl. de la Mercè, 10-12
08002-Barcelona
[nuria.bel|montserrat.marimon]@upf.edu

Resumen: Presentamos el proyecto CLARIN, un proyecto cuyo objetivo es potenciar el uso de instrumentos tecnológicos en la investigación en las Humanidades y Ciencias Sociales.
Palabras clave: Ciencias sociales, humanidades, servicios web, tecnología *grid*, tecnologías y recursos lingüísticos

Abstract: This article presents CLARIN, a project that aims to promote the use of technological tools in research in the fields of the Humanities and Social Sciences.
Keywords: Humanities, social sciences, grid technology, web services, language resources and **t**echnologies.

1.1.1 Introducción

Presentamos el proyecto CLARIN (*Common Language Resources and Technology Infrastructure*), un proyecto de colaboración entre 22 países europeos cuyo objetivo es potenciar el uso de instrumentos tecnológicos en la investigación en ámbitos de las Humanidades y Ciencias Sociales. CLARIN creará la infraestructura necesaria para dar acceso genérico a grandes bancos de datos lingüísticos (textos, diccionarios, ontologías, etc.), así como a los instrumentos de análisis y explotación de estos datos (segmentadores, etiquetadores, analizadores sintácticos, etc.). Para ello se implementará, en una estructura de red grid, y mediante tecnología de servicios web y de web semántica, una única interfaz de acceso a los datos y a los instrumentos de análisis, así como a procesadores y otros servicios necesarios.

CLARIN es uno de los 35 proyectos seleccionados por el Comité ESFRI y que figuran en la "Hoja de ruta" de las infraestructuras que han de ser construidas, por su importancia para la investigación, a diez años vista.

2 Antecedentes

CLARIN tiene sus antecedentes en los trabajos para la estandarización de datos lingüísticos y de los instrumentos que los analizan, para garantizar la reusabilidad y la interoperatibilidad: EAGLES, OLIF (Lieske et al., 2001), ISLE (Atkins et al., 2002) y LIRICS-ISO (Framcopoulo et al., 2006); así como implementaciones directas de estas directrices: MULTEXT (Ide y Véronis, 1994), PAROLE (Zampoli, 1997) y SIMPLE (Lenci et al., 2000).

Por otro lado, el ver la explotación de datos lingüísticos vino de la mano de proyectos de investigación como LAMUS (Broeder et al., 2007), en los que se necesitaba archivar y gestionar datos lingüísticos en el área de tipología lingüística.

Más recientemente, se han llevado a cabo proyectos que han usado el enorme potencial que tiene la integración virtual de recursos distribuidos y autónomos ya existentes y que han demostrado la viabilidad de formar colecciones digitales virtuales. Algunos ejemplos son: IMDI (Wittenburg et al., 2002) y DAM-LR (Broeder et al., 2006).

ISSN: 1135-5948

3 La Infraestructura CLARIN

El objetivo de CLARIN es crear una infraestructura estable y persistente para dar acceso a los recursos lingüísticos y a sus instrumentos de análisis y explotación.

La infraestructura CLARIN consiste en la aplicación de la tecnología grid, del concepto de metadatos y de servicios web para, en primer lugar, garantizar la interoperabilidad que haga de un conjunto de elementos sin relación, diferentes y remotos, un sistema estructurado de componentes funcionales interconectados, y, en segundo lugar, para facilitar la identificación, la ubicación, el acceso y la explotación de recursos lingüísticos, entendiendo por recursos lingüísticos cualquier colección de datos en forma textual (hablada o escrita) o con información sobre lenguas y donde el objetivo de la tecnología sea el procesamiento del material lingüístico.

Por una parte, la tecnología grid permite utilizar de forma coordinada todo tipo de recursos (datos, procesos, servicios, etc.) sin necesidad de estar sujetos a un control centralizado. Estos recursos pueden ser heterogéneos y estar distribuidos geográficamente, es decir, pueden ser propiedad y/o estar administrados por diferentes instituciones. Por otra parte, los metadatos son una definición estándar, utilizada por todos los componentes del grid, para describir los contenidos de forma que haga posible la identificación y búsqueda unificadas de recursos y funcionalidades.

4 Planificación de CLARIN

CLARIN se encuentra actualmente en su primera fase (2008-2010), una fase preparatoria en la que se realizará una planificación detallada de la construcción de la infraestructura, con una estimación de costes reales, la definición de uso de la red y la definición de centros, recursos y tecnología que aseguren su mantenimiento de forma estable. En una segunda fase (2011-2015), está prevista la construcción de la infraestructura, con la integración de recursos y tecnologías, y el desarrollo de aplicaciones piloto que la usarán. Y, finalmente, está prevista la fase de plena explotación, con el desarrollo de aplicaciones más complejas e innovadoras.

El proyecto que cubre la fase preparatoria ha sido aprobado por la Comisión de la Unión Europea y cuenta con la participación de 32 miembros de 22 Estados miembros de la Unión, además de un amplio apoyo internacional. CLARIN ha recibido también apoyo del Ministerio de Educación, Subdirección General de Promoción e Infraestructuras Tecnológicas y Grandes Instalaciones (CAC-2007-23). Además, El DIUE de la Generalitat de Catalunya y la UPF han firmado un convenio para la financiación del desarrollo de un demostrador catalán para CLARIN.

Bibliografía

Atkins, S. et al. 2002. From Resources to Applications. Designing the Multilingual ISLE Lexical Entry. En *Proceedings of LREC*. Las Palmas de Gran Canaria, España.

Broeder, D. et al. *LAMUS – the Language Archive Management and Upload System*. <http://www.lat-mpi.eu/papers/papers2006/lamus-paper-final2.pdf>.

Broeder, D. et al. 2006. A Grid of Language Resource Repositories. En *Proceedings of the 2nd IEEE International Conference on e-Science and Grid Computing*. Amsterdam, Holanda.

Francopoulo, G. et al. 2006. Lexical Markup Framework (LMF). En *Proceedings of LREC*. Génova, Italia.

Ide, N. y Véronis, J. 1994. MULTEXT: Multilingual Text Tools and Corpora. En *Proceedings of the 15th International Conference on Computational Linguistics*. Kyoto, Japón.

Lenci, A. et al. 2000. SIMPLE: A General Framework for the Development of Multilingual Lexicons, *International Journal of Lexicography*. Vol. 13, núm. 4., pág. 249-263.

Lieske, C. at al. 2001. The Open Lexicon Interchange Format (OLIF) Comes of Age. En *Proceedings of the MT Summit VIII*. Santiago de Compostela, España.

Wittenburg, P. et al. 2002. Metadata Proposals for Corpora and Lexica. En *Proceedings of LREC*. Las Palmas de Gran Canaria, España.

Zampoli, A. 1997. The PAROLE project in the general context of the European actions for Language Resources. En *Proceedings of the Second European Seminar: Language Applications for a Multilingual Europe*. IDS/VDU, Manheim/Kaunas.