



El proyecto CLARIN: una infraestructura de investigación científica para las humanidades y las ciencias sociales

Núria Bel

Investigadora Ramón y Cajal en el Institut Universitari de Lingüística Aplicada
(Universitat Pompeu Fabra)
nuria.bel@upf.edu

Santiago Bel

Técnico superior informático en el Institut Universitari de Lingüística Aplicada
(Universitat Pompeu Fabra)
santiago.bel@upf.edu

Sergio Espeja

Técnico superior informático en el Institut Universitari de Lingüística Aplicada
(Universitat Pompeu Fabra)
sergio.espeja@upf.edu

Montserrat Marimon

Investigadora en el Institut Universitari de Lingüística Aplicada
(Universitat Pompeu Fabra)
montserrat.marimon@upf.edu

Marta Villegas

Investigadora en el Institut Universitari de Lingüística Aplicada
(Universitat Pompeu Fabra)
marta.villegas@upf.edu

Fecha de presentación: enero de 2008

Fecha de aceptación: febrero de 2008

Fecha de publicación: mayo de 2008

Cita recomendada:

BEL, Núria; BEL, Santiago; ESPEJA, Sergio; MARIMON, Montserrat, VILLEGAS, Marta (2008). "El proyecto CLARIN: una infraestructura de investigación científica para las humanidades y las ciencias sociales". *Digithum*, n.º 10 [artículo en línea]. DOI: <http://dx.doi.org/10.7238/d.v0i10.501>



Resumen

En este artículo presentamos CLARIN (*Common Language Resources and Technologies*), un proyecto de colaboración europea a gran escala cuyo objetivo es potenciar el uso de instrumentos tecnológicos en la investigación en los ámbitos de las humanidades y las ciencias sociales.

CLARIN es uno de los treinta y cinco proyectos seleccionados por el Comité ESFRI (*European Strategy Forum on Research Infrastructures*) para la lista de las infraestructuras que han de construir, por su importancia para la investigación, a diez años vista. CLARIN pretende llevar a las humanidades y a las ciencias sociales los beneficios del acceso compartido y colaborativo a recursos digitales, así como el uso del cómputo intensivo con instrumentos específicos de análisis y explotación para el acceso inteligente a grandes bases de datos. Con este objetivo, CLARIN va a crear la infraestructura necesaria para poder dar un acceso genérico a grandes bancos de datos, así como a los instrumentos de análisis y explotación de estos datos mediante la utilización de tecnología. Para ello 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 herramientas de procesamiento y a otros servicios necesarios. Esta interfaz, al estar diseñada para servir a los objetivos comunes de la investigación en humanidades y ciencias sociales, facilitará su uso a investigadores de diferentes ámbitos sin necesidad de que tengan conocimientos sobre las tecnologías implicadas.

Palabras clave

humanidades, ciencias sociales, tecnología *grid*, servicios web, tecnologías y recursos lingüísticos

Abstract

This article presents the CLARIN (*Common Language Resources and Technologies*) project, a large-scale pan-European collaborative project that aims to promote the use of technological tools in research in the fields of the humanities and social sciences.

CLARIN is one of the 35 projects selected by ESFRI (*European Strategy Forum on Research Infrastructures*) to form a list of infrastructures that need to be built, due to their importance in terms of research, in the next ten years. CLARIN aims to bring the benefits of shared and collaborative access to digital resources to the humanities and social sciences and increase use of specific analysis and exploitation computing tools for intelligent access to large databases. With this in mind, CLARIN is to create the infrastructure needed to offer generic access to large databases, alongside technological tools for the analysis and exploitation of the data. To do so, the project envisions a grid structure using web service and semantic web technologies, a single interface for accessing data and analysis tools, as well as the processing tools and other services needed. This interface, due to the fact that is designed to meet the common research aims in the humanities and social sciences, is easy to use by researchers from different fields without any prior knowledge of the technology involved.

Keywords

humanities, social sciences, grid technology, web services, language resources and technologies

—
precisamente porque las máquinas nos proporcionan y nos ordenan... los datos que les pedimos, debemos dedicar a la tarea de reflexionar sobre ellos buena parte del tiempo que antes empleábamos en conseguirlos.

Ignacio Bosque, *Diccionario REDES*, pág. XXIV.

errónea, ya que muchos de estos trabajos tienen un componente de búsqueda, acceso y análisis de datos básicamente lingüísticos que es posible optimizar mediante instrumentos y herramientas basadas en tecnologías lingüísticas.

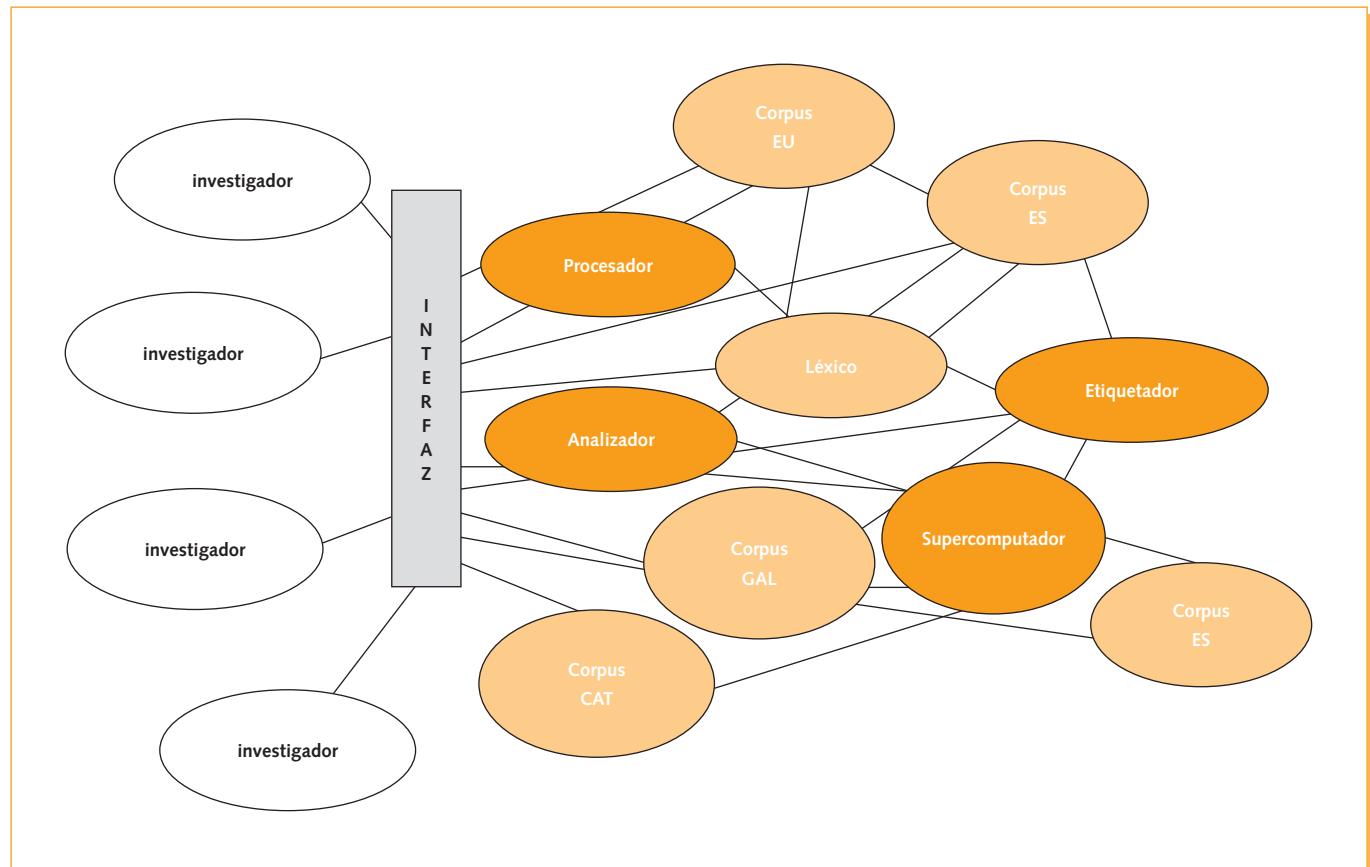
La disponibilidad de tecnología, de recursos computacionales y de cada vez más cantidad de datos digitalizados está llevando a que se planteen nuevas formas de enfocar la investigación en las humanidades y en las ciencias sociales. La posibilidad de crear y utilizar grandes colecciones digitales de recursos estructurados y no estructurados, y la disponibilidad de algoritmos y equipos de gran potencia computacional para procesar datos lingüísticos afectará profundamente a la investigación en disciplinas de estos ámbitos en los próximos años.

Introducción

La investigación en las áreas de las humanidades y las ciencias sociales aún se percibe en muchos sectores como un trabajo individual, más intelectual que experimental. Ésta es una percepción



Figura 1. Interfaz CLARIN de acceso a datos e instrumentos de análisis



En este contexto, presentamos el proyecto CLARIN (*Common Language Resources and Technologies*), un proyecto de colaboración entre veintidós países europeos cuyo objetivo es potenciar el uso de instrumentos tecnológicos en la investigación en ámbitos de las humanidades y las ciencias sociales. CLARIN pretende llevar a estos campos los beneficios del acceso compartido y colaborativo a recursos digitales, y el uso de cómputo intensivo con instrumentos específicos de análisis y explotación para que el acceso a grandes bancos de datos sea inteligente. Con este objetivo, CLARIN va a crear la infraestructura necesaria para poder dar acceso genérico a grandes bancos de datos lingüísticos (textos, grabaciones multimedia, diccionarios, ontologías, etc.), así como a los instrumentos de análisis y explotación de estos datos (segmentadores, etiquetadores, analizadores sintácticos, etc.) mediante la utilización de tecnología. 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. Esta interfaz, al estar diseñada para servir a los objetivos comunes de la investigación en las humanidades y en las ciencias sociales, facilitará

su uso a investigadores de diferentes ámbitos sin que se tenga la necesidad de poseer conocimientos sobre las tecnologías implicadas. Este factor es especialmente importante, ya que estos ámbitos han sido, tradicionalmente, impermeables a la innovación tecnológica.

CLARIN es uno de los 35 proyectos seleccionados por el Comité ESFRI (*European Strategy Forum on Research Infrastructures*) que figuran en la «hoja de ruta»^[www1] de las infraestructuras que se han de construir, por su importancia para la investigación, a diez años vista. La Comisión de la Unión Europea ha dispuesto la cofinanciación de una fase preparatoria para estos proyectos dentro del área Infraestructuras del VII Programa Marco, y el Ministerio de Educación y Ciencia, Dirección General de Política Tecnológica, Subdirección General de Promoción e Infraestructuras Tecnológicas y Grandes Instalaciones ha convocado un programa de «acciones complementarias» para cofinanciar esta fase preparatoria. El objetivo del proyecto es organizar la coordinación europea y elaborar un plan detallado de construcción de la infraestructura, con una evaluación de los costes y una propuesta de organización y gestión de la infraestructura en la que se incluirá también el estudio de todas aquellas cuestiones legales que puedan afectarle.

[www1]: <<http://cordis.europa.eu/esfri/roadmap.htm>>.



1. Antecedentes

CLARIN tiene sus antecedentes en los trabajos para la estandarización de datos lingüísticos, y de los instrumentos que los analizan para garantizar, primero, la reusabilidad de los recursos y, luego, la interoperabilidad. El *Expert Advisory Group on Linguistic Engineering Standards*^[www2] (EAGLES) fue el primer proyecto europeo sobre estándares para tecnologías lingüísticas. A este proyecto se le sumaron otros trabajos que perseguían también la interoperabilidad de recursos y herramientas como eran *Open Lexicon Interchange Format*^[www3] (OLIF; Lieske *et al.*, 2001), *International Standards for Language Engineering*^[www4] (ISLE; Atkins *et al.*, 2002) y *Linguistic Infrastructure for Interoperable Resources and Systems*^[www5] (LIRICS-ISO; Francopoulo *et al.*, 2006), así como implementaciones directas de estas directrices en los proyectos *Multilingual Text Tools and Corpora*^[www6] (MULTTEXT; Ide *et al.*, 1994), *Preparatory Action for linguistic Resources Organisation for Language Engineering* (PAROLE; Zampoli, 1997) y *Semantic Information for Multifunctional Plurilingual Lexicons*^[www7] (SIMPLE; Lenci *et al.*, 2000).

Por otro lado, el ver la explotación de datos lingüísticos como una de las necesidades del área de humanidades y ciencias sociales vino de la mano de proyectos de investigación en los que se necesitaba archivar y gestionar datos lingüísticos, por ejemplo, en el área de tipología lingüística, proyectos como *Language Archiving Technology*^[www8] (LAT) y *Language Archive Management and Upload System*^[www9] (LAMUS; Broeder *et al.*, 2007) que se encontraron con dificultades para reunir recursos de diferentes fuentes por la diversa estructuración y codificación de los datos.

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 a las que los investigadores puedan acceder sin necesidad de saber resolver los problemas que causan los formatos diferentes, las diversas bases de datos, o la codificación diferente de éstos. Algunos ejemplos de estos proyectos son: *ISLE Meta Data Initiative*^[www10] (IMDI; Wittenburg *et al.*, 2002) y *Distributed Access Management for Language Resources*^[www11] (DAM-LR; Broeder *et al.*, 2006).

2. Planificación de CLARIN

Como ya hemos mencionado, CLARIN se encuentra actualmente en su primera fase (2008-2010), una etapa preparatoria en la que se realizará una planificación detallada de la construcción de la infraestructura, con una estimación de costes reales de la infraestructura propuesta, 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 plataforma CLARIN, con la integración de recursos y tecnologías, y el desarrollo de aplicaciones piloto que usarán la infraestructura CLARIN. Finalmente, está prevista la fase de plena explotación de la infraestructura, con el desarrollo de aplicaciones más complejas e innovadoras.

El proyecto que cubrirá la fase preparatoria del proyecto CLARIN ha sido aprobado por la Comisión de la Unión Europea, cuenta con la participación de treinta y dos socios de veintidós estados miembros de la Unión y tiene un amplio apoyo internacional. Gracias al apoyo de usuarios y proveedores de recursos y tecnología potenciales que manifestaron su interés en el proyecto, CLARIN ha recibido también apoyo del Ministerio de Educación, Subdirección General de Promoción e Infraestructuras Tecnológicas y Grandes Instalaciones (CAC-2007-23).

Durante la fase preparatoria se desarrollará una maqueta, que asegurando su escalabilidad, sirva para verificar que CLARIN es viable, así como la metodología que se implementará para la construcción real de la infraestructura y para hacer una evaluación realista de los costes. Estos trabajos permitirán definir un proyecto detallado y fidedigno en el que se puedan basar las decisiones relativas a la construcción de la infraestructura y su financiación por parte de los estados miembros de la Unión Europea. Por otro lado, la fase preparatoria de CLARIN tendrá en cuenta la dimensión aplicada y específica de estos instrumentos a diferentes lenguas y áreas de investigación, y se propondrán demostradores en los que se trabaje con las diferentes lenguas de los estados participantes. En cada estado, además, el desarrollo de esta fase preparatoria ha de llevar a cabo estudios y acciones destinados a la identificación, formación y coordinación de usuarios y proveedores de recursos y tecnología a partir de los cuales se crearán comunidades para la identificación de sus necesidades, con especial hincapié en el español y en otras lenguas del Estado; además de la contribución

[www2]: <<http://www.ilc.cnr.it/EAGLES96/home.html>>.

[www3]: <<http://www.olif.net/>>.

[www4]: <http://www.ilc.cnr.it/EAGLES/isle/ISLE_Home_Page.htm>.

[www5]: <<http://lirics.loria.fr/>>.

[www6]: <<http://aune.lpl.univ-aiex.fr/projects/multext/>>.

[www7]: <<http://www.ub.es/gilcub/SIMPLE/simple.html>>.

[www8]: <<http://www.lat-mpi.eu/>>.

[www9]: <<http://www.lat-mpi.eu/tools/lamus>>.

[www10]: <<http://www.mpi.nl/IMDI>>.

[www11]: <<http://www.mpi.nl/DAM-LR>>.



con información relativa a cada estado para el proyecto común, y la participación con demostradores adaptada a los casos locales y su validación en el entorno de CLARIN.

Además de la definición técnica, otro objetivo de esta fase preparatoria en Europa es la definición del acuerdo entre todos los estados implicados en su construcción y su financiación. Se ha de definir el marco legal en el que se basará la construcción y explotación conjunta de la infraestructura. Más concretamente, el acuerdo regulará aspectos de la organización europea y nacional, siguiendo un modelo de federación de recursos, y el establecimiento y normativa de gestión de un registro que almacene y ubique todos los nodos de la red. También se ha de llevar a cabo el estudio de los problemas que se puedan derivar de los derechos de propiedad de los recursos que entren en la red, y una propuesta de modelo de uso.

3. La infraestructura CLARIN

El objetivo de CLARIN no es crear nuevos recursos lingüísticos ni más tecnología, sino establecer las condiciones necesarias; esto es, 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 éstos 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 hace posible la identificación y búsqueda unificadas de recursos y funcionalidades. De esta forma, se pretende evitar el problema actual que representa que estos recursos y tecnologías tengan diferente descripción y funcionamiento, que en último término son sólo conocidos por los mismos grupos e instituciones que los han creado, y que, por tanto, están infraexplotados.

4. La federación CLARIN

CLARIN se basa en la idea de federación, que, a su vez, parte de los siguientes principios básicos:

- En primer lugar, el despliegue de la red *grid* implica asignar identificadores únicos y persistentes a todos los recursos e instrumentos de análisis y explotación, y sincronizar los protocolos de identificación y autenticación de los participantes. Puede haber también diferentes niveles: federaciones nacionales agrupadas en una federación europea, por ejemplo.
- En segundo lugar, se ha de disponer de un registro de recursos y servicios, todos ellos descritos mediante metadatos estandarizados. Los metadatos se utilizan para la descripción tanto de los recursos y las herramientas de búsqueda y localización, como para la descripción de los contenidos lingüísticos, con el fin de construir herramientas de acceso, análisis y explotación que puedan interoperar. El registro único permite búsquedas exhaustivas de recursos y/o servicios que concuerden con los criterios de búsqueda expresados por el usuario. De esta forma, el usuario no ha de conocer la existencia del recurso y no depende tampoco de criterios de accesibilidad relacionados con la ubicación física del recurso.
- Finalmente, se ha de disponer de un sistema de autenticación única de los usuarios para el acceso a la infraestructura, de forma que el acceso a herramientas y servicios diferentes no implique tener claves de acceso también diferentes.

5. Algunos ejemplos

Por las características tan técnicas de CLARIN, puede ser difícil imaginar qué tipo de servicios podrá ofrecer, una vez desplegado, a los investigadores en humanidades y ciencias sociales. Pondremos, por tanto, algunos ejemplos de posibles usos de la infraestructura CLARIN. El lector ha de tener en cuenta que son ejemplos de lo que se podría hacer ahora si ya dispusiéramos de una infraestructura de red, como la que propone CLARIN.

El despliegue de la red *grid*, la estandarización de la descripción de recursos y la centralización del registro de metadatos permitirá desarrollar herramientas para identificar y acceder a datos y a servicios virtuales. Por ejemplo, será posible utilizar un servicio que, tomando los datos sobre las necesidades del investigador, identifique, ubique y acceda a estos datos automáticamente entregando listas o, de un modo directo, ficheros que contengan los datos buscados. También se podrán reunir recursos procedentes de distintos archivos, probablemente ubicados físicamente en diversos países, que cumplan ciertas condiciones para, por ejemplo, disponer de un servicio de «creación de un corpus de conversaciones



entre personas de 16 a 18 años con ejemplos etiquetados de abertura vocálica para marcar el plural en castellano peninsular». Y gracias a los estándares de codificación de recursos, en este caso de recursos de voz, todos éstos se podrán analizar por una única herramienta. También está previsto el alojamiento temporal de estos corpus de estudio virtuales para facilitar el análisis y la explotación.

La posibilidad de crear aplicaciones distribuidas y dinámicas (las que pueden acceder a diferentes recursos y solicitar servicios dependiendo de variables expresadas por el usuario) hará posible implementar buscadores con características especiales. Por ejemplo, los investigadores que trabajen con textos históricos podrán disponer de un servicio de acceso a un diccionario histórico que localice la forma antigua de una expresión actual, y que ésta dependa de la fecha del documento. De esta forma podremos encontrar documentos con la palabra actual «ejemplo» en textos del siglo XVI buscando por la forma «ensiemplo», contando con que el buscador añada esta forma a la consulta. Otro caso es la compilación de textos mediante búsqueda por palabras clave, que puede llegar a implicar el acceso a un servicio web de traducción de las palabras clave a diferentes lenguas, para conseguir datos independientemente de la lengua de consulta.

El desarrollo de este tipo de aplicaciones permitirá la creación de nuevos servicios y de aplicaciones más complejas, en las que los instrumentos de análisis y explotación de datos administrados por diferentes instituciones estarán concatenados, por ejemplo, para que investigadores que trabajen en el análisis de opinión puedan acceder a los datos de un tiempo particular de uno o más periódicos a través de sus hemerotecas, y puedan utilizar instrumentos automáticos –analizadores sintácticos, segmentadores, etiquetadores– para analizarlos y obtener la distribución de marcadores discursivos de opinión negativa sobre un tema determinado a lo largo del periodo de tiempo marcado.

6. Contactar con el equipo de CLARIN-ES

Como hemos mencionado, uno de los objetivos de la fase preparatoria es la identificación, formación y coordinación de usuarios y proveedores de recursos y tecnología a partir de los cuales se podrán identificar necesidades comunes a los usuarios de diferentes disciplinas, así como planear las aplicaciones que darán soluciones. Por lo tanto, invitamos a todos aquellos que estén interesados a ponerse en contacto con nosotros.

El proyecto CLARIN dispone de una página web (<http://clarin-es.iula.upf.edu>) en la que convocamos a proponer recursos lingüísticos e instrumentos de análisis y explotación para su integración en la infraestructura. Usuarios y proveedores pueden también ponerse en contacto con el Institut Universitari de Lingüística Aplicada, representante español en la fase preparatoria del proyecto CLARIN, a través del siguiente enlace: <http://clarin-es.iula.upf.edu/es/contacto>.

Bibliografía

ATKINS, S.; BEL, N.; BERTAGNA, F. [et al.] (2002). «From Resources to Applications. Designing the Multilingual ISLE Lexical Entry». En: *Proceedings of LREC*. Las Palmas de Gran Canaria.

BEL, N.; BUSA, F.; CALZOLARI, N. [et al.] (2000). «SIMPLE: A General Framework for the Development of Multilingual Lexicons». *International Journal of Lexicography*. Vol. 13. N.º 4., págs. 249-263.

BEL, N.; FRANCOPOULO, G.; GEORGE, M. [et al.] (2006). «Lexical Markup Framework (LMF)». En: *Proceedings of LREC*. Génova.

BROEDER, D.; CLAUS, A.; OFFENGA, F. [et al.] (2006). «LAMUS – the Language Archive Management and Upload System». [en línea]. <<http://www.lat-mpi.eu/papers/papers-2006/lamus-paper-final2.pdf>>.

BROEDER, D.; NATHAN, D.; STRÖMQVIST, S. [et al.] (2006). «A Grid of Language Resource Repositories». En: *Proceedings of the 2nd IEEE International Conference on e-Science and Grid Computing*. Ámsterdam.

BROEDER, D.; PETERS, W.; WITTENBURG, P. (2002). «Metadata Proposals for Corpora and Lexica». En: *Proceedings of LREC*. Las Palmas de Gran Canaria.

IDE, N.; VÉRONIS, J. (1994). «MULTEXT: Multilingual Text Tools and Corpora». En: *Proceedings of the 15th International Conference on Computational Linguistics*. Kioto.

LIESKE, C.; McCORMICK, S.; THURMAIR, G. (2001). «The Open Lexicon Interchange Format (OLIF) Comes of Age». En: *Proceedings of the MT Summit VIII*. Santiago de Compostela.

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*. Manheim / Kaunas: IDS/VDU.

**Núria Bel**

Investigadora Ramón y Cajal en el Institut Universitari de Lingüística Aplicada (Universitat Pompeu Fabra)
nuria.bel@upf.edu

Investigadora Ramón y Cajal en el Institut Universitari de Lingüística Aplicada (IULA) desde finales de 2003 y profesora de Procesamiento del lenguaje natural en el Departamento de Traducción y Filología de la Universidad Pompeu Fabra. Su actividad investigadora se ha desarrollado en los ámbitos de la traducción automática, la clasificación automática de documentos, los recursos lingüísticos para el procesamiento del lenguaje natural, su estandarización y las tecnologías relacionadas con ellos en el Grup d'Investigació en Lingüística Computacional de la Universitat de Barcelona (gilcUB) desde 1993 hasta 2003. Actualmente dirige el proyecto CLARIN en España y es investigadora principal del proyecto del Plan Nacional del Ministerio de Educación y Ciencia AAILE2: Adquisición Automática de Información Léxica.

**Santiago Bel**

Técnico superior informático en el Institut Universitari de Lingüística Aplicada (Universitat Pompeu Fabra)
santiago.bel@upf.edu

Licenciado en Ingeniería Informática por la Universitat Politècnica de Catalunya en 2004, ha participado en diferentes proyectos tecnológicos internacionales en el Reino Unido y en Grecia desde 2001. Actualmente es técnico contratado en el Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra (IULA).

**Sergio Espeja**

Técnico superior informático en el Institut Universitari de Lingüística Aplicada (Universitat Pompeu Fabra)
sergio.espeja@upf.edu

Licenciado en Ingeniería Informática por la Universitat Politècnica de Catalunya. Ha trabajado con tecnologías web desde el año 1999. Lidera varios proyectos de desarrollo *open-source*, entre otros: *Bayesian networks for Ruby* y *Coldic*. Ha realizado ponencias sobre desarrollo web en congresos como el RailsConf Europe 2007 o la Conferencia Rails Hispana 2006 y 2007. Desde 2005 es técnico superior de apoyo a la investigación en el Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra, cofinanciado por el Programa de Personal Técnico de Apoyo (PTA-CTE/1370/2003).

**Montserrat Marimon**

Investigadora en el Institut Universitari de Lingüística Aplicada (Universitat Pompeu Fabra)
montserrat.marimon@upf.edu

Licenciada en Filología Inglesa por la Universitat de Barcelona y doctorada en Ciencias Informáticas por la Universitat Politècnica de Catalunya. Empezó la investigación en procesamiento del lenguaje natural en el Grup d'Investigació en Lingüística Computacional de la Universitat de Barcelona (gilcUB) en 1994. Trabajó durante dos años como investigadora en el Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e.V. an der Universität des Saarlandes (Saarbrücken, Alemania). En 2004 se unió al Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra al haber sido seleccionada por el programa Juan de la Cierva del Ministerio de Educación y Ciencia. Actualmente trabaja como investigadora contratada en el IULA y coordina el proyecto CLARIN.



<http://digithum.uoc.edu>

El proyecto CLARIN: una infraestructura de investigación científica...



Marta Villegas

Investigadora en el Institut Universitari de Lingüística Aplicada (Universitat Pompeu Fabra)

marta.villegas@upf.edu

Licenciada en Filología Inglesa por la Universitat de Barcelona y doctora en Ciencias Informáticas por la Universitat Politècnica de Catalunya. Empezó la investigación en procesamiento del lenguaje natural en el Grup d'Investigació en Lingüística Computacional de la Universitat de Barcelona (gilcub) en 1993, donde trabajó hasta junio de 2004. Durante dos años (1998-1999), además, trabajó como investigadora en el Institut d'Estudis Catalans. Desde 2006 trabaja como investigadora contratada en el Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra.



Esta obra está sujeta a la licencia **Reconocimiento-NoComercial-SinObraDerivada 2.5 España** de Creative Commons. Puede copiarla, distribuirla y comunicarla públicamente siempre que cite su autor y la revista que la publica (*Digithum*), no la utilice para fines comerciales y haga con ella obra derivada. La licencia completa se puede consultar en: <http://creativecommons.org/licenses/by-nc-nd/2.5/es/deed.es/>