Universitat Oberta de Catalunya

DIGITHUM
Les humanitats en l'era digital

http://digithum.uoc.edu

El projecte CLARIN: una infraestructura de recerca científica per a les humanitats i les ciències socials

Núria Bel
Investigadora Ramón y Cajal de l'Institut Universitari de Lingüística Aplicada
(Universitat Pompeu Fabra)
nuria.bel@upf.edu

Santiago Bel
Tècnic superior informàtic de l'Institut Universitari de Lingüística Aplicada
(Universitat Pompeu Fabra)
santiago.bel@upf.edu

Sergio Espeja
Tècnic superior informàtic de l'Institut Universitari de Lingüística Aplicada
(Universitat Pompeu Fabra)
sergio.espeja@upf.edu

Montserrat Marimon
Investigadora de l'Institut Universitari de Lingüística Aplicada
(Universitat Pompeu Fabra)
montserrat.marimon@upf.edu

Marta Villegas
Investigadora de l'Institut Universitari de Lingüística Aplicada
(Universitat Pompeu Fabra)
marta.villegas@upf.edu

Data de presentació: gener de 2008
Data d'acceptació: febrer de 2008
Data de publicació: maig de 2008

Citació recomanada:

BEL, Núria; BEL, Santiago; ESPEJA, Sergio; MARIMON, Montserrat, VILLEGAS, Marta (2008). "El projecte CLARIN: una infraestructura de recerca científica per a les humanitats i les ciències socials". *Digithum*, núm. 10 [article en línia]. DOI: http://dx.doi.org/10.7238/d.v0i10.501

Resum

En aquest article presentem CLARIN (*Common Language Resources and Technologies*), un projecte de col·laboració europea a gran escala l'objectiu del qual és potenciar l'ús d'instruments tecnològics en la recerca en els àmbits de les humanitats i les ciències socials.

CLARIN és un dels trenta-cinc projectes seleccionats pel Comitè ESFRI (*European Strategy Forum on Research Infraestructures)* per a la llista de les infraestructures que s'han d'haver construït, per la seva importància per a la recerca, d'aquí a deu anys. CLARIN vol portar a les humanitats i a les ciències socials els beneficis de l'accés compartit i en col·laboració a recursos digitals, i també l'ús del còmput intensiu amb instruments específics d'anàlisi i explotació per a l'accés intel·ligent a grans bases de dades. Amb aquest objectiu, CLARIN crearà la infraestructura necessària per a poder donar un accés genèric a grans bancs de dades i als instruments d'anàlisi i explotació d'aquestes dades mitjançant la utilització de tecnologia. Per a això implementarà, en una estructura de xarxa *grid*, i mitjançant tecnologia de serveis web i de web semàntic, una única interfície d'accés a les dades i als instruments d'anàlisi, i també a eines de processament i altres serveis necessaris. Aquesta interfície, pel fet de ser dissenyada per a servir els objectius comuns de la recerca en humanitats i ciències socials, en facilitarà l'ús a investigadors de diferents àmbits sense necessitat de tenir coneixements sobre les tecnologies implicades.

Paraules clau

humanitats, ciències socials, tecnologia *grid*, serveis web, tecnologies i recursos lingüístics

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.

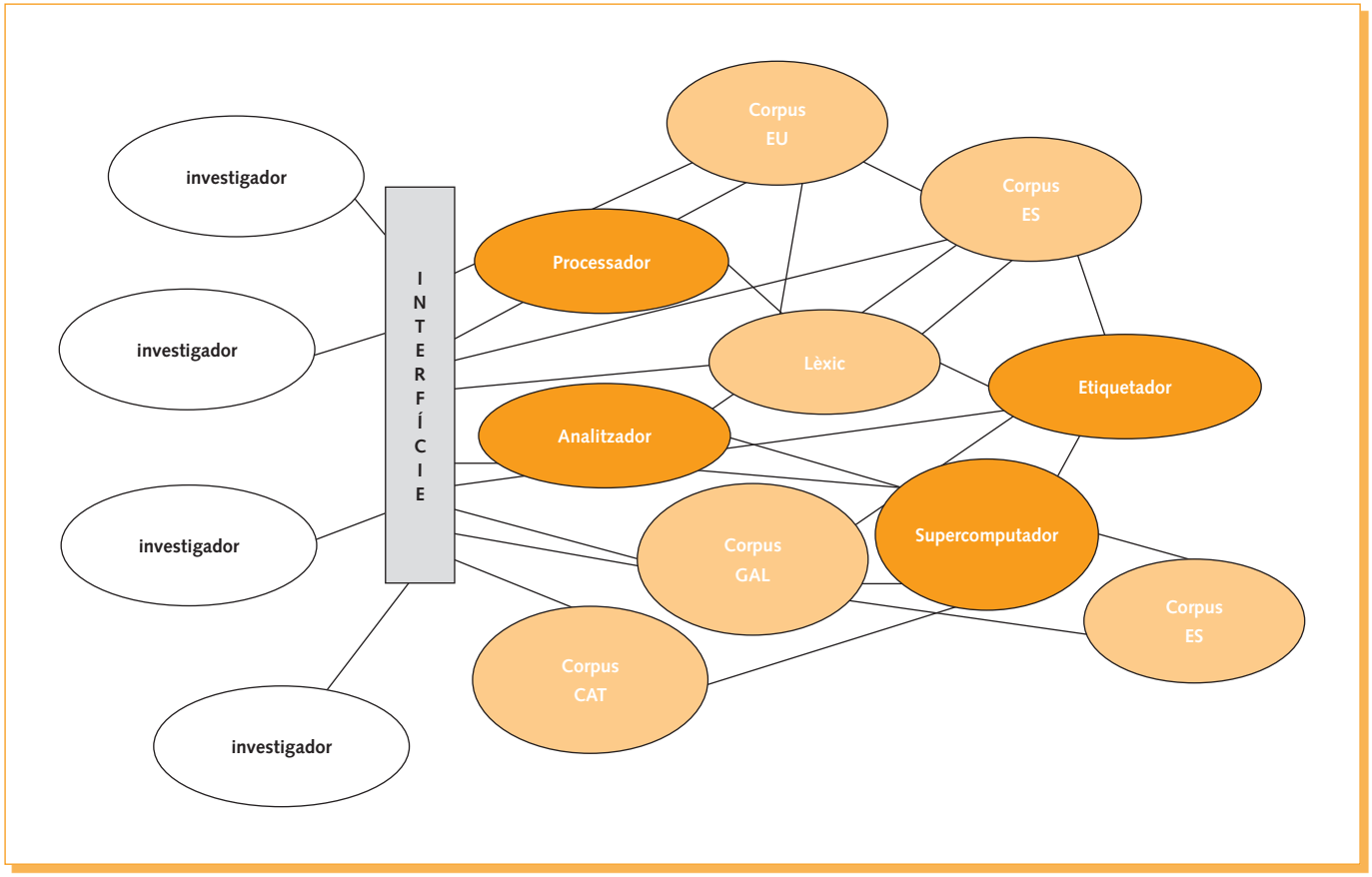
Introducció

La recerca en les àrees de les humanitats i les ciències socials encara es percep en molts sectors com un treball individual, més intel·lectual que experimental. Aquesta és una percepció errònia, ja que molts d'aquests treballs tenen un component de cerca, accés i anàlisi de dades bàsicament lingüístiques que és possible optimitzar mitjançant instruments i eines basades en tecnologies lingüístiques.

La disponibilitat de tecnologia, de recursos computacionals i de més i més quantitat de dades digitalitzades fa que es plantegin noves maneres d'enfocar la recerca en les humanitats i en les ciències socials. La possibilitat de crear i utilitzar grans col·leccions digitals de recursos estructurats i no estructurats, i la disponibilitat d'algoritmes i equips de gran potència computacional per a processar dades lingüístiques afectarà profundament la recerca en disciplines d'aquests àmbits en els pròxims anys.

En aquest context, presentem el projecte CLARIN (*Common Language Resources and Technologies*), un projecte de col·laboració entre vint-i-dos països europeus l'objectiu dels qual

Figura 1. Interfície CLARIN d'accés a dades i instruments d'anàlisi



és potenciar l'ús d'instruments tecnològics en la recerca en els àmbits de les humanitats i les ciències socials. CLARIN vol portar a aquests camps els beneficis de l'accés compartit i en col·laboració a recursos digitals, i l'ús de còmput intensiu amb instruments específics d'anàlisi i explotació perquè l'accés a grans bancs de dades sigui intel·ligent. Amb aquest objectiu, CLARIN crearà la infraestructura necessària per a poder donar accés genèric a grans bancs de dades lingüístiques (textos, enregistraments multimèdia, diccionaris, ontologies, etc.), i també als instruments d'anàlisi i explotació d'aquestes dades (segmentadors, etiquetadors, analitzadors sintàctics, etc.) mitjançant la utilització de tecnologia. Per a això s'implementarà, en una estructura de xarxa *grid*, i mitjançant tecnologia de serveis web i de web semàntic, una única interfície d'accés a les dades i als instruments d'anàlisi i a processadors i altres serveis necessaris. Aquesta interfície, pel fet de ser dissenyada per a respondre als objectius comuns de la recerca en les humanitats i en les ciències socials, en facilitarà l'ús a investigadors de diferents àmbits sense que hi hagi la necessitat de tenir coneixements sobre les tecnologies implicades. Aquest factor és especialment important, ja que aquests àmbits han estat, tradicionalment, impermeables a la innovació tecnològica.

CLARIN és un dels trenta-cinc projectes seleccionats pel comitè ESFRI (*European Strategy Forum on Research Infraestructures)* que figuren en el «full de ruta»[www1] de les infraestructures que s'han de construir, per la seva importància per a la recerca, a deu anys vista. La Comissió de la Unió Europea ha disposat el cofinançament d'una fase preparatòria per a aquests projectes dins de l'àrea Infraestructures del VII Programa Marc, i el Ministeri d'Educació i Ciència, Direcció General de Política Tecnològica, Sotsdirecció General de Promoció i Infraestructures Tecnològiques i Grans Instal·lacions ha convocat un programa d'«accions complementàries» per a cofinançar aquesta fase preparatòria. L'objectiu del projecte és organitzar la coordinació europea i elaborar un pla detallat de construcció de la infraestructura, amb una avaluació dels costos i una proposta d'organització i gestió de la infraestructura en la qual s'inclourà també l'estudi de totes les qüestions legals que la puguin afectar.

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

1. Antecedents

CLARIN té els seus antecedents en els treballs per a l'estandardització de dades lingüístiques i dels instruments que les analitzen per a garantir, primer, la reusabilitat dels recursos i, després, la interoperabilitat. L'*Expert Advisory Group on Linguistic Engineering Standards*[www2] (EAGLES) va ser el primer projecte europeu sobre estàndards per a tecnologies lingüístiques. A aquest projecte se li van sumar altres treballs que perseguien també la interoperabilitat de recursos i eines com *Open Lexicon Interchange Format*[www3] (OLIF; Lieske *et al.*, 2001), *International Standards for Language Engineering*[www4] (ISLE; Atkins *et al.*, 2002) i *Linguistic Infrastructure for Interoperable Resources and Systems*[www5] (LIRICS-ISO; Francopoulo *et al.*, 2006), i també implementacions directes d'aquestes directrius en els projectes *Multilingual Text Tools and Corpora*[www6] (MULTEXT; Ide *et al.*, 1994), *Preparatory Action for linguistic Resources Organisation for Language Engineering* (PAROLE; Zampoli, 1997) i *Semantic Information for Multifunctional Plurilingual Lexicons*[www7] (SIMPLE; Lenci *et al.*, 2000).

D'altra banda, el fet de veure l'explotació de dades lingüístiques com una de les necessitats de l'àrea d'humanitats i ciències socials va anar acompanyat de projectes de recerca en els quals es necessitava arxivar i gestionar dades lingüístiques, per exemple, en l'àrea de tipologia lingüística, projectes com *Language Archiving Technology*[www8] (LAT) i *Language Archive Management and Upload System*[www9] (LAMUS; Broeder *et al.*, 2007) que es van trobar amb dificultats per a reunir recursos de diferents fonts per la diversa estructuració i codificació de les dades.

Més recentment, s'han dut a terme projectes que han usat l'enorme potencial que té la integració virtual de recursos distribuïts i autònoms ja existents i que han demostrat la viabilitat de formar col·leccions digitals virtuals a les quals els investigadors puguin accedir sense necessitat de saber resoldre els problemes que causen els formats diferents, les diverses bases de dades o la seva diferent codificació. Alguns exemples d'aquests projectes són *ISLE Meta Data Initiative*[www10] (IMDI; Wittenburg *et al.*, 2002) i *Distributed Access Management for Language Resources*[www11] (DAM-LR; Broeder *et al.*, 2006).

2. Planificació de CLARIN

Com ja hem esmentat, CLARIN actualment és en la seva primera fase (2008-2010), una etapa preparatòria en què es farà una planificació detallada de la construcció de la infraestructura, amb una estimació de costos reals de la infraestructura proposada, la definició d'ús de la xarxa i la definició de centres, recursos i tecnologia que n'assegurin el manteniment de manera estable. En una segona fase (2011-2015), hi ha prevista la construcció de la plataforma CLARIN, amb la integració de recursos i tecnologies, i el desenvolupament d'aplicacions pilot que usaran la infraestructura CLARIN. I, finalment, hi ha prevista la fase de plena explotació de la infraestructura, amb el desenvolupament d'aplicacions més complexes i innovadores.

El projecte que cobrirà la fase preparatòria del projecte CLARIN ha estat aprovat per la Comissió de la Unió Europea, hi participen trenta-dos socis de vint-i-dos estats membres de la Unió i té un ampli suport internacional. Gràcies al suport d'usuaris i proveïdors de recursos i tecnologia potencials que van manifestar interès en el projecte, CLARIN ha rebut també suport del Ministeri d'Educació, Sotsdirecció General de Promoció i Infraestructures Tecnològiques i Grans Instal·lacions (CAC-2007-23).

Durant la fase preparatòria es desenvoluparà una maqueta que, assegurant-ne l'escalabilitat, serveixi per a verificar que CLARIN és viable, i també la metodologia que s'implementarà per a la construcció real de la infraestructura i per a fer una avaluació realista dels costos. Aquests treballs permetran definir un projecte detallat i fidedigne en el qual es puguin basar les decisions relatives a la construcció de la infraestructura i el seu finançament per part dels estats membres de la Unió Europea. D'altra banda, la fase preparatòria de CLARIN tindrà en compte la dimensió aplicada i específica d'aquests instruments a diferents llengües i àrees de recerca, i es proposaran demostradors en els quals es treballi amb les diferents llengües dels estats participants. A cada estat, a més, el desenvolupament d'aquesta fase preparatòria ha de dur a terme estudis i accions destinades a la identificació, la formació i la coordinació d'usuaris i proveïdors de recursos i tecnologia a partir dels quals es crearan comunitats per a la identificació de les seves necessitats, amb especial recalcament en l'espanyol i en altres llengües de l'Estat; a més de la contribució amb informació relativa a cada estat per al projecte comú i la participació amb

[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-aix.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/>.

demostradors adaptada als casos locals i la seva validació en l'entorn de CLARIN.

A més de la definició tècnica, un altre objectiu d'aquesta fase preparatòria a Europa és la definició de l'acord entre tots els estats implicats en la seva construcció i el seu finançament. S'ha de definir el marc legal en què es basarà la construcció i explotació conjuntes de la infraestructura. Més concretament, l'acord regularà aspectes de l'organització europea i nacional, seguint un model de federació de recursos, i l'establiment i normativa de gestió d'un registre que emmagatzemi i ubiqui tots els nodes de la xarxa. També s'ha de dur a terme l'estudi dels problemes que es puguin derivar dels drets de propietat dels recursos que entrin a la xarxa, i una proposta de model d'ús.

3. La infraestructura CLARIN

L'objectiu de CLARIN no és crear nous recursos lingüístics ni més tecnologia, sinó establir les condicions necessàries; és a dir, una infraestructura estable i persistent per a donar accés als recursos lingüístics i als seus instruments d'anàlisi i explotació.

La infraestructura CLARIN consisteix en l'aplicació de la tecnologia *grid*, del concepte de metadades i de serveis web per a, en primer lloc, garantir la interoperabilitat que faci d'un conjunt d'elements sense relació, diferents i remots, un sistema estructurat de components funcionals interconnectats, i, en segon lloc, facilitar la identificació, la ubicació, l'accés i l'explotació de recursos lingüístics, entenent per aquests recursos qualsevol col·lecció de dades en forma textual (parlada o escrita) o amb informació sobre llengües, i on l'objectiu de la tecnologia sigui el processament del material lingüístic.

D'una banda, la tecnologia *grid* permet utilitzar de manera coordinada tot tipus de recursos (dades, processos, serveis, etc.) sense necessitat d'estar subjectes a un control centralitzat. Aquests recursos poden ser heterogenis i ser distribuïts geogràficament, és a dir, poden ser propietat i/o ser administrats per diferents institucions. D'altra banda, les metadades són una definició estàndard, utilitzada per tots els components del *grid*, per a descriure els continguts de manera que fa possible la identificació i cerca unificades de recursos i funcionalitats. D'aquesta manera, es vol evitar el problema actual que representa que aquests recursos i tecnologies tinguin diferent descripció i funcionament, que en darrer terme són només coneguts pels mateixos grups i les mateixes institucions que els han creat, i que, per tant, estan infraexplotats.

4. La federació CLARIN

CLARIN es basa en la idea de federació, que, al seu torn, parteix dels principis bàsics següents:

- En primer lloc, el desplegament de la xarxa *grid* implica assignar identificadors únics i persistents a tots els recursos i instruments d'anàlisi i explotació, i sincronitzar els protocols d'identificació i autenticació dels participants. Hi pot haver nivells també diferents: federacions nacionals agrupades en una federació europea, per exemple.
- En segon lloc, s'ha de disposar d'un registre de recursos i serveis, tots ells descrits mitjançant metadades estandarditzades. Les metadades s'utilitzen per a la descripció tant dels recursos i les eines de cerca i localització com per a la descripció dels continguts lingüístics, a fi de construir eines d'accés, anàlisi i explotació que puguin interoperar. El registre únic permet cerques exhaustives de recursos i/o serveis que concordin amb els criteris de cerca expressats per l'usuari. D'aquesta manera, l'usuari no ha de conèixer l'existència del recurs i no depèn tampoc de criteris d'accessibilitat relacionats amb la ubicació física del recurs.
- Finalment, s'ha de disposar d'un sistema d'autenticació única dels usuaris per a l'accés a la infraestructura, de manera que l'accés a eines i serveis diferents no impliqui tenir claus d'accés també diferents.

5. Alguns exemples

Per les característiques tan tècniques de CLARIN, pot ser difícil imaginar quin tipus de serveis podrà oferir, una vegada desplegat, als investigadors en humanitats i en ciències socials. Posarem, per tant, alguns exemples de possibles usos de la infraestructura CLARIN. El lector ha de tenir en compte que són exemples del que es podria fer ara si ja disposéssim d'una infraestructura de xarxa, com la que proposa CLARIN.

El desplegament de la xarxa *grid*, l'estandardització de la descripció de recursos i la centralització del registre de metadades permetrà desenvolupar eines per a identificar i accedir a dades i a serveis virtuals. Per exemple, serà possible utilitzar un servei que, prenent les dades sobre les necessitats de l'investigador, identifiqui, ubiqui i accedeixi a aquestes dades automàticament lliurant llistes o, d'una manera directa, fitxers que continguin les dades buscades. També es podran reunir recursos procedents de diferents fitxers, probablement ubicats físicament en diversos països, que compleixin certes condicions per a, per exemple, disposar d'un servei de «creació d'un corpus de converses entre persones de 16 a 18 anys amb exemples etiquetats d'obertura vocàlica per a marcar el plural en castellà peninsular». I gràcies als estàndards de codificació de recursos, en aquest cas de veu, tots aquests recursos es podran analitzar amb una única eina. També hi ha previst l'allotjament temporal d'aquests corpus d'estudi virtuals per a facilitar l'anàlisi i l'explotació.

La possibilitat de crear aplicacions distribuïdes i dinàmiques (les que poden accedir a diferents recursos i sol·licitar serveis depenent

de variables expressades per l'usuari) farà possible implementar cercadors amb característiques especials. Per exemple, els investigadors que treballin amb textos històrics podran disposar d'un servei d'accés a un diccionari històric que localitzi la forma antiga d'una expressió actual, i que aquesta expressió depengui de la data del document. D'aquesta manera podrem trobar documents amb la paraula actual «ejemplo» en textos del segle XVI buscant per la forma «ensiemplo», comptant que el cercador afegeixi aquesta forma a la consulta. Un altre cas és la compilació de textos mitjançant la cerca per paraules clau, que pot arribar a implicar l'accés a un servei web de traducció de les paraules clau a diferents llengües, per a aconseguir dades independentment de la llengua de consulta.

El desenvolupament d'aquest tipus d'aplicacions permetrà la creació de nous serveis i d'aplicacions més complexes en què els instruments d'anàlisi i explotació de dades administrades per diferents institucions estaran concatenats, per exemple, perquè investigadors que treballin en l'anàlisi d'opinió puguin accedir a les dades d'un temps particular d'un o més diaris per mitjà de les seves hemeroteques, i puguin utilitzar instruments automàtics —analitzadors sintàctics, segmentadors, etiquetadors— per a analitzar-los i obtenir la distribució de marcadors discursius d'opinió negativa sobre un tema determinat al llarg del període de temps marcat.

6. Contactar amb l'equip de CLARIN-ES

Com hem esmentat, un dels objectius de la fase preparatòria és la identificació, formació i coordinació d'usuaris i proveïdors de recursos i tecnologia a partir dels quals es podran identificar necessitats comunes als usuaris de diferents disciplines, i també planejar les aplicacions que donaran solucions. Per tant, convidem tothom qui hi estigui interessat a posar-se en contacte amb nosaltres.

El projecte CLARIN disposa d'una pàgina web (http://clarin-es.iula.upf.edu) en la qual convoquem els visitants a proposar recursos lingüístics i instruments d'anàlisi i explotació per a la seva integració en la infraestructura. Usuaris i proveïdors poden posar-se també en contacte amb l'Institut Universitari de Lingüística Aplicada, representant espanyol en la fase preparatòria del projecte CLARIN, per mitjà del següent enllaç: http://clarin-es.iula.upf.edu/es/contacto.

Bibliografia

ATKINS, S.; BEL, N.; BERTAGNA, F. [*et al.*] (2002). «From Resources to Applications. Designing the Multilingual ISLE Lexical Entry». A: *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. Núm. 4. Pàgs. 249-263.

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

BROEDER, D.; CLAUS, A.; OFFENGA, F. [*et al.*] (2006). «*LAMUS – the Language Archive Management and Upload System*» [en línia]. <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». A: *Proceedings of the 2nd IEEE International Conference on e-Science and Grid Computing*. Amsterdam.

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

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

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

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

Núria Bel
Investigadora Ramón y Cajal de l'Institut Universitari de Lingüística Aplicada (Universitat Pompeu Fabra)
nuria.bel@upf.edu

Investigadora Ramón y Cajal de l'Institut Universitari de Lingüística Aplicada (IULA) des del final de 2003 i professora de Processament del llenguatge natural del Departament de Traducció i Filologia de la Universitat Pompeu Fabra. La seva activitat investigadora s'ha desenvolupat en els àmbits de la traducció automàtica, la classificació automàtica de documents, els recursos lingüístics per al processament del llenguatge natural, la seva estandardització i les tecnologies relacionades en el Grup d'Investigació en Lingüística Computacional de la Universitat de Barcelona (gilcUB) des de 1993 fins a 2003. Actualment dirigeix el projecte CLARIN a Espanya i és investigadora principal del projecte del Pla Nacional del Ministeri d'Educació i Ciència AAILE2: Adquisició Automàtica d'Informació Lèxica.

Santiago Bel
Tècnic superior informàtic de l'Institut Universitari de Lingüística Aplicada (Universitat Pompeu Fabra)
santiago.bel@upf.edu

Llicenciat en Enginyeria Informàtica per la Universitat Politècnica de Catalunya el 2004, ha participat en diferents projectes tecnològics internacionals al Regne Unit i a Grècia des de 2001. Actualment és tècnic contractat de l'Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra (IULA).

Sergio Espeja
Tècnic superior informàtic de l'Institut Universitari de Lingüística Aplicada (Universitat Pompeu Fabra)
sergio.espeja@upf.edu

Llicenciat en Enginyeria Informàtica per la Universitat Politècnica de Catalunya. Ha treballat amb tecnologies web des de l'any 1999. Lidera diversos projectes de desenvolupament de codi obert, entre d'altres, *Bayesian networks for Ruby* i Coldic. Ha fet ponències sobre desenvolupament web a congressos com el RailsConf Europe 2007 o la Conferència Rails Hispana 2006 i 2007. Des de 2005 és tècnic superior de suport a la recerca de l'Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra, cofinançat pel Programa de Personal Tècnic de Suport (PTA-CTE/1370/2003).

Montserrat Marimon
Investigadora de l'Institut Universitari de Lingüística Aplicada (Universitat Pompeu Fabra)
montserrat.marimon@upf.edu

Llicenciada en Filologia Anglesa per la Universitat de Barcelona i doctorada en Ciències Informàtiques per la Universitat Politècnica de Catalunya. Va començar la recerca en processament del llenguatge natural en el Grup d'Investigació en Lingüística Computacional de la Universitat de Barcelona (gilcUB) el 1994. Va treballar durant dos anys com a investigadora a l'Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e.V an der Universität des Saarlandes (Saarbrücken, Alemanya). El 2004 es va unir a l'Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra per haver estat seleccionada pel programa Juan de la Cierva del Ministeri d'Educació i Ciència. Actualment treballa d'investigadora contractada a l'IULA i coordina el projecte CLARIN.

Marta Villegas
Investigadora de l'Institut Universitari de Lingüística Aplicada (Universitat Pompeu Fabra)
marta.villegas@upf.edu

Llicenciada en Filologia Anglesa per la Universitat de Barcelona i doctora en Ciències Informàtiques per la Universitat Politècnica de Catalunya. Va començar la recerca en processament del llenguatge natural en el Grup d'Investigació en Lingüística Computacional de la Universitat de Barcelona (gilcUB) el 1993, on va treballar fins al juny de 2004. Durant dos anys (1998-1999), a més, va treballar com a investigadora a l'Institut d'Estudis Catalans. Des de 2006 treballa d'investigadora contractada a l'Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra.

UOC
Universitat Oberta de Catalunya
www.uoc.edu