Procesamiento de Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 185-188 recibido 14-04-12 revisado 30-05-12 aceptado 05-06-12

METANET4U: Enhancing the European Linguistic Infrastructure

METANET4U: Aumentar la Infraestructura Lingüística Europea

Núria Bel
Universitat Pompeu Fabra
Roc Boronat, 138
08028 Barcelona
nuria.bel@upf.edu

Asunción Moreno
Universitat Politècnica de Catalunya
Jordi Girona 1-3 Edifici D5
08034 Barcelona
asuncion.moreno@upc.edu

Resumen: El proyecto METANET4U está contribuyendo a la creación de una plataforma digital pan-europea que sustentará la distribución y el intercambio de recursos y servicios lingüísticos con el objetivo último de apoyar el desarrollo de aplicaciones basadas en tecnologías lingüísticas.
Palabras clave: recursos lingüísticos, infraestructuras

Abstract: METANET4U Project is participating in the creation of a digital pan-European platform that will support the distribution and the interchange of linguistic resources and services. The ultimate goal is to support the development of applications based on Language Technologies.
Keywords: Language Resources, infraestructures

1 Introducción

El proyecto METANET4U está contribuyendo a la creación de una plataforma digital pan-europea que sustentará la distribución y el intercambio de recursos y servicios lingüísticos con el objetivo último de apoyar el desarrollo de aplicaciones basadas en tecnologías lingüísticas. METANET4U es un proyecto co-financiado al 50% por el programa CIP-PSP (ICT *Policy Support Programme Competitiveness and Innovation framework Programme*) y por el consorcio participante. En él participan el *Institut Universitari de Lingüística Aplicada*[1], IULA, de la Universitat Pompeu Fabra y el *Centre de Tecnologies i Aplicacions del Llenguatge i la Parla*[2], TALP de la Universitat Politècnica de Catalunya. Coordinado por la Universidad de Lisboa, cuenta además con la participación de los siguientes centros: la Universidad de Manchester, la Universidad Alexandru Ioan Cuza, Insitutul de Cercetarri Pentru Inteligenta Artificiala y la Universidad de Malta.

El proyecto, que empezó en febrero de 2011, tiene prevista una duración de 24 meses y forma parte de la red de excelencia META-NET: *Multilingual Europe Technology Alliance*[3].

En este artículo presentamos en la sección 2 el contexto de iniciativas europeas que sirve de marco para entender las acciones de la red META-NET y de sus proyectos asociados. En la sección 3, se presenta la iniciativa META-SHARE para compartir recursos lingüísticos. En la sección 4, se presentan las contribuciones concretas realizadas en el proyecto METANET4U por las dos universidades españolas participantes y en la sección 5 se muestran las conclusiones.

2 METANET4U en contexto

El Séptimo Programa Marco de la Unión Europea (2007-2013) está teniendo una gran repercusión para el ámbito de las tecnologías lingüísticas en Europa. El Programa Marco ha

[1] http://www.iula.upf.edu
[2] http://www.talp.cat
[3] http://www.meta-net.eu

ISSN 1135-5948

dado financiación a un número importante de proyectos (hasta el momento más de 25 de ellos con participación española) en el ámbito de las tecnologías lingüísticas gracias a la creciente preocupación por la diversidad lingüística y las aplicaciones que estas tecnologías pueden tener en la llamada Europa digital. Además, varias acciones financiadas por la Comisión Europea han fomentado la creación de un espacio para la concertación y la cooperación entre todos los agentes y miembros de la comunidad de desarrolladores y de usuarios potenciales de estas tecnologías, con el objetivo de definir las prioridades para el óptimo desarrollo de este ámbito en el futuro. Este espacio se ha materializando en diferentes iniciativas, quizá las más conocidas han sido CLARIN[4], FLaReNet[5] y META-NET. Estas tres redes comparten el objetivo de fomentar el uso y la aplicación de tecnologías lingüísticas en diferentes aplicaciones, usos o audiencia. En particular han coincidido en proponer la visión de que es necesario crear una infraestructura que facilite la investigación y el desarrollo en este ámbito.

CLARIN (*Common Language Resources and Technology Infrastructure*, 2009-2011) ha sido la fase piloto de lo que es ya una infraestructura europea de investigación que pretende dar apoyo a los investigadores en humanidades para el acceso y explotación de textos gracias, entre otros factores, al uso de tecnologías lingüísticas. Ha desarrollado una red de instituciones colaboradoras de más de 200 nodos, y a partir de la participación de 9 Estados miembros de la Unión han constituido el CLARIN *European Research Infrastructure Consortium* (CLARIN-ERIC).

FLaReNet (Fostering Language Resources Network, 2009-2011) ha constituido una red de 99 miembros institucionales y 365 socios representando a 33 países diferentes, con la misión de consensuar y difundir las prioridades y objetivos estratégicos de la toda la comunidad relacionada con los la tecnología y recursos lingüísticos. Los resultados de los estudios y discusiones (Calzolari et al., 2012) se han hecho públicos y ya han sido enviados a representantes de instituciones europeas y nacionales para dar las claves de la futura coordinación del área.

En las primeras recomendaciones de FLaReNet ya se hacía énfasis en la necesidad de que la comunidad actúe de forma coordinada y como una unidad para conseguir mantener el apoyo a un área tan sensible para Europa: defender el multilingüismo que la caracteriza y desarrollar una industria que lo haga sostenible. Esta recomendación ha tenido ya un resultado en la constitución de la Red de Excelencia META-NET: *Multilingual Technology Alliance Network*, iniciada en 2010 gracias al proyecto T4ME financiado también por el 7PM. Esta red de excelencia ha puesto de manifiesto, ya en sus primeros estudios, la urgente necesidad de disponer de información, documentación y acceso a recursos y tecnologías de todas las lenguas de Europa y ha puesto en marcha META-SHARE, una plataforma para la creación de un catálogo y repositorio digital de recursos lingüísticos.

Para contribuir a poner en marcha la versión operativa de META-SHARE dotándola de contenidos que puedan acelerar su utilización, la Comisión coordinó la colaboración de tres proyectos del programa europeo *Information and Communication Technology Policy Support Programme* (ICT-PSP) y *Competitiveness and Innovation Framework Programme* (CIP): METANET4U[6], CESAR[7] y METANORD[8]. Estos proyectos tienen como misión: (i) extender los objetivos de META-NET difundiendo sus objetivos en el ámbito de los diferentes Estados europeos y (ii) compilar y aportar una masa crítica de recursos lingüísticos para la plataforma META-SHARE.

3 Acceso y disponibilidad de recursos lingüísticos en META-SHARE

En la actualidad hay un gran número de recursos lingüísticos que potencialmente pueden ser la base para diferentes tecnologías y aplicaciones: texto, datos lingüísticos, ficheros multimedia, herramientas, etc. Su descripción, archivo y mantenimiento a largo plazo tienen, no obstante, realizaciones diversas y heterogéneas.

Los repositorios digitales proporcionan ahora la infraestructura necesaria para hacer que la búsqueda y acceso a ellos sea no solamente posible sino también fácil. Estos repositorios

[4] http://www.clarin.eu
[5] http://www.flarenet.eu
[6] http://metanet4u.eu
[7] http://www.meta-net.eu/projects/cesar
[8] http://www.meta-nord.eu/

son una evolución del paradigma de las bibliotecas digitales en la que no solo se catalogan recursos, también se da acceso a los mismos recursos permitiendo la descarga, y se proporcionan nuevas capacidades de búsqueda basadas en descripciones formalizadas o metadatos.

META-SHARE tiene como objetivo ofrecer un repositorio digital distribuido de recursos y servicios lingüísticos que aporte estas nuevas funcionalidades y que sea un paso adelante en relación a otras iniciativas ya existentes de recopilación de información y distribución de recursos (el catálogo de ELRA, *European Language Resources Association*, o el modelo de distribución de ELDA, Evaluation and Language Distribution Agency, por ejemplo). META-SHARE quiere aportar también un marco de interoperabilidad entre recursos y tecnologías y ofrecer una infraestructura abierta que incluya recursos libres o bajo licencia, gratuitos o de pago. Además el proyecto está prestando atención a las cuestiones legales que a menudo comportan problemas para la distribución de estos recursos.

4 METANET4U

El objetivo del proyecto METANET4U es contribuir a la creación de esta plataforma digital pan-europea que sustentará la distribución y el intercambio de recursos y servicios lingüísticos con el objetivo último de apoyar el desarrollo de aplicaciones basadas en tecnologías lingüísticas. Este objetivo central se articula a partir de las siguientes iniciativas:

1) Recoger, organizar y difundir información sobre el estado de las actividades relacionadas con las tecnologías lingüísticas en las lenguas representadas por los participantes en el proyecto. Para ello ha secundado la iniciativa de META-NET de redactar una serie de libros blancos de las lenguas en Europa que identifiquen los beneficios que se pueden esperar de estas tecnologías y el nivel de disponibilidad para las diferentes lenguas europeas. METANET4U ha redactado 3 de estos informes: catalán, gallego y vasco. Antes de final del proyecto se editarán como libros independientes en la editorial Springer.

2) Recopilar recursos lingüísticos ya existentes y disponibles de las lenguas representadas en el proyecto para documentarlos siguiendo los esquemas de metadatos de la plataforma y, en su caso, convertirlos a los formatos que garanticen la interoperabilidad entre recursos y tecnologías. En este primer año, METANET4U ha hecho una primera entrega de 87 recursos que incluyen 27 corpus (monolingües, bilingües, alineados, etc.), 41 léxicos y diccionarios, 15 bases de datos de registros de habla y modelos de lenguaje, y 4 gramáticas y reglas de inferencia. En la mayoría de los casos las versiones de los recursos que se encuentran en los servidores META-SHARE son versiones revisadas, documentadas con metadatos para permitir la búsqueda inteligente, y, en su caso, convertidas a formatos estándar y a UTF8. Al final del proyecto, METANET4U habrá dado acceso a cerca de 300 recursos lingüísticos, entre datos y herramientas. Para conseguirlo, el proyecto tiene la misión de ponerse en contacto con desarrolladores de recursos externos al proyecto brindándoles la posibilidad de dar acceso a sus recursos desde META-SHARE además de aprovechar la posibilidad de que el proyecto lleve a cabo la adaptación a estándares y documentación específica necesarias. Así, en cuanto a recursos españoles, se han hecho versiones LMF de diccionarios tanto bilingües de acceso abierto como los de Apertium[9], como monolingües como el léxico PAROLE-SIMPLE catalán desarrollado por el Institut d'Estudis Catalans[10]; glosarios terminológicos multilingües, como el Banco de datos terminológico[11] da Universidade de Vigo, corpus anotados con información de dependencias, como los corpus AnCora[12] del grupo CLIC o el corpus Grial[13], del grupo GRIAL, ambos de la Universidad de Barcelona. Se ha abierto también una colaboración con la Universidade de Vigo y con la Universidad del País Vasco que, juntamente con el TALP han dado acceso a recursos lingüísticos relacionados con las tecnologías del habla, concretamente, se ha dado acceso a grandes bases de datos orales para entrenar sistemas de reconocimiento del habla en español, catalán, gallego y vasco, grabaciones preparadas para ser utilizadas en sistemas automáticos de conversión texto voz en español, catalán y vasco, tanto monolingües como bilingües, lexicones específicamente diseñados para ser incorporados en sistemas de

[9] http://www.apertium.org
[10] http://www.iec.cat
[11] http://sli.uvigo.es/TUVI/
[12] http://clic.ub.edu/corpus/AnCora
[13] http://grial.uab.es/sensem/download/main,es

reconocimiento automático del habla y en sistemas de conversión de texto a voz en las cuatro lenguas mencionadas, así como textos paralelos alineados y etiquetados para ser utilizados en el entrenamiento de sistemas de traducción automática entre el español y el catalán y el español y el gallego.

También se han incorporado a la plataforma META-SHARE programas para realizar automáticamente transcripciones fonéticas en catalán, español y gallego a partir de textos.

3) Difundir la existencia de estos recursos así como la disponibilidad de tecnologías que pueden hacer del espacio digital europeo un espacio multilingüe sin barreras. El público objetivo son investigadores, desarrolladores de aplicaciones y responsables políticos, sectores críticos para garantizar la diversidad lingüística en Europa.

METANET4U ha supuesto también la puesta en marcha de varios proyectos para mejorar y enriquecer recursos propios de los participantes. En el caso del IULA, se han estandarizado recursos que gracias a los grupos de investigación del IULA ya estaban a disposición de todos los usuarios para consulta *on line* en su página web. En particular, se está enriqueciendo el Corpus Tècnic de l'IULA (Cabré et al. 2006) hasta ahora solamente anotado con información morfosintáctica con información sintáctica y de dependencias. El IULA Treebank (Marimon et al. 2012) contribuirá así a la disponibilidad de más datos analizados sintácticamente del castellano que hasta ahora contaba con el corpus ANCORA (Taulé et al., 2008), y el UAM Spanish Treebank (Moreno and López, 1999).

Por su parte, el TALP, ha iniciado un proyecto para ampliar y mejorar recursos lingüísticos multimodales como son los seminarios interactivos CHIL, inicialmente consistentes en seminarios de una hora de duración grabados en inglés y transcritos ortográficamente. Se ha procedido a grabar seminarios en catalán y castellano y a extender la transcripción también a la parte de video incorporando información de movimientos, gestos, posiciones, y estados anímicos tanto del ponente como de los asistentes al seminario. Por otra parte, se ha realizado una mejora de los datos disponibles para realizar conversión texto a voz. Concretamente se ha realizado una compatibilización de los sistemas y anotaciones existentes en bases de datos en catalán y español (Bonafonte et al. 2008) para que puedan ser fácilmente integrados en el sistema de código abierto Festival.

5 Conclusiones

En este artículo se han presentado los proyectos que la Comisión Europea está cofinanciando para fomentar la participación del mayor número posible de grupos de investigación en la definición de estrategias para desarrollar tecnologías y recursos lingüísticos en Europa.

Se ha presentado también la iniciativa META-SHARE para crear una plataforma que facilite la disponibilidad pública de recursos y servicios lingüísticos así como el proyecto METANET4U gracias al cual se han puesto a disposición de los usuarios recursos relacionados con las lenguas oficiales habladas en España.

Bibliografía

Calzolari, N.; Quochi, V. y Soria, C. 2012, The Strategic Language Resource Agenda. En http://www.flarenet.eu/sites/default/files/FLaReNet_Strategic_Language_Resource_Agenda.pdf.

Marimon, M., Fisas, B., Bel, N., Arias, B., Vázquez, S., Vivaldi, J., Torner, S., Villegas, M. y Lorente, M. 2012. The IULA Treebank. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC2012)*, Estambul.

Moreno, A. and S. López. 1999. Developing a Spanish Tree Bank. In *Proc. Journées ATALA, Corpus annotés pour la syntaxe*. Paris, 18-19 June 1999.

Taulé, M.; M.A. Martí and M. Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. *Proceedings of 6th International Conference on Language Resources and Evaluation (LREC2008)*. Marrakesh.

Bonafonte, A.; J. Adell, I. Esquerra, S. Gallego, A. Moreno, J. Pérez . 2008. Corpus and Voices for Catalan Speech Synthesis. *Proc. of the Sixth Int. Conf. on Language Resources and Evaluation LREC 2008* Marrakech Marruecos