Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 151-154 recibido 24-03-2016 revisado 06-04-2016 aceptado 25-04-2016

CLARIN Centro-K-español

Spanish CLARIN K-Centre

Núria Bel
Universidad Pompeu Fabra
Roc Boronat, 138
08018 Barcelona
nuria.bel@upf.edu

Elena González-Blanco García
Universidad Nacional de
Educación a Distancia
Paseo Senda del Rey 7
28040 Madrid
egonzalezblanco@flog.uned.es

Mikel Iruskieta
Universidad del País Vasco
(UPV/EHU)
Sarriena auzoa z/g
48940 Leioa
mikel.iruskieta@ehu.eus

Resumen: Presentamos CLARIN Centro-K-español que forma parte de la infraestructura europea CLARIN, *Common Language Resources and Technology Infrastructure*, y cuyo objetivo es ofrecer los conocimientos y experiencia de los tres grupos que inicialmente lo componen en la utilización de tecnología para la investigación en humanidades y ciencias sociales.
Palabras clave: infraestructura lingüística, análisis de textos, humanidades, ciencias sociales.

Abstract: We introduce Spanish CLARIN Centre-K, a node of the European infrastructure CLARIN, Common Language Resources and Technology, whose objective is to share knowledge and experience of the three funding constituent groups for research in humanities and social sciences.
Keywords: Language infrastructure, text analytics, humanities, social sciences.

1 Introducción

CLARIN[1], la e-infraestructura europea de investigación para humanidades y ciencias sociales se ha ido desplegando desde hace 10 años como una red de centros que comparten misión, tecnología y recursos para ponerse a disposición de los investigadores que trabajan en el procesamiento y explotación de textos (escritos u orales) en áreas de humanidades y ciencias sociales. El objetivo es garantizar el acceso, integración y explotación de la gran cantidad de datos lingüísticos (o relacionados con las lenguas) y la tecnología relacionada. Actualmente CLARIN da servicios a investigadores en psicología, lingüística, filología, ciencias políticas, o sociología entre otros.

CLARIN fue uno de los proyectos seleccionados por el Comité ESFRI (*European Strategy Forum on Research Infraestructures*) y que figuran en la primera "Hoja de ruta"[2] de las infraestructuras que habían de ser construidas, por su importancia para la investigación, a diez años vista. Ya se han cumplido los diez años y su despliegue como infraestructura europea demuestra lo acertado de aquella decisión.

La Comisión de la Unión Europea dispuso la cofinanciación de una fase preparatoria para estos proyectos dentro del área Infraestructuras del VII Programa Marco. En España, el Ministerio de Educación y Ciencia, Dirección General de Política Tecnológica, Subdirección General de Promoción e Infraestructuras Tecnológicas y Grandes Instalaciones (CAC-2007-23) primero, y después el Ministerio de Ciencia e Innovación, Dirección General de Planificación y Coordinación (ICTS-2008-11) asumieron la cofinanciación de dicha fase preparatoria para la participación española. Por otra parte, el Departament d'Innovació, Universitats i Empresa de la Generalitat de Catalunya ha participado en la financiación del proyecto CLARIN-CAT, apoyando la presencia de datos y herramientas específicamente en y para la lengua catalana. El Centro de Competencias CLARIN-IULA-UPF se cofinanció mediante el programa FEDER Catalunya 2007-2013.

[1] www.clarin.eu
[2] http://cordis.europa.eu/esfri/roadmap.htm

ISSN 1135-5948

CLARIN está constituido formalmente como un consorcio europeo[3] de países y cuenta en la actualidad con diecisiete asociados y un país observador, Los asociados pueden incorporar centros de investigación específicos como miembros de la red. La red de centros CLARIN es actualmente de 137, entre los que se incluyen instituciones como universidades, centros de investigación, academias nacionales de ciencia y academias nacionales de la lengua.

Además, el consorcio CLARIN invita a incorporarse, como centro de conocimiento o *Centre-K*, a instituciones, de estados asociados o no –que es el caso de España–, que dispongan de servicios estables y que quieran ofrecer sus conocimientos y experiencia en el uso de infraestructuras lingüísticas en la investigación en una lengua o tema particular.

En 2015, CLARIN concedió el reconocimiento de Centre-K al centro distribuido especializado para España y tres de sus lenguas oficiales (castellano, catalán y euskera), el CLARIN Centro-K-español. Constituido como una asociación de centros ya existentes, en el CLARIN Centro-K-español se reúnen las especialidades del Centro de Competencias CLARIN IULA de la Universidad Pompeu Fabra (UPF), el Laboratorio de Innovación en Humanidades Digitales de la Universidad Nacional de Educación a Distancia (UNED) y del Grupo IXA de la Universidad del País Vasco/Euskal Herriko Unibertsitea (UPV/EHU).

2 *CLARIN, e-Infraestructura de Servicios*

CLARIN está concebida como una e-Infraestructura de tecnología lingüística que ofrece, entre otras cosas, un observatorio virtual de tecnología lingüística común para los países participantes, varios repositorios que almacenan, recogen y preservan proyectos y datos, clasificados con metadatos que conllevan índices de calidad (Thompson), y sistemas de identificación permanente, además de *login* unificado para las diferentes instituciones que participan. Todo esto, acompañado de un potente grupo de investigadores con una intensa actividad académica manifestada a través de congresos, seminarios, talleres y proyectos específicos en los diferentes grupos de interés, que han generado a su vez la creación e inclusión en otros proyectos europeos, como Europeana-DSI[4], LT-Observatory[5], EUDAT2020[6] o Parthenos[7].

3 *Descripción de los servicios del CLARIN Centro-K-español*

Los participantes en el Centro-K-español[8] tienen en común el objetivo de fomentar y asistir en el uso de la tecnología en la investigación en Humanidades y Ciencias Sociales. La experiencia y competencias de los tres grupos son complementarios, abarcando así buena parte del espectro de conocimientos necesarios para asesorar a los investigadores de estas áreas. IXA y IULA-UPF-CCC están especializados en *Text Analytics* y Tecnologías y Recursos del Lenguaje. LINHD está especializado en los beneficios del enriquecimiento de textos, Web Semántica, bases de datos y tecnologías de visualización para las Humanidades Digitales. Los tres ofrecen servicios a investigadores que trabajan con textos en español e inglés y, además, IXA aporta competencias en el manejo de textos en euskera y IULA-UPF-CCC de textos en catalán: disponen de recursos y herramientas para el análisis y la anotación automática de los textos en estas lenguas.

Así, el CLARIN Centro-K-español consta de los servicios que aportan sus grupos participantes y que en el Centro-K se asocian para constituirse en un punto de acceso único y actuar así de agente de distribución de solicitudes de servicios. Como Centro-K, suman las fortalezas de los tres centros participantes con el fin de ofrecer servicios a la comunidad de forma unificada y organizada.

Los servicios que el CLARIN Centro-K-español ofrece son:

– Consultoría virtual para asesorar y responder dudas sobre cuestiones prácticas relacionadas con estándares, uso de herramientas de procesamiento y acceso a recursos lingüísticos. El centro ofrece contacto por correo electrónico con garantía de respuesta en 24 horas. Se trata de un servicio de orientación personalizada y que también pone al servicio de los investigadores

[3] *European Research Infrastructure Consortium*.

[4] http://pro.europeana.eu/

[5] http://www.lt-observatory.eu/

[6] http://eudat.eu/

[7] http://www.parthenos-project.eu/

[8] http://www.clarin-es.org

documentación y casos de uso conocidos en los que encontrar buenas prácticas.
- Soporte para auto-aprendizaje con recursos especializados: catálogos especializados donde encontrar información sobre prácticas actuales y acceso directo a herramientas, en aplicaciones web, que facilitan el uso de las tecnologías existentes, videotutoriales, MOOCs, etc.
- Organización de programas de enseñanza y formación para investigadores, estudiantes, proyectos o grupos de interés.
- Servicios tecnológicos y de gestión y planificación de proyectos personalizados en función de las necesidades de los solicitantes.

4 Descripción de los grupos del Centro-K-español

El Centro de Competencias CLARIN IULA-UPF[9] está especializado en Análisis y Minería de Textos y en Tecnologías y Recursos del Lenguaje. Dirigido por el grupo Tecnología de los Recursos Lingüísticos, ha liderado la participación de España en CLARIN como proveedor de servicios web de procesamiento del lenguaje natural. Ha desarrollado aplicaciones web como ContaWords[10] con la que el usuario puede obtener información cuantitativa de textos (aportados por el usuario o de la web): frecuencia de palabras, por categorías morfosintácticas, reconocimiento y clasificación de entidades nombradas. Los resultados se obtienen en un formato fácilmente manipulable por el usuario. También ofrece acceso web a analizadores de dependencias sintácticas para el español[11], y a otras herramientas básicas de análisis populares como FreeLing (Padró y Stanilovsky, 2012) y MaltParser (Nivre et al., 2004). El centro de competencias, también cuenta con la experiencia del HDLab@UPF, un nuevo entorno de investigación y aprendizaje desarrollado por el Departamento de Humanidades de la UPF.

El Laboratorio de Innovación en Humanidades Digitales[12] (LINHD) fue fundado en 2014 gracias a la financiación inicial de la UNED. Se trata de un centro interdisciplinar de investigación en el que participan todas las facultades de humanidades y ciencias sociales, además de la ETSI de Ingeniería Informática de la UNED, la sección digital de la Biblioteca, el CEMAV e Intecca (medios audiovisuales de la UNED). El centro agrupa hoy día más de una veintena de proyectos, es un centro de referencia en formación en humanidades digitales en español, pues ofrece tres títulos propios, una escuela de verano con más de un centenar de alumnos y diferentes actividades colaborativas con otros centros e instituciones a nivel nacional e internacional (entre los que destaca el Secrit-Conicet, en Argentina).

El LINDH está especializado en el enriquecimiento de textos, la web semántica, las bases de datos y la edición digital. Se ocupa de velar por los estándares propios de las humanidades digitales, como el TEI-XML para el marcado de textos, además de ofrecer un acercamiento al mundo hispanohablante de los proyectos, recursos y actividades de otros países –principalmente anglófonos–. En este sentido, destacan las actividades colaborativas y de traducción del entorno virtual de edición Textgrid de Dariah-DE[13], del vocabulario semántico de TADIRAH[14], del índice de herramientas Dirt Directory[15] y de la gestión de los dos últimos eventos de blogging DayofDH 2015 y 2016. En investigación es, además un centro puntero, pues ha recibido recientemente un proyecto ERC y está trabajando en la creación de un entorno virtual de investigación para humanistas que se pondrá en breve a disposición del centro-K.

El Grupo IXA es un grupo de investigación multidisciplinario de la UPV/EHU que incluye miembros de cinco departamentos: Lenguajes y Sistemas Informáticos, Arquitectura y Tecnología de Computadores, Ciencia de la Computación e Inteligencia Artificial, Lengua Vasca y Comunicación (Filología Vasca), y Didáctica de la Lengua y la Literatura. El grupo IXA trabaja con el procesamiento del lenguaje natural desde el año 1988. Está especializado en el Análisis y Minería de Textos y en las Tecnologías y Recursos del Lenguaje. La lengua de estudio principal es el euskera, aunque también se han desarrollado productos para otras lenguas, como el inglés y el

[9] http://clarin-es-lab.org
[10] http://contawords.iula.upf.edu
[11] http://lod.iula.upf.edu/resources/278
[12] http://linhd.uned.es
[13] https://de.dariah.eu/
[14] http://tadirah.dariah.eu/vocab/index.php
[15] http://dirtdirectory.org/tadirah

castellano. Sus productos más reconocidos son el corrector ortográfico Xuxen, el traductor automático Matxin integrado en la plataforma Opentrad, Basque WordNet, el corpus de Ciencia y Tecnología (ZT), y el Corpus de Referencia para el Procesamiento del Euskera (EPEC). Asimismo, ha desarrollado más de 20 herramientas[16], algunas de estas herramientas han sido reutilizadas y rediseñadas para el Centro-K. Estas nuevas herramientas son útiles para un uso masivo e intuitivo, como por ejemplo, el programa ANALITZAK[17] basado en los IXA-PIPES (Agerri et al., 2014), con el que se obtienen frecuencias de morfemas, palabras por categorías o entidades de textos y webs tanto en euskera, como en inglés y en castellano.

Para la consulta de información lingüística de diferentes niveles se ofrece, además, acceso web a bases de datos y herramientas de uso sencillo: i) EDBL[18], base de datos lexical de euskera, en la que se detalla la información morfosintáctica de palabras y morfemas, utilizada en el corrector Xuxen. ii) Konbitzul[19], para consultar combinaciones (y su información morfosintáctica) de nombres y verbos útiles para la traducción de euskera a castellano o viceversa, útil para que el traductor automático Matxin traduzca adecuadamente dichas unidades fraseológicas. iii) e-ROLda[20], donde se puede realizar consultas para obtener información sintáctico-semántico de verbos (e incluso de nombres) en corpus anotado. iv) EusEduSeg[21], el segmentador discursivo automático para el euskera y diferentes bases de datos[22] con información discursiva basados en el trabajo desarrollado por Iruskieta et al. (2013), donde se pueden realizar consultas sobre relaciones de coherencia en lenguas tan dispares como el euskera, inglés, castellano y portugués (y muy pronto también el chino), útil para tareas de análisis de sentimiento, traducción y análisis de textos científicos, políticos y de crítica literaria.

[16] https://ixa.si.ehu.es/Ixa/Produktuak
[17] http://ixa2.si.ehu.es/clarink
[18] http://ixa2.si.ehu.es/edbl/
[19] http://ixa2.si.ehu.es/konbitzul/
[20] http://ixa2.si.ehu.es/e-rolda/en/
[21] http://ixa2.si.ehu.es/EusEduSeg/EusEduSeg.pl
[22] http://ixa2.si.ehu.es/rst/

5 Conclusión y trabajo futuro

Las herramientas brevemente descritas en este trabajo han sido desarrolladas para que la investigación en humanidades y ciencias sociales pueda tener una base lingüística fiable y de fácil uso, como puede ser el ejemplo de Villegas et al. (2012). Evidentemente las herramientas aquí descritas no satisfacen las necesidades existentes hoy en día, pero el proyecto CLARIN Centro-K-español ofrece la posibilidad de explorar la utilización de herramientas de interés general contando con servicios de consultoría o soporte para el aprendizaje, así como para diseñar y desarrollar mejores herramientas en colaboración con los grupos de investigación, empresas y agentes educativos que estén interesados en este proyecto y que pueden unirse al centro.

Referencias

Agerri, R., J. Bermudez, and G. Rigau. 2014. IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In Proceedings of the 9th Language Resources and Evaluation Conference, LREC2014.

Iruskieta, M., M.J. Aranzabe, A. Diaz de Ilarraza, I. Gonzalez, M. Lersundi, O. Lopez de Lacalle. 2013. The RST Basque TreeBank: an online search interface to check rhetorical relations. In the 4th Workshop RST and Discourse Studies, pp. 40-49, October, Fortaleza, Brasil.

Nivre, J., J. Hall, and J. Nilsson. 2004. Memory-Based Dependency Parsing. In Ng, H. T. and Riloff, E. (eds.) *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL)*, pp. 49-56

Padró, L., and E. Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In Proceedings of the 8th Language Resources and Evaluation Conference, LREC2012.

Villegas M., N. Bel, C. Gonzalo, A. Moreno, and N. Simelio. 2012. Using Language Resources in Humanities research. In Proceedings of the 8th Language Resources and Evaluation Conference, LREC2012.