Primera Jornada del Procesamiento Computacional del Catalán

First Workshop on the Computational Processing of Catalan

Gemma Boleda[(1)], **Montse Cuadros**[(1)], **Cristina España-Bonet**[(1)],
Maite Melero[(2)], **Lluís Padró**[(1)], **Martí Quixal**[(2)], **Carlos Rodríguez**[(2)]

[(1)] Centre de Recerca TALP
Dept. de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya

{gboleda,cuadros,cristinae,padro}@lsi.upc.edu

[(2)] Fundació Barcelona Media - Universitat Pompeu Fabra

{maite.melero,marti.quixal,carlos.rodriguez}@barcelonamedia.org

Resumen: Presentamos las conclusiones de la primera *Jornada del Processament Computacional del Català*, celebrado en Barcelona en marzo del 2009.
Palabras clave: PLN, catalán, jornada, código libre, política científica.

Abstract: We present the conclusions of the first *Jornada del Processament Computacional del Català*, held in Barcelona on March 2009.
Keywords: NLP, Catalan, workshop, open source, scientific politics.

1 Introducción

A partir de la constatación de que la comunidad de investigación de Procesamiento del Lenguaje Natural y del Habla en catalán precisaba mayor cohesión, se organizó una jornada (*Jornada del Processament Computacional del Català*, JPCC) que se celebró en el Palau Robert de Barcelona en marzo de 2009.[1]

Los objetivos de la jornada eran (1) mejorar la comunicación y la colaboración entre los diferentes grupos de investigación, empresas e instituciones que desarrollan herramientas y recursos computacionales para el catalán, (2) encontrar maneras de aprovechar de forma eficiente los recursos existentes y, (3) dar visibilidad a la investigación en el tratamiento computacional del catalán.

2 Participación

La primera JPCC reunió a 14 grupos o centros de investigación de Cataluña y de la Comunidad Valenciana. Asimismo, se presentaron 27 herramientas, recursos y aplicaciones para el catalán. Se inscribieron 166 personas, la mayoría de centros de investigación, pero también de otros agentes sociales, como son la empresa privada y la administración pública (38 inscritos). Creemos que el interés despertado por la Jornada, así como la cantidad y calidad de grupos, herramientas, recursos y ámbitos de aplicación que se presentaron, son una prueba de la vitalidad tecnológica del catalán.

3 Conclusiones y recomendaciones

Durante la Jornada tuvo lugar un debate moderado por Vicent Partal (Vilaweb) cuyos ponentes fueron Toni Hermoso (Softcatalà), M. Antònia Martí (U. Barcelona, Thera), Mikel Forcada (U. Alicante, Transducens), y Joan Soler (Institut d'Estudis Catalans). El presente resumen se restringe a dos de las cuestiones debatidas, que son relevantes para otras lenguas del estado español.

¿Qué se podría hacer para ampliar la presencia de la lengua catalana en los servicios y productos tecnológicos? ¿Y para aumentar la transferencia de la tecnología que se desarrolla a los centros de investigación?

Los ponentes estuvieron de acuerdo en que la salud computacional del catalán es buena; de hecho, los recursos y herramientas disponibles superan en cantidad y calidad a los que le corresponderían por demografía o poder político, y su situación de casi normalidad en este sentido es excepcional para una lengua sin estado. No obstante, se identificaron una serie

[1] http://sites.google.com/site/jornadacatala.

de problemas que dificultan la difusión de las herramientas computacionales del catalán tanto para objetivos de investigación como para aplicaciones de usuario final, y que consignamos a continuación.

Se apuntó como problema **la baja demanda social de recursos tecnológicos del catalán** por parte del mundo de la empresa y los servicios públicos (se mencionaron por ejemplo sectores como la banca o la sanidad) en que el catalán no es lengua vehicular real.

También se apuntó **la desconexión entre el desarrollador y el usuario final**. En este sentido, algunas intervenciones apuntaron que las aplicaciones distribuidas con licencias de código abierto (p.e. Apertium[2]) facilitan esta conexión gracias a las comunidades que se estructuran alrededor de este tipo de código. Estas comunidades pueden aglutinar grupos de investigación, usuarios y empresas orientadas a servicios.

Un punto crucial es también **la dificultad en el acceso a ciertas tecnologías y recursos por problemas de licencia**. Aquí, de nuevo se subrayó la importancia de que aumente el número de recursos de libre distribución. En el supuesto de que un recurso no se pueda distribuir libremente por problemas de propiedad o *copyright*, como un corpus con textos que sean propiedad de autores o editoriales, habría que facilitar su uso para objetivos de investigación. En este caso, es necesario que sus condiciones de uso estén bien especificadas y que el acceso a los recursos sea ágil. Actualmente se da todavía una situación muy compleja, con mucha diversidad de licencias y condiciones de uso.

Especialmente nocivas son **la duplicación y la dispersión de esfuerzos** en la construcción de corpus, herramientas, etc. El catalán es una lengua relativamente pequeña, los recursos lingüísticos son costosos y la financiación escasa. La impresión general es que, como comunidad, no nos podemos permitir derrochar esfuerzos de este modo. Hace falta pues mejorar la coordinación y fomentar la cooperación entre grupos de investigación y también con las empresas. Consideraciones similares podrían aplicarse al caso del euskera, del gallego o del español.

Se mencionó asimismo la relativamente **baja interacción entre lingüistas e informáticos** al crear y compartir aplicaciones y recursos. Es necesaria una mayor interdisciplinariedad, y ello se ve dificultado por las rígidas estructuras en la organización docente de la universidad, así como por la falta de centros de investigación híbridos.

Finalmente, se apuntó que debía favorecerse la presencia de recursos textuales para el catalán, por ejemplo corpus anotados, en competiciones internacionales para tareas computacionales. De nuevo, lo mismo podría aplicarse a las otras lenguas del estado.

¿Qué estrategias de financiación son necesarias? ¿Qué relación debe existir entre el origen de la financiación y las condiciones de accesibilidad y distribución de los recursos?

Desde la mesa de ponentes se apuntó que el empresario catalán tiende más hacia la seguridad que hacia el riesgo y la innovación. Por lo tanto, son necesarias las políticas de incentivación a las empresas. Las políticas actuales, sin embargo, están orientadas a proyectos demasiado grandes sólo asumibles por grandes empresas, como Microsoft. **Las acciones de la administración podrían ser más efectivas si se desarrollaran políticas orientadas a PYMES.**

Asimismo, para favorecer el código abierto, se propuso que en las convocatorias de proyectos competitivos los grupos tengan que **especificar las condiciones de uso de los recursos que se generarán**, para que los evaluadores lo puedan tener en cuenta a la hora de conceder el proyecto y de controlar su progreso. Desde el público se sugirió que esta demanda se podría hacer a nivel estatal a fin de que se incorpore a la nueva Ley de la Ciencia.

Un problema respecto a la libre distribución de los recursos generados por los proyectos de investigación es que hay que contemplar los **costos del empaquetado final** de la aplicación. Estos costos, así como la preparación de licencias y el mecanismo de distribución, se tendrían que prever en la definición inicial de los proyectos.

4 Patrocinadores

La JPCC fue patrocinada por la *Secretaria de Política Lingüística* y la *Secretaria de Telecomunicacions i Societat de la Informació* de la *Generalitat de Catalunya*.

[2] http://xixona.dlsi.ua.es/apertium