

Producción eficiente de recursos lingüísticos: el proyecto Victoria*

Efficient production of linguistic resources: the Victoria Project

Lionel Nicolas^{*}, Miguel A. Molinero[◇], Benoît Sagot[⊕],
Elena Sánchez Trigo[•], Eric de La Clergerie[⊕], Miguel Alonso Pardo[◇],
Jacques Farré^{*}, Joan Miquel Verges[•].

| | |
|---|---|
| Équipe RL, Laboratoire I3S, UNSA+CNRS, France | {lnicolas,jf}@i3s.unice.fr |
| ◇ Grupo LYS, Univ. de A Coruña, España | {mmolinero,alonso}@udc.es |
| ⊕ Projet ALPAGE, INRIA Rocquencourt + Paris 7, France | {benoit.sagot, Eric.De.La.Clergerie}@inria.fr |
| • Grupo Cole, Univ. de Vigo, España | {etrigo,jmv}@uvigo.es |

Resumen: La eficiencia de las herramientas dedicadas al **Procesamiento** de los **Lenguajes Naturales (PLN)** depende directamente de la calidad y la cobertura de los recursos lingüísticos sobre los cuales se basan. Presentamos un proyecto cuyo objetivo es mejorar las capacidades de producción de recursos lingüísticos.

Palabras clave: Palabras, palabras, palabras en castellano...

Abstract: In order to produce efficient **Natural Language Processing (NLP)** tools, reliable linguistic resources are a preliminary requirement. This paper presents a project whose ambition is to enhance the production capacities of linguistic resources.

Keywords: Palabras, palabras, palabras en inglés...

1. *Orígenes y objetivos*

La eficiencia de las herramientas PLN depende directamente de la calidad y la cobertura de los recursos lingüísticos sobre los cuales se basan. Sin embargo, cuando existen para una lengua concreta, estos recursos no suelen cumplir las expectativas en términos de calidad, cobertura o usabilidad que dichas aplicaciones requieren. Para idiomas de primer orden como el Español o el Francés, muchos de los recursos más usados están todavía en un estado de desarrollo precario. Para idiomas como el Gallego, son casi inexistentes.

Esta carencia resulta, principalmente, del alto coste que implican los desarrollos manuales: la complejidad y/o el tamaño de los recursos requieren el uso masivo de costosa mano de obra experta. Obviamente, estos esfuerzos manuales se podrían dividir entre las personas interesadas en obtener esos recursos. Pero las colaboraciones de largo plazo pueden ser complejas por razones de licencias, gestión, distancia, tiempo y/o dinero.

Dadas estas consideraciones, la construcción de recursos es un proceso largo, más o menos aislado, que no suele alcanzar resul-

tados visibles o usables. Por lo tanto, la adquisición y corrección de recursos lingüísticos sigue siendo un problema no resuelto y una tremenda necesidad para el dominio del PLN.

Con la ambición de enfrentarse a dicha necesidad, el proyecto Victoria se propone (I) desarrollar una cadena secuencial de herramientas semi-automáticas de adquisición y corrección de recursos lingüísticos, (II) explorar métodos de transferencia de conocimiento lingüístico entre recursos que describen lenguajes relacionados y (III) desarrollar una plataforma colaborativa de desarrollo de recursos lingüísticos.

En su primera fase, el proyecto se concentra en los recursos necesarios para construir analizadores sintácticos¹ para Francés, Español y Gallego.

Habiendo comenzado oficialmente en Noviembre de 2008, el proyecto está actualmente compuesto de investigadores del campo de la Informática y de la Traducción pertenecientes a dos equipos españoles (Grupo Cole y Lys) y dos franceses (Projet Alpage y Equipe RL).

* Agradecimientos

¹Reglas morfológicas, léxicos y gramáticas.

2. *Proyectos relacionados*

Numerosos proyectos se han centrado en el desarrollo de recursos lexicales. Entre ellos, MULTEXT² y su continuación MULTEXT-East³, GENELEX, EAGLES y PAROLE. A nivel sintáctico, los proyectos DELPHIN⁴ y AGFL⁵ han permitido producir varias gramáticas. Por otro lado, los proyectos CLARIN⁶ y FLARENET⁷ intentan agrupar bajo una plataforma común los recursos existentes.

Sin embargo, los recursos producidos han sido desarrollados casi siempre manualmente, con poca o ninguna ayuda de técnicas automatizadas. Esto suele llevar los recursos hasta un cierto estado donde encontrar manualmente sus carencias resulta difícil.

3. *Directrices*

Tras estudiar otros proyectos hemos establecido seis directrices de las cuales depende el éxito del proyecto.

(1) Disponer de herramientas que automaticen los procesos de extensión y de corrección de recursos es una necesidad **absoluta** para garantizar sus desarrollos y perfeccionamientos.

(2) El uso regular y “extendido” de un recurso es una fuente importante de información para su desarrollo. Por lo tanto, debe estar distribuidos libremente sin restricciones legales o económicas.

(3) Cualquier recurso existente puede ser considerado para ampliar otros, incluso aquellos que describen otros idiomas lingüísticamente próximos, tal como las Lenguas Romance.

(4) La edición colaborativa de un recurso puede ser limitada por varias razones : de distancia, si es imposible desde un lugar concreto, tecnológicas, si está limitada a un sistema de explotación particular, y de capacidad, si requiere el conocimiento de un formalismo y/o de reglas particulares. Aun así la edición colaborativa es necesaria para aumentar las posibilidades de desarrollo continuo de un recurso.

(5) Los formalismos usados para describir los recursos deben ser lo bastante flexi-

bles como para permitir el uso combinado de recursos y la descripción de idiomas variados.

(6) Las herramientas de corrección y extensión deben poder usarse de forma regular y continua para cualquier idioma, por lo que los datos de entrada en los que se basen deben estar disponibles en grandes cantidades.

4. *Estado actual*

Este proyecto toma como base las herramientas, la teoría y los formalismos realizados en el proyecto Alpage⁸.

Para generar recursos, nuestra estrategia se basa en la reutilización de recursos existentes seguida de la aplicación de técnicas que automaticen su corrección y extensión. En este sentido, si dos recursos son usados conjuntamente en una misma herramienta, es posible utilizar uno de ellos para detectar y corregir errores del otro. Este concepto, basado en la asunción de que los fallos de dos recursos no suelen ocurrir al mismo tiempo, nos ha permitido desarrollar dos técnicas semi-automáticas para corregir y extender un léxico y nos abre también posibilidades para otros tipos de recursos. Además, dichas técnicas toman como entrada texto plano digital, que es diariamente producido en grandes cantidades para un gran número de idiomas.

En este momento ya hemos construido varios recursos lingüísticos que próximamente serán liberados bajo licencia LGPL-LR⁹. Entre ellos disponemos, para el Español, de reglas morfológicas, un léxico de gran cobertura con información morfológica y sintáctica y una meta-gramática. Para el Gallego, contamos con reglas morfológicas y un léxico con información morfológica.

Para permitir la edición colaborativa de recursos, desarrollamos interfaces basadas en tecnologías web, ya que estas permiten el acceso desde cualquier plataforma o sistema. Una interfaz detallada para la gestión de léxicos ya está disponible en el recién creado sitio Internet del proyecto.¹⁰

²aune.lpl.univ-aix.fr/projects/MULTEXT/

³nl.ijs.si/ME/

⁴wiki.delph-in.net/moin/PageD\%27Accueil

⁵www.agfl.cs.ru.nl/

⁶www.clarin.eu, Abril 2009

⁷www.flarenet.eu, Abril 2009

⁸alpage.inria.fr, Abril 2009.

⁹Lesser General Public License for Linguistic Resources

¹⁰www.victoria-project.org/