OpenTrad: Traducción automática de código abierto para las lenguas del Estado español

Iñaki Alegría Loinaz[1], Iñaki Arantzabal[2], Mikel L. Forcada[3], Xavier Gómez Guinovart[4], Lluis Padró[5], José Ramom Pichel Campos[6], Josu Waliño[7]

[1]Euskal Herriko Unibersitatea (IXA Taldea), acpalloi@si.ehu.es
[2]Eleka Ingeniaritza Linguistikoa, komertziala@eleka.net
[3]Universitat d'Alacant (Transducens), mlf@ua.es
[4]Universidade de Vigo (Seminario de Lingüística Informática), sli@uvigo.es
[5]Universitat Politècnica de Catalunya (TALP Research Center), padro@lsi.upc.edu
[6]imaxin|software, jramompichel@imaxin.com
[7]Elhuyar Fundazioa, josu@elhuyar.com

Resumen: OpenTrad es un sistema de traducción automática basado en transferencia, de código abierto, y en funcionamiento para español, gallego, catalán/valenciano y euskara. En la página web www.opentrad.org es posible comprobar su funcionamento traduciendo texto, documentos o páginas web. Programas y datos se distribuyen a través de SourceForge.
Palabras clave: traducción automática, código abierto, recursos lingüísticos

Abstract: OpenTrad is an operating open-source and transfer-based machine translation system for Spanish, Galician, Catalan and Basque. It can be accessed in the URL www.opentrad.org for translating text, documents or web pages. Programs and data can be downloaded from SourceForge.
Keywords: machine translation, open source, linguistic resources

1 Datos del proyecto

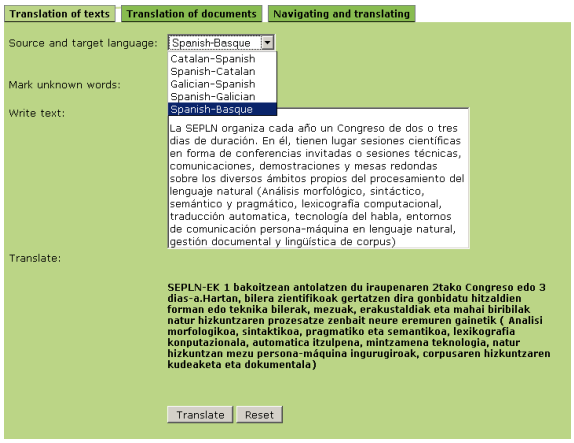
Título: OpenTrad: Traducción automática de código abierto para las lenguas del Estado español.
Pagina web y demo: www.opentrad.org
Descarga del sistema: apertium.sourceforge.net y matxin.sourceforge.net.
Institución financiadora: Ministerio de Industria, Turismo y Comercio, Plan Nacional de I+D+I, Programa de Fomento de la Investigación Técnica, 2004-2005 (refs. FIT-340101-2004-3 y FIT-340001-2005-2).
Universidades participantes: Euskal Herriko Unibersitatea (IXA Taldea), Universitat d'Alacant (Transducens), Universidade de Vigo (Seminario de Lingüística Informática), Universitat Politècnica de Catalunya (TALP).
Empresas participantes: Eleka Ingeniaritza Linguistikoa, imaxin|software, Elhuyar Fundazioa.
Coordinación del proyecto: Eleka Ingeniaritza Linguistikoa. Iñaki Arantzabal. Zelai Haundi kalea, 3. Osinalde Industrialdea. 20170 Usurbil. Tlf: (+34) 943377225. Fax: (+34) 943365923. info@eleka.net.

2 Descripción del proyecto

En este proyecto se ha logrado crear un sistema de traducción automática de código abierto para las principales lenguas oficiales del Estado español. Los pares de lenguas que cuentan ya con un prototipo operativo son:

- español (es) → catalán/valenciano (ca)
- catalán/valenciano → español
- español → gallego (gl)
- gallego → español
- español → euskara (eu)

El sistema se basa en dos motores de traducción de alta velocidad programados en C++: uno de transferencia sintáctica parcial (denominado *Apertium*) para pares de lenguas próximas (es-ca, es-gl) (Corbí et al. 2005), y otro de transferencia sintáctica completa (denominado *Matxin*) para pares de lenguas divergentes (es-eu) (Alegria et al. 2005).



Los datos lingüísticos monolingües y multilingües para todas las lenguas integrantes del sistema se almacenan en XML, con un diseño modular que facilita la interoperabilidad de los datos y su integración en el flujo del programa (Armentano et al. 2005).

La arquitectura del sistema de traducción es totalmente modular y básicamente la misma para todos los pares de lenguas:

1. **Desformateado**: separación de texto y marcas de formato (HTML, RTF, etc.); el formato se restituye al final.
2. **Análisis morfológico,** incluyendo la identificación de las locuciones.
3. **Desambiguación** de los homógrafos.
4. **Análisis sintáctico**: sólo para el par es-eu y basado en FreeLing (Atserias et al. 2005, Atserias et al. 2006).
5. **Transferencia estructural**: transformación de estructuras (superficiales o parciales para los pares es-ca y es-gl; profundas o completas para el par es-eu)
6. **Transferencia léxica**: traducción de las palabras (de los lemas).
7. **Generación sintáctica** (sólo para es-eu).
8. **Generación morfológica** de las formas flexionadas de las palabras del texto meta.
9. **Posgeneración**: transformaciones ortográficas, contracciones, apostrofado, etc.
10. **Reformateado**: restitución del formato (HTML, RTF, etc.) del texto original.

Toda la información relativa al desarrollo de OpenTrad se puede consultar en la web del proyecto (www.opentrad.org), donde también es posible comprobar el funcionamento del sistema de traducción, para los pares de lenguas anteriormente mencionados, a través de la traducción automática de texto, de documentos (HTML, RTF y TXT) o de páginas web.

Los motores de traducción y los datos lingüísticos necesarios para el funcionamiento del sistema se distribuyen como código abierto a través de SourceForge (sourceforge.net), bien con licencia GPL, bien con licencia Creative Commons 2.5. Los módulos correspondientes a es-ca, ca-es, es-gl y gl-es están disponibles en apertium.sourceforge.net; y los del par es-eu, en matxin.sourceforge.net.

Cualquier persona, empresa o institución puede disponer de este sistema para mejorarlo o adaptarlo a entornos específicos de aplicación, con sus propios recursos o en colaboración con los participantes de este proyecto.

Bibliografía

Alegria I., A. Díaz de Ilarraza, G. Labaka, M.Lersundi, A. Mayor, K. Sarasola, M. Forcada, S. Ortiz-Rojas y L. Padró. 2005. An open architecture for transfer-based machine translation between Spanish and Basque. En *MT Summit. A workshop at Machine Translation Summit X*. Phuket (Tailandia).

Armentano-Oller, C., A. M. Corbí-Bellot, M.L-Forcada, M. Ginestí-Rosell, B. Bonev, S. Ortiz-Rojas, J. A. Pérez-Ortiz, G. Ramírez-Sánchez y F. Sánchez-Martínez. 2005. An open-source shallow-transfer machine translation toolbox: consequences of its release and availability. En *MT Summit. A workshop at Machine Translation Summit X*. Phuket (Tailandia).

Atserias, J., E. Comelles y A. Mayor. 2005. TXALA: un analizador libre de dependencias para el castellano. *Procesamiento del Lenguaje Natural*, 35:455-456.

Atserias, J., B. Casas, E. Comelles, M. Gonzàlez, L. Padró y M. Padró. 2006. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. En *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*. Génova (Italia).

Corbí Bellot, A.M., M.L. Forcada, S. Ortiz Rojas, J. A. Pérez Ortiz, G. Sánchez Ramírez, F. Sánchez Martínez, I. Alegria, A. Mayor y K. Sarasola. 2005. An open-source shallow-transfer machine translation engine for the Romance languages of Spain. En *Proceedings of the European Association for Machine Translation, 10th Annual Conference*. Budapest (Hungría).