Desarrollo de un analizador morfológico de catalán antiguo basado en corpus textuales

Entidad financiera: Sin financiación directa.[1]

Grupos participantes: Grupo TICA (Tractament Informàtic del Català Antic) de la Univ. d'Alacant: Dept. Llenguatges i Sistemes Informàtics (Mikel L. Forcada, Alícia Garrido-Alenda, Patrícia Gilabert-Zarco); Dept. Filologia Catalana (Marinela Garcia-Sempere, Sandra Montserrat-Buendia); Amaia Iturraspe-Bellver.

Duración: 2 años aproximadamente (mayo 2003–abril 2005).

Responsable: Mikel L. Forcada, Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, E-03071 Alacant. Teléfono: 96 590 9776. Fax: 96 590 9326. Dirección electrónica: `mlf@ua.es`

Resumen: Este proyecto retoma el desarrollo de un analizador morfológico de catalán antiguo públicamente accesible por Internet, el primero de su tipo.[2] Pretendemos que este analizador se convierta en una herramienta útil para el tratamiento morfológico de corpus de catalán antiguo con fines didácticos y de investigación; además, será el primer módulo de un sistema futuro cuyo objetivo es ofrecer automáticamente una lectura en catalán moderno de cualquier texto antiguo en soporte informático. El analizador del que se parte, de cobertura limitada,[3] fue desarrollado por algunos de los autores tomando como base el vocabulario y los paradigmas de flexión antigua recogidos en un diccionario manual (Costa Clos y Tarrés Fernández, 1998). El sistema (extremadamente veloz gracias al uso de técnicas de estados finitos) se genera automáticamente a partir de los datos lingüísticos, cosa que permite una actualización continua y sencilla del programa. Además de reorganizar los diccionarios y completar los paradigmas de flexión y de variación gráfica del lematizador actual, el proyecto se propone usar corpus de textos catalanes antiguos públicamente disponibles (p.ej, RIALC,[4] Biblioteca Virtual Joan Lluís Vives[5]) para hacer el sistema más robusto frente a variaciones gráficas y criterios divergentes de transcripción y para mejorar la cobertura (fracción de texto analizado) mediante la inclusión de entradas según la frecuencia de aparición observada. El objetivo final es la construcción de un analizador morfológico de catalán antiguo que sea rápido, robusto, libremente accesible por Internet y fácilmente integrable en otras aplicaciones (como por ejemplo los buscadores o indexadores de bibliotecas digitales).

En este documento se describen brevemente las estrategias lingüísticas e informáticas que se están usando en el proyecto para conseguir estos objetivos.

Estrategias informáticas: El analizador se basa en ficheros denominados *diccionarios morfológicos* que describen la correspondencia entre las *formas superficiales* (p.ej. *tengua*) observadas en los textos antiguos (flexionadas, tal vez siguiendo algún paradigma antiguo y grafiadas siguiendo criterios de transcripción variables y muchas veces alejados de la grafía estándar catalana actual –*tinga*–) y sus correspondientes *formas léxicas* consistentes en un *lema* o *forma canónica* adecuada (*tenir*), una *categoría léxica* (verbo) e información sobre la flexión (presente de subjuntivo, primera o tercera persona del singular).

Estas correspondencias no se establecen como una simple lista de pares (forma superficial, forma léxica); las regularidades observadas en la flexión y en las variaciones gráficas se representan mediante *paradigmas* (declarados al inicio del diccionario) que agrupan correspondencias alternativas entre fragmentos de formas superficiales y formas léxicas.

El fichero resultante se convierte en un pro-

[1] Financiado indirectamente a través del proyecto CICYT TIC2000-1599 y de dos convenios universidad–empresa.

[2] Financiado en 1999 por la Conselleria de Educació i Ciència de la Generalitat Valenciana.

[3] `http://www.torsimany.ua.es/lematitzador/`

[4] `http://www.rialc.unina.it/`

[5] `http://www.lluisvives.com/`

grama ejecutable basado en transductores de estados finitos (con velocidades superiores a las 10.000 palabras por segundo en ordenadores de sobremesa actuales) usando compiladores similares a los usados en un proyecto de traducción automática (Canals-Marote et al., 2001; Garrido et al., 1999; Garrido-Alenda, Forcada, y Carrasco, 2002).

Por otro lado, nos planteamos estudiar e implementar:

- filtros para permitir el análisis de textos en los formatos más usuales en bibliotecas digitales (XML, HTML, etc.), basados en los usados en sistemas de traducción automática (Canals-Marote et al., 2001);
- varias modalidades de presentación del texto analizado, por ejemplo, preservando el formato y presentando sólo el análisis o la forma moderna de la palabra en un cuadro temporal ("Post-It") cuando el ratón está sobre ella;
- la utilización del analizador morfológico resultante para indexar los textos antiguos de manera que en una biblioteca digital se permita buscar textos por lemas en lugar de por formas superficiales (por ejemplo, de manera que se entregue el poema I de Ausiàs March que comienza "Així com cell que en lo somni es delita" al buscar *delitar*);
- una interfaz estándar para que otros servidores puedan invocar el analizador como un servicio *web*.

Estrategias lingüísticas: Para enriquecer la cobertura y mejorar la precisión del analizador morfológico se está siguiendo el siguiente plan de trabajo:

- Recogida, limpieza y fijación de textos antiguos ya digitalizados, sean cuales sean los criterios usados en su transcripción, de épocas anteriores a la normalización ortográfica del catalán (especialmente del siglo XIV al XVI).
- Estudio estadístico de las formas superficiales observadas en estos textos con el fin de añadir nuevos lemas o de refinar los paradigmas de flexión y de variación gráfica ya existentes en los diccionarios morfológicos, dando prioridad a aquellas formas superficiales más frecuentes con el fin de asegurar la máxima cobertura posible en cada momento.
- Completación de los paradigmas de flexión antigua con material procedente de fuentes adecuadas (Alcover y Moll, 1984), y con las variantes observadas en los corpus.
- Fijación de criterios para asignar lemas a formas antiguas (*aycels* → ¿*aicell* o *aquell*?) con el fin de facilitar (a) la lectura de los textos antiguos por parte de expertos y (b) la futura traducción de los mismos al catalán normalizado.
- Evaluación periódica de los analizadores morfológicos resultantes sobre los textos de referencia y sobre otros textos de las mismas épocas.

Conclusión: Nos proponemos desarrollar, partiendo de un prototipo ya existente, un analizador morfológico de catalán antiguo público, extensible, rápido, robusto y flexible, con el fin de que se pueda integrar en otros productos o servicios, tales como los ingenios de búsqueda en bibliotecas digitales o sistemas de lectura asistida de textos antiguos.

Bibliografía

Alcover, Antoni M. y Francesc de B. Moll. 1984. *Diccionari català–valencià–balear*. Ed. Moll, Palma de Mallorca. (10 vols.; disponible también a través de `http://dcvb.iecat.net`).

Canals-Marote, Raül, Anna Esteve-Guillen, Alicia Garrido-Alenda, Maribel Guardiola-Savall, Amaia Iturraspe-Bellver, Sandra Monserrat-Buendia, Sergio Ortiz-Rojas, Hermínia Pastor-Pina, Pedro M. Perez-Antón, y Mikel L. Forcada. 2001. El sistema de traducción automática castellano-catalán interNOSTRUM. *Procesamiento del Lenguaje Natural*, 27:151–156.

Costa Clos, Mercé y Maribel Tarrés Fernández. 1998. *Diccionari del català antic*. Edicions 62, Barcelona.

Garrido, Alicia, Amaia Iturraspe, Sandra Montserrat, Hermínia Pastor, y Mikel L. Forcada. 1999. A compiler for morphological analysers and generators based on finite-state transducers. *Procesamiento del Lenguaje Natural*, (25):93–98.

Garrido-Alenda, Alicia, Mikel L. Forcada, y Rafael C. Carrasco. 2002. Incremental construction and maintenance of morphological analysers based on augmented letter transducers. En *Proceedings of TMI 2002 (Theoretical and Methodological Issues in Machine Translation, Keihanna/Kyoto, Japan, March 2002)*, páginas 53–62.