SAÓ - Sistema de ayuda ortoépica para la lectura en voz alta del valenciano.

Entidad financiera: Acadèmia Valenciana de la Llengua

Grupos participantes: Departament de Llenguatges i Sistemes Informàtics de la Universitat d'Alacant (Mikel L. Forcada); Departament de Filologia Catalana de la Universitat d'Alacant (Vicent Beltran, Carles Segura, Jordi Colomina).

Duración: 1 año (junio 2003–mayo 2004).

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[1] se propone elaborar un programa informático que ayudará a los locutores de radio y televisión y, en general, a cualquier persona, a leer el valenciano correctamente en voz alta. El sistema resultante, denominado SAÓ (sistema de ayuda ortoépica), anotará automáticamente el texto con marcas sencillas que indicarán la pronunciación correcta en algunos casos difíciles, usando un ordenador de sobremesa común. En particular, podrá ser usado para enseñar la pronunciación del valenciano a quien lo esté aprendiendo.

SAÓ podrá leer textos llanos (ASCII, ANSI) y formateados (RTF, HTML, XML); estará basado en diccionarios y reglas de pronunciación, tanto para palabras problemáticas completas como para sufijos (p. ej., terminaciones como por ejemplo *-ori*, *-osi*, etc.). Estos diccionarios se convertirán, usando tecnologías derivadas de las producidas en el seno del proyecto de traducción interNOSTRUM[2] (Canals-Marote et al., 2001), en programas basados en técnicas de estados finitos que optimizan la búsqueda de las palabras o sufijos en el diccionario de pronunciación. En cuanto al tratamiento de los diversos formatos de texto, también será de aplicación la tecnología usada en interNOSTRUM; en particular, la que permitirá la anotación en línea (al vuelo) de los textos de Internet a los cuales se acceda mediante un programa navegador.

Objetivos: La naturaleza definitiva (la apariencia en la pantalla o en el papel) de las anotaciones se determinará durante el proyecto, después de un estudio de ergonomia y aprovechando los resultados de estudios similares en otros lenguas. En particular, se desea que las anotaciones: (a) sean sencillas de aprender, de identificar y de interpretar a las velocidades de lectura usuales;[3] (b) sean mínimas, es decir, modifiquen o anoten el texto donde sea estrictamente necesario para un locutor o lector concreto; (c) conserven la ortografía original del texto; (d) sean técnicamente compatibles con los medios de lectura y formatos usados por locutores; (e) se generen rápidamente para evitar retardos entre la generación y la lectura del texto, y (f) sean generadas mediante un sistema de anotación accesible desde cualquier ordenador conectado a la red Internet.

La información de pronunciación que ha de generar el sistema tiene que ser independiente de la presentación concreta que se haga de esta información. Esto nos permitirá presentarla en más de un estilo diferente y diseñar el componente de anotación fonética independientemente del de presentación; por ejemplo, para abordar problemas de ergonomia o de preferencias de los lectores o de los medios. Las anotaciones seran, con toda probabilidad, del estilo de las anotaciones *ruby* o *furigana* usadas en japonés para indicar la pronunciación correcta de los ideogramas (*kanji*): pequeños símbolos en un tipo de letra más pequeño sobre las grafías que precisan anotación, o como superíndice. He aquí un ejemplo de anotación del valenciano con superíndices:

[1] Surgido de una iniciativa del académico de la Acadèmia Valenciana de la Llengua y catedrático de Filología Catalana de la Universitat d'Alacant Jordi Colomina Castanyer.

[2] `http://www.interNOSTRUM.com`

[3] No pueden, por tanto, basarse en alfabetos fonéticos como el internacional, conocidos sólo por especialistas.

El$\underline{s}^{z}$ diputa$\underline{ts}^{dz}$ han decidit que cal que es po$\underline{o}^{ò}\underline{s}^{z}$en lími$\underline{ts}^{dz}$ al gov$\underline{e}^{è}$rn.

Necesidad de la anotación: Las anotaciones de pronunciación son necesarias en valenciano por diversas causas:

- Por la ambigüedad (residual) del sistema de escritura, en general razonablemente ajustado a la pronunciación: por ejemplo, la grafia *o* en posición tónica sin tilde puede ser abierta como en *cosa* ([ˈkɔza]) o cerrada ([ˈkopa]).
- Por divergencias entre una grafía estándar no ambigua y la pronunciación valenciana: por ejemplo la *è* de *perquè*, a pesar de llevar una tilde grave que la marca como abierta se pronuncia cerrada en valenciano ([perˈke]).
- Por la existencia de algunas secuencias mudas: *assumpte* ([aˈsunte]), *comprendre* ([komˈpendre]), etc.
- Para evitar pronunciaciones viciadas (normalmente por el castellano) a pesar de que la grafía sea clara: *oboé* ([oβoˈe]), *adequen* ([aˈðekwen]), Eufrates ([ewˈfrates])
- Para pronunciar correctamente grafías no estándar en apellidos valencianos: *Domenech* ([doˈmenek], no [doˈmenetʃ]), *Chordá* ([dʒorˈða], no [tʃorˈða]), etc.
- Para anotar cambios de pronunciación debidos al contacto con otras palabras: *quina casa* ([ˈkina ˈkaza]) pero *quina hora* ([ˈkin ˈɔra]), *els cotxes* ([els ˈkotʃes]) pero *els amics* ([elz aˈmiks]), etc.
- Para indicar pronunciaciones de una variedad concreta del valenciano.

Diseño informático: El sistema de anotación se basa en cuatro módulos: un *desformateador* que separa el texto a anotar de la información de formato existente en el documento original, un *preanotador* que determina anotaciones provisionales (dependientes del contexto) y definitivas (independientes del contexto) de la pronunciación de cada palabra, un *postanotador* que confirma o descarta las anotaciones provisionales de acuerdo con el contexto y un *reformateador* que restituye el formato al texto original a la vez que materializa la forma gráfica que tendrán las anotaciones de pronunciación. Todos los programas están basados en tecnología de estados finitos.

El programa anotador se genera, usando un compilador adecuado, a partir de un diccionario de preanotación en el que se especifican las anotaciones provisionales y definitivas de la pronunciación de cada palabra; dado que muchos fenómenos ocurren en las terminaciones de palabra, el uso de paradigmas de flexión anotados simplifica enormemente la generación de todas las formas. El postanotador se genera, usando otro compilador, a partir de un diccionario de reglas de anotación para la realización de las anotaciones provisionales dependiendo del contexto en el que ocurran. Tanto el diccionario de preanotación como el diccionario de reglas de anotación son archivos XML. Los compiladores están basados en los usados en el proyecto interNOSTRUM.

Prototipo y situación actual: Se ha completado un catálogo provisional de fenómenos que necesitan ser anotados (Lacreu et al., 2001; Segura Llopes, 2003). Un prototipo de demostración del programa de anotación (para textos llanos y HTML, compatible con navegadores capaces de ejecutar hojas de estilo CSS2) que conoce unas doscientas palabras muy frecuentes en valenciano está actualmente disponible en `http://www.internostrum.com/sao/`. En la actualidad se está a la espera de los primeros fondos de la Acadèmia Valenciana de la Llengua para contratar a los becarios que realizarán el sistema.

Bibliografía

Canals-Marote, R., A. Esteve-Guillén, A. Garrido-Alenda, M. Guardiola-Savall, A. Iturraspe-Bellver, S. Monserrat-Buendia, S. Ortiz-Rojas, H. Pastor-Pina, P.M. Perez-Antón, y M.L. Forcada. 2001. El sistema de traducción automática castellano-catalán interNOSTRUM. *Procesamiento del Lenguaje Natural*, 27:151–156. XVII Congreso de la Sociedad Española de Procesamiento del Lenguaje Natural, Jaén, Spain, 12-14.09.2001.

Lacreu, Josep, Rosanna Mestre, Francesc B. Salas, y Ofèlia Sanmartín. 2001. *Diccionari valencià de pronunciació*. Bromera.

Segura Llopes, Carles. 2003. Ressenya de Lacreu, J. (2001) *Diccionari valencià de pronunciació*, ed. Bromera. *Caplletra*, 32:202–206.