Llisterri, J., Machuca, M. J., Madrigal, N., Mancini, F., Massimino, P., de la Mota, C., . . . Ríos, A. (2004). Aspectos lingüísticos en el diseño de un conversor de texto en habla en castellano y en catalán: El sistema loquendo TTS®. In *VI Congreso de Lingüística General*. Universidad de Santiago de Compostela, Santiago. 3-7 de mayo de 2004. (pp. 521-2). Santiago de Compostela: Universidade de Santiago de Compostela, Facultade de Filoloxía, Área de Lingüística Xeral.

http://liceu.uab.cat/~joaquim/publicacions/Llisterri_et_al_04_Conversor_Texto_Habla_Castellano_Catalan_Loquendo.pdf

Aspectos lingüísticos en el diseño de un conversor de texto en habla en castellano y en catalán: el sistema Loquendo TTS®

JOAQUIM LLISTERRI
Universitat Autònoma de Barcelona

MARÍA MACHUCA
Universitat Autònoma de Barcelona

NATALIA MADRIGAL
Universitat Autònoma de Barcelona

FRANCA MANCINI
Loquendo SpA

PAOLO MASSIMINO
Loquendo SpA

CARME DE LA MOTA
Universitat Autònoma de Barcelona

MONTSERRAT RIERA
Universitat Autònoma de Barcelona

ANTONIO RÍOS
Universitat Autònoma de Barcelona

[Paneles de investigación]

Un sistema de conversión de texto en habla tiene como finalidad la transformación automática de cualquier texto escrito que esté disponible en formato electrónico en su correspondiente realización sonora. Dado que se trata de una aplicación informática que ha de reproducir el proceso que realizaría el lector de un texto al emitirlo en voz alta, es necesario dotar al sistema de toda la información lingüística que se precisa para efectuar la lectura. En el presente trabajo se describen los distintos módulos lingüísticos que emplea, para la síntesis de habla del castellano y del catalán, el sistema multilingüe Loquendo TTS® desarrollado por Loquendo SpA (Quazza *et al.*, 2001).

El sistema consta de dos grandes módulos, uno destinado al análisis lingüístico-prosódico y otro que lleva a cabo la síntesis. El módulo de análisis lingüístico-prosódico realiza el análisis léxico de las palabras del texto para determinar su categoría y para localizar las sílabas portadoras de acento, efectúa un análisis sintáctico-prosódico para asignar etiquetas de límite, transcribe fonéticamente el texto y calcula los parámetros prosódicos de duración y F_0 para cada sonido. Para generar una onda sonora a partir de los datos obtenidos, el módulo de síntesis concatena las unidades más largas que se ajustan a cada contexto, con lo que consigue reducir el número de concatenaciones y la necesidad de reajustes prosódicos. Las unidades se seleccionan a partir de una amplia base de datos acústica, una técnica que permite encontrar más fácilmente la unidad apropiada.

En el desarrollo del sistema Loquendo TTS® para el castellano y el catalán han trabajado tanto expertos en telecomunicación y en informática como lingüistas especialistas en cada una de las dos lenguas (Llisterri *et al.*, 2003). El papel del lingüista es aportar los datos específicos de cada lengua, proponer soluciones con conocimiento de causa a los problemas propios de su ámbito y, en las etapas finales, evaluar el resultado en función de su familiaridad con los patrones habituales que esperan encontrar los usuarios de una determinada comunidad lingüística. Esto significa que debe dotar de información categorial a las palabras del texto, intervenir en la definición del inventario de sonidos que se va a emplear para cada lengua, realizar la adaptación del módulo de preprocesado, diseñar las reglas de transcripción fonética, silabificación y acentuación, definir las reglas de segmentación prosódica, intervenir en el diseño de la base de datos de la que se van a extraer las unidades y en la selección del locutor, diseñar las reglas de modelado prosódico (Garrido *et al.*, 2000) y evaluar la acción de los módulos diseñados y la calidad del habla generada.

La versión actual del sistema (puede accederse a una demostración en http://loquendo.com), posee ya varias voces para el castellano y una voz femenina para el catalán. En la actualidad se está trabajando en la mejora de los sistemas ya existentes para el español peninsular y en el desarrollo de una nueva voz, ahora masculina, para el catalán.

REFERENCIAS:

GARRIDO, J.M.- ORTÍN, I.- QUAZZA, S.- SALZA, P.L.- MANCINI, F. (2000) "Desarrollo de un módulo de asignación de parámetros prosódicos para la versión en español del sistema de conversión texto-habla ACTOR®", *Procesamiento del Lenguaje Natural, Revista n°* 26: 183-190.

LLISTERRI, J.- CARBÓ, C.- MACHUCA, M. J.- de la MOTA, C.- RIERA, M.- RÍOS, A. (2003) "La conversión de texto en habla: aspectos lingüísticos", in MARTÍ, M. A. – LLISTERRI, J. (Eds.) *Tecnologías del texo y del habla.* Barcelona. Edicions de la Universitat de Barcelona – Fundación Duques de Soria. (to appear).

http://liceu.uab.es/publicacions/Linguistica_CTH_FDS02.pdf

QUAZZA, S.- DONETTI, L.- MOISA, L.- SALZA, P.L. (2001) "ACTOR: A multilingual unit-selection speech synthesis system", in *Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis.* August 29 - September 1, 2001. Pertshire, Scotland. http://www.isca-speech.org/archive/ssw4/ssw4_209

ASPECTOS LINGÜÍSTICOS EN EL DISEÑO DE UN CONVERSOR DE TEXTO EN HABLA EN CASTELLANO Y EN CATALÁN: EL SISTEMA LOQUENDO TTS®

Joaquim Llisterri[1], María Machuca[1], Natalia Madrigal[1], Franca Mancini[2], Paolo Massimino[2], Carme de la Mota[1], Montserrat Riera[1], Antonio Ríos[1]
[1]*Departament de Filologia Espanyola, Universitat Autònoma de Barcelona*
[2]*Loquendo SpA*

1. Introducción

Un conversor de texto en habla es, esencialmente, un programa informático que permite leer en voz alta de forma automática cualquier texto previamente almacenado en formato electrónico. Desarrollar una aplicación de este tipo y lograr una elocución lo más natural posible supone conjugar distintos tipos de saberes que, tradicionalmente, se han situado en ámbitos académicos y profesionales muy alejados. Por una parte, es obvio que se requiere un componente importante de programación que corre a cargo de especialistas en informática y, por otra, parece claro que, si el resultado es una señal sonora, deben intervenir también expertos en tratamiento digital de señales, área habitualmente ligada a la ingeniería de telecomunicación. Sin embargo, a menudo se olvida que la "materia prima" con la que trata un conversor es la lengua, tanto en su vertiente oral como en la escrita, y que, por lo tanto, la intervención del lingüista es también necesaria para alcanzar un buen resultado (Llisterri 2003, Llisterri *et al*. 2003a, b).

La estrategia seguida para el desarrollo de las versiones castellana y catalana del conversor de texto en habla de Loquendo SpA se ha basado, justamente por este motivo, en la estrecha colaboración entre un equipo orientado hacia los aspectos técnicos y otro centrado en los componentes lingüísticos. Tal como se muestra en este trabajo, la incorporación de reglas que reflejan el conocimiento de la estructura lingüística permite, en combinación con modelos de corte más estadístico, llegar a disponer de un conversor que, si bien es aún susceptible de mejoras, ofrece un elevado nivel de calidad que permite su distribución comercial.

En los siguientes apartados, tras una descripción general del sistema de conversión de texto en habla LoquendoTTS, se presenta la información lingüística que se ha integrado en cada uno de sus módulos, los métodos utilizados para la obtención del corpus de síntesis y los procedimientos empleados para evaluar las versiones castellana y catalana del sistema[1]. En concreto, se describe la información lingüística que se ha precisado para desarrollar las tres voces del español peninsular (dos masculinas –Jorge y Juan– y una femenina –Carmen–) y las dos voces del catalán (una masculina –Jordi– y una femenina –Montserrat–).

[1] Además de los firmantes de este trabajo, colaboraron también en diversas etapas del proyecto Beatriu Fernández, Juan M. Garrido e Isabel Ortín.

Llisterri, J., Machuca, M. J., Madrigal, N., Mancini, F., Massimino, P., Mota, C., et al. (2004).
Aspectos lingüísticos en el diseño de un conversor de texto en habla en castellano y en catalán: El sistema Loquendo TTS®.
In VI Congreso de Lingüística General. (pp. 521-2). Santiago de Compostela.
http://liceu.uab.cat/~joaquim/publicacions/Llisterri_et_al_04_Conversor_Texto_Habla_Castellano_Catalan_Loquendo.pdf

2. Descripción general del sistema LoquendoTTS

LoquendoTTS es un sistema comercial de conversión de texto en habla multilingüe y multilocutor basado en la concatenación de segmentos acústicos contextuales no uniformes directamente extraídos del habla natural (Quazza *et al*. 2001). Con este procedimiento se pretende preservar al máximo la naturalidad del timbre de la voz original.

El sistema LoquendoTTS se estructura según un esquema modular, compuesto por un núcleo cuyo funcionamiento es independiente del idioma y por módulos satélites que contienen información específica para cada lengua desarrollada. Tanto el núcleo central como los módulos lingüísticos son componentes construidos únicamente mediante *software* que no requieren el uso de tarjetas o de *hardware* especiales. Se incorporan tanto los conocimientos lingüísticos necesarios como los datos acústicos que se requieren para generar habla sintetizada mediante la concatenación de unidades acústicas. Estos datos –que forman el corpus de síntesis– consisten en señal vocal etiquetada de manera que se pueda recuperar toda la información segmental y suprasegmental, y son específicos para cada voz. De hecho, en la síntesis por concatenación la cantidad de datos acústicos disponibles incide de un modo muy importante en la calidad y la naturalidad de la voz.

Desde un punto de vista conceptual, LoquendoTTS se compone de dos módulos fundamentales: el primer módulo –que podemos llamar módulo de análisis textual– convierte el texto de entrada en una representación fonética detallada, acompañada de informaciones prosódicas. Este proceso, expuesto en el apartado 3, prevé diferentes fases de elaboración del texto antes de obtener la trascripción fonética: la delimitación de palabra y de frase, la expansión de números, acrónimos, direcciones de web y símbolos matemáticos, el etiquetado gramatical y prosódico, la asignación de acento léxico y, finalmente, la conversión de grafema a sonido. El proceso depende sustancialmente del idioma: para obtener el resultado deseado, el núcleo de elaboración principal cuenta con funcionalidades presentes en el módulo específico de cada lengua.

El segundo módulo –que llamaremos módulo de síntesis– se encarga de convertir la representación abstracta producida por el módulo de análisis textual en una señal vocal. Esto se consigue a través de la concatenación de una serie de unidades acústicas adecuadamente elegidas entre las que constituyen la base de datos vocal o corpus de síntesis.

Las unidades acústicas que se eligen no están determinadas a priori, sino que se seleccionan dinámicamente en función de las secuencias de fonemas que se quieran sintetizar y de su contexto (Balestri *et al*. 1999). La filosofía de la concatenación consiste en reducir al mínimo la manipulación de la señal vocal para no alterar su calidad, y en evitar discontinuidades en los puntos de concatenación, de modo que se obtenga una voz sintetizada con un buen nivel de naturalidad.

El algoritmo de selección de unidades acústicas compara la descripción fonético-prosódica del texto de entrada con las informaciones contenidas en la base de datos acústica. De esta manera se obtienen las secuencias más largas de difonemas o de unidades superiores al difonema que resultan idóneas para ser concatenadas y que corresponden a las partes de texto que se quiere sintetizar. Si las características prosódicas de las unidades acústicas seleccionadas son lo bastante parecidas a las requeridas por la selección como para permitir realizar cortes en segmentos consonánticos donde las discontinuidades de la frecuencia fundamental son limitadas, la simple concatenación de ondas es suficiente para obtener el resultado final. En caso

Llisterri, J., Machuca, M. J., Madrigal, N., Mancini, F., Massimino, P., Mota, C., et al. (2004).
Aspectos lingüísticos en el diseño de un conversor de texto en habla en castellano y en catalán: El sistema Loquendo TTS®.
In VI Congreso de Lingüística General. (pp. 521-2). Santiago de Compostela.
http://liceu.uab.cat/~joaquim/publicacions/Llisterri_et_al_04_Conversor_Texto_Habla_Castellano_Catalan_Loquendo.pdf

contrario, es necesario realizar modificaciones en los parámetros de duración y de frecuencia fundamental.

Una base de datos acústica alcanza habitualmente una dimensión de 200Mb cuando está codificada a 16kHz con PCM (*Pulse Code Modulation*) lineal. Se utilizan también otros formatos de codificación que permiten obtener una base de datos acústica de menor tamaño, como por ejemplo la codificación en 8 kHz PCM, 8kHz *μlaw* y *A-law*, y algunas técnicas de compresión *lossy*.

El sistema de síntesis de LoquendoTTS está disponible para sistemas que utilizan Microsoft Windows®, Sun Solaris® y Linux, pero se podría utilizar también con otras arquitecturas, ya que todo el software que lo compone está escrito en ANSI C. LoquendoTTS utiliza el lenguaje de marcas VoiceXML y acepta el texto de entrada tanto en formato ASCII como en UNICODE. Esta característica permite tratar cualquier tipo de texto con una codificación estándar y universalmente reconocida, lo que permite también el empleo del conversor en lenguas como el chino o el japonés.

El funcionamiento del sistema está completamente configurado a través de API (*Application Program Interfaces*) que permiten al usuario no sólo elegir la voz y el idioma deseados para la síntesis, sino también modificar algunos parámetros fundamentales, como el rango tonal y la velocidad de elocución.

Asociando el concepto de canal a una determinada voz, el sistema permite controlar de forma simultánea e independiente un número elevado de canales Por término medio, empleando un ordenador con CPU Intel® de 1.4GHz se pueden gestionar al mismo tiempo 160 canales.

Finalmente, debe señalarse que Loquendo TTS puede encontrarse ya en distintas lenguas y variantes geográficas: alemán, catalán, chino, francés, griego, inglés americano e inglés británico, italiano, español (castellano, argentino, chileno, mexicano), portugués de Brasil y portugués europeo, y sueco[2].

3. La información lingüística integrada en el sistema Loquendo TTS

La conversión de texto en habla, si se lleva a cabo del modo en que se describe en el apartado anterior, es un procedimiento complejo que requiere, además de una amplia base de datos acústica, la aplicación de conocimientos lingüísticos de distinto orden. La voz generada deberá reproducir un modelo de habla y de pronunciación, para lo que deberán tenerse en cuenta los estudios normativos, sociolingüísticos y dialectales. La elección de las unidades de síntesis deberá basarse en el conocimiento del sistema fonológico y de la descripción fonética de la lengua para la que se desarrolle el sistema (Quazza, van den Heuvel, 2000). Además, el modelado de la señal requerirá conocimientos de fonética acústica segmental y suprasegmental, algo que depende también de la información morfológica y sintáctica, especialmente en lo que respecta a la adecuada realización de las pausas y de los patrones melódicos en la lectura. A continuación se presentan en detalle los pasos necesarios para realizar las versiones castellana y catalana del conversor de texto en habla desarrollado por Loquendo SpA en colaboración con la Universitat Autònoma de Barcelona.

[2] En las páginas en la web de Loquendo (http://www.loquendo.com/) pueden escucharse demostraciones del funcionamiento del conversor en cada una de estas lenguas.

Llisterri, J., Machuca, M. J., Madrigal, N., Mancini, F., Massimino, P., Mota, C., et al. (2004).
Aspectos lingüísticos en el diseño de un conversor de texto en habla en castellano y en catalán: El sistema Loquendo TTS®.
In VI Congreso de Lingüística General. (pp. 521-2). Santiago de Compostela.
http://liceu.uab.cat/~joaquim/publicacions/Llisterri_et_al_04_Conversor_Texto_Habla_Castellano_Catalan_Loquendo.pdf

3.1. El preprocesamiento del texto

Un primer paso para iniciar la conversión de texto en habla consiste en la preparación del texto escrito para que el sistema sea capaz de transformar la representación ortográfica en lengua oral, independientemente de los símbolos que en él aparezcan y sea cual sea su procedencia.. El módulo de preprocesamiento consta de una serie de reglas y listas que pretenden sistematizar toda la información lingüística necesaria para que el conversor oralice de forma correcta los elementos cuya lectura puede ser ambigua —como sucede con las siglas y los acrónimos—, deba ser interpretada —como es el caso de los números—, o no siga las reglas de transcripción generales de la lengua específica del sistema —por ejemplo, los extranjerismos— (véanse los apartados 3.4, 3.5 y 3.7).

En el módulo de preprocesamiento del sistema de conversión Loquendo TTS se ha incluido la siguiente información lingüística, tanto para el castellano como para el catalán:

- interpretación de símbolos ASCII,
- reglas para la lectura de siglas,
- reglas para la detección y expansión de números y otras expresiones matemáticas,
- instrucciones para la lectura de las abreviaturas más comunes,
- extranjerismos acentuados y transcritos, dado que en muchas de sus grafías no siguen las reglas generales de acentuación y de transcripción, y
- palabras terminadas en *–ment* (catalán) o *–mente* (castellano) que no son adverbios y que, por lo tanto, no presentan acento secundario.

En el caso del catalán, además, se ha incorporado información sobre palabras derivadas por prefijación y sobre palabras compuestas. Se han señalado las cadenas gráficas iniciales tras las cuales existe un límite morfológico interno a consecuencia del cual se ve afectada la transcripción, la acentuación o la silabación de la palabra y se han indicado también las correspondientes excepciones. Por ejemplo, las diferencias de pronunciación que se observan en las palabras del catalán que empiezan por *bi-, contra-* y *sub* se explican si se tiene en cuenta la estructura morfológica: mientras que *bi*[s]*il·làbic* (“bisilábico”), *c*[o]*ntra*[r]*evolució* (“contrarrevolución”) y *su*[b]*lunar* (“sublunar”) son compuestos, *bi*[z]*ellar* (“biselar”), *c*[u]*ntra*[ɾ]*i* (“opinión”) y *su*[β̪]*lim* (“sublime”) no lo son.

3.2. La información categorial

Los categorizadores diseñados para el catalán y el castellano consisten en un conjunto de reglas que tienen como objetivo asignar la categoría gramatical a las palabras de los textos escritos en dichas lenguas. La asignación categorial es necesaria para el modelado prosódico, para la asignación correcta de las pausas e incluso para la transcripción fonética de los textos. Por ejemplo, en catalán la forma ortográfica *son* puede corresponder a dos palabras gramaticales distintas –un nombre ("sueño") y un determinante posesivo ("su")–, cada una con su propia forma fonológica: /ˈsɔn/ y /sun/.

Las reglas aplicadas no atienden a información semántica, sino únicamente a la forma gráfica de la palabra. Pueden ser de tres tipos: (1) reglas que asignan una categoría, de modo inequívoco, a aquellas palabras que no presentan ambigüedad

Llisterri, J., Machuca, M. J., Madrigal, N., Mancini, F., Massimino, P., Mota, C., et al. (2004).
Aspectos lingüísticos en el diseño de un conversor de texto en habla en castellano y en catalán: El sistema Loquendo TTS®.
In VI Congreso de Lingüística General. (pp. 521-2). Santiago de Compostela.
http://liceu.uab.cat/~joaquim/publicacions/Llisterri_et_al_04_Conversor_Texto_Habla_Castellano_Catalan_Loquendo.pdf

categorial; (2) reglas que asignan la categoría —o categorías— en función de las terminaciones de las palabras, y (3) reglas contextuales, que eliminan las ambigüedades generadas por las reglas anteriores o categorizan las palabras que aún no han sido etiquetadas, teniendo en cuenta la categoría de las palabras adyacentes.

3.3. El inventario de sonidos

Tanto en catalán como en castellano se han contemplado aquellos sonidos propios de estas lenguas, teniendo en cuenta los procesos fonológicos regulares que actúan en cada una de ellas: asimilación del punto y modo de articulación y de la sonoridad, debilitamiento o refuerzo. Este es el caso, por ejemplo, de las sonorizaciones, que tanto pueden darse en interior de palabra (por ejemplo la consonante fricativa alveolar de *mismo*), como entre sonidos de palabras distintas que entran en contacto (por ejemplo la consonante fricativa labiodental de la secuencia del catalán *tu*[v] *d'oli*, "tufo de aceite"). De modo análogo, existe para el castellano un alófono nasal velar que se emplea en aquellos contextos en los que se desencadena una velarización por asimilación del punto de articulación, como en la cadena *ni*[ŋ]*gú*[ŋ] *contacto.*

Se han incluido además unidades que no forman parte del sistema fónico propio de la lengua pero que pueden necesitarse para la pronunciación de extranjerismos. En catalán por ejemplo, se dispone del sonido fricativo interdental sordo [θ̟] para la pronunciación de apellidos castellanos como *Martínez*.

3.4. La silabación

Las reglas de separación en sílabas son necesarias –tanto en castellano como en catalán– para realizar otras tareas posteriores, por ejemplo, la acentuación y la transcripción fonética. En el primer caso, la separación en sílabas permite detectar los núcleos silábicos y determinar la sílaba tónica de cada palabra; en el segundo, la posición de un determinado segmento dentro de la sílaba puede conllevar la asignación a una determinada grafía de sonidos distintos. Por ejemplo, en catalán la grafía <i> constituye un núcleo silábico vocálico en la palabra *ti.a*, pero no en la palabra *no.ia*.

En ambas lenguas, la silabación de cada palabra se realiza a partir de una serie de reglas que determinan los elementos que constituyen el núcleo silábico (las vocales) y que permiten identificar, posteriormente, los elementos consonánticos de la sílaba que pueden aparecer en posición de ataque o de coda, teniendo en cuenta la estructura fonológica de cada lengua.

Aparte de las reglas que asignan a cada elemento la etiqueta de "ataque", "núcleo" o "coda", se han elaborado listas de palabras para la correcta silabación y transcripción de los extranjerismos (por ejemplo *top.less*), así como un listado de las cadenas iniciales de compuestos y de prefijos productivos que dan lugar a la aparición de un límite interno de palabra en el que no se aplican las reglas generales de silabación de ambas lenguas (por ejemplo, en catalán, *sub.al.pí*).

3.5. La acentuación

Llisterri, J., Machuca, M. J., Madrigal, N., Mancini, F., Massimino, P., Mota, C., et al. (2004).
Aspectos lingüísticos en el diseño de un conversor de texto en habla en castellano y en catalán: El sistema Loquendo TTS®.
In VI Congreso de Lingüística General. (pp. 521-2). Santiago de Compostela.
http://liceu.uab.cat/~joaquim/publicacions/Llisterri_et_al_04_Conversor_Texto_Habla_Castellano_Catalan_Loquendo.pdf

Mediante el componente dedicado a la acentuación se decide qué palabras son átonas y qué palabras son tónicas, y dentro de estas últimas, cuáles se pueden acentuar a partir de las reglas de asignación de acento y cuáles llevan acento secundario, por ejemplo, los adverbios terminados en *–mente* o los compuestos. La asignación del acento a las palabras tónicas se realiza a través de un sistema de aprendizaje automático que se entrena mediante listas de palabras de ambas lenguas, previamente silabificadas y acentuadas, según las conocidas reglas teóricas que tienen en cuenta los núcleos silábicos, la terminación de las palabras y las reglas de acentuación gráfica. Todos aquellos casos cuya acentuación no es regular y que, por tanto, no puede determinarse a partir de reglas, como sucede con las formas léxicamente átonas (*de, dels, pels,* en catalán) o los extranjerismos (*Washington*), se tratan mediante listas de excepciones como las descritas en el apartado 3.1.

3.6. La transcripción fonética

El transcriptor tiene el objetivo de representar mediante símbolos fonéticos la pronunciación de un texto escrito, de acuerdo con un modelo aceptado de pronunciación. . En el caso de las lenguas para las que se ha creado una voz, como es el caso del castellano y del catalán, se ha optado por escoger un modelo próximo al considerado estándar por la comunidad y se han tenido en cuenta tanto las indicaciones de carácter normativo como la información recogida en distintas obras de referencia sobre la lengua oral. Se han evitado los coloquialismos, las expresiones poco prestigiosas y la mezcla de formas propias de variantes geográficas distintas.

La transcripción se realiza a partir de un inventario de alófonos y de un conjunto de reglas contextuales con las correspondientes listas de excepciones. Por ejemplo, para transcribir la consonante nasal de la secuencia <con Carlos> se emplean dos bloques de reglas diferentes. En primer lugar operan las reglas que convierten grafemas en alófonos, cuyo ámbito es la palabra, y el grafema <n> que se encuentra al final de la palabra <con> queda transcrito con la nasal alveolar [n]. En segundo lugar, las reglas de transcripción entre palabras, que operan sobre palabras ya transcritas fonéticamente, se encargan de dar cuenta del proceso de asimilación del punto de articulación que sufre la nasal al encontrarse ante un sonido velar y cambian el alófono [n] por [ŋ]. El resultado es la transcripción correcta de una nasal velar debida a una asimilación del punto de articulación.

En el caso del catalán se han empleado tres bloques de reglas según su ámbito de aplicación: reglas para la conversión de grafía a alófono en posición interior de morfema, reglas para la transcripción de elementos separados por límites internos a la palabra, que operan únicamente sobre palabras prefijadas o compuestas, y reglas de transcripción de elementos separados por límite de palabra, que dan cuenta de los fenómenos de contacto, como en el ejemplo anterior. Se ha previsto también, como se indica en el apartado 3.1, la pronunciación de los extranjerismos.

Llisterri, J., Machuca, M. J., Madrigal, N., Mancini, F., Massimino, P., Mota, C., et al. (2004).
Aspectos lingüísticos en el diseño de un conversor de texto en habla en castellano y en catalán: El sistema Loquendo TTS®.
In VI Congreso de Lingüística General. (pp. 521-2). Santiago de Compostela.
http://liceu.uab.cat/~joaquim/publicacions/Llisterri_et_al_04_Conversor_Texto_Habla_Castellano_Catalan_Loquendo.pdf

3.7 La información prosódica

Buena parte de la información prosódica que debe emplearse al leer un texto es difícil de predecir a partir únicamente de los signos ortográficos y de puntuación. Los hablantes de español, por ejemplo, tienden a realizar pausas entre sujeto y verbo, especialmente si el sujeto es largo, pero esas pausas no están marcadas en un texto y se precisa un análisis lingüístico para poder predecir su aparición.

Una vez se ha determinado la categoría y el acento de las palabras es posible efectuar un análisis sintáctico-prosódico. Se definen las reglas de segmentación prosódica y se asignan las etiquetas de límite que han de permitir asignar las pausas y especificar las características prosódicas (Balestri *et al.* 1999). Se calculan los parámetros de duración y frecuencia fundamental para cada sonido de acuerdo con las particularidades de cada lengua, y se asignan oportunamente etiquetas con información sobre las modificaciones prosódicas que deben realizarse en las posiciones señaladas.

Para generar la onda sonora que corresponde al texto escrito, el módulo de síntesis concatena las unidades más largas que se ajustan a cada contexto, con lo que se consigue reducir el número de concatenaciones y la necesidad de reajustes prosódicos (Balestri *et al*. 1999), como se ha señalado en el apartado 2. Aunque las unidades necesarias se seleccionan a partir de una amplia base de datos acústica específica para cada lengua, podría suceder también que la unidad deseada no se encontrara registrada en ella. Podrían no haberse almacenado oraciones interrogativas o exclamativas, por ejemplo. En ese caso, las reglas de modelado prosódico específicas para cada tipo de modalidad oracional permiten asignar correctamente la prosodia. Estas reglas son específicas para cada lengua y en ellas se tiene en cuenta la agrupación lingüística en unidades, tanto sintácticas como acentuales y entonativas, y se predice la asignación y concatenación de movimientos melódicos. El sistema implementado está basado en el modelo desarrollado por el IPO y posee una configuración prosódica jerárquica (Garrido *et al*. 2000). Se dispone en la actualidad de modelos predictivos para modalidades oracionales distintas en varias lenguas y geolectos.

4. Los métodos utilizados para obtener el corpus de síntesis

El corpus de síntesis consiste, como se ha indicado en el apartado 2, en un conjunto de enunciados que contiene todas las unidades necesarias para llevar a cabo la conversión de texto en habla en una determinada lengua. Antes de nada es preciso determinar las características que deben tener estos enunciados, para proceder después a su grabación por parte del locutor seleccionado.

4.1. La selección de unidades y la creación de la base de datos de síntesis

Para construir los enunciados empleados posteriormente en la grabación, se recogió, para cada una de las lenguas, un corpus de 3 millones de palabras extraídas de diversas fuentes. Se han elegido textos periodísticos, literarios, culturales, científicos, económicos, didácticos e institucionales, con el fin de reflejar, de la forma más

Llisterri, J., Machuca, M. J., Madrigal, N., Mancini, F., Massimino, P., Mota, C., et al. (2004).
Aspectos lingüísticos en el diseño de un conversor de texto en habla en castellano y en catalán: El sistema Loquendo TTS®.
In VI Congreso de Lingüística General. (pp. 521-2). Santiago de Compostela.
http://liceu.uab.cat/~joaquim/publicacions/Llisterri_et_al_04_Conversor_Texto_Habla_Castellano_Catalan_Loquendo.pdf

equilibrada posible, la gama de temas y tipos textuales que deberá ser capaz de leer el conversor.

Aplicando un algoritmo a estos textos, se selecciona, en una primera etapa, una serie de secuencias cuya composición se acerque en la medida de lo posible al equilibrio fonético en lo que se refiere a la aparición de cada una de las unidades de síntesis en todos los contextos posibles. Mediante una edición manual, se modifica, si es necesario, la estructura gramatical de estas secuencias y se ajusta el equilibrio fonético hasta llegar a un corpus constituido por unas 2000 oraciones, enunciativas e interrogativas, para cada lengua.

El procedimiento utilizado permite al sistema LoquendoTTS generar enunciados usando unidades de tamaño diferente, ya que la técnica empleada para la conversión consiste en la selección de las unidades óptimas a partir de un corpus o base de datos (*unit selection)*. Esta técnica permite localizar en el corpus las unidades más largas posibles manteniendo su prosodia, como se ha indicado en el apartado 3.7.

4.2. La selección de locutores y la grabación del corpus de síntesis

Las sesiones de grabación tienen como finalidad obtener un corpus de voz homogéneo que contenga las unidades fónicas necesarias para desarrollar el sintetizador en una variante lingüística. Se precisa tanto la selección de locutores adecuados como el control lingüístico del proceso de grabación.

Para que la voz de un locutor sea seleccionada debe resultar perceptivamente clara y armónica, no puede dar la impresión de monotonía ni presentar cambios excesivos o bruscos de tono y de intensidad. Se tiene en cuenta también la variedad lingüística materna del locutor, así como su dicción y su plasticidad, que se evalúan en una sesión de pruebas. Debido a su naturaleza, resulta especialmente conflictiva la realización de las consonantes vibrantes, fricativas y africadas, y la serie de sonidos palatales. En cuanto a las vocales, deben realizarse nítidamente con el timbre que les corresponde, tanto en posición tónica como en posición átona.

La selección de locutores profesionales y bilingües ha resultado de utilidad para desarrollar las voces artificiales también bilingües de Loquendo. Durante las sesiones de grabación, no obstante, hay que controlar que no se den pronunciaciones debidas a interferencias lingüísticas, como la tendencia a pronunciar las vocales átonas del castellano siguiendo las pautas de la reducción vocálica del catalán, que induce cambios significativos de timbre, o la tendencia a modificar el grado de abertura de las vocales abiertas tónicas [ɛ] y [ɔ] del catalán, que pasan a pronunciarse como si fueran cerradas, por influencia del castellano. Deben reflejarse únicamente los procesos fonológicos propios de la lengua, y la pronunciación debe responder a la variante seleccionada, independientemente de que haya otras realizaciones posibles en el uso. La lectura debe realizarse con la entonación conveniente, colocando las pausas en el lugar previsto y manteniendo constante la velocidad de elocución. Hay que evitar también los golpes glotales o nasales en posición inicial absoluta, la tendencia a la nasalización y el uso inconveniente de la voz susurrada y de los suspiros. Durante las sesiones de grabación se controla tanto la locución como el resultado de la transcripción fonética que el sistema ha generado automáticamente, para asegurar la necesaria correspondencia entre una y otra.

Llisterri, J., Machuca, M. J., Madrigal, N., Mancini, F., Massimino, P., Mota, C., et al. (2004).
Aspectos lingüísticos en el diseño de un conversor de texto en habla en castellano y en catalán: El sistema Loquendo TTS®.
In VI Congreso de Lingüística General. (pp. 521-2). Santiago de Compostela.
http://liceu.uab.cat/~joaquim/publicacions/Llisterri_et_al_04_Conversor_Texto_Habla_Castellano_Catalan_Loquendo.pdf

5. Los procedimientos empleados para evaluar el sistema Loquendo TTS

La adaptación de los módulos y componentes del sistema de síntesis Loquendo TTS comentados ha dado lugar a la creación de varias voces en castellano y en catalán, tanto femeninas como masculinas. Una vez obtenida una primera versión del sistema, se han realizado dos pruebas, una objetiva y otra subjetiva, para evaluar su funcionamiento y corregir posibles errores en la elaboración de una versión definitiva. Las voces consideradas han sido las de Juan y Carmen para el castellano, y la de Montserrat para el catalán.

5.1. La prueba objetiva

La finalidad de la evaluación objetiva es comprobar si las formas lingüísticas presentes en un texto son convenientemente procesadas por el sistema. Si no es así, se indica cuál es el origen del problema y se sugieren, en la medida de lo posible, pautas para solventarlo. Se ha llevado a cabo, por un lado, una evaluación de la interpretación lingüística que el conversor realiza de las unidades de preprocesamiento, y por otro, una revisión de la lectura de textos de procedencia diversa realizada por el sistema, para poder detectar así problemas relacionados con todos los módulos que contienen información lingüística.

En el primer caso, para evaluar el preprocesamiento se ha constituido un corpus formado por extranjerismos, números, fechas, expresiones matemáticas, abreviaturas y siglas. En el segundo caso, se han empleado 31 textos escritos en lengua catalana para evaluar la voz de Montserrat y 31 en lengua castellana para evaluar la voz de Carmen. Con el fin de obtener un corpus equilibrado, los textos se han seleccionado teniendo en cuenta tanto su duración una vez oralizados (1 minuto aproximadamente) como el tema. Prácticamente todas las oraciones resultantes son enunciativas, con escasa presencia de secuencias de otras modalidades oracionales.

El resultado de la evaluación objetiva de las voces femeninas del conversor en castellano y en catalán indica que se producen todavía algunos errores debidos a problemas derivados del tratamiento de las unidades, del preprocesamiento y de la transcripción. En el sistema del castellano los errores más frecuentes se deben al preprocesamiento, mientras que en el sistema del catalán se deben especialmente al funcionamiento de las unidades seleccionadas por el conversor. En general, los errores de transcripción son una minoría en ambas lenguas.

5.2. La prueba subjetiva

Mediante las pruebas subjetivas se pretende evaluar la calidad, la naturalidad y la comprensión de cada una de las voces implementadas en el sistema de síntesis al generar la voz femenina del catalán (Montserrat) y las voces masculina (Juan) y femenina (Carmen) del castellano. Para evaluar la voz de Montserrat se han seleccionado 31 textos escritos en lengua catalana (los mismos empleados en la evaluación objetiva) y para evaluar la voz de Juan y la de Carmen se han escogido 16 en lengua castellana (parte de los usados en la evaluación objetiva).

Llisterri, J., Machuca, M. J., Madrigal, N., Mancini, F., Massimino, P., Mota, C., et al. (2004).
Aspectos lingüísticos en el diseño de un conversor de texto en habla en castellano y en catalán: El sistema Loquendo TTS®.
In VI Congreso de Lingüística General. (pp. 521-2). Santiago de Compostela.
http://liceu.uab.cat/~joaquim/publicacions/Llisterri_et_al_04_Conversor_Texto_Habla_Castellano_Catalan_Loquendo.pdf

Las pruebas de percepción se han llevado a cabo en 30 sujetos de lengua dominante castellana para la evaluación de las voces castellanas y en 30 de lengua dominante catalán para la evaluación de la voz en catalán. Las personas seleccionadas no tenían experiencia en evaluaciones de habla sintetizada. Todas ellas escucharon uno de los textos en su lengua dominante para valorar hasta qué punto les resultaba agradable la voz que escuchaban. Algunos realizaron la misma prueba, además, en la otra lengua, para así poder comparar sus preferencias en lenguas distintas.

Tanto los hablantes bilingües de dominancia castellana como los de dominancia catalana se muestran críticos en lo que respecta a la "agradabilidad" de las voces femeninas del castellano catalán y del catalán (Carmen y Montserrat). Se observa en ellos, además, una cierta tendencia a valorar de forma más estricta el sistema que se expresa en la lengua que mejor dominan. Si se comparan las dos voces del castellano, Carmen y Juan, los sujetos consideran más agradable la voz de Carmen, aunque las diferencias no son estadísticamente significativas.

A partir del análisis de los resultados, se puede concluir que aunque los sujetos hayan detectado en las voces de los conversores del castellano y del catalán unas leves carencias en la calidad con respecto a las voces naturales, este hecho no repercute en la comprensión de los textos leídos por el conversor. Hay que añadir que no se han llevado a cabo pruebas comparativas con otros sistemas de síntesis basados en las mismas técnicas.

6. Conclusiones

En el presente trabajo se ha intentado mostrar qué tipo de conocimiento lingüístico se integra en un conversor de texto en habla, partiendo de la experiencia adquirida en el trabajo conjunto entre Loquendo SpA y la Universitat Autònoma de Barcelona. Esta colaboración pone de manifiesto que es posible aunar en un proyecto común perfiles académicos aparentemente dispares, siempre y cuando los equipos sepan comprender las necesidades específicas del sistema y logren adaptar sus modelos, métodos y herramientas a los requisitos que impone el desarrollo de un producto, sin renunciar por ello a la calidad científica. Desde el punto de vista de la lingüística, las tareas desarrolladas han requerido un esfuerzo de sistematización de fenómenos conocidos y, en algunos casos, de recopilación y ordenación de nuevos datos o de planteamiento de problemas desde una óptica diferente a la habitual en los estudios filológicos. Se han revisado las descripciones fonéticas y fonológicas de la lengua, se han abordado las cuestiones gramaticales teóricas que necesariamente subyacen a la formulación de reglas lingüísticas implementables y se han contrastado las propuestas existentes para decidir el modelo de pronunciación que debe seguir el sistema. Este trabajo enriquece la perspectiva con la que se aborda el análisis lingüístico, ya que, de forma natural, repercute en nuevas investigaciones y en una visión más amplia y actualizada de la disciplina.

Llisterri, J., Machuca, M. J., Madrigal, N., Mancini, F., Massimino, P., Mota, C., et al. (2004).
Aspectos lingüísticos en el diseño de un conversor de texto en habla en castellano y en catalán: El sistema Loquendo TTS®.
In VI Congreso de Lingüística General. (pp. 521-2). Santiago de Compostela.
http://liceu.uab.cat/~joaquim/publicacions/Llisterri_et_al_04_Conversor_Texto_Habla_Castellano_Catalan_Loquendo.pdf

7. Referencias bibliográficas

Balestri, Marcello, Alberto Pacchiotti, Silvia Quazza, Pier Luigi Salza & Stefano Sandri (1999): "Choose the best to modify the least: a new generation concatenative synthesis system". *Eurospeech'99. 6th European Conference on Speech Communication and Technology. Budapest. v*ol. 5, 2291-2294.

Garrido, Juan María, Isabel Ortín, Silvia Quazza, Pier Luigi Salza & Franca Mancini (2000): "Desarrollo de un módulo de asignación de parámetros prosódicos para la versión en español del sistema de conversión texto-habla ACTOR®". *Procesamiento del Lenguaje Natural*. 26. 183-190. http://www.sepln.org/revistaSEPLN/revista/26/garrido-alminana.pdf [Consulta: 23/04/2004].

Llisterri, Joaquim (2003): "Las tecnologías del habla: Entre la ingeniería y la lingüística". *Actas del Congreso Internacional "La Ciencia ante el Público. Cultura humanística y desarrollo científico y tecnológico"*. Salamanca: Instituto Universitario de Estudios de la Ciencia y la Tecnología. 44-67 [CD-ROM.]
http://liceu.uab.es/~joaquim/publicacions/TecnolHab_Salamanca_02.pdf [Consulta: 23/04/2004].

Llisterri, Joaquim, Carme Carbó, María Jesús Machuca, Carme de la Mota, Montserrat Riera & Antonio Ríos (2003a): "El papel de la lingüística en el desarrollo de las tecnologías del habla". *VII Jornadas de Lingüística*. Cádiz: Servicio de Publicaciones de la Universidad de Cádiz. En prensa.
<http://liceu.uab.es/publicacions/Linguistica_TH_Cadiz02.pdf> [Consulta: 23/04/2004].

Llisterri, Joaquim, Carme Carbó, María Jesús Machuca, Carme de la Mota, Montserrat Riera & Antonio Ríos (2003b): "La conversión de texto en habla: aspectos lingüísticos". *Tecnologías del texo y del habla* ed. por M. A. Martí & J. Llisterri. Barcelona. Servei de Publicacions de la Universitat de Barcelona – Fundación Duques de Soria. En prensa.
<http://liceu.uab.es/publicacions/Linguistica_CTH_FDS02.pdf> [Consulta: 23/04/2004].

Quazza, Silvia & Henk van den Heuvel (2000): "The use of lexica in text-to-speech systems". *Lexicon Development for Speech and Language Processing* ed. por F. van Eynde & F. Gibbon. Dordrecht: Kluwer Academic Publishers. 207-234.

Quazza, Silvia, Laura Donetti, Loreta Moisa & Pier Luigi Salza (2001): "Actor®: a multilingual unit-selection speech synthesis system". *Proceedings of 4th ISCA Tutorial and Research Workshop on Speech Synthesis 2001*. Edimburgh.
<http://www.ssw4.org/system_papers/actor.pdf> [Consulta: 23/04/2004].