TRANSCRIPCIÓN FONÉTICA EN UN ENTORNO PLURILINGÜE

Tatyana Polyákova, Antonio Bonafonte

Universitat Politècnica de Catalunya (UPC), Barcelona, Spain

RESUMEN

España es un país plurilingüe, lo cual es sin duda una gran riqueza, pero a veces también es una dificultad añadida para las tecnologías del habla. Cada vez más las aplicaciones de voz tienen que ser adaptadas al ámbito multilingüe, ya que se pretende que tengan la mayor difusión y utilidad posibles.
La UPC está participando en el proyecto AVIVAVOZ cuyo objetivo es crear un sistema completo de traducción de voz a voz capaz de realizar traducciones entre las lenguas oficiales del Estado español. Nuestro grupo de síntesis tiene como objetivo conseguir que el sistema lea correctamente textos con muchas palabras de otras lenguas. Con ese fin hemos desarrollado sistemas de identificación de la lengua y de conversión fonética específicos para cada lengua en cuestión, junto a una serie de reglas de nativización. Los resultados obtenidos para la identificación de la lengua son buenos tanto para párrafos como para palabras aisladas (como nombres propios). La calidad de síntesis fue mejorada, utilizando sistemas de conversión fonética específicos para cada lengua y aplicando reglas de nativización.

1. INTRODUCCIÓN

En la era de la globalización y del multilingüismo nos encontramos con un abanico de nombres propios de diversos orígenes en todo tipo de ámbitos, como por ejemplo conversaciones, noticiarios, prensa, etc. La globalización de los medios de comunicación, la interacción económica a nivel mundial, el emergente desarrollo de tecnologías de alta gama, sin mencionar la movilidad internacional de recursos humanos, aportan una gran variedad lingüística a nuestro día a día. Los textos multilingües ya no sorprenden a nadie, pero, sin embargo, presentan una dificultad para muchos lectores. En los países de habla no sajona, el uso de anglicismos es creciente. En España cada vez más podemos oír nombres y apellidos procedentes del mundo entero.
El problema de la pronunciación de textos multilingües se puede descomponer en dos partes: (i) mejora de pronunciación de nombres propios, y (ii) mejora de pronunciación de palabras y frases extranjeras dentro del texto en otro idioma. Los nombres propios, sin embargo, son difíciles de pronunciar incluso para los humanos, dado que su pronunciación depende de una serie de factores tales como nivel de asimilación fonética y ortográfica, la popularidad de los mismos en un contexto dado, y también de las preferencias individuales de la persona portadora del nombre [7].
En [1], Font Llitjós probó que el hecho de saber el origen del nombre propio, su pertenencia a una de las familias lingüísticas o ambos, ayuda a mejorar su transcripción fonética.
El identificador de la lengua presentado por los autores consistía en modelos n-grama entrenados para cada lengua a partir de una base de datos que incluía los marcadores de principio y final de palabra. Para cada palabra de entrada, se buscaban todos los trigramas y se estimaba su probabilidad de pertenecer a cada una de las lenguas, de modo que la lengua con mayor probabilidad se elegía como lengua de origen de la palabra. La información de la lengua de origen se podía incorporar de manera directa o indirecta. La manera directa consistía en entrenar un sistema de conversión fonética independiente para cada lengua, mientras la indirecta permitía más flexibilidad incluyéndola sólo como característica adicional en un clasificador, usándola en los casos cuando era relevante. El objetivo en [1] era pronunciar todos los nombres propios correctamente, desde el punto de vista de la gramática americana, o, en otras palabras, americanizarlos. Los n-gramas también se usaron en [3] y [4] para identificar la lengua. En [3], aparte de los n-gramas, se propuso usar clústeres de letras basados en sílabas, siendo éstas unidades estables que contienen más información lingüística que las letras.
Los objetivos de este trabajo consisten en mejorar la transcripción fonética de palabras extranjeras y de nombres propios de origen extranjero. Con el fin de alcanzar un alto nivel de inteligibilidad del habla sintetizada proponemos adaptar la pronunciación de la palabra extranjera a la lengua del texto, sabiendo que la voz nativizada es más fácil de entender que la voz sintetizada a partir de fonemas extranjeros [5]. Para España este es un problema de gran importancia, debido a la existencia de al menos tres grandes regiones bilingües, donde las lenguas específicas de la región se usan con la misma naturalidad (frecuencia) y en los mismos ámbitos que el castellano. Nos vamos a centrar en lo relacionado con Cataluña.
En catalán, igual que en castellano, las palabras extranjeras tienden a adquirir una pronunciación "nativizada" basándose en los correspondientes conjuntos de fonemas. No obstante, los nombres españoles en catalán, la mayoría de las veces, se pronuncian con fonemas castellanos. Por ejemplo, el nombre Jorge en catalán se lee /x 'o r x e/ y no /dZ 'o r dZ e/ como sugerirían las reglas de fonética catalana, a pesar de que el fonema /x/ no existe en catalán.
Sirva como ejemplo ilustrativo el fenómeno del "Spanglish", cuando un hablante bilingüe de español e inglés en EE.UU. cambia de un idioma a otro de forma espontánea, en mitad de una frase, con una pronunciación impecable en cada lengua. Asimismo, en Cataluña se pueden oír frases bilingües y existen textos escritos en castellano con un gran porcentaje de nombres propios o palabras catalanas y viceversa. Tampoco podemos olvidar las palabras inglesas cuyo uso es elevado en todos los ámbitos. El fenómeno de multilingüismo, tan frecuente en la prensa, foros, programas de televisión, correos electrónicos, SMS, etc., necesita una atención especial. Se ha de precisar que en las sociedades bilingües españolas, se da más importancia en los medios de comunicación a la lengua "autonómica", de modo que existen más textos en castellano con muchas palabras catalanas que al revés.
Los siguientes ejemplos ilustran diferentes ocurrencias del multilingüismo. En la Figura 1 tenemos un fragmento de un foro de discusión de noticias del diario "AVUI" editado en

catalán. El primer comentario está en catalán e inglés, mientras el segundo está en castellano.

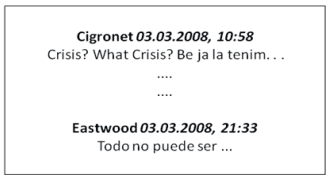

Figura 1. *Comentarios multilingües en el diario "AVUI".*

A continuación tenemos otro ejemplo de un texto en castellano con muchos nombres propios ingleses.

Figura 2. *Fragmento de un reportaje sobre las elecciones primarias en EE.UU.*

Existen muchos más ejemplos donde la transcripción fonética depende de si se conoce o no la lengua de la palabra.

2. SISTEMA DE TRASCRIPCIÓN FONÉTICA MULTILINGÜE

Tener una herramienta capaz de determinar la lengua del párrafo es muy importante a la hora de tratar información multilingüe procedente de diarios, foros, correos electrónicos, artículos científicos, manuales técnicos, páginas web, y otras fuentes donde la lengua del párrafo puede no saberse a priori o puede cambiar de forma repentina. En casos como esos, sabiendo la lengua podemos mejorar la calidad de síntesis considerablemente.

Una vez la lengua del párrafo quede determinada, es necesario identificar la lengua de cada una de las palabras de forma aislada. Esto es importante por dos razones: (i) para mejorar la pronunciación de las palabras extranjeras, y (ii) para la adaptación de esa pronunciación a la lengua del párrafo. Noticias internacionales y deportivas abundan en nombres propios de diversos orígenes mientras en los foros, comentarios en las páginas web, correos electrónicos, posters de publicidad de líneas aéreas (p.ej. Vueling), SMS, etc., hay mucha mezcla lingüística y la probabilidad de encontrar una palabra española o catalana en un texto en inglés, o al revés, es muy alta.

El diagrama abajo muestra nuestro sistema de identificación de la lengua y de nativización de la pronunciación. El primer paso del método consiste en determinar la lengua de cada párrafo. Pongamos que sea la lengua por defecto "LANG" o llamémosla "lengua destino".

La búsqueda de la transcripción fonética se hace para cada palabra de forma aislada. Primero se averigua si la palabra, incluyendo POS, está en el diccionario de la lengua destino. Si es así, la correspondiente transcripción fonética se considera válida; si no, el siguiente paso consiste en averiguar si la lengua de la palabra es distinta de la lengua destino del texto. Para ello, se efectúa una llamada al módulo identificador de la lengua.

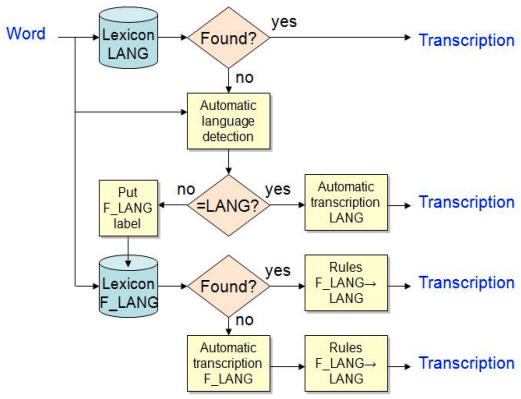

Figura 3. *Sistema de transcripción fonética multilingüe.*

Si efectivamente resulta ser distinta, la palabra se etiqueta con la lengua origen proporcionada por el identificador, F_LANG. En caso contrario, la pronunciación se deriva usando el sistema automático de transcripción fonética para la lengua destino. A continuación, la transcripción fonética de las palabras con la etiqueta de una cierta lengua origen se buscan en el diccionario de esa lengua. Antes de validar la pronunciación, se aplican las reglas de "nativización" entre la lengua origen y la lengua destino. Si la palabra no se encuentra en el diccionario de la lengua origen, se procesa con el sistema automático de transcripción fonética de esa lengua y luego, igual que antes, se aplican las reglas de "nativización". El proceso de nativización se explica en la sección 3.3. Es importante enfatizar que si la palabra es de origen extranjero pero se encuentra en el diccionario de la lengua destino, se considera nativizada. Por eso no hay necesidad de identificar la lengua antes del primer paso.

2.1. Identificación de la lengua

Para identificar la lengua hemos implementado el modelo n-grama estándar, previamente utilizado para la misma tarea en [1, 3, 4]. Se estima un n-grama para cada lengua en cuestión. Los n-gramas incluyen los marcadores del principio <s> y final <s/> de la palabra: <s>mi<s/><s>casa<s/>.

$$Lng^{*} = \arg\max_{Lng} p\left(l_1 \ldots l_n \middle| Lng\right)$$

La lengua del párrafo se determina teniendo en cuenta la lengua de cada una de las palabras aisladas.

2.2. Trascripción fonética

En nuestro anterior trabajo [6] obtuvimos una tasa de palabras correctas igual a 79.63%, utilizando traductores de estados finitos [8] en combinación con el método de aprendizaje a partir de errores [9], para el diccionario LC-STAR de inglés americano [10]. Los traductores de estados finitos constituyen un método poco costoso en tiempo pero eficaz en cuanto a resultados, lo que justifica su utilización para obtener la pronunciación de palabras desconocidas en inglés. La mayoría de los métodos necesitan un diccionario alineado para entrenamiento. En este caso el alineamiento es similar al de [11]. Para trascripción fonética de castellano y catalán

utilizamos transcriptores basados en reglas, ambos desarrollados en la UPC.

2.3. Nativización

La nativización es un proceso de adaptación de la pronunciación de la lengua origen a la pronunciación más cercana, y al mismo tiempo correcta, en la lengua destino.
Es importante destacar que hay una gran diferencia entre el habla nativizada y el habla no nativa (o con acento extranjero) [5]. En muchos aspectos el habla nativizada tiene ventajas sobre la no nativa. El habla no nativa se diferencia del habla nativa en los puntos articulación, distribución de la pausas, elección de las palabras, conducta en las fronteras entre palabras, errores de pronunciación, existencia de fonemas suspendidos entre pronunciación correcta e incorrecta, etc., mientras que el habla nativizada conserva el punto de articulación de la lengua origen, no implica errores de pronunciación, ya que sólo pretende moldear de la mejor manera la pronunciación de la palabra para que encaje suavemente en las oraciones en la lengua destino. La transcripción nativizada está basada en reglas concretas y no da lugar a fonemas medio-nativos o mal pronunciados.
Se ha trabajado en la síntesis tanto en catalán como en castellano, dado que se disponía de sintetizadores en estas dos lenguas. Antes de llevar a cabo los experimentos, se definieron los conjuntos de fonemas (*phonesets*), para cada lengua. En el caso del catalán, se definió un nuevo conjunto enriquecido al que se llamó "CAT+", que además de los fonemas propios del catalán contiene 5 fonemas propios del castellano: /x/, /T/, y las tres vocales átonas /a/, /o/ y /e/, que no existen en catalán. Esto fue posible gracias a que los locutores cuyas voces se emplearon en el sintetizador eran perfectamente bilingües. De hecho, es un fenómeno propio de las sociedades bilingües el uso de un *phoneset* que combine los fonemas de las dos lenguas coexistentes, como ocurriría también en el caso del Spanglish. El *phoneset* "CAT+" es ventajoso para la adaptación al catalán de palabras tanto inglesas como procedentes de otras lenguas oficiales del estado español, a que tiene más fonemas en común con cada una de ellas que el *phoneset* básico tanto de catalán como de castellano. En el caso del castellano, en cambio, los locutores cuyas voces se usaron para la construcción del sintetizador solamente dominaban este idioma, de modo que el *phoneset* utilizado fue el estándar.
Se desarrollaron manualmente tablas de nativización para adaptar el inglés americano (EN-US), euskera (EU), castellano (ES), y gallego (GA) al catalán enriquecido (CAT+). Del mismo modo, se desarrollaron tablas para nativizar al castellano los idiomas EN-US, EU, CA (catalán con *phoneset* estándar), y GA. En nuestros experimentos hemos utilizado sólo las tablas correspondientes a EN-US → CAT+, ES → CAT+, CAT → ES, EN-US → ES. Se prevé trabajar con otros pares de lenguas en el futuro.
La nativización se realiza fonema a fonema, utilizando las tablas correspondientes. En los casos ambiguos, donde utilizando la tabla se obtiene un error evidente de manera reiterada, se usa como ayuda la transcripción fonética de la palabra extranjera obtenida para la lengua destino. Por ejemplo, si la palabra en la lengua origen es *talent* con transcripción / t ‘ae l @ n t`/ y la tabla dice que la shwa /@/en inglés pasa a ser una /a/ en castellano, pero la transcripción por reglas en castellano dice que en esa posición hay una /e/, la transcripción nativizada va a tener una /e/.

3. EXPERIMENTOS Y RESULTADOS

Los experimentos se hicieron para la identificación de la lengua de párrafos y de nombres propios. Para la evaluación de la transcripción nativizada se utilizó el sintetizador Ogmios [12] basado en la selección de unidades, entrenado con 10 horas de voz. Se sintetizaron varias frases antes y después de aplicar las reglas de nativización. La calidad de las mismas fue puntuada por oyentes sin experiencia en la síntesis del habla, para obtener una valoración lo menos influenciada posible.

3.1. Descripción de la base de datos.

Para el experimento de la identificación de la lengua del párrafo se consideraron las siguientes lenguas: catalán, castellano, euskera, gallego e inglés.
En el marco del proyecto AVIVAVOZ se desarrollaron varios corpus bilingües para la traducción estadística, disponible para catalán, euskera, gallego y castellano. Las frases y sus traducciones se extrajeron de la revista de consumidores Eroski. La presencia de nombres propios es insignificante en comparación con la de nombres comunes. El modelo de lenguaje para inglés se entrenó a partir de los nombres comunes del diccionario LC-STAR.
Para los experimentos de identificación de la lengua de nombres propios (palabras aisladas), creamos una base de datos de nombres y apellidos gallegos de tamaño de ~3600 palabras, y otra base de datos de nombres y apellidos vascos (alrededor de 11200 palabras). Para castellano y catalán sólo consideramos los nombres propios de personas, marcados como tales en los diccionarios de nombres propios creados en el marco del proyecto LC-STAR para estas lenguas [10] (alrededor de 20-27 mil palabras). Hay que tener en cuenta que cada diccionario fue filtrado para eliminar los nombres presentes en cualquier otro diccionario, de modo que todas las entradas son únicas. Eso fue necesario para “pulir” los modelos de los nombres que se escriben igual en varias lenguas, p.ej. David /d a B ‘i D/ en español y /d ‘ae v I d/ en inglés. Los modelos de lenguaje se entrenaron a partir del 90% del corpus Eroski, y el 10% restante fue reservado para la evaluación. Se emplearon los mismos porcentajes para entrenar los modelos de nombres propios.

3.2. Resultados de identificación de la lengua.

La Tabla 1 es una tabla de confusión para las cuatro lenguas oficiales del Estado español (las comunidades autónomas) y además el inglés.

Lang. es	ca	es	eu	ga	en
ca	95.2	1.2	2.1	1.5	-
es	1.9	90.9	3	0.1	-
eu	0.6	0.8	92.6	8	-
ga	1.9	0.8	2.1	89.2	-
en	0.6	0.3	0.3	1.2	-

Tabla 1. *Resultados de identificación de la lengua del párrafo para el corpus Eroski.*

Los mejores resultados fueron obtenidos para catalán y euskera, y los peores para el gallego, lo que puede deberse a su alto grado de parecido con el castellano y catalán, que a su vez tienen algunas palabras en común.

La mayoría de los errores se deben a los siguientes factores:
- Presencia de palabras extranjeras como Kodak, Kellogg's, Fuji, etc.
- A las palabras y fragmentos de frases que pueden pertenecer a varias lenguas al mismo tiempo como "agua mineral natural".
- Dígitos y abreviaciones: 10g, rayos UVA, etc.

En la Tabla 2 se muestran los resultados de identificación de la lengua de nombres propios para 200 palabras elegidas aleatoriamente dentro del corpus de evaluación.

Languages	ca	es	eu	ga	en
ca	77	33	5	25	9
es	46	98	12	29	11
eu	6	12	170	6	3
ga	37	24	6	121	9
en	31	33	7	19	168

Tabla 2. *Resultados de identificación de la lengua usando modelos de lenguaje "pulidos".*

Aquí los mejores resultados se obtuvieron para el euskera, seguidos por el inglés y el gallego.

3.3. Test perceptual

Para validar la metodología se ha realizado una primera evaluación de las reglas de nativización mediante un test perceptual. Primero se crearon dos corpus multilingües, cada uno de los cuales constaba de 1000 palabras, uno en castellano con ~50 palabras inglesas y ~50 catalanas y otro en catalán con ~ 50 palabras españolas y ~50 inglesas.

Dado que los oyentes que participaron en la evaluación del sistema tenían conocimientos de castellano, catalán e inglés pero no de euskera ni gallego, se diseñó un test para palabras procedentes de los tres primeros idiomas. Para evaluar la inteligibilidad de la síntesis en catalán y castellano siguiendo el esquema presentado en la Figura 3, se eligieron aleatoriamente y se sintetizaron 10 frases, 5 de cada corpus, Las etiquetas F_LANG fueron añadidas a mano. Se evaluaron la inteligibilidad y la naturalidad de las frases sintetizadas. Entre 10 oyentes, para las 5 frases en castellano, el 70% consideró las frases nativizadas más naturales el 62%, más inteligibles, el 16 % no notó diferencia significativa en la naturalidad y el 28% en la inteligibilidad. El 14% de los oyentes opinó que las frases nativizadas eran menos naturales y el 12% que eran menos inteligibles que la síntesis básica.

Para 5 frases en catalán de los mismos diez oyentes, el 44% opinó que la síntesis nativizada sonaba más natural que la básica, el 36% que sonaba igual de natural que la otra y el 20% que sonaba peor. En cuanto a la inteligibilidad el 26% de oyentes decidió que se entendía mejor, el 48% que se entendía igual, y el 28% que se entendía peor. Los resultados para castellano se pueden considerar bastante buenos. El hecho de que muchos oyentes prefirieron la síntesis básica en catalán se debe principalmente a la ausencia de algunos difonemas en la voz en catalán, que aunque la transcripción fonética fuera la adecuada, hicieron que aparecieran algunos artefactos muy molestos al oído, a su vez afectando a la opinión de los oyentes. Otras fuentes de errores son el alineamiento de los diccionarios de entrenamiento, que a veces introduce ambigüedad, y las reglas de nativización, que en el caso de palabras con comportamiento excepcional no siempre dan los resultados esperados.

4. CONCLUSIONES

De los resultados descritos en el apartado anterior se puede concluir que el sistema de identificación de lengua funciona bien para párrafos y ligeramente peor para palabras aisladas. Una evaluación preliminar indica que sabiendo la lengua y usando el módulo de transcripción fonética multilingüe se consigue mejorar la inteligibilidad y la naturalidad del habla sintetizada en castellano y en algunos casos en catalán. En el futuro se trabajará en la integración del módulo de identificación de la lengua con el módulo de nativización de la pronunciación.

5. AGRADECIMIENTOS

Este trabajo ha sido financiado con el proyecto AVIVAVOZ (TEC2006-13694-C03) y la beca FPU (AP2005-4526).

6. BIBLIOGRAFÍA

[1] Llitjós, A., and Black, A., "Knowledge of language origin improves pronunciation of proper names", Proceedings of EuroSpeech-01, 1919-1922, 2001.

[2] Lewis, S., and McGrath, K., "Language identification and language specific letter-to-sound rules", Colorado Research in linguistics, 17 (1), 1-8, 2004.

[3]. Chen, Y., You, J. Chu. M., Zhao, Y., Wang, J., "Identifying language origin of person names with n-grams of different units", ICASSP 2006, Toulouse, France

[4] Litjós, A., "Improving pronunciation accuracy of proper names with language origin classes", Master Thesis, CMU-LTI-01-169, Carnegie Mellon University, Aug 2001.

[5] Schultz, T. and Kirchhoff, K., "Multilingual speech synthesis", Elseiver, USA 2006

[6] Polyákova, T., Bonafonte, A., "Learning from errors in grapheme-to-phoneme conversions", International Conference on Spoken Language Processing, Pittsburg, USA, 2006

[7] Lindström, A., "English and other foreign linguistics elements in spoken Swedish", Linköping University Linköping, 2004

[8] Galescu L., J. Allen, "Bi-directional Conversion Between Graphemes and Phonemes Using a Joint N-gram Model", In Proc. of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Perthshire, Scotland, 2001

[9] Brill E., "Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging", Computational linguistics 21(4), pp. 543-565, 1995

[10] http://www.lcstar.org

[11] Damper R. I., Marchand Y., Marsterns J.-D. and Bazin A., "Aligning letters and phonemes for speech synthesis" in Proceedings of the 5thISCA Speech Syntesis Workshop, Pittsburgh, 209-214., 2004

[12] Bonafonte A., Adell J., Agüero, Erro D., Esquerra I., Moreno A., Pérez J., Polyákova T., "The UPC TTS System Description for the 2007 Blizzard Challenge", 6th ISCA Workshop on Speech Synthesis, Bonn, Germany, August '07.