Transcriptor ortográfico-fonético para el castellano*

María José Castro†, Salvador España†, Andrés Marzal‡ e Ismael Salvador†

mcastro@dsic.upv.es, sespana@dsic.upv.es, amarzal@lsi.uji.es, issalig@doctor.upv.es

† Departament de Sistemes Informàtics i Computació,
Universitat Politècnica de València, Valencia, España

‡ Departament de Llenguatges i Sistemes Informàtics,
Universitat Jaume I, Castelló, España

Resumen El trabajo describe un sistema para transcribir automáticamente texto ortográfico en idioma español en una cadena de símbolos de tipo fonético. El transcriptor ortográfico-fonético se basa en una serie de reglas que indican cómo se deben transcribir los grafemas en unidades de tipo fonético atendiendo al contexto en que se presentan. La utilidad de este transcriptor ortográfico-fonético es, principalmente, el entrenamiento de sistemas de reconocimiento de voz. Se ha incluido la opción de pronunciaciones múltiples (posibilidad de que un sonido no se pronuncie o se pronuncie de diferentes formas). Finalmente, se ha desarrollado una herramienta para crear modelos léxicos de un sistema de reconocimiento automático del habla.

1 *Introducción*

El objetivo del presente trabajo es el diseño de un transcriptor ortográfico-fonético de fácil manejo para el idioma español. Su uso principal es el entrenamiento de sistemas de reconocimiento del habla, si bien también encuentra aplicación en síntesis de voz. El transcriptor se basa en una serie de reglas que indican cómo se deben transcribir los grafemas en unidades de tipo fonético atendiendo al contexto en que se presentan.

El objetivo de un sistema de reconocimiento automático del habla es convertir una señal acústica en una secuencia de palabras. Los sistemas de síntesis de voz, por su parte, tienen como entrada una secuencia de palabras escritas que se convierten en un mensaje acústico. Una parte importante de los dos tipos de sistemas es cómo se pronuncian las palabras en términos de unidades de habla individuales. Así, un sistema de reconocimiento de voz necesita la pronunciación de cada palabra que debe reconocer, y un sistema de síntesis necesita la pronunciación de cada palabra que puede generar.

Por lo tanto, si el objetivo de la transcripción es utilizarla para síntesis de habla, una transcripción detallada en alófonos (con o sin archifonemas) resultará adecuada, pues el objetivo de la síntesis es la generación de pronunciaciones lo más naturales posible.

Si el objetivo de la transcripción es su utilización para análisis del habla, sería necesaria, en principio, una transcripción *estrecha* que reflejara las realizaciones concretas de cada uno de los sonidos en sus variaciones alofónicas. Sin embargo, en tareas concretas de reconocimiento automático del habla, con conjuntos de datos de entrenamiento limitados, puede no resultar adecuado el tener que representar y reconocer todos los sonidos con ese nivel de detalle. En particular, experimentaciones previas para tareas semánticamente restringidas en el idioma español [9, 5] nos han decidido a utilizar un conjunto de unidades de tipo fonético (que, abusando del lenguaje, llamaremos fonemas).[1] El conjunto de unidades de tipo fonético que genera el transcriptor se ilustra en la tabla 1. Dado que los caracteres estándar del *Alfabeto Fonético Internacional* [7] no se ajustan al juego de caracteres ASCII nos hemos visto en la necesidad de cambiar su representación. Hemos procurado en la medida de lo posible que los caracteres seleccionados fueran lo más parecidos gráficamente a sus equivalentes fonéticos estándar, y/o se asemejen a

* Trabajo subvencionado parcialmente por el proyecto CICYT TIC98/0423-CO6-02 y el contrato 1FD97-2055-C02-01 del gobierno español.

[1] De igual forma, muchos de los sistemas de fonetización creados para síntesis de voz en español suelen trabajar con una transcripción fonética ancha. Se añaden a lo sumo al repertorio de fonemas los alófonos sonoros de los fonemas /b/ y /d/, el alófono sonoro de /g/ y /s/ y las semiconsonantes /j/ y /w/ [6].

Tabla 1: Unidades de salida del transcriptor. Las vocales tónicas se representan con mayúsculas. "0" indica "Sordo" y "1", "Sonoro".

CONSONANTES

	Bilabial		Labio-dental		Linguo-dental		Linguo-interdental		Linguo-alveolar		Linguo-palatal		Linguo-velar	
	0	1	0	1	0	1	0	1	0	1	0	1	0	1
Oclusivas	p	b			t	d							k	g
Fricativas			f				z		s			y	x	
Africadas											c			
Nasales		m								n		h		
Laterales										l		H		
Vibrante simple										r				
Vibrante múltiple										@				

VOCALES

	Anterior	Central	Posterior
Semiconsonantes	j		w
Cerrada	i I		u U
Media	e E		o O
Abierta		a A	

caracteres ortográficos de modo que una persona no tenga excesiva dificultad en leer la salida proporcionada por el transcriptor. Las diferencias del juego de caracteres empleado por el transcriptor con respecto a los símbolos estándar son ⟨z⟩ = ⟨Θ⟩, ⟨h⟩ = ⟨η⟩, ⟨H⟩ = ⟨ʎ⟩, ⟨@⟩ = ⟨r̄⟩, ⟨y⟩ = ⟨j̆⟩.

Adicionalmente, en análisis sigue existiendo un importante problema: en contextos muy determinados o en zonas de la geografía hispano-hablante, algunos sonidos tienden a no pronunciarse en absoluto o a pronunciarse de un modo muy diferente. Nos referimos a los fenómenos asociados al habla espontánea (pronunciación relajada) o a fenómenos regionales tales como el seseo, yeísmo, etc. Una transcripción que no indique explícitamente esta posibilidad conducirá a un mayor número de errores. Debido a ello, se ha incorporado al transcriptor la opción de múltiples pronunciaciones, que incluye la posibilidad de que un fonema no se pronuncie o se pronuncie de más de un modo.

Finalmente, se ha desarrollado una herramienta para crear el léxico de un sistema de reconocimiento automático del habla. Esta herramienta crea un autómata que representa una transcripción (múltiple) para cada entrada de una tarea dada.

2 *Trabajos relacionados*

Existen dos aproximaciones a la conversión de grafemas en fonemas: los sistemas basados en reglas y los métodos inductivos, que intentan aprender automáticamente las reglas fonológicas a partir de ejemplos.

Aunque el castellano es un idioma relativamente transparente en su relación ortográfica-fonética, se necesita algún tipo de conversión tanto en los sistemas de síntesis de voz como en sistemas de reconocimiento. En la tesis de Ríos [6, capítulo 2] se realiza una revisión de los trabajos realizados para el castellano.

En cambio, la transcripción de grafemas ha obtenido mucha más atención en otras lenguas, tales como el inglés. Los sistemas de síntesis utilizan normalmente transcriptores basados en reglas y un diccionario de excepciones (como, por ejemplo, el sistema MITalk [2]), aunque también se han propuesto aproximaciones inductivas, como el sistema conexionista NETtalk de Sejnowski y Rosenberg [8].

3 *El transcriptor ortográfico-fonético*

El transcriptor ortográfico-fonético (herramienta `ort2fon`) ha sido programado en Python. Se puede encontrar una descripción detallada del código en el informe técnico [3].

Además, la versión *bytecode* de la herramienta se encuentra públicamente disponible en `http://www.dsic.upv.es/users/rfia/transcriptor/transcriptor.html`.

3.1 Módulo de acentuación

Las reglas de acentuación se han obtenido, principalmente, de las normas descritas en [1, 4]. Este módulo devuelve el texto acentuado a partir de la representación ortográfica. Es decir, encuentra la sílaba tónica (una sílaba por palabra excepto en los adverbios terminados en *−mente*). Estas vocales tónicas se representan con sus respectivas mayúsculas (por ejemplo, *papel* se transcribe como ⟨papEl⟩).

El proceso de acentuación se ha realizado como si fuéramos hablantes no nativos y tuviéramos que leer una palabra que desconocemos (nunca la hemos oído). Así, sólo disponemos de información gráfica, es decir la presencia o ausencia de tilde. De esta forma se ha aplicado la inversa de las reglas de acentuación (suficientes para poder "leer" una palabra desconocida). Por ejemplo, la regla para detectar las palabras *agudas* es: "una palabra que acaba en consonante, excepto en *n* o *s* se acentúa en la última sílaba".

3.2 Conversión grafema-fonema

Dada la transcripción ortográfica del texto, ya acentuado, se traduce a una transcripción fonética. Las reglas para la conversión de grafemas a fonemas han sido extraídas de [7].

En castellano, existen muchas reglas que son una correspondencia 1-a-1 de grafemas a fonemas. Por ejemplo: *b* siempre se pronuncia como ⟨b⟩, *barco* → ⟨bArko⟩.

Sin embargo, la mayoría de reglas son dependientes del contexto, tales como las reglas asociadas al grafema *c*: *c* se pronuncia como ⟨k⟩ antes de *a*, *o*, *u*; antes de *e*, *i* se pronuncia como ⟨z⟩; y antes del grafema *h* como el fonema africado ⟨c⟩. Algunos ejemplos son: *casa* → ⟨kAsa⟩, *cero* → ⟨zEro⟩, *chino* → ⟨cIno⟩.

Otra característica importante es que las reglas se deben aplicar en un cierto orden. Por ejemplo, es necesario convertir el grafema *ch* antes de eliminar el sonido mudo del grafema *h*, para no procesar, por ejemplo, *che* como ⟨ze⟩.

El transcriptor también trata las concurrencias (secuencias de vocales iguales, que se pronuncian como una sola). Adicionalmente, se soportan los modos de transcripción intra-palabra (palabra-a-palabra) e inter-palabra (frase-a-frase). En el modo intra-palabra, las palabras se transcriben de forma independiente; en el modo inter-palabra, la frase completa se procesa como una única palabra. Así, en el modo intra-palabra *costa azahar* será ⟨kOsta azAr⟩ y ⟨kOstazAr⟩ en modo inter-palabra.

4 Pronunciaciones múltiples

Como se ha comentado anteriormente, algunas frases se pueden pronunciar de formas diversas, conduciendo a más de una transcripción fonética. Estas variaciones dependen de varios tipos de fenómenos, tales como variaciones regionales, velocidad y tipo de pronunciación (lectura, habla espontánea, ...). El transcriptor recoge, como principales fenómenos asociados a pronunciaciones múltiples, los siguientes [7]:

- Borrado de sonidos debido a fenómenos de pronunciación relajada. Ejemplo: *abogado* será ⟨abogAdo⟩ pero también ⟨abogAo⟩.
- Inserción de sonidos debido, especialmente, a fenómenos de pronunciación enfática. Ejemplo: *psicólogo* será ⟨sikOlogo⟩ o ⟨psikOlogo⟩.
- Sustitución de una unidad fonética por otra, debido, fundamentalmente, a fenómenos regionales. Por ejemplo:
  - *Seseo* [7, 4]: fenómeno de identificación del fonema asociado a *z* con ⟨s⟩. Ejemplo: pronunciación de *azul* como ⟨asUl⟩.
  - *Yeísmo* [7, 4]: fenómeno de identificación del fonema asociado a *ll* con el fonema asociado a *y*. Ejemplo: pronunciación de *calle* como ⟨kAye⟩.

En determinados casos, los fenómenos de concurrencia entre palabras puede conducir a pronunciaciones múltiples. Por ejemplo, la transcripción de *ciudad de* puede ser ⟨zjudAde⟩ o ⟨zjudAd.de⟩.

Con la opción de transcripciones múltiples se necesita listar todas las variantes. Para evitar un listado de todas las posibilidades, se han utilizado expresiones regulares para indicar elección "(opt1|opt2|...|opt*n*)" y opción "[opt]". Por ejemplo, ⟨abogA[d]o⟩ y ⟨ka(H|y)e⟩. En la primera opción siempre aparece la transcripción "correcta" (la más aceptada como tal).

Adicionalmente, la salida del transcriptor puede incluir etiquetas que marcan el tipo de fenómeno asociado a la salida múltiple. Por ejemplo, la salida informada para *azul* es ⟨a(z|<sso>s)Ul⟩, con la etiqueta "<sso>" para denotar el fenómeno de seseo.

Ejemplos de la salida del transcriptor se pueden consultar en el apéndice A.

5 Conversión transcripción-autómata

Para generar un autómata a partir de la salida del transcriptor ortográfico-fonético se ha implementado un traductor (`fon2lex`), también programado en Python. Esta herramienta facilita el uso de la construcción del léxico de una cierta tarea en sistemas de reconocimiento automático del habla.

La figura 1 muestra un ejemplo del autómata creado para una transcripción múltiple generada por el fenómeno del *yeísmo*.

6 Conclusiones

El objetivo de este trabajo ha sido el diseño de un transcriptor ortográfico-fonético de fácil manejo para el idioma español. El transcriptor está especialmente diseñado para tareas de reconocimiento automático del habla, e incluye la opción de generar pronunciaciones múltiples que pueden mejorar un sistema de reconocimiento. Adicionalmente, debido al formato de salida informado (con etiquetas), es fácilmente adaptable a fenónemos regionales.

Una versión preliminar de este transcriptor (sin la opción de pronunciaciones múltiples) se ha utilizado y probado de forma extensiva en nuestros laboratorios. En un futuro próximo se realizarán experimentos de voz utilizando este transcriptor para medir las mejoras que introduce esta nueva herramienta (reconocimiento utilizando un léxico sin y con pronunciaciones múltiples). Otras mejoras que se plantean incorporar son aspectos de preproceso sobre los datos de entrada (transcripción automática de números, fechas, acrónimos, etc...). Finalmente, se asignarán probabilidades a las pronunciaciones múltiples.

A Ejemplos de transcripción

A continuación se muestra la salida del transcriptor para conjuntos de palabras y frases con diferentes opciones. La opción `-p` indica una transcripción palabra a palabra; `-m`, una transcripción múltiple; `-r`, una transcripción relajada; `-e`, una transcripción con salida informada. En los ejemplos, el texto de entrada se muestra en *itálica* y el de salida sin serif.

A.1 Palabras

Original	-p	-m	-mr
papel	papEl	papEl	papEl
barco	bArko	bArko	bA(r|@)ko
casa	kAsa	kAsa	kAsa
cero	zEro	|s)Ero	(z|s)Ero
chino	clno	clno	clno
che	ce	ce	ce
abogado	abogAdo	abogA[d]o	abogA[d]o
psicólogo	sikOlogo	[p]sikOlogo	[p]sikOlogo
azul	azUl	a(z|s)Ul	a(z|s)Ul
calle	kAHe	kA(H|y)e	kA(H|y)e
cielo	zjElo	(z|s)jElo	(z|s)jElo
cerilla	zerlHa	(z|s)erl(H|y)a	(z|s)erl(H|y)a
abogado	abogAdo	abogA[d]o	abogA[d]o
Madrid	madrld	madrl[(d|t|z)]	madrl[(d|t|z)]
apto	Apto	Apto	A[(p|b)]to
atlas	Atlas	Atlas	Atlas
acta	Akta	Akta	A[(k|g)]ta
casa	kAsa	kAsa	kAsa
caza	kAza	kA(z|s)a	kA(z|s)a
Israel	is@aEl	is@aEl	i[s]@aEl

A.2 Frases

```
python ort2fon.py
```

Insultad directamente, sin tapujos sucios ni mentiras.
insultAdirEktamEnte.sintapUxosUzjosnimentIras.
La abeja picó al abogado sin gran éxito pues es peor que un áspid.
labExapikOalabogAdosingranEksitopwesespeOrkeunAspid.
Ata la jaca a la reja.
AtalaxAkala@Exa.

```
python ort2fon.py -p
```

Insultad directamente, sin tapujos sucios ni mentiras.
insultAd dirEktamEnte . sin tapUxos sUzjos ni mentIras .
La abeja picó al abogado sin gran éxito pues es peor que un áspid.
la abExa pikO al abogAdo sin gran Eksito pwes es peOr ke un Aspid .
Ata la jaca a la reja.
Ata la xAka a la @Exa .

```
python ort2fon.py -m
```

Insultad directamente, sin tapujos sucios ni mentiras.
insultAd[.d]irEktamEnte[.]sin[.]tapUxos[.s]U(z|s)jos[.]ni[.]
mentIras[.]
La abeja picó al abogado sin gran éxito pues es peor que un áspid.
l[a.]abExa[.]pikO[.]al[.]abogA[d]o[.]sin[.]gran[.]Eksito[.]
pwes[.]es[.]peOr[.]ke[.]un[.]Aspid[.]
Ata la jaca a la reja.
Ata[.]la[.]xAk[a.]a[.]la[.]@Exa[.]

```
python ort2fon.py -mr
```

Insultad directamente, sin tapujos sucios ni mentiras.
insultAd[.d]irE[(k|g)]tamEnte[.]sin[.]tapUxos[.s]U(z|s)jos[.]ni[.]
mentIras[.]
La abeja picó al abogado sin gran éxito pues es peor que un áspid.

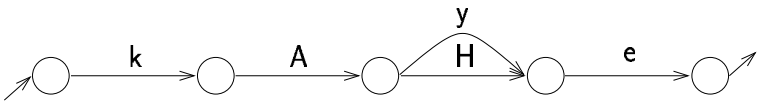

Figura 1: Ejemplo de autómata generado por `fon2lexa` partir de la salida múltiple ⟨kA(H|y)e⟩. La palabra *calle* se puede pronunciar como ⟨kAHe⟩ o ⟨kAye⟩.

```
l[a.]abExa[.]pikO[.]al[.]abogA[d]o[.]sin[.]gran[.]E[(k|g)]sito[.]pwes
[.]es[.]peOr[.]ke[.]un[.]Aspid[.]
Ata la jaca a la reja.
Ata[.]la[.]xAk[a.]a[.]la[.]@Exa[.]

python ort2fon.py -mre
Insultad directamente, sin tapujos sucios ni mentiras.
insultAd[.d]irE[(k|g)]tamEnte[<pe>.]sin[.]tapUxos[.s]U
(z|<sso>s)jos[.]ni[.]mentIras[<pe>.]
La abeja picó al abogado sin gran éxito pues es peor que
un áspid.
l[a.]abExa[.]pikO[.]al[.]abogA[<mr>d]o[.]sin[.]gran[.]E
[(k|g)]sito[.]pwes[.]es[.]peOr[.]ke[.]un[.]Aspid[<pe>.]
Ata la jaca a la reja.
Ata[.]la[.]xAk[a.]a[.]la[.]@Exa[<pe>.]
```

Referencias

[1] La página del idioma Español. `http://www.el-castellano.com/acentos.html`.

[2] J. Allen, S. Hunnicutt, and D. Klatt. *From Text to Speech: The MITalk System.* Cambridge University Press, 1987.

[3] M. J. Castro, S. España, A. Marzal e I. Salvador. Transcriptor ortográfico-fonético para el castellano. Informe técnico DSIC II/37/00, Dep. de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, 2000.

[4] Real Academia Española. *Ortografía de la Lengua Española.* Espasa Calpe, 1999.

[5] I. Galiano, E. Sanchis, I. Torres, and F. Casacuberta. Acoustic-phonetic decoding of Spanish continuous speech. *IJPRAI*, 8(1):155–180, 1994.

[6] Antonio Ríos Mestre. La transcripción fonética automática del diccionario electrónico de formas simples flexivas del español: estudio fonológico en el léxico. *Estudios de Lingüística Española*, 4, 1999.

[7] A. Quilis y J. A. Fernández. *Curso de Fonética y Fonología Españolas para Estudiantes Angloamericanos.* CSIC, Inst. Miguel de Cervantes. Madrid, 1979.

[8] T. J. Sejnowski and C. R. Rosenberg. NETtalk: A parallel network that learns to read aloud. Technical Report 86-01, Dep. of Electrical Engineering and Computer Science, Johns Hopkins Univ., Baltimore, MD, 1986.

[9] I. Torres. Selección de unidades subléxicas para la decodificación acústico-fonética del habla en castellano. Informe técnico DSIC II/25/92, Dep. de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, 1992.