

NAME

Saga – transcribe fonéticamente textos en castellano

SINOPSIS

saga [Opciones] (TextIn | -) [NomOut]

DESCRIPCIÓN

Saga es una herramienta para transcribir fonéticamente textos en castellano. Para ver el modo de funcionamiento basta con ejecutar el programa sin ningún argumento, que mostrará la información de uso más actualizada.

En el momento de escribirse esta pagina, la salida de *saga* sin ningún argumento es:

Empleo:

saga [opciones] (TxtIn | -) [NomOut]

donde:

TxtIn : texto ortográfico de entrada ('-' indica entrada estándar)

NomOut : nombre, sin extensión, de los ficheros de salida (por defecto, salida estándar)

y opciones puede ser:

-L dialecto: Establece todas las opciones para un dialecto (argentina, castilla, chile, colombia, mexico, peru o venezuela)

-f : Transcripción fonética (ext. .fon)

-F : Transcripción en fonemas (ext. .fnn)

-p : Transcripción en fonemas por palabras (ext. .fnp)

-s : Transcripción en semi-sílabas (ext. .sem)

-y : Transcripción en semifonemas (ext. .sef)

-a : Transcribir palabras aisladas

-b : Transcribir cada línea de forma aislada

-Y ExtPal[,FinPal] : Marca de inicio y/o final de palabra

-S : Conservar los silencios en la salida

-d DicExc : Diccionario de excepciones ortográficas

-T TrnPal : Diccionario de transcripción de palabras

-t TrnFon : Diccionario de transcripción de grafemas

-x DicSus : Diccionario de substitución de fonemas

-g DicGrp : Diccionario de substitución de grupos fonéticos

-v NovVoc : Lista de grafemas vocálicos introducidos

-c NovCns : Lista de grafemas consonánticos introducidos

-l NovFon : Lista de fonemas introducidos

-MS : Aplicar seseo

-MX : Transcribir [x] como [ks]

-MH : Transcribir [s] implosiva como [h] siempre

-Mh : Transcribir [s] implosiva como [h] salvo final o enlace entre palabras

-MK : Transcribir [sT] como [ks]

-MA : Transcribir [b], [d] y [g] andinas

-My : Transcribir siempre vocálica la conjunción /y/

-MEb : Eliminar [b] entre vocales y al final de palabra

-MEd : Eliminar [d] entre vocales y al final de palabra

-MEg : Eliminar [g] entre vocales y al final de palabra

-MN : Velarizar las [n] implosivas

-MM : Velarizar todas las nasales implosivas

- M~ : Nasalizar vocales
- MP : Representar [pbdtkgx] implosivas internas como [G]
- MR : Transcribir [r] implosiva como [R]
- M@ : Transcribir [l] y [r] tras [pbd...] como [@l] [@r]
- M: : Marcar consonantes implosivas con [:]
- M_ : Marcar vocales postónicas con [_]
- M. : Marcar con . los fonemas inicial y final de palabra
- MC : Separar las oclusiones (p: pcl p t: tcl t ...)

-e FicErr : Fichero de errores

TxtIn = "-" implica lectura de entrada estándar (teclado). Por defecto, (FonOut no indicado) se escribe en salida estándar (pantalla)

Por defecto se realiza solo la transcripción fonética (ext. .fon)

La transcripción fonética siempre conserva los silencios. Las otras, si no se indica la opción -S, no los conservan.

Los diccionarios de excepciones ortográficas y substitución de grupos fonéticos permiten el uso de '*' al principio y/o final de la palabra

Para encadenar mas de un diccionario, usar: FicDic1+FicDic2...

INVOCACIÓN

saga, sin mas opciones y/o argumentos, proporciona una descripción concisa de las opciones reconocidas y los argumentos necesarios. **saga** reconoce dos posibles argumentos:

TextIn es el fichero de entrada que contiene el texto del cual se desea obtener la transcripción fonética. Si en lugar del nombre de un fichero se indica el carácter '-', el texto de entrada se tomara de la entrada estándar, permitiendo la entrada por teclado o la concatenación en un *pipe*.

El texto de entrada debe estar codificado en UTF-8 y debe estar normalizado, de modo que no haya mayúsculas, números o emoji. Saga ignora las palabras con mayúsculas, dígitos o caracteres especiales.

NomOut es el nombre raíz del fichero o ficheros con la transcripción fonética. El nombre completo se forma concatenando este nombre raíz con la extensión correspondiente al tipo o tipos de transcripción deseados. En el caso de no indicarse **NomOut**, las distintas transcripciones se escribirán en la salida estándar (en general, en la pantalla).

Las opciones que reconoce **saga** se dividen en opciones del tipo de transcripción, modo de realizarla, aplicación de diccionarios y modificadores del comportamiento del transcriptor.

Tipos de Transcripción

saga permite cinco tipos distintos de transcripción. Pueden indicarse uno o mas de ellos en cada invocación al programa. Por defecto solo realiza la transcripción fonética (opción -f). En este tipo de transcripción, cada palabra aparece transcrita en sus fonemas, con estos concatenados sin espacios entre ellos y las silabas separadas por guiones ('-'). El resultado de la transcripción se almacena en el fichero formado por el nombre raíz indicado en *NomOut*, y la extensión .fon. Por ejemplo, si el texto contenido en el fichero *TextIn* es:

mi casa no es blanca

El resultado escrito en *NomOut.fon* sera:

mi k'a-sa n'o 'ez Bl'aN-ka

La opción -F permite obtener la transcripción en fonemas, en la cual cada fonema va separado por espacios, y el resultado se almacena en el fichero *NomOut.fnm*:

mi k 'a s a n 'o 'e z B l 'a N k a

La opción **-p** permite obtener la transcripción en fonemas por palabras, en la cual cada fonema va separado por espacios, pero se incluyen guiones para separar las sílabas, y barras para separar las palabras. El resultado se almacena en *NomOut.fnp*:

mi / k 'a - s a / n 'o / 'e z / B l 'a N - k a

La opción **-s** proporciona en *NomOut.sem* el resultado de transcribir el texto de entrada en semi-sílabas:

mi- -i ka- -'a sa- -a no- -'o e- -'ez Bla- -'aN ka- -a

Finalmente, la opción **-y** proporciona en *NomOut.sef* la transcripción en semifonemas:

.-m m+i m-i i+. .-k k+'a k-'a 'a+s 'a-s s+a s-a a+. .-n n+'o n-'o 'o+. .-'e 'e+z 'e-z z+. .-B B+l B-l l+'a l-'a 'a+N 'a-N N+k N-k k+a k-a a+.

Modos de Transcripción

Por defecto, **saga** realiza la transcripción considerando los efectos de la coarticulación entre palabras. Así, en el ejemplo anterior, la /s/ de la palabras 'es' es sonora ([z] en vez de [s]), y la /b/ de la palabra 'blanca' es aproximante ([B] en vez de [b]). Este comportamiento puede modificarse con la opción **-a** que indica que cada palabra ha de ser transcrita como si estuviera rodeada de pausas:

mi k 'a-sa n 'o 'es bl 'aN-ka

La opción **-b** permite hacer lo mismo pero solo línea a línea. Es decir, considera los saltos de línea equivalentes a una pausa.

La opción **-Y ExtPal[,FinPal]** hace que la marca usada por **saga** en la transcripción en semifonemas para indicar el fonema inicial y final de cada palabra sea **ExtPal** en vez de la opción por defecto que es un punto '.'. Si solo se especifica **ExtPal**, esta será la cadena usada para indicar tanto el principio como el final. Si se indica también **FinPal**, **ExtPal** será la cadena que marque el principio, y **FinPal** la que marque el final.

La opción **-S** permite conservar los signos de puntuación y otras indicaciones de silencio en la salida generada. Estos solo son conservados por defecto en el caso de la transcripción fonética.

Diccionarios

En su operación por defecto, **saga** realiza la transcripción fonética estándar de cualquier palabra patrimonial del español. Las reglas seguidas en esta descripción pueden encontrarse en [Marinho95a]. Siendo el español una lengua de escritura fonética, esta conversión, salvo error, es siempre correcta. Existen, sin embargo, diversas situaciones en las que es necesario modificar el modo normal de operación del programa: palabras no patrimoniales, como 'ping-pong'; cultismos, como 'quo'; palabras extranjeras, como 'whisky'; etc. También es necesario modificar el comportamiento por defecto a la hora de reflejar las distintas variantes dialectales como, por ejemplo, las de los dialectos hispanoamericanos descritos en [MM98a]. El modo más inmediato es la utilización de modificadores de la transcripción, explicados en la siguiente sección, y de diccionarios específicos, que se describen a continuación.

Los distintos diccionarios son ficheros de texto en los que cada línea refleja una modificación al comportamiento estándar. Cada línea está compuesta por dos partes separadas por tabulador. En general, aquello que antecede al tabulador será sustituido por lo que le sigue. Varios diccionarios del mismo tipo pueden ser usados al mismo tiempo uniendo sus nombres con el carácter '+'. Así, *Dic1+Dic2* será equivalente al diccionario resultante de concatenar las líneas de *Dic1* y *Dic2*.

La opción **-d DicExc** permite el uso de un diccionario de excepciones ortográficas. Estas excepciones son palabras cuya transcripción estándar es incorrecta y para las que se proporciona una *ortografía simulada* que si permite su correcta transcripción. Casos típicos son las palabras extranjeras 'whisky' o 'football'. Para conseguir una transcripción correcta, podemos usar un diccionario de excepciones con las líneas:

whisky g"uiski
football f'utbol

Nótese que la ortografía figurada ha de ser indicada en formato de 7 bits, con las vocales acentuadas indicadas como tilde+vocal.

La opción **-T TrnPal** permite indicar la transcripción fonética concreta de cada una de las palabras que aparecen en la primera columna del diccionario *TrnPal*. Así, para el ejemplo anterior, podríamos usar un diccionario de transcripción con las líneas:

whisky gw'is-ki
 footbal fuD-B'ol

Es posible indicar modificaciones que afecten sistemáticamente a las palabras empezadas o terminadas con una determinada cadena usando el carácter comodín '*'. Así, para transcribir las palabras empezadas por 'ps' podemos usar:

ps* s

La opción *-t TrnFon* permite indicar cual es la transcripción correcta de un grafema o grupo de grafemas. Así, si queremos que el dígrafo 'ph' sea transcrito siempre como /f/, podemos incluir la línea siguiente:

ph f

La opción *-x DicSus* permite sustituir todas las apariciones de un fonema por otro distinto. Por ejemplo, un modo de indicar la aspiración del sonido /x/ (grafema [j]) sería:

x h

La opción *-g DicGrp* permite convertir cadenas de fonemas en otras cadenas. Como en el caso del diccionario de excepciones, es posible indicar cadenas iniciales o finales de palabra. Así, para indicar que el final /NG/, que aparece en la transcripción de palabras extranjeras como 'parking' o 'ping-pong', introduciríamos una línea del tipo

*NG N

Al forzar transcripciones ortográficas con las opciones anteriores pueden aparecer grafemas o fonemas nuevos. Si es el caso, y para posibilitar la silabización y la división en fonemas, es necesario indicar estos nuevos grafemas y/o fonemas en las listas correspondientes:

-v NovVoc para indicar los nuevos grafemas vocálicos.

-c NovCns para indicar los nuevos grafemas consonánticos.

-l NovFon para indicar los nuevos fonemas aparecidos.

Modificadores de la transcripción

La opción **-M** seguida de una o dos letras permite modificar el modo estándar de transcripción para adaptarse a distintos dialectos o para señalar situaciones concretas como la implosividad de las consonantes o la tonicidad de las vocales. Las opciones disponibles, y una corta descripción de las mismas, son:

- MS Aplicar seseo
- MX Transcribir [x] como [ks]
- MH Transcribir [s] implosiva como [h] siempre
- Mh Transcribir [s] implosiva como [h] salvo final o enlace entre palabras
- MK Transcribir [sT] como [ks]
- MA Transcribir [b], [d] y [g] andinas
- My Transcribir siempre vocálica la conjunción /y/
- MEb Eliminar [b] entre vocales y al final de palabra
- MEd Eliminar [d] entre vocales y al final de palabra
- MEg Eliminar [g] entre vocales y al final de palabra
- MN Velarizar las [n] implosivas
- MM Velarizar todas las nasales implosivas
- M~ Nasalizar las vocales
- MP Representar [pbdtkgx] implosivas internas como [G]
- MR Transcribir [r] implosiva como [R]
- M@ Transcribir [l] y [r] tras [pbd...] como [@l] [@r]
- M: Marcar consonantes implosivas con [:]
- M_ Marcar vocales postónicas con [_]
- M. Marcar con . los fonemas inicial y final de palabra
- MC Separar las oclusiones (p: pcl p t: tcl t ...)

DIAGNÓSTICOS

saga proporciona tres tipos de mensaje de diagnostico: mensajes de error en la ejecución del programa, mensajes de error en la transcripción y aparición de palabras extrañas.

Los errores de ejecución pueden ser debidos a la utilización de opciones incorrectas, en cuyo caso se muestra el modo de funcionamiento y termina la ejecución, o a errores internos del programa, en cuyo caso puede que se proporcione un mensaje explicativo.

Los mensajes de error en la transcripción pueden deberse a diversas causas, pero siempre se refieren a palabras que **saga** no sabe como transcribir. Un caso típico es la de palabras sin ninguna vocal, o con demasiadas consonantes, en cuyo caso **saga** no sabe como realizar la silabificación. Estos errores se muestran en la salida de error estándar, pero la ejecución continua con el resto del texto.

Los mensajes referidos a palabras extrañas se producen cuando una palabra es sospechosa de no ser una palabra propia del español. Eso ocurre, por ejemplo, si aparecen muchas vocales o consonantes seguidas. En caso de encontrar una palabra de este tipo, la ejecución prosigue pero, al acabar, se muestra la lista de palabras extrañas encontradas. Esta lista suele ser de utilidad para localizar palabras que requieren un tratamiento especial, probablemente con el uso de diccionarios.

La opción *-e FicErr* permite redirigir la información de errores de transcripción y palabras extrañas al fichero *FicErr*.

BIBLIOGRAFÍA

[LM93a] J. Llisterri y J.B. Mariño, "Spanish Adaptation of SAMPA and Automatic Phonetic Transcription", internal report SAM-A/UPC/001/V1, 1993.

[Marinho95a] J.B. Mariño, "SAGA. Reglas de Transcripción del Español Standard", internal report UPC, 1995.

[MM98a] A. Moreno y J.B. Mariño, "Spanish Dialects: Phonetic Transcription", Proc. ICSLP98, pp. 189-192, Sydney, Australia, 1998.

LICENCIA

saga: Un transcriptor fonético para el idioma español

Copyright (C) 1993-2009 Albino Nogueiras Rodríguez y José B. Mariño
TALP - Universitat Politècnica de Catalunya, ESPAÑA

This program is free software: you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

You should have received a copy of the GNU General Public License along with this program. If not, see <<http://www.gnu.org/licenses/>>.