Procesamiento del Lenguaje Natural, boletín nº 17, Septiembre de 1995

MORFOLOGIA/ETIQUETADORES

UNA CONTRIBUCIÓN AL PROCESAMIENTO DEL PLURAL EN ESPAÑOL: UN ALGORITMO DE SINGULARIZACIÓN

Álvaro Sánchez Ladrón de Guevara, Francisco García Jumela
Laboratorio de Inteligencia Artificial
Facultad de Informática de Madrid, UPM
e-mail: alvaro@alcala.dia.fi.upm.es, paco@alcala.dia.fi.upm.es

RESUMEN

Este trabajo presenta una forma original de procesar los plurales en español: la singularización. El algoritmo de singularización recibe como entrada una palabra en plural y devuelve como salida la forma singular. Las reglas del algoritmo, divididas según las terminaciones de las palabras, permiten reconstruir el singular de cualquier plural castellano. La consulta de un diccionario de formas singulares sirve para comprobar el éxito del proceso de singularización. El artículo describe en detalle el algoritmo, ofreciendo ejemplos, datos estadísticos y todas las posibles excepciones.

Palabras clave: morfología, verificación ortográfica

1.- INTRODUCCIÓN

En la lengua española la forma del número plural presenta tres variantes: en *-s* (*toros*), en *-es* (*árboles*) y en grado cero (*atlas*).

Aparentemente la variación del número es un fenómeno uniforme y fácil de describir o estudiar. Sin embargo, en no pocas ocasiones, el hablante duda a la hora de formar algunos plurales. El propio lector puede hacer la prueba con la siguiente lista de palabras: *déficit, yo, fagot, champú, i, álbum, cañí, cenit, club, convoy.* [1]

[1] Los plurales son: *fagotes, ies, álbumes, cañís, convoyes, clubes.* Con doble plural: *yos/yoes, champús/champúes.* Invariables: *déficit, cenit.*

Tampoco los procesadores de texto se salvan de los errores al tratar las palabras en plural. WordPerfect 5.1 no es capaz de reconocer, entre otros, los siguientes plurales (pero sí sus formas singulares): *aes, adioses, ballets, bises, cabarés, ces, clubes, doses, fes, filmes, ruines, esquíes, íes, síes, cafés, complots, pluses, seises, bengalíes, canadienses*. Microsoft Word 6.0 no reconoce, entre otras palabras: *aes, bises, cabarés, ces, doses, fas, íes, síes, seises*.

Los sistemas de procesamiento de lenguaje natural suelen emplear dos estrategias básicas para el tratamiento automático de los plurales:

- La pluralización. Consiste en generar el plural a partir de la forma singular. Algunas gramáticas (Bello (1984), Alcina/Blecua (1987), RAE (1991)) describen las reglas de formación de plurales, añadiendo (aunque nunca de forma exhaustiva) sus numerosas excepciones.
- Disponer de un diccionario con la descripción morfológica completa de cada palabra.

Otra estrategia posible para el tratamiento de los plurales es la singularización, que presentamos en este trabajo. Por singularización entendemos el proceso que permite reconstruir el singular de una forma plural dada. No se trata, como pueda parecer a primera vista, de una tarea sencilla. La singularización automática presenta algunas dificultades de cierta importancia. El principal obstáculo lo constituyen los plurales terminados en *-vocal+consonante+es*, cuando coinciden los dos posibles morfemas de plural (*-s* y *-es*), como en *vorágines*. En estos casos caben dos singularizaciones, *vorágine/voragin*, puesto que no existe criterio o norma para determinar el singular correcto. Otro problema es realizar los cambios ortográficos necesarios para

singularizar adecuadamente algunas formas. Estos cambios pueden ser: la eliminación de acento ortográfico (*gérmenes→germen*), la adición de acento ortográfico (*volcanes→volcán*) y la sustitución de la letra *C* por *Z* (*voces→voz*). En el siguiente apartado describiremos el algoritmo de singularización así como las soluciones a los problemas mencionados.

2.- DESCRIPCIÓN DEL ALGORITMO DE SINGULARIZACIÓN

El objetivo del algoritmo es obtener el singular de cualquier palabra en plural. El procedimiento recibe como entrada un plural, por ejemplo *leones*, y devuelve como salida la forma singular correspondiente, *león*.

El proceso de singularización se realiza en tres pasos:

1) Se comprueba la terminación de la palabra en plural. Distinguimos dos grandes grupos según las palabras terminen en *-ES* o en *-S* (excluidas las acabadas en *ES*). Estos dos grupos se dividen a su vez en subgrupos con condiciones de terminación más restrictivas. La terminación más compleja que contempla el algoritmo es -LTR+CON+CON+VOC+*SES* (ej.: *c-ipreses*) y la más sencilla VOC+*S* (ej.: *t-és*); siendo LTR una letra cualquiera, CON consonante y VOC vocal.

2) Dependiendo del tipo de terminación, se efectúan las operaciones necesarias para singularizar la palabra. El conjunto de operaciones posibles es: quitar *-S*, quitar *-ES*, quitar acento, poner acento en la última vocal, sustituir *C* por *Z*.

3) Finalmente se busca en el diccionario la forma singular resultante. Este diccionario no es más que una lista de palabras en singular. Debe contener

[illegible] aquellas palabras susceptibles de recibir morfemas de plural. [illegible], por tanto, las formas verbales (menos los participios), las [illegible], las conjunciones, los adverbios y cualquier otra palabra invariable [illegible] que no admita plural[2].

Son varias las razones que justifican la necesidad de contar con un diccionario en el proceso de singularización. En primer lugar, la consulta del diccionario sirve para comprobar que el singular resultante es una palabra correcta, perteneciente a nuestro léxico. De esta manera se rechazan las singularizaciones de formas inexistentes o extrañas al vocabulario español (ej.: *kkdles→kkdle, haxwyoces→haxwyoz).* En segundo lugar, utilizando un diccionario como este, que sólo incluye el singular de las palabras que admiten morfemas de plural, evitamos dar por buenas singularizaciones de palabras sin número o invariables (ej: *teneres→tener, arribas→arriba, teces→tez*).

Por último, y no menos importante, el diccionario también resulta imprescindible para resolver algunas singularizaciones. Como mencionábamos antes, no existe ningún criterio o regla capaz de precisar la forma singular correspondiente a las palabras terminadas en -VOC+CON+*ES* (como *jades, abades, nenes, genes, coces* o *roces*). En estos casos resulta imposible determinar si el morfema de plural es *-ES* o *-S*. Por consiguiente, la singularización es ambigua: la palabra puede finalizar tanto en *-e* como en consonante. Así tenemos:

jades→jade / abades→abad
nenes→nene / genes→gen
roces→roce / coces→coz

---

[2] En este grupo se ecuentran también los sustantivos invariables: los acabados en *-S* (ej.: *martes, oasis, crisis, tesis...*) y otros como *cariz, tez, fénix, hazmerreír, salud, sed,* etc., que no tienen forma plural.

La única solución es realizar una segunda singularización en los casos en que la primera fracase (cuando la palabra no se encuentra en el diccionario de singulares). Por ejemplo, si se singulariza *vorágines* como *voragin* (véase la regla 5.1 del algoritmo), al no estar incluida esta forma en el diccionario se intentará una segunda singularización (véase la regla 5.1bis), ya definitiva, como *vorágine*. Las reglas denominadas bis en el algoritmo responden a esta finalidad: efectúan una segunda singularización para dar cuenta de las palabras que deben acabar en *-e*. El número de palabras españolas terminadas en esta vocal es muy inferior al de palabras que finalizan en consonante[3]. Por esta razón, las reglas *bis* sólo se disparan para obtener los singulares en *-e*, tras fracasar el primer intento de singularización. Así se consigue reducir al mínimo los casos en que es necesario singularizar dos veces para encontrar la forma correcta.

El algoritmo que se presenta a continuación está ordenado de acuerdo con el número total de palabras que cumplen las condiciones de terminación exigidas por las reglas. Para contabilizar el número de palabras según sus terminaciones se ha consultado el muy útil *"Diccionario inverso del español"* de H. de la Campa (1987), que comprende todas las entradas del diccionario de la RAE (edición 20ª).

Las escasas excepciones a las reglas se señalan y enumeran en las notas a pie de página. Se entiende por excepción aquella palabra que cumpliendo la condición de la regla no se singulariza del modo que la acción de la regla indica. Las palabras que no son contempladas por ninguna regla también se consideran como excepciones.

[3] Compárense las cifras totales (abreviatura **T**) que se ofrecen en el algoritmo.

Los plurales especiales de los latinismos y las palabras compuestas no se tratan en el algoritmo ya que presentan formas muy variadas o irregulares[4].

Los símbolos y abreviaturas empleados son:

VOC: vocal sin acento; si la vocal puede ir acentuada se añade **(con/sin acento)**.

CON: consonante.

LTR: cualquier letra (vocal o consonante).

(X|Y): o bien *X*, o bien *Y* (siendo *X* e *Y* letras).

Un guión **(-)** situado delante de las terminaciones indica que en esa posición puede aparecer cualquier cosa (una o varias letras, o bien ninguna letra).

[T:*cifra***]** contabiliza el número total de palabras que cumplen la condición de la regla.

Las palabras de los ejemplos **(Ej.)** van en cursiva, apareciendo en negrita la terminación por la que se pregunta en la regla.

[4] Invariables. Latinismos: *accésit, campus, superávit, junior, senior, quorum, tándem...* Palabras compuestas: *paracaídas, tragaperras, sacacorchos...*

Plural único. Palabras compuestas: *coche cama-coches cama, guardia marina-guardias marinas, cualquiera-cualesquiera, hijodalgo-hijosdalgo...*

Varias formas de plural. Latinismos: *curriculum-currícula/currículos, referéndum-referendos/referéndums/referéndum, ultimátum-ultimatos/ultimátum...*