Corrector ortográfico de libre distribución basado en reglas de derivación

Santiago Rodríguez, Jesús Carretero
Facultad de Informática
Universidad Politécnica de Madrid, 28660 Madrid, España
e-mail: {srodri, jcarrete}@fi.upm.es

Septiembre 1999

Resumen

La carencia de herramientas software de libre distribución que permitan la corrección de textos escritos en castellano llevó a los autores a la construcción de un diccionario de castellano basado en el programa de libre distribución *ispell*. Este artículo presenta la integración del diccionario al entorno de la herramienta *ispell*, haciendo especial hincapié en los aspectos de la especificación formal de las reglas de derivación, etiquetado de las palabras que componen el diccionario raíz y la generación final del diccionario. El diccionario se distribuye como una herramienta de libre distribución desde 1994 bajo los términos de la *General Public License* de *Free Software Foundation*.

Palabras Clave: Lenguaje Natural, Especificación Formal, Software de Libre distribución, LaTeX.

1 Introducción

La introducción de los computadores en el procesamiento de textos ha demostrado la carencia de herramientas especializadas (correctores ortográficos, correctores gramaticales, etc.) para el castellano. Ésta es la tercera lengua más extendida y, sin embargo, la disponibilidad de este tipo de herramientas está muy lejos de otras lenguas mucho menos extendidas, pero con mayor influencia tecnológica que los países de habla hispana.

La importancia potencial del castellano en un futuro próximo llevó a los autores a desarrollar algunas herramientas gramaticales software. Además, para promocionar la utilización de estos programas, se pensó en su distribución totalmente gratuita. Una de estas herramientas es el diccionario de castellano para el corrector ortográfico *ispell* desarrollado por Geoff Kuenning que permite

la incorporación de distintos diccionarios (Ispell se puede obtener mediante anonymous ftp de `ftp.math.orst.edu` en
`/pub/ispell-3.1/ispell-3.1.20.tar.gz`).

Uno de los principales objetivos impuestos en el desarrollo del diccionario de castellano fue su distribución totalmente gratuita para permitir su utilización al mayor número de usuarios posible. Por otra parte el diccionario debe ser exhaustivo, es decir, debe contener el mayor número posible de palabras aceptadas en castellano, así como la mayor parte de sus derivaciones. Puesto que es una herramienta de libre distribución debe ser fácil de mantener ya que se espera de la colaboración de los usuarios finales para actualizar y mejorar tanto el diccionario raíz como el conjunto de reglas de derivación. Por último, debido a la diversidad de vocabulario del castellano, dependiendo de la zona geográfica del usuario que utilice la herramienta se utiliza un conjunto de palabras ligeramente distinto del resto. Por tanto el proceso de generación del diccionario debe permitir al usuario seleccionar el conjunto de palabras que mejor se adapte al que se utiliza en su zona geográfica.

El principal problema encontrado en el desarrollo del diccionario fue la adaptación de las reglas gramaticales castellanas a una especificación formal. A diferencia del inglés, el castellano contiene un número muy elevado y complejo de reglas de derivación a partir de una palabra raíz. Las principales tareas que se realizaron fue la formalización de las reglas de derivación gramaticales y la generación de un conjunto de palabras etiquetadas (palabras con su conjunto de reglas a aplicar). El desarrollo de esta herramienta se inició a comienzos de 1994. El primer prototipo estuvo finalizado a mediados de 1994 y se dedicó a su uso interno para detectar errores. Se distribuye de forma gratuita desde finales de 1994[1]. Actualmente se distribuye la versión 1.6 (Abril 1999).

2 Características Morfológicas del Castellano

El castellano es una lengua que derivó del latín y tiene una gramática compleja. Para llevar a cabo la construcción de cualquier plataforma léxica la primera tarea que se debe llevar a cabo es el estudio de la gramática castellana [2]. Este estudio debe permitir formalizar el conjunto mínimo de reglas necesario que permita extraer el conjunto de palabras reconocidas por la lengua castellana a partir de un conjunto de palabras raíces mínimo. Para alcanzar dicho objetivo se construyó un árbol de derivación a partir de una palabra raíz. Una versión simplificada de dicho árbol se muestra en la figura 1.

Los principales problemas que se encontraron en la construcción de este conjunto de reglas de derivación vinieron originados por las características del castellano. Los más relevantes se exponen a continuación:

[1] Se puede obtener mediante anonymous ftp en `ftp.fi.upm.es` en `pub/unix/espa~nol.tar.gz` o mediante el URL `http://www.datsi.fi.upm.es/~coes/`

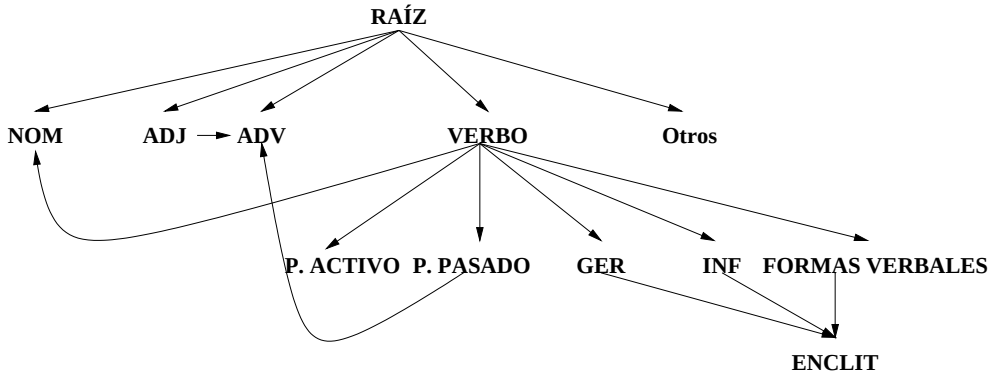

Figura 1: Estructura Morfológica simplificada del castellano

Derivaciones de género y número. Los adjetivos (ADJ) y sustantivos (NOM) tienen género (masculino o femenino) y número (singular y plural). La situación habitual es que un adjetivo o nombre tenga derivación tanto en género como en número. Por ejemplo: *perro* → *perra* y (*perros, perras*). Otros casos únicamente tienen un género y, por tanto sólo admiten derivación en número. Es el caso de un sustantivo masculino como *álamo, álamos*, o femenino como *casa, casas*.

Conjugación verbal. Cada una de las tres conjugaciones verbales del castellano tienen 40 derivaciones temporales (P. ACTIVO, P. PASADO, GER, INF y FORMAS VERBALES). Como es sabido, los verbos regulares tienen un conjunto estricto de reglas de derivación que son idénticas para todos los de una misma conjugación. Los verbos irregulares tienen al menos una derivación diferente que las derivaciones regulares correspondientes a su conjugación. Las derivaciones irregulares se agrupan en alrededor de 100 tipos diferentes de irregularidades [2].

Formas enclíticas. Algunas derivaciones verbales se generan añadiendo una forma pronominal al final de una forma verbal (ENCLIT). En el castellano escrito se pueden encontrar dos formas enclíticas diferentes: los verbos pronominales, cuyas formas enclíticas se generan añadiendo los sufijos *-me, -te, -se, -nos* y *-os* en el infinitivo y en el gerundio (*amar* → *amarte*), y los verbos transitivos, cuyas formas pronominales se generan añadiendo las terminaciones *-lo, -la, -los, -las, -le* y *-les* (*amar* → *amarla*). Ambas formas enclíticas se pueden combinar para formar enclíticos más complejos (*ajustar* → *ajustármelo*). Esto genera un conjunto de reglas de complejidad $O(n^2)$. Además, estas formas enclíticas se ven afectadas por las irregularidades que presentan algunos verbos en su gerundio (*vestir* → *vistiéndote*), lo que incrementa el grado de complejidad para las formas enclíticas.

Nombres derivados de verbos. Algunos sustantivos son formas derivadas

de un verbo como *imaginar* → *imaginación* o *abatir* → *abatimiento* (VERBO → NOM).

Adverbios derivados de adjetivos. Gran parte de los adverbios modales se generan añadiendo el sufijo *-mente* a un adjetivo (ADJ → ADV): *tranquilo* → *tranquilamente*.

Superlativos y diminutivos. Las formas regulares de superlativos se forman añadiendo el sufijo *-ísimo* a un adjetivo (*grande* → *grandísimo*). Los diminutivos se forman añadiendo los sufijos *-ico, -ito* y *-illo* a un adjetivo o nombre.

Vocales acentuadas. Hay muchas particularidades relacionadas con las derivaciones en género y número que se han tenido en cuenta al hacer el estudio del modelo. Algunas palabras pierden una vocal acentuada sustituyéndola por su equivalente no acentuada: *gañán, gañanes*.

Teniendo en cuenta las características descritas en los párrafos anteriores se ha desarrollado un conjunto de reglas formales que comprende un extenso subconjunto de las que conforman la gramática castellana. Cada una de las entradas del diccionario que contiene las palabras raíces tiene una etiqueta que representa una lista de reglas de derivación que se deben aplicar a dicha palabra para obtener sus formas derivadas.

3 Implementación del modelo.

Las características gramaticales estudiadas en el apartado anterior han llevado a la realización del modelo ajustándose a las restricciones que imponía la herramienta *ispell*. Estas restricciones se basan en agrupar un conjunto de reglas (clase) al que se asocia una etiqueta que será referenciada en las etiquetas del diccionario raíz.

Puesto que el uso del castellano se basa en gran parte en las formas derivadas (un verbo castellano tiene 55 formas derivadas), las reglas de derivación del diccionario incluyen todas las derivaciones de los verbos regulares. Además, se han incorporado reglas adicionales para tener en cuenta la mayor parte de los patrones por los que se rigen las derivaciones de los verbos irregulares. La inclusión de las formas derivadas de los verbos *ser*, *ir*, *haber* y *estar* se han incluido en su totalidad en el diccionario raíz al no adecuarse fácilmente a ninguno de los patrones considerados. En resumen, el conjunto de reglas implantado para especificar la gramática castellana contiene alrededor de 3.300 reglas agrupadas en 57 macroreglas o clases.

Cada macroregla que se describe a continuación refleja un aspecto particular de la gramática castellana que se ha descrito en la sección anterior. Cada una de las reglas de derivación que componen una clase o macroregla trata un

caso particular [6, 7, 1]. La condición que se muestra en la primera columna de cada uno de los ejemplos representa la aceptación de una palabra para ejecutar la acción que se muestra en la segunda columna. Si una palabra termina con el sufijo especificado, se realiza la acción subsiguiente. Esta acción se basa en sustituir un morfema de la palabra raíz por otro, o simplemente añadir un morfema adicional.

Derivaciones de género y número. Se han incluido dos macroreglas que realizan estas derivaciones. La derivaciones en número incluyen 11 reglas. La regla a aplicar depende de la terminación de la palabra raíz sobre la que se aplica la regla. La macroregla de derivación en género y número se compone de 20 reglas. A continuación se muestran algunos ejemplos de derivación de estas clases:

Derivaciones en número			**Derivaciones en género y número**		
Condición	Acción	Ejemplo	Condición	Acción	Ejemplo
[AEIOU]	S	vacas	O	-O, A	amiga
Z	-Z, CES	arroces	O	S	amigos
ÚN	-ÚN,UNES	atunes	[^AONS]	ES	pastores

La última regla del ejemplo anterior se muestra la generación del plural masculino y femenino para aquellas palabras que no acaban en *a, o, n* ni *s*.

Conjugación Verbal. Las reglas de derivación que permiten generar las formas verbales se agrupan en cuatro clases: dos de ellas se aplican a verbos regulares y las otras dos a verbos irregulares. Alrededor de 200 reglas componen las dos clases que derivan los verbos regulares mientras que el conjunto de derivaciones de los verbos irregulares se compone de unas 2.500 reglas. En este último aspecto es donde se ha dedicado la mayor parte del esfuerzo. Sin embargo su formalización ha sido factible puesto que las formas irregulares del castellano siguen patrones de derivación bien definidos: *-ontar* → *-uento*, *-oder* → *-uedo*, *-ervir* → *-irvo*, etc. Algunas reglas de derivación para verbos regulares e irregulares se muestran a continuación:

Verbos Regulares			**Verbos Irregulares**		
Condición	Acción	Ejemplo	Condición	Acción	Ejemplo
AR	-AR, O	amo	IAR	-IAR, ÍO	envío
CER	-CER, ZO	venzo	OÑAR	-OÑAR, UEÑO	sueño
CIR	-CIR, ZO	zurzo	SABER	-ABER, É	sé

Estas reglas no han tenido en cuenta los verbos *ser, estar, ir* y *haber* puesto que no existen un conjunto de patrones que permitan derivar to-

das sus formas verbales a partir del infinitivo. Todas sus formas derivadas se han incluido explícitamente en el diccionario de palabras raíces.

Formas enclíticas. Los verbos regulares incluyen alrededor de 200 reglas de derivación para generar las formas enclíticas, mientras que los verbos irregulares incorporan en torno a 400. Estas reglas representan los enclíticos generados por las derivaciones pronominales, transitivas y combinadas de ambas. Estas reglas únicamente se aplican a las formas del gerundio e infinitivo.

Nombres derivados de verbos. Se han tenido en cuenta los nombres acabados en *-miento* y *-ción* que se derivan de verbos a partir de dos macroreglas.

Adverbios derivados de adjetivos. Se ha considerado una macroregla (clase) que genera los adverbios terminados en *-mente*.

Superlativos y diminutivos. Actualmente únicamente los superlativos regulares se han considerado y constituyen una clase.

4 Generación del Diccionario

El léxico para esta plataforma ha sido extraído de un *Corpus de Español* compilado por los autores. Este corpus, que contiene más de 20 millones de palabras, incluye textos extraídos de las siguientes fuentes: Textos de periódicos españoles (ABC Cultural, El Mundo, El Periódico, etc.); Libros seleccionados, como la Biblia; Textos técnicos (informes técnicos, artículos, proyectos de fin de carrera, diccionarios técnicos y libros); Corpus oral [5]; Versión concisa del diccionario Español-Inglés Collins [8].

El *léxico básico* resultante contiene más de 80.000 palabras distintas, 53.000 de las cuales han sido ya etiquetadas de acuerdo a las reglas de derivación mostradas en la sección anterior. Las restantes 27.000 están en proceso de etiquetado. El etiquetado se hace de forma semiautomática mediante una herramienta que extrae los morfemas de cada palabra y propone una o varias *etiquetas tentativas*. Sin embargo, las formas derivadas son comprobadas manualmente para verificar la corrección o no del etiquetador. El *léxico de referencia* usado para desarrollar COES es el diccionario de la *Real Academia Española* (RAE), la institución oficial que vela por la pureza del lenguaje español y admite las nuevas palabras del mismo.

La versión de libre distribución de COES incluye un fichero de afijos y varios ficheros de léxico:

- `espa~nol.words` contiene una lista de palabras del diccionario oficial de Español [3].

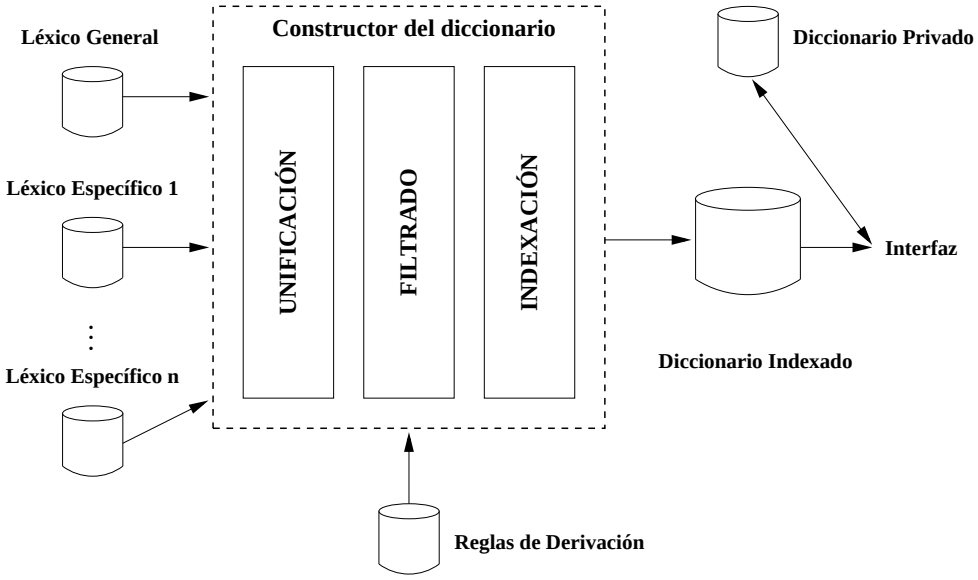

Figura 2: Generación del Diccionario

- `espa~nol.comp` contiene una lista de palabras que no aparecen en el diccionario oficial de la Lengua Española, pero de uso habitual en los textos técnicos.
- `antiguas.words` contiene una lista de palabras que aparecen en el diccionario oficial de Español, pero etiquetadas como palabras en desuso.
- `espa~nol.nofl` contiene una lista de palabras que no aparecen en el diccionario oficial de Español, pero que han sido frecuentemente encontradas en el corpus.
- `espa~nol.propios` contiene una lista de nombres propios.

Cuando se aplican las reglas de derivación al léxico básico usado en COES actualmente se crea un diccionario que contiene más de 650.000 palabras. El diccionario de español se construye usando la herramienta *ispell*, que aplica las reglas de derivación elaboradas por los autores, siguiendo el formato de esta herramienta, al léxico básico. *ispell* sigue cuatro pasos básicos para generar del diccionario (figura 2):

1. Generación de las reglas de derivación a partir del fichero de reglas.
2. Unificación de las entradas del diccionario para evitar redundancias y formas ilegales.
3. Interpretación de las reglas de derivación para calcular las formas derivadas.

4. Construcción de un árbol indexado para conseguir una búsqueda eficiente en el diccionario.

Existe la posibilidad de que los usuarios puedan generar diccionarios *particularizados* mezclando varios ficheros de léxico cuando se construye el diccionario. Esta opción permite a cada usuario incluir sus propios léxicos (que deberían estar etiquetados). Además, los usuarios pueden particularizar el diccionario eligiendo el formato en que se codificarán los caracteres especiales (ü, ñ y letras acentuadas), que no se encuentran definidos en el conjunto básico de caracteres ASCII de siete bits. Para permitir esta particularización, se proporciona a los usuarios la codificación de estos caracteres en los formatos más habituales cuando se definen las reglas.

Acutalmente se proporcionan los siguientes formatos distintos:

latin1	Formato TeX	Formato LaTeX	Html
á	\' a	'a	á
é	\' e	'e	é
í	\' {\i}	'i	í
ó	\' o	'o	ó
ú	\' u	'u	ú
ñ	\' n	'n	ñ
ü	\" u	"u	ü
Á	\' A	'A	Á
É	\' E	'E	É
Í	\' {\I}	'I	Í
Ó	\' O	'O	Ó
Ú	\' U	'U	Ú
Ñ	\' N	'N	Ñ
Ü	\" U	"U	Ü

Formato **msdos**: Las letras acentuadas se codifican utilizando el código ASCII MS-DOS extendido.

Para ejecutar el ispell con un determinado formato:

```
ispell -T <formato> -d español <fichero>
```

El diccionario se puede generar en cualquier sistema operativo para el que exista una versión de ispell. Esto incluye cualquier computador que ejecute alguna versión de Unix y, además los sistemas Windows NT y Windows 95/8.

5 Conclusiones y Trabajo Futuro

En este trabajo se ha presentado un diccionario de español desarrollado para la herramienta *ispell*, diccionario que está siendo usado por una comunidad creciente de usuarios. La versión actualmente existente de COES puede ser mejorada en los aspectos siguientes: Elaboración de diccionarios locales y temáticos,

para dar cabida a palabras usadas en áreas restringidas de la comunidad hispanohablante o en entornos lingüísticos especiallizados (leyes, medicina, etc.); optimización de reglas; incrementar el léxico básico para reducir la tasa de error de COES y aumentar su eficiencia.

La difusión de este trabajo como una herramienta de libre distribución ha permitido una rápida extensión de la misma, encontrándose actualmente plenamente integrada con herramientas de libre distribución que usan ispell (por ejemplo *emacs*). Además se están manteniendo conversaciones con representantes del proyecto Lucas (Linux en Castellano) para mejorar la distribución y la calidad de COES.

El mantenimiento del diccionario y su depuración es una tarea que han asumido los autores casi en su totalidad. Sería interesante contar con la colaboración de los usuarios para detectar palabras erróneas, ausentes, reglas con erratas, etc., aunque la experiencia demuestra que los usuarios son poco colaboradores. En este sentido existe una dirección de correo a la que se pueden enviar cualquier tipo de sugerencia o error detectado:

`espanol-bugs@datsi.fi.upm.es`

Actualmente están en desarrollo algunas nuevas utilidades para COES. Un *tesauro*, que usará intensivamente las reglas de derivación, estará disponible pronto. Además se está llevando a cabo un estudio preliminar de las reglas y modelos sintácticos y gramaticales del español [4, 9] con el propósito de construir un corrector sintáctico en un futuro próximo.

Referencias

[1] J. Carretero, S. Rodríguez. Building lexical tools to manage information written in Spanish. *Journal of Information Science*, 22(5):391–399, Octubre 1996.

[2] Real Academia Española de la Lengua. *Esbozo de una Nueva Gramática de la Lengua Española*. Espasa Calpe, 1991.

[3] Real Academia Española de la Lengua. *Diccionario de la Lengua Española*. Espasa Calpe, 21ª edición, 1992.

[4] J.Hallebeek. *Morfología y Sintaxis del Español: Introducción al Análisis Oracional*. Playor, Madrid, España, 1994.

[5] F. Marcos, A. Ballester, C. Santamaría, E. Pertierra, O. Brandeo, P. Díez. Corpus oral de referencia de la lengua española contemporánea. Technical report, Universidad Autónoma de Madrid, 1992.

[6] S. Rodríguez, J. Carretero. Building a Spanish speller. *Taller sobre Software de Libre Distribución*. Universidad Carlos III de Madrid, España, 1995.

[7] S. Rodríguez, J. Carretero. A formal approach to Spanish morphology: the COES tools. *SEPLN'96 Conference Proceedings*, pp. 118–126. SEPLN, Sevilla España, 1996.

[8] C. Smith. *Collins English-Spanish Dictionary.* Collins, 1988.

[9] E. Tzoukermann, M. Liberman. A Finite-State Morphological Processor for Spanish. *Proceedings of the 13th International Conference on Computational Linguistics (COLING 90)*, pp. 277–281, 1990.