NOTAS E INFORMACIÓN

FORMALIZACIÓN DE REGLAS MORFOLÓGICAS PARA UN NUEVO CORRECTOR ORTOGRÁFICO EN ESPAÑOL

1. *Introducción*

La utilización de los computadores como herramientas de procesamiento de texto ha mostrado la carencia de herramientas especializadas (correctores ortográficos completos, correctores gramaticales, etc.) para corregir textos en español. Esta deficiencia se aprecia especialmente en el entorno de programas de libre disposición, donde ya existen herramientas para la corrección de textos en lengua inglesa desde hace tiempo (programa *ispell* desarrollado por Geoff Kuenning). Para tratar de resolverla, se planteó la construcción de un corrector ortográfico completo de la lengua española válido para su utilización por cualquier tipo de sistema informático, pero cuya primera versión estaría construida sobre *ispell* Dicho corrector se ha desarrollado a partir de la especificación gramatical publicada por la Real Academia Española de la Lengua (RAEL) expuesta en [2] y un conjunto de palabras (50.000) a las que se aplican dichas normas gramaticales. El principal problema que se encontró en la construcción del diccionario fue adaptar la gramática española a una especificación formal utilizable por un sistema informático. A diferencia del inglés, el español es una lengua derivada del latín y tiene muchas y muy complejas reglas gramaticales que implicaban aportar un gran esfuerzo en dicha formalización.

Los principales objetivos impuestos al comienzo del desarrollo fueron los siguientes:

1. El conjunto de reglas gramaticales debe ser exhaustivo, es decir, debe incluir la mayor parte de las reglas de derivación de nuestra lengua.

2. Será de libre disposición. La distribución de la herramienta será gratuita para favorecer su utilización por un gran número de usuarios. Esto implica que la información que se va a obtener de los usuarios de la herramienta va a permitir la mejora de sus prestaciones. Por esto tanto las reglas gramaticales como su aplicación a las palabras raíces deben ser fácilmente mantenibles.

3. Puesto que será de gran difusión, el usuario tipo del diccionario no será homogéneo especialmente en los países americanos de habla hispana. Por esta razón el conjunto de palabras de las que puede disponer el usuario debe ser fácilmente configurable a sus necesidades.

4. La herramienta debe proporcionar altas prestaciones, puesto que se concibe como una plataforma de explotación más que como un estudio de investigación en lingüística.

El desarrollo del diccionario comenzó a primeros de 1994. Las principales tareas en este trabajo fueron la obtención de las reglas de derivación españolas y un conjunto de palabras raíces a las que habría que aplicar dichas reglas.

2. *Características gramaticales del español*

La construcción de un diccionario para un corrector ortográfico automático exige un estudio preliminar de las características gramaticales del lenguaje que permitan obtener todas las palabras derivadas del español a partir de un pequeño conjunto de palabras raíces.

A continuación se describirán los problemas detectados en la identificación de las reglas de derivación:

Derivaciones de género y número. Los adjetivos y sustantivos tienen género y número. Hay que tener en cuenta que en español existen palabras que tienen su derivación femenina como *perro* → *perra* y que ambas tienen sus derivaciones del plural *perros* → *perras*. Además hay palabras que tienen género, pero no tienen su correspondiente derivación en género. Por ejemplo *álamo* o *casa*, pero sí tienen sus derivaciones de número: *álamos, casas.*

Ésta es la razón por la que se han incluido dos conjuntos de reglas que separan las palabras que tienen derivaciones en género y número y aquéllas que únicamente se pueden aplicar derivaciones en número.

Conjugaciones verbales. Como es bien sabido los verbos españoles se dividen en tres conjugaciones: *-ar*, *-er* e *-ir*. Además cada conjugación tiene más de 30 derivaciones temporales. Dependiendo de dichas formas verbales la lengua española tiene dos tipos de verbos:

1. Verbos Regulares. Son aquéllos que se conjugan sin alterar su raíz ni la derivación.

2. Verbos Irregulares. Son aquéllos que sufren una alteración en la raíz o en las derivaciones en al menos una de sus formas. Se han identificado 100 formas distintas de verbos irregulares. ([2]).

Formas enclíticas. Algunas derivaciones verbales se generan añadiendo un pronombre al final de la forma verbal. En el español actual se han encontrado dos tipos de derivaciones enclíticas:

1. Verbos pronominales. Las formas enclíticas se forman añadiendo las partículas *-me*, *-te*, *-se*, *-nos*, *-os* a una forma verbal: *amar* → *amarme.*

2. Verbos transitivos. Las formas enclíticas se forman añadiendo las partículas *-lo*, *-la*, *-los*, *-las*, *-le*, *-les* a una forma verbal.

El español escrito actual únicamente utiliza formas enclíticas derivadas del infinitivo y del gerundio. En algunas regiones de España como por ejemplo Asturias, forman enclíticos en todas sus formas verbales: *comer* → *comió* → *comiólo* (*lo comió*). Estas derivaciones son poco usuales y no se han tenido en cuenta en las reglas de derivación expuestas en este trabajo. Las derivaciones que sí han sido tenidas en cuenta son los verbos que presentan irregularidad en la derivación del gerundio: por ejemplo *vestir* → *vistiéndome*.

Además de los dos tipos de enclíticos descritos anteriormente se pueden crear formas verbales combinando las dos formas enclíticas ya explicadas: *ajustar* → *ajustármelo, ajustándomelo*.

Otros afijos. Además de las derivaciones descritas en los párrafos anteriores, hay muchas palabras que se pueden formar como derivadas de adjetivos, sustantivos y verbos. Por ejemplo los superlativos regulares se forman añadiendo la partícula *-ísimo*. Los adverbios acabados en *-mente* se generan a partir de sus adjetivos correspondientes. Muchos verbos permiten formar un participio activo o de presente: *cantar* → *cantante*.

Letras acentuadas. Muchas palabras al formar sus derivaciones exigen cambiar una letra con tilde por su correspondiente sin ella o viceversa. Es el caso de *gañán* → *gañanes* o *régimen* → *regímenes*. El primer caso se ha tenido en cuenta en las reglas de derivación puesto que se presenta muy frecuentemente. El segundo caso es muy poco frecuente y la palabra presenta un cambio en su raíz y se han dado de alta como palabras raíces tanto el singular como el plural.

Teniendo en cuenta las características descritas en los párrafos anteriores, se han desarrollado una serie de reglas formales que abarcan la mayor parte de las reglas gramaticales españolas.

3. *Definición formal*

La definición formal de las reglas de derivación se han realizado utilizando la herramienta *ispell* y siguiendo el lenguaje formal impuesto por ella. El conjunto de reglas que especifican la gramática española se compone de alrededor de 3500 reglas. Una regla se compone de una condición seguida de una acción. Si la condición especificada en la regla se cumple se ejecuta la acción para formar una palabra derivada. En el ejemplo mostrado a continuación la condición es que la palabra a la que se va a aplicar la regla termine en AR. Si se cumple la condición se ejecuta la regla de derivación consistente en eliminar la partícula AR y añadir la terminación O.

```
A R          > -AR, O      # amar amo
# condición  > acción      # comentario
```

Estas reglas están agrupadas en 57 macroreglas. Una macroregla es un conjunto de reglas que se aplican como si de una sola se tratase. En el ejemplo que se muestra a continuación la macroregla etiquetada como V se compone de tres reglas que definen la conjugación de la primera persona del singular del presente de indicativo para algunos verbos regulares. Si esta macroregla se aplica al verbo *amar* se realiza la derivación *amar* [-ar+o], puesto que cumple la condición de la primera regla y no las condiciones asociadas a la segunda y tercera. Si se aplicara a los verbos comer y vivir se realizarían las derivaciones *comer* [-er+o] y *vivir* [-ir+o] respectivamente por la razón anteriormente expuesta.

```
flag *V:                    # Presente indicativo
                            #1.ª persona singular
    A R     > -AR, O        # amar amo
    E R     > -ER, O        # comer como
    I R     > -IR, O        # vivir vivo.
```

Cada palabra raíz del diccionario tendrá un conjunto de macroreglas aplicadas que definirán las palabras derivadas de la palabra raíz. En el ejemplo anterior las palabras *amar*, *comer* y *vivir* tendrían, entre otras, aplicada la macroregla V.

De las 57 macroreglas, 41 de ellas generan derivaciones relativas a prefijos. Estas 41 reglas no son operativas puesto que se presentan problemas aún no resueltos. Algunos verbos que son transitivos dejan de serlo al aplicarles un prefijo. Otras palabras cambian radicalmente su significado al aplicarles un prefijo. Es el caso de *alzar* → *realzar*.

Cada una de las 16 macroreglas restantes describen un aspecto de nuestra gramática previamente discutido en la sección anterior.

Derivaciones en género y número. Se han definido dos macroreglas: derivación en número y derivación en género y número. La derivación en número se compone de 14 reglas dependiendo de la terminación de la palabra sobre la que se debe aplicar. A continuación se muestran las reglas que componen la derivación en número.

```
[AEIOU]         > S             # vaca vacas
Á               > S             # sofá sofás
É               > S             # café cafés
Ó               > S             # café cafés
[ÚÍDJLRY]       > ES            # tabú tabúes
Z               > -Z, CES       # audaz audaces
[^[É] S         > ES            # gas gases
[^ÁÉÍÓÚ] N      > ES            # can canes
Á N             > -ÁN, ANES     # volcán volcanes
É N             > -ÉN, ENES     # retén retenes
Í N             > -ÍN, INES     # orín orines
Ó N             > -ÓN, ONES     # camión camiones
```

Ú N	> -ÚN, UNES	# atún atunes
É S	> -ÉS, ESES	# aragonés aragoneses.

La derivación en género y número se resuelve con 18 reglas teniendo en cuenta terminaciones en: [áió]n, o, és y las que no se ajustan a estas terminaciones.

Conjugaciones verbales. Las derivaciones verbales se han resuelto construyendo 4 macroreglas que generan las conjugaciones de los verbos regulares e irregulares:

1. Derivación regular para las tres conjugaciones verbales. Se generan todas las formas derivadas de un verbo excepto los participios y los gerundios. En las reglas de derivación se entienden por verbos regulares aquéllos que son fonéticamente regulares. Por ejemplo el verbo {em zurcir} no sigue estrictamente los patrones de derivación del verbo {em vivir}, pero se considera un verbo fonéticamente regular. Se compone de 140 reglas aproximadamente. A continuación se muestra como ejemplo la derivación de la primera persona del singular del presente de indicativo.

A R	> -AR, O	# amar amo
[^CG] E R	> -ER, O	# comer como
C E R	> -CER, ZO	# vencer venzo
G E R	> -GER, JO	# coger cojo
[^CGU] I R	> -IR, O	# vivir vivo
C I R	> -CIR, ZO	# esparcir esparzo
G I R	> -GIR, JO	# fingir finjo
G U I R	> -UIR, O	# distinguir distingo
Q U I R	> -QUIR, CO	# delinquir delinco

2. Derivación regular para el gerundio y participios de las tres conjugaciones verbales. Se compone de 11 reglas. A continuación se muestran todas las reglas de derivación.

[AI] R	> -R, DO	# amar amado
[AI] R	> -R, DOS	# amar amados
[AI] R	> -R, DA	# amar amada
[AI] R	> -R, DAS	# amar amadas
E R	> -ER, IDO	# comer comido
E R	> -ER, IDOS	# comer comidos
E R	> -ER, IDA	# comer comida
E R	> -ER, IDAS	# comer comidas
A R	> -R, NDO	# amar amando
E R	> -ER, IENDO	# comer comiendo
I R	> -R, ENDO	# vivir viviendo

Derivación irregular para las tres conjugaciones verbales. Se generan todas las formas derivadas de un verbo excepto los participios y los gerundios. Para derivar

todos los verbos irregulares se han identificado las terminaciones de los verbos que generaban alguna desviación frente a la forma regular de su conjugación. Se han tenido en cuenta todos los casos que se exponen en [2] entre los cuales se encuentran los siguientes: *-iar, -ebrar, -edrar, -egar, -elar, -evar, -emblar, -caer, -raer, -aler, -ecer, -alir, -ebir, -egir, -eguir,* etc.

Téngase en cuenta que de esta lista de patrones se han excluido verbos como *ser, estar, ir* y *haber*. Estos verbos tienen irregularidades que no se pueden asociar a ninguno de los patrones anteriormente expuestos y se han conjugado explícitamente en el diccionario. Se compone de 2500 reglas aproximadamente.

4. Derivación irregular en participios y gerundio. Se compone de 140 reglas aproximadamente y se aplicará a todos aquellos verbos que presenten una irregularidad bien en el participio o bien en su gerundio. Se han tenido en cuenta los siguientes patrones entre otros: *-facer, -hacer, -aer, -oer, -eer, -oder, -olver, -oner, -abrir, -ebir, -ecir, -edir, -egir, -emir, -eir, -entir, -uir,* etc.

Formas enclíticas. Se han desarrollado 6 macroreglas que permiten realizar las derivaciones de enclíticos pronominales, transitivos y combinados para verbos regulares e irregulares en su gerundio. A continuación se describen cada una de las macroreglas:

1. Derivación regular de verbos pronominales. Se incluyen las siguientes reglas:

```
[AEI] R     > ME          # amar amarme
[AEI] R     > TE          # amar amarte
[AEI] R     > SE          # amar amarse
[AEI] R     > NOS         # amar amarnos
[AEI] R     > OS          # amar amaros
A R         > -AR, ÁNDOME # amar amándome
A R         > -AR, ÁNDOTE # amar amándote
A R         > -AR, ÁNDOSE # amar amándose
A R         > -AR, ÁNDONOS # amar amándonos
A R         > -AR, ÁNDOOS # amar amándoos
E R         > -ER, IÉNDOME # comer comiéndome
E R         > -ER, IÉNDOTE # comer comiéndote
E R         > -ER, IÉNDOSE # comer comiéndose
E R         > -ER, IÉNDONOS # comer comiéndonos
E R         > -ER, IÉNDOOS # comer comiéndoos
I R         > -R, ÉNDOME  # evadir evadiéndome
I R         > -R, ÉNDOTE  # evadir evadiéndote
I R         > -R, ÉNDOSE  # evadir evadiéndose
I R         > -R, ÉNDONOS # evadir evadiéndonos
I R         > -R, ÉNDOOS  # evadir evadiéndoos
```

2. Derivación regular de verbos transitivos. Las reglas incluidas son análogas a las anteriores, pero con las partículas *-lo, -la, -los, -las.*

3. Derivación regular de enclíticos combinados. A continuación se muestran algunos ejemplos:

```
A R      > -AR, ÁRMELO     # solucionar solucionármelo
A R      > -AR, ÁRTELO     # solucionar solucionártelo
A R      > -AR, ÁRSELO     # solucionar solucionárselo
                :
E R      > -ER, ÉRNOSLO    # merecer merecérnoslo
E R      > -ER, ÉROSLO     # merecer merecéroslo
E R      > -ER, ÉRMELA     # merecer merecérmela
                :
I R      > -R, ÉNDOMELO    # partir partiéndomelo
I R      > -R, ÉNDOTELO    # partir partiéndotelo
I R      > -R, ÉNDOSELO    # partir partiéndoselo
                :
```

4. Derivación irregular de verbos pronominales. Se tienen en cuenta los patrones de los verbos que presentan irregularidad en su gerundio: *-aer, -eer, -oer, -uir, -eguir, -egir, -eguir y -eír.* Se compone de 120 reglas aproximadamente.

5. Derivación irregular de verbos transitivos. Se tienen en cuenta los mismos patrones que en el caso anterior. Se incluyen del orden de 150 reglas.

6. Derivación irregular para enclíticos combinados. Se compone de 180 reglas.

Adverbios derivados de adjetivos. Es una macroregla compuesta de 2 reglas. Se incluyen las siguiente reglas:

```
O          > -O, AMENTE     # tonto tontamente
[ELNRSZ]   > MENTE          # virtual virtualmente
                            # audaz audazmente
```

Superlativos. Se realiza extensión en género y número del superlativo para los adjetivos acabados en: *-e, -o* y *-l.*

Participio presente. Se realiza extensión en número del participio presente de un verbo. Este conjunto de reglas comprende los casos regulares de derivación de participios presentes. Esto nos permite reducir el volumen del diccionario final, lo que permite aumentar las prestaciones del corrector. Las formas irregulares, menos comunes, no se han tenido en consideración.

```
A R      > -R, NTE         # amar amante amantes
A R      > -R, NTES
E R      > -ER, IENTE      # temer temiente temientes
E R      > -ER, IENTES
I R      > -R, ENTE        # vivir viviente vivientes
I R      > -R, ENTES
```

Las reglas que se acaban de describir se han aplicado a un conjunto de alrededor de 50.000 palabras raíces que se expanden a 500.000 palabras derivadas.

4. *Explotación*

Como resultado de este proceso de formalización se han obtenido un conjunto de reglas y varios diccionarios de palabras con las reglas que les son aplicables. Este paquete se puede obtener mediante *anonymous ftp* en verb-ftp.fi.upm.es- en /pub/unix/espa~nol.tar.gz-. La distribución está compuesta por los siguientes ficheros:

1. espa~nol.words. Es un fichero que contiene una lista de palabras que son reconocidas oficialmente por el Diccionario de la RAEL ([3]).

2. espa~nol.comp. Es un fichero que contiene una lista de que no están reconocidas oficialmente por la academia, pero que tienen un uso muy frecuente en los textos informáticos.

3. antiguas.words. Es un fichero que contiene una lista de palabras oficialmente reconocidas que están marcadas como palabras antiguas o en desuso en el Diccionario de la RAEL [3].

4. espa~nol.nofl. Es un fichero que contiene una lista de palabras muy utilizadas en español que no están aceptadas por la Academia.

5. *Conclusiones y trabajos futuros*

Se ha desarrollado un corrector ortográfico de la Lengua Española que se está usando en una amplia comunidad de usuarios. La información recibida de estos usuarios nos hace pensar que el diccionario funciona correctamente y es exhaustivo. La razón por la que se ha utilizado la herramienta *ispell* es que es de gran difusión en el ámbito informático, aun cuando no sea el instrumento ideal para una lengua de morfología tan compleja como el español. Sin embargo, tiene la ventaja de proporcionar buenas prestaciones, lo que la hace muy adecuada para el filtrado de grandes cantidades de información.

Las líneas de trabajo que permanecen abiertas son:

1. *Diccionarios de la América de habla hispana.* La validación de las palabras se ha basado en el diccionario oficial de la Academia. El principal problema que presenta es que aunque hay un gran número de palabras que son correctas, se aplican en un ámbito geográfico muy reducido. Para conseguir un proceso de corrección más rápido y eficiente habría incluir en el diccionario únicamente aquellas palabras que se vayan a usar potencialmente. Por ello es interesante separar las palabras usadas en el entorno hispanoamericano y permitir su inclusión a los usuarios de estas comunidades. Según esto se crearían diccionarios para Chile, Colombia, Perú, etc.

2. *Diccionarios de palabras «científicas».* En un entorno especializado como Medicina, Derecho, Lingüística, etc. se utilizan palabras que no están reconocidas por la Academia, pero tienen una importancia fundamental en dicho entorno. La creación

de diccionarios especializados permitiría la configuración del diccionario que se desea utilizar, incluyendo todas las palabras oficiales más aquéllas del entorno especializado en el que se va a trabajar.

3. *Optimización de Reglas*. Debido a la forma de trabajar de la herramienta *ispell*, la especificación de reglas contiene algunas reglas repetidas en varias macroreglas. El fichero de reglas debe reducirse en la medida de lo posible para conseguir un tamaño del diccionario menor que permita un aumento de sus prestaciones en la corrección ortográfica.

REFERENCIAS

[1] E. Akkerman, P. Masereeuw, W. Meijs, *Designing a Computerized Lexicon for Linguistic Purposes,* Amsterdam, Rodopi, 1985.

[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, 21 edition, 1992.

[4] R. Garside, G. Leech, G. Sampson, *The Computational Analysis of English. A Corpus-Based Approach,* Londres, Longman,1987.

[5] J. González, J. M. Goñi, A. Nieto, «Aries: a ready for use platform for engineering spanish-processing tools», en *Digest of the Second Language Engineering Convention,* págs. 219-226, October 1995.

[6] J. Hallebeek, «A Formal Approach to Spanish Grammar», *Language and Computers: Studies in Practical Linguistics,* 1992.

[7] S. Rodríguez, J. Carretero, «Building a spanish speller», en *Taller sobre Software de Libre Distribución,* Universidad Carlos III de Madrid, 1995.

[8] F. Sánchez, A. Nieto, «Development of a Spanish Version of the Xerox Tagger», Technical Report CRATER/WP6/FR1, CRATER Project, CEC, 1995.

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

SANTIAGO RODRÍGUEZ-JESÚS CARRETERO
Universidad Politécnica de Madrid