Generación automática de familias morfológicas mediante morfología derivativa productiva*

Jesús Vilares, Miguel A. Alonso
Departamento de Computación
Universidad de La Coruña
Campus de Elviña s/n
15071 La Coruña
jvilares@mail2.udc.es
alonso@dc.fi.udc.es

David Cabrero
Departamento de Informática
Universidad de Vigo
Edificio Politécnico, As Lagoas
32004 Orense
cabrero@uvigo.es

Resumen En este artículo se propone la utilización de mecanismos de morfología derivativa productiva con el fin de agrupar en una misma familia morfológica a todas aquellas palabras que se derivan de una misma raíz gramatical. En particular, se propone la creación de un sistema automático para la generación de dichas familias, el cual utiliza escasos recursos lingüísticos y cuya ejecución conlleva un bajo coste computacional. Dicho sistema es aplicado en tareas de normalización de términos simples para así mejorar la eficiencia de los sistemas de recuperación de información que trabajan con colecciones de textos en español.

1 *Introducción*

La riqueza léxica y morfológica del español queda reflejada en la gran productividad y flexibilidad que presentan sus mecanismos de formación de palabras, lo que conlleva una morfología derivativa rica y compleja. En efecto, el mecanismo preferido para la formación de nuevas palabras en español es la derivación, a diferencia de otros idiomas como el inglés que hacen un uso mayor de la composición.

Definimos una *familia morfológica* como el conjunto de palabras obtenidas a partir de una misma raíz morfológica mediante la aplicación de mecanismos de derivación. Es de esperar que exista una relación semántica entre las palabras de dicho conjunto, relación que normalmente es de tipo proceso-resultado (por ejemplo *fijación-fijado*), proceso-agente (por ejemplo *inhibición-inhibidor*), y similares. A la hora de obtener unos patrones regulares de formación de palabras podemos valernos de las llamadas *reglas de formación*, basadas en teorías tales como la Gramática Generativa Transformacional y el desarrollo de la denominada Fonología Derivativa. Aunque dicho paradigma no es completo, supone un avance considerable puesto que nos permite diseñar un sistema de generación automática de familias morfológicas con un grado aceptable de corrección y completud.

El resto de esta sección se dedica a describir brevemente los elementos involucrados en la formación de palabras en español. Los mecanismos de derivación y las condiciones fonológicas que conllevan se analizan en las secciones 2 y 3, respectivamente. En la sección 4 se describe la implementación de un sistema para la generación automática de familias morfológicas. La sección 5 considera la aplicación de este sistema a la tarea de normalización de términos simples en sistemas de recuperación de infomación y muestra los resultados obtenidos con diferentes motores de indexación. Finalmente, la sección 6 expone las conclusiones obtenidas.

1.1 La formación de palabras en español

Definimos un *morfema* como una unidad gramatical mínima distintiva que ya no puede ser significativamente subdividida en términos gramaticales. Los morfemas antepuestos al primitivo se denominan *prefijos*, y los pospuestos, *sufijos*. En español existen además *infijos*, elementos que aparecen intercalados en el interior de la estructura de un derivado.

Podemos también clasificar los morfemas en *flexivos* y *derivativos*. Los morfemas flexi-

* Deseamos expresar nuestro agradecimiento a Margarita Alonso por sus valiosos comentarios y sugerencias. Este trabajo ha sido financiado en parte por el Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (TIC2000-0370-C02-01), los fondos FEDER de la EU (1FD97-0047-C04-02) y la Xunta de Galicia (PGIDT99XI10502B).

vos representan conceptos gramaticales tales como género, persona, modo, tiempo y aspecto. Los morfemas derivativos producen un cambio semántico respecto al lexema base, y frecuentemente también un cambio de categoría sintáctica. El elemento común a ambos es conocido como el morfema léxico constituyente de la raíz.

Tradicionalmente la formación de palabras ha sido dividida en *composición* y *derivación*. Hablamos de composición cuando combinamos lexemas independientes, y de derivación cuando alguno de los componentes (un morfema derivativo) no puede aparecer como tal lexema independiente, incluso en el caso de que se le pueda asignar un contenido semántico. Se trata en ambos casos de procedimientos morfológicos, unión de morfemas individuales o grupos de morfemas en unidades superiores para formar lexemas complejos. Los nuevos lexemas obtenidos pueden actuar, a su vez, como bases para la formación de nuevas palabras.

La estructura morfémica es, de hecho, fundamental para el análisis de los procesos de formación de palabras. Debemos tener en cuenta sin embargo que en las lenguas románicas, y especialmente en español, la propia palabra antes que cualquiera de sus componentes morfémicos constituye la base de la derivación, siguiendo la mayor parte de sus palabras la estructura *morfema léxico* + *prefijo* o *sufijo*.

2 Mecanismos de derivación

Los mecanismos básicos de derivación en español son la prefijación, la sufijación apreciativa, la sufijación no apreciativa, la parasíntesis y la derivación regresiva. La prefijación consiste en la adición de un prefijo a una forma base, la sufijación en la adición de un sufijo y la parasíntesis en la adición simultánea de un prefijo y un sufijo[1].

En el caso de la sufijación, debemos distinguir entre sufijación apreciativa, que altera semánticamente el lexema base de un modo subjetivo emocional pero sin cambiar su categoría gramatical[2], y sufijación no apreciativa, que involucra un cambio fundamental más que marginal en el significado del lexema base, frecuentemente acompañado de un cambio de categoría sintáctica.

El repertorio de sufijos no apreciativos del español cuenta con cientos de morfemas derivativos, cuyo inventario no está fijado, ni tampoco sus restricciones, extensión o cambios de toda índole. Uno de los problemas con el que nos encontramos debido al elevado número de sufijos existentes, es el de su clasificación. En nuestro caso, el criterio que hemos tomado es doble. Por una parte, en función de la categoría gramatical del derivado, podemos hablar de: *nominalización* para dar sustantivos (el más común), *adjetivización* para dar adjetivos y *verbalización* para obtener verbos. El segundo criterio es conforme a la categoría gramatical de la base: sufijos *denominales* (a partir de sustantivos), *deadjetivales* (a partir de adjetivos) y *deverbales* (a partir de verbos y los más frecuentes).

Desde el punto de vista semántico todos los sufijos no apreciativos son significativos dado que el significado del derivado es siempre diferente del que poseía la base. El problema viene dado porque muchos sufijos involucran polisemia[3].

En lo referente a la derivación regresiva, esta juega un importante papel en el español contemporáneo como mecanismo para la formación de sustantivos a partir de verbos. Su característica principal radica en que, en lugar de incrementar el tamaño del lexema base, provoca una reducción del mismo, al añadir tan solo una vocal a la raíz verbal[4].

Un fenómeno importante a tener en cuenta es la existencia de *alomorfos*, variantes de un mismo morfema derivativo[5]. El alomorfo a utilizar en cada caso puede estar determinado por la fonología o venir impuesto por convención o por la etimología.

3 Condiciones fonológicas

No debe descuidarse el análisis de las condiciones fonológicas que preside el proceso de

[1] Por ejemplo, la derivación de *enrojecer* a partir de *rojo*, que sólo puede realizarse en un único paso al no existir las formas intermedias **enrojo* ni **rojecer*.

[2] Los sufijos apreciativos pueden subdividirse en diminutivos, que transmiten una idea de pequeñez o afectividad; aumentativos, que implican amplia dimensión, fealdad o grandiosidad; y peyorativos, que implican desagrado o ridiculez.

[3] Por ejemplo, el sufijo *-ada* involucra los significados de "nombre colectivo" y "acción propia de la base". Así, a partir de *muchacho* obtenemos el derivado *muchachada*, que puede significar "grupo de muchachos" o "acción propia de muchachos".

[4] Por ejemplo, a partir de *deteriorar* obtenemos el derivado *deterioro*.

[5] Por ejemplo: *innecesario*, *imprudente*, *irreal*.

formación de palabras, puesto que toda operación morfológica implica a su vez una alteración fonológica de la base. Dichas alteraciones pueden ser regulares o aparentemente irregulares. Consideremos los siguientes ejemplos:

a) *responder* → *respondón*
vencer → *invencible*
grosero → *grosería*

b) *pan* → *panadero*
agua → *acuatizar*
carne → *carnicero*

En a) los sufijos se adjuntan a sus bases o raíces de manera morfológicamente regular, con resultados previsibles tanto morfológica como fonológicamente, mientras que en b) este proceso semeja irregular, puesto que el resultado no es el esperado (**pandero*, **agüizar*, **carnero*). Sin embargo algunos de esos fenómenos son lo suficientemente frecuentes como para que deban ser forzosamente incluidos en cualquier estudio riguroso sobre morfología léxica de la lengua española.

El esfuerzo por intentar explicar situaciones como las indicadas conllevó la creación de las llamadas *reglas de reajuste*[6]. Aronoff [Lang 1992] clasifica dichas reglas en dos grupos: alomórficas (relativas a aspectos de alomorfia) y de truncamiento (aquellas reglas que suprimen un morfema terminal ante la adjunción de un nuevo sufijo).

Las condiciones fonológicas concretas que se han considerado en este trabajo se detallan en la sección 4.3.

4 *Implementación*

Tomando como base los fundamentos teóricos expuestos en las secciones precedentes, hemos desarrollado un sistema para la generación automática de familias morfológicas.

Llegados a este punto, debemos llamar la atención sobre uno de los grandes problemas a los que deben hacer frente las técnicas de Procesamiento del Lenguaje Natural (PLN) cuando son aplicadas al español: la escasa disponibilidad de grandes recursos lingüísticos libremente accesibles (corpora etiquetados, bancos de árboles, diccionarios avanzados, etc.). Trataremos de sobreponernos a estas limitaciones creando un sistema lo más simple posible, afrontando nuestra tarea desde un nivel léxico. Los recursos lingüístico-computacionales requeridos son mínimos, empleando únicamente un lexicón, cada una de cuyas entradas contiene una forma base, su etiqueta y su lema correspondiente.

4.1 Descripción del algoritmo

Para facilitar la comprensión del algoritmo explicaremos su funcionamiento mediante un ejemplo práctico correspondiente a la generación de una supuesta familia morfológica {*rojo*, *enrojecer*} a partir del término base *rojo*.

Por cada nuevo lema que va a ser procesado, creamos una familia morfológica F, que inicialmente contendrá únicamente ese lema. En nuestro ejemplo, F={*rojo*}. En este momento, F es la familia activa. El lema *rojo* es apilado en S, una pila que mantiene los componentes todavía no procesados de la familia activa. En nuestro ejemplo, S=[*rojo*] en este instante.

Mientras S no se vacíe y existan componentes no procesados en F, se realizan las siguientes acciones:

1. Se extrae el lema situado en la cima de la pila S y se le aplican los mecanismos de derivación acordes con su categoría gramatical. La validez de las palabras derivadas (en sus respectivas categorías) es contrastada por medio del lexicón: sólo si están presentes en el mismo se consideran válidos. Si un derivado no es válido, se le aplican las condiciones fonológicas para tratar de obtener uno que sí lo sea. En nuestro ejemplo, *rojo* es extraído de la pila (con lo que ésta queda vacía) y mediante la parasíntesis *en- -ecer* se deriva el verbo *enrojecer*, que es identificado como un derivado correcto puesto que pertenece al lexicón.
2. Si se ha obtenido un derivado válido:
   (a) Si el derivado no había sido previamente procesado, se incluye en F y se apila en S para ser procesado posteriormente. En nuestro ejemplo, F ={*rojo, enrojecer*} y S =[*enrojecer*] en este momento.

[6] Por ejemplo, de *pan* /pan/ obtenemos el derivado *panadero* /panaδero/ mediante la inserción de /aδ/ entre la raíz y el sufijo. En el caso de *agua* /aγua/ derivamos *acuatizar* /akwatiθar/ a partir de la inserción de /t/ antes del morfema de infinitivo.

(b) Si el derivado ya había sido procesado previamente y además pertenece a una familia $F' \neq F$, ambas, F y F', se refieren a subconjuntos de una misma familia morfológica. En tal caso, todos los lemas de la familia activa F son asignados a F' y F' pasa a ser la familia activa. Denominamos a este fenómeno *transitividad derivativa*. En nuestro ejemplo tendría lugar si, por ejemplo, el lema *enrojecer* hubiese sido procesado antes que *rojo*. En este caso, la familia morfológica que inicialmente se habría obtenido sería $F' = \{enrojecer\}$. Más tarde, el lema *rojo* sería procesado en F, obteniendo entonces *enrojecer* como derivado de *rojo*. Como consecuencia, F y F' se fusionarían en una única familia $\{rojo, enrojecer\}$.

Podemos observar que el algoritmo opta por sobregenerar, es decir, aplica todos los sufijos posibles obteniendo todos los derivados morfológicamente válidos, los cuales son filtrados mediante el lexicón. De este modo resolvemos el problema de la decisión sobre la validez y aceptación del término derivado a través únicamente de la forma léxica y de su etiqueta, sin considerar otros aspectos.

La derivación regresiva se implementa de modo indirecto por medio de la transitividad derivativa: en vez de derivar el sustantivo a partir del verbo, se espera a que el sustantivo sea procesado para obtener el verbo mediante verbalización denominal.

4.2 Tratamiento de la alomorfia

Muchos morfemas derivativos presentan formas variables, alomorfos, en ocasiones determinados fonológicamente, pero en otras impuestos léxicamente por motivos etimológicos o convencionales. Existen variantes que son excluyentes entre sí, como pueden ser aquellas cuya elección viene dada por la vocal temática[7]. Existen otras que no lo son, como *-ado* y *-azgo* (variantes arcaica y culta de una misma forma), dando lugar a veces a resultados diferentes aplicados a la misma base[8]. La variante a aplicar en cada caso dependerá de cada sufijo en particular, e incluso influirán factores como la vocal temática y la forma de la base. Por lo tanto hemos considerado e implementado los diferentes casos para cada sufijo particular.

4.3 Tratamiento de las condiciones fonológicas

Hemos considerado las siguientes reglas de ajuste fonológico [2, 3]:

- *Supresión de la vocal final átona*: constituye el comportamiento por defecto del sistema. A la hora de concatenar los sufijos el sistema trabaja en principio sobre el término base, eliminado la vocal final en el caso de sustantivos y adjetivos; en caso de finalizar en consonante no se modifica[9]. En todo caso, el término original siempre está disponible.
- *Eliminación de cacofonías*: en ocasiones, al concatenar el sufijo a la raíz obtenida mediante el proceso anterior, dos vocales iguales quedan adyacentes. Ambas se fusionan para eliminar la cacofonía resultante.
- *Vocal temática*: en el caso de que el término primitivo sea un verbo, basta con comprobar si acaba en *-ar/-er/-ir/-ír* para conocer la vocal temática y así ser tenida en cuenta, por ejemplo, a la hora de escoger la variante alomórfica a utilizar. Una muestra es el caso de *-miento/-amiento/-imiento/-mento*, donde *-amiento* sólo se emplea con vocal temática *a* e *-imiento* con las vocales *e* e *i*.
- *Monoptongación de la raíz diptongada*: se sustituye el diptongo por la forma pertinente. Se considera la monoptongación de *ie* en *e*[10] y de *ue* en *o*[11].
- *Cambio en la posición del acento*: puesto que los sufijos producen generalmente un cambio en la acentuación, dicha situación debe ser considerada, ya que puede conllevar cambios ortográficos debidos a la aparición o desaparición de tildes. La práctica totalidad de los sufijos son tónicos, con lo que es inmediato saber si debemos introducir o eli-

[7] Por ejemplo, *-amiento* para la vocal *a* e *-imiento* para las vocales *e* e *i*.

[8] Por ejemplo, *líder* da lugar a *liderato* y *liderazgo*.

[9] Por ejemplo, *arena* $\rightarrow$ *aren-* $\xrightarrow{-oso}$ *arenoso*. Análogamente, *temor* $\rightarrow$ *temor* $\xrightarrow{a- \; -izar}$ *atemorizar*.

[10] Por ejemplo al derivar *dental* a partir de *diente*.

[11] Por ejemplo al derivar *forzudo* a partir de *fuerza*.

minar una tilde aplicando las reglas ortográficas pertinentes.

- *Mantenimiento de los fonemas consonánticos finales*: conociendo el fonema, podemos deducir la ortografía final[12]. Los fonemas y cambios cubiertos son:

$$\begin{array}{ll} /\mathrm{k}/ & \mathrm{c} \rightarrow \mathrm{qu} \\ /\gamma/ & \mathrm{g} \rightarrow \mathrm{g\ddot{u}} \\ /\gamma/ & \mathrm{g} \rightarrow \mathrm{gu} \\ /\theta/ & \mathrm{z} \rightarrow \mathrm{c} \\ /\theta/ & \mathrm{c} \rightarrow \mathrm{z} \end{array}$$

- *Reglas ad-hoc*: nos referimos a ajustes varios tales como modificaciones en la consonante final de la raíz en casos como la derivación de *concesión* a partir de *conceder*. Estos casos se resuelven mediante reglas ad-hoc, es decir, que operan para un sufijo dado. Frecuentemente vienen dados por la presencia de fonemas dentales $/\delta/$ o $/t/$.

4.4 Evaluación

El sistema de generación automática de familias morfológicas ha sido aplicado sobre un lexicón de 995.859 formas, 92.125 de las cuales corresponden a lemas de palabras con contenido (sustantivos, adjetivos, verbos), obteniendo como resultado 54.243 familias morfológicas. En la tabla 1 se muestra el número de palabras que componen cada una de las familias.

Para evaluar el sistema se tomó una muestra aleatoria de familias de dos o más componentes, las cuales fueron examinadas manualmente una por una, comprobando con ayuda de diccionarios no sólo si las palabras involucradas pertenecían o no a la misma familia, sino también si la relación semántica era lo suficientemente fuerte como para ser útiles a nuestros propósitos. No se emplearon diccionarios especiales en el proceso, sino diccionarios comunes. La razón para ello vino motivada por la naturaleza del lexicón empleado, muy completo pero abundante en palabras de uso marginal, americanismos, etc. De este modo aquellas palabras que no tenían entrada en los diccionarios consultados se consideraron de uso infrecuente y no se tuvieron en cuenta a la hora de la evaluación. El porcentaje de dichas palabras ascendió a un 20%.

[12] Por ejemplo, la *z* en *cerveza* corresponde a $/\theta/$ y por consiguiente la *c* de cervecería corresponde también a $/\theta/$ y no a /k/.

	Familias		Palabras	
Tamaño	Número	%	Número	%
1	43.007	79,29	43.007	46,68
2	4.470	8,24	8.940	9,70
3	2.314	4,27	6.942	7,54
4	1.405	2,59	5.620	6,10
5	904	1,67	4.520	4,91
6	501	0,92	3.006	3,26
7	368	0,68	2.576	2,80
8	270	0,50	2.160	2,34
9	223	0,41	2.007	2,18
10+	781	1,43	13.347	14,49
Total	54.243	100	92.125	100

Tabla 1: Composición de las familias

Un familia se consideró incorrecta cuando alguno de sus miembros pertenecía a otra familia. En la tabla 2 se muestran los resultados obtenidos con respecto a las familias de dos o más componentes y con respecto al conjunto global de familias.

	2+	Global
correctas	79 %	95,59 %
incorrectas (2 fam.)	7 %	1,47 %
incorrectas (3 fam.)	12 %	2,52 %
incorrectas (4+ fam.)	2 %	0,42 %

Tabla 2: Evaluación de familias morfológicas

Los resultados pueden considerase más que aceptables teniendo en cuenta que estamos ante una aproximación léxica a un problema cuya solución completa requeriría información semántica e incluso contextual, con el consiguiente incremento del coste computacional.

Hemos analizado las fuentes de error, determinando que las más frecuentes son:

- El agrupamiento de diferentes familias morfológicas mediante transitividad derivativa, siendo los casos de riesgo:
  1. Palabras con grafía similar, especialmente aquellas muy cortas[13].
  2. Monoptongación de diptongos[14].

[13] Por ejemplo, de *ano* obtenemos *anal*, palabra también obtenida desde *ana* (medida de longitud).

[14] Por ejemplo, de *fuel* (carburante) podríamos obtener *folía* (baile).

3. Formaciones parasintéticas[15].

- Existencia de más de una acepción en el significado de una palabra[16].
- Especialización de significados[17].
- Sentidos figurados[18].

Existen mecanismos mediante los cuales podríamos reducir los casos de error, tales como la utilización de información etimológica o semántica para comprobar si la palabra candidata a derivado guarda o no relación con la palabra primitiva. Esto supondría unos costes notablemente mayores, sin que pudiésemos garantizar la desaparición total de errores. Consideremos por ejemplo el caso de palabras que aun manteniendo una relación etimológica, sus significados hayan variado considerablemente a lo largo del tiempo[19]. En cuanto a la utilización de información semántica, en el caso de palabras con más de una acepción, habría que desambiguar el sentido de cada palabra a partir del contexto en el que ocurre, lo que elevaría considerablemente los costes computacionales.

5 *Aplicación a la recuperación de información*

En tareas de Recuperación de Información (RI), los documentos de una colección son representados en forma de conjuntos de términos índice o palabras clave. Aunque las nuevas generaciones de ordenadores están haciendo posible la representación de un documento por su conjunto completo de palabras, al trabajar con grandes colecciones de documentos todavía debemos seguir limitando el conjunto de palabras clave representativas. Para ello se recurre a operaciones tales como la eliminación de *stopwords* (palabras excesivamente frecuentes y sin significación aparente) o técnicas de *stemming* (las cuales reducen las palabras a una supuesta raíz gramatical). A dicho tipo de operaciones se las denomina *operaciones de texto*, y generan una *vista lógica* del documento procesado.

En efecto, los sistemas de RI normalizan los documentos antes de su indexación para decrementar su variabilidad lingüística mediante la agrupación de términos referentes a conceptos similares explotando para ello similitudes gráficas, thesaurus, etc. [1, 6]

En el presente trabajo analizaremos un nuevo mecanismo de normalización, el empleo de familias morfológicas. Para ello, primeramente obtendremos las etiquetas y lemas correspondientes al texto a indexar por medio de un etiquetador-lematizador. A continuación, sustituiremos cada uno de los lemas obtenidos por un identificador de su familia morfológica. Como consecuencia, todos los componentes de una familia morfológica se representarán mediante un único término en el índice.

La evaluación de un sistema de recuperación de información implica calcular las medidas de *precisión* P y *cobertura* R, donde:

$$P = \frac{\text{nº de documentos relevantes recuperados}}{\text{nº total de documentos recuperados}}$$

$$R = \frac{\text{nº de documentos relevantes recuperados}}{\text{nº total de documentos relevantes}}$$

Hemos evaluado tres métodos de indexación diferentes:

pln: indexación del texto original sin stopwords.

lem: indexación de las palabras con contenido del texto lematizado.

fam: el texto es primero lematizado, y posteriormente cada lema de una palabra con contenido es sustituido por el identificador de su familia morfológica.

El corpus de referencia empleado para la evaluación está formado por 21.899 documentos periodísticos (artículos de nacional, internacional, economía, cultura, ...) abarcando la totalidad del año 2000. La longitud media de los documentos es de 447 palabras. Para realizar los experimentos se ha considerado un conjunto de 14 consultas en lenguaje natural. La longitud media de cada consulta es de 7,85 palabras, de las cuales 4,36 son palabras con contenido.

En la tabla 3 se muestran algunas medidas que caracterizan al corpus utilizado. La

[15] Por ejemplo, de *plasta* podríamos obtener *aplastar* (aplanar).

[16] Por ejemplo, *ranchero* debería derivarse de *rancho* (granja) pero no de *rancho* (comida).

[17] Por ejemplo, *golpeador* (que golpea) puede obtenerse en general de *golpe*, excepto cuando esta palabra se refiere a un golpe de estado, en cuyo caso debería derivarse *golpista*.

[18] Por ejemplo, derivar *lincear* (advertir lo oculto) a partir de *lince* (animal).

[19] Es el caso de *Morfeo* (dios del sueño) y *morfina* (analgésico)

	Total	Únicos
original	9,780,513	154,419
pln	4,526,058	154,071
lem	4,625,579	111,982
fam	4,625,579	105,187

Tabla 3: Características del corpus utilizado

primera y segunda columna muestran, respectivamente, el número total de términos y el número de términos únicos obtenidos para los documentos indexados, tanto para su texto original como para sus diferentes representaciones normalizadas. Como se puede observar en la primera columna, la lematización y las familias morfológicas producen una reducción de más del 50% en número de términos a indexar. Con respecto al número de términos diferentes en el índice, mostrado en la segunda fila, se observa que la reducción resultante de la utilización de stopwords es despreciable, mientras que la utilización de lematización produce una reducción del 27% y la utilización de familias morfológicas proporciona una reducción del 32%, con el consiguiente ahorro de espacio y disminución del tiempo de acceso a los índices.

En los experimentos se han empleado tres motores de indexación diferentes, Altavista Search Development Kit[20], un motor de búsqueda comercial; SWISH-E[21], un motor de búsqueda basado en el modelo booleano [1]; y SMART[22], un motor de búsqueda basado en el modelo vectorial [1]. En las figuras 1, 2 y 3 se muestran los resultados obtenidos para cada uno de dichos motores de indexación.

Podemos observar que los métodos *fam* y *lem* mejoran considerablemente los resultados, tanto en precisión como en cobertura medias, respecto al método *pln*. Como sería de esperar, *fam* proporciona unos niveles más elevados de cobertura en Altavista y SWISH-E, mientras que su precisión es ligeramente inferior a la obtenida por *lem*. No ocurre lo mismo en el caso de SMART, donde *lem* aventaja ligeramente a *fam* tanto en precisión como en cobertura.

En cuanto a las gráficas de evolución de la precisión respecto a la cobertura podemos ver que en el caso de utilizar SWISH-E como

[20] `http://solutions.altavista.com/`

[21] `http://sunsite.berkeley.edu/SWISH-E/`

[22] `ftp://ftp.cs.cornell.edu/pub/smart/`

	pln	*lem*	*fam*
Precisión media	0,1982	0,2125	0,2232
Cobertura media	0,5956	0,6321	0,6769

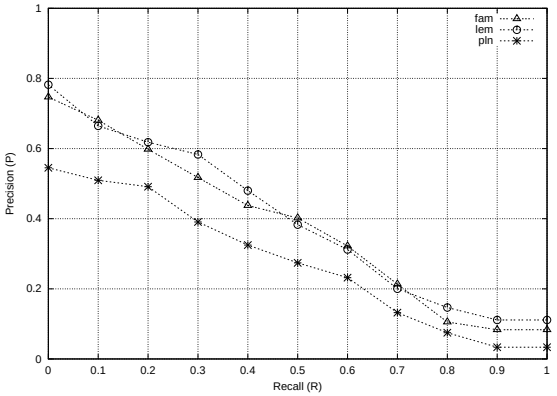

Figura 1: Precisión y cobertura con Altavista

	pln	*lem*	*fam*
Precisión media	0,3875	0,5171	0,4190
Cobertura media	0,1296	0,4104	0,4526

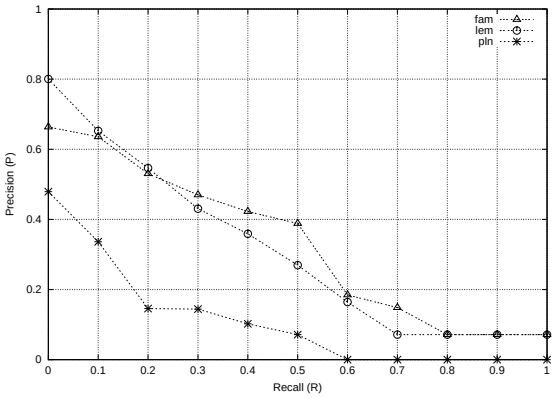

Figura 2: Precisión y cobertura con SWISH

motor de indexación, para los primeros documentos devueltos la lematización da unos resultados ligeramente superiores, sin embargo estos disminuyen rápidamente debido a que la cobertura global del método basado en familias morfológicas es mayor, devolviendo en general mayor número de documentos relevantes. En el caso de Altavista, ambos métodos están muy próximos en toda la gráfica, alternándose como el mejor en diversos segmentos de cobertura. En el caso de SMART, la lematizacióm muestra un rendimiento ligeramente superior en todo el recorrido de las curvas.

Es interesante remarcar que el comporta-

	pln	*lem*	*fam*
Precisión media	0,1714	0,2018	0,1982
Cobertura media	0,5515	0,6316	0,6028

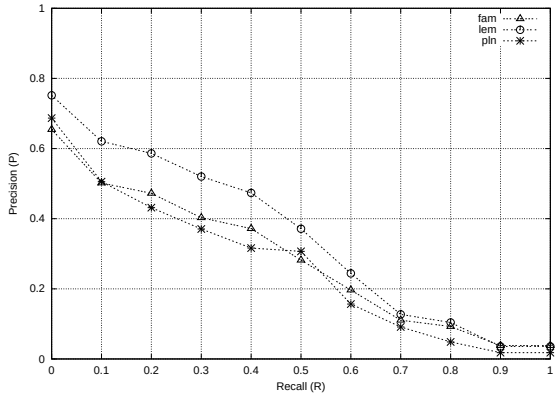

Figura 3: Precisión y cobertura con SMART

miento de los diferentes métodos de indexación varía de una consulta a otra. La indexación mediante familias morfológicas muestra claramente un mejor comportamiento en consultas que contienen palabras pertenecientes a familias morfológicas con cierta varibilidad lingüística. De este modo, se obtienen muy buenos resultados para todos los motores de búsqueda con consultas tales como *negociaciones del PP con el PSOE sobre el pacto antiterrorista*, puesto que en la familia morfológica de *negociación* encontramos palabras como *negociar*, *negociador* o *negociable*, mientras que en la familia de *pacto* encontramos palabras como *pactar* o *pactable*. Por contra, la lematización funciona mejor con consultas en las cuales aparecen involucradas palabras cuyas familias morfológicas gozan de poca variabilidad lingüística.

6 Conclusiones

En este artículo hemos mostrado que los mecanismos de derivación para la formación de palabras pueden ser formalizados y utilizado en la construcción de un sistema para la generación automática de familias morfológicas, conjuntos de palabras que comparten una misma raíz, utilizando para ello muy pocos recursos lingüísticos, tan solo un lexicón. Las familias obtenidas, aunque no son correctas al cien por cien, muestran un grado de corrección lo suficientemente alto como para poder ser aplicadas en tareas como la recuperación de información de textos en español, puesto que su bajo coste computacional y su independencia del motor de indexación permite que sean aplicadas a sistemas reales en este ámbito. Los resultados prácticos obtenidos en experimentos realizados con diversos motores de indexación sobre una colección de textos periodísticos avalan la utilidad de las familias morfológicas en dicho ámbito.

Referencias

[1] Baeza-Yates, R., Ribeiro-Neto, B.: Modern information retrieval. Harlow, England: Addison-Wesley (1999)

[2] Bajo Pérez, E.: La derivación nominal en español. Madrid: Arco Libros, Cuadernos de lengua española (1997)

[3] Fernández Ramírez, S.: La derivación nominal. Madrid: Real Academia Española, Anejos del Boletín de la Real Academia Española **40** (1986)

[4] Jacquemin, C.: Syntagmatic and paradigmatic representations of term variation. Proc. of 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), University of Maryland, USA

[5] Jacquemin, C., Tzoukermann, E.: NLP for Term Variant Extraction: A Synergy of Morphology, Lexicon and Syntax. In T. Strzalkowski, editor, Natural Language Processing Information Retrieval, páginas 25–74. Boston: Kluwer (1999)

[6] Kowalski, G.: Information retrieval systems: theory and implementation. Boston: Kluwer (1997)

[7] Lang, Mervyn F. Formación de palabras en español: morfología derivativa productiva en el léxico moderno. Madrid: Cátedra, Lingüística (1992)

[8] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. J.: Introduction to WordNet: An On-line Lexical Database. International Journal of Lexicography **3** (1990) 235–244

[9] Strzalkowski, T. (editor): Natural language information retrieval. Dordrecht: Kluwer (1999)