# Análisis morfológico como ayuda a la recuperación de información

M.Meya
Siemens S. A. -CDS
Barcelona

Resumen:

El análisis morfológico realizado con una red de estados finitos es un medio para poner a disposición de un usuario de retrieval ayudas que le capaciten a encontrar los términos para él relevantes. El sistema MARS (=Morphological Analisis for Retrieval Support) pone a la disposición de un usuario de un Banco de Datos la posibilidad de acceder a los datos mediante una búsqueda que es asistida por un paquete lingüístico. Este paquete lingüístico descompone las palabras que el usuario da la ordenador a la vez que le ofrece todas las palabras que lingüísticamente estén emparentadas con la originaria.

La enorme ventaja de un sistema así frente a cualquier otro convencional que opere con "máscaras" o plantillas, es que en los sistemas convencionales se pueden dar comodines para ciertas posiciones dentro de una palabra, pero con el inconveniente de que el ordenador en la mayoría de los casos responde a su vez en exceso "excesivo recall", mientras que en otros casos, segun se trunquen las palabras el sistema accede a entradas erroneas.

Con un procedimiento lingüístico como lo es MARS un usuario de retrieval recibe solo las acepciones que ciertamente están emparentadas con el término preguntado.

El presente artículo presenta la gramática que descompone las palabras, y mediante la cual se segmentan en sus morfemas correspondientes.

Como efecto secundario de la aplicación de MARS a varios textos y como ayuda a la recuperación de la información se obtuvieron tipos de combinaciones de morfemas de acuerdo con el tipo de texto investigado.

## 1 Problemática de la descomposición morfológica automática

El objetivo del sistema MARS es descubrir las relaciones morfológicas entre los términos de un corpus almacenado. Estas relaciones se ponen a disposición del usuario o erudito que quiera cuestionar a una base de datos léxica, bibliográfica, de referencias científicas, etc..

Todo sistema que quiera analizar palabras debe decidirse por una de las siguientes alternativas:

a) trabajar sin diccionario y sólo con reglas heurísticas
b) trabajar con un diccionario de morfemas
c) trabajar con un diccionario de formas base - lema - y hacer sólo el análisis de plurales y de algunas formas de conjugación regulares
d) trabajar con un diccionario de formas de palabra (ejemplares)

Las posibilidades aquí mencionadas se mueven en la escala que va de mayor procesamiento y reducida capacidad de memoria, a sólo búsqueda de

diccionario y cantidades enormes de listas.

El que se haya decidido en favor de la posibilidad (b) viene dado por dos razones:
- un diccionario de morfemas es mucho más reducido que uno de lemas o de palabras.
- la búsqueda para la recuperación de la informació es mucho más precisa ya que entran en consideración los aspectos léxicos de derivación y composición.

Uno de los aspectos más interesantes que ofrece el español frente a otras lenguas, es la alternancia vocálica según el acento prosódico, y otro es la alternancia consonántica debida a aspectos diacrónicos de la lengua. Esto hace que un morfema determinado tenga un montón de morfos - variantes - , sobre todo los morfemas verbales : contar/cuento, sentir/siento/sintió,construir/construyo/ constructor....etc

Estas transformaciones causadas por la prosodia y otras que son causa de la ortografía (entreg-ar/entregu-e) son totalmente regulares y pueden ser tratadas adecuadamente.

**2 Datos del sistema**

El sistema posee bases de datos de diversa índole como:

a) ficheros invertidos con la BD de referencia .(son los datos a los que el usuario accede)
Estos datos se han extraido del siguiente corpus:
- glosario terminológico de telefonía (Telefónica)
- artículos del periódico el PAIS (Pais)
- entrevistas habladas del proyecto "el habla de Sevilla"(E)
- Texto literario: historia corta.(Short-story)

He aquí los datos desglosados en ejemplares y muestras.

| Tipo de texto | Token | Type |
|---|---|---|
| Telefonía | 140.000 | 8.135 |
| el Pais | 9.628 | 2.053 |
| short-story | 10.900 | 1.593 |
| entrevista | 3.983 | 753 |
| Total | 164.511 | 12.534 |

Sobre la lista de ejemplares (type) se realizó la descomposición morfológica automática y se fijaron dinámicamente punteros entre las palabras que pertenecían al mismo morfema aunque estas presentaran diferentes morfos.

b) Datos lingüísticos:

Como datos del sistema se consideran todos aquellos listados con información lingüística que son necesitados para realizar la descomposición morfológica. Estos son:
- una gramática de estados finita
- una lista con los morfemas del castellano (actualmente ca. 7.000)
- una lista de las transformaciones grafémicas
- una tabla de la jerarquía entre segmentaciones alternativas

c) Autómata de estados finitos.

La descomposición de las palabras se hace mediante un autómata de estados finitos. El proceso va agrupando grafías hasta que encuentra en el diccionario de morfemas una entrada igual. En este punto el automáta ha reconocido una posible segmentación. Si luego los rasgos que tiene en el diccionario tal morfema, le reconocen como adecuado para el tipo de estado en que se encuentra actualmente el autó - mata se consume tal estado y se pasa al siguiente.
Para el español se especificaron 15 estados que corresponden a a las clases de morfemas que se presentarán más adelante. El autómata tiene, pues, 16 estados: estado inicial,y 15 estados morfológicos de los cuales algunos están marcados como posibles estados finales.

La representación de reglas morfológicas como un automata de estados finito tiene la ventaja, de su fácil implementación y de que se logra un procesamiento altamente rApido por lo que se le debe considerar como un método óptimo para procesamientos morfológicos de cualquier índole.Así es usado para Text-to-speech conversion (Allen 1980 ,Hunnicutt 1976), Recuperación de la Información (MARS, 1984) y es usada por numerosos procesadores morfológicos (Kay 1986),(Kaplan 1985), Kartunnen, etc...

**3 Gramática morfológica del español**

La descomposición de una palabra en sus morfemas se hace a partir de 15 estados y las transiciones entre estos. Cada estado corresponde a una clase de morfemas distinta.
La gramática usa las siguientes clases morfológicas.

- Pre-prefijo
- prefijo no adverbal
- prefijo adverbal
- prefijo denominal
- morfema léxico no-verbal
- morfema léxico ligado
- morfema léxico verbal
- pre-sufijo no-adverbal
- pre-sufijo adverbal
- sufijo no-adverbal
- sufijo adverbal
- flexión no verbal
- flexión verbal
- enclíticos

Adverbal = morfema contiguo a una raíz verbal.
Denominal= morfema que transforma la categoría nominal de la raíz.

Una subcategorización de este tipo presupone saber lingüístico, es decir conocimientos acerca de la capacidad de combinación de los morfemas entre sí. Cuanto más refinada y exacta sea una subclase de morfemas, tanto más se reducen las posibles segmentaciones de las palabras. Con los datos procesados se obtuvio en un 40% de los casos una sola segmentación, y en un 20 % dos segmentaciones, ambas correctas ya que se trataba de alternativas para homográfos. En los restantes casos el ordenador arrojaba diversas segmentaciones, originadas por distintas motivaciones - entre otras por razones de diacronía - , pero entre las cuales siempre está la correcta.

La gramática se esquematiza en:

Prefijo /β/ raíz /β/ Sufijo /β/ Flexión/β/ Enclítico

en donde /β/ es exponente de la combinación permisible entre morfemas.

Así por ejemplo, cuando dentro de una segmentacion morfológica tenemos un segmento que es del tipo 9, estado 9, lo que es equivalente a decir que es un segmento no verbal por lo que entonces el morfema siguiente sólo puede ser del tipo 11, sufijo del tipo 11 y que además tiene que cumplir las condiciones pertinentes de posición en la palabra; Esta restricción le viene dada en la gramática.

Segun esto los sufijos están subcategorizados de acuerdo con determinadas reglas de combinación con los morfemas precedentes, y de acuerdo con su posible **posición dentro de la derivación**. Esta subcategorización especifica si los sufijos pueden ir inmediatamente después de la raíz, en posición intermedia, final o si aparecen sólo solos.

La gráfica 1 recoge las posibles combinaciones morfémicas en español ,de acuerdo con las categorías:

P 2 : Prefijo ante prefijo
P 3 : prefijo anterior a raíz no verbal
P 4 : prefijo anterior a raíz no verbal
P 5 : prefijo denominal
L 6 : morfema léxico no verbal
L 7 : morfema léxico ligado ( no independiente)
L 8 : morfema léxico verbal
S 9 : sufijo anterior a sufijo no verbalizador
S 10: sufijo anterior a sufijo y verbalizador
S 11: sufijo no verbal
S 12: sufijo verbal
F 13: flexión no verbal
F 14: flexión verbal
PR15: enclíticos (pronombres)

En esta gráfica las **líneas discontinuas significan afijo transformadores de la categoria anterior** . Las líneas continuas significan que la categoría existente continúa después de tener el afijo en cuestión.

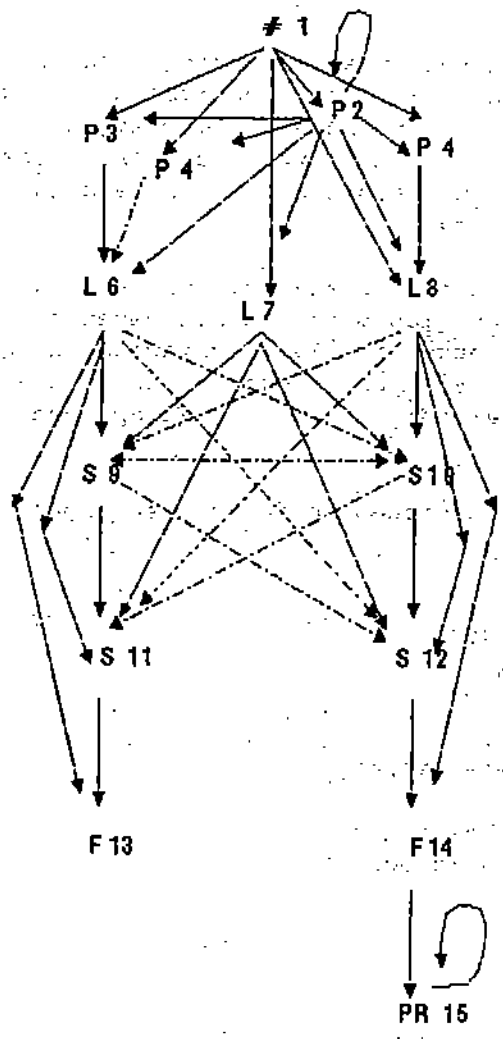
Por ejemplo, el sufijo "os" da como resultado una base que continua siendo no verbal;por consiguiente, este es un sufijo de tipo 9 que puede tener a su vez otro sufijo detrás suyo:

**"carin- os - o" / "carin-os-isim-o"**
**"estudi-os-a" / "estudi-os-ament-e"**

La gramática morfológica aquí presentada se basa fundamentalmente en dos principios: Restricciones lingüisticas, que se derivan de las distintas categorias morfosintácticas que resultan de la combinación de morfemas, y por otro lado en **restricciones acerca de la distribución de los morfemas**.

Una tarea muy interesante sería definir un modelo de Markov que operara sobre los resultados de segmentaciones morfologicas obtenidas con MARS. A partir de la frecuencia de ciertos morfemas en ciertas posiciones se podría especificar la frecuencia de aparición de un morfema en una lengua dada.

Fig. 1 : Diagramm of the Spanish Grammar



**4 Listado de morfemas**

El listado de morfemas de espanol contiene actualmente 7000 morfemas. esto cubre todas las formas del "Diccionario de Frecuencias del español" (Juilland 1974) y los correspondientes morfemas de la mitad del diccionario de Slavy (Slavy & Grossmann 1975).

La lista contiene las siguientes clases morfémicas:

- morfemas libres
  **morfemas léxicos** :con y sin flexión . Por ejemplo :"árbol"(/arbol/)
  "nena"(/nen/).
  Contiene las siguientes subclases: raíz nominal (NOM), verbal(VRB), adjetiva (ADJ), adverbial (ADV), numerales (NUM),palabras funcionales (FW).

- morfemas ligados :
  **morfemas lexicos** : por ej.:. "fragancia" (/frag/)
  **Afijos : Prefijos, sufijos.**
  estos incluyen información acerca de su posición en la cadena, y acerca de la categoria originaria y resultante.
  **Flexión**: tipo de formante de la flexión verbal o no verbal.

Un aspecto importante que se mencionó más arriba es el de las transformaciones grafemáticas que se dan sistemáticamente en la lengua según el tipo de morfo de que se trate. Estas transformaciones son básicamente:

- variantes del morfema raíz cuando este termina en "z" y se le añade sufijos o flexiones que empiezan con vocales débiles:
  **"luz" --> "luc-es"**
  **"feliz" --> "felic-isimo" / "felic-es"**

- determinadas variantes de la raíz verbal en ciertos tiempos:
  **"cont-ar" --> "cuent-o"**
  **" perd-er" --> "pierd-o"**
  **" produc-ir"--> "produzc-o"**

Se han catalogado ocho tipos de transformaciones gráficas que son comprobadas en la cadena de entrada a partir de la información que lleva el morfema raiz, según este acepte transformaciones o no. Este procedimiento tiene dos ventajas:

- se descarga al diccionario de morfemas de entradas innecesarias
- se reduce al mínimo las entradas de morfos.

La lista de las formas fuertes, las supletivas queda así reducida a aquellas que no pueden ser captadas por reglas, o a aquellos casos en que procesar la sistematica de la regla seria muy costoso. Estos casos que aludimos aqui son casos de alternancia consonántica. Por ejemplo,

**absorb -- absor :absorber / absorcion**
**constru -- construct :construir/constructor**

Los ficheros invertidos contienen punteros que señalan, ligan las variantes entre sí.

## 5 Tabla jerárquica de la combinatoria de morfemas

El proceso de descomposición morfológica de las palabras lleva algunas veces a varias descomposiciones por palabra; esto se debe, como se apunto más arriba a:

- la evolución diacrónica de la palabra
- yuxtaposiciones de los morfemas entre sí

Ya que el sistema trabaja sin semántica se tenía que encontrar alguna forma de reducir estas descomposiciones alternativas. La reducción que permite esto se hace a partir de una tabla que contiene las posibles combinatorias de morfemas ponderadas. Esto quiere decir que la tabla expresa prioridades de rango de una descomposición morfológica sobre otra. Esta tabla ha sido creada a partir de saber lingüístico y no estadístico, por lo que tiene la ventaja que una vez creada es aplicable a cualquier tipo de texto, sea técnico o no. El que una combinación de morfemas tenga mayor prioridad que otra es un hecho que viene dado por las restricciones lingüísticas que se conocen acerca de la combinatoria de los morfemas.

Lo que el sistema hace es reordenar las descomposiciones alternativas en aquellos casos que se den y se decide por la primera. No obstante, si el usuario quiere ver todas las descomposiciones y buscar en el banco de datos con otra alternativa, puede hacerlo igualmente.

Para los linguistas y lexicógrafos es interesante constatar las distintas frecuencias que se dan de las combinatorias de los morfemas según la tipología de los textos. Así es sumamente interesante observar la distinta distribución que se da en los textos analizados,que son el periódico el PAIS, lengua hablada, una novelita y lenguaje de telefonía. No obstante, para poder hacer afirmaciones acerca de la tipología morfológica de los textos cara a su tipo de derivación o composición se tendrían que analizar muchos textos más. A pesar de todo el lector puede observar la siguiente tabla con los resultados de la descomposición de 165.000 muestras que corresponden a 12.534 ejemplares (types).

| EL PAIS | ENTREVISTA | SHORT-STORY | TELEFONIA |
|---|---|---|---|
| 713: 6 13 | 278: 6 13 | 489: 6 13 | 2705: 6 13 |
| 539: 8 14 | 240: 8 14 | 395: 8 14 | 2264: 8 14 |
| 159: 6 9 13 | 59: 6 | 74: 6 9 13 | 747: 6 9 13 |
| 151: 6 | 50: 6 9 13 | 34: 5 8 14 | 438: 6 |
| Total types | | | |
| 2053 | 753 | 1593 | 8135 |

La descomposición 6 & 13 significa raíz nominal (=6) y flexión nominal (=13). 8 es raíz verbal y 14 flexión verbal. Consultese tabla 1. Como puede observarse la derivación es más acusada en el lenguaje técnico y en el de la prensa. Si este fenómeno se cumple a grandes rasgos es algo que se tendría que comprobar con mayor cantidad de datos.

## 6 Caracteristicas especificas del español

A la vista de los resultados obtenidos a partir de la aplicación de MARS a datos de tres lenguas, el alemán, el inglés y el español, se pudieron observar fenomenos lingüisticos interesantes que tienen consecuencias no solo lingüísticas sino también para las técnicas de la recuperación de la información. Estas caracteristicas se han de tener en cuenta a la hora de diseñar sistemas que guien al usuario en su búsqueda para la recuperación de la información alamacenada en una base de datos.

Las características observadas quedan agrupadas en las siguientes rúbricas:

- **Marca del género y número**
  al contrario del inglés o alemán, en español los morfemas están marcados respecto a estos rasgos. Observense los ejemplos:

| palabra | morfemas | palabra | morfemas |
|---|---|---|---|
| "casa" | cas/a | Haus | Haus |
| "trabajo" | trabaj/o | work | work |

  es decir, en español no coinciden tanto como en otras lenguas los morfemas con las formas de palabras.

- **Marca del tipo de conjugación**

  Esta marca contiene información sobre el tiempo, persona, modo y aspecto

- **Transformaciones regulares del morfema raíz.**
  Estas transformaciones son básicamente: diptongación, altenancia consonantica y marca de acento. Las transformaciones están intimamente relacionadas con la lexicogénesis.

- **Fuerte derivación**
  En las lenguas románicas la derivación es el fenómeno base de la lexicogéneis, frente al proceso composicional de las lenguas germánicas. Esto tiene consecuencias a la hora de formular la lógica de búsqueda en un sistema de recuperación de la información. Mientras que inglés y alemán hay que buscar con operadores boolianos la combinatoria de dos términos, es español es suficiente, en muchos casos, buscar con un sólo término. Obsérvense los ejemplos:

| | | |
|---|---|---|
| **Haustuer** | **house door** | **puerta de la casa** |
| **Hausmeister** | **portier** | **portero** |
| **Hausmeisterwohnung** | **portier lodge** | **portería** |

- **Frecuente sincretismo**
  Se da el fenómeno de sincretismo cuando la forma de una palabra expresa diversos significados:

  **"control" --> el que controla, lo controlado, el lugar de control**
  **"dirección"--> el hecho de dirigir, la calidad de ser director, el lugar de despacho**

- **Capacidad generativa con neologismo "derivacionales"**
  "sandwich" "sandwicheria"
  "crepes" "creperia" etc...

- **Conversión ( derivación impropia)**
  La conversión es uno de los fenomenos más importantes para la forma-

ción de substantivos y adjetivos. La categoria puede cambiarse sin añadir formantes derivacionales. Por ejemplo:
"coser" -- "el coser"
"bello" -- "lo bello"

- **Frecuente derivaciones del tipo cero**
Se denominan derivaciones tipo cero aquellas cuyo determinado no tiene forma morfológica. Ej.:
"el corte" -- cort/e
"la caza" -- caz/a
"el empleo" -- emple/o

- **Gran cantidad de formas supletivas**
Este fenomeno es debido a la gran cantidad de latinismos en nuestra lengua. Ej.:
/hues/ vs. /os/ "hues-ud-o", "hues-o"
/nariz/ vs. /narig/ "nariz" vs "narigudo"
/fiel/ vs. /fidel/

## 7 Consideraciones cara a la recuperación de la información. Formulación y valorización de la petición.

El sistema MARS guia al usuario de un sistema de recuperación de la información a formular la petición de búsqueda lo más correcta posible, siempre dentro de la morfología aislada de las palabras - sin considerar las especificaciones de las palabras entre si, tal como hacen otras herramientas lingüisticas más complejas de Siemens (COPSY,REALIST)-. El sistema pretende darle al usuario referencias de otras formas lingüisticas que están relacionadas con el término en cuestión. En definitiva lo que se hace es ofrecer al usuario una relación de la familia de palabras que tienen que ver con un término determinado, y también de decirle en que pools y ficheros se encuentran para que este decida si el término que él pensaba es el adecuado o no para iniciar una busqueda.

En los sistemas de recuperación de información existentes hasta ahora apenas se hallan incorporados componentes lingüisticos. Si por el contrario la búsqueda de información se hace con herramientas lingüísticas obviamente estamos optimando el sistema, ya que entran en juego las caracteristicas especificas de cada lengua. Por ejemplo el concepto "Schreibmaschinenkoffer" se expresa en español mediante un sintagma : "maleta de máquina de escribir". Eso quiere decir que con un sistema como MARS que sólo opera con palabras aisladas el sistema le ofreceria las familias de palabras:

| | | | | | |
|---|---|---|---|---|---|
| /malet/ | maletín | /máquin/ | maquinista | /escribir/ | escribiente |
| | maletero | | maquinaria | | escritorio |
| | maletón | | maquinaje | | escritura |
| | | | | | escriba |

Siemens posee actualmente un sistema más complejo que consta del analizador morfológico de MARS y de otro componente que descubre las relaciones especificador/especificado dentro de sintagmas nominales.(REALIST 1987)

Si se contempla el analizador aislado, cuando un usuario quiere recuperar información de su base de datos este tiene que preguntar al sistema con el término que según él sea más relevante. En la mayoría de los casos se tratará del especificado; así en el ejemplo anterior seria

"maleta".

Lo que hace MARS es ayudar al usuario a buscar de la siguiente manera:
a) la demanda del usuario - palabras aisladas- son descompuestas
b) el sistema recoge de sus listas de terminos invertidos aquellos que corresponden al término de la búsqueda.
c) se le presentan al usuario todos los términos emparentados junto con la referencia de en qué fichero están y cuál es su frecuencia
d) el usuario puede ir almacenando sus demandas y las referencias ofrecidas por el sistema en su fichero : "user file".

El español por ser una lengua con una derivación muy marcada tiene la ventaja de que a partir de una raiz, el usuario puede acceder a multitud de variantes que ciertamente tienen que ver con el concepto de busqueda. En otras lenguas como el inglés o el alemán en la lógica de búsqueda se han de combinar las raices con operadores boolianos.
Por ejemplo:
"escritorio" , "Schreibtisch" en alemán, este significado es la conjunción de dos conceptos : "schreiben" /schreib/ (=escribir) y
"Tisch" /tisch/ (=mesa)

En español, por el contrario, dado lo productivo de la derivación, las raices tienen asociadas en el fichero invertido variedad de formas . Por ejemplo la raíz /escrib/ puede generar las siguientes derivaciones:

escribir --- schreiben
escriba --- Schriftgelehrter
escribano --- Notar
escribanía --- Schreibzeug, Notariat
escribiente --- Schreiber
escrito --- Schrifstueck
escritor --- Schriftsteller
escritorio --- Schreibtisch

**8 Lematización en español**

En el paquete de MARS se incluyó un componente que obtenía el lema a partir de la forma de la palabra. La lematización se hace mediante un proceso de pattern-matching, comparación de los patrones con información que tiene cada morfema -asiento en fichero de morfemas-.

La lematización es una rutina que aprovecha la información que se ha ido obteniendo a lo largo del proceso de descomposición morfológica de las palabras. Para llevar a cabo la lematización de las apariciones de las palabras, el sistema necesita la siguiente información, que se obtiene en la descomposición :

a) información acerca de las transformaciones realizadas sobre la cadena para obtener los morfemas
/ric/ isim/ as/ ---> lema: riquísimo

b) información acerca de las caracteristicas de los dos últimos morfemas en que se ha descompuesto la palabra.

Para decidir acerca de la categoría morfosintáctica de las bases de que consta una palabra se sigue un proceso idéntico.

La rutina que creo el lema de una palabra construye en un primer paso los componentes del lema a excepción de la flexión. En un segundo paso se selecciona la flexión adecuada, la cual se obtiene de la unión del

conjunto de rasgos del morfema prefinal y de la información del morfema de flexión si este existe.

Esto significa que el algoritmo de lematización y de decisión acerca de la categoria morfosintáctica de las bases es un proceso sencillo de pattern- matching sobre los arrays de bits que corresponden a la información respectiva de los morfemas en juego. Este proceso es muy sencillo ya que todos los morfemas han sido categorizados de forma unificada, a partir de criterios comunes.

Para ejemplificar esto tomaremos la palabra "riquísimas". El elativo /isím/ se considera sufijo, y las flexiones de plural son morfemas complejos que contienen información acerca del género y número.

| MORFEMA | CATEGORIZACION | | | | TIPO | | TRANSFORMACION |
|---|---|---|---|---|---|---|---|
| a: /ric/ | :LEX | | | | ADJ | o a | qu |
| b: /isím/ | :AFF | SUF | | | ADJ | o a | |
| c: /as/ | :AFF | END | NOM | VRB | ADJ | a | ar er ir |
| | | | | | categoria | Flexión del lema | |

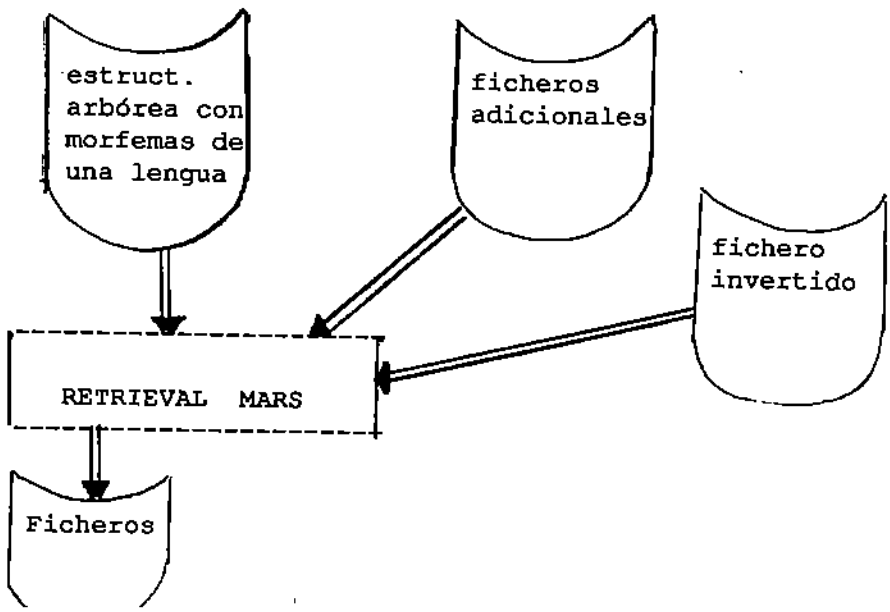
Categoría morfosintáctica= categ. de (b) ∩ categ. de (c) --> ADJ
Lema = CONCAT (morfema (a) con transformación,
morfema (b),
tipo del morfema (b) tipo del morfema (c) )

La sustitución en esta función da : CONCAT (riqu, isím, a)

**9 Resumen**

El siguiente esquema recoge de forma gráfica la funcionalidad del sistema MARS como herramienta de ayuda a la recuperación de la información a partir de un analizador morfológico y de listas invertidas que relacionan las familias de palabras entre sí.

El mecanismo usado para la descomposición morfológica es un automáta de estados finito,y la gramática utilizada juega por un lado con, la categorización de los morfemas a partir de: la combinatoria de la categoría originaria del morfema y categoria resultante de la base obtenida; por otro lado, con la información acerca de las restricciones distributivas de los morfemas entre sí.



Referencias

A. Juilland : Frequence dictionary of Spanish Words. The Hague. Mouton 1974

Jens Lüdtke : Praedikative Nominalisierungen mit Suffixen im Franzoesischen, spanischen und Katalanischen. Niemayer Verlag Tuebingen 1978

Ch.Schwarz & G. Thurmair (ed) :" Informationslinguistische Texterschliessung". Editorial Olms. Hildesheim, N.York. 1986