

OCTAVIO SANTANA SUÁREZ
FRANCISCO JAVIER CARRERAS RIUDAVETS
JOSÉ RAFAEL PÉREZ AGUIAR
GUSTAVO RODRÍGUEZ RODRÍGUEZ

Grupo de Estructuras de Datos y Lingüística Computacional
Departamento de Informática y Sistemas
Universidad de Las Palmas de Gran Canaria
España
[osantana,fcarreras,jperez,grodriguez]@dis.ulpgc.es

Una aplicación para el procesamiento de la sufijación en español

1 Introducción

El objetivo principal de este trabajo es automatizar una parte importante de la morfología del español: la **sufijación**. A través de la sufijación, unas palabras dan lugar a la formación de otras, y éstas a su vez a la de otras; al aplicar sucesivamente este proceso de formación se establecen vínculos familiares entre vocablos. Las familias de palabras que se relacionan son de gran utilidad en aplicaciones de procesamiento del lenguaje natural: buscadores automáticos, correctores ortográficos, analizadores de estilo, generadores automáticos de texto, etc. En un estudio sincrónico de la automatización de la morfología con medios informáticos, los aspectos formales o teóricos no tienen por qué coincidir con los estrictamente lingüísticos. Es obvio que, para el hablante, y por lo tanto debe serlo para la informática, *acuario*, *portuario* y *campanario* son lugares igualmente relacionados con *agua*, *puerto* y *campana*. Es necesario, por tanto, situarse en otro nivel al del problema intrínseco que conlleva este tipo de estudios —la morfología—, para solventar barreras lingüísticas que impedirían tratar aspectos de interés para el procesamiento del lenguaje natural más allá de la derivación.

2 Lexicón

Para la realización de este trabajo se ha creado un corpus de palabras a partir del léxico de distintos diccionarios: el *Diccionario de la Lengua Española* (DRAE), el *Diccionario General de la Lengua Española* (VOX), el *Diccionario de Uso del Español* (MARÍA MOLINER), el *Gran Diccionario de la Lengua Española* (LAROUSSE), el *Diccionario de Uso del Español Actual* (CLAVE SM), el *Diccionario de Sinónimos y Antónimos* (ESPASA CALPE), el *Diccionario Ideológico de la Lengua Española* (JULIO CASARES) y el *Diccionario de Voces de Uso Actual* (MANUEL ALVAR EZQUERRA).

Definida la **forma canónica** como todo vocablo con identidad propia susceptible de aplicársele o de habersele aplicado en su formación algún mecanismo de derivación, en el corpus de referencia se consideran como tales las palabras resultantes de la unión de todas las entradas de la fuentes consultadas, siempre que posean un significado institucionalizado, independientemente de que en su formación entrara un sufijo apreciativo: *aguilón* se ha consolidado como ‘brazo de una grúa’ aunque también es aumentativo de *águila*, pero no se considera *animalucho* por no haber consolidado algún significado distinto del aportado por el sufijo.

Se añaden además, unos 1240 nombres propios, no incluidos como entradas en las fuentes consultadas, relacionados con gentilicios y otros adjetivos y sustantivos del tipo ‘seguidor de’ o ‘doctrina’ —*marxista*, *marxismo* de *Marx*— entre otros significados; y unos 9000 adjetivos procedentes de participios verbales que tampoco han sido recogidos como entradas en dichas fuentes.

El universo de formas canónicas analizadas para conseguir el objetivo de la aplicación que se presenta se compone de 148798 formas canónicas —134645 no verbales y 14153 verbales.

3 La sufijación

La sufijación constituye el procedimiento de formación de palabras más importante en español. Entre la palabra original y el sufijo se desarrolla una interacción dinámica en los ámbitos semántico, funcional y formal que da lugar a una nueva palabra vinculada con la original. Un sufijo responde a “una secuencia fónica que se añade a la base de un vocablo, en posición posterior a él y anterior a las desinencias —cuando las hay—; carece de existencia propia fuera del sistema de palabras; está incapacitado para unirse a otro morfema para formar un derivado; es conmutable por otro morfema analizable como sufijo y cuya base es igualmente conmutable por otra”¹. Con esta definición se puede detallar el conjunto de sufijos que se han considerado en el tratamiento automatizado —existe sufijo si aparece en tres o más vocablos distintos. Así pues, aparecen terminaciones como *-arra*, *-aste*, *-ello*, *-ingo*, *-uz* donde es discutible la condición de sufijo en el sentido gramatical, pero que responden a la definición de sufijo descrita.

Una palabra original puede ser cualquier forma canónica susceptible de añadirsele un sufijo para obtener otra palabra —palabra formada. También se tratan en este trabajo las regresiones deverbales con adición de las terminaciones *-a*, *-e*, *-o* y el sufijo cero o vacío, así como los plurales que se consolidan semánticamente y aquellos sufijos apreciativos que han creado una nueva entidad nocional con independencia de si categorizan o

¹ *Procedimientos de formación de palabras en español*, RAMÓN ALMELA PÉREZ, 1999.

cambian la clase semántico-gramatical de la original. Ya que forman parte del estudio de la composición, se dejan para otro estudio los llamados elementos sufijales —sufijos con una fuerte carga semántica sobre la palabra original o que poseen existencia propia.

Si bien es cierto que la mayoría de los procesos de formación coinciden con una derivación formal —sobre todo las regulares—, ocurre que la coincidencia gráfica con un sufijo concreto puede dar lugar a interpretar como sufijación un proceso ajeno a éste desde el punto de vista lingüístico; por ejemplo, se considera *psiquiatra* y *psiquiátrico* formados a partir de *psiquiatría*, mediante las terminaciones *-a* e *-íco*, a sabiendas de la existencia de *-iatría* e *-iatra*, y es que se pretende tratar la sufijación entre vocablos en español y no con raíces como es el caso de *psiqui-*.

Se consideran los vínculos estudiados entre formas canónicas a partir de cada una de las alteraciones sufijales que produce una terminación al añadirse a una palabra para formar otra, mediante las correspondientes reglas generales de unión o la suya propia en su caso. Una palabra formada puede a su vez ser utilizada para dar lugar a otra si se le añade un nuevo sufijo:

quej-a → *quej-ica* → *quejic-oso*.

Uno de los vínculos que se establece entre la palabra original y la formada, como consecuencia de la sufijación, es la **transcategorización**, —más productiva entre las categorías más frecuentes—; se observa una fuerte tendencia a modificar la categoría gramatical, salvo entre los sustantivos. La mayoría de las derivaciones deverbales son adjetivos y, en menor medida, sustantivos; igual ocurre con las desustantivales, aunque existe un gran número de verbos; en cambio, las deadjetivales presentan una peculiar inclinación a generar adverbios, figura 1.

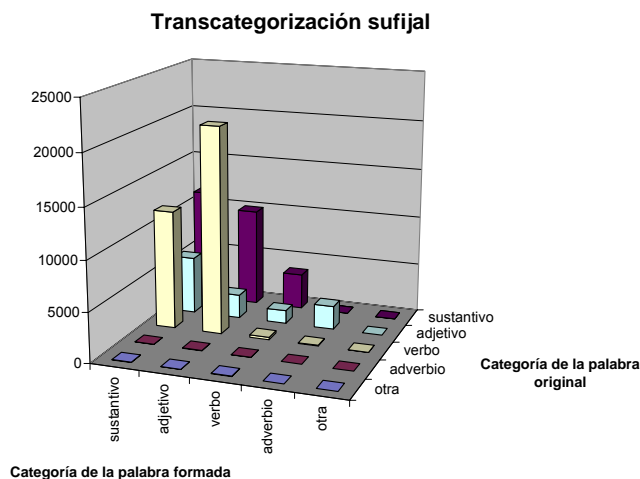


Figura 1

Se presenta la lista en orden alfabético de los 176 sufijos, con sus variantes, contemplados para el tratamiento automático de la sufijación:

Sufijo	Variantes del sufijo	Sufijo	Variantes del sufijo	Sufijo	Variantes del sufijo
-a	-a	-el	-el	-is	-is
-áceo	-áceo,-ácea	-ello	-ello,-ella	-ísimo	-ísimo,-ísima
-acho	-acho,-acha	-elo	-elo,-ela	-ismo	-ismo
-aco	-aco,-aca	-én	-én	-ista	-ista
-ada	-ada,-ade,-ades	-enco	-enco,-enca,-engo,-enga	-ístico	-ístico,-ística
-ago	-ago,-aga	-eno	-eno,-ena	-ita	-ita
-aino	-aino,-aina	-ense	-ense,-iense	-íte	-íte
-aje	-aje	-enta	-enta	-ítimo	-ítimo,-ítima
-ajo	-ajo,-aja	-ento	-ento,-enta,-iento,-ienta,-ulento,-ulenta	-ito	-ito,-ita
-al	-al	-eño	-eño,-eña	-'ito	-'ito,-'ita
-ales	-ales	-eo	-eo,-ea	-itud	-itud
-allo	-allo,-alla,-alle	-'eo	-'eo,-'ea	-ivo	-ivo,-iva,-ativo,-ativa.itivo,-itiva
-amen	-amen	-er	-er,-ier	-iz	-iz,-iza
-án	-án	-'er	-'er	-izar	-izar
-anco	-anco,-anca	-ería	-ería	-izo	-izo,-iza
-áneo	-áneo,-ánea	-erio	-erio	-ma	-ma,-ema
-ango	-ango,-anga	-erío	-erío	-mbre	-ambre,-imbre,-umbre
-ano	-ano,-ana,-iano,-iana	-erizo	-erizo,-eriza	-mente	-mente
-anza	-anza,-enza	-erno	-erno,-ema	-mento	-amento,-imento,-amiento,-imientto
-año	-año,-aña,-uño	-ero	-ero,-era	-ncho	-ancho,-ancha,-oncho,-oncha,-encho,-encha

Sufijo	Variantes del sufijo	Sufijo	Variantes del sufijo	Sufijo	Variantes del sufijo
-ar	-ar	-érrimo	-érrimo	-ncia	-ancia,-encia
-ardo	-ardo,-arda	-és	-és,-esa	-ndero	-andero,-andera,-endero,-endera
-ario	-ario,-aria	-esa	-esa	-ndo	-ando,-anda,-endo,-enda,-ondo,-onda,-iondo,-ionda
-'aro	-aro,-ara	-ésimo	-ésimo,-ésima	-nte	-ante,-ente,-iente
-arra	-arra	-estre	-estre	-o	-o
-aste	-aste	-ete	-ete,-eto,-eta	-ojo	-ojo
-astro	-astro,-astra,-astre	-euta	-euta	-ol	-ol
-atario	-atario,-ataria	-ez	-ez,-eza	-olo	-olo,-ol,-ola
-ate	-ate	-ezno	-ezno,-ezna	-ón	-ón,-ona
-átil	-átil	-ezo	-ezo	-ongo	-ongo,-onga
-ato	-ato,-ata	-grama	-grama	-or	-or,-ora
-avo	-avo,-ava	-í	-í	-'ora	-'ora
-az	-az	-íaco	-íaco,-íaca,-iaco,-iaca	-orio	-orio,-oria
-azgo	-azgo,-azga,-adgo,-adga	-icio	-icio,-icia,-icie	-oso	-oso,-osa
-azo	-azo,-aza	-ico	-ico,-ica	-ote	-ote,-ota,-oto
-azón	-azón	-'ico	-'ico,-'ica	-rragia	-rragia
-bilidad	-bilidad	-'ide	-'ide	-rro	-arro,-arra,-orro,-orra,-arrio,-arria,-orrio,-orria
-ble	-able,-ible	-'ido	-'ido,-'ida	-s	-os,-as,-es
-bundo	-abundo,-ebundo,-ibundo	-ificar	-ificar	-sco	-asco,-esco,-isco,-usco
-ción	-ción,-ación,-ición	-'igo	-'igo,-'iga	-tano	-itano,-ítana,-etano,-etana,-tano,-tana
-'culo	-culo,-cula,-'iculo,-'icula,-'áculo,-'ácula	-iguar	-iguar	-ticio	-aticio,-iticio
-dad	-dad,-tad,-edad,-idad	-ijo	-ijo,-ija	-'tico	-tico,-ático,-ético,-fítico,-ótico
-dero	-adero,-edero,-idero	-il	-il	-torio	-torio,-toria
-dizo	-adizo,-edizo,-idizo	-'il	-'il	-triz	-triz
-do	-ado,-ada,-ido,-ida	-illo	-illo,-illa	-ucho	-ucho,-ucha
-dor	-ador,-edor,-idor	-ilo	-ilo	-uco	-uco,-uca
-dumbre	-edumbre,- idumbre	-imonio	-imonio,-imonia	-udo	-udo,-uda
-dura	-adura,-edura,-idura	-ín	-ín,-ina	-uelo	-uelo,-uela
-duría	-aduría,-eduría,-iduría	-ina	-ina	-ujo	-ujo,-uja
-e	-e	-inche	-inche	-ullo	-ullo,-ulla
-é	-é	-'ine	-'ine	-'ulo	-'ulo,-'ula
-ear	-ear	-íneo	-íneo,-ínea	-uno	-uno,-una
-ecer	-ecer	-ing	-ing	-ura	-ura
-echo	-echo,-echa	-ingo	-ingo.inga	-urno	-urno
-eco	-eco,-eca	-ino	-ino,-ina	-uro	-uro
-edo	-edo,-eda	-iño	-iño,-iña	-uto	-uto,-uta
-ego	-ego,-ega,-iego,-iega	-'io	-'io,-'ia	-uz	-uz
-ejo	-ejo,-eja	-ir	-ir	-uzo	-uzo,-uza

El sufijo más utilizado en la formación de palabras es *-do*, *-da* cuya frecuencia es de 16320. En la figura 2, se muestra la frecuencia de aparición en el corpus del resto de las terminaciones más utilizadas para establecer los vínculos sufijales entre la palabra formada y la palabra original.

Frecuencia de vínculos sufijales

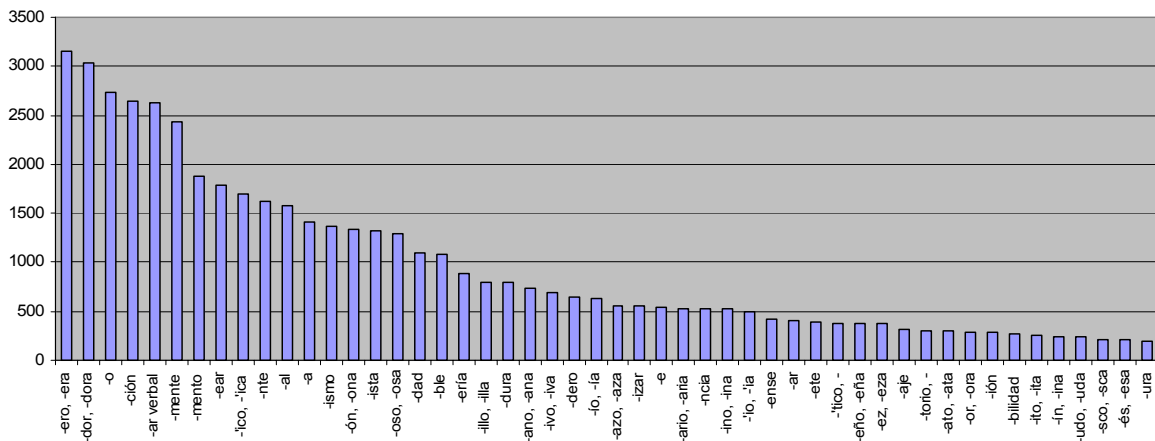


Figura 2

4 La aplicación

Como resultado del trabajo de investigación realizado, se ha desarrollado una aplicación informática capaz de interpretar y manejar con versatilidad los aspectos más relevantes derivados de la sufijación en español. La aplicación representa más una forma de mostrar la potencialidad de un Sistema Computacional de Gestión Morfológica Sufijal del Español que una herramienta finalista. Este Sistema se añade a otra herramienta desarrollada por GEDLC² para dar lugar a un prototipo de uso personal, sin menoscabo de su integración en otras herramientas útiles para el procesamiento del lenguaje natural como corrección ortográfica, búsqueda avanzada de información, analizadores de texto, desambiguadores, estación lexicográfica, analizadores sintácticos, extracción de información, generación automática de texto, corrección sintáctica y extracción de resúmenes, entre otras.

La aplicación constituye una herramienta de interfaz gráfica, amigable, realizada en lenguaje de programación C++, preparada para ejecutarse en ordenadores personales con sistema operativo Windows 95 o superior y exportable a otros sistemas operativos como Linux y Macintosh. La ocupación de memoria física que demanda es de 1,7 Mbytes y la ocupación en disco de los datos necesarios para su funcionamiento es de 41,7 Mbytes.

La base de conocimiento referente a la sufijación se compone de: 1) el vocablo original con el que se forma una palabra, 2) la transcategorización que se produce, 3) el sufijo utilizado en el proceso, 4) la regularidad lexicográfica y 5) la familia genealógica a la que pertenece. Esta información se preprocesa con el fin de obtener un formato adecuado para su uso automatizado mediante un dispositivo informático; se generan dos ficheros binarios que disponen los datos en memoria secundaria: **índice de palabras** y **catálogo de relaciones**, figura 3.

El **índice de palabras** se utiliza para acceder directamente a la familia a la que pertenece cualquier palabra mediante una función de dispersión; almacena: 1) la palabra con su categoría gramatical, 2) el detalle sobre las colisiones, 3) la posición de comienzo de su familia en el catálogo de relaciones, 4) el número de elementos que componen su familia, 5) información de si la palabra es o no la palabra original de la familia y 6) una clave numérica que identifica la familia a la que pertenece; los apartados 5) y 6) ahorran accesos a disco en las operaciones de búsqueda y recorrido, con lo que aumenta la velocidad de respuesta del sistema. Existe un registro con estas características por cada forma canónica perteneciente a una familia.

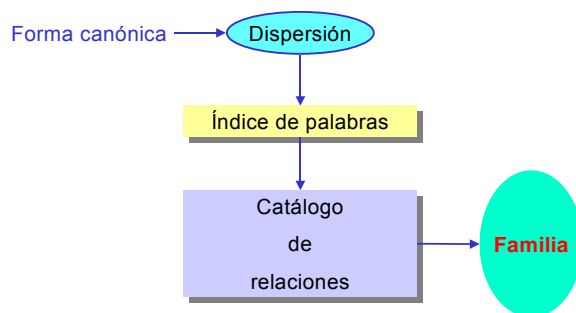


Figura 3

El **catálogo de relaciones** soporta: 1) la palabra con su categoría gramatical, 2) el sufijo con el que se establece la relación y 3) la regularidad lexicográfica.

Del tratamiento de cualquier forma canónica por la función de dispersión, se logra la dirección del registro del índice de palabras que contiene la información necesaria para recuperar su familia del catálogo de relaciones. La familia consta de todas las palabras, con toda su información almacenada en el catálogo de relaciones, que

² FLAPE: Flexionador y Lematizador Automático de Palabras del Español por el Grupo de Investigación de Estructuras de Datos y Lingüística Computacional de la Universidad de Las Palmas de Gran Canaria <http://gedlc.ulpgc.es>

obtiene el sustantivo *profesionista*, los vocablos horizontales de nivel dos son el sustantivo *profesorado* y el adjetivo *profesoral*.

4.1.2.4 Descendencia.

Se entiende por descendencia todas las palabras que han sufrido alteraciones sufijales a partir de una palabra original dada. La descendencia de nivel dos incluye los vocablos que poseen una relación previa con una misma palabra original: recupera los descendientes de cada uno de los descendientes de la voz original. En la familia de *profeso*, los descendientes del adjetivo *profesional* son los sustantivos *profesionalidad* y *profesionalismo*, el verbo *profesionalizar* y el adverbio *profesionalmente*. Los descendientes de nivel dos del adjetivo *profesional* son el sustantivo *profesionalización* y el adjetivo *profesionalizado*.

4.2 Filtros

Las respuestas derivadas de los distintos tipos de navegación a partir de un vocablo concreto pueden, en ocasiones, aportar tal volumen de información que dificulte encontrar las palabras que se buscan y los vínculos que se desean observar. Estos filtros permiten la discriminación selectiva de la respuesta de la navegación. Todos los resultados como consecuencia de los distintos tipos de navegación son susceptibles de ser sometidos a filtros de distinta índole —funcional, regularidad y por sufijos.

4.2.1 Funcional.

Se entiende por filtro funcional la selección por categoría gramatical de las palabras que componen el resultado de una determinada navegación. Si se quieren explorar los sustantivos descendientes de un vocablo, se aplica la navegación descendente y se seleccionan los sustantivos exclusivamente; y si se desean los descendientes no adjetivales, se desciende por todas las posibilidades menos por los adjetivos. Así, en la familia de *profeso*, el único descendiente verbal del adjetivo *profesional* es *profesionalizar*, la respuesta sin filtro se ha reducido en tres vocablos.

4.2.2 Afijo.

Se puede aplicar un filtro basado en el tipo de afijo utilizado en la formación de la palabra. Se ejecuta sobre la selección de las palabras que componen el resultado de una determinada navegación. Esta opción, se enriquece notablemente si se complementa con: 1) la información referente a los atributos de significado, 2) las categorías gramaticales que forman y 3) las categorías gramaticales a las que se aplican. Esta ampliación permite seleccionar una determinada respuesta, por categoría gramatical y por afijos que producen una cierta semántica, lo cual hace muy provechosa su aplicación. Es aplicable la discriminación por uno o por varios afijos simultáneamente de los estudiados en este trabajo, por lo que las combinaciones y posibilidades son notables. Por ejemplo, si se quieren explorar las palabras formadas a partir de *profesar* con sufijos exclusivamente sustantivadores muy frecuentes: *-ción*, *-dura*, *-aje*, *-miento*, *-ancia*, *-azo*, *-bilidad*, *-dad*, *-ería*, *-ez*, *-illo*, *-ión*, *-ismo* y *-ura*, el resultado sería el sustantivo *profesión*, al filtrar *profesor*, dado que la terminación *-or* es sustantivadora y adjetivadora.

4.2.3 Regularidad.

Se puede establecer un filtro en función de la regularidad en el proceso de formación de la palabra. Se aplica sobre la selección de las palabras que componen el resultado de una determinada navegación. Si se quieren explorar las formaciones irregulares horizontales de un vocablo, se aplica la navegación horizontal y se seleccionan exclusivamente las palabras que se hayan establecido como irregulares. En la familia de *vejiga*, la respuesta horizontal irregular del adjetivo *vejigoso* sería el sustantivo *vesícula* y el adjetivo *vesical* —la respuesta sin filtro se ha reducido en tres vocablos, *vejigatorio*, *vejigazo* y *vejiguilla*, figura 6.

Familia de palabras vinculadas

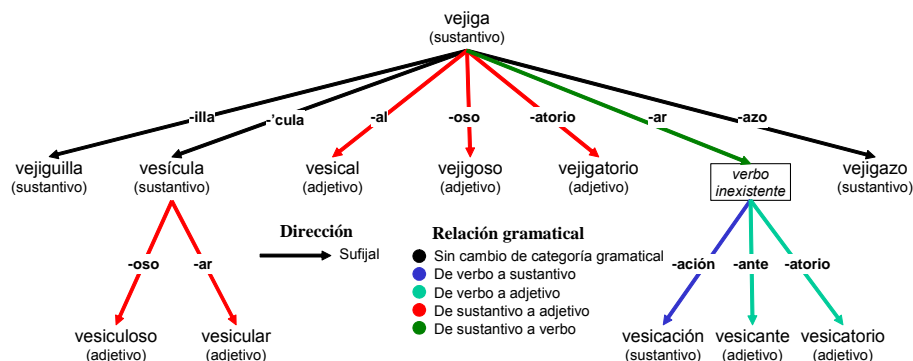


Figura 6

Bibliografía

- Ait-Mokhtar, S., Rodrigo Mateos, J. L. 1995. "Segmentación y análisis morfológico de textos en español utilizando el sistema SMORPH". *Boletín de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)*, Nº 17: 29/41, Septiembre.
- Alcoba Rueda, S. 1992. "Tema verbal y formación de palabras en español", *Actas do XIX Congreso Internacional de Lingüística e filoloxía románicas*, Volumen II, Universidad de Santiago de Compostela, La Coruña.
- Almela Pérez, R. 1999. *Procedimientos de formación de palabras en español*, Barcelona, Ariel.
- Alvar Ezquerro, M. 1993. *La formación de las palabras en español*, Cuadernos de lengua española, Arco/Libros, Madrid.
- Bajo Pérez, E. 1997. *La derivación nominal en español*, Madrid, Arco/Libros.
- Faitelson-Weiser, S. 1993. "Sufijación y derivación sufijal: sentido y forma", *La formación de palabras*, Varela (ed.), Taurus, Madrid.
- González Collar, A. L., Goñi Menoyo, J. M., González Cristóbal, J. C. 1995. "Un Analizador Morfológico para el castellano basado en Chart". *Actas de la VI Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA'95)*, Alicante, Noviembre.
- Lang, Mervyn F. 1992. *Formación de palabras en español. Morfología derivativa productiva en léxico moderno*, Madrid, Cátedra.
- Malkiel, Y. 1993. "El análisis genético de la formación de palabras", *La formación de palabras*, Soledad Varela (ed.), Taurus, Madrid.
- Pilleux, M. S. 1980. *Análisis morfológico, funcional y semántico de los sufijos en español*, Universidad Austral de Chile, Valdivia.
- Santana, O., Pérez, J., Hernández, Z., Carreras, F., Rodríguez, G. 1997. "FLAVER: Flexionador y lematizador automático de formas verbales", *Lingüística Española Actual*, 19-2, Ed. Arco/Libros, S.L., , págs. 229/282.
- Santana, O., Pérez, J., Carreras, F., Duque, J., Hernández, Z., Rodríguez, G. 1999. "FLANOM: Flexionador y lematizador automático de formas nominales", *Lingüística Española Actual*. 21-1, Ed. Arco/Libros, S.L., .
- Soledad Varela (ed.): 1993. *La formación de palabras*, Taurus, Madrid.