Software Application for Parasynthesis in Spanish Automatic Processing

Octavio Santana, Francisco J. Carreras, José R. Pérez, Juan Carlos Rodríguez
Grupo de Estructuras de Datos y Lingüística Computacional.
Departamento de Informática y Sistemas, Campus Universitario de Tafira,
Universidad de Las Palmas de Gran Canaria,
35017 Las Palmas de Gran Canaria, Spain

({OSantana, FCarreras, JPerez, JCRodriguez}@dis.ulpgc.es)
http://www.gedlc.ulpgc.es

Abstract - This work is about a computer application for parasynthesis in Spanish automatic processing, which works with 3800 parasynthetical morpholexical relationships deduced from a corpus of 148798 canonical forms. The developed computational tool is capable of solving and answering to any morpholexical aspect of a Spanish word because of it includes the suffixation and prefixation processing. The tool encompasses everything related with derivation, prefixation, parasynthesis and other nearby aspects. It allows the recognition, the generation and the manipulation of morpholexical relationships of any word and of its related words, includes the recovery of all its lexicogenetic information until arriving at a primitive, the management and the control of the affixes in the treatment of its relationships, as well as the irregularities and exceptions of lexicon, which are many in a highly inflected language.

Keywords: Spanish morphology, parasynthesis, derivation, computational linguistic, natural language processing.

1.0 Introduction

The primary target of this work is to automate the parasynthesis by its importance in the Spanish morphology and with the purpose of completing the program ―developed by the Group of Data Structures and Computational Linguistic (GEDLC) http://www.gedlc.ulpgc.es― that processes two mechanisms of formation of words in Spanish: the suffixation and the prefixation [13], [14], [16], [17], [18] y [19]. Through these mechanisms of formation, words give rise to others, and these as well to the one of others; when applying successively these processes of formation take familiar bonds between words. The families of words that are related are very useful in applications of the natural language processing; for example, for the journalist which it writes up his chronicle hastily and it needs a corrector style, or for who wants to change to the forms of treatment and the consequent agreements in a text, or for the publicist that looks for associations of words of diverse nature, or for the professor of Spanish that it wants to teach to foreign students the most frequent structures in a type of texts, or the most frequent voices in a certain scope, or to look for words of a same semantic field, etc.

From the computational linguistic point of view, we began from several works made by the GEDLC on the morpholexical relationships of the Spanish that have given rise to a program able to manage the most frequent mechanisms of formation of words in Spanish: the suffixation and the prefixation. The study of the parasynthesis qualitatively complements the results obtained until now and improves specifically the developed software application. This program represents the suffixal, prefixal and parasynthetical morpholexical relationships and it allows to solve and to respond to the aspects of these relations between the words in the recognition of same like functional and morphologic elements ―the origin word the derived word, grammatical transcategorization, affix implied and its meaning, type of relationship and the set of family of related words. It allows finding words related to different criteria from functionality, morphology and proximity, as well as selection of the result based on the

type of relationship, of the grammatical category and the chosen affixes. The developed program is available in Internet for free use, in addition, exists a local version that shows, from the point of view of an expert, all its linguistic potentiality and, most important, it is a module can be incorporated, like basic tool, to other software applications of high level ―intelligent morphologic search engine, checker of style, etc.

The inherent problem in the works of investigation on the recognition of the lexical morphology, as an essential and autonomous component of the grammar, is to explain the derivative properties of the lexicon across the relationships that are established between constituent morphemes. The amount of existing morphemes ―bases and affixes― and the excessive number of allowed combinations make difficult that study still more. However, it is true that the words, respect to others, have common patterns in its morphologic behaviour ―field of which takes care this work. Another controversial question in the research on morphology consists of the synchronous boundary respect to the diachronic one; words are totally tied to their history, in both the morphologic and semantic fields. The history generally determines its current lexicography and semantics, reason that make its cataloguing laborious without considering its etymological references. Some words, already old-fashioned or belonging to mother tongues like Latin or Greek, have relevant information of the historical morphologic process, which completes the connections between the different formation types of the current lexicon in a generational sequence of words. In this work, a wide corpus of Spanish words have been selected, without spreading to other languages to minimize the absence of connectors between morpholexically related words by the formative processes ―excluding the suffixation and the prefixation whose articles already have been published by Santana et al. [13], [14], [16], [17], [18] y [19]― and to define the synchronous study which is claimed, without obviating the necessary abuses of archaic etymological processes. In addition, it is necessary to consider that a synchronous study of the automation of the morphology with means of computer science, the formal or theoretical aspects many not coincide with those strictly linguistics; barriers are saved therefore that they would prevent to deal with aspects interest for the processing of the natural language beyond these processes of formation.

2.0 Lexicon

The corpus handled in this work has been created from: the *Diccionario de la Lengua Española* (DRAE), the *Diccionario General de la Lengua Española* (VOX), the *Diccionario de Uso del Español* (María Moliner), the *Gran Diccionario de la Lengua Española* (LAROUSSE), the *Diccionario de Uso del Español Actual* (Clave SM), the *Diccionario de Sinónimos y Antónimos* (Espasa Calpe), the *Diccionario Ideológico de la Lengua Española* (Julio Casares) and the *Diccionario de Voces de Uso Actual* (Manuel Alvar Ezquerra).

A canonical form is defined as any word with its own identity susceptible of enduring derivational processes to form other words. Such a word could be formed from another by similar processes. In the reference corpus a canonical form is any entry word of consulted sources having own meaning ―those entries that are appreciative forms of others and do not add any substantial meaning variation are discarded. The universe of words analyzed in this work is composed of 148798 canonical forms.

3.0 The Parasynthesis in Spanish

Some words in Spanish have been formed suffering the processes of suffixation and prefixation simultaneously ―parasynthesis. The theoretical discussion about being a derivative or compositive process or both still remains. Following the criterion of other authors, like Mervin F. Lang [11], David Dolader [21] and some others, these alterations are studied from a synchronous point of view ―the discussion synchronic-diachronic is irrelevant with respect to those applications aimed to improve the natural language processing. This mechanism of formation deserves a separate study due to its singular characteristics. This type of formation does not have to be considered like two types of

consecutive alterations in time, since each one does not exist separately. For example, *abrochar* is directly related to *broche* by means of parasynthetical alteration; its semantic corroborates such consideration: 'close, join or fit with fasteners' —neither the canonical form **abroche* nor **brochar* exist.

To give an example, the parasynthetically related verbs with *loco* are *alocar*, *enloquecer* and *aloquecer*, which notably diminishes the possible response of a mask search "*loc?r": *aclocar*, *allocar*, *alocar*, *bilocar*, *blocar*, *clocar*, *colocar*, *descolocar*, *desflocar*, *dislocar*, *enclocar enllocar*, *locar*, *recolocar*. Although they are not in this group the response of the verbs which have suffered spelling changes as consequence of the phonetic adjustments or of any other kind: *enloquecer* and *aloquecer*.

The extrapolation of a parasynthetical process in a mother tongue common to two primitive ones is applied as a parasynthesis to cause a morpholexical relationship between them. Just like in the suffixal or prefixal alteration, in the current state of the Spanish language the relationship existing between those primitive words presents a strong parallelism in their morphological, semantic, and grammatical aspects with the parasynthetical process between Spanish forms. This way, for the substantive *cadena* and the verb *concadenar* —both of them primitive Spanish words and directly derived from Latin— *concadenar* is likely to be considered as a parasynthetical verbalization of *cadena*.

Spanish words formed from a derivative base exist where its morphological, semantic and grammatical behaviour is identical to those of the secondary derived form. In this case, the considered morpholexical relationship follows the one which would exist if the corresponding primary derived had been formed. For example, *exinanición* and *exinanido* derive from *inane* but their morpholexical relationship is analogous to the one of *embarcación* and *embarcado*, which derives from *barco* through the verb *embarcar*. However, when there is not a possibility of losing the suffixal morpholexical relationship with a word which is under these circumstances, the parasynthetical morphology does apply —*defonologización* is considered as parasynthetical of *fonología*.

There are words of which is difficult to know their real morphological situation —parasynthetical or prefixal— sometimes due to the lack of historical-morphological information and some others due to the character of the word. In the last cases, the meaning has been used for more accuracy. For instance, *embetunar* establishes a prefixal relationship with *betunar*, or parasynthetical with *betún*; since the same semantic relationship is established between *betunar* and *betún* as between *embetunar* and *betún* —they are semantically the same—, the parasynthetical one is preferred because the sources do not clarify the origin of *embetunar*, and *betunar* is an archaic word —morpholexical relationships with grammatical category change are mainly searched.

There are Spanish words which have a close semantic and functional relationship which cannot be directly established by means of a process morphologically classified in the parasynthesis field. The relationships characterized by having a point in common in the etymologic history and by incorporating a prefix and an ending with the right functional and semantic contribution are included. As a model, *coetáneo* is considered to be parasynthetically related to *edad*, though *coetáneo* is primitive and has a different root from *edad*.

One of the bonds that take between the original word and the formed one, as a result of the parasynthesis, is the transcategorization, which is more productive between the most frequent categories. A strong tendency is observed to transform the grammatical category into verb. Most of the verbal formation came from substantives and, to a lesser extent, from adjectives.

Table 1. The frequency of appearance in corpus of the pairs prefix-suffix used to establish the parasynthetical relationships between the formed word and the original word is shown. In order to distinguish the verbal suffix *-ar* of the substantive suffix *-ar*, the nomenclature *-arv* and *-ar* has been used respectively. The apostrophe in front of a suffix indicates an atonic suffix.

Pair	Freq.	Pair	F.	Pair	F.	Pair	F.
a-...-arv	1064	hipo-...-'io	6	trans-...-ción	3	hipo-...-ío	2
en-...-arv	990	per-...-arv	6	a-...-iguar	3	di-...-ecer	2
des-...-arv	406	e-...-ción	6	des-...-dor	3	res-...-arv	2
a-...-do	261	des-...-ecer	5	des-...-mente	3	pro-...-e	2
en-...-ecer	104	es-...-ecer	5	anti-...-nte	3	poli-...-'io	2

Pair	Freq.	Pair	F.	Pair	F.	Pair	F.
en-...-do	67	re-...-ecer	5	sin-...-ío	3	sub-...-áneo	2
es-...-arv	63	hiper-...-'io	4	ad-...-arv	3	poli-...-o	2
de-...-arv	52	re-...-do	4	re-...-ón	3	por-...-arv	2
a-...-ear	45	dis-...-arv	4	mono-...-'ico	3	sub-...-'io	2
re-...-arv	42	di-...-arv	4	re-...-ir	3	pre-...-ano	2
des-...-do	37	con-...-áneo	4	en-...-ear	3	poli-...-'ico	2
in-...-arv	36	e-...-ecer	4	ex-...-ecer	2	archi-...-o	2
trans-...-arv	31	ob-...-arv	4	mono-...-al	2	re-...-o	2
a-...-ecer	31	a-...-o	4	dis-...-'io	2	mono-...-ío	2
con-...-arv	28	a-...-ir	4	eu-...-'io	2	re-...-izar	2
in-...-ble	26	a-...-mento	4	e-...-ble	2	re-...-'io	2
so-...-arv	17	in-...-ción	4	di-...-'io	2	anti-...-mento	2
ex-...-arv	15	con-...-'io	3	e-...-'io	2	re-...-ear	2
a-...-izar	13	extra-...-arv	3	con-...-ción	2	so-...-do	2
in-...-e	12	bi-...-o	3	za-...-arv	2	a-...-ate	2
a-...-'io	11	de-...-ción	3	con-...-ecer	2	per-...-ecer	2
e-...-arv	11	in-...-do	3	en-...-ción	2	uni-...-'ico	2
en-...-izar	10	pro-...-arv	3	en-...-mento	2	uni-...-ar	2
entre-...-arv	8	iso-...-'ico	3	des-en-...-arv	2	multi-...-e	2
in-...-o	8	ex-...-e	3	in-...-'ito	2	pan-...-'io	2
en-...-ir	7	con-...-al	3	entre-...-'io	2	trans-...-al	2
des-...-ción	7	inter-...-'io	3	hiper-...-ismo	2	mono-...-o	2
des-...-izar	6	a-...-mente	3	en-...-mente	2	es-...-ir	2
bi-...-e	6	di-...-o	3	hipo-...-ismo	2	158 pairs more	1

4.0 Computer Application

As a result of the research done, a computer application capable of interpreting and handling with versatility the most relevant aspects of the mechanisms of formation of words in Spanish has been created. This computer application adds another tool already developed —FLAPE: Automatic Inflectioner and Tagger of the Spanish Words [15] y [20]— to cause a version of personal use, without detriment of its integration in other useful tools for the natural language processing as spell-checking, advanced search engine, parser, word sense disambiguator, lexicographical station, extraction of information, automatic text generation, among others.

The application is a tool of graphic user interface, friendly, made in the programming language C++, prepared to be carried out in personal computers with the operative system Windows 95 or higher and exportable to the other operative systems as Linux and Macintosh. The memory occupation needed is 3 Mbytes and the occupation in the disk of the data needed for the proper functioning is 54.2 Mbytes.

4.1 Implementation

The knowledge base with the information referred to the word formation mechanisms consists of:

1) The original word with which a word is formed.
2) The transcategorization produced.
3) The affixes used in the process.
4) The lexicographic regularity.
5) The genealogical family it belongs to.

This information is processed in order to obtain a suitable format for its automated use by means of a computer device. Two binary files are generated. These two files have data in secondary memory: word index and relationship catalogue.

The word index is used to accede directly, by means of a hash function, to the family to whom any word belongs; and it stores:

1) The word with its grammatical category.
2) The detail about the collisions.

3) The beginning position of its family in the relationship catalogue.
4) The number of elements which make its family.
5) Information about whether the word is the family original word.
6) A numerical key which identifies its family.

Information 5 and 6 saves accesses to disk in the operations of search and route, this way the system answer speed increases. There is a register with those characteristics for each canonical form belonging to a family.

The relationship catalogue stores:

1) The word with its grammatical category.
2) The type of relationship with its original form (suffixal, prefixal, parasynthetical).
3) The pair prefix-suffix with which the relationship is established.
4) The lexicographic regularity.

From the processing of any canonical form by the hash function, we achieve the direction of the word index record which contains the information necessary to retrieve its family from the relationship catalogue. The family consists of all the words, with all its information stored in the relationship catalogue, that are tie direct or indirectly —through other words— with the introduced canonical form. If a word belongs to different families, one register per family is obtained.

4.2 Interface

The computer application takes advantage of some of the services provided by FLAPE with the purpose of facilitating its use and avoiding advanced knowledge of the language. The interface of *relationships*, figure 1, facilitates the exploration of the parasynthetical morpholexical relationships which has a canonical form from any Spanish word. The *entry* allows the user to enter any Spanish word which is recognized morphologically in an automatic way and at the same time as the user is keying. As result of the recognition three possibilities can be considered:

1) If the entry is not recognized as a Spanish word, the buttons of the interface cannot be used.
2) If the entry is recognized but the canonical forms which belong to it do not have morpholexical relationships, the buttons of the interface can neither be used.
3) If the entry is recognized and at lest one of its canonical forms has morpholexical relationships, it is permitted to work with the interface.

Only the canonical forms which have relationships with other canonical forms, as a result of the suffixation, prefixation or parasynthesis are placed in the control *canonical form* and it can be used at any time. In the window *inflection*, it is shown at any time the inflection of the entry with respect to the selected canonical form which is visible and below it we can see its *grammatical category*.

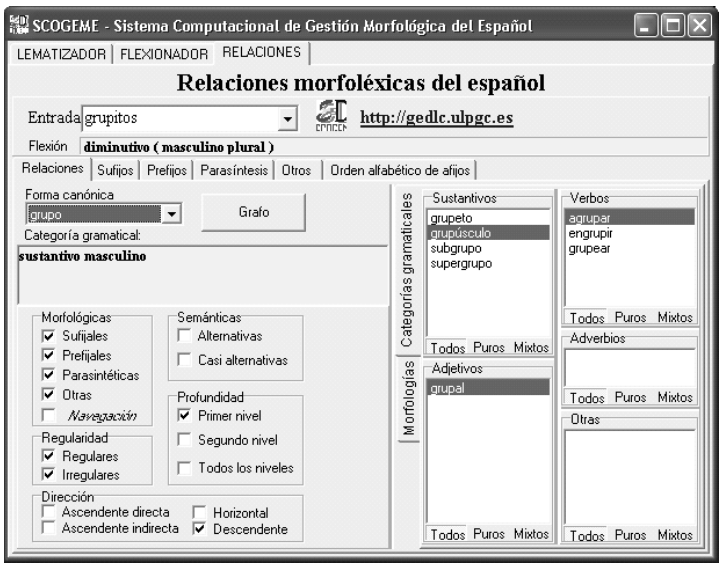

Figure 1

A group of buttons lets the navigation in the family of the selected word and the majority directly filters the result without having to generate new searches. The result of the morpholexically related words are shown in the windows placed on the right of the interface organized by grammatical category or by kind of morphology —*result windows*.

The buttons grouped in *regularity* filter the result words according to the regularity in their formation process. Either the regular or irregular relationships or both at the same time are obtained with this option —the buttons are independent.

The buttons of *direction* and *depth* establish the words morpholexically related to the canonical form. These two groups of buttons are linked since, once established the direction of the search in the graph, it is necessary to specify the words related to each of the directions marked with each level of deepness selected. Any change in the selection of the buttons of *direction* or *depth* generates a new search of words to satisfy the request.

Together with the result classified by grammatical categories, we find three buttons in each window —*all*, *pure* and *mixed*— which let us select the words of that window only by the grammatical function of the window —*pure*— or because they have another grammatical function apart from the one which defines the window —*mixed*—, or to show all the words which have alt least the grammatical function which defines the window —*all*. With the button *all* pressed in the substantive window and in the adjective one, a word with grammatical category 'adjective used as substantive' appears in the two windows at the same time. If in any of them the button *pure* is pressed, that word will disappear from that window since it does not show words with only one grammatical function —substantive or adjective exclusively. And if the button *mixed* is pressed, the words which share another grammatical function from the one expressed in the window will be selected.

The *mixed* substantive related to the word *chino* is: *achinado* —with grammatical function 'adjective also used as substantive—; the *pure* substantives are: *chinaje*, *chinería* and *chinerío*.

Within the interface of *relationships* is the flap *parasynthesis*, which allows configure the computer application for the operation of the parasynthesis. All pairs prefixes-suffixes considered are shown, classified by the grammatical category that produce —verbs, substantives, adjectives and adverbs— and by its frequency of appearance in the data base; within each one of these groups they alphabetically appear to facilitate its location: *More frequent*, those that appear more than 200 times, *Frequent*, those that appear between 20 and 200 times, *Less frequent*, those that appear less than 20 times. The meanings of the selected affix are shown at the bottom of the window.

5.0 Conclusions

A taxonomic, exhaustive and systematic study is made about affixes used in the derivative, prefixal and parasynthetical morphology of the Spanish on corpus sufficiently wide that it ensures all the casuistry of each one of the affixes existing in this language. The way for use the affixes, the transcategorization, the meaning and the lexicographic regularity in the relation provides a overview of the formative behaviour of the Spanish words, since the affixes implied in the main processes of formation appear in this computer application —suffixation, prefixation and parasynthesis. It is important to emphasize that all the irregularities and exceptions of lexicon of the section 2 have been studied, which are many in a highly inflected language —20% of irregularity in the suffixation, 7% of irregularity in the prefixation and 15% of irregularity in the parasynthesis.

The software application integrates the recognition and generation of the Spanish morpholexical relationships, by using the same logical and physical design —synchronous recovery of the words called *origin* of one morpholexical relationship. This process can be applied in iterative way to obtain the *origins* of the *origins* with morpholexical relationships.

The computer application is designed to be of utility to those who works with documents in Spanish: lexicologists, analysts of style, extractors of textual information, translators, etc. An intuitive graphical user interface with dialog windows, buttons and other graphical tools facilitates the human-machine interaction. This supposes a first step towards the multiple possibilities in computer science and specialized programs that they must be developed on this knowledge base.

The obtained results and the FLAPE tool —FLAPE: Automatic Inflectioner and Tagger of the Spanish Words— are integrated to take a Computational System of the Morphologic Management of the

Spanish able to manage the inflected morphology and the derivation, the prefixation and the parasynthesis, and the morpholexical relationships between Spanish words. This program easily turns out an integral able product in tools from aid to the oriented document treatment to solve problems of the natural language processing.

6.0 References

[1] Alarcos Llorach, E. Gramática de la Lengua Española. Espasa-Calpe, Madrid, Spain, 1995.
[2] Almela Pérez, R. Procedimientos de formación de palabras en español. Ariel, Barcelona, 1999.
[3] Alvar Ezquerra, M. La formación de las palabras en español. Cuadernos de lengua española, 5ª ed. Arco/Libros, Madrid, 1993.
[4] Alvar Ezquerra, M. Nuevo diccionario de voces de uso actual. Cuadernos de lengua española. Arco/Libros, Madrid, 2003.
[5] Bajo Pérez, E. La derivación nominal en español. Arco/Libros, Madrid, 1997.
[6] Bosque, I., Demonte, V., Lázaro Carreter, F. Gramática descriptiva de la lengua española. Espasa, Madrid, 1999.
[7] Dee, J. H. A lexicon of latin derivatives in Italian, Spanish, French and English, Vol. I Introduction and Lexicon. Olms-Weidmann, New York, 1997.
[8] Dee, J. H. A lexicon of latin derivatives in Italian, Spanish, French and English, Vol. II Index. Olms-Weidmann, New York, 1997.
[9] Faitelson-Weiser, S. Sufijación y derivación sufijal: sentido y forma. La formación de palabras. Varela (ed.), Taurus, Madrid, 1993.
[10] García-Medall, J. Formaciones prefijales en español: morfología derivativa del verbo. Ph. Degree Thesis. University of Valencia, Valencia, 1991.
[11] Lang, Mervyn F. Formacion de palabras en español. Morfología derivativa productiva en léxico moderno. Cátedra, Madrid, 1992.
[12] Malkiel, Y. El análisis genético de la formación de palabras. La formación de palabras. Soledad Varela (ed.), Taurus, Madrid, 1993.
[13] Santana, O., Carreras, F., Pérez, J. Relaciones morfoléxicas sufijales para el procesamiento del lenguaje natural. Mileto, Madrid, 2004.
[14] Santana, O., Hernández, Z., Rodríguez, G. *Morphoanalysis of Spanish Text: Two Applications for Web Pages*. Lecture Notes in Computer Science, Vol. 2722, 511-514,. Springer-Verlag, Berlin Heidelberg New York, 2003.
[15] Santana, O., Pérez, J., Carreras, F., Duque, J., Hernández, Z., Rodríguez, G. FLANOM: *Flexionador y lematizador automático de formas nominales*. Lingüística Española Actual, Vol. 21-1, 253-297. Arco/Libros, S.L., Madrid, 1999.
[16] Santana, O., Pérez, J., Carreras, F., Rodríguez, G. *Relaciones morfoléxicas prefijales del español*. Procesamiento de Lenguaje Natural, nº 32, 9-36. SEPLN, Madrid, 2004.
[17] Santana, O., Pérez, J., Carreras, F., Rodríguez, G. *Relaciones morfoléxicas sufijales en español*. Procesamiento de Lenguaje Natural, nº 30, 1-73. SEPLN, Madrid, 2003.
[18] Santana, O., Pérez, J., Carreras, F., Rodríguez, G. *Suffixal and Prefixal Morpholexical Relationships of the Spanish*. Lecture Notes in Artificial Inteligente, Vol. 3230, 407-418. Springer-Verlag, Berlin Heidelberg New York, 2004.
[19] Santana, O., Pérez, J., Hernández, Z., Carreras, F., Rodríguez, G. *The Spanish Morphology in Internet*. Lecture Notes in Computer Science, Web Engineering, Vol. 2722, 507-510. Springer-Verlag, Berlin Heidelberg New York, 2003.
[20] Santana, O., Pérez, J., Hernández, Z., Carreras, F., Rodríguez, G. *FLAVER: Flexionador y lematizador automático de formas verbales*. Lingüística Española Actual, Vol. 19-2, 229-282. Arco/Libros, S.L., Madrid, 1997.
[21] Serrano Dolader, D. Las formaciones parasintéticas del español. Arco/Libros, Madrid, 1995.
[22] Soledad Varela (ed.) La formación de palabras. Taurus, Madrid, 1993.
[23] Varela, S., Martín, J. "La prefijación", I. Bosque, V. Demonte (eds.), Gramática descriptiva de la lengua española. Espasa-Calpe, Madrid, 4993-5040, 1999.