Parasynthetic Morpholexical Relationships Of The Spanish: Lexical Search Beyond The Lexicographical Regularity

Octavio Santana Suárez
Departamento de Informática y Sistemas
Universidad de Las Palmas de Gran Canaria
Las Palmas de Gran Canaria, 35017, Spain
osantana@dis.ulpgc.es

Francisco J. Carreras Riudavets
Departamento de Informática y Sistemas
Universidad de Las Palmas de Gran Canaria
Las Palmas de Gran Canaria, 35017, Spain
fcarreras@dis.ulpgc.es

Jose R. Pérez Aguiar
Departamento de Informática y Sistemas
Universidad de Las Palmas de Gran Canaria
Las Palmas de Gran Canaria, 35017, Spain
jperez@dis.ulpgc.es

Juan Carlos Rodríguez del Pino
Departamento de Informática y Sistemas
Universidad de Las Palmas de Gran Canaria
Las Palmas de Gran Canaria, 35017, Spain
jcrodriguez@dis.ulpgc.es

ABSTRACT

This work talks about parasynthesis of the Spanish language. This formative process of Spanish words is useful for the establishment of morpholexical relationships. From a lexicon of over 4 million different words, around 6 million parasynthetic morpholexical relationships are established. All the irregularities and exceptions found in referenced lexicon have been considered, which are many in a highly inflected language. These relationships turn out to be useful because they allow, between other possibilities, doing semantic searches, offering alternative sentences in the correction of style or summarization and finding semantically synonymous sentences. The principal main function of this application is that it allows lexical searches beyond the lexicographical regularity. The developed web tool is capable of solving any morpholexical aspect of a Spanish word. This tool includes the suffixation and prefixation processing and also shows the graph of morpholexical relationships. This tool is only one way to show the potentiality of the system which can be incorporated to other tools of high linguistic level.

KEYWORDS

Morpholexical Relationships, Computational Linguistic, Information Retrieval.

1. INTRODUCTION

The problem in the works of investigation on recognition of the lexical morphology, as an essential and autonomous component of the grammar, is to face the derivative properties of the lexicon across the relationships that are established between the constituent morphemes. The existence of a lot of morphemes ―bases, roots, and affixes― and its excessive number of allowed combinations makes its study difficult. However, the words possess a common pattern in its morphologic behaviour, which is the aim of present

work. Another controversial question, in the research on morphology, consists of the synchronous definition; words are fully tied to their history, in both the morphologic and semantic fields. The history generally determines its current lexicography and semantic. Therefore, it is necessary to take into account the etymological references of the words, which complicates the tagging. Some words, already old-fashioned or belonging to mother tongues like Latin or Greek, have relevant information about its historical morphologic process. This information increases in number the connections between the different formation types of the current lexicon in a generational sequence of words. In this work, a wide lexicon of Spanish words have been considered, without spreading to other languages, to minimize the absence of connectors between morpholexically related words by the formative processes. This wide lexicon makes a synchronous study possible without obviating some necessary information about archaic etymological processes. The resulting system is characterized by simple and flexible algorithms with 100% success in the recognition and generation of morpholexical relationships of the language independently of the complexity of the data they are handling —in another highly inflectional languages like Greek (Socrates, 2005) or French (Polguère, 2000) similar works have been developed, the latter emphasizes the usefulness of these relations for natural language processing applications.

2. LEXICON

The lexicon handled in this work has been created from: the *Diccionario de la Lengua Española* DRAE (Real Academia Española y Espasa Calpe, 1995), the *Diccionario General de la Lengua Española* VOX (Bibliograf, s. a., 1997), the *Diccionario de Uso del Español* (María Moliner, 1996), the *Gran Diccionario de la Lengua Española* (Larousse, 1996), the *Diccionario de Uso del Español Actual* (Clave SM, 1997), the *Gran Diccionario de Sinónimos y Antónimos* (Espasa Calpe, 1991), the *Diccionario Ideológico de la Lengua Española* (Julio Casares, 1990) and the *Diccionario de Voces de Uso Actual* (Manuel Alvar Ezquerra, 1994).

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 lexicon 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 consists of 148,798 canonical forms that provide about 4 million different inflected words, among which 6 million parasynthetic morpholexical relationships have been established, of which 3,800, between canonical forms. A morpholexical relationship between canonical forms is projected to all the words that are generated from each of the canonical forms, through any of the processes of push-up or appreciation of Spanish: gender, number, augmentative, diminutive, pejorative, superlative, conjugation... (Santana et al., 1997 and Santana et al., 1999).

3. AIM

This work aims at obtaining a set of parasynthetic morpholexical relationships among Spanish words useful for automatic applications in natural language processing, such as automatic searchers, spelling correctors, style analyzers, automatic text generators, automated ideological dictionaries, etc. It would be of great usefulness for those who deal with documents in Spanish as lexicographers, analysts of style, text information retrievers, translators, and many others.

To give an example, the parasyntheticly 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 verbs which have suffered spelling changes as consequence of the phonetic adjustments or of any other kind: *enloquecer* and *aloquecer*.

Having morpholexical relationships allows doing lexical searches, which by means of searching mechanisms based on the graphic word are notably improved: exact searches, with wildcards, masks or truncations. On the one hand, introducing words whose relationship with the search pattern is merely graphical is avoided and, on the other hand, the abysses which are producing the strong irregularities of inflected languages, as the Spanish one, are avoided. This way, about any word, it can be figured-out the

lemma from which it comes, the relationships it supports with other lemmas and the words to which these lemmas give place. All this is applied in the context of interest; for example: a paragraph, a text, a web address, an electronic library or a textual corpus.

This searching system allows the whole flexibility that could be needed: from an exact search *casero*, to include changes of genre and/or number *casero, casera, caseros, caseras*, to include the appreciative and/or the superlative forms with or without changes in gender and number *caserote, caserito, caserejo, caserísimo...* to include morpholexically related canonical forms in the first descending suffixing level *caserillo, caseramente*, prefixing or parasynthetic level *acaserar*, ascending casa, in other levels *acaserado, acaseramiento, casariego, caserío, caserón, caseta, casona...* up to consider all the inflections for every canonical form (in this example, there would be 542 different words with some kind of relationship with the considered entry).

On the other hand, the search where the emphasis is more semantically than morphological can be done from the meanings that the joined affixes provide. For example, if it is desirable to look for adjectives derived from verbs that mean 'unable to carry out the action of the verb', using parasynthetic relations, it would be enough to discriminate for the affixes *in-...-ble: incansable, inconmovible, inconvencible, infatigable, irreprimible...*

A parasynthetic morpholexical relationship, between two words, exists when one of them has been formed from the other one by simultaneous addition of an affix on the left and another on the right —usually a prefix and a suffix— and also the grammatical categories and the semantics are just right. By way of example, *agrupar* possesses a semantic and functional relationship with *grupo*.

In a synchronous study, with automated analysis of the morphology in mind, formal or theoretical aspects may be not coincide with those strictly linguistics. There are Spanish words which have a strong functional and semantic relationship like the parasynthesis, but it is not. Though, there is a formal relationship through other stages in the evolution of the languages, so it is indeed considered necessary to include them —*cabeza* with *acapiza*, *pez* with *empegar*, *edad* with *coetáneo*. This concept must be restricted to avoid reaching the concept of related idea —which exceeds the objectives of this work, *nuevo* with *remozar*, *color* with *anaranjado*, *pelo* with *encabellecer*. Therefore, a historic-etymological meeting criterion is applied. It is obvious that for the speaker *aovado, arrocado* and *amelonado* are forms equally related with *huevo, rueca* and *melon* respectively —it must also be so for the automatic data processing. In order to solve the linguistic boundaries preventing to treat relationships beyond the strict parasynthesis, it is necessary to be located at a different level of the morphological level. Thus, the concept of morpholexical relationships is extended to improve the quality of the language processing in this aspect.

4. PARASYNTHETIC RELATIONSHIPS

A word might have suffered, with respect to another one, suffixal and prefixal alterations —parasynthesis. These alterations have been studied to establish the morpholexical relationships between two words from a synchronic point of view and applying the extended criteria explained in the previous section. When it comes to considering this type of relationships, the coincidence in the application of both affixes on the original word is significant in the proposed examples. The words **afrancés / *francesar, *encajón / *cajonar, *amujer / *mujerado* do not exist, therefore, the principle of simultaneity is conclusive.

```
francés    --> a-frances-ar
cajón      --> en-cajon-ar
mujer      --> a-mujer-ado
```

Consequently, a morpholexical relationship between an original word and one related by means of both a prefixal process and a suffixal one is established, when none of these processes have caused separately the existence of an intermediate word. Those words which, by prefixal alteration, have caused a new word with its own identity, and have also suffered a suffixal alteration or vice versa are not estimated as sharing those relationships, since the principle of simultaneity is lost. The intermediate word is called in this work as **previous-word**.

```
salir      --> sobre-salir --> sobresal-iente
América --> americ-ano --> anti-americano
verbo     --> verb-al      --> de-verbal
mar        --> mar-ino     --> sub-marino
músculo --> muscul-ar  --> intra-muscular
```

The lexicogenesis of some words —*deverbal*, *submarino*, *intramuscular*— cast doubt and has been doubly interpreted by different authors as parasynthetic —mainly by means of semantic characterization— or as prefixal —mainly by means of a formal characterization in the continuity of the formation. In those cases, which are likely of this double analysis, the prefixal interpretation is preferred, since the following criterion will be applied to establish the relationships: if there is not a previous-word between two forms, a parasynthetic relationship is established, otherwise, a suffixal or prefixal relationship, depending on the suitability.

It is obvious to exclude from the previous-word concept the coinciding words lexicographically which do not have the suitable semantics. A parasynthetic morpholexical relationship is established between *agua* and *enaguar*, in spite of the existence of the word *enagua* which does not have to be confused with a previous-word of *enaguar* because it does not has neither a semantic nor an etymological relationship neither with *agua* nor with *enaguar*. The same happens with *broma* and *abromar* with respect to *abroma*.

Those which have a suffixal or prefixal relationship parallel to the parasynthetic are also excluded. For example, since *desgarbo* has the same meaning as *desgarbado* ('without grace'), it is preferred to relate morpholexically the two words directly with *garbo* —prefixal and parasynthetic respectively— and *desgarbo* is not considered previous-word of desgarbado. The same happens with *atóxico* and *atoxicar* with respect to *tóxico*. The prefix *a-* in *atóxico* ('not toxic') has the meaning of negation, whereas in *atoxicar* ('poison with something toxic'), the prefix does not have a meaning of negation.

Certainly, we consider as parasynthetic relationship the one where there could be a possible previous word which, however, has not been consolidated in the lexicon studied, unless there are several words in the same situation with respect to the possible previous-word. In this case, a non existent word is kept to maintain the relationship among them, as in the example of *hipotiroideo* and *hipotiroidismo* with respect to *tiroides*.

Some words, irrespective of the formation process which they have suffered, keep a close semantic and functional relationship, as explained in the previous section, and therefore, they have to be considered in the applications which are expected to be developed. To offer an illustration, *entenebrecer* is closely related to *tiniebla* and there are not previous-words: **tenebrecer*, **entenebro*, **tineblecer*, **entinieblar*, nor **tenebro*. In computer applications, it is essential to consider this type of relationships if it is expected to cover a homogeneous set of families of words related by the same objective. This type of relationships, which have a common point in the etymologic history of the words involved and an analogy with parasynthetic relationships as for the affixes used and the semantics which they add, are treated as extended morpholexical relationships classified as parasynthesis in this work. Some examples of parasynthetic formation analogous to the case of *entenebrecer* are *endurecer* from *duro*, *embellecer* from *bello*, *embrutecer* from *bruto*, among others.

5. PARASYNTHETIC FAMILIES

Once we have established the parasynthetic extended morpholexical relationships between two words, it can be considered two groups of words which have the same kind of extended morpholexical relationship as regards as a common word called *original word*. They are words which belong to the same morpholexical field. This way, all the words which are parasyntheticly related with a word given will be call parasynthetic family, figure 1.
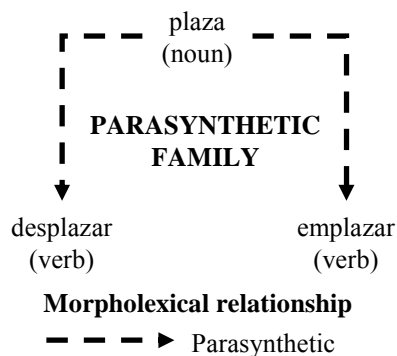
Figure 1. Parasynthetic family of *plaza*.

Since a word can be related to an original word and at the same time to be the original word in relationships linked to other words, a family tie is established between different families through this word. All the families related in this way make a **clan**.

5.1 Logical Structure

The word formation richness of Spanish along time, its irregularities and its peculiarities make it difficult —not impossible— to represent the extended morpholexical relationships between the elements of the lexicon in a non-diachronic way. In order to represent the different types of relationships which are deduced from the Spanish word formation rules and from the applied extended criteria, we have opted for a directed graph. The nodes identify the Spanish words, the arcs show that there is an extended morpholexical relationship between them; the direction of each arc corresponds to the relationship between the nodes, and the labels of the arcs classify the type of extended morpholexical relationship. The words in Spanish are, in this way, grouped together by disjunctive sets of interconnected elements —connected components of the graph.

Exceptionally, there are nodes with more than one arc directed towards it, which breaks the possible structure of a tree and the representation by means of a directed graph is necessary. This way, the word *incomodar* is parasynthetic of *cómodo* and a verbalization of *incómodo* which is, at the same time, a prefixation of *cómodo* —there are two ways to go through the graph to get to *incomodar* from *cómodo*, figure 2.
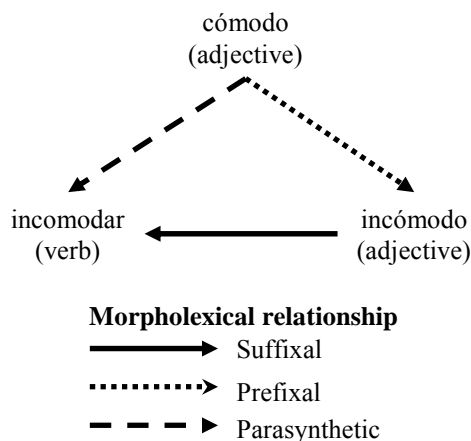

Figure 2. Graph that represents the morpholexical relationships with *incomodar*.

Exceptionally, two groups of extended morpholexical relationships can lose its connection —in a solely synchronic analysis—, because of the non-existence between them of the link node. This happens when extended morpholexical relationships have been established with other words from a non-existent original

word in the present Spanish. A word is considered non-existent when it is not mentioned in the reference sources because its use has been lost throughout time, the word has been substituted by another word with a different morphology or it simply has not been consolidated in Spanish language. In order to not to lose these relationships without incorporating new elements —archaisms or neologisms not considered by the reference sources—, the existence of the link node —labelled as non-existent— between the two groups of related nodes is established and its label is not visible for the final user of the application, although it keeps the conceptual interconnection in a connected component. In this way, we can establish morpholexical relationships between *nihilidad*, *nihilismo*, *nihilista*, *aniquilar* and *anihilar* through a non-existent word **nihil* —the Greek word that means 'nothing'—, figure 3.
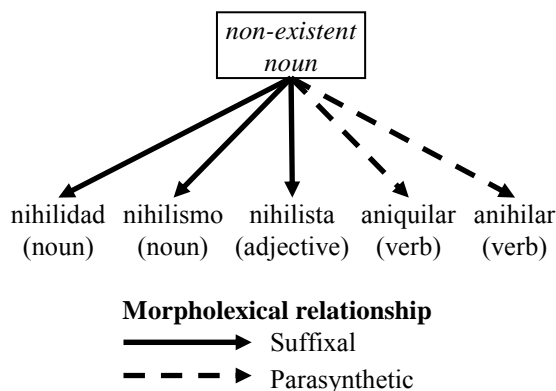
Figure 3. Morpholexical relationships through a non-existent word.

Here we show the connected component of a graph which represents a clan of extended morpholexical relationships. The suffixal and prefixal morpholexical relationships (Santana et al., 2004) are also displayed in order to understand better the structures and paths, figure 4.
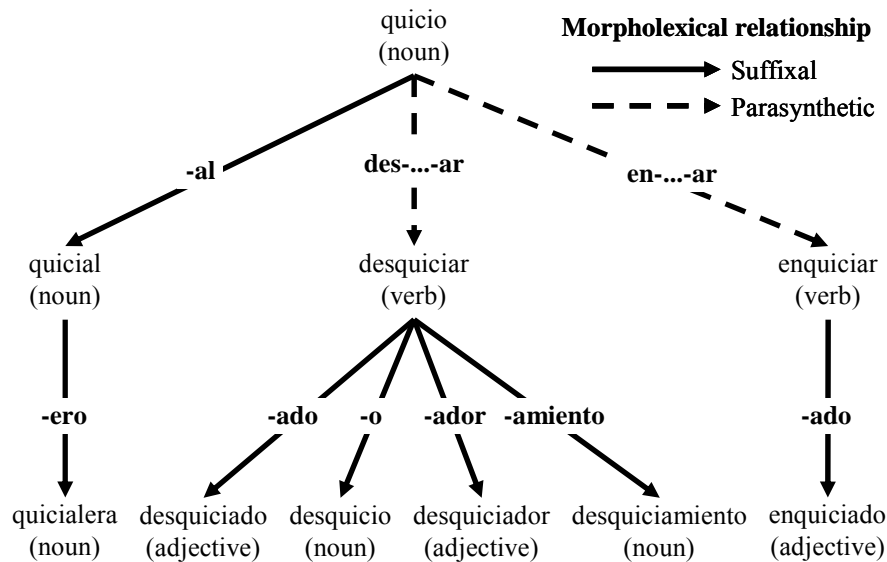
Figure 4. Morpholexical relationships clan of *quicio*.

5.2 Navigation

The information has been properly structured and catalogued in a way that lets effective access. The graph can be covered in any direction. From any node, we can get to any other with same related component of the

graph knowing at any time the extended morpholexical relationships —arcs— which are crossing to get to the destination. Starting from this premise, the different linguistic possibilities which are clear from this system are detailed. It has to be stressed that the non-existent nodes are kept internally to let the navigation among morpholexically related words.

5.2.1 Direction

It is linguistically interesting to know the family of words morpholexically related with a word given at certain proximity. According to the path which we follow in the graph —upward, downward, or horizontal—, we classify both proximity and morphology. From a word we obtain the ones which have suffered less number of derivative processes —upward—, the ones which have suffered the same number of alterations —horizontal—, and the ones which have suffered more derivative processes from it —downward. Now the different types considered are detailed.

5.2.2 Direct Ancestry

It is regarded as so, the reverse method to the parasynthesis, the process of obtaining the original word with which a specific word has been related to. To recognize this process in the words, all that needs to be done is following up the graph a level, if it exists. Therefore, the direct ancestry of the verb *desquiciar* is the substantive *quicio*. If the direct ancestry is applied twice, we obtain the original of the original of the current node. This way, the two levels direct ancestry of the adjective *desquiciador* is *quicio*.

5.2.3 Indirect Ancestry

The words morpholexically related to the direct ancestry and which are in the same level in the graph are regarded as indirect ancestries. The related words which have suffered one alteration less than the current word can be retrieved. In the *quicio* clan, the indirect ancestries of the adjective *desquiciador* are the verb *enquiciar* and the substantive *quicial*. Just like in the case of the direct ancestry, several levels of morpholexical relationships would be applied here.

5.2.4 Horizontal Direction

The words morpholexically related with the same original word and, therefore, which have suffered the same number of alterations, are regarded as horizontal direction. We manage to get them by retrieving the direct ancestry and going only one level down all the arcs of that node. This option retrieves all the members of a family from one of them —it does not include the original word. The verbs *desquiciar* and *enquiciar* are obtained from the substantive *quicial*. However, the words directly related to some indirect ancestry of level one are regarded as horizontal of second level. The substantive *quicialera* and the adjective *enquiciado* are obtained from the substantive *desquicio*.

5.2.5 Descendents

The descendents are the morpholexical relationships of a family from a given original word. The descendents of level two include the words that have a previous relationship with a same original word: it recovers the descendent family of each one of the members of the descendent family of the original word. Of the *quicio* clan, the descendent of the substantive *quicial* would be the substantive *quicialera*. The descendents of level two of the substantive *quicio* is all the substantives and adjectives of the nodes of the base of the graph of morpholexical relationships shown previously.

5.3 Filters

The output derived from the different kinds of navigation from a specific word can, sometimes, provide such a volume of information that it could make it difficult to find the words which are being looked up and the relationships which we want to see. These filters, in the extended morpholexical relationships, allow the selective discrimination of the navigation output. All the results, as consequence of the different types of navigation, are liable to be subject to different kinds of filters —functional, regularity, and by affixes.

6. RESULTS

Because of the non-existence of similar tools for Spanish, this system has been tried with a textual corpus containing more than 8 thousand texts, both literary and not literary (narrative, theatre, poetry, law, politics, history...), containing more than three hundred million words, of which more than half a million are different, proceeding from more than one hundred thousand canonical different forms.

It is proceeded to recognize morphologically the words to study the incidence of the morpholexical relationships that handles this system, on the lexical searches in this corpus. To every recognized word the processor of morpholexical relationships is applied to and its canonical morpholexical relationships are obtained. More than 3,300 millions of canonical morpholexical relationships from the first level are identified —150,243 are different—, of which more than 164 million are parasynthetic ones —6,603 are different. With extending the canonical morpholexical relationships to all their inflected words, generated by our morphologic processor, more than 50 billion of relationships are obtained —suffixal, prefixal and the parasynthetic ones but only those from first level.

With the treatment of prefixes and enclitic pronouns, witch can be carried out by our morphologic processor; the identified morpholexical relationships would overcome the trillion. Although with weaker relations, their number can keep on rising, as the considered levels increase, until getting to include the whole clan —the whole morpholexical paradigm of a word, see the example of the previously exhibited clan. This makes clear the potential of this system, for example, to locate words in a text.

7. CONCLUSIONS

A taxonomic, exhaustive and systematic study is made about affixes used in the derivative, prefixal and parasynthetic morphology of the Spanish on lexicon 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 lexicographical 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 a web application —suffixation, prefixation and parasynthesis. The web 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. 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 inflectional language —20% of irregularity in the suffixation, 7% of irregularity in the prefixation and 15% of irregularity in the parasynthesis.

The resulting system is characterized by simple and flexible algorithms with 100% success in the recognition and generation of morpholexical relationships of the language independently of the complexity of the data they are handling. This system can be incorporated easily to other tools from aid to the oriented document treatment to solve problems of the natural language processing. The principal potential of this application is that it allows lexical searches beyond the lexicographical regularity.

This kind of systems supposes a first step towards the multiple possibilities in computer science and specialized programs that they must be developed on this knowledge base.

REFERENCES

Alcoba Rueda, S. 1992. Tema verbal y formación de palabras en español. *Actas do XIX Congreso Internacional de Lingüística e filoloxía románicas*, Vol. II. Universidad de Santiago de Compostela, La Coruña, Spain, pp. 323-346.

Almela Pérez, R. 1999. *Procedimientos de formación de palabras en español*. Ed. Ariel, Barcelona.

Alvar Ezquerra, M. 2002. *La formación de las palabras en español*. 5th edn. Ed. Arco/Libros, Madrid.

Alvar Ezquerra, M. 2003. Nuevo d*iccionario de voces de uso actual*. Ed. Arco/Libros, Madrid.

Bajo Pérez, E. 1997. *La derivación nominal en español*. Ed. Arco/Libros, Madrid.

Biblograf, s.a. 1997. *Diccionario General de la Lengua Española VOX* en CD-ROM. Biblograf, s.a. Barcelona.

Casares, J. 1990. *Diccionario Ideológico de la Lengua Española*, 2ª Edición. Ed. Gustavo Gili, s.a. Barcelona.

Clave SM.1997. *Diccionario de Uso del Español Actual*. Clave SM, edición en CD-ROM. Madrid.

Dee, J. 1997. *Volume I Introduction and Lexicon. A lexicon of latin derivatives in Italian, Spanish, French and English.* Olms-Weidmann, New York.

Dee, J. 1997. *Volume II Index. A lexicon of latin derivatives in Italian, Spanish, French and English.* Olms-Weidmann, New York.

Espasa-Calpe. 1991. *Gran Diccionario de Sinónimos y Antónimos*, 4ª edic. Espasa-Calpe, Madrid.

Faitelson-Weiser, S. 1993. Sufijación y derivación sufijal: sentido y forma. *La formación de palabras.* Varela (ed.), Taurus, Madrid, pp. 117-161.

Lang, M. 1992. *Formación de palabras en español. Morfología derivativa productiva en léxico moderno.* Cátedra, Madrid.

Larousse Planeta, s.a. 1996. *Gran Diccionario de la Lengua Española.* Larousse Planeta, s.a., Barcelona.

Moliner, M. 1996. *Diccionario de Uso del Español*, edición en CD-ROM. Gredos, Madrid.

Polguère, A. 2000. Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French. *Proceedings of EURALEX'2000.* Stuttgart, pp. 517-528.

Real Academia Española and Espasa-Calpe. 1995. *Diccionario de la Lengua Española*, edición electrónica, versión 21.1.0. Real Academia Española y Espasa-Calpe, Madrid.

Santana O., Carreras, F., Pérez, J. and Rodríguez, G. 2003. Relaciones morfoléxicas sufijales del español. *Procesamiento de Lenguaje Natural,* Vol. 30, Ed. SEPLN, Madrid. pp. 1-73.

Santana, O., Pérez, J., Carreras, F., Duque, J., Hernández, Z. and Rodríguez, G. 1999. FLANOM: Flexionador y lematizador automático de formas nominales. *Lingüística Española Actual*, XXI, Vol. 2, Ed. Arco/Libros, S.L. Madrid. pp. 253-297.

Santana, O., Pérez, J., Carreras, F. and Rodríguez, G. 2004. Relaciones morfoléxicas prefijales del español. *Procesamiento de Lenguaje Natural,* 32, Ed. SEPLN, Madrid. pp. 9-36.

Santana, O., Pérez, J., Hernández, Z., Carreras, F. and Rodríguez, G. 1997. FLAVER: Flexionador y lematizador automático de formas verbales. *Lingüística Española Actual*, XIX, Vol. 2, Ed. Arco/Libros, S.L. Madrid. pp. 229-282.

Santana, O., Pérez, J., Carreras, F. and Rodríguez, G. 2004. Suffixal and Prefixal Morpholexical Relationships of the Spanish, *Lecture Notes in Artificial Inteligence,* 3230, Ed. Springer-Verlag. pp. 407-418

Seco, M. 1991. *Diccionario de dudas y dificultades de la lengua española*, 9ª Edición. Espasa-Calpe, Madrid.

Socrates, D., Baldzis, S., Kolalas, A. and Eumeridou, E. 2005. The Computational Modern Greek Morphological Lexicon ―An Efficient and Comprehensive System for Morphological Analysis and Synthesis. *Literary and Linguistic Computing*, Vol 2, No. 20, pp. 153-187.

Varela, S. (ed.). 1993. *La formación de palabras.* Taurus, Madrid.