The Spanish Morphology in Internet

Octavio Santana, José Pérez, Francisco Carreras,
Zenón Hernández and Gustavo Rodríguez

Departamento de Informática y Sistemas,
Universidad de Las Palmas de Gran Canaria,
Campus Universitario de Tafira,
35017 Las Palmas de Gran Canaria, Spain
{OSantana, JPerez, FCarreras,
ZHernadez, GRodriguez}@dis.ulpgc.es
http://www.gedlc.ulpgc.es

Abstract. This Web service tags morpholexically any Spanish word and it gets the corresponding forms starting from a canonical form and from the flexion asked for. In the verbs, it deals with the simple and compound conjugation, the enclitic pronouns, the flexion of the participle like verbal adjective and the diminutive of the gerund. With the nonverbal forms, this web service considers: gender and number, heteronomy for change of sex, superlative degree, adverbiation and the appreciative derivation. In the tag and in the generation the prefixation is taken into account. It allows the manipulation of morpholexical relationships. It offers a global vision of the behavior and productivity of the Spanish words in the principal processes of formation (sufixation, prefixation, parasinthesis, suppression, regression, zero-modification, apocopation, metathesis and others which are unclassifiable and that generate alternative graphical forms). It includes the principal Spanish lexicographic repertoires. It considers 151103 canonical forms that produce more than 4900000 flexioned and derived forms and about 90000 morpholexical relationships are established.

1 Introduction

As result of the work done by the Group of Data Structures and Computacional Linguistic of the University of Las Palmas de Gran Canaria (http://www.gedlc.ulpgc.es), a Web application has been created able to interpret and to handle with versatility different relevant aspects of the Spanish morphology. The application represents more a form to show the potentiality of a Web System of management of the morphology of the Spanish language than a finalist tool. This system allows its integration in other useful tools for the natural language processing like orthographic correction, advanced search information, text analysers, disambiguousers, lexicographical station, parsers, information extraction, text automatic generation, syntactic correction and extraction of summaries, among other

applications, all of them aimed at offering interactive services distributed through Internet.

The automatized treatment of the Spanish morphology arouses great interest because it constitutes the first touchstone on which to construct any natural language processor and starts the way towards future Web services more specialized in the handling, learning and control of that great human potential –the language.

The available system in the Web tags any Spanish word identifying its canonical form, grammar category and the flexion or derivation that produces it, and is able to produce the corresponding forms from a canonical form and from the flexion or derivation asked for; both the recognition and the generation operate on a same data structure, to cross it in opposite senses implies that the tool works in one or another modality.

In the verbs, it deals with the simple and compound conjugation, the enclitic pronouns, the flexion of the participle like verbal adjective (gender and number) and the diminutive of the gerund. With the nonverbal forms, it considers: the gender and the number in the nouns, adjectives, pronouns and articles; heteronomy by sex change in the nouns; the superlative degree in the adjectives and adverbs; the adverbiation and the superlative adverbiation in the adjectives; the appreciative derivation in the nouns, adjectives and adverbs; the graphical variants in all the grammar categories and the invariant forms such as preposition, conjunctions, exclamations, words of other languages and locutions or phrases. As much in the tag as in the generation the incorporation of prefixes is discretionarily considered.

In addition it allows the recognition, the generation and the manipulation of the morpholexical relations of any word, it includes the recovery of all its lexicogenetic information until arriving at a primitive one, the management and control of the affix in the treatment of its relations, as well as the regularity in the established relation. It provides a global vision of the behavior and productivity of the Spanish words in the main processes of formation (sufixation, prefixation, parasinthesis, suppression, regression, modification-zero, apocopation, metathesis and nonclassifiable others that generate alternative graphical forms).

For the accomplishment of this work, a corpus of Spanish words with the lexicon of different dictionaries has been created: the *Diccionario de la Lengua Española de la Real Academia*, the *Diccionario General de la Lengua Española VOX*, the *Diccionario de uso del Español de María Moliner*, the *Gran Diccionario de la Lengua Española de Larousse*, the *Diccionario de Uso del Español Actual Clave SM*, the *Diccionario de Sinónimos y Antónimos de Espasa-Calpe*, the *Diccionario Ideológico de la Lengua Española de Julio Casares* and the *Diccionario de voces de uso actual* directed by Manuel Alvar Ezquerra. From 151103 canonical forms (which include about 15000 proper names and 9000 adjectives coming from participles of verbs not registered in the previous repertoires), they obtain more than 4900000 flexionned and derived forms (without adding the inherent extension to the prefixes and the enclitic pronouns) and around 90000 morpholexical relations are established.

2 The Tagger

This service (http://www.gedlc.ulpgc.es/investigacion/scogeme02/lematiza.htm) allows the user to tag any Spanish word •it identifies its canonical form, grammar category and the flexion, derivation and prefix that affects to it. The tag is the first advisable step for the users of the system: on the one hand, it makes the entry flexible when allowing to have access to the rest of the available services without previously having neither morphologic nor grammar knowledge and, on the other hand, it identifies without ambiguity the different morphologic interpretations of the inserted form. Thus, for example, the answer for the entry *habiéndose sobreaterrado* is:

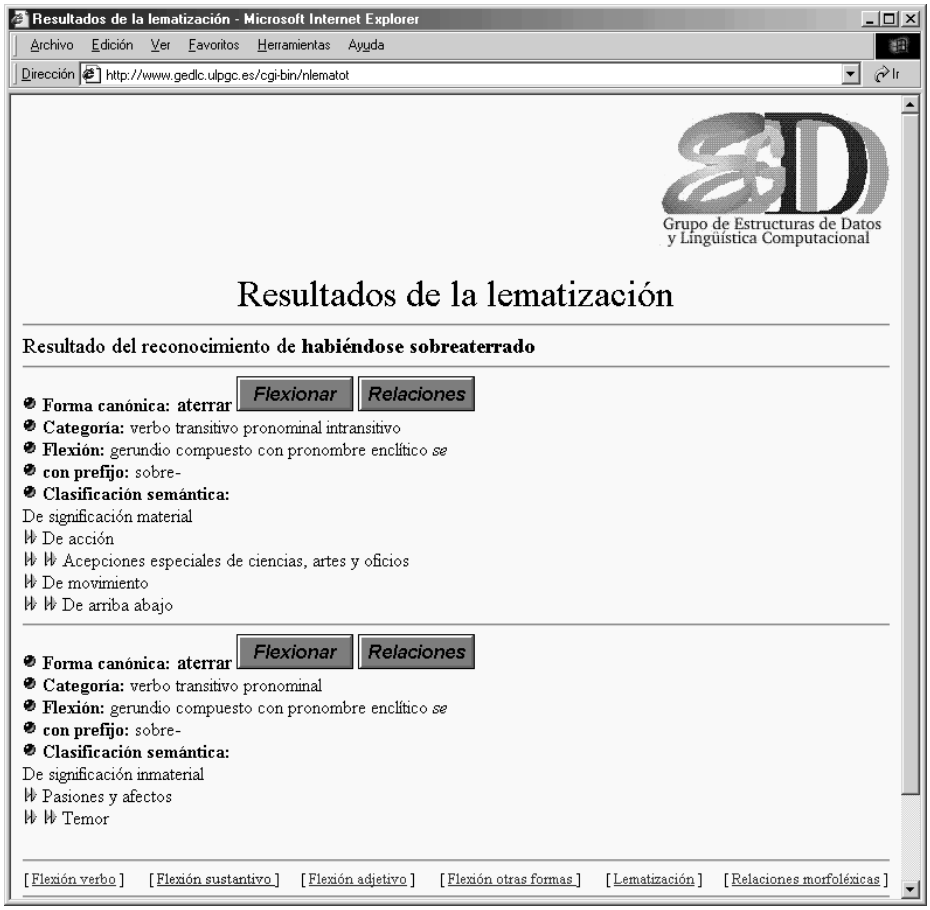

Fig. 1. Example of tagger

Next to each canonical form two links are provided that allow to flexion it and to obtain their morpholexical relations. These links lead in an intelligent way to the suitable services: for the verbs, the conjugater, and for the nonverbal forms the

possible flexions are supplied according to the grammar category of the canonical form associated •the prefixation is also transferred. Additionally, in the verbs, semantic information useful in several contexts is shown and this can help to identify better the infinitive and, consequently, the entry.

The abstraction that this system offers on the morphologic irregularities of the Spanish language makes its use easier and, in no case, diminishes its precision. Thus, for example, for the entries with irregular flexions *cuélguemelo*, *bendigo*, *óyelo*, *fue*, *habiéndole dicho*, *nazca*, *degüellan*, *muévete*, *riño*, *tengo* and *vi*, the corresponding infinitives are provided: *colgar*, *bendecir*, *oír*, *ir* or *ser*, *decir*, *nacer*, *degollar*, *mover*, *reñir*, *tener* and *ver*; either to identify the forms •irrespective of the existence of enclitic pronouns in the words•, or to have access later to the conjugation service *Flexioner: verb*, with which all previous morphologic knowledge of the language on the part of the user is eluded. And in the nonverbal examples, also with irregular flexion: *paupérrimo*, *sapientísimo*, *bestezuelo*, *blanquecino*, *boyazo*, *ovecico*, *hijosdalgo*, *cualesquiera*, *robaliza*, *flámines*, *princesa*, the service provides all the canonical forms from which these words can come from with the corresponding information of the flexion applied in each case: *pobre*, *sabio*, *bestia*, *blanco*, *buey*, *huevo*, *hijodalgo*, *cualquiera*, *róbalo* or *robalo*, *flamen* and *príncipe*. This allows to accede to the suitable service *Flexioner*.

Special mention deserves the incorporation of enclitic pronouns to the verbal forms. The system admits the appearance of up to three pronouns for a verbal form •either simple or compound•, thus *comerme*, *habérselo comido* or *comiéndosenosla* lead us to the infinitive *comer*; and, of course, the service takes into account the necessary rules of union, not only as far as the incorporation of accents is concerned, but also, as far as the modificationns in original forms is concerned •*temeos* is *temed* with the enclitic pronoun *os* and *amémonos* is *amemos* with the enclitic pronoun *nos*.

Another important flexibility is the recognition of the existence of prefixes, discretionarily incorporated in the form to tag. It avoids, therefore, knowledge in this area •as to which prefixes exist in Spanish and which the rules of union and their irregularities are• for the use of the system if on starts from words such as: *predigestión*, *reatacar*, *irrealización*, *coadmitido*, etc.

3 The verbal flexioner.

This service (http://www.gedlc.ulpgc.es/investigacion/scogeme02/flexver.htm) allows the user to obtain, from a verb, the conjugation of a simple or composed tense, nonpersonal simple or compound forms, the flexion of the participle as a verbal adjective and the diminutive of the gerund. In addition, it is possible to select from three pull-down lists the valid combinations of one, two or three enclitic pronouns, so that the system incorporates them correctly to the required conjugated forms. Thus, for example, the past perfect of indicative of the infinitive *temer* with the enclitic pronoun *le* is:

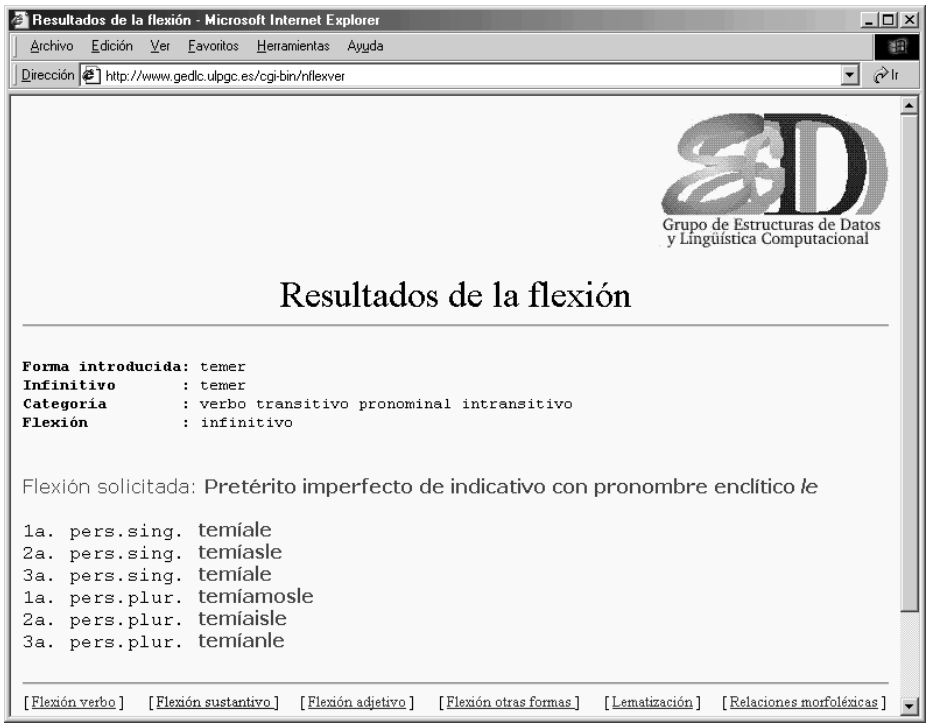

Fig. 2. Example of verbal flexioner

Like it happens with the tagger, in the generation of flexioned forms, the irregularities that appear in the conjugation models are considered, in addition to all the changes in the graphical form as a result of the incorporation of the pronouns to the conjugated forms. Thus, for example, the indicative present of the infinitive *oler* with the enclitic pronouns *se* and *lo* is presented in Table 1.

Table 1. For input *oler*

The indicative present of *oler* with *se* and *lo*	
First person of the singular	*huéloselo*
Second person of the singular	*huéleselo*
Third person of the singular	*huéleselo*
First person of the plural	*olémoselo*
Second person of the plural	*oléiselo*
Third person of the plural	*huélenselo*

It is possible to add up to three prefixes discretionarily to the conjugated forms, selecting them of three pull-down lists alphabetically organized. The system considers for each prefix the general rules of application in this word formation process and the specific ones. Thus, for example, if the prefix is gotten up *re-* to the first and third person of the singular of the indefinite past of the verb *hacer •hice* and *hizo• rehíce*

and *rehízo* are obtained; and the simple nonpersonal forms for *emitir* with the prefixes *trans-*, *re-* and *co-* are: *corretransmitir*, *corretransmitiendo* and *corretransmitido*.

Since any Spanish word is admitted like input to the conjugater, all the requests are previously treated by the tagger in order to apply the flexions on the infinitives from which the inputs come. If the input is a infinitive is conjugated, if is a conjugated form its infinitivo is flexioned and if the input does not come from any verb is the result of its tag with intelligent links to the flexion pages. In all the cases it can be implied more than one canonical form. Thus, for example, if it is the present of indicative asked for *aterrar*, *fue* and *óptimas*, the different answers are presented in Table 2, Table 3 and Fig. 3.

Table 2. For input *aterrar*

Of the infinitive *aterrar* (transitive pronominal intransitive)	
First person of the singular	*atierro*
Second person of the singular	*atierras*
Third person of the singular	*atierra*
First person of the plural	*aterramos*
Second person of the plural	*aterráis*
Third person of the plural	*atierran*
Of the infinitive *aterrar* (transitive pronominal intransitive)	
First person of the singular	*aterro*
Second person of the singular	*aterras*
Third person of the singular	*aterra*
First person of the plural	*aterramos*
Second person of the plural	*aterráis*
Third person of the plural	*aterran*

Table 3. For input *fue*

Of the infinitive *ser*	
First person of the singular	*soy*
Second person of the singular	*eres*
Third person of the singular	*es*
First person of the plural	*somos*
Second person of the plural	*sois*
Third person of the plural	*son*
Of the infinitive *ir*	
First person of the singular	*voy*
Second person of the singular	*vas*
Third person of the singular	*va*
First person of the plural	*vamos*
Second person of the plural	*vais*
Third person of the plural	*Van*

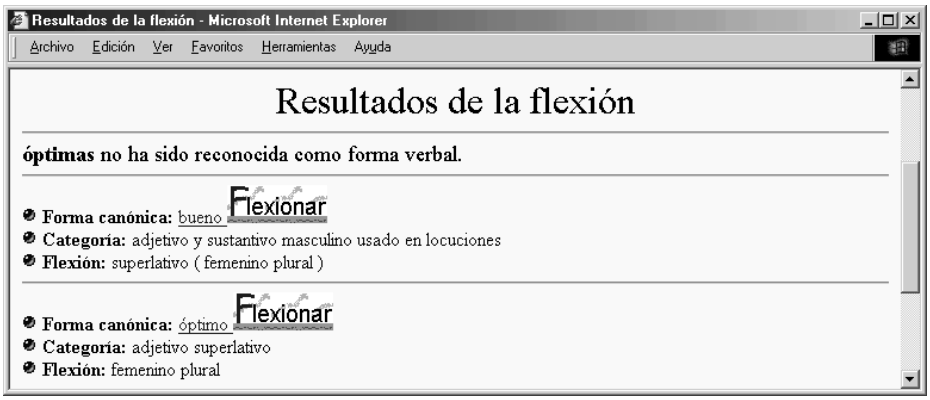

Fig. 3. For input *óptimas*

4 The nominal flexioner

This service allows the user to obtain, from a noun, the changes of gender and number and the appreciative derivation for the selected gender and number. Thus, for example, the diminutives in masculine plural of *pez* are: *pececitos*, *pececillos*, *pececicos*, *pececines* and *pecezuelos*.

In the generation of flexionned forms, those irregularities that appear in the formation of feminine and plural are considered, apart from a great variety of irregularities that the appreciative suffixation in Spanish has. Someone outstanding examples are: the feminine of *rey* is *reina*, the feminines of *actor* are *actora* or *actriz*, the plural of *régimen* is *regímenes*, the plurals of *maniquí* are *maniquíes* or *maniquís*, *velón* exists as augmentative masculine of the feminine noun *vela* and the irregular appreciative forms of *chico* are *chicorrotín*, *chiquitín*, *chicorrotico*, *chicorrotillo*, *chicorrotito*, *chiquirritín*, *chiquilín*, *chiquirritico*, *chiquirritillo* and *chiquirritito*. If one asks for the plural feminine form of *toro*, the service indicates that it does not admit it and offers *vacas* like an option •heteronomy.

Like in the verbal conjugator, it is also possible to add prefixes to the flexioned forms. Thus, for example, if the preffix *anti-* is joined to the form *son*, we obtain *antisón*; if we join *nutación* to the preffix *circun-* we read *circumnutación*; and if the prefix *exo-* is joined to the form *ósmosis* it generates *exosmosis* and *exósmosis*.

Since the flexioner admits in any Spanish word, the requests are previously examined by the tagger in order to apply the corresponding flexions on the canonical forms, just like the conjugater does. If the entry is a nominal canonical form, it is flexioned; if it is a flexioned nominal form, the canonical form from which it comes is flexioned; and if the entry does not come from any nominal form, the result of its tag is shown with intelligent links to the suitable flexion pages. Thus, for example, if the masculine plural of *libro*, *barcucho* and *bebió* are asked for, the different answers

are: *libros*, *barcos* •from the canonical form *barco*• and *bebió* is not recognized as a noun •the system offers to conjugate the verb *beber*.

5 The adjectival flexioner

This service (http://www.gedlc.ulpgc.es/investigacion/scogeme02/flexadj.htm) is similar to the one for nouns, although it adds the specific flexions of the adjectives: the superlative degree, the adverbiation and the superlative adverbiation. In addition to the formation rules, the existing irregularities for these flexions are considered. Thus, for example, the superlative in masculine plural of *blanco* is *blanquísimo*, the adverbiation of *pobre* is *pobremente* and the adverbiation of the superlative is *pobrísimamente* and *paupérrimamente*.

The treatment of prefixes and the flexibility in the processing of the entries are completely analogous to the one offered by the *Flexioner: noum* service.

6 The flexioner of other forms

This service (http://www.gedlc.ulpgc.es/investigacion/scogeme02/flexotra.htm) allows the user to obtain, from pronouns, articles, adverbs, prepositions, conjunctions, exclamations, words and expressions from other languages, changes in gender and number and appreciative forms when possible. In addition to the formation rules, the existing irregularities for these flexions are considered. Thus, for example, the neutral singular of the personal pronoun *él* is *ello*, the feminine plural of the article *el* is *las* and the diminutive forms of the adverb *cerca* are *cerquita* and *cerquininga*.

The treatment of prefixes and the flexibility in the processing of the entry are completely analogous to those offered by the previous services, although in these forms the incorporation of prefixes is not admitted.

7 The morpholexical relations

The primary target of this service (http://www.gedlc.ulpgc.es/investigacion/scogeme02/relmorfo.htm) consists of obtaining a set of morpholexical relations between useful Spanish words for applications of processing of the natural language. In a synchronous study, and with the glance put in the automatic treatment of the morphology with computer science means, the formal or theoretical aspects do not have to agree with the strictly linguistic ones. There exist Spanish words that maintain a strong semantic and functional relation •the same that appears in the derivative or prefix level•, and that cannot take derivation or prefixation, although yes a formal relation through other stages in the evolution of the languages exists, that is why it is considered necessary to include them •*agua* with *acuoso*, *vejiga* with *vesical*, *conejo* with *cunicular*. This

concept must be restricted to avoid to arrive at the concept of related notion •*blanco* with *albura*, *sólido* with *endurecer*, *niño* with *pueril*•, that is why a criterion of historic-etimologic confluence is applied. It is obvious that, for the speaker, and therefore for computer science *acuario*, *portuario* and *campanario* must be places equally related to *agua*, *puerto* and *campana*. So, it is necessary to pass to another level beyond morphology level, in order to resolve linguistic barriers that they would prevent to treat relations beyond the derivation or the prefixation; the concept of morpholexical relation is extended in this way.

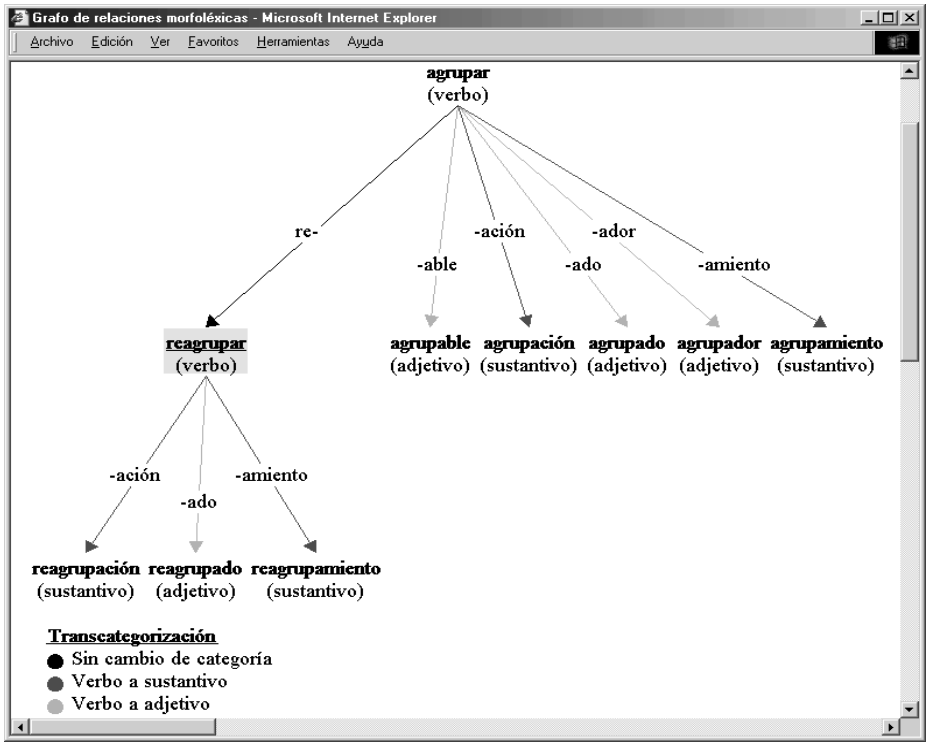

Fig. 4. Example for morpholexical relations

Derivative forms by grammar category •nouns, adjectives, verbs, adverbs and other categories• can be obtained from a canonical form. Thus, for example, the nouns derived from *peso* are *pesero*, *peseta* and *pesillo*. Also it is possible to ask for the primitive form and the words that have undergone some process of formation from the entry •suffixal, preffixal, parasinthetic and less frequent others like suppression, regression, zero-modification, apocope, metathesis, aphaeresis and some nonclassifiable ones that generate alternative spellings. Thus, for example, the primitive form of *cartero* is *carta* and the formed ones from *tubo* are: *tuba*, *tubería*, *tubular*, *tubuloso*, *multitubo*, *entubar* and *intubar* •they are showed according to the applied process of formation. All the requests allow filtering the words by its process of formation •regular or irregular.

There is an option that facilitates to obtain the set of words morpholexically closer to the given one •its primitive, its derivatives and the derivatives from its primitive. Another option serves to obtain the complete family of words with any morpholexical relation. These two possibilities allow in addition their graphical presentation, Fig. 4.

Due to any Spanish word is admited as input by the morpholexical relations service, the request is previously treated by the tagger in order to apply the adequate relations on the canonical forms. If the entry comes from more than one canonical form with relations, the requested relations for each canonical form are obtained.

References

1. Alsina, R.: Todos los Verbos Castellanos Conjugados. 17th edn. Teide, Barcelona (1990)
2. Alvar Ezquerra, M.: La formación de palabras en español. 5th edn. Arco/Libros, Madrid (2002)
3. Alvar Ezquerra, M.: Nuevo diccionario de voces de uso actual. Arco/Libros, Madrid (2003)
4. Carreras, F.: Sistema Computacional de Gestión Morfológica del Español (SCOGEME). Ph. Degree Thesis directed by O. Santana and J. Pérez, Universidad de Las Palmas de Gran Canaria (2002)
5. Casares, J.: Diccionario Ideológico de la Lengua Española. 2nd edn. Gustavo Gili, Barcelona (1990)
6. Corripio Pérez, F.: Diccionario Práctico. Incorrecciones: Dudas y Norma Gramatical. Larousse Planeta, Barcelona (1995)
7. Diccionario de la Lengua Española. Edición electrónica. Electronic edn. 21.1.0, Real Academia Española and Espasa Calpe, Madrid (1995)
8. Diccionario de Uso del Español Actual. Clave. Electronic edition, SM, Madrid (1997)
9. Diccionario de Uso del Español de María Moliner. 2nd edn. Electronic edn., Gredos, Madrid (2001)
10. Diccionario General de la Lengua Española Vox. Electronic edn. Biblograf, Barcelona (1997)
11. Gran Diccionario de la Lengua Española. Larousse Planeta, Barcelona (1996)
12. Gran Diccionario de Sinónimos y Antónimos. 4th edn. Espasa Calpe, Madrid (1991)
13. Gómez Torrego, L.: Manual de Español Correcto, 10th edn. Arco/Libros, Madrid, (2000)
14. Real Academia Española: Esbozo de una nueva gramática de la lengua española. 1st edn. Espa-sa Calpe, Madrid (1989)
15. Santana, O., Carreras, F., Hernández, Z., Pérez, J., Rodríguez, G.: Manual de la conjugación del español. 12790 verbos conjugados. Arco/Libros, Madrid (2002)
16. Santana, O., Pérez, J., Carreras, F., Duque, J., Hernández, Z., Rodríguez, G.: FLANOM: Flexionador y lematizador automático de formas nominales. Lingüística Española Actual, Vol. XXI-2, Arco/Libros, (1999) 253–297
17. Santana, O., Pérez, J., Hernández, Z., Carreras, F., Rodríguez, G.: FLAVER: Flexionador y lematizador automático de formas verbales. Lingüística Española Actual, Vol. XIX-2, Arco/Libros, Madrid (1997) 229–282
18. Santana, O.; Pérez, J.; Losada, L.: Generación automática de respuestas en análisis morfológico. Estudios de lingüística, Universidad de Alicante, Vol. 14, (2000) 245–257
19. Seco, M.: Diccionario de dudas y dificultades de la lengua española. 9th edn. Espasa Calpe, Madrid (1991)