Automatic conjugation and identification of regular and irregular verb neologisms in Spanish

Luz Rello and Eduardo Basterrechea
Molino de Ideas s.a.
Nanclares de Oca, 1F
Madrid, 28022, Spain
{lrello, ebaste}@molinodeideas.es

Abstract

In this paper, a novel system for the automatic identification and conjugation of Spanish verb neologisms is presented. The paper describes a rule-based algorithm consisting of six steps which are taken to determine whether a new verb is regular or not, and to establish the rules that the verb should follow in its conjugation. The method was evaluated on 4,307 new verbs and its performance found to be satisfactory both for irregular and regular neologisms. The algorithm also contains extra rules to cater for verb neologisms in Spanish that do not exist as yet, but are inferred to be possible in light of existing cases of new verb creation in Spanish.

1 Introduction

This paper presents a new method consisting of a set of modules which are implemented as part of a free online conjugator called *Onoma*[1].

The novelty of this system lies in its ability to identify and conjugate existing verbs and potential new verbs in Spanish with a degree of coverage that cannot completely be achieved by other existing conjugators that are available. Other existing systems do not cope well with the productively rich word formation processes that apply to Spanish verbs and lead to complexities in their inflectional forms that can present irregularities. The operation of these processes mean that each Spanish verb can comprise 135 different forms, including compound verb forms.

[1] *Onoma* can be accessed at http://conjugador.onoma.es

Several researchers have developed tools and methods related to Spanish verbs. These include morphological processors (Tzoukermann and Liberman, 1990), (Santana et al., 1997), (Santana et al., 2002), semantic verb classification (Esteve Ferrer, 2004) or verb sense disambiguation (Lapata and Brew, 2004). Nevertheless, to our knowledge, ours is the first attempt to automatically identify, classify and conjugate new Spanish verbs.

Our method identifies new and existing Spanish verbs and categorises them into seven classes: one class for regular verbs and six classes of irregular verbs depending on the type of the irregularity rule whose operation produced it. This algorithm is implemented by means of six modules or transducers which process each new infinitive form and classify the neologism. Once the new infinitive is classified, it is conjugated by the system using a set of high accuracy conjugation rules according to its class.

One of the advantages of this procedure is that only very little information about the new infinitive form is required. The knowledge needed is exclusively of a formal kind. Extraction of this information relies on the implementation and use of two extra modules: one to detect Spanish syllables, and the other to split the verb into its root and morphological affixes.

In cases where the neologism is not an infinitive form, but a conjugated one, the system generates a hypothetical infinitive form that the user can corroborate as a legitimate infinitive.

Given that the transducers used in this system are easy to learn and remember, the method can be employed as a pedagogic tool itself by students of

Proceedings of the NAACL HLT 2010 Second Workshop on Computational Approaches to Linguistic Creativity, pages 1–5,
Los Angeles, California, June 2010.

Spanish as a foreign language. It helps in the learning of the Spanish verb system since currently existing methods (e.g. (Puebla, 1995), (Gomis, 1998), (Mateo, 2008)) do not provide guidance on the question of whether verbs are regular or irregular. This is due to the fact that our method can identify the nature of any possible verb by reference only to its infinitive form. The application of other kinds of knowledge about the verb to this task are currently being investigated to deal with those rare cases in which reference to the infinitive form is insufficient for making this classification.

This study first required an analysis of the existing verb paradigms used in dictionary construction (DRAE, 2001) followed by the detailed examination of new verbs' conjugations (Gomis, 1998), (Santana et al., 2002), (Mateo, 2008) compiled in a database created for that purpose. For the design of the algorithm, in order to validate the rules and patterns, an error-driven approach was taken.

The remainder of the paper is structured as follows: section 2 presents a description of the corpora used. In Section 3, the different word formation processes that apply to Spanish verbs are described, while Section 4 is devoted to the detailed description of the rules used by the system to classify the neologisms, which are evaluated in Section 5. Finally, in Section 6 we draw the conclusions.

2 Data

Two databases were used for the modeling process. The first (named the DRAE Verb Conjugation Database (DRAEVC-DB)) is composed of all the paradigms of the verbs contained in the 22nd edition of the Dictionary of the Royal Spanish Academy (DRAE, 2001). This database contains 11,060 existing Spanish verbs and their respective conjugations. The second database (named the MolinoIdeas Verb Conjugation Database (MIVC-DB)), created for this purpose, contains 15,367 verbs. It includes all the verbs found in the DRAE database plus 4,307 conjugated Spanish verbs that are not registered in the Royal Spanish Academy Dictionary (DRAE, 2001), which are found in standard and colloquial Spanish and whose use is frequent on the web.

The MIVC-DB contains completely conjugated verbs occurring in the Spanish Wikipedia and in a collection of 3 million journalistic articles from newspapers in Spanish from America and Spain[2].

Corpus	Number of verbs
DRAE	11,060
MolinoIdeas	15,367

Table 1: Corpora used.

Verbs which do not occur in the Dictionary of the Royal Spanish Academy (DRAE, 2001) are considered neologisms in this study. Thus 4,307 of the 15,367 verbs in the MIVC-DB are neologisms. The paradigms of the new verbs whose complete conjugation was not found in the sources were automatically computed and manually revised in order to ensure their accuracy. The result of this semi-automatic process is a database consisting only of attested Spanish verbs.

3 Creativity in Spanish verbs

The creation of new verbs in Spanish is especially productive due to the rich possibilities of the diverse morphological schema that are applied to create neologisms (Almela, 1999).

New Spanish verbs are derived by two means: either (1) morphological processes applied to existing words or (2) incorporating foreign verbs, such as *digitalizar* from *to digitalize*.

Three morphological mechanisms can be distinguished: prefixation, suffixation and parasynthesis. Through prefixation a bound morpheme is attached to a previously existing verb. The most common prefixes used for new verbs found in our corpus are the following: *a- (abastillar), des- (desagrupar), inter- (interactuar), pre- (prefabricar), re- (redecorar), sobre- (sobretasar), sub- (subvaluar)* and *super- (superdotar)*. On the other hand, the most frequent suffixes in Spanish new verbs are *-ar (palar), -ear (panear), -ificar (cronificar)* and *-izar (superficializar)*. Finally, parasynthesis occurs when the suffixes are added in combination with a prefix (bound morpheme). Although parasynthesis is rare in other grammatical classes, it is quite relevant in the creation of new Spanish verbs (Serrano,

[2]The newspapers with mayor representation in our corpus are: *El País, ABC, Marca, Público, El Universal, Clarín, El Mundo* and *El Norte de Castilla*

1999). The most common prefixes are *-a* or *-en* in conjunction with the suffixes *-ar, -ear, -ecer* and *-izar* (*acuchillear, enmarronar, enlanguidecer, abandalizar*).

In this paper, the term derivational base is used to denote the immediate constituent to which a morphological process is applied to form a verb. In order to obtain the derivational base, it is necessary to determine whether the last vowel of the base is stressed. When the vowel is unstressed, it is removed from the derivational base while a stressed vowel remains as part of the derivational base. If a consonant is the final letter of the derivational base it remains a part of it as well.

4 Classifying and conjugating new verbs

Broadly speaking, the algorithm is implemented by six transduction modules arranged in a switch structure. The operation of most of the transducers is simple, though Module 4 is implemented as a cascade of transduction modules in which inputs may potentially be further modified by subsequent modules (5 and 6).

The modules were implemented to determine the class of each neologism. Depending on the class to which each verb belongs, a set of rules and patterns will be applied to create its inflected forms. The proposed verb taxonomy generated by these transducers is original and was developed in conjunction with the method itself. The group of patterns and rules which affect each verb are detailed in previous work (Basterrechea and Rello, 2010). The modules described below are activated when they receive as input an existing or new infinitive verb form. When the infinitive form is not changed by one transducer, it is tested against the next one. If not adjusted by any transducer, then the new infinitive verb is assumed to have a regular conjugation.

Module 1: The first transducer checks whether the verb form is an auxiliary verb (*haber*), a copulative verb (*ser* or *estar*), a monosyllabic verb (*ir*, *dar* or *ver*), a Magnificent verb[3], or a prefixed form whose derivational base matches one of these aforementioned types of verbs. If the form matches one of these cases, the verb is irregular and will undergo the rules and patterns of its own class. (Basterrechea and Rello, 2010).

Module 2: If the infinitive or prefixed infinitive form finishes in *-quirir (adquirir)* or belongs to the list: *dormir, errar, morir, oler, erguir* or *desosar*, the form is recognized as an irregular verb and will be conjugated using the irregularity rules which operate on the root vowel, which can be either diphthongized or replaced by another vowel (*adquiero* from *adquirir*, *duermo* and *durmió* from *dormir*).

Module 3: The third transducer identifies whether the infinitive form root ends in a vowel. If the verb belongs to the second or third conjugation (*-er* and *-ir* endings) (*leer, oír*), it is an irregular verb, while if the verb belongs to the first conjugation (*-ar* ending) then it will only be irregular if its root ends with an *-u* or *-i* (*criar, actuar*). For the verbs assigned to the first conjugation, diacritic transduction rules are applied to their inflected forms (*crío* from *criar*, *actúo* from *actuar*); in the case of verbs assigned to the second and third conjugations, the alterations performed on their inflected forms are mainly additions or subtitutions of letters (*leyó* de *leer*, *oigo* de *oír*).

There are some endings such as (*-ier*, *-uer* and *-iir*) which are not found in the MIVC-DB. In the hypothetical case where they are encountered, their conjugation would have followed the rules detailed earlier. Rules facilitating the conjugation of potential but non-existing verbs are included in the algorithm.

Module 4: When an infinitive root form in the first conjugation ends in *-c, -z, -g* or *-gu* (*secar*, *trazar*, *delegar*) and in the second and third conjugation ends in *-c, -g, -gu* or *-qu* (*conocer*, *corregir*, *seguir*), that verb is affected by consonantal orthographic adjustments (irregularity rules) in order to preserve its pronunciation (*sequé* from *secar*, *tracé* from *trazar*, *delegué* from *delegar*, *conozco* from *conocer*, *corrijo* from *corregir*, *sigo* from *seguir*).

In case the infinitive root form of the second and third conjugation ends in *-ñ* or *-ll* (*tañer*, *engullir*), the vowel *i* is removed from some endings of the paradigm following the pattern detailed in (Basterrechea and Rello, 2010).

Verbs undergoing transduction by Module 4 can undergo further modification by Modules 5 and 6. Any infinitive form which failed to meet the trig-

[3]There are 14 so-called Magnificent verbs: *traer, valer, salir, tener, venir, poner, hacer, decir, poder, querer, saber, caber, andar* and *-ducir* (Basterrechea and Rello, 2010).

gering conditions set by Modules 1-4 is also tested against 5 and 6.

Module 5: This module focuses on determining the vowel of the infinitive form root and the verb's derivational base. If the vowel is *e* or *o* in the first conjugation and the verb derivational base includes diphthongs *ie* or *ue* (*helar*, *contar*), or if the vowel is *e* in the infinitive forms belonging to the second and third conjugation (*servir*, *herir*), then the verb is irregular and it is modified by the irregularity rules which perform either a substitution of this vowel (*sirvo* from *servir*) or a diphthongization (*hielo* from *helar*, *cuento* from *contar* or *hiero* from *herir*).

Module 6: Finally, the existence of a diphthong in the infinitive root is examined (*reunir*, *europeizar*). If the infinitive matches the triggering condition for this transducer, its paradigm is considered irregular and the same irregularity rules from module 3 -inserting a written accent in certain inflected forms- are applied (*reúno* from *reunir*, *europeízo* from *europeizar*).

Any verb form that fails to meet the triggering conditions set by any of these six transducers has regular conjugation.

It is assumed that these 6 modules cover the full range of both existing and potential verbs in Spanish. The modules' reliability was tested using the full paradigms of 15,367 verbs. As noted earlier, there are some irregularity rules in module 3 which predict the irregularities of non existing but possible neologisms in Spanish. Those rules, in conjunction with the rest of the modules, cover the recognition and conjugation of the potential new verbs.

5 Evaluation

The transducers have been evaluated over all the verbs from the DRAEVC-DB and the 4,307 new verbs from MICV-DB.

In case a new verb appears which is not similar to the ones contained in our corpus, the transduction rules in Module 3 for non existing but potential verbs in Spanish would be activated, although no examples of that type have been encountered in the test data used here. As this system is part of the free online conjugator *Onoma*, it is constantly being evaluated on the basis of users' input.

Every time a new infinitive form absent from MIVC-DB is introduced by the user[4], it is automatically added to the database. The system is constantly updated since it is revised every time a new irregularity is detected by the algorithm. The goal is to enable future adaptation of the algorithm to newly encountered phenomena within the language. So far, non-normative verbs, invented by the users, such as *arreburbujear, insomniar, pizzicatear* have also been conjugated by *Onoma*.

Verb neologism type	**Verb neologism class**	**Number of neologisms**
regular	regular rules	3,154
irregular	module 1 rules	27
irregular	module 2 rules	9
irregular	module 3 rules	39
irregular	module 4 rules	945
irregular	module 5 rules	87
irregular	module 6 rules	46
Total verb neologisms		**4,307**

Table 2: New verbs evaluation

Of all the new verbs in MIVC-DB, 3,154 were regular and 1,153 irregular (see Table 2). The majority of the irregular neologisms were conjugated by transducer 4.

6 Conclusions

Creativity is a property of human language and the processing of instances of linguistic creativity represents one of the most challenging problems in NLP. Creative processes such as word formation affect Spanish verbs to a large extent: more than 50% of the actual verbs identified in the data set used to build MIVC-DB do not appear in the largest Spanish dictionary. The processing of these neologisms poses the added difficulty of their rich inflectional morphology which can be also irregular. Therefore, the automatic and accurate recognition and generation of new verbal paradigms is a substantial advance in neologism processing in Spanish.

In future work we plan to create other algorithms to treat the rest of the open-class grammatical categories and to identify and generate inflections of new

[4]Forms occurring due to typographical errors are not included.

words not prescribed by dictionaries.

Acknowledgments

We would like to express or gratitude to the Molino de Ideas s.a. engineering team who have successfully implemented the method, specially to Daniel Ayuso de Santos and Alejandro de Pablos López.

References

Ramón Almela Pérez. 1999. *Procedimientos de formación de palabras en español*. Ariel, Barcelona, España.

Eduardo Basterrechea and Luz Rello. 2010. *El verbo en español. Construye tu propio verbo*. Molino de Ideas, Madrid, España.

Eva Esteve Ferrer. 2004. Towards a semantic classification of Spanish verbs based on subcategorisation information. *Proceedings of the ACL 2004 workshop on Student research*, 13.

Pedro Gomis Blanco and Laura Segura. 1998. *Vademécum del verbo español*. SGEL. Sociedad General Española de Librería, Madrid, España.

Mirella Lapata and Chris Brew. 2004. Verb class disambiguation using informative priors. *Computational Linguistics*, 30(1): 45–73.

Francis Mateo. 2008. *Bescherelle. Les verbes espagnols*. Hatier, Paris, France.

Jorge Puebla Ortega. 1995. *Cómo conjugar todos los verbos del español*. Playor, Madrid, España.

Real Academia Española. 2001. *Diccionario de la lengua española*, 22 edición. Espasa, Madrid, España.

David Serrano Dolader. 1999. La derivación verbal y la parasíntesis. *Gramática descriptiva de la lengua española*, I. Bosque, V. Demonte, (eds.), (3): 4683-4756. Real Academia Española / Espasa, Madrid, España.

Evelyne Tzoukermann and Mark Y. Liberman. 1990. A Finite-State Morphological Processor for Spanish. *Proceedings of the 13th conference on Computational linguistics*, (1): 277-282.

Octavio Santana Suárez, José Rafael Pérez Aguiar, Zenón José Hernández Figueroa, Francisco Javier Carreras Riudavets, Gustavo Rodríguez Rodríguez. 1997. FLAVER: Flexionador y lematizador automático de formas verbales. *lingüística española actual XIX*, (2): 229-282. Arco Libros, Madrid, España.

Octavio Santana Suárez, Francisco Javier Carreras Riudavets, Zenón José Hernández Figueroa, José Rafael Pérez Aguiar and Gustavo Rodríguez Rodríguez. 2002. *Manual de la conjugacin del español. 12 790 verbos conjugados.* Arco Libros, Madrid, España.