Building a Spanish Speller

S. Rodríguez, J. Carretero
Universidad Politécnica de Madrid, España

Facultad de Informática
Campus de Montegancedo
28660 Boadilla del Monte, Madrid, Spain
e-mail: srodri@fi.upm.es, jesus@fi.upm.es

Abstract
The lack of spelling tools for the Spanish language in the Unix Operating System, led us to build a speller based on the Unix tool ispell. *The main task of this work was to develop an extensive set of Spanish grammatical rules, which was very complex due to the big number of rules. The speller built has been integrated with* ispell *software and it is being distributed for free since the end of 1994.*

Keywords *spelling, Spanish, formal specification*

1 Introduction

The introduction of computers in text processing has shown the lack of some specialized tools such as spelling checker, grammar checker, etc. This lack can be especially seen in the free software environment and in the Spanish context. For many years we have been using the *ispell* built by Geoff Kuenning for English documents, but we didn't have a Spanish tool to check Spanish documents. As a result, we studied the documentation enclosed to *ispell* and we saw that extensions for other languages were possible by using this tool. We thought to add a Spanish dictionary to this tool and distribute it for free.

The main problem found in the building of this dictionary was to adapt the Spanish grammatical rules to a formal specification. Contrary to English, Spanish is a language evolved from Latin and it has many and complex grammatical rules which made a big effort to formalize the language.

Three main objectives were established for the Spanish speller:

- It should be exhaustive. It had to include most of rules of the Spanish language.
- It should be free. So, we expected to receive a lot of feedback including bugs, non-existent words, ... Moreover it should be easy to maintain and to update with new suggested improvements.
- It should allow the user to modify the configuration to include particularities in the language. This is very important due to the several variations of the Spanish language, especially in America.

The development of the speller was started at the start of 1994 and the main tasks related with this work was to obtain a root words file and to generate the derivation rules. The first beta prototype was ready for internal use by mid 1994 and it is distributed for free since December 1994.

2 Spanish Grammatical Features

As stated above, Spanish is a Latin derived language and its grammar has many features that English does not. In order to construct a Spanish speller we first studied the Spanish grammar to formalize the derivation rules in order to get a complete dictionary from a reduced set of root words. Due to the intrinsic features of the language, many problems were detected building the derivation rules, namely:

Gender and number derivations. The adjectives and substantives have gender (male or female) and number (singular and plural).

- Some of them have its particular female derivation such as *perro* → *perra*. Both derivations have their plural derivations (*perros, perras*).
- Some other have only one gender: masculine such as *álamo*, feminine such as *casa*. These words have their plural derivations (*casas, álamos*).

This is the reason why we have included two set of rules splitting the gender and number derivations, and only number derivations.

Verb conjugation. Spanish verbs are grouped in three kind of conjugations depending on the two ending characters of the infinitive form: -ar, -er, -ir. Moreover each conjugation has more than 30 temporal derivations. In addition, Spanish has two types of verbs:

- Regular verbs. They have a very strict set of derivation rules which is the same for all of them in the conjugation.
- Irregular verbs. These are verbs that have at least one different derivation from the regular verbs. So this set of derivation have to be taken into account. These verbs are grouped in 100 different sets of irregularities ([2]).

Enclitic forms. Some verb derivations are generated by adding a pronominal form at the end of a verbal form. Two different kind of enclitics are found in the written Spanish:

- Pronominal verbs. The enclitic forms are generated by adding the terminations -te, -me, -se at the infinitive and gerund forms: *amar* → *amarte.* A special important way of the enclitic forms are the reflexive verbs: *amar* → *amarse.*
- Transitive verbs. The enclitic forms are generated by adding the terminations -lo, -la, -los, -las, -le, -les: *amar* → *amarle.*

The two rules exposed above can be combined together: *ajustar* → *ajustármelo.* This generates a set of rules of complexity $O(2)$. Moreover the enclitics have to take into account the irregular forms in the gerund such as *vestir* → *vistiéndote*, which increases the complexity to $O(3)$.

Affixes. There are many words that allow modifications by adding prefixes or suffixes. We have found around 20 different prefixes accepted in the Spanish grammar. The main suffixes present in Spanish are derived from nouns and adjectives by adding comparatives (-ísimo), gender, number, and adverbs ending in -mente, and from verbs by adding active participle termination such as *cantar* → *cantante.*

Acute characters. There are many particularities related with gender and number. Some words lose its acute characters and change it by the non-acute one: *gañán, gañanes.*

Taking into account the above features, we developed a set of formal rules comprising an extensive subset of the Spanish grammar.

3 Formal Definition

The formal definition of the derivation rules has been implemented using the ispell tool and following the formal language imposed by it. The set of rules used to specify the Spanish grammar is around 3500 derivation patterns grouped in 57 macrorules, 41 of them are related with prefixes specification and they are not currently working because of the complexity of these rules and the implications they have on the rest of the definitions. For example, some transitive verbs are not longer transitive whether a prefix is applied, and some other words have different meaning if a coherent prefix rule is applied (i. e. *alzar* → *realzar*).

Every macrorule to be described below reflects one particular aspect of the Spanish grammar shown in the previous section.

Gender and number derivations. Two macrorules have been used to implement these features. *Number derivations* include 14 single rules depending on the termination of the word to be applied. Three examples are shown below:

```
[AEIOU] > S         # vaca vacas
Z       > -Z, CES   # arroz arroces
'U N    > -'UN,UNES # at'un atunes
```

Gender and number derivations include 18 single rules. For example:

```
O       > -O, A     # amigo amiga
O       > S         # amigo amiga
O       > -O, AS    # amiga amigas
[^AONS] > A         # pastor pastora
[^AONS] > ES        # pastor pastores
[^AONS] > AS        # pastor pastoras
```

Verb conjugation has been defined using four macrorules describing regular and irregular verbs. Around 200 rules compose the regular verbs derivations and 2700 the irregular ones. Note that irregular verbs are the most consuming effort part in the development of the derivation rules. However we considered very important to formalize them because Spanish has a lot of irregular forms following a set of patterns well defined in its grammar: -ontar → -uento, -oder → -uedo, -ervir → -irvo, etc. Some derivation rules for phonetically regular verbs are:

```
A R        > -AR, O    # amar amo
[^CG] E R  > -ER, O    # comer como
C E R      > -CER, ZO  # vencer venzo
G E R      > -GER, JO  # coger cojo
[^CGU] I R > -IR, O    # vivir vivo
C I R      > -CIR, ZO  # zurcir zurzo
G I R      > -GIR, JO  # fingir finjo
G U I R    > -UIR, O   # extinguir extingo
Q U I R    > -QUIR, CO # delinquir delinco
```

A few examples for irregular verbs are shown below:

```
I A R     > -IAR, 'IO     # enviar env'io
E G A R   > -EGAR, IEGO   # regar riego
O 'N A R  > -O'NAR, UE'NO # so'nar sue'no
E R E R   > -ERER, IERO   # querer quiero
O D E R   > -ODER, UEDO   # poder puedo
S A B E R > -ABER, 'E     # saber s'e
E 'I R    > -E'IR, 'IO    # re'ir r'io
E N I R   > -IR, GO       # venir vengo
U C I R   > -UCIR, UZCO   # lucir luzco
```

Excluded from these rules are *ser, estar, ir,* and *haber* because no way to derive the different forms of these verbs from the infinitive has been found. Instead, every derivation has been explicitly included in the root words file.

Enclitic forms. Regular verbs include around 200 derivation rules and irregular ones around 450. These rules specify the behavior of pronominal, transitive, and combined derivations. All the rules are applied only to infinitive and gerund forms.

```
---------------- REGULAR -------------
[AEI] R > ME            # amar amarme
[AEI] R > TE            # amar amarte
[AEI] R > SE            # amar amarse
[AEI] R > NOS           # amar amarse
[AEI] R > OS            # amar amarse
A R     > -AR, 'ANDOME  # amar am'andome
E R     > -ER, I'ENDOME # comer comi'endose
I R     > -R, 'ENDOME   # vivir vivi'endome

--------------- IRREGULAR -------------
[AEO] E R > -ER, Y'ENDOME # caer cayendo
[AEO] E R > -ER, Y'ENDOTE # caer cayendo
[AEO] E R > -ER, Y'ENDOSE # caer cayendo
[AEO] E R > -ER, Y'ENDONOS# caer cayendo
[AEO] E R > -ER, Y'ENDOOS # caer cayendo
```

Affixes. Suffixes have been grouped in 4 macrorules with 13 rules. Some examples are shown below:

```
A R  > -R, NTE     # amar amante
A R  > -R, NTES    # amar amantes
E    > -E, 'ISIMO  # grande grand'isimo
O    > -O, AMENTE  # loco locamente
```

Prefixes can not be grouped, so each one has been defined as a macrorule, resulting in a set of 20 macrorules for the prefixes most common used.

```
. > SEMI # curado semicurado
```

Acute characters. There are no specific macrorules for these derivations. They have been included in the former rules.

The previously described rules have been applied to a root words file including more than 50000 Spanish words to build a dictionary of around 500000 words.

4 Exploitation

Once the dictionary was ready, we issued a first beta prototype for internal use. The first problem found in this prototype was the different formats used to represent the acute characters ü and ñ, which are not included in the character set managed by the original *ispell*. To solve this problem, the four most useful formats are supported by the rules file. The codification of some special characters for each format is shown below:

- Default format. It is the format used for the development of the dictionary and the format used by Babel.

’a	á
’e	é
’i	í
’o	ó
’u	ú
’n	ñ
”u	ü

- TeX format. It uses the representation of this characters as LaTeXdoes:

\’a	á
\’n	ñ
\”u	ü

- plainTeX format. It uses the representation of this characters as TeX does:

\’{a}	á
\’{n}	ñ
\”{u}	ü

- latin1 format. The acute characters are coded as specified in the iso_8859_1 character set.

Once this problem was solved, the internal prototype was validated and it was ready to be distributed as public domain software, so we contacted Geoff Kuenning to include it in the *ispell* distribution. This software can be obtained by anonymous ftp to `ftp.fi.upm.es` in `pub/unix/espa~nol.tar.gz`. The distribution is composed by the affix file and three word files:

- `espa~nol.words` contains a list of words that appears in the official Spanish dictionary ([3]).

- `espa~nol.comp` contains a list of words not appearing in the official dictionary but being used in computer related texts.

- `antiguas.words` contains a list of words that appears in the official Spanish dictionary although they are old and not currently in use.

Maintenance is performed by collecting bugs and reports about correct words that do not appear in the dictionary. Reports have to be sent to the following Email address: `espanol-bugs@datsi.fi.upm.es`. However, our experience has not been very positive on this respect, because we have received very few feedback messages, even when the users of this tools is increasing.

Some efforts have been made to use this tool in other environments than Unix. We have tried to port it to WordPerfect, but the results were not satisfactory because of the limitations of the dictionary size of this tool. However, there are no limitations related with the speller, because we could port a subset of it successfully.

5 Conclusions and Future Works

A Spanish Speller has been developed and it is being used by a large community. Our feeling is that the dictionary works properly and it is very exhaustive. The error rate is approximately 3 % of the words present in a document. This error rate is mainly due to: prefixes, enclitics, comparatives, and local expressions.

The work can be improved in the following aspects:

- **American Spanish words.** The Spanish Speller uses the official dictionary to validate a word. Anyway there are many valid words that are only used in one area of the Spanish language community specially in Center and South America. Our proposal is to separate the words that are only used in a special area and to generate a word list for this area. In that line, several dictionaries should be set up for countries such as Chile, Perú, Colombia, etc.

- **Scientific words.** A specialized environment such as law, medicine, etc. uses a word set that does not appear in the official dictionary used to validate the words. Anyway, these words have to be considered as correct ones. Therefore, several files can be created to distribute several scientific specialized word lists.

- **Rules Optimization.** The way used to build the affix file (*español.aff*) is not completely optimum. There are many rules that are replicated in several macrorules. The affix file will be reduced for the next version.

Finally, negotiations with the *Real Academia Española de la Lengua* are going on to get an official root words list, because the `espa~nol.words` file is not exhaustive.

References

[1] P. Abrahams and B. Larson. *UNIX for the Impatient.* Addison Wesley, 1992.

[2] Real Academia Espaola de la Lengua. *Esbozo de una Nueva Gramática de la Lengua Española.* Espasa Calpe, 1991.

[3] Real Academia Espaola de la Lengua. *Diccionario de la Lengua Española.* Espasa Calpe, 21 edition, 1992.

[4] D. Dougherty. *sed & awk.* O'Reilly & Associates, 1990.