GRAMPAL: A Morphological Processor for Spanish implemented in Prolog

Antonio Moreno
Dep. de Lingüística
Universidad Autónoma de Madrid
Cantoblanco, Madrid, SPAIN
Phone: +34 1 397.41.09
Fax: +34 1 397.39.30
sandovalo@cc.uam.es

José M. Goñi
Dep. Matemática Aplicada a las TT. II.
Universidad Politécnica de Madrid
Ciudad Universitaria, Madrid, SPAIN
Phone: +34 1 336.72.87
Fax: +34 1 336.72.89
jmg@mat.upm.es

Abstract

A model for the full treatment of Spanish inflection for verbs, nouns and adjectives is presented. This model is based on feature unification and it relies upon a lexicon of allomorphs both for stems and morphemes. Word forms are built by the concatenation of allomorphs by means of special contextual features. We make use of standard Definite Clause Grammars (DCG) included in most Prolog implementations, instead of the typical finite-state approach. This allows us to take advantage of the declarativity and bidirectionality of Logic Programming for NLP.

The most salient feature of this approach is simplicity: A really straightforward rule and lexical components. We have developed a very simple model for complex phenomena.

Declarativity, bidirectionality, consistency and completeness of the model are discussed: all and only correct word forms are analysed or generated, even alternative ones and gaps in paradigms are preserved. A Prolog implementation has been developed for both analysis and generation of Spanish word forms. It consists of only six DCG rules, because our *lexicist* approach —i.e., most information is in the dictionary. Although it is quite efficient, the current implementation could be improved for analysis by using the non logical features of Prolog, especially in word segmentation and dictionary access.

Keywords: Applications of Logic Programming to NLP, Computational Morphology.
1 Introduction

The successful treatment of morphological phenomena in some languages by means of finite state automata appears to have led to the idea that this model is the most efficient and universal way to deal with morphology computationally. Although there exist good finite-state processors for Spanish-like [Marzi, 1986], [Meyer, 1986] or [Tsoukermann & Liberman, 1990], we think that some phenomena can be handled more elegantly using a context-free approach, particularly if the morphological component is to be included as a part of a syntax grammar. Our model has been implemented in standard DCG using a logic programming approach instead of a plain finite-state one.

It is well-known that the so-called non-concatenative processes are the most difficult single problem that morphological processors must deal with. Experience has shown that it is not easy for any approach. Unification-based morphology uses suppletion (i.e. alternative allomorphs for a lemma) and feature description as a general mechanism for handling those processes. Two-level morphology uses instead rules that match lexical representations (lemmas) with surface representations (actual spelling forms). The latter has been claimed to be more elegant, but it is obvious that often the two-level model contains many rules needed for only a very few cases.

The pure two-level/finite-state automata model is not very adequate for treating certain non-concatenative processes, and in such cases one is required to depart from this approach, for example by adding an extension in which two-level rules are retained under the control of feature structures [Trout, 1990]. Moreover, every language has irregularities that can only be treated as suppletive forms, e.g. say (I am), are or was (I was). Since suppletion is needed anyway, and since it is a much simpler approach than rules, we consider that the ‘elegance’ objection is not well-founded.¹

On the other hand, our goal is to generate and recognize all (and only) well-formed inflected forms, and thus we do not accept “missing forms” for defective verbs (see below), but do accept duplicate but correct forms.

2 Major issues in Spanish computational morphology

Spanish morphology is not a trivial subject. As an inflectional language, Spanish shows a great variety of morphological processes, particularly non-concatenative ones. We will try to summarize the outstanding problems which any morphological processor of Spanish has to deal with:

²See the next section for further discussion of the adequacy of the two-level model for Spanish, including defective forms (i.e. null forms in the conjugation) and alternative correct forms.

1. A highly complex verb paradigm forms (see Table 1), excluding the duplicate Imperfect Past forms for compound tenses, their

2. The frequent irregularity of both such as tener (to have), poner etc., have up to 7 different stem forms, una as regular -i.e. only one stem for love.

3. Gaps in some verb paradigms, are missing or simply not used lover, never (no. to go) singular. Others are more first, second and third singular present subjunctive forms, and i verbs, the compound tenses are do usually.

4. Duplicate past participles: a nur correct, like impreso, imprimir treat both.

5. There exist some highly irregula many of their forms directly in 1

6. Nominal inflection can be of two concatinating the gender morp (i.e. without gender morpheme the first class, but nouns and with a different distribution: 49 92% have inherent gender, while Some nouns and adjectives pres for bambu (bamboo), bambu-s an
nal phenomena in some languages by have led to the idea that this model with morphology computationally, issues for Spanish-like (Martí, 1986), 1990)—we think that some phenomena next-free approach, particularly if the as a part of a syntax grammar. Our C using a logic programming approach

catenative processes are the most dif-

f-essors must deal with. Experience has
ification-based morphology uses sup-
is) and feature description as a general-
level morphology uses instead rules
with surface representations (actual
claimed to be more elegant, but it is
many rules needed for a very few
model is not very adequate for treating
such cases one is required to depart
an extension in which two-level rules
ctures (Tront, 1990). Moreover, every
treated as suppletive forms, e.g. see
n is needed anyway, and since it is a
that the "elegance" objection is not
and recognize all (and only) well-
accept "missing forms" for defective
t correct forms.

h computational mor-

As an inflexional language, Spanish
cases, particularly non-catenative
problems which any morphological
degacy of the two-level model for Spanish;

1. A highly complex verb paradigm. For simple tenses, we consider 53 inflected
forms (see Table 1), excluding the archaic Future Subjunctive, but including
the duplicate Imperfect Past Subjunctive (6 forms). If we add the 45 possible
forms for compound tenses, then 98 inflected forms are possible for each verb.

2. The frequent irregularity of both verb stems and endings. Very common verbs,
such as tener (to have), poner (to put), poder (to be able to), hacer (to do),
etc., have up to 7 different stems: hac-er, ha-g-o, hag-o, ha-re, ha-r-e, hiz-o, hiz-
hech-o. This example shows internal vowel modification triggered by different
morphemes having the same external form: hag-o, hiz-o, hech-o. The first
/-o/ is first person singular present indicative morpheme; the second /-er/
is third singular pretérito indicativo morpheme; and the third /-er/ is past
participle morpheme—an irregular one, by the way. As well as these non-
catenative processes, there exist other, very common, kinds of internal
variation, as illustrated by the following example.

\[ e \rightarrow [ie]: \quad \text{quer-er (to want)} \rightarrow \text{quier-o (I want)} \]

2,300 out of 7,600 verbs in our dictionary are classified as irregular, and 5,300
as regular—i.e. only one stem for all the forms as in an-ar—an-o, etc. (to
love).

3. Gaps in some verb paradigms. In the so-called defective verbs some forms
are missing or simply not used. For instance, meteorological verbs such as
llorar, nevar (to rain, to snow), etc., are conjugated only in third person
singular. Other ones are more peculiar, like abolcir (to abolish) that lacks
first, second and third singular and third plural present indicative forms, all
present subjunctive forms, and the second singular imperative form. In other
verbs, the compound tenses are excluded from the paradigm, like in soler (to
do usually).

4. Duplicate past participles: a number of verbs have two alternative forms, both
correct, like impres-o, imprimiento (printed). In such cases, the analysis has to
be treated both.

5. There exist some highly irregular verbs that can be handled only by including
many of their forms directly in the lexicon (like ir (to go), ser (to be), etc.).

6. Nominal inflection can be of two major types: with grammatical gender (i.e.
catenating the gender morpheme to the stem) and with inherent gender
(i.e. without gender morphemes). Most pronouns and quantifiers belong to
the first class, but nouns and adjectives can be in any of the two classes,
with a different distribution: 4% of the nouns have grammatical gender and
92% have inherent gender, while 70% of the adjectives are in the first group.
Some nouns and adjectives present alternative correct forms for plural—e.g.
for bambú (bamboo), bambús and, bambúes.
7. There is a small group (3%) of invariant nouns with the same form for singular and plural, e.g. crísis. On the other hand, 30% of the adjectives present the same form for masculine and feminine, e.g. azul (blue). There exist also singularia tantum, where only the singular form is used, like astrón (stress); and pluralia tantum, where only the plural form is allowed, e.g. matemáticas (mathematics).

8. In contrast with verb morphology, nominal processes do not produce internal change in the stem caused by the addition of a gender or plural suffix, although there can be many allomorphs produced by spelling changes: luz, luc-en (light, lights).

For a detailed description of all verb and nominal phenomena, including a classification into paradigmatic models, see [Moreno, 1991].

All these phenomena suggest that there is no such a universal model (e.g. two-level, unification, or others) for (surface) morphology. Instead, we have approaches more suited for some processes than others. The computational morphologist must decide which is more appropriate for a particular language. We support the idea that unification and feature-based morphology is more adequate for languages, such as Spanish and other Latin languages, that have alternative stems triggered by specific suffixes, missing forms in the paradigm, and duplicate correct forms.

3 The model

It is well known that morphological processes are divided into two types: processes related to the phonological and/or graphic form (morpho-phonemics), and processes related to the combination of morphemes (morpho-syntax). Each model treats these facts from its particular perspective. Two-level morphology uses phonological rules and combination classes (in the lexical component). Mixed systems such as [Bear, 1986] or [Ritchie et. al., 1987] have different sets of rules.

As we stated before, our model relies on a context-free feature-based grammar, that is particularly well suited for the morpho-syntactic processes. For morpho-phonemics, our model depends on the storage or computation of all the possible allomorphs both for stems and endings. This feature permits that both analysis and synthesis be limited to morpheme concatenation, as the general and unique-mechanism. This simplifies dramatically the rule component.

We present some examples of dictionary entries: two verbal ending entries (allomorphs) for the past participle morphemes and two allomorph stems for imprimir, compatible with those endings.

\[\text{vm(no,part,nofin, [2,3], 99, reg)} \rightarrow [ido].\]

\[\text{vm(no,part,nofin, [2,3], 99, part1)} \rightarrow [o].\]
nouns with the same form for singular
hand, 30% of the adjectives present
e.g. azul (blue). There exist also
lar form is used, like entrés (stress);
form is allowed, e.g. matemáticas.
nal processes do not produce internal
of a gender or plural suffix, although
by spelling changes: luz, luc-as
nominal phenomena, including a clas-
no, 1991],
no such a universal model (e.g. two-
ology. Instead, we have approaches
computational morphologist must
language. We support the idea that
more adequate for languages, such as
ative stems triggered by specific
uplicate correct forms.
uses are divided into two types: pro-
form (morpho-raphemics), and
omorphs (morpho-syntactic). Each model
e. Two-level morphology uses phono-
cial component). Mixed systems such
different sets of rules.
context-free feature-based grammar,
syntactic processes. For morp-
or computation—of all the possible
feature permits that both analysis
ation, as the general and unique
le component.
ies: two verbal ending entries (allo-
d two allomorph stems for imprimir,

\[
\begin{align*}
\text{vl} & : \text{imprimir}, v, 3, [100], [reg] \rightarrow \text{imprimir}, \\
\text{vl} & : \text{imprimir}, v, 3, [98], [parti] \rightarrow \text{impress}.
\end{align*}
\]

Where \( v \) and \( l \) stands for the values of the “morphological category” that we are using to drive the DCG rule invocation. All the dictionary entries are coded with a predicate that corresponds to its morphological category. The full inventory of such categories follows:

\[ w \text{ For complete inflected word forms.} \]
\[ vl \text{ For words (nouns and adjectives) that can accept a number morpheme.} \]
\[ nl \text{ For nominal lexemes (stems).} \]
\[ vm \text{ For nominal gender morphemes.} \]
\[ ng \text{ For nominal gender morphemes.} \]
\[ nn \text{ For nominal number morphemes.} \]

For reference, and to check the meaning of the examples, a short self-description of the arguments of those predicates follows:

\[ w(\text{Lemma, Category, Pers_Num, Tense_Mood}). \]
\[ w(\text{Lemma, Category, Gender, Number}). \]
\[ vl(\text{Lemma, Category, Number_Type_List, Gender, Number}). \]
\[ vl(\text{Lemma, Category, Conjugation, Stem_Type_List}. \\
\text{Suffix_Type_List}). \]
\[ vm(\text{Pers_Num, Tense_Mood, Finiteness, Conjugation_List}. \\
\text{Stem_Type, Suffix_Type}). \]
\[ nl(\text{Lemma, Category, Gender_Type_List, Number_Type_List}. \\
\text{Gender, Number}). \]
\[ ng(\text{Gender_Type, Gender, Number}). \]
\[ nn(\text{Number_Type, Number}). \]

We have introduced some contextual atomic features that impose restrictions on
the concatenation of morphemes through standard unification rules. Such features
are never percolated up to the parent node of a rule. Multi-valued atomic features
are permitted in the unification mechanism, being interpreted as a disjunction
of atomic values. We represent this disjunction as a Prolog list. Disjunction of values
is used only for contextual features (stem_type, suffix_type, conjugation, gender_type

\[ \text{[ido].} \]
\[ \rightarrow \text{[o].} \]
and number type) just to improve storage efficiency, since this device is actually not needed if different entries were encoded in the lexicon.

In the conjugation table (Table 1), the stem type values of the grammatical features person-number and tense-mood are displayed in boldface. For example, sing.1 means first person, singular number; while pres.ind means present tense, indicative mood.

<table>
<thead>
<tr>
<th></th>
<th>sing.1</th>
<th>sing.2</th>
<th>sing.3</th>
<th>plu.1</th>
<th>plu.2</th>
<th>plu.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>pres.ind</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>impf.ind</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
<td>25</td>
<td>26</td>
</tr>
<tr>
<td>indf.ind</td>
<td>31</td>
<td>32</td>
<td>33</td>
<td>34</td>
<td>35</td>
<td>36</td>
</tr>
<tr>
<td>fut.ind</td>
<td>41</td>
<td>42</td>
<td>43</td>
<td>44</td>
<td>45</td>
<td>46</td>
</tr>
<tr>
<td>pres.subj</td>
<td>51</td>
<td>52</td>
<td>53</td>
<td>54</td>
<td>55</td>
<td>56</td>
</tr>
<tr>
<td>impf.subj</td>
<td>61</td>
<td>62</td>
<td>63</td>
<td>64</td>
<td>65</td>
<td>66</td>
</tr>
<tr>
<td>cond</td>
<td>71</td>
<td>72</td>
<td>73</td>
<td>74</td>
<td>75</td>
<td>76</td>
</tr>
<tr>
<td>imper</td>
<td>82</td>
<td>83</td>
<td>84</td>
<td>85</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>inf</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>90</td>
</tr>
<tr>
<td>ger</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>99</td>
</tr>
<tr>
<td>part</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Conjugation table.

Each of the 49 inflected forms is represented by a numeric code, and the additional value 100 is used as a shorthand for the disjunction of all of them (used for regular verbs; see the entry for imprin above). The contextual feature stem type (att) is used to identify the verb stem and ending corresponding to each form, and the contextual feature suffix type (att) distinguishes among several allomorphs of the inflectional morpheme by means of a set of values:

- reg
- pres
- pret1
- pret2
- fut.cond
- imp.subj
- imper
- infin
- ger
- part1
- part2

Since this value set is much smaller than the stem type set, we have chosen an alphabetic code. With the combination of both features, and the addition of a third feature conj (conjugation), we can state unambiguously which is the correct sequence of stem and ending for each case (see examples above, where imprin only matches 100 for all features, and imprres matches 0, thus preventing ill-formed concatenations—for these morphemes—such as imprino or imprres-ido).

In the same fashion, we have two inflection, nut (number-type) and gnomorph for the plural and gender in nominal stems. The following examplary morphemes are in nominal 1.

/* NOUN AND ADJECTIVE */

- zg(mas1, masc, sing) $\rightarrow$ [s]
- zg(mas2, masc, sing) $\rightarrow$ [s]
- zg(fem, fem, sing) $\rightarrow$ [es]
- zn(plu1, plu) $\rightarrow$ [es]
- zn(plu2, plu) $\rightarrow$ [es]

/* SOME ENTRIES FOR NOUN */

- nl(presidente, n, [mas2, fe], [plu2], masc)
- nl(doctor, n, [fem], [no], [plu2], masc)
- nl(bambul, n, [plu1, plu2], masc)

These entries allow the analysis, presidenta, presidente and pres doctora, doctores and doctores f for bambú.

The grammatical features (categ) gender are the only features that are can be used by a syntactic DCG gra.

A unification-based system relies a robust and large dictionary, prone on the accessibility of all possible a. Fortunately, there is no need for typi impractical, time consuming and error is quite regular and we have formalized the allomorphs of any verb from the.

The formalized description of the mentioned in [Moreno, 1991], where some formalizations are made: Paradigms for the inflectional behaviour of the Spa.

---

2The codes 93 and 86 stand for the courtesy imperative: imprin usted, imprin ustedes. These word forms are the same as the 53 and 55 ones.

3Actually, this number encodes a particular combination of person, number, tense and mood features.
efficiency, since this device is actually not the lexicon.
ne stem_type values of the grammatical
are displayed in boldface. For example, n:
while pres.ind means present tense.

<table>
<thead>
<tr>
<th>g.3</th>
<th>plu.1</th>
<th>plu.2</th>
<th>plu.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>14</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>25</td>
<td>26</td>
</tr>
<tr>
<td>3</td>
<td>34</td>
<td>35</td>
<td>36</td>
</tr>
<tr>
<td>3</td>
<td>44</td>
<td>45</td>
<td>46</td>
</tr>
<tr>
<td>3</td>
<td>54</td>
<td>55</td>
<td>56</td>
</tr>
<tr>
<td>3</td>
<td>64</td>
<td>65</td>
<td>66</td>
</tr>
<tr>
<td>3</td>
<td>74</td>
<td>75</td>
<td>76</td>
</tr>
<tr>
<td>3</td>
<td>85</td>
<td>86</td>
<td></td>
</tr>
</tbody>
</table>

| 60  |
| 90  |
| 99  |

the disjunction of all of them (used for we). The contextual feature stem_type
ending corresponding to each form, and
linguistic among several allomorphs of
of values:

pret1 pret2
imper infin
part2

the stem_type set. We have chosen an
these features, and the addition of a third
which is the correct sequence
ies above, where imprim only matches
preventing ill-formed concatenations
imprim-ido.

creative: imprimas ustedes.

In the same fashion, we have two special contextual features for the nominal
infection, num (number_type) and gen (gender_type), to identify the various allo-
morphs for the plural and gender morphemes, and associate them with the proper
nominal stems. The following examples show those contextual features both in
nominal morphemes and in nominal lexeme entries:

/* NOUN AND ADJECTIVE MORPHEMES */
ng(masi,masc,sing) --> [0].
ng(mas2,masc,sing) --> [s].
ng(fem,fem,sing) --> [a].

zn(plu1,plu) --> [s].
zn(plu2,plu) --> [es].

/* SOME ENTRIES FOR NOUNS */
zl(presidente, n, [mas2,fem], [plu1], .., .., ..) --> [presidente].
w1(doctor, n, [plu2], masc, sing) --> [doctor].
z1(doctor, n, [fem], [mas], masc, sing) --> [doctor].
w1(bambú, n, [plu1, plu2], masc, sing) --> [bambú].

Those entries allow the analysis/generation of the word forms presidente,
presidenta, presidentes, and presidentas for the lemma presidente:doctor,
doctora, doctores and doctoras for doctor; and bambú, bambús and bambús for
bambú.

The grammatical features (category, lemma, tense, mood, person, number and
gender are the only features that are delivered to the w node, and from this level
can be used by a syntactic DCG grammar.

A unification-based system relies very much on the lexical side. It is needed
a robust and large dictionary, properly coded. Additionally, our model depends
on the accessibility of all possible allomorphs, so their storage is also necessary.
Fortunately, there is no need for typing all of them by hand, since this would be an
impractical, time-consuming and error-prone task. Morpho-phonemics for Spanish
is quite regular and we have formalized and implemented the automatic computation
of the allomorphs of any verb from the infinitive form.

The formalized description of the morphological phenomena of Spanish was
presented in [Moreno, 1991], where some interesting and well founded linguistic
generalizations are made: Paradigms for verbs are described to capture regularities in
the inflectional behaviour of the Spanish verbs, and the same is done with nouns.

These are not the traditional ones, since they capture the problems arising in written language,
such as diacritical marks, different surface letters for the same phoneme, etc.
Nominal inflection is an important component in the inflectional morphemes (gender and number rules) shown to handle this. It is necessary to have a gender suffix as well as a plural suffix to determine the gender and number of the singular noun and further concatenation:

```plaintext
/\* NOUN AND ADJECTIVE INFLECTION RULES \*/

w(Lex, Cat, [plu], Gen, Num) --> nl(Lex, Cat, Gen)

member(Cat, 

w(Lex, Cat, Gen, Num) --> nl(Lex, Cat, Nut)

member(Nut, 

w(Lex, Cat, Gen, Num) --> nl(Lex, Cat, Nut)

Nut = plu2,

member(Nut, 

w(Lex, Cat, Gen, Num) --> nl(Lex, Cat, Gen)

The predicate member included in the implementation of the disjunction in atomic context is associated with a different encoding of the

5 The Processor

The grammar rules are stated using implementations, thus we have used the

Nominal inflection is a bit more complicated, because of the combination of two
inflectional morphemes (gender and number) in some cases. Our model needs
the 4 rules shown to handle this. The first one is for singular words, when the stem
has to be concatenated to a gender suffix (niñ-o, niñ-a); the second is for plural
words, where an additional number suffix is added (niñ-o-s); the third builds plurals
from an allomorph stem and a plural morpheme (león / león-es); and the fourth
rule validates as words the singular forms (wl) obtained from the first rule without
further concatenation:

```/*
/* NOUN AND ADJECTIVE INFLECTION RULES */
/**

wl(Lex, Cat, [plu1], Gen, Num)
    --> ni(Lex, Cat, GetL, _, _, ), ng(Get, Gen, Num),
        { member(Get, GetL) }

w(Lex, Cat, Gen, Num) -->
    wl(Lex, Cat, NutL, Gen, _), nn(Nut, Num),
    { member(Nut, NutL) }

w(Lex, Cat, Gen, Num) -->
    ml(Lex, Cat, _, NutL, Gen, _), nn(Nut, Num),
    { Nut = plu2, member(Nut, NutL) }

w(Lex, Cat, Gen, Num) -->
    wl(Lex, Cat, _, Gen, Num).
```

The predicate member included in the procedural part of the DCG rule imple-
ments disjunction in atomic contextual features, although it could have been elimi-
nated with a different encoding of the lexical entries.

5 The Processor

The grammar rules are stated using the DCG formalism included in most Prolog
implementations, thus we have used the DCG interpreter both for parsing and gen-
erating word forms. Since the interpreter is supplied with morphemes included in
the dictionary for its proper operation, a segmenter has to be included to provide
the parser with candidate word segmentations. This is achieved by means of a non-
deterministic predicate that finds all the possible segmentations of a word. This is
one of the efficiency drawbacks of the current implementation of GRAMPAL.

To avoid such inefficiency the system could be augmented with a letter trie index
or trie- [Aho et al., 1983] to the lexical entries. With this device, segmentation
will be no longer non-deterministically blind and the search would be efficiently
guided. Generation does not have those efficiency problems, and the system is
bidirectional without any change in the rules.

6 Conclusions

A Prolog prototype, GRAMPAL, was developed to intensively test the model,
both as analyzer and as generator. This processor implemented in Prolog has shown
that logic programming can be used successfully to handle the Spanish inflectional
morphology. We have also implemented a C version of GRAMPAL, but it needs
separate components for analysis and generation, due to the lack of reversibility
that Logic Programming has provided us with.

The model presented is based on two basic principles:

- Empirical rigour: all and only correct forms are analysed and generated,
  whether regular or not; gaps in verb paradigms are observed; suppletive forms
  are considered valid, and so on. It is important to stress that GRAMPAL does
  not overgenerate or overanalyse.

- Simplicity and generalization: GRAMPAL employs a really straightforward
  rule component, that captures the logical generalization of the combination
  of a stem and an ending to form a inflected word. "Standard scientific con-
  siderations such as simplicity and generality apply to grammars in much the
  same way as they do to any other theories about natural phenomena. Other
  things being equal, a grammar with seven rules is to be preferred to one with
  63 rules" [Gazdar and Mellish, 1989].

The current dictionary has a considerable size: 43,000 lemma entries\(^4\), including
24,400 nouns, 7,600 verbs, and 11,000 adjectives. The model could be used for
derivative morphology and compounds as well, but this has not been done yet, since
further linguistic analysis must be done to specify the features needed to permit
derivatives and compounds.

\(^4\)In these figures are neither included closed categories, nor allomorphs for verb and nominal
morphemes.

References

Data Structures and Algorithms, Addison Wes

Linguistics. Dallas, TX.

the International Conference on Computational Linguistics (COLING 86).


Generación de la Información. Universidad Autónoma de Madrid.

Procesamiento del

Universidad Autónoma de Madrid.

[Ritchie et al., 1987] Ritchie, G.: Ful
Computational Framework for Lex

[Tsoukermann & Liberman, 1990] To
Finite State Morphological Process
n Congres de la Generación de la Información.

ence on Computational Linguistics
References


