A Paradigm-and Feature-Based Computational Morphology for Spanish[1]

ANTONIO MORENO SANDOVAL
Universidad Autónoma de Madrid

1.Computational treatment of Spanish morphology: state of the art

Computational Morphology can be divided into three types:

-Finite-state morphology and Two-Level morphology (Koskenniemi 1983).

-Concatenative morphology, with phrase structure rules and stem allomorphs in the lexicon.

-Paradigmatic morphology, with templates.

1.1. TWO-LEVEL MORPHOLOGY

Two-level Morphology, theoretically based on IP (Item-and-Process), is the most widespread descriptive model, especially if we take into account that it was adopted by Generative Morphology. In this approach, the relationships between words are seen as derivative processes: the item *took* /tθk/ is derived from the item *take* /teik/ by a process involving a vowel change /ei / → /θ/. Along the same line, Koskenniemi(1983) developed Two-level Morphology, which has a Lexical Representation Level (for abstract units or morphemes as *take*) and a Surface Representation Level (for actual realizations or allomorphs, such as *took*). Finite-state transducers are used to implement the rules for morphological alternations during the derivation process between those levels. Possible combinations of stems and affixes (*continuation classes* in Koskenniemi's terminology) are also encoded as a finite-state network.

This two-stage model is based on *The Sound Pattern of English* (Chomsky & Halle 1968), a book written in the early days of Transformational Grammar when phonology held a more important role than it does now. Koskenniemi's major contribution was to replace the very powerful ordered rules with automaton transducers, whose formal properties are much less powerful, but well-understood. The result was an efficient model for the description of morphophonological changes in the final surface phonetic representations.

The appeal of the Two-Level Model is due to its successful implementation in different languages, especially in those with a long string of morphemes and a high number of inflected forms, e.g. Finnish, Turkish or Basque. In general, this computational model has shown its efficiency when applied to agglutinative languages with a many morphophonological changes in the morpheme boundaries. However, as Sproat (1992) remarks:

> "Finnish morphology is purely concatenative, and this is the simplest kind of morphological combination [...] the length of words should not be taken as a measure of the complexity of the morphology."

The principal limitation of this model is when it comes to nonconcatenative morphology, of which Semitic languages are undoubtedly the best example. In certain environments, morphological information is not presented in a linear form (i.e. suffixation or preffixation), but rather discontinuously. In this type of environment, in order to describe a given morphological process, one has to take into account information not directly available in the most immediate context, i.e. the previous morpheme. Examples of nonconcatenative phenomena are:

- infixation, as in Tagalog verbs
- template morphology, such as can be found in Semitic verb paradigms or the broken plurals of Arabic nouns
- reduplication, as in Warlpiri or Tagalog
- circumfixation, as in German participles in *ge-t* and *ge-en*

To solve this type of problem, some researchers have proposed using descriptions containing several intermediate levels, with lexical transducers in cascade, or implementing several finite state machines (see Sproat 1992 for a critical review). By substantially modifying the original Two-Level Model, such systems are able to describe the above phenomena[3]. As a result, the Two-Level Model is generally considered to be the universal model for Computational Morphology[4], and in a parallel way, IP is regarded as the universal descriptive model. Neither from a computational perspective nor from a8 theoretical one, can we agree with this view for the simple reason that a successful universal model does not as yet exist. Our point is that, at least for certain types of languages, the finite state-IP approach is not the most appropriate in terms of elegance and simplicity. Of course, this does not imply that that the paradigmatic model is universal.

Given that it is the most popular and widely-used computational approach for morphology, it is only natural that Two-Level Morphology has been applied to Spanish in different implementations (e.g. Marti 1986; Meya 1987; Tzoukerman and Liberman 1990[5]; Badia *et al* 1996; Carulla and Oosterhoff 1996).

1.2. **The concatenative model (IA)**

The Item-and-Arrangement (IA) model assumes that words are linear sequences (*arrangements*) of morphs (*items*). In Word Grammar, there are morph concatenation rules, and a lexicon, which is a list of morphs (allomorphs where the morpheme has more than one realization). The most complex task that morphological rules have to perform is to set the conditions for a correct string of morphs where there is more than one possible candidate. In other words, where co-occurrence is possible, distributional statements are needed for complementary allomorphs.

The Spanish verb *poder* (to be able) has two stem allomorphs: /pod/ (as in *pod-er*) and /pued/ (as in *pued-o*). In order to generate the proper infinitive form, the system has to decide between the two options for the stem, and in this case only /pod/ is the correct one, since the concatenation of /pued/ and /er/ (the infinitive suffix) is ill-formed. In contrast, in the first-person singular present indicative, /pued/ must be selected.

Moreno Sandoval (1991) gives a full description of the Spanish inflection within a unification-and-feature approach. The most interesting characteristic of our implementation is the fact that the rule-component is reduced to only seven rules. One verb formation rule generates every wordform in the paradigm, including irregular ones. This rule states that an inflected verb is composed of a verb stem and a suffix, and special features are used to check each combination. Highly irregular verbs, such as *ser*, *ir* or *estar*, with many suppletive forms, are directly included in the dictionary as full inflected forms (in the same way as in Two-Level Morphology). This strategy is extremely effective because the reduced number of rules avoids the occurrence of complex interactions between them (as is the case of certain Two-Level rules).[6]

GRAMPAL, a prototype in Prolog, was designed as a demonstrator. An enhanced and tested version of this implementation is presented in Moreno & Goñi (1995), where more than 40,000 lexical entries are handled. This model has also been used for the morphosyntactic knowledge of the lexical platform ARIES (Goñi, González and Moreno (in press)).

Its main drawback is the size of the dictionary and the redundancy in the lexical entries, in comparison with dictionaries in Two-Level Morphology. This is a consequence of the lexicalist approach adopted, in which the syntactic component is reduced to the minimum. In contrast, it is a considerable improvement over considerable redundancy of the messy Two-Level rule component[7].

In our view, both Item-and-Arrangements implementations and Two-Level Morphology are suitable for morpheme concatenation[8]. The basic difference is in regards to the rule component: which is syntactically oriented in the first model and phonologically oriented in the second one. It seems to be the case that certain languages have a greater proportion of phonologically-conditioned allomorphy, and others have more grammatically-conditioned allomorphy. In the first case, a phonologically oriented approach, such as Two-Level Morphology, seems to be more appropriate, whereas in the second case, a more syntactic approach, such as Unification and Feature Morphology, is more suitable. In both cases, they are reductionist approaches to Morphology, since they do not recognize the following:

-Morphology can be an independent domain with its own structure and organizational principles (the centrality of the notion of paradigm).

-Most languages, if not but all, have a mixture of different morphological processes (specially concatenative and non-concatenative, but also suppletion, defectiveness, and periphrastic formation). Consequently, focusing only on one aspect of the problem can lead to the unsatisfactory treatment of other areas.

The second question is addressed in some interesting implementations by means of a combination of phonological and syntactic rules (Bear 1986; Trost 1990). Still, in our opinion, more elegant results can be obtained by using an autonomous morphological approach, as some theoretical morphologists have suggested. In the next section, we shall examine certain proposals from the paradigm and feature perspective.

2. Introduction to the descriptive framework

The first part of this section describes the paradigmatic model and the second deals with features and inheritance.

2.1 Paradigmatic morphology

Our proposal is based on the WP model for morphological description. This model has been progressively developed by different scholars over the last three decades (i.e. Robins 1959; Matthews 1972;

Wurzel 1984, 1989; Van Marle 1985; Carstairs 1987; Stump 1991; Anderson 1992). The starting point for our approach is in the work of Carstairs (1987) and Wurzel(1984, 1989), both of whom have formulated extremely productive theories based on the concept of inflectional paradigm. These claim that whenever different morphological information is conveyed by the same morph (the characteristic *syncretism* of inflecting or fusional languages), it is better to use a paradigm rather than a morpheme concatenation (IA) or the application of processes to an underlying form (IP). This assumption is generally accepted in theoretical morphology (Spencer 1991: 50), since it seems to be the simplest approach to the description of morphological syncretism. This is so, despite the fact that it renders the model less universal (Beard 1995: 11), and makes its expressiveness too powerful, thus allowing non-existing models to be described (Spencer 1991: 52; Bauer 1988). Notwithstanding, we are willing to accept these theoretical limitations for the sake of achieving a better characterization of Spanish morphology.

2.1.1. Macroparadigm, subparadigm and the Paradigm Economy Principle

Following Carstairs (1987) we define *an inflectional paradigm* as a set of wordforms or inflectional realizations for a given part of speech. A *macroparadigm* consists of either two or more similar paradigms, which have different phonologically-conditioned allomorphs, or any paradigm which cannot be combined with another. Sometimes several realizations in a given paradigm make use of the same signifier. This is called *inflectional homonymy* within the paradigm, or "paradigmatic many-to-one relationships between morphosyntactic properties and their exponents"(Carstairs 1987:87). We can thus say that a subparadigm is obtained when two or more "slots" in the paradigm have the same exponent. A case in point is the past imperfect indicative in Spanish:

CLASS -AR	CLASS -ER	CLASS -IR

	SG	PL	SG	PL	SG	PL
1st person	am*aba*	amábamos	tem*ía*	tem*íamos*	part*ía*	part*íamos*
2nd person	amabas	amábais	temías	tem*íais*	part*ías*	part*íais*
3rd person	am*aba*	amaban	tem*ía*	tem*ían*	part*ía*	part*ían*

From this data, two macroparadigms are obtained: Class AR and Class ER-IR (taking into account that this example is a subset of the whole verb paradigm of 55 wordforms, and that Class ER and Class IR differ in other parts of the paradigm). The cases of paradigmatic homonymy define several subparadigms (in boldface in the table):

-The 1st and 3rd person singular in Class AR: /-aba/
-The 1st and 3rd person singular persons in Class ER and Class IR: /-ía/
-The 2nd person singular in Class ER and Class IR: /-ías/
-The 1st person plural in Class ER and Class IR: /-íamos/
-The 2nd person plural in Class ER and Class IR: /-íais/
-The 3rd person plural in Class ER and Class IR: /-ían/

The Paradigm Economy Principle (Carstairs 1987) offers a possible explanation for this redundancy: the inflectional resources of a given word-class must be organized into as few paradigms as possible in order to reduce memory load in the learning of paradigms. These notions will be applied to the verb inflection of Spanish in section 3.

2.2. Features and inheritance: DATR

DATR is a formal language specifically designed to represent the lexical knowledge in feature and unification-based linguistic theories. These theories, which have recently been developed within generative grammars, are extremely interesting for the insights they provide. In the last 15 years they have enjoyed considerable success in the field of Theoretical Linguistics as well as in Computational Linguistics.

A *feature* is an association of an attribute with its value. Such features are the principal carriers of information. Almost every unit, linguistic or not, can be decomposed into features. Morphemes can also be described as feature structures, and this fact is especially useful in describing syncretism, where theories based on the morphemic unit have problems determining how the same signifier can bear more than one signified[9].

When Gazdar and Evans first began to design DATR (in the late 80's), they were looking for a formalization that had sufficient expressive power to encode complex lexical entries (using feature structures), and which were also capable of expressing generalizations about the implicit information contained in such entries (Gazdar 1989; Evans and Gazdar 1989, 1996).

Cahill and Gazdar (1997) present their theory of German inflection in DATR. We will not follow that implementation because, although it is concerned with inflectional morphology, and its description is stated in terms of phonological concepts, such as the syllable and the segment. Cahill and Gazdar (1997: 214) state:

> "DATR itself is no more than a very general LANGUAGE for lexical description and therefore does not commit or restrict the linguist using it to any particular linguistic framework [...] and [DATR] is perhaps best thought of as a programming languagethat can be used to implement and test linguistic theories."

Using DATR as a formal framework, Corbett and Fraser (1993, 1995) have developed a morphological theory called *Network Morphology*. The core idea is that *"irregular paradigms are basically the same as regular ones except that "they deviate in one or two characteristics"*. For this reason, they prefer the term "subregularity" instead of "irregularity". Of course, there are always some paradigmatic exceptions that can only be dealt with by means of suppletion (*voy, fui, ir*). To capture this generalization, they use *inheritance* and *overwriting*, two concepts first developed in the field of Artificial Intelligence (Knowledge Representation).

In this method, a hierarchical classification of the lexical paradigm to be represented is first elaborated with the information organized in a network. Each node is a compendium of related information represented by features. In the context of a lexical description, a node can correspond to an inflected form, a lexeme, or a lexeme type. Nodes in lower positions in the hierarchy inherit the default features from higher nodes except when the lower node already has a value assigned to a particular feature. In that case, the default value is overwritten. For instance, in a classification of the concept BIRD, the positive features "it has feathers" and "it can fly" are inherited both by SEAGULL and by EAGLE, whereas PENGUIN will bring the value "no" overwriting the default values inherited from BIRD (example taken from Fraser and Corbett 1995):

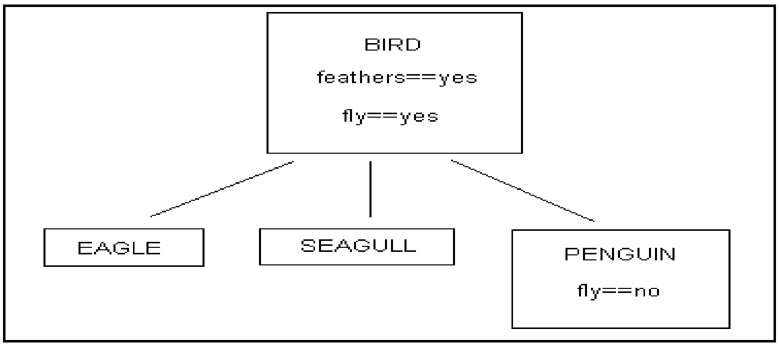

Figure 1: Inheritance network

In this type of lexical organization, generalizations only have to be expressed once at the upper level, since the lower dependent levels automatically acquire it through inheritance. This type of representation has two main advantages: Firstly, the same basic procedure is used for coding both regularities and subregularities as well as exceptions. Secondly, exceptions are explicitly marked as such.

Fraser and Corbett (1995) make use of such an inheritance network for the Russian nominal declension. Figure 2 shows a macroparadigm, NOMINAL, from which the macroparadigms ADJECTIVE and NOUN can inherit. Eventually, nouns are organized in four declensions, from N_I to N_IV:

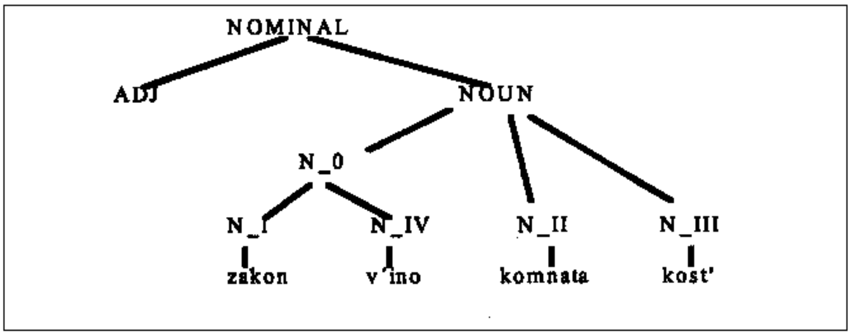

Figure 2: Inheritance network for the Russian noun declension (Fraser and Corbett 1995)

Fraser and Corbett provide an elegant account of gender, animacy and declensional assignment in Russian, with a considerable simplification of lexical entries. We have also used the idea of a default inheritance to express paradigms in fusional languages. Network Morphology is particularly well suited for describing relationships between different levels of linguistic information (prosody, phonology, morphology, syntax), and offers a sophisticated account of paradigms (Brown *et al* 1996).

3. A feature and paradigm account of the Spanish morphology

3.1. Inflectional morphology

Our description of verb inflection in Spanish is based on the paradigms proposed in Moreno (1991). It shows that theoretical concepts such as macroparadigm or subparadigm can be expressed in DATR, providing an elegant account of Spanish morphology.

3.1.1. Verb paradigms

The full paradigm consists of 55 verb forms (which do not include the obsolete future subjunctive):

-present indicative (6 forms)

-imperfect indicative (6 forms)
-past indicative (6 forms)
-future indicative (6 forms)
-present subjunctive (6 forms)
-imperfect (past) subjunctive (6 forms in *-ra*, 6 forms in *-se*)
-conditional (6 forms)
-imperative (4 forms: second and third persons)
-non-finite forms (3, infinitive, participle, gerund)

There are three classes according to the theme vowel which appears with the infinitive stem: Class AR (C1), Class ER(C2), Class IR(C3). Traditional descriptions distinguish between REGULAR and IRREGULAR paradigms. Most Spanish verbs follow a single paradigm of stem plus suffix, in which several cases of homonymy occur:

-HOMONYMY 1: The 1st person singular present indicative suffix is the same for the three classes: /-o/

-HOMONYMY 2: The suffixes for the 2nd and 3rd person singular, and 3 person plural of the present indicative suffixes are the same for C2 and C3.
-HOMONYMY 3: All the wordforms for the imperfect indicative, past indicative, present subjunctive, and imperfect subjunctive are the same for C2 and C3.

-HOMONYMY 4: The 2nd person singular imperative is the same for C2 and C3.

-HOMONYMY 5: The gerund (imperfect participle) and (perfect) participle forms are the same for C2 and C3.

-HOMONYMY 6: The polite imperatives (3rd person singular and plural) are the same forms as the corresponding 3rd person present subjunctive in all three classes.

These cases of paradigmatic homonymy can be formalized in an inheritance network, as shown in Figure 3.

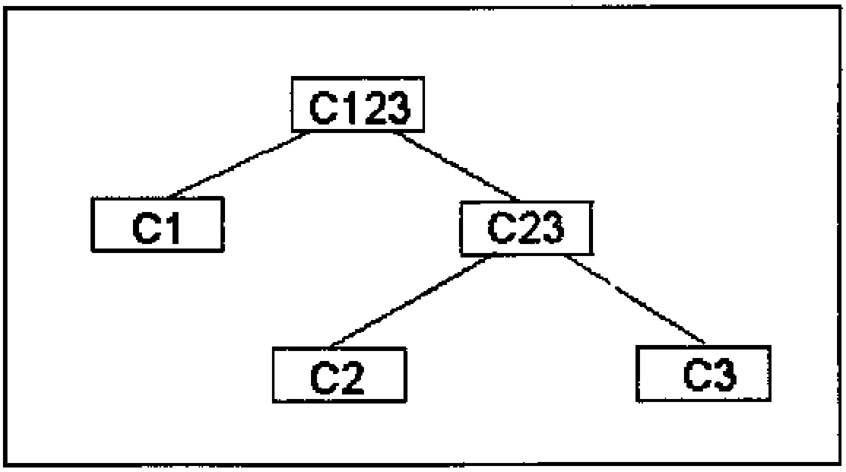

Figure 3: Spanish verb conjugation network

The above schema shows that there is a general macroparadigm (C123), which reflects Homonymies 1 and 6, as well as macroparadigm C23, which captures Homonymies 2 to 5. Finally, regular paradigms C1, C2, C3. C1 inherit directly from C123, while C2 and C3 inherit from C23, and indirectly from C123. All these facts can be expressed in DATR[10]:

```
C123:
        <suf  reg ind pres 1 sing>    == o.        (Homonymy 1)
        <suf imper pol sing>          == <suf subj pres 3 sing>  (Hom.
                                      6)
        <suf imper pol plu>           == <suf subj pres 3 plu>  (Hom.
                                      6)

C23:

        <suf>                         == C123: <suf>
        <suf reg ind pres 2 sing>             == es        (Homonymy 2)
        <suf reg ind pres 3 sing>             == e (Homonymy 2)
        <suf reg ind pres 3 plu>              == en        (Homonymy 2)
        <suf reg ind impf 1 sing>     == ía        (Homonymy 3)
```

```
<suf reg ind impf 2 sing>     == ías       (Homonymy 3)
<suf reg ind impf 3 sing>     == < suf reg ind impf 1 sing>
...
<suf reg imper 2 sing>           == e(Homonymy 4)
<suf reg ger>                 == iendo(Homonymy 5)
<suf reg part>                == ido       (Homonymy 5)
```

Each statement is called a *path*, containing one or more *attributes* (‘< >‘) and a *value* associated with that path. The attributes are atoms representing linguistically relevant features. There are no formal constraints on the number and the class of attributes that may be used in a path[11], but order of attributes is significant. Therefore, the first statement in C123 means: the regular suffix for first singular present indicative is “o”.

From the examples we can see other characteristics of the formalism. For instance, “<suf> == C123: <suf>” means that, by default, every suffix unspecified in C23 inherits form the path <suf> of C123 (i.e. the three statements for the 1st person singular present indicative and the polite imperatives). If more statements had been included in C123, and were not specified in C23 or at any other lower level (i.e. C2 or C3), then these added paths would have also been inherited from the general macroparadigm.

Another property of this formalism is shown in the following statement:

“<suf reg ind impf 3 sing>   == < suf reg ind impf 1 sing>“.

This is the way paradigmatic homonymy (section 2.1.1) can be represented. The suffix of the 3rd person singular imperfect indicative is the same as that of the 1st person singular imperfect indicative. The same mechanism is also used for polite imperatives.

According to this analysis, regular (i.e. by default or unmarked) macroparadigms are expressed this way:

```
C1:
        <suf>                       == C123: <suf>
        <suf  reg  inf>         == ar
        ... (and 51 more)
C2:
        <suf>                       == C23: <suf>
        <suf reg inf>               == er
        ... (and 15 more)
C3:
        <suf>                       == C23: <suf>
        <suf reg inf>               == ir
        ... (and 15 more).
```

Although the three paradigms have the same number of verb forms, the specifications for C2 and C3 are much shorter than for C1, since the former ones inherit from C23, and C1 inherits from C123 with only three statements.

We now need the rule which governs the formation of inflected words, i.e. the concatenation of a root and a suffix. In the regular paradigms the stem does not change. The basic default rule is as follows:

```
REG:
        <>                == VERB
        <root>            == "<stem>"
        <suf>                     == "suf reg>"
        <wf>                      == "<root>" "<"suf>".
```

REG inherits from VERB, where one can state any kind of information different from the morphological one. Usually, it is used for encoding syntactic information such the part of speech.

The last statement is the concatenation rule: a wordform (wf) is consists of a root and a suffix. We distinguish between *root* and *stem*: The latter is used to designate an allomorph stem (“stem” is the most basic one, “stem 2”, “stem 3”, etc.). *Root* is used to select the correct allomorph stem; in the regular paradigm, it is “<stem>“. The same can be said of *suf*, since it differentiates regular endings (“<suf reg>“) from irregular ones (“<suf irreg>“, such as /-o/ in *pus-o*, instead of **pon-ió*).

This rule is used for all the verbs, except for the extremely irregular ones (such as *ser* or *ir*) and defective verbs (those with missing forms in their paradigms). Lexical entries, (i.e. the lowest level in the network) are so simple because they inherit most of their information from the upper levels:

```
Amar:
  <>          == REG
  <des>       == C1
  <stem>      == am.
Temer:
  <>          == REG
  <suf reg>          == C2
  <stem>      == tem.
Partir:
  <>         == REG
  <suf reg>  == C3
  <stem>     == part.
```

Regarding the so-called IRREGULAR paradigms, Spanish shows a great variety of irregularities (80 paradigms in the *American Heritage Larousse Spanish Dictionary;* 62 paradigms in Seco (1986), 62 classes in Tzoukermann and Liberman (1990)). All of these forms would be difficult to learn and process if there were not a good deal of redundancy and generalization, as the Paradigm Economy Principle suggests. Our proposal specifies a very reduced number of default macroparadigms, which the rest of the subparadigms inherit from[12]. Moreover, all the paradigms can be ordered in an inheritance network (Figure 4), which captures these generalizations:

> -All paradigms inherit directly or indirectly from the regular one, REG.
> -Most subparadigm (those with the name of a representative verb) follow either pattern 2-STEM_A or pattern 2-STEM_B.
> -Most irregular paradigms inherit directly from 2-STEM_A

Interestingly enough, our proposed inheritance network validates Cartairs' *Macroparadigm Uniqueness Claim*: each paradigm belongs to (i.e. inherits from) one macroparadigm and one only.

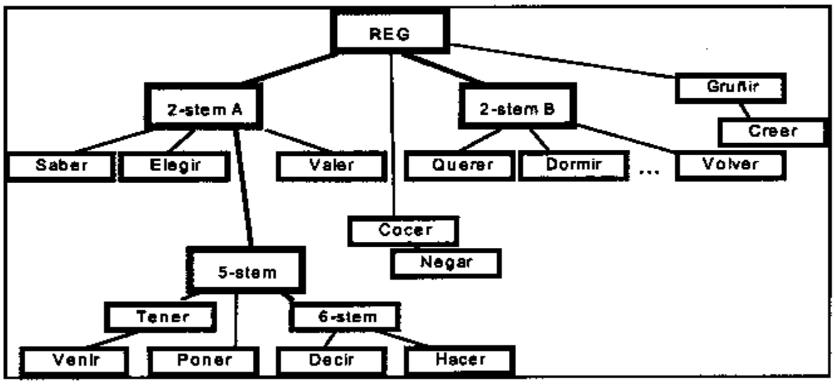

Figure 4: Verb inheritance network

Irregularity in Spanish verbs is produced mainly by a modification in the stem. Subparadigms can thus be classified in terms of their number of alternate stems:

TWO-STEM verbs such as *cazar* [13](/caz-/ and /cac-/), *despertar* (/despert/, /despiert-/), coger (/cog-/, /coj-/)

THREE-STEM verbs: *querer* (/quer-/, /quier-/, /quis-/), *valer* (/val-/, /valg/, /vald-/).

FOUR-STEM verbs: *cocer* (/coc-/, /cuec-/, /coz-/, /cuez-/), *negar* (/neg-/, /nieg-/, /negu-/, /niegu-/) *saber* (/sab-/, /sep-/, /sup-/, /sé/)

FIVE-STEM verbs: *venir* (/dec-/, /dig-/, /dij-/, /di-/, /dic-/, /dich-/), *poner* (/pon-/, /pong-/, /pus-/, /pond-/, /puest-/), *tener* (/ten-/, /teng-/, /tuv-/, /tend-/, /tien-/)

SIX-STEM verbs: *decir* (/dec-/, /dig-/, /dij-/, /di-/, /dic-/, /dich-/) y *hacer* (/hac-/, /hag-/, /hic-/, /ha-/, /hiz-/, /hech-/)

We shall now examine the two main irregular macroparadigms: 2-STEM_A and 2-STEM_B. The tables below illustrate the distribution of the stems within the paradigm:

Table 1: Macroparadigm 2-STEM_A

	1 sg	**2 sg**	**3 sg**	**1 pl**	**2 pl**	**3 pl**	**nofin**
pres_ind	stem 2	stem 1	stem 1	stem 1	stem 1	stem 1	
impf_ind	stem 1	stem 1	stem 1	stem 1	stem 1	stem 1	
pret_ind	stem 1	stem 1	stem 1	stem 1	stem 1	stem 1	
future	stem 1	stem 1	stem 1	stem 1	stem 1	stem 1	
pres_subj	**stem 2**	**stem 2**	**stem 2**	**stem 2**	**stem 2**	**stem 2**	
imp_subj	stem 1	stem 1	stem 1	stem 1	stem 1	stem 1	
cond	stem 1	stem 1	stem 1	stem 1	stem 1	stem 1	
imper		stem 1			stem 1		
infin							stem 1
ger							stem 1
part							stem 1

Table 2: Macroparadigm 2-STEM_B

	1 sg	2 sg	3 sg	1 pl	2 pl	3 pl	nofin
pres_ind	**stem 2**	**stem 2**	**stem 2**	stem 1	stem 1	**stem 2**	
impf_ind	stem 1	stem 1	stem 1	stem 1	stem 1	stem 1	
pret_ind	stem 1	stem 1	stem 1	stem 1	stem 1	stem 1	
future	stem 1	stem 1	stem 1	stem 1	stem 1	stem 1	
pres_subj	**stem 2**	**stem 2**	**stem 2**	stem 1	stem 1	**stem 2**	
imp_subj	stem 1	stem 1	stem 1	stem 1	stem 1	stem 1	
cond	stem 1	stem 1	stem 1	stem 1	stem 1	stem 1	
imper		**stem 2**			stem 1		
infin							stem 1
ger							stem 1
part							stem 1

```
2-STEM_A:
        <>                          == REG
        <root ind pres 1 sing>      == “<stem 2>“
        <root subj pres>            == “<stem 2>“.

2-STEM_B:
        <>                          == REG
        <root ind pres 1 sing>      == “<stem 2>“
        <root ind pres 2 sing>      == “<stem 2>“
        <root ind pres 3 sing>      == “<stem 2>“
        <root ind pres 3 plu>       == “<stem 2>“
```

```
<root subj pres>              == “<stem 2>“
<root subj pres 1 plu>        == “<stem 1>“
<root subj pres 2 plu>        == “<stem 1>“
<root imper 2 sing>           == “<stem 2>“
```

Generalizations can be expressed very compactly in DATR. The three statements in 2-STEM_A say that this paradigm inherits from REG by default, and that the 1st person singular present indicative and all present subjunctive wordforms use stem 2, which thus overrides the default stem.

The subparadigm 4-stem, which is the base for the most complex paradigms, is interesting because in it, one can find not only five or six stem allomorphs, but also suffix allomorphs:

Table 3: Macroparadigm 4-STEM

	1 sg	**2 sg**	**3 sg**	**1 pl**	**2 pl**	**3 pl**	**nofin**
pres_ind	stem 2	stem 1	stem 1	stem 1	stem 1	stem 1	
impf_ind	stem 1	stem 1	stem 1	stem 1	stem 1	stem 1	
pret_ind	stem3	stem3	stem3	stem3	stem3	stem3	
future	stem4	stem4	stem4	stem4	stem4	stem4	
pres_subj	stem 2	stem 2	stem 2	stem 2	stem 2	stem 2	
imp_subj	stem3	stem3	stem3	stem3	stem3	stem3	
cond	stem4	stem4	stem4	stem4	stem4	stem4	
imper		stem 1			stem 1		
infin							stem 1
ger							stem 1
part							stem 1

```
4-STEM:
  <>                        ==  2-STEM_A
  <suf ind indf>            == PRET1-13
  <suf ind fut>             == FUTCOND
  <suf cond>                == FUTCOND
  <root ind indf>           == "<stem 3>"
  <root ind fut>            == "<stem 4>"
  <root subj impf>          == "<stem 3>"
  <root cond>               == "<stem 4>"
  <wf imper 2 sing>         == "<stem 1>".
```

PRET1-13 stands for the irregular suffix allomorphs for the 1st and 3rd person singular past indicative (/-e/ and /-o/, as in *tuv-e* and *tuv-o*). Analogously, FUTCOND stands for the irregular allomorphs for future and conditional (*ten-DRÉ*, *ten-DRÁS*, etc.)

Interestingly enough, the five multi-allomorphs paradigms (TENER, VENIR, PONER, DECIR, and HACER) do inherit from 4-STEM. For instance, TENER (including its derivatives *contener, retener*, etc.) has the following paradigm:

```
TENER:
  <>                        == 4-STEM
  <root ind pres 2 sing>    == "<stem 5>"
  <root ind pres 3>         == "<stem 5>".
```

Its lexical entry would be:

Tener:

```
  <>                 == TENER
  <suf>              == C2
  <stem>             == ten
  <stem 2>           == teng
  <stem 3>           == tuv
  <stem 4>           == tend
  <stem 5>           == tien
```

To add new verbs to the lexicon it is only necessary to assign the new verb to an existing paradigm and specify the stems. Of course, to add regular verbs is even simpler:

Ablanablar[14]:

```
 <>              == REG
 <suf>           == C1
 <stem>          ==  ablanabl
```

This approach has further advantages over a non-paradigmatic one. First of all, *suppletive forms* are naturally encoded in paradigmatic morphology, since it is only necessary to specify the suppletive form(s) in the lexical entry or the paradigm. A case in point is the paradigm of *ser*:

Ser:

```
    <>                       == VERB
    <wf ind pres 1 sing>     == soy
    <wf ind pres 2 sing>     == eres
    <wf ind pres 3 sing>     == es
    ...
```

Secondly, *defective verbs*, which lack certain wordforms in their paradigms, are extremely difficult to encode in a typical two-level or IP model. In a straightforward paradigmatic framework, only the existing forms are defined. In DATR, this can be encoded by means of a special paradigm for each defective verb, in which the statement '<> == REG' is eliminated, something that also happens with highly suppletive verbs. In other words, defective and highly irregular verbs are the only cases where there is no inheritance from any (macro)paradigm.

Finally, we provide an estimation of the productivity of the main paradigms, based on the ARIES lexicon containing about 7,500 verbs:

Paradigm	*% in ARIES lexicon*
REG	71.4 %
2-STEM_A	3.0 %
2-STEM_B	4.7 %
4-STEM based paradigms (TENER, VENIR, PONER, DECIR, HACER)	0.9
CAZAR	15.5 %
NEGAR	0.7 %

The above figures show that an overwhelming majority of verbs in Spanish follow the REGULAR paradigm. If we take the phonological approach, then the CAZAR paradigm (which includes all verbs with only spelling modifications in the stem) should be considered as a member of REGULAR. That implies that around 87 % of the Spanish verbs are regular.

The most difficult paradigms to learn are obviously those with 5 or 6 stem allomorphs. However, they represent less than 1% of the total. On the other hand, they are among the most frequently used. Difficult learning is evidently compensated by productivity, or otherwise the system would eventually try to minimize their irregularities. Of the 13% of irregular verbs, almost 8% are two-stem verbs, representing two different patterns that are followed by the majority of the remaining 5% of irregular verbs.

This estimation, obviously, is not based on a closed inventory of verbs, but while we are certain that virtually all irregular verbs are included in it, many regular ones are not. New verbs are productively created following the regular paradigms. We can say confidently that the real proportion is nine to one in favor of the regular paradigms.

4. Conclusions

We have presented an account of the inflectional morphology of Spanish verbs, using the concepts of paradigm, feature and inheritance. We believe that its explanatory capacity and simplicity makes it more satisfactory than Two-Level/finite-state implementations, as well our own unification approach (Moreno 1991).

This article is a contribution to the WP model, especially to the theoretical concepts proposed by Carstairs, name that of macroparadigm, and the Paradigm Economy Principle. We think that for Spanish (and other similar languages) the paradigmatic approach is both descriptively and psychologically well-founded,. In this respect, we believe that for a learner of Spanish as a foreign language, it is easier to acquire a paradigm than a rule. The inheritance network presented suggests that in spite of the apparently high number of irregular paradigms, which are the product of "flat" descriptions, all of them are well structured. This description seems to agree with the fact that any competent speaker of Spanish can manage its rich verb morphology.

We believe that our analysis provides useful information for teaching Spanish as second language. Firstly, it is possible to teach irregular paradigms by focusing on the macroparadigms and their connections with subparadigms. Secondly, the quantification of the paradigm productivity can provide information concerning the best areas to focus on. Although further research on this topic is necessary, García and Ambadiang (1996) present preliminary within a paradigmatic approach.

Finally, we have shown the suitability of DATR and Network Morphology frameworks for a rich description of morphology. In future papers we will implement the Spanish nominal inflection (nouns, adjectives, determiners, pronouns, some quantifiers), and try to apply the model to derivative morphology.

Notes

1. A shorter version of this paper was presented at *XXVI Simposio de la Sociedad Española de Lingüística*, Madrid, December 1996. Part of the new information included here has been taken from an unpublished work by Moreno and Goñi.

2. In this respect, insights from Paradigmatic Morphology (e.g. Matthews, Carstairs, Stump) and feature-based theories (e.g. LFG) have been especially helpful. We would particularly like to acknowledeg the influence of DATR (Evans & Gazdar 1996; Cahill & Gazdar 1997) and Network Morphology (Corbett & Fraser 1993; Fraser & Corbett 1995; Brown et *al.* 1996).

3. We will omit the discussion on the complexity and efficiency of the Two-Level morphology, that Sproat (1992:174) sums up this way:

> "it is possible to encode very hard problems in the system, and (not surprisingly) the system performs generally poorly on such problems [...] The KIMMO framework as it stands is inadequate as a model of computational description of natural language morphology because the formalism does not allow one to distinguish genuinely computationally difficult problems [...] from the typically easier problems that one finds in natural language."

4. The most extreme position is sustained by Karlsson and Karttunen (1996): "The quest for an efficient method for the analysis and generation of wordforms is no longer an academic research topic". In contraposition, Cahill and Gazdar (1997) express that they "do not embrace the traditional notion of level of description, nor its concomitant notion of rule type (or process) mapping from one level to another".

5. This work does not follow the Koskenniemi model, but is a finite-state implementation.

6. As for the complexity of rules interaction, Sproat (1992: 147) notes:
Koskenniemi is "cautious" with the rules in that, unless an alternation is very productive, he would rather describe that alternation by listing two allomorphs of the same morpheme in the lexicon than describe it using a rule [...] although there is no question that the two-level model renders unwieldy a description involving many rules with complex interactions.

7. This criticism only applies to Spanish (and morphologically similar languages) implementations of Two-Level Morphology, since for other languages it has shown itself to be an elegant account.

8. The continuation classes in Two-Level Morphology are in fact a type of morpheme concatenation. On the other hand, string concatenation is one of the two parts in the unification operation.

9. Analyses of the use of features and unification in morphology within the LFG framework can be found in Andrews (1990) and Vincent and Börjars(1996).

10. Abbreviations used:

suf	= suffix
reg	= regular
ind	= indicative
pol	= polite
pres	= present
impf	= imperfect
ger	= gerund
par	= participle.

[11] "However, one of the aims of the network morphology enterprise is to construct the minimal superset of atoms [...] required for a theory of language[...]" (Brown *et al* 1996: 62)

[12] This observation dates back to Bello (1898), who distinguished 13 classes of irregularities based on the wordform distribution in the paradigm.

[13] We have classifed different spelling forms as stem allomorphs. Obviously, from a phonological point of view, they are regular verbs.

[14] Non-existing but possible verb in Spanish.

References

Anderson, S. (1992): *A-Morphous Morphology*. Cambridge, CUP.

Andrews, A.(1990): "Unification and morphological blocking". *Natural Language & Linguistic Theory*, 8, pp. 507-557.

Badia, T., Egea, M.A. & Tuells, A.(1996): "SEGMORF: un formalismo para analizadores morfológicos de dos niveles", *SEPLN*, 19, pp. 63-71.

Bauer, L. (1988): *Introducing Linguistic Morphology*. Edinburgh University Press

Bear, J. (1986): "A morphological recogniser with syntactic and phonological rules". *COLING-86*, pp. 272-276.

Beard, R. (1995): *Lexeme-Morpheme Base Morphology*. State University of New York.

Bello, A. (1898): *Gramática de la lengua castellana*. Madrid, Edaf (reimp. 1984).

Brown, D., Corbett, G., Fraser, Hippisley, A. & Timberlake, A. (1996): "Russian noun stress and network morphology". *Linguistics*, 34, pp.53-107.

Cahill, L. & Gazdar, G. (1997): "The inflectional phonology of German adjectives, determiners, and pronouns." *Linguistics*, 35, pp. 221-245.

Carstairs, A.(1987): *Allomorphy in inflexion*. Beckenham, Croom Helm.

Carulla, M. & Oosterhoff, A.(1996): "El tratamiento de la morfología flexiva del castellano mediante reglas de dos niveles en una gramática de unificación", *SEPLN*, 19, pp. 72-80

Corbett, G.& Fraser, N. (1993): "Network Morphology: A DATR account of Russian inflectional morphology." Journal of Linguistics, 29, pp. 113-142.

Evans, R. & Gazdar, G. (1989): *The DATR papers.*. CSRP 139, Univ. of Sussex.

Evans, R. & Gazdar, G. (1996): "DATR: A language for lexical knowledge representation." Computational Linguistics, 22 (2), pp. 167-216.

Fraser, N. & Corbett, G. (1995): "Gender, animacy, and declensional class assignment," in Booij & van Marle (eds.) *Yearbook of Morphology 1994*. Kluwer, pp. 123-150.

García, I. and Ambadiang, T. (1996): "Lo paradigmático en el sistema verbal de los estudiantes de español L2: implicaciones teóricas y pedagógicas", en Pérez Pereira (ed.)*Primer Congreso Internacional sobre Adquisición de las lenguas del Estado*. Univ. de Santiago.

Gazdar, G. (1989): "An introduction to DATR", in Evans & Gazdar (1990), pp. 1-14.

Goñi, J.M, González, J.C. & Moreno, A.(in press): "ARIES: A Lexical Platform for Engineering Spanish Processing Tools"" *Natural Language Engineering*

Koskenniemi, K. (1983): *Two-level Morphology: A general computational model for wordform recognition and production*. Ph.D. thesis, University of Helsinki.

Marti, M.A. (1986): "Un sistema de análisis morfológico por ordenador""*SEPLN*, 4, pp. 91-103.

Matthews, P.H. (1972): *Morphology*. Cambridge, Cambridge University.

Meya, M. (1987): "Morphological analysis of Spanish for retrieval"" *Literary & Linguistic Computing*, 2(3), pp. 166-170.

Moreno Sandoval, A. (1991): *Un modelo computacional basado en la unificación para el análisis y generación de la morfología del español*. Ph.D. Thesis. Universidad Autónoma de Madrid.

Moreno, A. & Goñi, J.M.(1995): "GRAMPAL: a Morphological Processor for Spanish implemented in Prolog", in *Proceedings of GULP-PRODE '95, Joint Conference on Declarative Programming*

Robins, R.H. (1959): "In defence of WP", reprinted in *Diversions of Bloomsbury: Selected writings on linguistics*. Amsterdam, North-Holland, pp. 47-77.

Spencer, A.(1991): *Morphological theory*. Oxford, Blackwell.

Sproat, R. (1992): *Morphology and Computation*. Cambridge, The M.I.T. Press

Stump, G.(1991): "A paradigm-based theory of morphosemantic mismatches"" *Language*, 67, 4, pp. 675-725.

Trost, H. (1990): "The application of two-level morphology to non-concatenative German morphology"" *Proc. of COLING-90*, vol. 2, pp. 371-376.

Tzoukermann E. & Liberman, M. (1990): "A finite-state morphology processor for Spanish". *Proc. of COLING-90*, Helsinki, vol. 3, pp. 277-281.

van Marle (1985): *On the paradigmatic dimension of morphological creativity*. Dordrecht, Foris.

Wurzel, W. (1984, 1989): *Inflectional morphology and naturalness*. Dordrecht, Reidel.