

On dealing with morphographemic and morphotactical interaction phenomena in SEGMORF

Toni Badia

IULA

Universitat Pompeu Fabra
La Rambla 30-32, Barcelona
Catalonia, Spain
tbadia@upf.es

Antoni Tuells

IULA

Universitat Pompeu Fabra
La Rambla 30-32, Barcelona
Catalonia, Spain
tuells@upf.es

Abstract

Within two-level morphology, morphotactics is modelled either in continuation classes or in unification word grammars. Though for the first approach very efficient systems exist, the latter provides more elegant morphosyntactic parsing. If the latter approach is taken, the present two-level formalisms have some problems in dealing with morphological phenomena involving interaction between two-level rules (TLR's) and the word grammar (WG). On the one hand their WG is not modular, since it cannot be design according to linguistic criteria only. On the other hand, they do not provide a clean and general solution for this kind of morphological phenomena in Romance languages. We therefore put forward a new implemented formalism, *SEGMORF*, that allows the linguist to express in an alternative way the interaction between TLR's and the WG. We discuss the benefits and drawbacks of our proposal.

Keywords: morphological analysis

1 INTRODUCTION

Morphology (especially recent Computational Morphology) sees word formation as composed of two main processes:

- morphographemics
- morphotactics

In the TLM framework ((Koskenniemi, 1984) see also (Ritchie et al., 1992) and (Kaplan and Kay, 1994))), the first process is modelled in two level rules, and the second one is usually modelled either in continuation classes or in unification word

grammars. For the first approach, very efficient systems exist (Karttunen et al., 1992) and (Karttunen, 1994), though the latter provides more elegant morphosyntactic parsing (see for example (Ritchie et al., 1992)). Both approaches aim at providing the linguists with a formalism that enables them to express in a natural way both the morphographemic and the morphotactic relations existing in words of the language being described. However, since morphographemics aims at finding a lexical decomposition from a given surface form and since morphotactics aims at building a word structure out of that lexical decomposition, several morphological phenomena pose interesting problems, as can be seen below:

- (a) (German) *Männer* (noun,pl)
(eng:men)
⇒ *Mann* (noun,sing) + *er* (pl)
- (b) (German) *Mütter* (noun,pl)
(eng:mothers)
⇒ *Mutter* (noun,sing) + *null* (pl)
- (c) (Spanish) *puedo* (verb,1st,sing,present)
(eng:can)
⇒ *pod* (verb_stem) + *o* (1st,sing,present)
- (d) (Spanish) *pudo* (verb,3rd,sing,past)
(eng:could)
⇒ *pod* (verb_stem) + *o* (3rd,sing,past)

On the one hand, the morphotactical component appears to require information on the application of TLRs (i.e., in order to select the appropriate decomposition, it should bear information on which rule has been applied). On the other hand, as it is well known, several morphological phenomena, like German Umlautung, restrict their appropriate morphological context (i.e., certain rules occur only in particular word formation processes). Both points are the two sides of the same coin: several morphological phenomena should be dealt with according to

well-defined, linguistically motivated interaction between the morphographemic and the morphotactical components of a two-level system.

In section 2 we review former approaches to these morphological phenomena. In section 3 we put forward an alternative two-level formalism, *SEGMORF*, which is meant to deal in an alternative way with morphographemic and morphotactical interaction phenomena within the WG approach. We also discuss the benefits and drawbacks of our proposal. In section 4 we describe *SEGMORF* more formally, and in section 5 our conclusions are reported.

2 Morphographemic and morphotactical interaction phenomena. Former approaches

This section is devoted to review former approaches to morphographemic and morphotactical interaction phenomena, though we will concentrate on the WG approach (i.e the approach in which morphotactics is modelled using an unification word grammar).

2.1 The continuation classes approach

In this section we review very briefly how morphographemic and morphotactical interaction can be accounted for in the continuation classes approach adopted in (Karttunen et al., 1992) and (Karttunen, 1994). For expository purposes we will concentrate on cases *puedo* and *pudo* (cases (c) and (d) in section 1).

The necessary lexical entries in their respective sublexicons may be the following:

```
V_stem
  p0d

V_suffix_present
  ^Diph o

V_suffix_past
  ^clos o
```

Through a cascade of TLRs one could arrive at the following lexical decompositions:

- p0d + ^Diph o
- p0d + ^clos o

Note that the diacritic symbols present in the sublexicons mark somehow which rule has been applied. The consequences of this approach are the following:

- Introduction of special symbols in lexicons and sublexicons

- No possibility of underspecification of “conflictive” suffixes. See section 3 for more details.
- However the implemented systems that follow this approach are reported to be very efficient.

2.2 The WG approach

In the WG framework, the most accepted way of stating this interaction has been Trost’s proposal. In his own words (Trost, 1990):

- Information about the application of a rule needs to be transferred to the morphosyntactic grammar.¹
- It must be possible to restrict the application of TLRs to certain classes of morphs.

The transfer of information is thought of as unification of the feature structure associated to the rule with the feature structure associated to the morph found in the lexicon. If unification succeeds, the resulting feature structure is transferred to the word grammar. Though this proposal certainly works, the first requirement poses problems which are well worth mentioning:

- The WG (which ideally should provide a linguistically motivated description of word structure) is not uniquely motivated on strict linguistic criteria. It is often seen as a process that validates the application of TLRs and filters out wrong decompositions. Its formulation cannot be independent of the TLRs.
- The WG is difficult to handle, since it has to be tuned to the interaction of different rules selecting their own morphological context.

Both points collapse into a single one: the WG is not modular. Consider for example what happens if we take this approach in order to account for the following phenomena in Catalan and Spanish:

- (e) (Sp) *canta* (verb,3rd,sing,present)
(eng:*sings*)
⇒ *cant* (verb_stem) + *a* (3rd,pres)
- (f) (Sp) *cántame* (verb,2nd,sing,imper)
(eng:*sing_to_me*)
⇒ *cant* + *a* (verbal_suffix) + *me* (clitic)

¹This is the approach taken in (Trost, 1990) for dealing with german Umlautung and the approach suggested in (Carulla and Oosterhoff, 1996) for dealing with diphthongization and obscuration (cases like *puedo* and *pudo* in section 1)

- (g) (Sp) *cántaselo* (verb,2nd,sing,imper)
(eng: *sing_it_to_him*)
⇒ *cant* + *a* (v_suffix) + *le* (clitic) + *lo* (clitic)
- (h) (Sp) *cantádmelo* (verb,2nd,pl,imper)
(eng: *sing_it_to_me*)
⇒ *cant* + *ad* (v_suffix) + *me* (clitic) + *lo* (clitic)
- (i) (Ca) *espès* (adj,masc,sing)
(eng: *thick*)
- (j) (Ca) *espessos* (adj,masc,pl)
(eng: *thick*)
⇒ *espès* + *s* (pl)
- (k) (Ca) *espesset* (adj,masc,sing)
(eng: *thickish*)
⇒ *espès* + *et* (affix)
- (l) (Ca) *espessets* (adj,masc,pl)
(eng: *thickish*)
⇒ *espès* + *et* (affix) + *s* (pl)
- (m) (Ca) *abús* (noun,masc,sing)
(eng: *abuse*)
- (n) (Ca) *abusos* (noun,masc,pl)
(eng: *abuses*)
⇒ *abús* + *s* (pl)
- (o) (Ca) *abusets* (noun,masc,pl)
(eng: *insignificant_abuses*)
⇒ *abús* + *et* (affix) + *s* (pl)

The rules that put the graphical accent or duplicate the “s” are clearly morphologically motivated. To put it another way, the graphical accent (and similarly for “s”-duplication) appears whenever a word formation process occurs. According to Trost’s proposal, the word grammar should handle the interaction between rules, which means that it should take into account whether the rules that put the graphical accent and duplicate the “s” have been applied. Even worse, since this validation has to be done for several word formation processes, the complexity of this task increases.²

3 OUR PROPOSAL

To overcome the difficulties previously shown, the handling of morphological phenomena which involve interaction between the morphographemic and morphotactical components within the two-level approach imposes the following requirements:

²Of course, another possibility is to write rules which specify all surface contexts; since this has to be done for many word formation processes, this task becomes also very complex

1. The WG should be kept as modular (independent) as possible.
2. It must be possible to restrict the application of TLRs to certain classes of morphs. (This is identical to the second point of Trost’s proposal).
3. It should be possible to specify TLRs in a way that the morphological context is taken into account (along with the morphographemic one).

TLRs in *SEGMORF* have been designed according to these requirements (see section 4 for more technical details). In what follows, our main concern are points 1 and 3 (given that point 2 has already been accounted for in previous work (Bear, 1988); (Trost, 1990)). With respect to them, we propose rules for handling cases (e)-(h) and rules for cases (c)-(d) (*puedo* and *pudo*). These rules will show the reader the expressivity of TLRs in *SEGMORF*.³ We thus propose the following rules for handling cases (e)-(h).⁴

Rule `stress_in_verb`:

```

á <=> A
morphemes_to_be_found =
(morf(_,class(verbal_stem))
+ morf(lema(a),class(verbal_suffix))
+ morf(_,class(clitic)))

```

Rule `clitic_before_clitic`:

```

se <=> le
morphemes_already_found =
(morf(_,class(verbal_stem))
morphemes_to_be_found =
(morf(_,class(clitic))
+
morf(_,class(clitic))
)
}

```

Informally, the first rule says that a surface vowel is converted into a lexical stressed one if the morpheme decomposition includes a verbal stem plus one suffix and one clitic, where the first one must be “a”. This rule would account for the following word forms of every verb :

- `verb_stem + a + me` (ex: *cántame*)

³This is an em syntactic sugar version of the actual rules implemented in our system

⁴Note that `morphemes_to_be_found` includes the morph to which the rule applies.

- verb_stem +a + me + lo
(ex: *cántamelo*)

- verb_stem +a + se (ex: *cántase*)

Note also that the incorrect word form *cántadmelo* (*cant* + *ad* + *me* + *lo*) is not allowed, since once we find morpheme “ad”, the application of the first rule is not validated.

The second rule says that a surface “se” is converted into a lexical “le” if “le” is a clitic which appears before another clitic, and a verbal stem has already been found (case (g)). It is important to point out that the proposed TLRs interact with the morphotactic component; they put constraints on the already found and expected morphemes, which intuitively corresponds to what a linguist understands as a morphological context restricting the application of a rule.

From the procedural point of view, the following data structures are needed:

- Surface Left and Right morphographemic contexts
- Lexical Left and Right morphographemic contexts
- Morphological Left and Right contexts
- Application context (i.e., a feature structure which keeps trace of the application of rules. For example [umlautung:y]). This feature structure must unify with the application-FS associated to every morph found in the lexicon.⁵

Thus the application of a rule is validated with respect to all contexts in a homegenous way. Note also that though one needs to write more rules, they are very simple to specify, and the morphological contexts help to rule out their application as soon as possible, thus avoiding overgeneration. Note also that the WG does not have to deal with the application of TLRs. Its main concern here is the validation of lexical decompositions; no information on rules is needed, thus making the grammar as modular (simple) as possible. Thus, to handle the cases of Spanish *puedo* (*can*) and *pudo* (*could*) the following rules are needed:

```
Rule diphtongization: {
  ue <=> 0
  morphemes_to_be_found =
    (morf(_,class(verbal_stem)))
```

⁵A morph disallowing Umlautung would have the following application-FS: [umlautung:n]. This FS corresponds to requirement 2 of our proposal.

```
+
  morf(_,class(verbal_suffix,present))
)
}
```

```
Rule closing: {
  u <=> 0
  morphemes_to_be_found =
    (morf(_,class(verbal_stem))
    +
    morf(_,class(verbal_suffix,past))
    )
}
```

Note that the rules themselves select their appropriate morphemes; disambiguation is done as soon as possible (in fact, when looking up at the lexicon); the WG does not have to check which rule was applied. Note also that specifying morphological contexts within TLRs allow underspecification of morphemes (in our last example, one could have an underspecified verbal suffix “o”) and its immediate disambiguation.

Lastly, the following rule accounts for cases (i)-(o). Recall that in those cases the graphical accent is not dependent on the surface context, but on word formation processes: it appears whenever a nominal suffix is added to the stem. We only show the rule that puts the graphical accent:

```
Rule put_accent : {
  Vowel<=>Stressed_vowel
  surface_right_context = sX
  /* X is a character variable */
  lexical_right_context = s+
  /* "+" = morpheme boundary */
  morphemes_to_be_found =
    (morf(_,class(noun))
    +
    morf(_,class(nominal_suffix))
    )
}
```

Note that rule `put_accent` neither needs to specify complex surface contexts nor involves a complex WG.

3.1 Benefits and drawbacks of our proposal

The advantages and drawbacks of our proposal are the following:

- One expects a system that deals with Natural Language to be as modular as possible, where modules are designed according to linguistic criteria. Former TLM formalisms failed to ac-

compish this objective with respect to the WG module, since its formulation was dependent on the TLRs. In *SEGMORF*, the WG is a true independent module, since TLRs select the appropriate morphemes whenever a disambiguation conflict may arise. Thus the WG can be designed under word formation considerations only.

- Obviously, under our proposal, TLRs are not formulated under morphographemic considerations only. TLRs become somewhat more complex, but in return the WG is clearly more simple; and it is easy to show that this shift in complexity adapts naturally to the way linguists look at the problem: although it is wholly natural to look at the morphotactical contexts in which TLR apply when writing them, it is not natural at all to specify which TLR have to be applied in the word-building rules.
- In terms of efficiency, our TLRs decrease the global performance of the system. Firstly, the application of TLRs have to take into account not only morphographemic contexts but also morphotactical ones. Secondly, TLRs become less specific in their surface contexts; as a result, more rules could be applied for a given surface context. In many cases we could alleviate this problem by using diacritic symbols; this is arguable for stress marking and diphthongization, for example.
- Our approach allows underspecification of morphemes and immediate disambiguation after application of TLRs.
- Our approach does not provide a better treatment for German Umlautung, as there are too many possible morphotactical contexts. In such cases, *SEGMORF* allows the linguist to adapt Trost's strategy.
- A requirement one would put on a morphological formalism is that of linguistic felicity. *SEGMORF* provides the linguists with more linguistically motivated ways of expressing morphographemic and morphotactical relations, thus avoiding some unnatural approaches.

4 SEGMORF. The Formalism

This section describes briefly the TLRs component of *SEGMORF*, which is an extension of the Alep morphographemic segmentation formalism

(Pulman, 1991)⁶. In *SEGMORF*, rules specify the morphographemic and morphotactical context constraining its application.

4.1 The Orthographical (morphographemic) Context

There are four types of orthographical contexts of the following form: (this is a reformulation of (Pulman, 1991), p.161)

1. $SLC\ LHS\ SRC \implies LLC\ RHS\ LRC$
2. $SLC\ LHS\ SRC \impliedby LLC\ RHS\ LRC$
3. $SLC\ LHS\ SRC \iff LLC\ RHS\ LRC$
4. $SLC\ LHS\ SRC\ opt\ LLC\ RHS\ LRC$

where each of *SLC*, *LHS*, *SRC*, *LLC*, *RHS* and *LRC* is a sequence of zero or more characters or character-variables (i.e. variables ranging over single characters) with:

- *SLC* abbreviating Surface Left Context
- *LHS* abbreviating Left Hand Side
- *SRC* abbreviating Surface Right Context
- *LLC* abbreviating Lexical Left Context
- *RHS* abbreviating Right Hand Side
- *LRC* abbreviating Lexical Right Context

Variables are interpreted via unification. It is possible to constraint the value a variable may take; for instance, the user may specify that a given variable must be a vowel.

It is well worth mentioning that Contexts, as well as the LHS and RHS of rules, do not need to have identical number of characters, so rules may affect more than one character at a time.⁷ The rules with an \implies operator mean that if the contexts (both lexical and surface) are satisfied, then its LHS corresponds to its RHS. The rules with an \impliedby operator mean that if the contexts are satisfied, then its RHS corresponds to its LHS. The rules with an \iff mean that if the contexts are satisfied, then its LHS can only correspond to its RHS, and its RHS can only correspond to its LHS. The rules with an *opt* operator mean that if the contexts are satisfied,

⁶It thus belongs to the family of *partition* formalism (Evans et al., 1996)

⁷An important difference between *SEGMORF* and the Alep morphographemic formalism is that in *SEGMORF* RHS of rules can cross the morpheme boundary. See examples in next section.

then its LHS may correspond to its RHS, and conversely, its RHS may correspond to its LHS. Thus, \Rightarrow rules enforce correspondences from Surface to Lexical strings, \Leftarrow rules enforce correspondences from Lexical to Surface strings and \Leftrightarrow rules enforce correspondences in both directions; these are actually obligatory rules. *opt* rules are optional rules, since they only allow correspondences (without enforcing them).

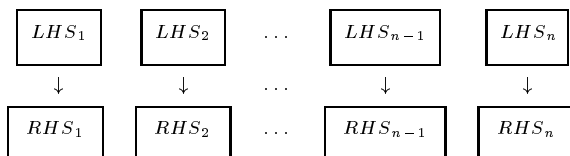
There are two characters which have a special meaning for *SEGMORF*:

- *+* is an explicit morpheme boundary marker.
- *#* is an explicit surface string end marker.

Lexical entries (morphemes) may disallow the application of certain rules via negative rule features (Bear, 1988); (Trost, 1991).

Taken directly from (Pulman, 1991), p.162:

The interpretation of the rules is in terms of partitioning of the surface and lexical string, where the surface string is the input for analysis, and the lexical string is some sequence of morphemes to be found in the lexicon. A partitioning is some division of the strings into concatenated sequences of zero or more characters, such that there are identical numbers of sequences in both strings and the *n*'th surface sequence is regarded as corresponding to the *n*'th lexical sequence:



4.2 The Morphological Context

The morphotactical context is of the following form:
LMC RMC

where each of LMC and RMC is a sequence of zero or more morphotactical descriptions with:

- LMC abbreviating Left Morphotactical Context
- RMC abbreviating Right Morphotactical Context

A morphotactical description includes very simple morphosyntactic information (POS, agreement, number of vowel groups. . .) of a morpheme (see next section for motivation).

The interpretation of the morphotactical context is the following: When applying the rule, the morphemes already found must satisfy the LMC context, and the morphemes to be found must satisfy the RMC context. Examples of two level rules are the following:

```
Rule identity: {
  [] [X] [] opt [] [X] []
  {X not in [+,#]}
  [] []/* No Morphotactic Context */
}
```

```
Rule a_absortion: {
  [] [a] [#] <= [] [e,+,a] []
  [] [adj + fem]
}
```

Note how in rule *a_absortion* the incorrect Catalan form *menyspra* (*menyspre* (verb_stem, eng: *despise*) + *a* (3rd,present)) is not recognized nor generated.

The analyses found by the application of the rules must be allowed by the Orthographical and the Morphotactical Contexts.

4.3 The Word Grammar

The Word Grammar is kept as independent as possible with regards to the morphographemic component. It builds a word structure out of the lexical decomposition found by TLRs, and in principle it does not need information on which rules have been applied. For reasons of efficiency, a partial word grammar that rules out wrong partial segmentations could be very useful (this has also been suggested, among others, by (Trost, 1991)).

4.4 Comments on the expressivity of SEGMORF

The expressivity in *SEGMORF* is limited by efficiency considerations. Although efficiency was not the ultimate criterion in our case, we aimed at an implementation that is as much efficient as is reasonable. And this seeking for efficiency has led to restrict the introduction of complementary tools to the basic formalism. For example, for a particular set of phenomena (as the determination of the quantity of syllables to the end of the word, which is relevant for the presence of the graphical accent in Catalan/Spanish) it might be useful to allow the use of regular expressions within the rule format. But the same effect of determining the presence of an accented syllable can be obtained if groups of vowels are identified; since groups of vowels can be lexically marked and are predictable from the word form, we

decided not to incorporate the use of regular expressions within TLRs. Similarly, for the use of morphosyntactic information of the lexical entries, we decided not to incorporate the use of types. Given the simplicity of the resulting WG, which can be easily modelled using a PATR-like grammar, it seemed to us that the introduction of types would turn *SEGMORF* into an inefficient system in terms of compilation and run times.

5 CONCLUSIONS

We have presented a new two-level formalism that allows the user to express in an alternative way morphological phenomena involving interaction between TLR's and the WG. We have also shown that our approach keeps the WG as modular as possible, thus simplifying its construction. *SEGMORF* has been used in the implementation of a wide-coverage morphological analyzer for Catalan.

References

- CEC. 1994. The Alep Linguistic System.
- Bear, J. 1988. Morphology with Two level rules and Negative Rule Features. In *Proceedings of the 12th International Conference on Computational Linguistics (COLING-88)*, Budapest, Hungary
- Carulla, M. and Oosterhoff, A. 1996. El Tratamiento de la Morfología Flexiva del Castellano mediante Reglas de dos Niveles en una Gramática de Unificación. In *Actas del XII Congreso de la SEPLN*, pages 72-80, Sevilla, Spain.
- Evans, E., Kiraz, G. and Pulman, S. 1996. Compiling a Partition-Based Two-Level Formalism. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-96)*, Copenhagen, Denmark
- Kaplan, R. and Kay, M. 1994. Regular models of phonological rule systems. In *Computational Linguistics*, 20(3):331-78
- Karttunen, L., Kaplan, R.M. and Zaenen, A. 1992. Two-Level Morphology with Composition. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, Nantes, France
- Karttunen, L. 1994. Constructing lexical transducers. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-94)*, pages 406-411, Kyoto, Japan.
- Kiraz, G. 1994. Multi-tape two-level morphology: a case-study in Semitic non-linear morphology. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-94)*, pages 180-6, Kyoto, Japan.
- Koskenniemi, K. 1984. A General Computational Model for Word-form Recognition and Production. In *Proceedings of the 10th International Conference on Computational Linguistics (COLING-84)*, Stanford, U.S.A
- Pulman, S. 1991. Two level morphology. In Alshawi et. al, *ET6/1 Rule Formalism and Virtual Machine Design Study*, chapter 5. CEC, Luxembourg
- Ritchie, G., Black, A., Russell, G., and Pulman, S. 1992. *Computational Morphology: Practical Mechanisms for the English Lexicon*. MIT Press, Cambridge Mass.
- Trost, H. 1990. The Application of two-level morphology to non-concatenative German morphology. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING'90)*, pages 371-376, Helsinki, Finland.
- H. Trost. 1991. X2MORF: A Morphological Component Based on Two-level Morphology. Research Report DFKI-RR-91-04, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Saarbrücken, Germany.