

# CATMORF: Multi two-level steps for Catalan morphology

**Toni Badia**

IULA

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

**Àngels Egea**

IULA

Universitat Pompeu Fabra  
La Rambla 30-32, Barcelona  
Catalonia, Spain  
ANGELS@slc.ub.es

**Antoni Tuells**

IULA

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

## Abstract

In computational morphology the two-level paradigm is regarded as a standard; in this paper we describe *CATMORF*, the first wide-coverage multi two-level steps morphological analyzer for Catalan which has been implemented in *SEGMORF* - a two-level morphological formalism. We discuss the benefits and drawbacks of the application of the two-level paradigm to Catalan morphology and compare our results to those obtained for other Romance languages, like Spanish. We therefore put forward a slightly different two-level framework - the Multi two-level steps framework - for dealing with Catalan and Spanish morphology. The paper also illustrates the acquisition of lexical entries for the analyzer's lexicon out of a machine readable dictionary (MRD). What made this task not so trivial is also reported.

**Keywords:** morphological analysis, lexicons

## 1 Brief Introduction

Within computational morphology, the two-level paradigm, either based on that introduced by (Koskeniemi, 1984) (see also (Ritchie et al., 1992) and (Kaplan and Kay, 1994)) or on variants (for example, (Pulman, 1991), (Kiraz, 1994)), is regarded as something of a standard.<sup>1</sup>

In this paper we present *CATMORF*, the first wide-coverage multi two-level steps morphological analyzer for Catalan which is the central module of a tagger intended to deal with free input. In section 2 we give an overview of the system. Section

<sup>1</sup>For example, the E.C. ALEP morphographemic formalism has been used in the description of 9 EU languages in the LSGRAM project (LRE 61-029).

3 describes the initial internal structure of the analyzer: the formalism in which the two-level rules (TLR) have been implemented is presented, along with some examples of TLRs and the main characteristics of the word grammar (WG) and lexical entries. Section 4 shows how we have constructed the lexicon of the system out of a MRD and the problems we encountered. In section 5 we give a summary of the main characteristics of the system: coverage, number of TLRs, number of lexical entries, etc. Section 6 discusses on the applicability of the two-level paradigm to Catalan morphology and compares our results to those obtained in other Romance languages, like Spanish. We therefore put forward a slightly different two-level framework for dealing with Catalan and Spanish morphology: the Multi two-level steps framework. Finally, further work is presented in section 7.

## 2 An Overview of the system

*CATMORF* is the central module of a tagger which is intended to deal with free input. It operates on texts structurally marked with SGML tags and yields its results as SGML tags attached to every word of the text.<sup>2</sup> The general structure of the system is as shown in fig. 1.

The text-handling module (THM) receives as input the plain text between SGML marks. The primary aim of the text-handler is to recognize those textual items which cannot be handled by *CATMORF*: numbers, dates, proper names, multi-word expressions, abbreviations..., and to assign to them a tag (i.e, the usual pre-process). The text-handler then returns the text with the new codes that it has been able to assign. The input to *CATMORF* is thus the set of textual items to which the text-handler has not been able to assign a tag.

<sup>2</sup>The marking on which we actually work is at paragraph and sentence level.

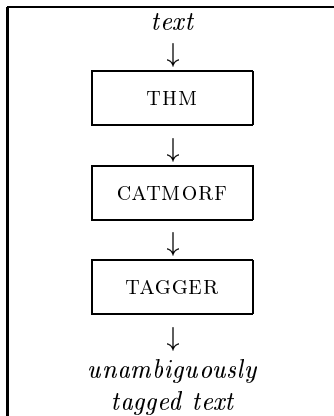


Figure 1: General overview of the system

*CATMORF* is then the second specific module of the system: it assigns as many tags to each word-form in the text as morphological analyses are allowed by its 70000 items dictionary and its two-level and word-grammar rules. The result takes the form of a set of SGML tags.<sup>3</sup> The output of *CATMORF* is returned to the master program which assigns a dummy tag to the still unrecognised words and passes over the tagged text to the third module.

In the current stage of our work the tagger is an adaptation to Catalan of the Multext tagger (Armstrong et al., 1996).

### 3 The internal structure of CATMORF

In this section we present the initial internal structure of the system (figure 2); we also discuss on the formalism we chose for implementing the TLRs and on the main characteristics of the WG and lexical entries. In section 5 and 6 we will discuss on the final internal structure of the system and on what led us to choose it. In the two-level framework, as it is well known, morphographemics is modelled in two-level rules (TLR) and morphotactics either in continuation classes or in unification word grammars (WG). For the first approach, very efficient systems exist (Karttunen et al., 1992) and (Karttunen, 1994), though the latter provides more elegant morphosyntactic parsing, as shown in the work by Trost (Trost, 1990) (see (Ritchie et al., 1992) as well). Our system models morphotactics in a (DCG-like) WG and morphographemics in SEGMORF, a variant of the ALEP morphographemic formalism (Pul-

<sup>3</sup>They basically follow the directions of the Corpus and Lexicon Morphosyntax Subgroups of the Eagles Project (Monachini and Calzolari, 1994).

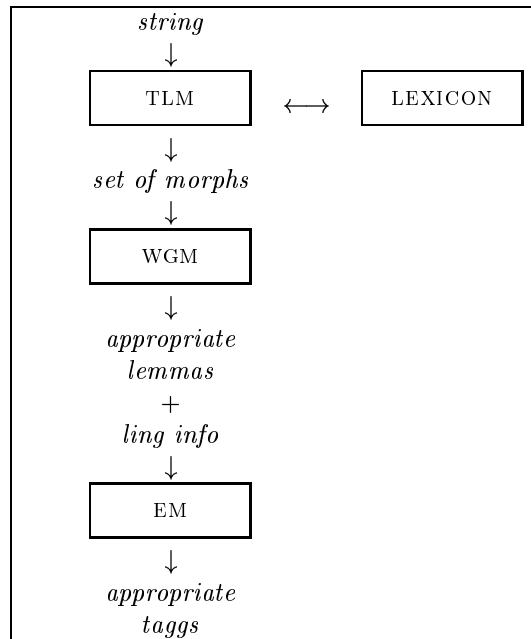


Figure 2: Initial internal structure of CATMORF

man, 1991). Although SEGMORF (Segmorf, 1996) closely follows the strategies used in the Alep morphographemic formalism, there are two aspects that induced us to use this formalism. Firstly, the general architecture envisaged for our system is different from that of the Alep platform; no general parsing is foreseen in our system, whereas in ALEP the word building module is intermixed with the general syntactic parsing, and secondly the expressivity of the Alep system (like that of most current two-level formalisms) does not enable the independence of the word grammar with respect to the two-level rules (Segmorf, 1996).

#### 3.1 The TLR module

This section is devoted to show a few interesting phenomena of Catalan morphology that are going to provide a general overview of the rules and of the expressive power of the formalism in which they have been implemented. TLRs have been implemented in *SEGMORF* (Segmorf, 1996), a two-level morphological formalism which is an extension of the Alep morphographemic segmentation formalism (Pulman, 1991) (also known as a *partition* formalism (Evans et al., 1996)). The main characteristics of the formalism is that it allows the linguist to express the morphographemic and morphotactical contexts constraining the application of TLRs. Typically, the specification of morphotactical contexts within TLRs makes sense when constraining the application

of TLRs to certain classes of stems or when dealing with morphological phenomena which involve interaction between the morphographemic and the morphotactical contexts (see (Segmorf, 1996) for a discussion of benefits and drawbacks of this approach and for further considerations).

Consider the following facts:

- *ample* (adj,sing,masc)
- *ampla* (adj,sing,fem) → *ample* (adj) + *a* (fem)

These examples show the fact that a couple of characters and a morpheme boundary (on the lexical side) are mapped onto a single character on the surface. This phenomenon in addition is restricted to adjectival inflection; just as the following shows:

- *\*menyspra* (v,pres,3rd,sing) → *menyspre* (v) + *a* (pres,3rd,sing)
- *menysprea* (v,pres,3rd,sing) → *menyspre* (v) + *a* (pres,3rd,sing)

The rule that accounts for this fact is the following:<sup>4</sup>

```
Rule a_absortion: {
  [] [a] [#] <= [] [e,+,a] []
  /* Next, morphotactic contexts */
  [] /* Left morphotactic context*/
  [
    lex(_,class(adj),_),
    lex(_,class(gender_suffix),_)
  ]
}
```

This rule is obligatory provided that the lexical contexts are satisfied; that is to say, if there is on the lexical side a morpheme ending with “e” followed by a morpheme “a” (which are an adjective and a feminine suffix, respectively) then on the surface side the word has to end with a single character “a”. Note that this result is easily obtained thanks to the possibility to map single characters onto multiple ones and to specify the morphotactical contexts<sup>5</sup>. Obviously, other rules are more general and do not need to specify morphotactical contexts. For example, the *identity* rule will look as follows:

```
Rule identity: {
  [] [X] [] opt [] [X] []
  {X not in [+,#]}
  [] []
}
```

<sup>4</sup>The analyses found by the application of the rules must be allowed by the Orthographical and the Morphotactical Contexts.

<sup>5</sup>Both these facilities were not available in the Alep formalism, nor in the most known TL formalisms.

### 3.2 The WG

Thanks to the increase in expressivity of the TL-formalism the Word Grammar can be kept very simple. It is a DCG-style grammar, which builds a word out of the morphemes into which the surface string has been divided.

For inflection the WG is divided into two groups: the nominal categories and the verbal ones. Nominal rules apply to nouns and adjectives, and verbal rules apply only to verbs.

The verbal grammar is very simple, since all verb forms are composed of a verbal stem and a single verbal suffix. This results in a single rule, which has the following form:

```
verb(STEM, MS-INFO) -->
  v_stem(STEM,MS-INFO),
  v_suffix(MS-INFO).
```

The nominal grammar is more complex, since nominals may be simply composed of the stem or of the combination of the stem with one or two suffixes (or more, if derivation is taken into account). Thus at least the following three rules are needed:

```
nom(CAT, STEM, MS-INFO) -->
  n_stem(CAT, STEM, MS-INFO).

nom(CAT, STEM, MS-INFO) -->
  n_stem(CAT, STEM)
  n_suffix(MS-INFO).

nom(CAT, STEM, (MS-INFO1, MS-INFO2)) -->
  n_stem (CAT, STEM)
  n_suffix(MS-INFO1),
  n_suffix(MS-INFO2).
```

Note that in this set of rules the control of the way in which the morphosyntactic information is obtained at the word level is only indicative.

As for the other categories, a new set of rules is entered only when a productive way is found of creating words out of other word classes; in the current stage of development a productive rule for creating adverbs out of feminine adjectives by appendig to them the suffix *ment* ('ly') has also been written.

### 3.3 The information to be found in CATMORF's lexicon

Since our system deals with morphotactics using a WG (hence, no continuation classes), the necessary information to be found in our lexicon is the following<sup>6</sup>:

- word form

<sup>6</sup>Obviously, additional information is present in the lexicon entries, but this is not relevant here

- lemma
- the inflection paradigm of verbs, nouns and adjectives (this information concerns both the WG and the TLR components)
- the blocking of rules by several classes of stems. As it is well known, some graphemic changes are optional for some stems but obligatory for others (Bear, 1988), (Trost, 1990); this information needs marking in the appropriate lexical entries.

Usually, word form and lemma coincide, but sometimes it was necessary to lexicalize word forms which are unpredictable (according to our model). An example of a lexical entry is given below:

```
lex(absolut,[rule_t_d_change:n],
    morf(lema(absolut),class(adj),
        flex(gender(yes),number(yes))
    )
).
```

Note that *absolut* blocks the application of rule *t\_d\_change* (which changes the ending “t” into a “d” –see below). It can also be seen that it inflects for gender and number.

## 4 The Construction of the Lexical Component of CATMORF

Since we aimed at building a true wide-coverage or corpus-oriented morphological analyzer, it was clear from the beginning that a large number of lexical entries should be made available to *CATMORF*. No available very large computational lexicons existed for Catalan, so we decided to use as a primary source of lexemes an MRD (a conventional “human-reader-in-mind” dictionary available in electronic form).<sup>7</sup> Though similar work has been already done for obtaining morphological, syntactic and semantic information from MRD’s (see for example (Boguraev and Briscoe, 1989)), the acquisition of the necessary information required by the lexical component of a two-level morphological analyzer poses its own problems, thus making the whole process far from trivial. To make this point clear, we first sketch the type of information contained in each MRD entry. Next, we show what kind of morphosyntactic and spelling variations on inflection information should be obtained from our MRD in order to build a lexicon for a two-level morphological analyzer, and lastly, we show how we obtained that information.

<sup>7</sup>Our MRD is the IEC dictionary (IEC, 1996), which is a recent normative dictionary for Catalan. No examples nor definitions were available to us in its electronic form.

### 4.1 The information found in our MRD entries

The MRD entries specify basically the following information:

- headword
- different part of speech (POS)
- plural (PL) and feminine (FM) fields (both fields mark not only which entries undergo inflection processes but also variations on spelling or inflection)

Additional associated information which might be to do with glosses on use of the word (for example, formal, old, etc. . .) was not available. An example is given below:

```
EN: aperitiu
FM: -iva
CG1: adj
...
CG11: m
CG11: m
```

This entry shows that word *aperitiu* has adjective as main POS, and that it undergoes inflectional processes (feminine is explicitly marked, but plural is implicitly). It can also be seen that *aperitiu* can be used as a noun (fields CG11 and CG12, which correspond to different senses of the word) with no feminine in these cases (otherwise, it would be explicitly marked).

It is also important to point out that the FM field is always present if a noun or adjective inflects for feminine. If not present, then the word can be used both as masculine and feminine. The PL field is always present if the word has an “irregular” plural.

### 4.2 Extracting the relevant information

This section is devoted to show how we obtained information concerning inflection paradigms and blocking of rules semi-automatically. Everything we say in the next subsections about information extraction from MRD concerns only nouns and adjectives. With respect to verbs, verbs belonging to the second and third conjugations were encoded by hand, and verbs belonging to the first conjugation were added to our lexicon with no additional information (no blocking rules had to be specified), since we do not follow the traditional declension patterns of verbs that was found in our MRD.<sup>8</sup>

<sup>8</sup>Around 8000 verbs of the first conjugation were added automatically, and around 3000 of the second and third conjugations (including lexicalized word forms) were added by hand.

### 4.2.1 Extracting the inflection paradigm

In using the MRD as a primary source of lexemes for constructing our computational lexicon, our main concern has been to generate the minimal set of lexical entries whenever possible:

- We have tried to collapse homographs that do not differ in POS in a single lexical entry.
- Since the WG deals with nouns and adjectives in exactly the same manner, we have tried to collapse homographs used as nouns and adjectives into a single lexical entry if they belonged to the same inflection paradigm.

A couple of examples will make these points clear:

```
EN: espanyol
FM: -a
CG1: adj,
CG2: m
CG3: f
...
CG11: adj
CG12: m
...
```

From the entry above it is clear that *espanyol* can be used as a noun and as an adjective with the same inflection paradigm (both POS admit a feminine form). Accordingly, our lexicon contains the following entry for *espanyol*.<sup>9</sup>

```
lex(espanyol,
    morf(lema(espanyol),
        class(nom-adj),
        flex(gender(yes), number(yes)))).
```

In other cases, it was necessary to generate two lexical entries given a single MRD entry:

```
EN: percussor
FM: -a
CG1: adj
CG11: m
...
```

*percussor* admits a feminine form if used as an adjective, but it can only be used in its masculine form if used as a noun. The corresponding two lexical entries in our system are the following:

```
lex(percussor,
    morf(lema(percussor),
        class(adj),
        flex(gender(yes), number(yes)))).
```

<sup>9</sup>This is not the actual entry present in our system; we only show the relevant characteristics.

```
lex(percussor,
    morf(lema(percussor),
        class(noun),
        flex(gen(yes), nom(yes)))).
```

Though the information contained in the MRD entries is represented in a moderately formal way, some entries had to be added to our system by hand. A case in point are those entries which correspond to taxonomic entities (animal and plants species, basically). For historical (lexicographical) reasons they are represented in their plural form, although they admit a singular one. We have added those entries (around 800) to our system in their singular form.

### 4.2.2 Extracting TLR blocking information

It is well known that some morphographemic phenomena (for instance, german Umlaut) are optional for some stems but obligatory for others; i.e, the application of some TLRs is not predictable from the morphographemic context. Since these phenomena are very common in Catalan, and they involve a large number of stems, it seemed reasonable to deduce this information from the MRD, instead of encoding it by hand. For example, “t-d” change is not predictable from the morphographemic context and it affects 5126 adjectives of the MRD:

- (a) *absoluta* (adjective,fem)  $\Rightarrow$  *absolut* (adj) + *a* (fem)
- (b) *perduda* (adjective,fem)  $\Rightarrow$  *perdut* (adj) + *a* (fem)

Another interesting example is the “n” change, which affects 600 entries in the MRD:

- (a) *camins* (noun,pl)  $\Rightarrow$  *camí* (noun) + *s* (pl)
- (b) *agamís* (noun,pl)  $\Rightarrow$  *agami* (noun) + *s* (pl)

Note that in the last example the application of two rules has to be blocked in order to obtain the right inflected form: the accent supression rule and the “n”-change one).

Fortunately, we could obtain this kind of information for all cases out of the FM and PL fields of each entry.<sup>10</sup> FM (femenine) and PL (plural) fields are of the form “-ending”, which gives enough information for knowing whether the rules have to be applied. For the first case, the corresponding endings are the followings:

- *absolut* – *a*
- *perdut* – *da*

<sup>10</sup>We have found 10 cases of TLR blocking information.

In the second case, only the entry for *agami* marks its plural, since it is considered “irregular”. The rest of blocking cases were dealt with in a very similar way; we have (luckily) found that with respect to this type of information the MRD was built in a very homogenous way.

### 4.2.3 Automatic detection of gaps in inflection coverage

We have used the MRD itself as a corpus for detecting gaps in the coverage of the TLR component or for lexicalizing unpredictable (not predictable according to our model) word forms; i.e. in some cases our TLR were incapable of analyzing specific word forms (built out of the Headword field and the PL or FM fields).<sup>11</sup> We then had to choose whether to refine our set of TLR or to lexicalize word forms. In general, we have only lexicalized some (around 150) unpredictable feminine word forms. For example, given an entry for *doctor* (noun,masc), *doctora* (noun, fem) was predictable for our system, and therefore we did not lexicalize *doctora*. Unfortunately, we found other cases to be unpredictable for our system; given an entry for *jutge* (noun, masc), there was no easy way to infer *jutgessa* (noun, fem), so we decided to lexicalize the feminine word form.

In sum, we have acquired the lexicon of our system and developed the set of TLR in parallel, and we have usually preferred to refine our set of TLR rather than lexicalize a large number of word forms.

## 5 Technical details

In this section we summarize the main technical characteristics of our analyzer:

- The system has been implemented in Sicstus Prolog.
- The system covers nominal and verbal inflection fully. A few nominal derivation processes are also covered.
- 114 rules cover nominal inflection; 10 rules cover verbal inflection.
- The WG has 1 rule for verbal inflection and 15 rules for nominal inflection and some nominal derivation processes.
- The original MRD contains 67567 entries.
- Our lexicon contains 70543 entries; 11092 verbs (around 9000 stems and 2000 lexicalized verb

<sup>11</sup>We could only use the headword, PL and FM fields, since no definitions nor examples were available within the MRD entries.

forms), 386 verbal suffixes, 56275 nouns and adjectives, 3 nominal suffixes and 2555 adverbs. The rest of the entries are prepositions, conjunctions, etc.

- Only around 800 nouns and around 2000 verb forms have been added to the system by hand. The rest of the entries (around 60000) have been added automatically.
- The system is currently being used in the analysis of Catalan newspapers.

## 6 Is the two-level paradigm appropriate for Catalan morphology (and in general, for other Romance Languages)? The Multi two-level steps framework

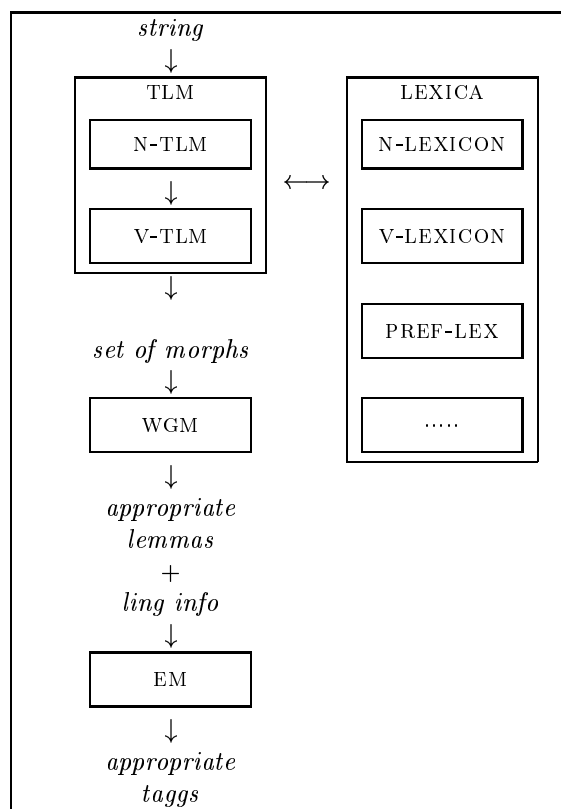


Figure 3: Final internal structure of CATMORF

As shown in the previous section, in Catalan, two-level rules depend on word formation processes. 114 rules cover nominal inflection and derivation processes, whereas only 10 rules cover verbal inflection; that is to say, few rules can be considered as applicable to both inflections (similar results have been

shown for Spanish (Carulla and Oosterhoff, 1996)). In terms of efficiency this result shows that Catalan morphology can be better accounted for in terms of multi two-level steps, which is slightly different than the original philosophy of the two-level paradigm. That is to say, one should have different TLR and WG rule sets, depending on the type of word formation process to cover. For example, given a surface form, first one should try to use the TLR and WG rules for nominal inflection in order to find lexical decompositions. Secondly, TLRs and WG rules for verbal inflection should be used in order to find alternative lexical decompositions. In sum, by separating TLR sets and WG rule sets we aim at narrowing down the search space. This is the approach which has been followed in *CATMORF*.

A case in point is rule `a_absortion`, which we repeat here for expository purposes:

```
Rule a_absortion: {
  [] [a] [#] <= [] [e,+,a] []
/* Next, morphotactic contexts */
[] /* Left morphotactic context*/
[
  lex(_,class(adj),_),
  lex(_,class(gender_suffix),_)
]
}
```

In the initial framework (as depicted in figure 2 above) this rule would initially apply to verbal stems; and its application would have to be rejected later, since it is only applicable to adjectives. In the current, final approach the rule only applies to stems listed in the nominal lexicon (thus restricting its application range). Note that this approach is more flexible than continuation classes; as shown in figure 3 more than one sublexicon is available when dealing with nominal or verbal inflection. The point is that morphemes (prefixes, noun stems, verbal stems, etc.) do not direct to continuation classes (or sublexicons); instead, word formation processes (according to the WG) select their appropriate sublexicons. This can be easily implemented in an object-oriented framework using specialized lexical lookup and TLR application functions. The final internal structure of the system is shown in figure 3. It is worth mentioning that the Multi two-level steps framework does not avoid the specification of morphotactical contexts for those morphographemic changes which involve interaction between TLRs and the WG. It simply specifies that for some word formation processes only a subset of TLRs should be considered. See (Segmorf, 1996) for further considerations.

## 7 Further work

We are currently working on improving the system; on the one hand we are increasing the coverage of the analyzer in order to cover derivation and composition processes. On the other hand, we are implementing a faster version of the analyzer<sup>12</sup>. We are also planning to implement a C/C++ version in the mid-term.

## References

- CEC. 1994. The Alep Linguistic System.
- Armstrong, S.; G. Robert; and P. Bouillon 1996. Building a Language Model for POS Tagging (ms.)
- 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
- Boguraev, B. and Briscoe, T. 1989. *Computational Lexicography for Natural Language Processing*. Longman, 1989.
- 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*, pp. 72-80, Sevilla, Spain.
- Institut d'Estudis Catalans. 1996. Diccionari de la Llengua Catalana.
- 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)*, pp. 406-11, Kyoto, Japan.

<sup>12</sup>Currently, the analyzer works very slowly; it analysis only around 20 words/sec

- 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)*, pp. 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
- Monachini, M. and Calzolari, N. (eds.) 1994 Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Applications to European Languages. Eagles Document.
- 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.
- Segmorf. 1996. . On dealing with morphographemic and morphotactical interaction phenomena in SEGMORF. Submitted to ANLP-97.