CATMORF: Multi two-level steps for Catalan morphology

Toni Badia, Àngels Egea & Antoni Tuells
IULA — Universitat Pompeu Fabra
La Rambla 30-32; Barcelona 08002; Catalunya – Spain
tbadia@upf.es ANGELS@slc.ub.es tuells@upf.es

1 An Overview of the system

CATMORF (Badia, Egea & Tuells, 1997) is the central module of a tagger intended to deal with free input. It operates on texts structurally marked with SGML tags and attaches SGML tags to every word of the text.

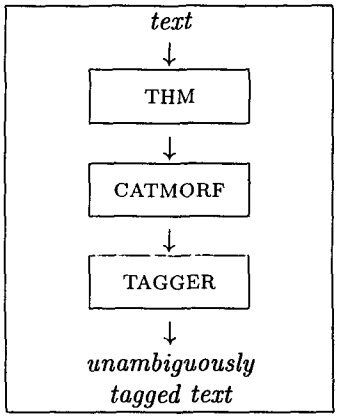

Figure 1: General overview of the system

The text-handling module (THM) receives as input the plain text between SGML marks. It assigns a tag to the textual items that be handled by *CATMORF*: numbers, dates, proper names...(i.e, the usual pre-process). The input to *CATMORF* is the set of textual items to which the text-handler has not been able to assign a tag.

CATMORF assigns as many tags to each word-form as morphological analyses are allowed by its 70000 items dictionary and its two-level and word-grammar rules. The output of *CATMORF*, in the form of a set of SGML tags, 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.

Currently the tagger is an adaptation to Catalan of the Multext tagger (Armstrong et al., 1996).

2 Internal structure of CATMORF

CATMORF's internal structure (figure 2) conforms to the two level paradigm. 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). Our system models morphotactics in a (DCG-like) WG and morphographemics in SEGMORF (Badia & Tuells, 1997), a variant of the ALEP morphographemic formalism (CEC, 1994).

2.1 The TLR module

The main characteristics of the formalism is that it allows the linguist to express both the morphographemic and morphotactical contexts thus constraining the application of TLRs.

Thus a rule in *CATMORF* may make use of the following data structures: the Surface Left and Right morphographemic contexts; the Lexical Left and Right morphographemic contexts; the Morphological Left and Right contexts; and the Application context (i.e, a feature structure which keeps trace of the application of rules and which must unify with the application-FS associated to every morph found in the lexicon).

As is customary the surface and lexical descriptions in rules are related by four types of operators. Note that some of the facilities in SEGMORF were not available in the Alep formalism: the specification of the morphotactical context, the possibility of mapping single characters onto multiple ones, and the ability to cross morpheme boundaries.

2.2 The Word Grammar

Due to the expressivity of the TLRs the WG can be very simple: it is a DCG-style grammar, which builds a word out of the morphemes into which the surface string has been divided and provides the morphosyntactic information at the word level.

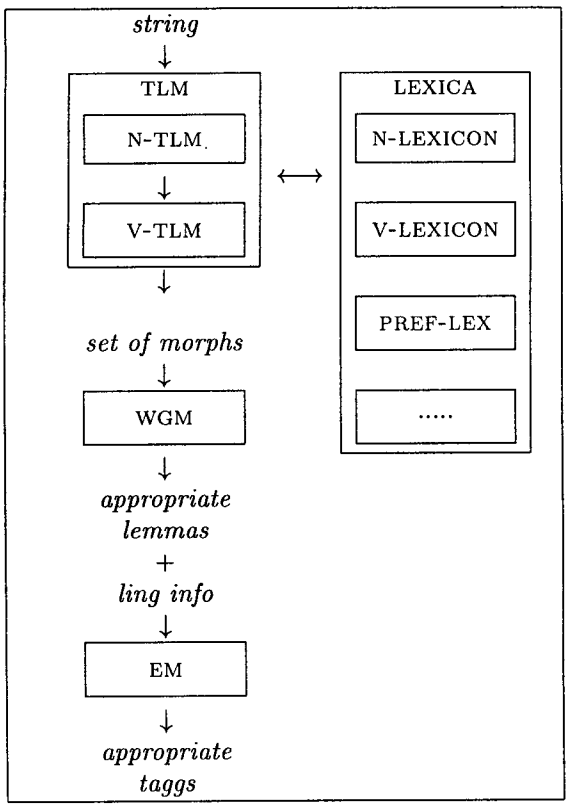

Figure 2: Internal structure of CATMORF

2.3 CATMORF's lexicon

The items in our lexicon contain information on the word form and lemma; the inflection paradigm of verbs, nouns and adjectives (needed for both the WG and the TLR components); and the blocking of rules by several classes of stems.

All this information, including the one concerning the inflection paradigms and the blocking of rules has been obtained semi-automatically from a MRD (a conventional "human-reader-in-mind" dictionary available in electronic form): the IEC dictionary (IEC, 1996), which is a recent normative dictionary for Catalan.

3 Technical details

The main technical characteristics of our analyzer:

- The system has been written 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 processes.
- The original MRD contains 67567 entries. Our lexicon contains 70543 entries; 11092 verbs (around 9000 stems and 2000 lexicalized verb 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.

4 The Multi two-level steps framework

In Catalan, TL-rules depend on word formation processes. 114 rules cover nominal inflection and derivation processes, whereas only 10 rules cover verbal inflection; thus, few rules can be considered as applicable to both inflections.

This shows that Catalan morphology can be more efficiently accounted for in a multi two-level steps framework, in which different TLR and WG rule sets are available, depending on the type of word formation process to cover (as depicted in figure 2). 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 appropiate sublexicons.

Note that this 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 (Badia & Tuells, 1997) for further considerations.

References

CEC. 1994. The Alep Linguistic System.

Armstrong; Robert; and Bouillon 1996. Building a Language Model for POS Tagging (ms.).

Badia; Egea & Tuells 1997. CATMORF: Multi-two-level steps for Catalan morphology IULA Working Paper.

Badia & Tuells 1997. SEGMORF: An extension of the Alep morphographemic segmentation formalism. 3rd Alep User Workshop. Saarbrücken.

Institut d'Estudis Catalans. 1996. Diccionari de la Llengua Catalana.