CATCG: a general purpose parsing tool applied

Alex Alsina, Toni Badia, Gemma Boleda, Stefan Bott, Àngel Gil, Martí Quixal, Oriol Valentín

Universitat Pompeu Fabra
Rambla 30-32
E-08002 Barcelona
{alex.alsina,toni.badia,marti.quixal}@trad.upf.es
{gemma.boleda,stefan.bott,angel.gil}@iula.upf.es

Abstract

This paper focuses on the language processing tool being developed at our centre and briefly describes two of its applications. CATCG, our morphosyntactic analyser, is designed to deal with general written Catalan text. In CATCG the whole processing task has been divided into specific subtasks and for each one of them we try to apply the best strategy available. The most relevant properties of our system are its robustness, the fact that we have given reusability a very high priority, and the goal of acquiring linguistic information by fully automatic means.

The paper is structured as follows: section 1 and 2 explicate and describe the global architecture of CATCG. Section 3 shows the output of CATCG and gives data on its performance. Section 4 describes two projects to which CATCG is being applied: BancTrad and PrADo. Section 5 presents our plans for future work. Section 6 closes the paper with some conclusions.

1. CATCG: A modular architecture

CATCG aims at providing an automatic analysis of free running text in Catalan. It is a modular system that allows the possibility to choose the best strategy available for each specific task. The system has evolved from being a tagging tool to being a partial parser, and will include semantic and pragmatic information in the future.

Our interest is to tag texts with linguistic information, so that operations that are performed on them can be based not only on surface information (basically word forms) but also on their linguistic structure. CATCG is being developed to achieve a linguistic parsing of running text as precise and detailed as possible. It is a general purpose parser, because we have conceived its tag sets as non-task specific as well as model independent. We can think of a wide range of further applications for it (from grammar checking to information extraction).

Deep linguistic analysis has proven to be impossible in one shot, since neither the resources nor the techniques are fully available at one given moment in time. We therefore gave priority to being able to (1) process and extract information from texts from the very beginning; and (2) add new modules if and when they were available.

Figure 1. Architecture of CATCG

2. A description of CATCG

CATCG consists of five modules, most of them divided in several sub-modules (some of which are in turn further modularised, see Figure 1): pre-processing, morphological tag mapping, morphological tag disambiguation, syntactic tag mapping and syntactic tag disambiguation. Each of the modules can be modified without affecting the rest. Furthermore, a progressive improvement of the whole processing can be obtained as new modules are available.

Technically speaking, it incorporates two programming languages: Prolog and Perl; and two formalisms: SEGMORF (an extension of the ALEP formalism, Badia & Tuells 1997) and the Constraint Grammar (CG) formalism (Karlsson *et al.* 1995, Tapanainen 1996). Perl is used both as a glue language and to perform data extraction and detection tasks in the pre-processing and the morphological modules.

The pre-processing phase is performed by a text handler that detects dates, abbreviations, entities and figures. It also verticalises the text and decomposes verb+clitic combinations. Its output is a verticalised text, with mark-up tags for the above mentioned elements as well as for sentences and paragraphs.

The SEGMORF formalism and SWI-Prolog have been used to develop CATMORF, a two-level morphology based analyser for Catalan developed in our group a few years ago (see section 2.1.1 for further details).

The CG formalism was used to develop three constraint grammars: one for morphological disambiguation, another one for syntactic mapping and a third one for syntactic disambiguation. Karlsson *et al.* (1995) states that CG is 'a language-independent formalism for surface-oriented, morphology-based parsing of unrestricted text. [...] The constraints discard as many alternatives as possible [...] with the proviso that no genuine ambiguities should be obliterated'. Therefore it seemed to us that CG fit best for our purposes. Morphological analysis

As for morphology, CATCG employs the tag set proposed in Morel *et al.* 1997, which follows in great part the EAGLES standards. It amounts to ca. 350 tags, though actually only about 200 are being used. The tag set allows for underspecification both in the main categories (noun, verb, etc.) and in the category features (gender, number, aspect, etc.). It also provides partial subcategorisation information for verbs, once their lemmata have been identified. For further details on the tag set, see Badia et al. 2001 and Morel *et al.* 1997.

2.1.1. Morphological tag mapping

This task is realised by a word form dictionary. The dictionary is generated by CATMORF (the old morphological analysis tool), which has been converted into a form generator. The system is actually 10 times faster than it used to be when we used CATMORF as a run-in-time analyser. Besides, we still profit from the advantages of the old analyser, using it to create and update the word form dictionary. The morphological mapping is not context-sensitive (in contrast to the syntactic mapping; see section 2.2.1), so every word form receives all of its possible readings.

CATMORF, written in SWI-Prolog, was the first wide-coverage two-level morphological analyser for Catalan, see Badia *et al.* 1998. It models morphotactics in a (DCG-like) unification word grammar, and morphographemics in SEGMORF, an extension of the ALEP (Advanced Language Engineering Platform) morphographemic formalism. The lexicon was semi-automatically built out of the DIEC (see DIEC), see Tuells 1998 for details. This MRD was based on a recent general purpose dictionary for Catalan. The information extracted was each headword, its part of speech, and the inflectional paradigm of nouns, adjectives and verbs. Around 68000 lexical entries were automatically added this way, and only around 2800 (800 nouns, 2000 verbs) were added manually.

2.1.2. Morphological tag disambiguation

As mentioned above, we have developed a CG-based morphological disambiguation engine for Catalan. DeMCat (Desambiguador Morfològic per al Català, DeMCat) includes over 1000 rules. The basic strategy is to select or remove certain tags according to the constraints imposed by the surrounding context. Rules have been developed on a trial-and-error basis and the development and test corpora have been collected from the web or, occasionally, manually built.

In the rules, there is a TARGET-TAG on which the rule is going to operate. There is also an OPERATOR that indicates whether the target tag is going to be selected or removed. And finally a CONTEXT that specifies the surrounding words and/or tags needed in order for the rule to apply. Context positions are indicated with positive (right of target) or negative (left of target) integers. Zero is the target word (usually a set of words or tags).

The CG formalism provides also other devices such as Kleene's star, the possibility to work with relative or absolute positions, and careful modes to control the rule application. It also makes it possible to use heuristic disambiguation by means of weighted rules. For example, this Rule 1 states that the reading Pron(oun) must be removed from words that can be read as Det(erminant) and have a Prep(osition) right at their left side. Of course, using Kleene's stars, careful mode and related contexts can make such rules pretty complex.

(Rule 1)
REMOVE (Pron) IF (0 DET) (NOT -1C PREP);

Up to now there are still some remaining ambiguities that DeMCat does not resolve. These concern mainly:

- participle readings vs. adjective or noun ones
- conjunction vs. reflexive or relative pronoun readings of "si" (if/whether) and "que" (that), respectively
- finite verb readings vs. noun readings

For further details on its output and performance see section 3.

2.2. Syntactic analysis

The syntactic analysis provides each word with a tag indicating its syntactic function: it can be either a syntactic function at the sentence level (like subject, object or main verb) or function dependent on a lexical head (like noun modifier or determiner). Especially this second kind of tag can lead to sort of unrealistic tags. For instance, modifiers of prepositions would be heads of a constituent introduced by a preposition, independent of their morphological tag. That is, they would be assigned a tag such as @<P independently of being a noun, and adjective or a verb (the angle bracket points to the direction in which the phrasal head should be found).

The principal function tag is assigned to the main word: for instance, in a sentence like *El noi és alt* ('the-MASC-SG boy-MASC-SG is tall-MASC-SG'), it is the word *noi* that will be assigned the tag @Subj(ect). The word *el* will be assigned @DN> (noun determiner): we can see here that the angle bracket indicates the idea that *el* depends on the head of a subject phrase.

As for the tag set, it presently consists of 35 items. It has been created following several traditional grammars, (Fabra 1956) and (Badia i Margarit 1994), adapting them to our needs: basically we have tried to make tags more practical and theoretically sounder (according to a modern view of grammar).

2.2.1. Syntactic tag mapping

The syntactic tag mapping, for which we use a CG-module with 227 rules, is the first stage of the syntactic analysis. In this module, all the possible syntactic functions for a word in a particular context will be projected in accordance with the morphological reading resulting from the morphological tag disambiguation module. For instance, *noi* (in the previous sample sentence) would be assigned the tags @Subj, @Atr, @Advl and @Pred. The morphosyntactic tags available allow us to control the process, so that some impossible ambiguities are avoided (and hence make the following disambiguation task easier). In this case, *noi* would never be assigned the tag @CD (direct object) because the main verb of the sentence is the copulative one *ser* ('to be').

Rule 2 illustrates how controlled mapping avoids unnecessary ambiguity in our example. It states that determiners (DET) are going to be assigned the tag @Subj unless there is a common noun (NOM) at their right. This is coherent with our linguistic approach, in which determiners are heads of NPs unless they specify a noun.

(Rule 2)
MAP (@Subj) IF (0 DET) (NOT *1 NOM
BARRIER Q_MOT/MGN);

It does not apply to the sentence *El noi és alt*, because *el* (a determiner) had *noi* (a noun) at its right side. However, it would fail if there were a preposition or a verb or any other kind of word different from an adjective or a numeral or another determiner (any typical element of a nominal group except for nouns). Without this kind of rules, the determiner, which, as we just saw, can be the head of a NP or a determiner of a head of a NP, would be mapped the @Subj and @DN> tags and the ambiguity would have to be solved in the following CG-module.

2.2.2. Syntactic tag disambiguation

Our CG-based syntactic disambiguation module has 1387 rules. In this module, when possible, one of the syntactic tags mapped in the previous module is selected. The decision is made according to the morphosyntactic and syntactic context of each word. The strategy adopted is to remove as many readings as possible, and rely only on tag selection in very specific and compelling contexts. For this task, we use the morphological information available from the previous steps, together with the progressively obtained syntactic information.

As one might expect, some ambiguities still remain after the application of this module. Some are due to the fact that the module is still under development, some

(Rule 3)
SELECT (@Subj) IF (0 NOM) (NOT *1 SUBJ)
(NOT *-1 SUBJ);

others are due to limitations of the formalism, because it is a surface-oriented approach and can simply not completely deal with constituency. An example of this kind of ambiguity is the systematic ambiguity between direct object and subject, as will be explained in section 3.

Rule 3 exemplifies the kind of rule that builds up this module. It states that a noun should be selected as subject if it has @Subj tag, and no other elements of the sentence are candidates for this function. This is the case of *noi* in *El noi és alt*, since neither *el* nor *alt* can play such role.

3. CATCG: output and performance

Figure 2 gives an example of the input and output of our system. The columns list word form, lemma, part of speech tag, complete morphological information in a compressed tag, and syntactic function (in order of appearance). The results are shown in a tabular format for clarity of exposure.

This example shows several things. First of all this sentence is morphologically disambiguated while three of the syntactic functions remain ambiguous. This reflects the more advanced state of our morphological disambiguation tool. In addition, morphological disambiguation is easier to achieve. The systematic ambiguity between direct object (CD) and subject (Subj) is due to the fact that Catalan has a relatively free order in the realisation of arguments and allows for topicalised objects and post-verbal subjects. For this reason it is not sufficient to assign the subject function to the preverbal NP.

Further on, the preposition *contra* cannot be disambiguated between a noun-dependent and a verb-dependent reading. PP-Attachment remains a systematic problem within Constraint Grammar. Usually purely

La	fi	de	la	guerra	va	suposar	la
the	*end*	*of*	*the*	*war*	*AUX-ed*	*entail*	*the*
	fi	de	la	luita	contra	el	règim
	end	*of*	*the*	*fight*	*against*	*the*	*regime*

```
<s id="1">
la        el        Det     AFS      DN>
fi        fi        Nom     N5-6S    CD_Subj
de        de        Prep    P        <NA
la        el        Det     AFS      DN>
guerra    guerra    Nom     N5-FS    <P
va        anar      Verb    VDR3S-   VAux>
suposar   suposar   Verb    VI----   VPrin
la        el        Det     AFS      DN>
fi        fi        Nom     N5-6S    CD_Subj
de        de        Prep    P        <NA
la        el        Det     AFS      DN>
lluita    lluita    Nom     N5-FS    <P
contra    contra    Prep    P        <NA_Advl
el        el        Det     AMS      DN>
règim     règim     Nom     N5-MS    <P
.         .         .       .        PT
</s>
```

Figure 2. Input and output of CATCG

syntactic information is not sufficient to determine the function of PPs. A PP-disambiguation module on the basis of lexical semantics is currently in development (Badia *et al.* 2001).

In contrast to *contra*, the preposition *de* could be successfully disambiguated in the example above because of the lucky coincidence that *de* is usually noun-dependent unless it appears right after a verb. Another lucky coincidence is that this is the most frequent preposition in Catalan. At present most problems for syntactic disambiguation still occur in co-ordinate and subordinate clauses.

The technical evaluation data of our system are summarised in Table 1. We weighted precision and recall equally ($\alpha = 0.5$) in the calculation of the F-measures. Note that the recall scores are much higher than precision scores which means that there remains a certain amount of ambiguity and that the disambiguation that was carried out discarded a few truly correct readings. It should further be noted that this is both meant and entailed in our approach. Remember we rather guarantee textual ambiguity than a one tag reading.

	morphological disambiguation	syntactic disambiguation
precision	0.83	0.71
recall	0.98	0.93
F (α= 0.5)	0.90	0.80

Table 1. Recall and precision for CATCG

However, for the accuracy estimation, we assumed the target is only one correct tag per word form (and did not count genuine ambiguity in the manual evaluation of morphological tag). The amount of syntactic ambiguity was measured automatically and at present we have no reliable measure figures about genuine syntactic ambiguity. As a consequence the real precision should be slightly higher than the value given in Table 1.

There is still a considerable potential to improve the scores at both levels. Parts of the disambiguation rule module are still incomplete and some of these areas affect common phenomena like subordinate conjunctions and clitics (pronouns). Syntactic disambiguation is harder to achieve, but it is also the part of the system which is least developed. In addition, many of the wrongly assigned tags (false positives) provided by the morphological disambiguation module cause further errors in the syntax one. An improvement in morphological tagging, which is still possible to a certain extent, will also reduce the ambiguity at this level.

The average speed is 1440 words per second on a Linux server with a Pentium III 733 MHz processor. We are confident that we can increase this speed considerably by changing some of the algorithms used in the pre-processing phase.

4. CATCG applied

4.1. BancTrad: a web interface to parallel annotated corpora

The goal of BancTrad[i] (see Badia *et al.*, *in preparation*) is to offer the possibility to access and search through parallel annotated corpora via the Internet. This helps students in our School in Translation and Interpreting looking for parallel texts or for evidence of previous translation decisions. Of course, other uses of BancTrad, such as research in translation theory, discourse studies, or (cross-)linguistics, are easily conceivable.

The languages we work with are Catalan, Spanish, English, German and French. Queries are possible from any of these languages to Spanish and Catalan and vice versa. The texts in the corpus have both extra-linguistic information (such as genre, type of text or topic) and linguistic information (tags indicating lemma, part of speech, etc.).

The web interface of BancTrad allows queries on both kinds of annotation, and this in three expertise levels, from simple string queries to expert ones. In fact, the major advantage of the interface is to allow for an intermediate query level, in which the user can search by lemma, POS tag or syntactic function, without need to know about the internal query syntax[ii] or about the tag set actually used. The user only has to type a form or lemma, or else choose a POS tag or syntactic function from a list (containing standard names such as "Verb", "Preposition" or "Subject", "Object"), and an external interface program (based on the common gateway interface, CGI) interfaces with a server providing a table with the matching contexts.

The use of a unique interface and a unique query language (CQP; see note ii) allows users to have access to corpora that might have been parsed with different techniques, or even to have access to monolingual corpora in a familiar interface. For instance, we are now using freely available stochastic parsers for German, French and English (developed with the TreeTagger software, see Schmid 1995, 1997), but we could switch to others if needed/possible. Another example, the task we are currently undertaking of making available through BancTrad two freely available corpora: the BNC and the Frankfurter Rundschau corpus.

The role CATCG plays in this project is, of course, to perform the linguistic tagging of the Catalan texts in the corpus. Up to now, Catalan is the only language which, in addition to lemma, POS tag and morphological features, such as gender and number, receives syntactic information, in the form described in section 2.2. Its surface-based and word-oriented tagging makes it adequate for the characteristics of the interface, both externally (the actual form that the user fills in) and internally (the interaction with CQP).

4.2. PrADo: an environment for the preparation of electronic documents

Another project in which CATCG is currently being applied is PrADo[iii], a project oriented to develop two grammar checker prototypes for Spanish and Catalan and to establish a linguistic and computational framework which will allow the further development of the prototypes into style checkers.

The grammar checkers will be specially focused on linguistic interferences between Catalan and Spanish (differences in subcategorisation frames, clitics, etc.) and between each of them and English (mainly false friends, both lexical and structural). Of course, the role CATCG plays in this project is to parse the documents to be corrected. The correction module (still to be implemented) will act upon this output and will perform two kinds of operations: either modify the text or issue a warning message, depending on the degree of certainty the system attains regarding a particular mistake.

Up to now, the main effort has been in evaluating existing grammar checkers, in developing CATCG and in elaborating a user model which takes the interferences mentioned above into account. As for the Spanish checker, it will be built based on the model of the Catalan one.

5. Future work

Different actions are foreseen in order to continue our work on CATCG, which are intended to (1) improve the accuracy of the tool, (2) augment its linguistic capacities, and (3) broaden its application range.

To start with, the accuracy of CATCG can be improved by modifying the rule files when needed, as well as by optimising the lexical information available to the rules. We have started preparing a database in which all sorts of lexical information (POS, morphosyntactic features, subcategorisation and even semantic information) can be appropriately stored and kept. Some remaining ambiguities may be resolved by means of statistical techniques, which could apply only when needed after the CATCG has performed all the linguistically sensitive actions.

We are also currently carrying out research in order to take advantage of feedback techniques combining electronic resources (dictionaries, lexical databases, ontology) and corpora tagged with morphosyntactic information. The first step in this direction is the

development of a specific PP-attachment disambiguation module, which is going to be attached to the CATCG and improve one of the areas in which its performance is quite poor.

We are investigating the augmentation of the linguistic capacities of CATCG in two directions. On the one hand we are developing a strategy for combining the shallow lexical morphosyntactic tagging explained with phrase structure syntactic parsing, which reflects constituency and dependency, see (Badia & Egea 2000). To this end we are implementing a unification-based grammar that takes as its input the output of CATCG.

On the other hand we are starting to explore the possibility of introducing semantic information directly to the CATCG output, in order to improve its performance in tasks such as information retrieval.

Eventually, we are thinking of further applications for the resulting grammar, such as Machine Translation, Information Retrieval, or its adaptation to a speech recognition system.

6. Conclusions

CATCG is a highly modular parsing architecture that is currently used to automatically tag large corpora with partial morphosyntactic information. This architecture reveals the advantages of modularity and reusability. Furthermore there is still room for improvement by (1) tuning the grammar rules and (2) optimising the processes to enhance speed.

One of our main goals to was to be able to exploit linguistic information as soon as it was available. This is reflected in the different stages of development of the various modules: the preprocessing and morphological mapping modules are finished and in an optimisation phase; the morphological disambiguation module is currently a revised and tuned; and the syntactic module is to be completed by the end of summer.

Besides, CATCG is already being used for purposes other than the tagging task itself. On the one hand, it is used to annotate a parallel text databank, BancTrad (so that users can perform equivalence queries by restricting word forms, lemmata, POS tag or syntactic function). On the other hand, the project PrADo will benefit form it, since a style checking tool is unable to detect mistakes without morphosyntactic information.

Finally, we definitely favour the continuation and enhancement of the tool by creating further modules that might improve or extend the kind of NLP tasks CATCG can be implemented in.

7. References

Badia i Margarit, A.M. (1994) Gramàtica de la llengua catalana: descriptiva, normativa, diatòpica, diastràtica. Barcelona: Enciclopèdia Catalana.

Badia, T. & T. Tuells (1997) On dealing with morphographemic and morphotactical interaction phenomena in SEGMORF. In Proceedings of the 3rd ALEP user group workshop. Saarbrücken, 1997

Badia, T., À. Egea & T. Tuells (1997) CATMORF: Multi-two level steps for Catalan morphology. In Demo Proceedings of the Conference on Applied Natural Language Processing. Washington

Badia, T., M. Pujol, A. Tuells, J. Vivaldi, L. de Yzaguirre & T. Cabré (1998) "IULA's LSP Multilingual Corpus: compilation and processing". Presented at the 1st ELRA Conference, Granada, 1998. URL: http://www.iula.upf.-es/corpus/corpubca.htm

Badia, T., Boleda, G., Bofias, E. & Quixal, M. (2001) A modular architecture for the processing of free text. Proceedings of the Workshop on 'Modular Programming applied to Natural Language Processing' at EUROLAN 2001. Iasi, Romania

Christ, Oliver (1994) "A modular and flexible architecture for an integrated corpus query system", *COMPLEX'94*, Budapest

Christ, Oliver, Schulze, Bruno M. and König, Esther (1999) *Corpus Query Processor (CQP). User's Manual*, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Stuttgart

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

Fabra i Poch, P. (1956) Gramàtica catalana (amb Prefaci de Joan Coromines). Barcelona: Teide.

Karlsson, F. *et. al.* (1995) Constraint Grammar: A Language-Independent Formalism for Parsing Unrestricted Text. Mouton de Gruyter: Berlin/New York

Morel, J. et al. (1997) El corpus de l'IULA: etiquetaris. IULA Papers: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona. Sèrie Informes, 18. (2nd edition: revised and amended)

Schmid, Helmut (1995) Improvements in Part-of-Speech Tagging with an Application to German, in *Proceedings of the ACL SIGDAT-Workshop*, pp. 47-50

Schmid, Helmut (1997) Probabilistic Part-of-Speech Tagging Using Decision Trees, in Daniel Jones and Harold Somers, editors, *New Methods in Language Processing Studies in Computational Linguistics*, UCL Press, London, pp. 154-164

Tapanainen, P. (1996) The Constraint Grammar Parser CG-2. Department of General Linguistics, University of Helsinki, Helsinki. Publications, no. 27

Tuells, T. (1998) "Constructing and Updating the Lexicon of a Two-Level Morphological Analyzer from a Machine-Readable Dictionary". In Proceedings of the First International Conference on Language Resources & Evaluation, Granada 1998.

---

[i] This project is running under the auspices of the *Programa d'Innovació Docent* (Educational Innovation Program) sponsored by our university (Universitat Pompeu Fabra) and has also been partially financed by the Spanish Government and by the 2001FI 00582 grant from the autonomous Government of Catalonia.

[ii] The syntax corresponds to that of CQP (Christ 1994; Christ *et al*. 1999), which is the query tool we use for the corpora. It allows full regular expressions and complex queries over several kinds of attributes, and interaction with the web interface through a CGI-module. See the web page of the project: http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

[iii] This project is founded by the Spanish Ministerio de Ciencia y Tecnología (ref. TIC2000-1681-C02-01). It started on January 2001 and will finish on December 2003.