

FrAG, a Hybrid Constraint Grammar Parser for French

ECKHARD BICK

University of Southern Denmark
Institute of Language and Communication
Rugbjergvej 98, DK-8260 Viby J
eckhard.bick@mail.dk

Abstract

This paper describes a hybrid tagger/parser for French (FrAG), and presents results from ongoing development work, corpus annotation and evaluation. The core of the system is a sentence scope Constraint Grammar (CG), with linguist-written rules. However, unlike traditional CG, the system uses hybrid techniques on both its morphological input side and its syntactic output side. Thus, FrAG draws on a pre-existing probabilistic Decision Tree Tagger (DTT) before and in parallel with its own lexical stage, and feeds its output into a Phrase Structure Grammar (PSG) that uses CG syntactic function tags rather than ordinary terminals in its rewriting rules. In a recent test run on Parliamentary debate transcripts, FrAG achieved F-scores of 98.7 % for part of speech (PoS) and 92.6 % for syntactic function tags.

1 CG with probabilistic input

This paper describes a hybrid tagger/parser for French, the *French Annotation Grammar* (FrAG), and presents preliminary results from ongoing development work, corpus annotation and evaluation. The core of the system is a sentence scope Constraint Grammar (CG), with linguist-written rules modelled on similar systems for Portuguese and Danish (Bick 2000). However, unlike traditional CG, the system does not compute all lexico-morphological analyses for later disambiguation. Rather, it uses as a point of departure unambiguous PoS/lemma input from a probabilistic Decision Tree Tagger (DTT, Schmid 1994), thus bypassing a labour-intensive step in grammar building and jump-starting the system without a full lexicon. This way, during the first phase of the project, lexicon development could be carried out in parallel with, rather than before the CG rule writing work.

Ordinarily, CG rules select or remove, in a context dependent way, word/token based readings that have been - ambiguously - provided either by the morphological analyser or later tag-mapping CG modules (for syntactic and other higher order

tags). However, confronted with morphological input that is at the same time unambiguous and potentially erroneous, FrAG's first CG-module employs *replacement* rules to correct possible PoS errors made by the probabilistic module, and *mapping* rules to add further "lexical" categories (like auxiliary/main verb, or adjectival/verbal status for participles).

In the current phase of the project, a full lexicon look-up was also added as a second stage, and all PoS-readings are now enriched with inflexional information, as well as – where available - valency potential and semantic prototypes (e.g. <Hprof> profession, <Aorn> bird, <food>, <tool> etc.).

lexemes with information on	
PoS, paradigmatical	58.700
verbal valency	6.218
nominal valency	230
semantic class (nouns)	13.781

Table 1: Lexical information types

At the same time, inflexional analysis and lexicon look-up are used to introduce alternative second readings in the case of nominal-verbal ambiguity, participle ambiguity, sentence initial upper case words etc., relying on the DTT-tags as (statistical) preference indicators rather than absolute, unambiguous tags, and allowing context based disambiguation rules as a supplement to existing category replacement rules.

2 Constraint Grammar Syntax

FrAG's second, syntactic level of analysis is a classical Constraint Grammar, consisting of currently 1266 context sensitive mapping and disambiguation rules, where each token is assigned a function tag like subject, auxiliary, predicative etc., in combination with a shallow "directional" dependency arrow (e.g. @ACC> for fronted direct object). Subclause function is tagged on head verbs (e.g. @FS-N< for a postnominal (relative) finite subclause). A typical CG rule, implementing the uniqueness principle, would for instance discard direct object readings to the right of a verb,

if there already is a (safe) pronominal, relative or interrogative direct object to the left of the verb. An example of a more semantically inspired rule is the selection of a subject tag for a noun of the semantic prototype “human professional” <Hprof> before or after a speech-verb without interfering clause boundaries.

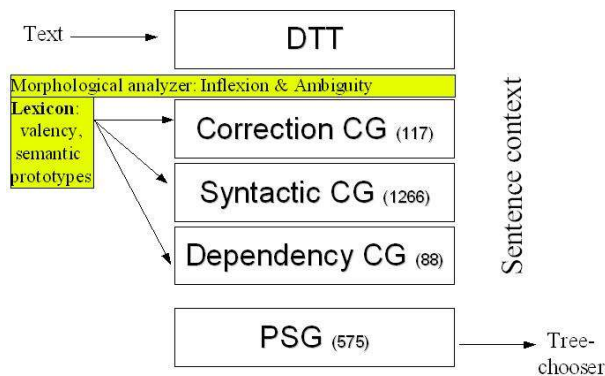


Fig. 1: A modular grammar

3 Tree structures

Like its morphological input-side, the top end output-side of FrAG's Constraint Grammar core uses hybrid methods as well, feeding its tags into an add-on phrase structure grammar (PSG) to generate syntactic tree structures (Fig. 1 and 2), a technique originally suggested for Danish and English in (Bick 2003), and now employed in a growing number of treebank projects.

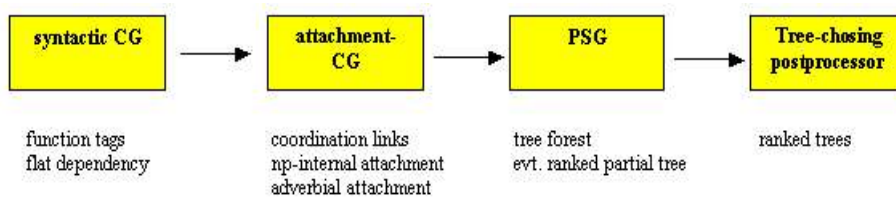


Fig. 2: From CG to Treebank

Instead of words, the French PSG uses syntactic CG function tags as terminals, in conjunction with certain CG-mapped dependency markers and form/PoS attributes. This way, the PSG module does not need a lexicon, is more language independent and ultimately profits from the robustness of the CG stage. Since Constraint Grammar underspecifies certain dependencies (e.g. of postnominal, non-adverbial pp's), and treats coordination in a flat way, an intermediate CG module was added in order to limit structural ambiguity ("forest size"), adding information about exactly which type of heads coordinaters coordinate, and whether to choose close or long attachment for postnominal dependents.

.....
Od:fcl
=S:np
==DN:art('le' <def> F S) *La*
==H:n('télévision' F S) *télévision*
==DN:fcl
===Od:pron-rel('que' <rel> INDP ACC) *que*
===S:pron-pers('nous' PERS 1P nC) *nous*
===P:vp
====Vaux:v-fin('avoir' PR 1P IND) *avons*
====Vm:v-pcp2('proposer' F S AKT) *proposée*
===fa:pp
===H:prp('à' <sam->) *à*
===DP:np
====DN:art('le' <-sam> M S) *le*
===H:prop('CSA' M S) *CSA*
=P:vp
==Vaux:v-fin('être' FUT 3S IND) *sera*
==Vm:v-pcp2('mettre' F S PAS) *mise*

Fig. 3: VISL source format (compatible with PENN and TIGER treebank formats)

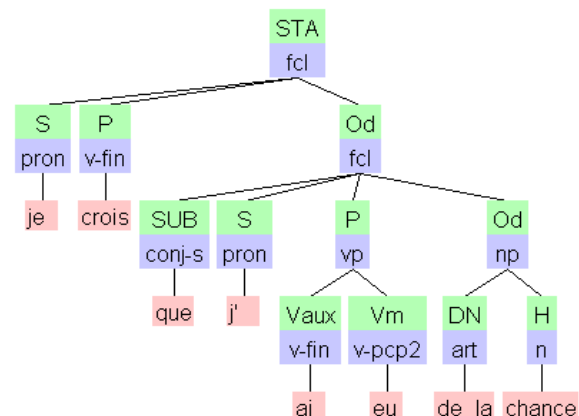


Fig. 4: VISL graphical format (adapted for teaching purposes)

Finally, a tree-chooser program ranks complete trees, adding negative and positive weights¹ to specific tags and structures in an attempt to judge, for instance, coordination depth, discontinuity, argument closeness etc.

4 Evaluation

Since Constraint Grammars are labour intensive and improve incrementally, it is too early (March 2004) for a comprehensive evaluation of the system. Current work on the Europarl corpus²

¹ These weightings are, for the moment, linguist-assigned preference ratings rather than statistical derived probability indices. At a later stage, information from FrAG-annotated corpora could be fed back into the system to bootstrap probabilistic markers as such.

² European Parliament debate transcripts, jf. chapter

suggests, however, a robust performance at both the CG- and PSG-levels. Thus, in from-scratch automatic runs without intervening revision, the system produces 40% complete PSG trees for entire sentences, though of course the vast majority of individual noun phrases or subclauses will be correctly chunked even in trees with incomplete global analyses.

In order to measure tagging accuracy, a chunk of 1.790 words from the Europarl corpus was automatically analysed in a small pilot study and manually evaluated at the CG-level with the following results:

	Recall	Precision	F-score
Part of speech ³	98.7 %	98.7 %	98.7
Syntactic function ⁴	93.7 %	92.5 %	93.1

Table 2: DTT+CG Performance

For a hybrid system, the relative performance of the different modules may be of interest, too. Thus, an inspection of error types showed that the baseline performance of the DTT-stage alone would have given an F-score of 97.5% for PoS. In other words, the added CG correction stage, though also making errors of its own, led to a marked overall increase in PoS recall.

In an earlier evaluation of a more immature version of the system (October 2003) - without a module to add lexical alternatives to DTT-readings - another, larger test run was performed against a newspaper benchmark text (17.500 words, average sentence length 28 words). Here, an F-score of 97.0 was achieved for PoS as opposed to 95.7 for the DTT module alone, translating into a 30% error reduction resulting from the PoS-correction CG. For syntactic function tags (including subclause function, but with a simplified adverbial set), recall was 83.9% and precision 80.0%, corresponding to an F-score of 81.9%.

These numbers, in particular the older newspaper results, are not as good as for other CG's and Finite State Parsers (FSP), which for some languages report syntactic accuracy of over 95% (cf. Chanod & Tapanainen 1997 for French FSP and Bick 2003 for Portuguese/Danish CG), but on the other hand syntactic performance is heavily dependent on correct PoS input, and here the probabilistically based FrAG is still at a disadvantage in comparison with mature, all-

linguist-written CG's, whose morphological modules prepare the field for syntax with PoS F-scores of about or above 99 %.

However (though this will have to be corroborated in further studies), it can be hoped that the increase in performance between the older newspaper run and the recent Europarl run reflects not only the lower degree of structural and lexical complexity in the Parliament transcripts, but also a larger and more mature grammar, as well as the effects of adding alternative lexical/ morphological readings to the DTT-input for later CG-disambiguation.

5 Applications

The applicative context of FrAG, for the time being, is on the one hand internet based grammar teaching (VISL, <http://beta.visl.sdu.dk>), and on the other hand syntactic corpus annotation (<http://corp.hum.sdu.dk>). In particular, the system has been used in a joint project⁶ to annotate French news texts, among these the ANANAS-corpus (Salmon-Alt 2002), which - among other things - targets coreference-research. Part of this material has been revised manually⁷ in tree-bank format and consistency-checked in a tree-viewer (Fig. 4). Apart from this "Botanical Garden", a larger treebank (L'Arboraire/ Freebank) is planned (Salmon-Alt & Bick 2003) and will include also sections with only partial ("Plantation Forest") or no revision ("The Jungle") of the automatic parse. In this context, the French part (28 million words) of the multilingual Europarl parallel corpus (<http://www.isi.edu/~koehn/europarl/>) has recently been annotated with the FrAG parser.

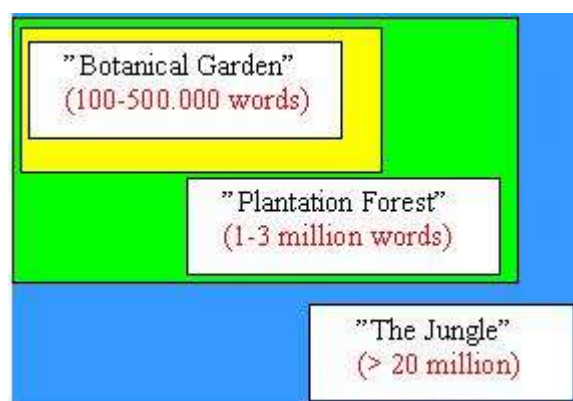


Fig. 5: Treebank revision levels

FrAG's immediate, "native" PSG-format is the VISL-format (Fig. 3), a kind of CG-extension with line based form & function nodes and indentation

³ Separately counting tenses, participles, infinitive.

⁴ Including subclause function, but without making a distinction between free and valency bound adverbials.

⁵ (Schmid 1994) reports 96.36% accuracy for English/Penn-Treebank data.

⁶ A corpus annotation initiative launched jointly by ATILF (Susanne Salmon-Alt, Nancy) and the University of Southern Denmark (the author, Odense).

⁷ Work by Ane Dybro Johansen.

for encoding depth and constituent borders. The format avoids crossing branches by using a special discontinuity notation, marks dependency heads inside constituents and handles, for instance, undefined coordination constructions. VISL's inventory of grammatical categories follows a cross-language standardisation scheme (<http://beta.visl.sdu.dk/visl2/cafeteria.html>) used for teaching treebanks in 22 languages at the University of Southern Denmark. Both GUI tools and format filters are available for end-users, among them TIGER-treebank XML and PENN-treebank bracketing format. The latter has been used as an intermediary stage to create a tgrep-based corpus search interface, which is accessible password-free on the internet. For the CG-versions of FrAG-annotated corpora, a special menu-based search interface has been built targeting "non-technical" users with a linguistic interest only.

The image shows a screenshot of a web-based search interface titled "Search". It consists of two side-by-side panels, each with a set of search filters. At the top of each panel are five radio buttons labeled 1, +, ?, and *. Below these are input fields for "Word:", "Base:", and "Extra:". The left panel has "Extra:" empty, while the right panel has "Extra:" containing "<Hprof>". Both panels have a "Part of Speech" section with a "Neg" checkbox and a list of categories: Noun, Proper Noun, Adjective (checked), Pronoun, Verb, Adverb, and Others. Below this is a "Morphology" section with a "Neg" checkbox and a "Function" section with a "Neg" checkbox. The right panel's "Function" section is expanded, showing a list of functions: Subject (checked), Object, Predicative, Adverbial, Arg. of prep., Adnominal, Apposition, and Adverbial Adject.

6 Outlook

Different schemes for hybridizing the Decision Tree Tagger, Constraint Grammar modules and a PSG module are of course feasible, and should be investigated. Profiting from a growing parsing lexicon, it should be possible to (a) integrate a from-scratch PoS CG with DTT choices to guide heuristic CG-rules, or (b) - assuming the two types of grammars make different types of errors - restrict human revision or specialist replacement rules to cases where the different systems disagree. However, it has to be born in mind that integrating probabilistic methods *between* CG-levels can also *decrease* performance, as reported by Chanod & Tapanainen (1995, p.153) for the statistical Xerox-tagger. Ultimately, it can be hoped, that FrAG-annotated (and, even better, revised) corpora will help to calibrate the interaction between different modules in a statistical way, allowing a task-based choice of methodology, as well as rule weighting and a differentiated way of tag conflict arbitration.

References

- Eckhard Bick. 2000. *The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press, Århus
- Eckhard Bick. 2003. "A CG & PSG Hybrid Approach to Automatic Corpus Annotation". In: Kiril Simow & Petya Osenova: *Proceedings of SProLaC2003*, pp. 1-12. Corpus Linguistics 2003, Lancaster
- Jean-Pierre Chanod & Pasi Tapanainen. 1995. "Tagging French - comparing a statistical and a constraint-based method". In: *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics (EACL'95)*. pages 149-156. Association for Computational Linguistics, Dublin, 1995.
- Susanne Salmon-Alt. 2002. "Le projet Ananas: Annotation Anaphorique pour l'Analyse Sémantique de Corpus". In: *Workshop sur les Chaînes de référence et résolveurs d'anaphores, TALN 2002*, 28/06/02, Nancy.
- Susanne Salmon-Alt, Eckhard Bick, Laurent Romary and Jean-Marie Pierrel. 2004. "La FREEBANK: Vers une base libre de corpus annotés". In: *Proceedings of TALN2004*. Fes, Morocco. (forthcoming)
- Schmid, Helmut. 1994. "Probabilistic Part-of-Speech Tagging Using Decision Trees". In *Proceedings of the International Conference on New Methods in Language Processing*, September 1994. Manchester, UK.
- Chanod, Jean-Pierre & Tapanainen, Pasi. 1997. "A Robust Finite-State Grammar for French." In: *ESSLLI'96 Workshop on Robust Parsing*. pages 16-25. Prague.