The implementation of a computational grammar of French using the Grammar Development Environment

Louisette Emirkanian
Université du Québec à Montréal
emirkanian.louisette@uqam.ca

Lyne Da Sylva
UQAM and
Université de Montréal
dasylva@iro.umontreal.ca

Lorne H. Bouchard
Université du Québec à Montréal
bouchard.lorne_h@uqam.ca

Abstract

The design and implementation of a large-coverage computational grammar of French is described. This grammar is compared to a comprehensive computational grammar of English which was implemented using the same computer workbench. Although many similarities may be observed in the two grammars, there are important structural differences which can be traced back to features specific to the French language, notably agreement and cliticization.

1 Introduction

We present a large-coverage computational grammar of French (CGF)[1] and discuss some implementation aspects of this grammar by comparing it to the Alvey Natural Language Tools Grammar (ANLT Grammar) (Grover et al., 1993) which was also implemented using the Grammar Development Environment (GDE) (Boguraev et al., 1988). The grammar of French was implemented as part of a larger project, the goals of which are the development and application of Generalized Phrase Structure Grammar (GPSG) (Gazdar et al., 1985) in general and the design and implementation of a comprehensive computational grammar for French in particular.

In this paper, we highlight the many similarities between the two grammars but especially differences due to phenomena particular to French. This leads to a description of the major structural differences between the two grammars.

We shall not justify the reasons for our choice of GPSG but will simply point out that GPSG is a formal theory which is unification-based and hence well-suited for computational linguistics.

The Grammar Development Environment is a computer workbench for the development and evaluation of computational grammars of natural language which are described in a style close to that of GPSG. It embodies a modified version of GPSG which is easier to implement than its theoretical counterpart. The GDE was developed as part of the Alvey Natural Language Tools project in the UK.

2 Comparison of the two grammars

The ANLT grammar was used as a model and initial source of inspiration for the design and implementation of the grammar of French, despite obvious differences between the two languages. Both grammars strive to account for the same properties of natural language although they differ on points of detail, the differences being essentially structural.

2.1 Similarities

X-bar theory is used in (Gazdar et al., 1985) to characterize constituent structure. In our grammar, as in the ANLT grammar, X-bar schemata are respected in general, although there are differences of detail. There is no specifier in the verb phrase (VP), thus the V2 immediately dominates the V; complex specifiers in noun phrases (NP) and adjectival and adverbial phrases (AdjP and AdvP) are given special treatment: there are X2 level specifiers of X2 constituents, as for example in:

(1) `N2 → N2[POSS +], H2`
the man's black hat
`N2 → N2[TYPN coll], H2`
une foule de ces étudiants (a crowd of those students)

Furthermore, some constituents have a specifier at level X1 even if there is a specifier at level X2 :

(2) `N2 → Spec, H2`
tous les enfants, all the children
`N2 → Det, H1`
les enfants, the children

Adjectival, adverbial and prepositional phrases are treated in a similar fashion in both grammars.

[1] This research is funded by SSHRCC (40-93-0607) and by FCAR (95ER1198)

Thus in many respects our grammar follows closely the ANLT grammar.

2.2 Differences

The structure of the NP and that of the VP in the CGF differ from those in the ANLT grammar. This is a reflection of phenomena which are characteristic of French: cliticization, present in French but not in English, and agreement, which is more limited in English than in French. We overview these phenomena and then examine the structures.

2.2.1 Agreement

There are many more cases of agreement in French than in English and many more lexical items exhibit agreement features (determiners, adjectives, past participles and conjugated verbs). In particular, two sets of rules are required to account for AdvP and AdjP constituents in CGF, since only the latter is subject to agreement. The agreement patterns are more complex as well. In the VP for example the past participle can agree with a direct object when it is anteposed (3), or with its subject when it is conjugated with *être* (for a non-pronominal verb); otherwise it remains invariant.

(3) la table que Paul a faite (the table that Paul made)
il les a vus (he saw them)

In quantified NPs, we examine two cases:

1. When the subject of a verb is an NP quantified by an adverb or an adjective, the complement of the quantifier determines agreement:

   (4) beaucoup de garçons sont arriv*és* (many boys have arrived(masc. pl.))
   beaucoup de filles sont arriv*ées* (many girls have arrived(fem. pl.))

2. When the subject of a verb is a collective noun, the verb can agree either with the collective or with the element it quantifies:

   (5) une foule de garçons est arriv*ée* (a crowd of boys has arrived(fem. sg.))
   une foule de garçons sont arriv*és* (a crowd of boys have arrived(masc. pl.))

2.2.2 Clitics

Cliticization refers to the phenomenon where a verbal complement can be pronominalized and adjoined to the verb:

(6) Pierre voit le garçon (Peter sees the boy)
Pierre le voit ("Peter him sees")

This phenomenon is interesting because it bears a superficial resemblance to long distance dependencies, which are dealt with by SLASH in GPSG.

3 Structural differences

The data relative to cliticization and agreement have dictated the structure of the French VP and NP.

3.1 The structure of the French VP

The structure for the VP in English developed in the ANLT grammar corresponds to that in (Gazdar et al., 1982): a binary branching structure where each verbal element takes a VP complement of a specific type (the "cascading structure"). The same type of structure for the French VP is generally assumed in transformational or generative grammar and also in GPSG analyses (Miller, 1991). We have departed from these traditional analyses and have implemented a flat structure for the compound tenses, while retaining the cascading one for the passive. For example, the structure for a direct transitive verb is expressed by the rule in (7) and the passive by the rule in (8):

(7) `SV → H[AUX +], V[AUX -], N2`

(8) `SV → H[SUBCAT être], X2[PRD]`

In rule (8), the X2 can be realized as a passive participle or any predicative complement, as in the GPSG analysis of passives.

3.1.1 Arguments for a flat structure

The arguments in favour of distinguishing the two structures are numerous. Many have been discussed in (Abeillé & Godard, 1994) and also in (Emirkanian & Da Sylva, 1995). Directly relevant to our discussion is the fact that only the passive participle may be cliticized (9a),[2] not the participle involved in compound tenses (9b):

(9) a. il est aimé par Marie (he is loved by Mary)
il l'est ("he it is")
b. il a mangé une pomme (he ate an apple)
* il l'a ("he it has")

Thus in the passive structure, the copula behaves like control verbs such as the modal *vouloir* (to want) which take a V2 complement: *il veut partir* (he wants to leave) yields *il le veut* ("he it wants").

Our analysis of the French VP also provides an account of past participle agreement. In GPSG, agreement is handled by the Control-Agreement Principle (CAP).[3] However, we have been unable to account for past participle agreement in French using the CAP (see (Emirkanian et al., in press));[4]

[2] With or without its complements. See (Abeillé & Godard, 1994) for more details.

[3] In the GDE implementation of the grammar of English (Grover et al., 1993), the CAP seems satisfactorily transposed: it is implemented by a relatively small set of propagation rules, which respect the generality of the CAP.

[4] This article shows on the other hand how, for the past participle, the CAP can account for agreement in predicative structures, for example the passive.

the latter account makes use of other devices, such as the Feature Cooccurrence Restrictions and the Feature Specification Defaults.

3.1.2 GDE Implementation of a flat VP

The insertion of the auxiliary into the VP structure to produce a flat VP is done by a metarule. However, implementing this flat structure in the GDE was found to be highly problematic. In the GDE, grammars are pre-compiled into ordered phrase structure rules, and the number of these rules necessary to account for the VP turned out to be extremely large. To give an idea of the size of the resulting grammar, consider the lexical ID rule for a verb requiring an NP and a PP complements:

(10) `V2 → H[3], N2, P2[à]`
Paul donne un livre à Marie (Paul gives a book to Mary)

The following metarules operate on this ID rule: passive[5]; direct object extraction (`SLASH N2`); direct object cliticization (accusative); direct object cliticization (oblique); indirect object extraction (`SLASH P2[à]`); indirect object cliticization (dative); indirect object cliticization (locative); direct object extraction (`SLASH N2[de]`); adverb insertion; auxiliary insertion; supercompound aux insertion; negative adverb (pas) insertion; subject clitic insertion.

Although not all rules are compatible with each other (in particular, no direct object extraction rule can apply to the result of the passive rule), the combinatorics are complex: in a test grammar, we found that the number of phrase structure rules corresponding to this ID rule was of the order of 2^{13}, or over 8000. This is of course unacceptable, as the CGF comprises 45 different lexical ID rules. Not all of them give rise to so many rules, but the compilation time and size of the output grammar made this solution impractical.

Instead, we implemented the structure of the VP as a verbal complex of the auxiliary and the past participle which is sister to the complements. The above metarules apply either to the verbal complex or to the complements, thus reducing considerably the combinatorics. The resulting structure is sufficiently equivalent from a theoretical perspective (i.e., the past participle and its complements do not form a constituent) and it allowed us to implement the bulk of our grammar. The size of our grammar is now as follows: 185 ID rules and 81 metarules. After compilation, these 185 ID rules expand to 2630 expanded ID rules and 3053 phrase structure rules.

Let us now turn to our account of the structure of the NP.

3.2 The structure of the French NP

The data relative to cliticization and agreement have shaped the structure of the NP in various ways. Our study of the French NP has been influenced mainly by the work of (Milner, 1978). We will concentrate here on quantitative structures (such as *beaucoup d'enfants*, many children) and partitive structures (such as *beaucoup de ces enfants*, many of these children). These involve a quantifier (adverb, pronoun, collective or adjective) whose "complement" contains respectively a determiner-less NP with *de* (*de filles*) or an NP with a definite determiner (*de ces filles*).

The ANLT grammar's treatment of partitive NPs such as "all of the children", akin to *trois de ces filles* and also to *beaucoup de filles*, assumes a three-way branching structure, given by the following rule:

(11) `N2[+SPEC] → A2, P[`*of*`], N2[+SPEC]`

This structure must be rejected for French based on data from cliticization: the *de* and what follows must form a constituent, which may be cliticized as *en* like a `P2[PFORM de]` can be :

(12) je parle de Marie (I speak of Mary)
j'en parle ("I of-her speak")
je vois beaucoup de filles (I see many girls)
j'en vois beaucoup ("I of-them see many")

But rather than treat this constituent as a P2, we treat it as an N2 with [PFORM de], because of agreement data: agreement features from the complement must be allowed to percolate upwards to explain possible agreement patterns with it, as in *une foule d'hommes sont venus* (a crowd of men have come). This can be achieved only if the complement is allowed to be an N2 head of the higher NP, and not a P2 (there is no justification for agreement features on a PP in French). (Miller, 1991) also argues in favour of treating *de* and *à* in French as markers on N2, rather than as prepositions, yielding `N2[PFORM de]` and `N2[PFORM à]`. The following rules for a subset of French quantitative constructions highlight the prevalence of this N2:[6]

(13) a. `N2[PFORM nil] → Adv2[+QTE], H2[`*de*`]`
beaucoup de filles (many girls)
b. `N2[PFORM nil] → N2[Coll], H2[`*de*`]`
une foule de filles (a crowd of girls)
c. `N2[PFORM nil] → A2[+QTE], H2[`*de*`]`
trois filles (three girls)
d. `N2[PFORM nil] → H2[`*de*`]`
des filles (girls)

[5] The optionality of the agent gives rise to two ID rules from the output of this metarule.

[6] We have omitted the description of the internal structure of this `N2[`*de*`]`. Our grammar accounts for the absence of an explicit *de* in certain contexts (i.e.,*trois filles*) as well as its presence in extraposed contexts (*j'en ai trois, de filles*) and its contraction with the determiner *des* to produce simply *de* (*de filles - règle de cacophonie*).

The following examples featuring cliticization and dislocated structures, where the *de* reappears, also argue for the N2[*de*]:

(14) a. il voit beaucoup de filles
("he sees many of girls")
il en voit beaucoup ("he of-them sees many")
il en voit beaucoup, de filles
("he of-them sees many, of girls")

(14) b. il voit une foule de filles (he sees a crowd of girls)
il en voit une foule, de filles ("he of-them sees a crowd, of girls")

(14) c. il voit trois filles
(he sees three girls)
il en voit trois ("he of-them sees three")
il en voit trois, de filles
("he of-them sees three, of girls")

We posit an analogous structure for NPs with the determiner *des* in rule (d), i.e., a partitive structure where the quantifier is not specified. Indeed, cliticization in *en* is possible (14d).

(14) d. il voit des filles (he sees girls)
il en voit ("he of-them sees")
il en voit, des filles ("he of-them sees, girls")

In all structures described above, we postulate an N2[*de*] which may be cliticized. In that case, it leaves behind a "stranded" quantifier (except in the case of rule (d), where that quantifier is null).

4 Conclusion

Overall, the coverage of the CGF is comparable to that of the ANLT grammar, although some of the phenomena treated are different. Presently, the CGF can parse isolated sentences using a limited, though representative, hand-compiled lexicon. In order to analyse sentences occurring in real text, a large coverage lexicon is required. Research in progress aims at automating the lexicon building process.

References

Anne Abeillé & Danièle Godard. 1994. *The complementation of French auxiliaries.* Ms. UFRL & CNRS, Université de Paris 7.

Bran Boguraev, John Carroll, Ted Briscoe & Claire Grover. 1988. Software Support for Practical Grammar Development. In *COLING'88*, Budapest, pp. 54–56, John Von Neumann Society for Computing Sciences.

Louisette Emirkanian, Lyne Da Sylva & Lorne Bouchard. In press. L'accord dans une grammaire computationnelle du français. In Emirkanian, L. & L. Bouchard (eds), *Traitement automatique du français écrit*, Cahiers scientifiques de l'ACFAS.

Louisette Emirkanian & Lyne Da Sylva. 1995. *Quelques phénomènes d'accord dans les grammaires syntagmatiques.* Ms. Département de linguistique, Université du Québec à Montréal.

Gerald Gazdar, Ewan Klein, Geoffrey Pullum & Ivan Sag. 1985. *Generalized Phrase Structure Grammar.* Cambridge MA: Harvard University Press, 276 pp.

Gerald Gazdar, Geoffrey Pullum & Ivan Sag. 1982. Auxiliaries and Related Phenomena in a Restrictive Theory of Grammar. In *Language*, 58(3), pp. 591–638.

Claire Grover, John Carroll & Ted Briscoe. 1993. *The Alvey Natural Language Tools Grammar (4th Release).* Tech. Report, 162 (revised), Computer Laboratory, University of Cambridge.

Philip Miller. 1991. *Clitics and constituents in Phrase Structure Grammar.* Doctoral dissertation, University of Utrecht. Published (1992), New York: Garland Publishers.

Jean-Claude Milner. 1978. *De la syntaxe à l'interprétation : quantités, insultes, exclamations.* Paris: Editions du Seuil.