LKB grammar development: French and beyond

Jesse Tseng

A Workshop on

Ideas and Strategies for Multilingual Grammar Development

Emily Bender, Dan Flickinger, Frederik Fouvry and Melanie Siegel (editors)

25–29 August 2003

Taking place during the
Fifteenth European Summer School for Logic,
Language and Information
18–29 August, Vienna, Austria

LKB grammar development: French and beyond

Jesse Tseng
MoDyCo (UMR 7114)
Université Paris X – Nanterre
jesse.tseng@linguist.jussieu.fr

Abstract

This paper presents some aspects of an HPSG-style grammar for French currently under development using the LKB platform, with particular emphasis on a number of technical solutions and linguistic analyses that may be relevant to the implementation of constraint-based grammars of other languages. The main issues discussed are lexical rule management, the implementation of argument composition, and the analysis of phrasal affixes.

The French grammar described in this paper has been in development since early 2001.[1] As is the case for many projects of this type, the grammar was built more or less from scratch, without direct reuse of existing grammar resources. It was however heavily influenced by the developers' previous experiences with other implementation projects for a variety of languages. The objective of this paper is to give a general overview of the current state of the grammar, focusing on a number of specific issues of wider applicability, beyond the monolingual treatment of French.

1 General presentation

We use the LKB platform for typed feature structure-based grammar development (Copestake, 2002). In this discussion we assume that the reader is familiar with this platform and the necessary distinction between a theoretical HPSG grammar (with its almost abitrarily expressive description language) and an HPSG-style grammar implemented with the LKB (or any other development platform). Specific instances of this divergence include: the unavailability of disjunction/negation, the absence of constraints with complex antecedents, the need to specify explicitly the number of daughters in every rule, and to fix their linear order. None of these issues will be addressed here.

In the initial stages, we have concentrated our efforts on a subset of grammatical phenomena of linguistic interest, rather than striving for wide coverage of constructions and lexical items. The grammar currently provides a treatment of the following phenomena of (written) French:

- verbal conjugation and complementation classes
- realization of bound pronominal clitics as verbal affixes (Miller and Sag, 1997), or as independent syntactic words
- argument extraction for interrogatives and relative clauses(Bouma et al., 2001)
- auxiliary and causative constructions allowing "clitic climbing" (Abeillé and Godard, 1997; Abeillé and Godard, 2002)
- past participle agreement (Abeillé and Godard, 1996)

[1] Most of the work reported here was carried out in the research groups TALaNa/Lattice (UMR 8094) and LLF (UMR 7110) at Université Paris 7. The author wishes to thank the members of these groups for fruitful collaboration.

- orthographic alternations reflecting elision, contraction, and liaison (Tseng, 2003b; Tseng, 2003a)

As indicated in the list above, the grammar incorporates the insights of a great deal of theoretical work on the formal analysis of French in HPSG. In fact, the main objective of the grammar at this stage is to evaluate the technical adequacy and the successful interaction of recent theoretical proposals. When a choice between efficiency and theoretical adequacy presents itself, therefore, we generally decide in favor of the more linguistically justified analysis. For example, strict binary branching is a useful constraint on processing, but empirical evidence points to a flat structure for complex VPs in French, and so this is the analysis the implementation produces.

We will discuss three main topics in this talk: our approach to lexical rule management, some problems associated with the implementation of argument composition (for auxiliary and causative constructions), and the analysis of certain "weak" elements as phrasal affixes.

2 Lexical rule application

In many cases, the formal linguistic proposals that the grammar is based on can only be implemented in modified form, since the LKB system does not fully support the considerable expressive power of the HPSG formalism, and because the underlying type logic of the LKB differs from that commonly assumed by HPSG linguists. Consequently, certain constraints that are easily stated in the HPSG description language require a more complex procedural implementation. In our grammar, this is especially apparent at the lexical level: in order to construct the rich lexical descriptions needed to drive the syntactic derivation, the grammar makes use of a large network of lexical rules, most of which have no theoretical counterpart.[2]

[2]In a grammar incorporating the recent construction-based proposals for HPSG (e.g., Ginzburg and Sag (2001)), similar effects would be seen at the phrasal level, with a multiplication of non-branching syntactic rules. At present, the French grammar makes no use of phrasal subtypes.

2.1 Rule ordering

We adopt the partition of lexical objects into the subtypes *lexeme* and *word*; only words can be input to syntactic rules. This allows the divison of lexical rules into three types (*lexeme-to-lexeme*, *lexeme-to-word*, and *word-to-word*) with a rudimentary built-in ordering: *l2l* rules apply first, followed by exactly one *l2w* rule, and then any *w2w* rules. This means that *l2w* rules do not need to be relatively ordered, and they can never apply recursively. For *l2l* and *w2w* lexical rules, on the other hand, both of these problems have to be dealt with.

As an example, consider the inflection of French adjectives for gender and number. In the lexicon, adjectival entries are of type *lexeme*, and they have underspecified GEN and NUM values. These lexemes must undergo GEN instantiation (by one of two *l2l* lexical rules) followed by NUM instantiation (by one of two *l2w* lexical rules). Iteration of the *l2l* gender instantiation rules has to be blocked, for example by adding an ad hoc feature whose value changes so that the output no longer unifies with the input.

This is a very common situation, and so instead of inventing a new feature for every set of *l2l* rules, we use a general list-valued attribute STACK (appearing only on lexemes) to manage lexical rule application. Adjectival lexemes, for example, are listed in the lexicon with a single STACK element, linked to the (underspecified) INDEX value. The presence of this STACK element triggers either the masculine or the feminine GEN instantiation rule, which produces as output a lexeme with a new STACK list, containing just the (underspecified) NUMBER value. This STACK value triggers the singular (or plural) *l2w* NUM instantiation rule, which produces a word (with no STACK feature). The derivation of the feminine plural forms `grandes` is shown in Figure 1.

The STACK feature corresponds to nothing in theoretical HPSG, but it is a practical, transparent way of keeping track of the many "layers" of lexical rules. Technically, the STACK value is a list of objects of any type (including arbitrary strings), and every lexical rule looks for lexemes bearing a specific type of first STACK element. The rule modifies the STACK appropriately to send the output lexeme to the next lexical rule (or in the case

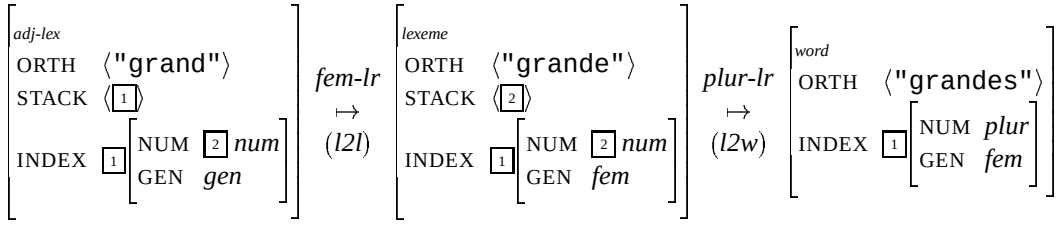

Figure 1: STACK for adjective inflection

of an *l2w* rule, the output is a word and can appear in the syntax without further manipulation).

2.2 List recursion

The STACK is also useful in situations where *l2l* rules do have to apply iteratively for list processing, a very common operation in HPSG. For instance, argument realization in French involves the resolution of each *synsem* object on a verb's ARG-ST list to a specific subtype (*canon-*, *affix-*, or *gap-synsem* for canonically expressed, cliticized, and extracted arguments, respectively). In the LKB grammar, this resolution must be achieved explicitly, with a distinct lexical rule for each *synsem* subtype. We assume that verbal lexemes copy their arguments onto STACK, thereby triggering the application of one of the three kinds of argument realization lexical rule. Each rule resolves the type of the first *synsem* on the input STACK, and makes the appropriate changes to the lexical entry (i.e., the argument is added to the COMPS list, to the SLASH set, or to one of the clitic lists). The resolved *synsem* object is then popped off the STACK list in the output of the lexical rule. When the verb's STACK is empty, all of its arguments have been processed, and the verb passes on to the next set of lexical rules (e.g., clitic prefixation). An example of this procedure applied to the verb *donnent* is given in Figure 2.

Another example of STACK recursion is the treatment of French compound tenses (a form of the auxiliary *être* or *avoir* plus a part participle). The selection of the auxiliary and presence or absence of agreement inflection on the past participle are determined by inspection of the elements of the participle's ARG-ST list. If any of the arguments is reflexive, the the auxiliary must be *être* instead of *avoir*, and if there is a non-canonical (i.e., affix or gap) accusative argument, the participle must agree in gender and number with this argument (Abeillé and Godard, 1996). See (1c) below, for example. In the LKB implementation, again, this kind of list inspection requires a chain of recursive lexical rules. Past participles copy their ARG-ST list onto STACK and undergo a special set of STACK-popping lexical rules that determine auxiliary selection and agreement.

3 Underspecified arguments

For most verbs, the list of arguments in STACK is defined lexically (in direct correspondance to the ARG-ST list) and so the argument realization procedure described in the previous section will eventually terminate, once the STACK list is empty. A problem arises, however, in the treatment of the French temporal auxiliary verbs *être* and *avoir*, and the causative verb *faire*. These verbs have a lexically underspecified ARG-ST list, which is instantiated by the infinitival or participial V complement to the right (Abeillé et al., 1998; Abeillé and Godard, 2002). The result is a complex predicate construction with argument composition, an analysis that allows a straightforward treatment of the phenomenon of "clitic climbing," where a complement selected by a lower verb is realized as a pronominal clitic attached to a higher verb:

(1) a. Jean les voit. 'Jean sees them'
(object realized as clitic on verb)

b. J veut les voir. 'J wants to see them'
(no climbing)

c. Jean les a vus. 'J has seen them'
(clitic climbing onto auxiliary)

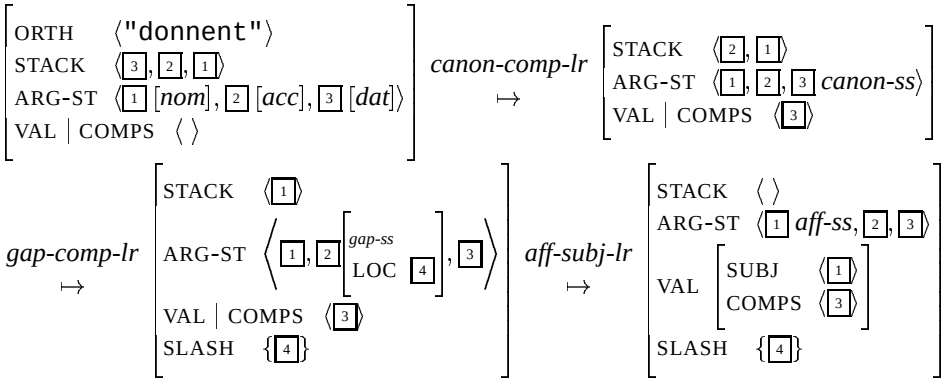

Figure 2: STACK for argument realization

In our grammar, a verb with an underspecified ARG-ST list will also have an underspecified STACK. Like other verbs, auxiliary/causative verbs must undergo argument realization to determine VALENCE, SLASH, and cliticization properties. The problem is that the parser works from left to right, and so it tries to process the underspecified auxiliary/causative verb before identifying its participial/infinitival complement. The application of the STACK-popping argument realization rules to a variable STACK leads to infinite recursion.

One solution would be to modify the parsing procedure to delay lexical rule application until the STACK is instantiated, and meanwhile to let the parser look ahead to identify the following verbal complement.[3] A second possibility would be to prevent the application of the normal argument realization principles, but this is linguistically unmotivated, since auxiliary/causative verbs behave exactly like other verbs in this respect.

At present, we have implemented a lexical rule based solution that simply fixes the length of the underspecified STACK, from zero to four elements. This ensures termination, but it impairs the performance of the parser, because every occurrence of an auxiliary or causative verb triggers the generation of hundreds of chart edges. For any given STACK element, up to 8 different lexical rules can apply, so in principle the number of possible lexical entries generated is $8^0 + 8^1 + 8^2 + 8^3 + 8^4$. In fact, the actual number of entries is much lower, because not all rule sequences are allowed (for example, at most one argument can be extracted, and there are various constraints on clitic cooccurrence). Nevertheless, the grammar is significantly slower when processing any input involving argument inheritance. And it should be noted that the arbitrary upper limit of four inherited arguments on STACK is certainly too low.

This is a part of the implementation that calls for reexamination. The problem described here is by no means specific to French. The same mechanism of argument composition has been proposed for the corresponding constructions in the other Romance languages (Abeillé and Godard, in press), and for the treatment of non-finite complementation in German and Dutch (Hinrichs and Nakazawa, 1994; van Noord and Bouma, 1996; Meurers, 2000).

4 Phrasal affixes

The French LKB grammar incorporates a treatment of certain "weak form function words"(such as pronominal clitics) as bound morphological affixes, rather than syntactically independent words. In the case of clitics, this analysis is relatively easy to implement: an ARG-ST element of a verb, instead of being mapped to a VALENCE list, is speci-

[3] This is the approach adopted in the B8 grammar of German, implemented using Trale; see (Kordoni, 1999; Tseng, 2000) for some documentation. The possibility of "delaying" in the LKB should be explored.

fied as an *affix-synsem* and triggers the application of a lexical rule that realizes the appropriate prefix (or suffix) on the verb (Miller and Sag, 1997). This approach is possible because clitics are always arguments selected by another element, and they are always realized morphologically on this element (the verb).

We extend the affix analysis to another class of "weak" forms in French, including the definite article "le", and the prepositions "à" and "de", adopting and further elaborating the GPSG proposals of Miller (1992). In these cases, the analysis is less straightforward, because the prefix is a functor syntactically and semantically, with grammatical scope over an entire phrase. This is why authors like Miller call these elements "phrasal affixes." At the same time, however, their status as prefixes implies that they must be realized morphologically at the lexical level, attached to a single word. (Any approach based on post-syntactic affixation would be inconsistent with the lexicalist foundations of HPSG.)

4.1 Implementation

The key to the analysis is the dissociation of the morphological realization of the affix (at the lexical level) and the incorporation of its syntactic and semantic interpretation (in the syntax) (Tseng, 2003b). Concretely, a set of *word-to-word* lexical rules allows the prefixation of "le," "les," "de," etc. to more or less any existing word. This is where phenomena such as haplology, liaison/elision, and idiosyncratic contracted forms like "du" and "aux" are dealt with. At this point, the syntactic function and semantic contribution of the prefix are not yet activated. Instead, the presence of the prefix is encoded as a positive (left) EDGE feature that can only propagate from the left-most daughter (not necessarily the head) in a syntactic combination (non-peripheral daughters are constrained to have no positive EDGE specifications). In the syntax, when the appropriate phrase has been constructed, a unary syntactic rule can apply to add in the grammatical effects of the prefix, at the same time "discharging" the positive EDGE specification. A well-formed maximal projection cannot bear any positive EDGE specifications corresponding to uninterpreted prefixes.

The phrasal affix analysis yields rather untraditional structures (see Figure 3). The proposed treatment provides a straightforward account of contracted forms like "aux" (and the ungrammaticality of uncontracted *"à-les") and of elision and liaison phenomena, all of which call into question the more traditional analysis of these elements as syntactic words.

4.2 Further issues

The notion of phrasal affix is useful for many languages besides French. Candidate phenomena include possessive "-'s" in English, and perhaps the "a"/"an" alternation, case suffixes in Japanese, Korean, and Turkish, and consonant mutation effects in Celtic. In general, phrasal affixes are the remnants of syntactic functors (determiners, prepositions, complementizers) that have lost their morphosyntactic independence, and are now bound morphophonologically to the first (or last) word in a phrase.

All of the phenomena mentioned here involve elements that occupy the frontier between syntax and morphology. In many of these cases, the proposed affixal status of a given element may be in conflict with its orthographic status in the written language. For instance, this is true of the non-elided forms of French "le," "à" and "de" (and their contractions): formally they are analyzed as prefixes but in French text they are written as independent words. At the moment, in our grammar, the lexical rule responsible for "aux" prefixation in (ii), for example, produces a single word `aux_anciens` but ideally we would prefer for the output of the rule to be two (orthographic) words. The same remark applies to pronominal clitics, treated as verbal affixes—throughout the Romance languages, in fact (Miller and Monachesi, in press). The problem of syntax-orthography mismatch recurs in many languages, and calls for a general, multilingual solution.

5 Conclusion

Based on the preliminary results of our monolingual French grammar implementation project, we hope to have identified some issues of multilingual relevance, including some ideas for lexical rule management, and a demonstration of an HPSG

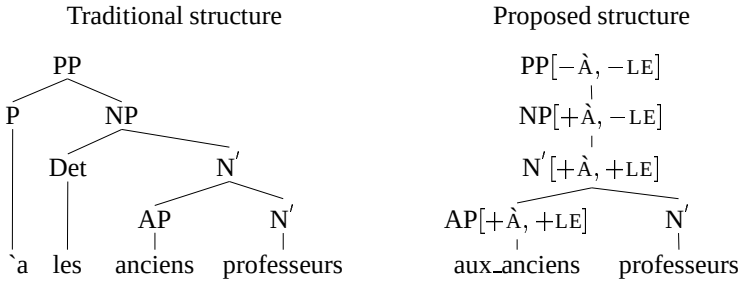

Figure 3: French phrasal affixes

implementation of phrasal affixes. We have also pointed out a few of the problematic aspects of our grammar, since related problems may exist in grammars for other languages, and a multilingual perspective may yield general solutions. It is our hope that the insights gained from our ongoing efforts, and from the discussion generated by this workshop, will be of benefit to similar projects, particularly those devoted to other Romance languages.

References

Anne Abeill´e and Dani`ele Godard. 1996. La compl´ementation des auxiliaires français. *Langages*, 122:32–61.

Anne Abeill´e and Dani`ele Godard. 1997. Les causatives en français: un cas de comp´etition syntaxique. *Langue Française*, 115:62–74.

Anne Abeill´e and Dani`ele Godard. 2002. The syntactic structures of French auxiliaries. *Language*, 78:404–452.

Anne Abeill´e and Dani`ele Godard. (in press). Les pr´edicats complexes dans les langues romanes. In Godard (in press).

Anne Abeill´e, Dani`ele Godard, and Ivan A. Sag. 1998. Two kinds of composition in French complex predicates. In Erhard Hinrichs, Andreas Kathol, and Tsuneko Nakazawa, editors, *Complex Predicates in Nonderivational Syntax*, volume 30 of *Syntax and Semantics*, pages 1–41. Academic Press, New York.

Gosse Bouma, Rob Malouf, and Ivan A. Sag. 2001. Satisfying constraints on extraction and adjunction. *Natural Language and Linguistic Theory*, 19:1–65.

Ann Copestake. 2002. *Implementing Typed Feature Structure Grammars*. CSLI Publications, Stanford, CA.

Jonathan Ginzburg and Ivan A. Sag. 2001. *Interrogative Investigations: The Form, Meaning and Use of English Interrogatives*. CSLI Publications, Stanford, CA.

Dani`ele Godard, editor. (in press). *Les langues romanes. Problèmes de la phrase simple*. CNRS Editions, Paris.

Erhard Hinrichs and Tsuneko Nakazawa. 1994. Linearizing AUXs in German verbal complexes. In John Nerbonne, Klaus Netter, and Carl Pollard, editors, *German in Head-Driven Phrase Structure Grammar*, volume 46 of *CSLI Lecture Notes*, pages 11–37. CSLI Publications, Stanford, CA.

Valia Kordoni, editor. 1999. *Tübingen Studies in Head-Driven Phrase Structure Grammar*. Number 132 in Arbeitspapiere des Sonderforschungsbereichs 340. Universität Tübingen, Tübingen.

Detmar Meurers. 2000. Lexical generalizations in the syntax of German non-finite constructions. Technical Report 145, SFB 340, Universität Tübingen.

Philip Miller and Paola Monachesi. (in press). Les pronoms clitiques dans les langues romanes. In Godard (in press).

Philip H. Miller and Ivan A. Sag. 1997. French clitic movement without clitics or movement. *Natural Language and Linguistic Theory*, 15:573–639.

Philip H. Miller. 1992. *Clitics and Constituents in Phrase Structure Grammar*. Garland, New York.

Jesse Tseng, editor. 2000. *Aspekte eines HPSG-Fragments des Deutschen*. Number 156 in Arbeitspapiere des Sonderforschungsbereichs 340. Universität Tübingen, Tübingen.

Jesse Tseng. 2003a. Edge features and French liaison. In Frank van Eynde, Lars Hellan, and Dorothee Beermann, editors, *Proceedings of the 9th International HPSG Conference*. CSLI Publications, Stanford, CA.

Jesse Tseng. 2003b. Un traitement lexical des affixes syntagmatiques du français. In Bernard Fradin, Georgette Dal, Françoise Kerleroux, Nabil Hathout, Marc Plénat, and Michel Roché, editors, *Les Unités morphologiques / Morphological Units*. SILEX, Villeneuve d'Ascq.

Gertjan van Noord and Gosse Bouma. 1996. Dutch verb clustering without verb clusters. In P. Blackburn and M. de Rijke, editors, *Specifying Syntactic Structures*, pages 1–31. CSLI Publications, Stanford, CA.