FTAG : current status and parsing scheme

A. Abeillé, M. Candito, A. Kinyon
TALANA, University Paris 7
{abeille,candito,kinyon}@linguist.jussieu.fr

Introduction

As far as electronic syntactic resources go, one can distinguish rule-based versus statistics-based grammars, as well as program-dependent versus reusable grammars. Lexicalized Tree adjoning grammars (LTAGs) have been used to develop reusable wide-coverage rule-based grammars for different languages (cf. Doran et al. 1994, 1998 for English, Abeillé 1991 and Candito 1999 for French). We describe here the current status and organization of the French LTAG (FTAG), developed over the past 10 years. The grammar is intended to model speaker competence, and is both application- and domain-independent. It can be used for syntactic tagging, for parsing and for generation. We present a parsing scheme, including a POS tagger, a parser and a parse ranker.

1. Organization of the French LTAG

A LTAG comprises morphological and syntactic lexicons and a vast repository of elementary trees. We present here the general principles on elementary trees, their factorization and an automatic generation tool used for grammar maintenance and development (MG).

1.1 Linguistic principles on elementary trees

We assume some familiarity with the LTAG formalism. We recall that elementary units of an LTAG are lexicalized constituent trees, which encode all the surface constructions available for a given language. Within FTAG, elementary trees respect the following linguistic well-formedness principles: (Kroch and Joshi 1985, Abeillé 1991, Frank 1992, Candito & Kahane 1998):

- Strict Lexicalization : all elementary trees are anchored by at least one lexical element, the empty string cannot anchor a tree by itself.
- Surfacism: an elementary tree encodes all word order variations, all basic syntactic phenomena (passive, extraction...) and crossing of phenomena.
- Semantic Consistency : no elementary tree is semantically void (this ensures the compositionality of the syntactic analysis),
- Semantic Minimality : no elementary tree correspond to more than one semantic unit (modulo lexicalism : lexical anchors are not broken down into morphemes).
- Predicate Argument Cooccurrence Principle (PACP): the elementary tree is the minimal syntactic structure that includes a leaf node for each realized semantic argument of the anchor(s).

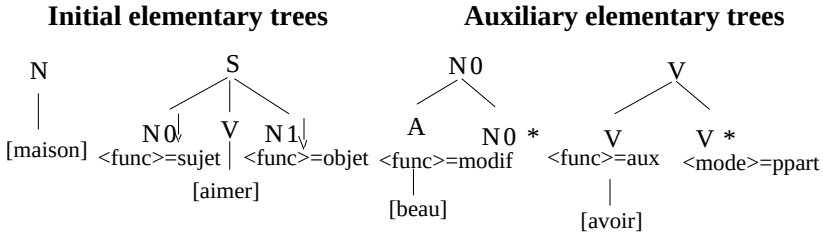


Figure 1. Examples of elementary trees (anchoring resp. *house, love, handsome, have*)

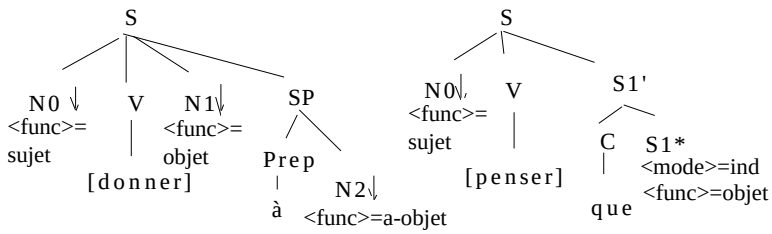


Figure 2. Elementary trees with functional co-anchors

Some examples of elementary trees are shown on Figure 1 [1].

Semantic minimality and consistency imply that functional words appear as co-anchors (cf. Figure 2, the relevant syntactic and semantic units are donner-à (give to) or penser-que (think that)). The elementary trees are combined by substitution or adjunction, and the features of nodes in contact must unify. They thus directly represent all the syntactic rules of the language.

1.2. Factorization of lexicalized elementary trees

Strict lexicalization at execution time does not prevent from internally compacting the common parts of the elementary trees. This compacting is required for any reasonably sized grammar, since for instance a verbal form may anchor dozens or hundreds of elementary trees. In practice, lexicalized elementary trees are compiled out of three sources of information:

- a set of tree sketches ("pre-lexicalized" structures, whose lexical anchor(s) is not instantiated)
- a syntactic lexicon, where each lexeme (or list of lexemes) is associated with the relevant tree sketches
- a morphological lexicon, where inflected forms point to a lemma associated to morphological features

Lexical selection of tree sketches is controlled by features from the syntactic and morphological lexicons, and uses the notion of tree families, grouping sets of tree sketches that share the same (initial) subcategorization frame. The tree sketches of a family show all possible surface realizations of arguments (pronominal clitic realization, extraction, inversion...) as well as all possible transitivity alternations (impersonal, passive, middle..).

A lexeme selects one or several families (corresponding to one or several initial subcat frames) and with the help of features selects exactly the relevant tree sketches : features may rule out some tree sketches in a given family, either because of morphological clash (eg. the passive trees are only selected by past participles) or because of « idiosyncrasies » (eg. the transitive verb *avoir* -to have- disallows passive).

Figure 3 shows the canonical elementary tree anchored by *parlait* (talked)[2] and Figure 4 the three sources of information associated with its internal representation.

The inflected form *parlait* points to the lemma PARLER, and the lexeme /PARLER/ selects in turn the n0Van1 family, where the preposition appears as a co-anchor (except when argument 1 is cliticized).

Morphological lexicons are already available for French (cf DELAF, Multext or ABU lexicon), syntactic lexicons are also available as lexicon-grammar tables (Gross 1975,Leclère 90).Therefore, we focus here on tree sketches which constitute the bulk of our system.

1.3. Semi-automatic generation of the elementary tree sketches

Although tree sketches constitute an abstraction over lexicalized structures, they represent combinations of syntactic constraints (valence, word order...) that are encoded separately in other formalisms such as HPSG. Since these crossings of phenomena typically lead to several hundreds of tree sketches, several authors have stressed the need for a more abstract representation (Vijay-Shanker & Schabes 1992, Becker 1994, Evans et al 1995).

To represent in a more compact and organized way the set of tree sketches for the French LTAG, we use an additional layer of linguistic description, called the metagrammar (MG) (Candito 1996, 1999). MG imposes a general organization for syntactic information and formalizes the well-formedness conditions on elementary trees. It provides a general overview of the grammar and makes it possible for a tool to automatically generate the desired tree sketches from the combination of smaller descriptions. MG takes up the proposal of Vijay-Shanker & Schabes 1992 to represent a TAG as a multiple inheritance network, whose classes specify syntactic structures as partial descriptions of trees (Rogers & Vijay-Shanker, 1994). While trees specify for any pair of nodes either a precedence relation or a path of parent relations, these partial descriptions of trees are sets of constraints that may leave the relation between two nodes underspecified. This relation between two nodes may be further specified by adding constraints in sub-classes or in lateral classes in the inheritance network. Inheritance of partial descriptions is monotonic.

Structures that share the same initial subcategorization frame are grouped in a tree family. For verbal predicates, the final (surface) subcategorization may differ from the initial one (e.g. for passive or causative). A given final (final) subcategorization frame allows for different surface realizations of its arguments (e.g. for extraction, cliticization). MG represents this repartition of information by imposing a three-dimensional inheritance network:

- Dimension 1: initial subcategorization
- Dimension 2: redistribution of functions and transitivity alternations
- Dimension 3: surface realization of arguments, clause type and word order

[1] The trees are simplified and not all features are shown. Standard LTAG conventions are used : ↓ for substitution node, * for foot node, and [] for lemmas.

[2] Information coming from the lexicon appears in bold characters.

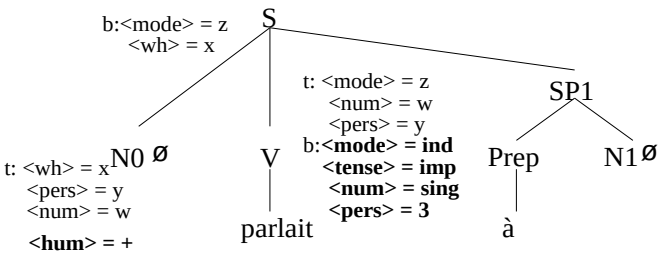


Figure 3. An example of an elementary tree

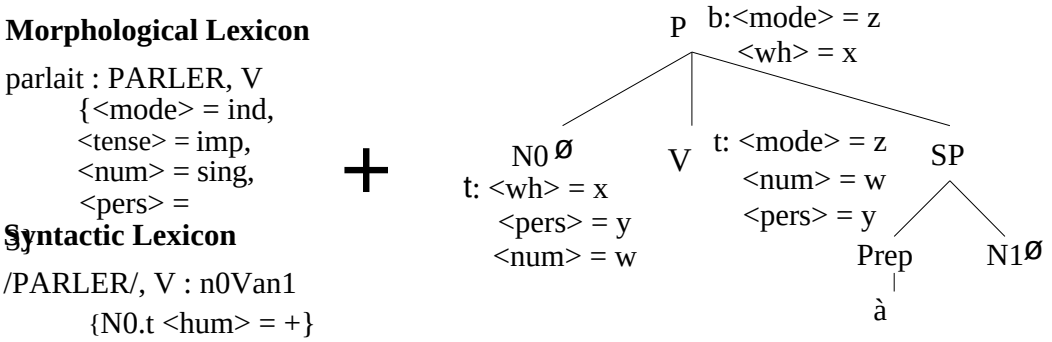


Figure 4. Three sources of information for the tree of Figure 3

Each terminal class in dimension 1 describes a possible initial subcat and partially describes the verbal morpho-syntax (for verbs with inherent clitics). Each terminal class in dimension 2 describes a list of ordered redistributions (including the case of no-redistribution). The redistributions may impose a verbal morphology (eg. the auxiliary for passive). Finally, each terminal class in dimension 3 represents the surface realization of a (final) function (independently of the initial function).

This organization of syntactic knowledge allows to alleviate the task of the linguist. The grammar developer needs to specify only the partial descriptions of the upper classes, without bothering about all crossings of phenomena or about fine-grained details of any elementary tree sketch. The 3 dimension hierarchy is handwritten. Elementary trees are automatically generated through a two-step process. First the compiler automatically creates additional classes in the inheritance network : the "crossing classes". Then each crossing class is translated into one or more tree sketches. During step 1, crossing classes are automatically built as follows:

- a crossing class inherits one terminal class from dimension 1
- then, the crossing class inherits one terminal class from dimension 2
- then, the crossing class inherits classes from dimension 3, representing the realization of every function of the final subcat.

Furthermore, for a crossing class to be well-formed, all unifications involved during the inheritance process must succeed, either for feature structures or for partial descriptions.

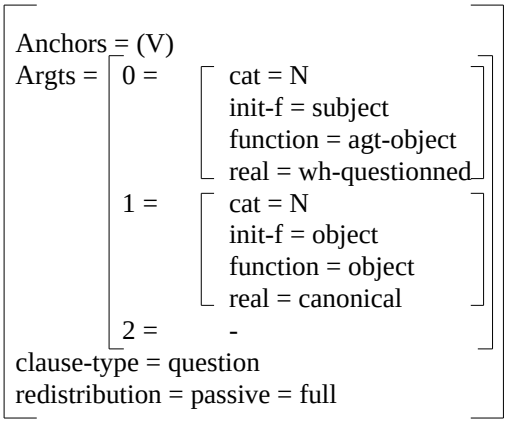
Suppose we want to represent a tree sketch for French passive transitive with a fronted agent phrase, as in : *Par qui sera accompagnée Marie* ? (By whom will Mary be accompanied ?) The crossing class corresponding to this example (see tree sketch Figure 5) inherits the terminal class strict transitive from dimension 1, which declares an initial nominal subject (arg0) and an initial nominal direct objet (arg1). It inherits the terminal class full personal passive from dimension 2, which assigns the final function agt-object to arg0 and the final function subject to arg1. For each final function (here subject and agt-object), the crossing class inherits a terminal class from dimension 3, here inverted-nominal-subject and agt-object-questionned.

With this automatic generation tool, we can maintain a repository of all surface realizations of all basic constructions for French (and their crossing) with a guarantee of consistency over thousands of elementary sketches, and a possibility of parametrization (if for a given text or application, we want to use only a subpart of the whole grammar).

1.4 Characterizing tree sketches as structures of metafeatures

The need to refer to tree sketches in a less rigid way than with a bare name was stressed for instance by Danlos (1998) who uses metafeatures to represent tree sketches within the G-TAG generation system. We take advantage

of the hierarchical representation of tree sketches within a metagrammar to characterize tree sketches as feature structures that we call t-feature structures, as is done in G-TAG. An example is the following :



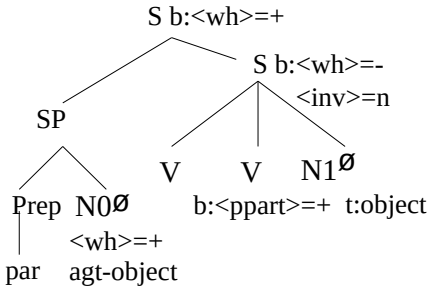


Figure 5. Tree sketch for a passive with wh-questionned agt-object, and associated t-features (Par qui sera accompagnée Marie ?)

Characterizing tree sketches as a combination of t-features allows to refer to a set of tree sketches simply by underspecifying a t-feature structure. This simple idea has several applications. It is already used in G-TAG. It could also be used to express the lexical selection of tree sketches : instead of pre-defining families of tree sketches, with the possibility for a lexical item to select a subset of a family (e.g. transitive family without trees for passive), the lexical item could directly be associated with an underspecified t-feature structure, allowing to select any matching tree sketch. It can also be used for supertagging (see below).

2. Current status of the French LTAG

2.1 General choices

Most of our linguistic analyses follow those of Abeillé 1991 (except that clitic arguments are substituted and not adjoined), complemented by Candito 1999. We dispense with most empty categories, especially in the case of extraction. Semantically void (or non autonomous) elements, such as complementizers, argument marking prepositions or idiom chunks are coanchors in the elementary tree of their governing predicate.

2.1.1 A minimal tagset

We depart from traditional part of speech wherever the modern linguistic analyses have better to propose, especially in the generative tradition. Thus, we distinguish a special category for Clitics (weak pronouns), and for Complementizers. We collapse proper names, common nouns and pronouns into one category N, with distinct features. We do not have a tag for subordinating conjunctions which are either Prepositions (followed by a complementizer: *pendant que* (during)) or (full) Complementizers (*si* (if), *comme* (as)...). Sentential structures are 'flat' (no internal VP). The tagset we use is the following:

- **Lexical categories**: D (determiners), N (nouns, names, pronouns), V (verb), Cl (clitic pronoun), Prep (preposition), A (adjective), Adv (adverb), Conj (Coordinating conjunction), C (complementizer, subordinating conjunction),

- **Non lexical categories**: SP (prepositional phrase), S (sentence). A and N are also used for nominal or adjectival phrases.

2.1.2 A rich set of grammatical functions

Tree sketches in the French TAG are compiled out of the French metagrammar, which expresses subcategorization in terms of grammatical functions. The functions used in the French MG for verbs are the following: *subject, object, dat-object, obl-object, gen-objet, locative, source-locative, manner, goal-infinitive, perception-infinitive, interrogative clause, and predicative complement (attribute of the subject, attribute of the object).*

All these functions can be both initial or final. An additional function "agt-object" is used as a final function only (for agent in passive).We use several "complement" functions for complements of adjectives, prepositions, nouns, adverbs. And these categories may bear the function "modifier" with respect to the element they modify (cf Figure 1).

2.1.3 A parsimonious use of features

With LTAGs, the topology of elementary trees directly captures most of the syntactic properties handled by feature structures in unification-based linguistic theories (LFG or HPSG). We do not need to resort to valence or slash features to ensure subcategorization requirements or filler-gap relations, nor to feature passing principles, apart from unification.

We only rely on atomic valued features (which prevents any cyclic structure). We distinguish:

- **Morphological features**, which are used in the morphological lexicon, in the syntactic lexicon to constrain an argument (eg *trouver* only takes indicative sentential complements) and in the elementary tree sketches for agreement.
- **Syntactic features**, used in the syntactic lexicon (e.g. to disallow passive for a verb) and in the tree sketches (to distinguish trees in the same family or to further constrain tree combinations).
- **Semantic features** : these are gross classifications used for arguments (human, locative etc) and should be further refined.

We are currently using approx. the following 40 features :

morphological features: <det>, <card>, <case>, <el>, <mode>, <num>, <ord>, <pers>, <P-num>, <P-pers>, <tense>.
syntactic features: <ant>, <ant-s>, <ant-v>, <aux>, <cq>, <det>, <extrap>, <gen>, <inv>, <modif>, <neg>, <nom>, <passive>, <part-num>, <part-gen>, <pred>, <princ>, <pro>, <quant>, <san1>, <san2>, <suj-gen>, <suj-pers>, <suj-num>, <sym>, <tense>, <wh>.
semantic features: <conc>, <degre>, <hum>, <loc>, <man>.

2.2 The lexicon

Contrary to the English LTAG which reuses existing dictionaries (Collins 1979 for morphology, Oxford English Dictionary and COMLEX for syntax), we developed our lexicons from scratch. Therefore, they are medium size and currently being extended:[3]

- **The Morphological lexicon**: over 50 000 (inflected) forms: 43000 for verbs, 4700 for nouns and pronouns, 1400 for adjectives and 90 for determiners.
- **The Syntactic lexicons** : over 6000 (disambiguated) entries: 2800 for verbs, 260 for prepositions and adverbs, 300 for adjectives, 50 for determiners, 2500 for nouns, 350 for idioms

The most frequent lexical items were initially extracted from the frequency lists of Julliand 1970 and Catach 1984, except for idioms where one had to rely on personal intuition. They were disambiguated (and separated into different syntactic entries) with standard dictionaries as well as LADL lexicon-grammar tables (Gross 1975,Leclère 90). The morphological lexicon was automatically generated, using PC-Kimmo adapted to French. Both lexicons are organized in lexical databases, and features are normalized with templates.

The morphological lexicon is standard and associates lemmas, inflected forms and relevant morphological features. The syntactic lexicon associates lemmas with constructions (elementary trees or tree families with features) and performs some meaning disambiguation (based on different syntactic constructions, for example for the French verb *abattre* - knock down, shoot down) :

INDEX: abattre/1(physical meaning)
ENTRY: abattre
POS: V
FAM: n0Vn1
FS:

INDEX: abattre/2 (psychological meaning, possible sentential subject)
ENTRY: abattre
POS: V
FAM: s0Vn1
FS: #N1_HUM+, #N0_HUM-

2.3 The elementary trees

FTAG comprises 5280 elementary tree sketches (not counting trees for causative constructions)[4]. Currently, all but 40 are compiled from the French metagrammar. These 40 tree sketches are trees for determiners (plain and complex), nouns used as arguments, coordination conjunctions, clitics and "minimal" trees for deficient verbs such as raising verbs and auxiliaries.

The French MG comprises the description of tree sketches anchored by full verbs, prepositions, full complementizers (subordinating conjunctions), adverbs, adjectives, and nouns (when used as modifiers). The vast majority of the compiled tree sketches are for verbs (over 2000 topologically distinct tree sketches, adding up to over 4000 final tree sketches once all features are taken into account).

2.3.1 Elementary trees for verbs

Within dimension 1, the French metagrammar defines 54 initial subcat frames for verbs (which means there are 54 tree families for verbs in FTAG).

Within dimension 2, the French metagrammar covers the following types of redistribution: passive, middle, causative, reflexive and impersonal, with some interactions with morphology.[5] Following the lines of GB and LFG, we decompose these valence alternations into more abstract redistributions : most valence alternations in French involve subject demotion and promotion to subject.

---

[3] We are currently integrating morphological information from Multext, ABU and INTEX lexicon (which amount to more than 400 000 forms already used by our POS tagger); for syntactic entries we are currently converting some of the LADL descriptions of verbs as well (cf. Abeillé, Clément 1999).

[4] Some tree sketches are counted twice if they belong to different tree families (eg. transitive trees for strict transitive verbs and for ditransitive verbs with an optional indirect object); some only differ by some feature equation (eg. auxiliary selection for intransitive trees). Trees for causatives are not currently used, nor generated, for efficiency reasons. Since the causative construction adds an argument to the subcategorization frame, this more or less doubles the number of verbal tree sketches.

[5] Reflexive is treated as a redistribution in French because it triggers auxiliary change and interacts with causative and impersonal (a verb with a reflexivized direct object behaves as an intransitive).

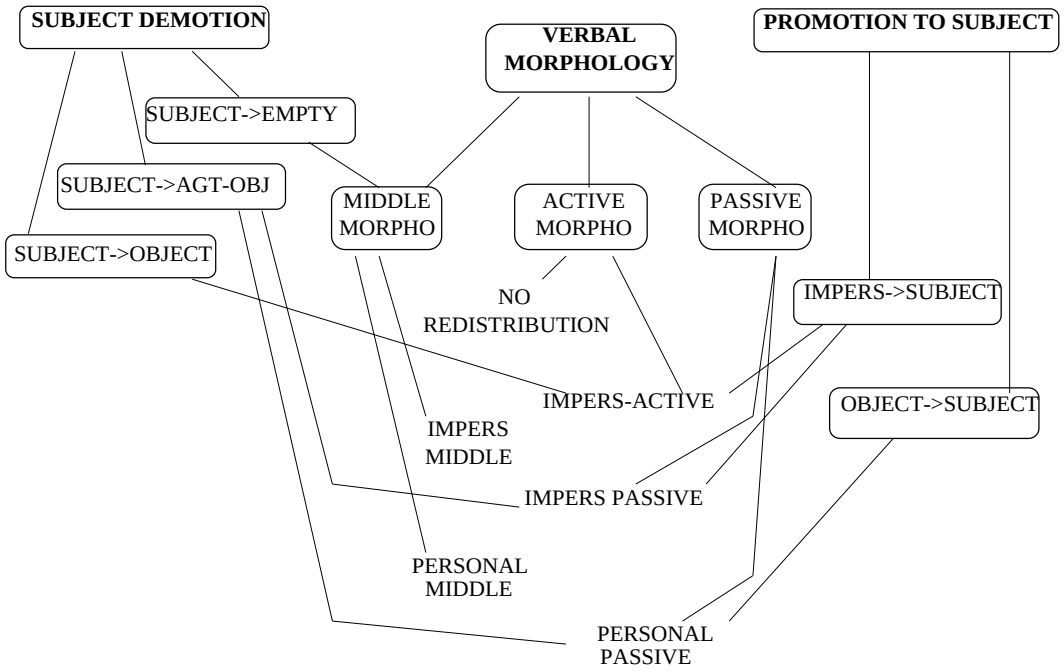


Figure 6. Redistributions for verbs in the French metagrammar (dimension 2)

For active impersonal (eg *Il est arrivé trois lettres pour vous*. 'There arrived three letters for you'), the subject is demoted to object (class SUBJECT _ OBJECT), and the impersonal *il* is introduced as subject (class IMPERS _ SUBJECT).

Passive is characterized by a particular morphology (with a substitution node for the auxiliary verb) and the demotion of the subject (deleted or demoted to a par-phrase), with either the promotion of the object to subject (class OBJECT __SUBJECT) or the introduction of an impersonal subject (class IMPERS _ SUBJECT) for the impersonal passive (cf. Comrie 1977):

Le film sera projeté mardi. (the movie will be shown on tuesday)

Il sera projeté un film mardi prochain (there will be a movie shown next tuesday)

Due to their specific syntactic properties (among which clitic climbing), causative constructions are analyzed as complex predicates, with a flat structure (cf. Abeillé et al 1998) and thus appear in the family anchored by the infinitive (with the causative verb - *faire* or *laisser*-substituted). The subject of the infinitive is demoted to direct-object, par-object or à-object, depending on the transitivity of the infinitive.

Figure 7 partially shows the inheritance links for dimension 2 (simplified, without causative, reflexive). Terminal classes are shown without frame.In dimension 3, the syntactic realizations covered are canonical position, extraction (cleft, relativized or questioned[6]), clitic or non-realized (with a non finite verb). We distinguish nominal and sentential realizations of functions and, for the latter, finite and infinitival realizations.

2.3.2 Metagrammar for other categories

The French metagrammar also describes, in a less sophisticated way though, the tree sketches for adjectives, adverbs, prepositions, subordinating conjunctions and nouns used as modifiers.

Within dimension 1, the French metagrammar describes 4 initial subcategorization frames for adjectives, 6 for adverbs, 5 for prepositions and subordinating conjunctions, and one for nominal modifiers. It thus covers negation placement and negative concord, np adjuncts (prepositionless locative and temporal adverbials), simple and complex determiners (*beaucoup de N* 'a lot of N'), pre- and post- adjectival modifiers, word order variations for most adverbs and adjunct clauses.

6 Thanks to TAG's extended domain of locality, the extracted element is realized as a node in the elementary tree of its predicate.

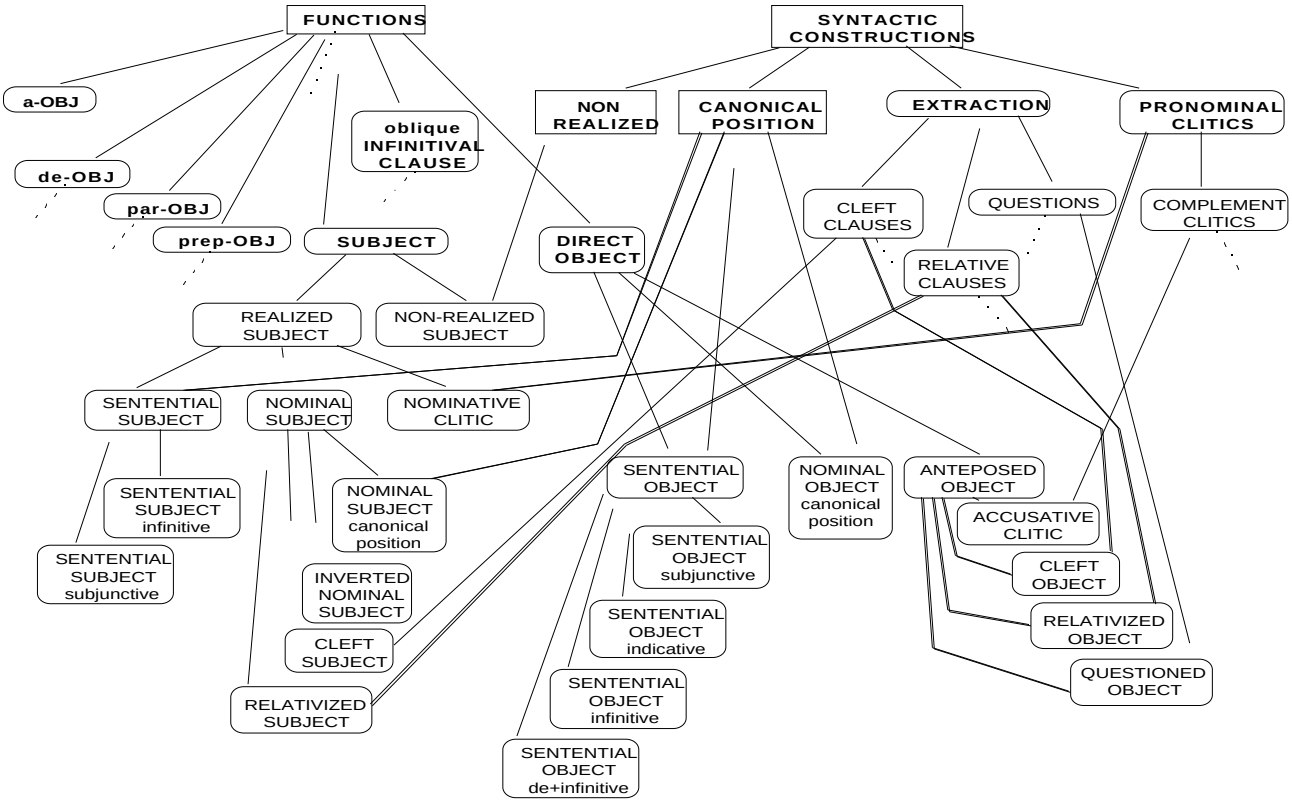


Figure 7. Dimension 3 for French verbs (simplified)

3. Evaluation and comparison

3.1. Evaluation

Evaluating a wide coverage grammar is a difficult task, especially in the absence of a reference syntactically annotated corpus (treebank) for French. We performed a quantitative evaluation using the French test suite developed in the TSNLP project (Estival & Lehmann 1996). The test suit aims at covering the major syntactic phenomena for each language, using a minimal vocabulary (a few hundred words). We extracted all the French items of the TSNLP database, classified by grammatical status (we only took 0 and 1). For all grammatical items, with the 1998 version of our grammar, more than 80% received at least one correct parse (with a lower percentage for complex coordinations). The ambiguity rate was an average of 2,86 analyses per sentence parsed (with number of parses increasing with the length of the sentence; see below)

We checked that there were no unknown words, and performed some automatic segmentation (au => à le). Failures were essentially due to :

- missing lexical encoding (transitive verb without object, transitive use of "parler" : *parler français*...),
- missing elementary trees (causative trees, postverbal clitics with imperatives),
- feature unification clash (agreement with politeness forms: *vous êtes belle* 'you-pl are-pl pretty-sg', or with coordination : *deux bandes bleue et jaune* 'two blue-sg and yellow-sg strips),
- missing phenomenon (tough construction, gapping...).

In order to evaluate also the precision (or discrimination power) of our grammar, we also evaluated it on all agrammatical items (marked 0 in TSNLP database). More than 82% of the agrammatical sentences were correctly rejected. Cases of overanalysis either came from a disputable TSNLP encoding[7], or from the incompleteness of our representation (e.g. we overgenerate for coordination and negation).

3.2. Comparison with other syntactic resources

The lexicon-grammars developed at the LADL for more than 20 years constitute an unrivaled source of knowledge fully reusable. However, it cannot be directly used to analyze (or generate) a text since it only lists some basic constructions (with their lexical head). It does not encode the crossing of constructions nor the productive phenomena which are not clearly lexically sensitive (such as causative, quantifier floating or argument extraction for simple verbs). Thus, even though it is crucial to know that

[7] For example the constraints on 'concordance des temps' (sequence of time) are too strict, ruling out *Demain l'ingénieur vient* (Tomorrow is the engineer coming) which is perfectly grammatical.

transitive *voler* (to steal) must be distinguished from intransitive *voler* (to fly), more general grammatical rules are needed to know that it is the transitive *voler* which is instanciated in examples (1)-(2) without a postverbal np, or that it is the intransitive *voler* which is instanciated in examples (3)-(4) (even though there is a postverbal np):

(1) Ils veulent tout voler (they want to steal everything)

(2) les bijoux qu'ils ont finalement avoué avoir volé... (the jewels that they finally confessed to have stolen)

(3) Lufthansa fait voler ses avions 5 jours sur 7 (L has its planes fly 5 days out of 7)

(4) A une altitude à laquelle ne vole normalement aucun avion ... (an altitude where normally no plane flies)

M. Salkoff (1973, 1979)'s string grammar has listed numerous grammatical strings representative of French syntax but has never been associated with a sizable lexicon and cannot be reused independently of the parser it was made for. The GB grammar developed for French at LATL (Laenzlinger et Wehrli 1991, Wehrli 1997), is more modular and associated with a sizable dictionnary. But it is not clearly separated from the program that uses it (extraction or passive phenomena are not handled as grammatical data but as types of action - attachment, trace creation... - performed by the parser) and thus cannot be easily reused as such.

4. A robust parsing scheme using FTAG

Our grammar is being used for text generation (Danlos 1998, Meunier 1999). When used for parsing, it has the same drawbacks as other rule-based approaches : it fails to analyze some grammatical sentences and usually assigns more than one analysis to the sentences it analyses. We describe here a robust parsing scheme which tries to improve on both aspects. To avoid failure, we have added a part-of-speech tagger developed independently (Abeillé et al. 1998) and a default mechanism for unknown words, to avoid spurious ambiguities, we have added a parse ranker (Kinyon 1999). The general architecture is shown in Figure 8.

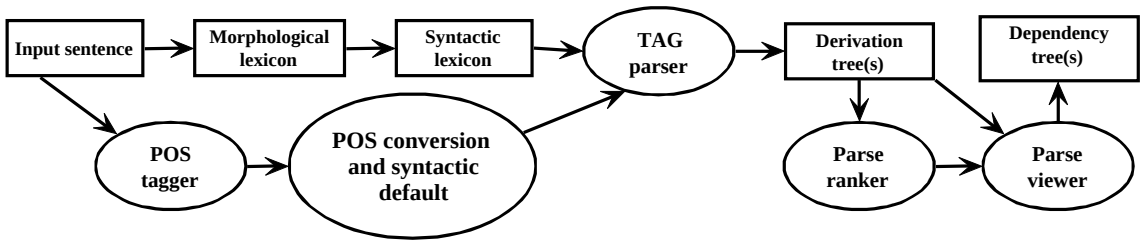


Figure 8. A robust parsing scheme using FTAG

Our parsing scheme is not dependent on a particular TAG parser. The TAG parser we use is that of Schabes 1994, also incorporated in the English XTAG system (Doran et al. 1994, 1998).

4.1. Coping with unknown words

Even with sizable lexicons, real texts are deemed to comprise unknown words. For these we use a POS tagger, developed independently by R. Réyès and L. Clément, which assigns the best possible part of speech to any word absent from the FTAG lexicon. This tagger has a lexicon of over 400 000 forms, more than 50 rules for unknown words and more than 200 contextual rules to handle POS ambiguity. Once the unknown word receives one POS, a default table is used which assigns it a default tree family (for a verb) or a default tree (for a noun).

4.2. Shallow parsing

Obviously not all sentences in real texts can be represented by a full syntactic tree, mainly because of unexpected phenomena (parenthesis, ellipsis) which are difficult to all capture in a structure-based formalism. For cases where no complete parse can be found, we are developing a shallow parser, or supertagger (similar to that of Srinivas 1997) which tries to associate the best elementary tree for each word, and computes some dependency links between them.[8]

Yet the task of assigning the correct elementary tree is much more difficult than assigning the correct POS, especially for verbs, which can be assigned hundreds of trees. We are exploring a first phase which assigns t-features (picked-up in sets of mutually exclusive t-features) to each (head) word. For a given word, only t-features for which there is good evidence would be kept, their combination pointing to a set of matching tree sketches.

4.3. Ranking parses

The output of a TAG parser can be viewed as a derived tree (encoding phrase structure) or as a derivation tree (encoding dependencies). Since it is both more compact and more informative (than the derived tree), we choose the derivation tree for postprocessing.

8. This technique which uses only local information (and cannot fail) resembles that of Roche 1993 based on finite state automata.

Our postprocessing consists in a parse ranker and a parse viewer.

4.3.1. Disambiguation principles based on derivation trees

Our parse ranker is based on empirical (i.e. corpus-based) and psycholinguistic-based preferences, so that a user can get only one parse per sentence if he wants to. It only resorts to lexical and syntactic sources of information (whereas a true disambiguator should also use semantic and discourse information). Since we work on derivation trees, which exhibit the lexicalized trees used for parsing, it is easy to mix lexical and syntactic preferences. Thus, our parse ranker uses 3 types of preferences (in this order):

- lexical preference (such as valence preference for verbs),
- grammatical preferences (i.e. construction types),
- general principles, which are structure-based, domain, language and application independent.

Lexical preferences encode either a category preference or a valence principle. They have to be computed for each word, but we rely on the general tendency for French to favor grammatical categories over lexical category (for example clitics over strong pronouns for *elle* in (1)). Another tendency is to prefer an auxiliary over a full valence verb (*est* as a tense auxiliary and not as a copula in (2))[9].

(1) Elle court (She runs > SHE runs)

(2) Elle est allée de Pau à Paris (She went from Pau to Paris > She is allée de Pau in Paris)

(3) Il est venu une nuit (he came one night > there came a night)

Grammatical preferences encode a preference for a given construction, for example active over passive or personal over impersonal (as in (3)).

These general principles assume the existence of a universal preference for economy and therefore favor analysis that needs to perform fewer and less costly operations (adjunction is more costly than substitution). Contrary to Srinivas & al. 1995, we formulate our structural preference principles in terms of derivation trees, and not constituent trees. This allows to capture widely accepted preferences, such as idiomatic over literal interpretations, and arguments over modifiers. These general principles are the following (Kinyon 1999a,b):

1- Prefer the derivation tree with the fewer number of nodes

2- Prefer to attach an initial tree low in a derivation tree[10]

3- Prefer the derivation tree with the fewer number of auxiliary trees

Principle 1 favors the idiomatic interpretation of a sentence over its literal interpretation (a). It also favors the attachment prepositional phrases as arguments rather than modifiers (b).

Principle 2 favors the low attachment of arguments, when several alternative attachments are possible : in (c) d*e la manifestation* is argument of *organisateur* rather than of *soupçonne*. In (d), *Jean* is argument of *dit* rather than of *parle*.

Principle 3 favors the derivation tree involving the fewer number of adjunctions (i.e. modifiers) : in (e) *le matin* could be a modifier, but the attachment as an argument is preferred.

(a) *Jean brise la glace* (J. cools things off / breaks the ice)

(b) *Jean pense à la reunion* (J. thinks of the reunion / thinks at the reunion)

(c) *Jean soupçonne l'organisateur de la manifestation* (J. suspects the organizer of the demonstration)

(d) *C'est à Jean que Marie dit que Paul parle* (It's to J. that M. says that John thinks /it's of J that ...)

(e) *Jean attend le matin* (J. awaits the morning / waits in the morning)

4.3.2 Application to TSNLP

As mentioned above, we parsed French sentences from the TSNLP test suite. In order to test our parse ranker, we kept the 1074 grammatical ones that received at least one parse (with an average of 2.85 derivations / sentence). There were very few categorial ambiguity, and most feature ambiguities were handled via underspecification (eg *les enfants* 'the children' feminine or masculine). The remaining (structural) ambiguities are the following (not all of these are spurious):

- modifier adjoined to S ou V after an intransitive verb (*l'ingénieur viendra volontiers*),
- prepositional phrase analysed as complement or modifier (L'ingénieur préfère le vin à l'eau; Il passe pour un spécialiste),
- passive with or without agent (the par-PP can be analysed as an agent phrase or as a modifier)
- several adjunction sites in case of multiple modifiers.[11]

A human picked one or more «correct» derivations for each sentence. We performed an experiment using only the general disambiguation principles, plus 5 lexical and grammatical preferences:

1- Prefer to attach an adverb to a main verb rather than to an auxiliary
2- Prefer clitics to nouns
3- Prefer auxiliaries to lexical verbs
4- Prefer determiner to numeral adjectives
5- Disprefer impersonal

9. In the examples, the prefered reading precedes the > sign.

10 In this view, sentential complements are substituted when no extraction takes place.

11 For NPs with a determiner and a relative clause, this ambiguity (relative clause or determiner higher) can be interpreted as encoding the distinction between appositive and restrictive relative interpretation (*la réunion qui aura lieu demain; les ingénieurs qui sont convoqués*).

If we apply only the general principles, we are left with 2.17 derivations / sentence, while the number of sentences for which a «correct» parse is present only marginally decreased (from 99,6% to 98,1%). If we apply the general principles plus the 5 above principles, we are left with 1,5 derivation / sentence. It turns out that one of the main sources of spurious ambiguities lies in adverbial attachment. Again, the principles and the heurisitcs aim at attaching arguments as accurately as possible, and pruning out spurious ambiguity. They do not deal yet with strict modifier attachment.

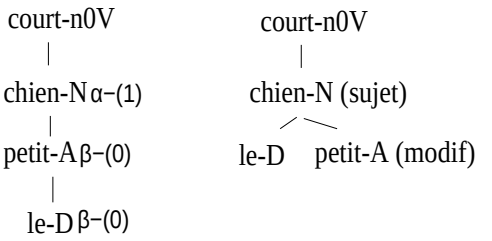
Future work (in order to obtain 1 derivation / sentence.) will incorporate more lexical preferences (i.e. subcategorization preferences for each verb) as well as grammatical preferences (i.e. preferred trees inside each tree families), computed on a sizable tagged French newspaper corpus (Abeillé et al. 1999).

4.4. Parse viewing

The output of a TAG parser can be difficult to interpret, and does not always show the desired dependency relations. In order to make the derivation trees easier to read, we are using a parse viewer (originally developed by L. Clément) which restores the correct dependencies in case of cascade modifiers (following a suggestion by Shieber and Schabes 1994) and replaces the tree adresses by more meaningful function names (cf. Candito 1999). An example is the following :

Le petit chien court (the little dog runs)

derivation tree (parser output) => dependency tree **(viewer output)**



Further work involves computing more dependency relations (adding arcs) for extraction or control phenomena (cf. Candito and Kahane 1998).

References

A. Abeillé, 1991. *Une grammaire lexicalisée d'arbres adjoints pour le français.* PhD Thesis, University Paris 7. (to appear Editions du CNRS)

A. Abeillé, M-H. Candito, 1999. FTAG A lexicalized tree adjoining grammar for French, *in* Abeillé & Rambow (eds) *Tree Adjoining Grammar*, CSLI, Stanford. (to appear)

A. Abeillé, L. Clément, R. Réyès, 1998. Talana annotated corpus for French, the first results, *Proceedings LREC*, Granada.

A. Abeillé, L. Clément, 1999. A reference tagged corpus for French, *Proceedings LINC99*, EACL, Bergen.

A. Abeillé, D. Godard, P. Miller, 1997, Les constructions causatives en français : un cas de compétition syntaxique", *Langue française*, 115, pp. 62-74.

T. Becker, 1994. Patterns in metarules, *Proceedings TAG+3 Workshop*, TALANA, Paris 7.

M-H Candito, 1996. A principle-based hierarchical representation of LTAG, *Proceedings 15th COLING*, Kopenhagen.

M.-H. Candito, 1999. Un outil multilingue de construction semi-automatique de grammaires d'arbres adjoints: application au français et à l'italien, *TAL* 40:1, pp. 87-124.

M.-H. Candito, 1999. Représentation modulaire et paramétrable de grammaires électroniques lexicalisées. Application au français et à l'italien. PhD dissertation. University Paris 7.

M-H. Candito, S. Kahane, 1998. Can the TAG derivation tree represent a semantic graph ? an answer in the light of MTT, *Proceedings TAG+4 Workshop*, Philadelphia.

N. Catach, 1984. *Les listes orthographiques de base du français*, Paris, Nathan.

B. Comrie, 1977. In defense of spontaneous demotion, in Cole and Sadock (eds) *Grammatical relations, Syntax and semantics*, New York, Academic Press.

L. Danlos, 1998. GTAG : un formalisme lexicalisé pour la génération inspiré de TAG, *TAL* 39:2. pp%

C. Doran, D. Egedi, Hockey, B., Srinivas, B., Zaidel M. 1994. The XTAG system : a wide coverage grammar for English, *Proceedings 15th COLING*, Kyoto, pp. 922-928.

C. Doran, B. Hockey, Sarkar A, Srinivas, B., Xia F. 1998. Evolution of the XTAG system, *in* Abeillé & Rambow (eds) *Tree Adjoining Grammar*, CSLI, Stanford.

E. Evans et al. 1999. Lexical rules are just lexical rules, in Abeillé & Rambow (eds) *Tree Adjoining Grammar*,CSLI.Stanford.

D. Estival, S. Lehmann, 1997. TSNLP: des jeux de phrases test pour le TALN, *TAL*, 38:1, pp. 155-172.

M. Gross, 1975. Méthodes en syntaxe, Hermann, Paris.

A. Kinyon 1999a : Parsing preferences and lexicalized Tree Adjoining Grammars : exploiting the derivation tree, Proceedings ACL'99. Maryland.

A. Kinyon 1999b : Hiérarchisation d'analyses basée sur des informations dependancielles pour les LTAGs, Proceedings TALN'99. Cargèse.

A. Kroch, A. Joshi, 1985. The linguistic relevance of Tree Adjoining grammars, Technical Report, Univ. of Pennsylvania.

A. Joshi, D. Weir, K Vijay-Shanker, 1991. The convergence of mildly context-sensitive formalisms, in P. Sells, S. Shieber, T. Wasow (eds), *Foundational issues in NLP*, Cambridge: MIT Press.

A. Julliand, D. Brodin, C. Davidovitch, 1970. *Frequency dictionnary of French words: 5000 words*, Mouton, La Haye.

C. Leclère 1990, Organisation du lexique-grammaire des verbes français. Langue Française 87, Paris: Larousse.

F. Meunier, 1999.Modélisation des ressources linguistiques d'une application industrielle. Proceedings TALN'99, Cargèse.

E. Roche (1993) Analyse syntaxique transformationnelle par transducteurs et lexique grammaire, Thèse de doctorat, Université Paris 7.

J. Rogers, K. Vijay-Shanker, 1994. Obtaining trees from their descriptions, an application to tree adjoining grammar, *Computational Intelligence*, 10:4.

M. Salkoff, 1973. *Une grammaire en cha»ne du français*, analyse distributionnelle, Paris: Dunod.

M. Salkoff, 1979. *Analyse syntaxique du français, grammaire en cha»ne* J. Benjamin, Amsterdam.

Y. Schabes, 1994. Left-to-right parsing of lexicalized Tree Adjoining grammars, *Computational Intelligence*, 10:4,pp.506-524.

Y. Schabes, S. Shieber, 1994. An alternative conception of Tree Adjoining derivation, *Computational Linguistics*, 20:1,91-124.

M. Silberztein, 1993. *Dictionnaires électroniques pour l'analyse lexicale : le système INTEX*, Masson, Paris.

B. Srinivas, Doran C., Kulick S. 1995 : Heuristics and Parse Ranking. 4th workshop on Parsing Technologies (IWPT). Prag.

B. Srinivas, 1997. Complexity of lexical descriptions and its relevance for partial parsing, PhD thesis, Univ. of Pennsylvania.

K. Vijay-Shanker, Y. Schabes, 1992. Structure sharing in lexicalized tree adjoining grammar, Proceedings 14th COLING.

E.Wehrli, 1997. *L'analyse syntaxique des langues naturelles : techniques et méthodes*, Paris: Masson.