Building Parallel LTAG for French and Italian

Marie-Hélène Candito
TALANA & UFRL, Université Paris 7, case 7003, 2, place Jussieu 75251 Paris Cedex 05 France
marie-helene.candito@linguist.jussieu.fr

Abstract

In this paper we view Lexicalized Tree Adjoining Grammars as the compilation of a more abstract and modular layer of linguistic description : the metagrammar (MG). MG provides a hierarchical representation of lexico-syntactic descriptions and principles that capture the well-formedness of lexicalized structures, expressed using syntactic functions. This makes it possible for a tool to compile an instance of MG into an LTAG, automatically performing the relevant combinations of linguistic phenomena. We then describe the instantiation of an MG for Italian and French. The work for French was performed starting with an existing LTAG, which has been augmented as a result. The work for Italian was performed by systematic contrast with the French MG. The automatic compilation gives two parallel LTAG, compatible for multilingual NLP applications.

1. Introduction

Lexicalized Tree Adjoining Grammars (LTAG) is a formalism integrating lexicon and grammar (Joshi, 87; Schabes et al, 88) : its description units are lexicalized syntactic trees, the *elementary trees*. The formalism is associated with a tree-rewriting process that links sentences with syntactic structures (in either way), by combining the elementary trees with two operations, adjunction and substitution. We assume the following linguistic features for LTAG elementary trees (Kroch & Joshi, 85; Abeillé, 91; Frank, 92):

- lexicalization : elementary trees are anchored by at least one lexical item.
- semantic coherence : the set of lexical items on the frontier of an elementary tree forms exactly one semantic unit[1].
- large domain of locality : the elementary trees anchored by a predicate contain positions for the arguments of the predicate.

This last feature is known as *the predicate-argument cooccurrence principle* (PACP). Trees anchored by a predicate represent the minimal structure so that positions for all arguments are included. These argumental positions are extended either by receiving substitution or by adjoining at a node. Adjunction is used to factor out recursion.

Figure 1 shows two elementary trees anchored by the French verbal form *mange* (eat-pres-sg), whose arguments in the active voice are a subject NP and a direct object NP[2]. The first tree shows all arguments in canonical position. The second tree shows a relativized subject and a pronominal object (accusative clitic). The argumental nodes are numbered, according to their oblicity order, by an index starting at 0 in the unmarked case (active). So for instance in passive trees, the subject is number 1, not 0.

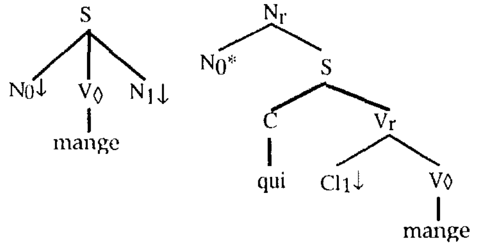


Figure 1: 2 elementary trees anchored by *mange*

Though LTAG units used during derivation are lexicalized trees, the LTAG internal representation makes use of "pre-lexicalized" structures, that we will call tree sketches, whose anchor is not instantiated and that are shared by several lexicalized trees. The set of tree sketches thus forms a syntactic database, in which lexical items pick up the structures they can anchor.

Families group together tree sketches that are likely to be selected by the same lexeme: the tree sketches may show different surface realization of the arguments (pronominal clitic realization, extraction of an argument, subject inversion...) or different diathesis —matchings between semantic arguments and syntactic

[1] Thus semantically void lexical forms (functional words) do not anchor elementary trees on their own. And words composing an idiomatic expression are multiple anchors of the same elementary tree.

[2] The trees are examples from a French LTAG (Abeillé, 91), with no VP node (but this is irrelevant here). The ↓ means the node must receive substitution. The * means the node must adjoin in another tree.

functions— (active, passive, middle..) or both. The lexical forms select their tree sketches by indicating one or several families, and features. The features may rule out some tree sketches of the selected family, either because of morphological clash (eg. the passive trees are only selected by past participles) or because of idiosyncrasies. For instance, the French verb *peser* (to weight) can roughly be encoded as selecting the transitive family, but it disallows the passive diathesis.

It remains that tree sketches are large linguistic unit. Each represents a combination of linguistic descriptions that are encoded separately in other formalisms. For instance, a tree sketch is in general of depth ≥ 1, and thus corresponds to a piece of derivation in a formalism using CF rewrite rules (cf (Kasper et al, 95) for the presentation of an LTAG as a compiled HPSG). This causes redundancy in the set of tree sketches, which makes it difficult to write or maintain an LTAG. Several authors (Vijay-Shanker et al, 92- hereafter (VSS92) - ; Becker, 93; Evans et al, 95) have proposed practical solutions to represent in a compact way an LTAG. The idea is to represent canonical trees using an inheritance network and to derive marked syntactic constructions from base tree sketches using lexico-syntactic rules.

(Candito, 96), building on (VSS92), defines an additional layer of linguistic description, called the metagrammar (MG), that imposes a general organization for syntactic information and formalizes the well-formedness of lexicalized structures. MG not only provides a general overview of the grammar, but also makes it possible for a tool to perform automatically the combination of smaller linguistic units into a tree sketch.

This process of tree sketch building is comparable to a context-free derivation - in the generation way - that would build a minimal clause. A first difference is that CF derivation is performed for each sentence to generate, while the tree sketches are built out of an MG at compile time. Another difference is that while CF derivation uses very local units (CF rules), MG uses partial descriptions of trees (Rogers et Vijay-Shanker, 94) more suitable for the expression of syntactic generalizations.

MG offers a common, principle-based frame for syntactic description, to fill in for different languages or domains. In section 2 we present the linguistic and formal characteristics of MG (in a slightly modified version), in section 3 the compilation in an LTAG, and in section 4 we describe the instantiation of the MG for French and Italian. Finally we give some possible applications in section 5.

2. The metagrammar

Formally the MG takes up the proposal of (VSS92) to represent grammar as a multiple inheritance network, whose classes specify syntactic structures as partial descriptions of trees (Rogers & Vijay-Shanker, 94). 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 underspecified the relation existing between two nodes.

The relation between two nodes may be further specified, either directly or by inference, by adding constraints, either in sub-classes or in lateral classes in the inheritance network.

In the MG, nodes of partial descriptions are augmented with feature structures : one for the feature structures of the future tree sketches and one for the features that are specific to the MG, called meta-features. These are, for instance, the possible parts of speech of a node or the index (cf Section 1) in the case of argumental nodes.

So a class of an instantiated MG may specify the following slots :

- the (ordered) list of direct parent classes
- a partial description of trees
- feature structures associated with nodes[3]

Contrary to (VSS92) nodes are global variables within the whole inheritance network, and classes can add features to nodes without involving them in the partial description. Inheritance of partial descriptions is monotonic.

The aim is to be able to build pre-lexicalized structures respecting the PACP, and to group together structures likely to pertain for the same lexeme. In order to achieve this, MG makes use of syntactic functions to express either monolingual or cross-linguistic generalizations (cf the work in LFG, Meaning-Text Theory or

[3] Actually the tree description language —that we will not detail here— involves constants, that name nodes of satisfying trees. Several constants may be equal and thus name the same node. The equality is either infered or explicitly stated in the description.

Relational Grammar (RG) - see (Blake, 90) for an overview). Positing syntactic functions, characterized by syntactic properties, allows to set parallels between constructions for different languages, that are different in surface (for word order or morpho-syntactic marking), but that share a representation in terms of functional dependencies. Within a language, it allows to abstract from the different surface realizations of a given function and from the different diathesis a predicate can show.

So in MG, subcategorization (hereafter subcat) of predicates is expressed as a list of syntactic functions, and their possible categories. Following RG, an *initial* subcat is distinguished, namely the one for the unmarked case, and is modifiable by *redistribution* of the functions associated with the arguments of the predicate. Technically, this means that argumental nodes in partial descriptions bear a meta-feature "initial-function" and a meta-feature "function". The "function" value is by default the "initial-function" value, but can be revised by redistribution. Redistributions, in a broad sense, comprise :

- pure redistributions that do not modify the number of arguments (eg. full passive).
- reductions of the number of arguments (eg. agentless passive)
- augmentations of the number of arguments (mainly causative).

In MG, structures sharing the same initial subcat can be grouped to form a set of structures likely to be selected by the same lexeme. For verbal predicates, a minimal clause is partly represented with an ordered list of successive subcats, from the *initial* one to the *final* one. Minimal clauses sharing a final subcat, may differ in the surface realizations of the functions. The MG represents this repartition of information by imposing a three-dimension inheritance network[4]:

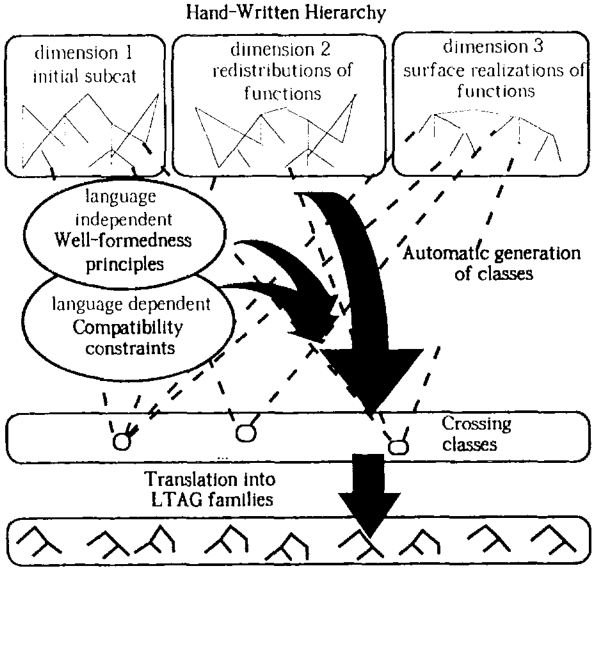
- **dimension 1**: initial subcat
- **dimension 2**: redistributions of functions
- **dimension 3**: surface realizations of syntactic functions.

In an instantiated MG for a given language, each terminal class of dimension 1 describes a possible initial subcat and describes partially the verbal morpho-syntax (the verb may appear with a frozen clitic, or a particle in English). Each terminal class of 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). Each terminal class of dimension 3 represent the surface realization of a function (independently of the initial function). For some inter-dependent realizations, a class may represent the realizations of several functions (for instance for clitics in romance languages).

Terminal classes of the hand-written hierarchy are pieces of information that can be combined to form a tree sketch that respects the PACP. For a given language, some of the terminal classes are incompatible. This is stated either by the content of the classes themselves or within an additional set of language-dependent constraints (*compatibility constraints*). For instance a constraint is set for French, to block cooccurrence of an inverted subject with an object in canonical position (while this is possible for Italian).

3. Compilation of MG to LTAG

The compilation is a two-step process, illustrated figure 2. First the compiler automatically creates additional classes of the inheritance network : the "crossing classes". Then each crossing class is translated into one or several tree sketches.


Hand-Written Hierarchy
dimension 1
initial subcat
dimension 2
redistributions of functions
dimension 3
surface realizations of functions
language independent
Well-formedness principles
language dependent
Compatibility constraints
Automatic generation of classes
Crossing classes
Translation into LTAG families

[4] More precisely a hierarchy is defined for each category of predicate. Dimension 2 is primarily relevant for verbal predicates. Further, remaining structures, for instance for argument-less lexemes or for auxiliaries and raising verbs are represented in an additional network, by classes that may inherit shared properties, but that are totally written by hand.

Figure 2 : Compilation of MG to LTAG

3.1 Automatic extension of the hierarchy

A crossing class is a linguistic description that must fulfill the PACP. Using syntactic functions and the three-dimension partition, MG makes more precise this well-formedness principle. A crossing class is a class of the inheritance network that is automatically built as follows:

- a crossing class inherits exactly one terminal class of dimension 1
- *then*, a crossing class inherits exactly one terminal class of dimension 2

These two super-classes define an ordered list of subcat, from the *initial* one to the *final* one.

- *then*, a crossing class inherits classes of dimension 3, representing the realizations of every function of the final subcat.

Further, 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. Clashes between features or inconsistencies in partial descriptions are used to rule out some irrelevant crossings of linguistic phenomena. Finally, the compatibility constraints must be respected (cf Section 2).

3.2 Translation into LTAG families

While crossing classes specify a partial description with feature structures, LTAG use trees. So the compiler takes the "representative" tree(s) of the partial description (see Rogers & Vijay-Shanker, 94 for a formal definition). Intuitively these representative trees are trees minimally satisfying the description. There can be several for one description. For example, the relative order of several nodes may be underspecified in a description, and the representative trees show every possible order.

A family is generated by grouping all the trees computed from crossing classes that share the same class of dimension 1.

4. Metagrammars for French and Italian : a contrast

We have instantiated the metagrammar for French, starting with an existing LTAG (Abeillé, 91). The recompilation MG→LTAG insures coherence (a phenomena is consistently handled through the whole grammar) and completeness (all valid crossings are performed). The coverage of the grammar has been extended[5].

Then we have adapted the French MG to Italian, to obtain a "parallel" LTAG for Italian, close with respect to linguistic analyses. The general organization of the MG gives a methodology for systematic syntactic contrast. We describe some pieces of the inheritance network for French and Italian, with particular emphasis on dimension 2 and, in dimension 3, on the surface realizations of the subject.

4.1 Dimension 1

We do not give a description of the content of this dimension, but rather focus on the differences between the two languages. A first difference in dimension 1 is that for Italian, there exist verbs without argument[6] (atmospheric verbs), while for French, a subject is obligatory, though maybe impersonal.

Another difference, is known as the unaccusative hypothesis (see (Renzi, 88, vol I) for an account). It follows from syntactic evidence, that the unique argument of *avere*-selecting intransitives (eg. (1)) and *essere*-selecting intransitives (the unaccusatives, eg. (2)) has different behavior when post-verbal:

(1) *Ne hanno telefonato tre.
(of-them have phoned three)
Three of them have phoned

(2) Ne sono rimaste tre.
(of-them are remained three)
Three of them have remained.

We represent unaccusatives as selecting an initial object and no initial subject. A redistribution in dimension 2 promotes this initial object into a special subject (showing subject properties and some object properties, like the *ne*-licensing shown in (2))[7]. This redistribution is also used for specifying passive and middle, which both trigger unaccusative behavior (see next section).

[5] The number of tree sketches passed from 800 to 1100 (without causative trees).

[6] An alternative analysis would be to consider that these verbs select a subject pronoun, that is not realized in Italian (pro-drop language).

[7] We take a simpler approach than RG, which accounts for most of the Italian data. Unhandled are the auxiliary change for verbs, when goal-phrases are added (see (Dini, 95) for an analysis in HPSG).

4.2 Dimension 2

The MG for French and Italian cover the following types of redistribution[8] : passive, middle, causative and impersonal (only for French). Causative verbs plus infinitives are analysed in Romance as complex predicates. Due to a lack of space will not describe their encoding in MG here. Figure 3 shows the inheritance links of dimension 2 for French (without causative). Terminal classes are shown without frame.

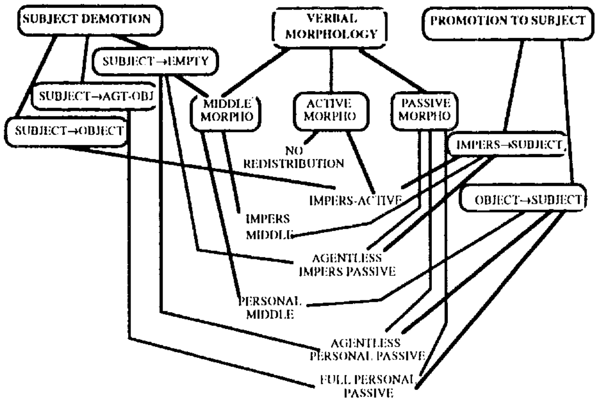


Figure 3 : Dimension 2 for French (without causative)

The verbal morphology is affected by redistributions, so it appears in the hierarchy. The hierarchy comprises the case of no-redistribution, that inherits an active morphology : it simply states that the anchor of the future tree sketch is also the verb that receives inflexions for tense, agreement...

Refering to the notion of hierarchy of syntactic functions (à la Keenan-Comrie), we can say that the redistributions shown comprise a subject *demotion* (which can be a deletion) and a promotion of an element to subject.

For active impersonal (3), the subject is demoted to object (class SUBJECT→OBJECT), and the impersonal *il* is introduced as subject (class IMPERS→SUBJECT).

(3) Il est arrivé trois lettres pour vous.
(IL is arrived three letters for you)
There arrived three letters for you.

Passive is characterized by a particular morphology (auxiliary bearing inflections + past participle) and the demotion of subject (which is either deleted, class SUBJECT→EMPTY, or demoted to a by-phrase, class SUBJECT→AGT-OBJ), but not necessarily by a promotion of the object to subject (class OBJECT→SUBJECT) (cf (Comrie, 77)). In French, the alternative to object promotion is the introduction of the impersonal subject (class IMPERS→SUBJECT)[9]. This gives four possibilities, agentless personal (4), full personal (5), agentless impersonal (6), full impersonal, but this last possibility is not well attested.

(4) Le film sera projeté mardi prochain.
The movie will be shown next tuesday.

(5) La voiture a été doublée par un vélo.
The car was overtaken by a bike.

(6) Il a été décrété l'état d'urgence.
(IL was declared the state of emergency)
The state of emergency was declared.

Middle is characterized by a deletion of the subject, and a middle morphology (a reflexive clitic *se*). Here also we have the alternative OBJECT→SUBJECT (7) or IMPERS→SUBJECT (8) . The interpretation is generic or deontic in French.

(7) Le thé se sert à 5h.
(Tea SE serves at 5.)
One should serve tea at 5.

(8) Il se dit des horreurs ici.
(IL SE says horrible things here)
Horrible things are pronounced in here.

Now let us contrast this hierarchy with the one for Italian. Figure 4 shows dimension 2 for Italian.

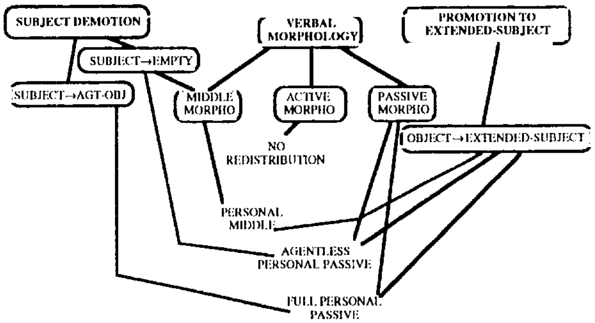


Figure 4 : Dimension 2 for Italian (without causative)

In Italian, what is called impersonal (9a) is a special realization of subject (by a clitic *si*), meaning either *people*, *one* or *we*. (cf Monachesi, 95). The French equivalent is the

[8] The locative alternation (John loaded the truck with oranges/John loaded oranges into the truck), is not covered at present time, but can easily be added. It requires to choose an initial subcat for the verb.

[9] So we do not analyse impersonal passive as passive to which apply impersonal. This allows to account for the (rare) cases of impersonal passives with no personal passive counterpart.

nominative clitic *on* (9b).

(9a) it. Si partì.
(SI left) People / we left.

(9b) fr. On partit.

This impersonal *si* is thus coded as a realization of subject, in dimension 3, and we have no IMPERS→SUBJECT promotion for the Italian dimension 2. The impersonal *si* can appear with all redistributions except the middle. The Italian middle is similar to French, with a reflexive clitic *si*. Indeed impersonal *si*, with transitive verbs and singular object (10), is ambiguous with a middle analysis (and subject inversion).

(10) Si mangia il gelato.
(SI eat-3sg the ice-cream)
The ice-cream is eaten.

With a plural nominal object, some speakers do not accept impersonal (with singular verb (11a)) but only the middle (with verb agreement (11b)).

(11a) Si mangia le mele.
(SI eat-3sg the apples)

(11b) Si mangiano le mele.
(SI eat-3pl the apples)

Another difference with French redistributions, is that when the object is promoted, in passive or middle, it is as a subject showing unaccusative behavior (eg. *ne*-licensing, cf section 4.1). To represent this, we use the class OBJECT→EXTENDED-SUBJECT, which is also used for the spontaneous promotion of initial object of unaccusative intransitives (cf section 4.1). So for Italian, passive (agentless or full) and middle (11b) comprise a subject demotion (a mandatory deletion for middle) and the promotion OBJECT→EXTENDED-SUBJECT, while for intransitive unaccusatives, this promotion is spontaneous.

Other differences between French and Italian concern the interaction of causative with other redistributions : passive and middle can apply after causative in Italian, but not in French.

4.3 Dimension 3

We describe in dimension 3 the classes for the surface realizations of subject. This function is special as it partially imposes the mode of the clause. The subject is empty for infinitives and imperatives[10]. Adnominal participial clauses are represented as auxiliary trees that adjoin on a N, the subject is the foot node of the auxiliary tree (we do not detail here the different participial clauses).

For French (Figure 5), when realized, the subject is either sentential, nominal or pronominal (clitic). Nominal subjects may be in preverbal position or inverted, relativized or cleft. These last two realizations inherit also classes describing relative clauses and cleft clauses. Sentential subjects are here only preverbal. Clitic subjects are preverbal (post-verbal subject clitics are not shown here, as their analysis is special). Note that in dimension 2, the class IMPERS→SUBJECT specifies that the subject is clitic, and dominates the word *il*. This will only be compatible with the clitic subject realization.

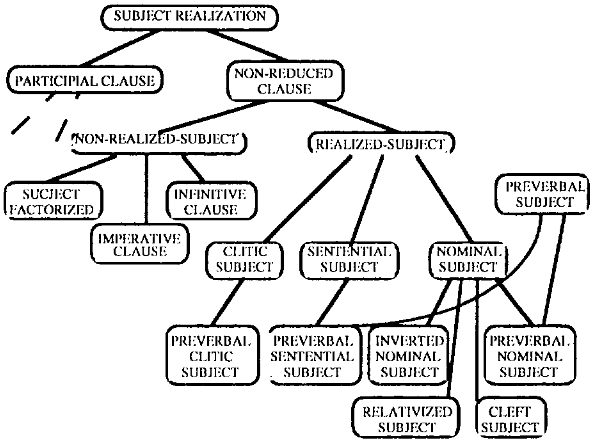


Figure 5 : Subject realizations for French

For Italian, (Figure 6), the hierarchy for subjects is almost the same : a class for non-realized subjects is added, since Italian is a pro-drop language, and pronominal subjects are not realized. But we mentioned in section 4.2 the special case of the impersonal subject clitic *si*. To handle this clitic, the Italian class for clitic subject introduces the *si*.

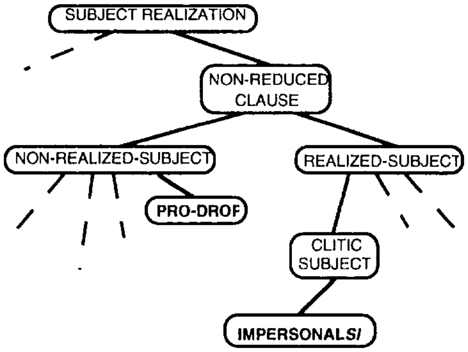


Figure 6 : Subject realizations for Italian (differences with French in bold)

[10] See (Abeillé, 91) for the detail of the linguistic analyses chosen for French. We describe here the hierarchical organization.

5. Applications

The two LTAG for French and Italian are easy to maintain, due to the hierarchical representation in MG. They can be customized for language domains, by cutting subgraphs of the inheritance network in MG.
The MG for French is currently used to maintain the French LTAG. It has also been used to generate the tree sketches for the text generator G-TAG (Danlos & Meunier, 96), based on TAG. The generator makes use of tree sketches characterization as a set of features —called t-features— such as <passive>, <infinitival-clause>... This characterization has been straightforward to obtain with the representation of the tree sketches in MG.

Further, the two MG for French and Italian can provide a basis for tranfer between syntactic structures for Machine Translation. LTAG elementary trees correspond to a semantic unit, with (extendable) positions for the semantic arguments if any. (Abeillé, et al, 90) propose to pair elementary trees for the source and target languages and to match in these pairs the argumental positions of the predicate. Once these links are established, the synchronous TAG procedure can be used for translation.
The argumental positions correspondance is straightforward to state within the MG framework. We plan to define an automatic procedure of tree-to-tree matching using MG representations for source and target languages, once the initial functions of arguments are matched for pairs of predicates. This procedure will make use of sets of t-features to characterize tree sketches (as in G-TAG) derived at the MG→LTAG compilation time. Correspondances between t-features or sets of t-features have to be defined.

References

A. Abeillé, 1991 : Une grammaire lexicalisée d'arbres adjoints pour le français. Ph.D. thesis. Univ. Paris 7.

A. Abeillé, Y. Schabes, A. Joshi, 1990 : Using Lexicalized TAG for Machine Translation. COLING'90.

T. Becker, 1993 : HyTAG : a new type of Tree Adjoining Grammars for Hybrid Syntactic representation of Free Order Languages, Ph.D. thesis, Univ of Saarbrücken.

M-H. Candito, 1996 : A principle-based hierarchical representation of LTAG. COLING'96.

B. Comrie, 1977 : In defense of spontaneous demotion : the impersonal passive. Syntax and semantics « Grammatical functions » Cole & Saddock.

Danlos, L Meunier, F, 1996 : G-TAG, un formalisme pour la génération de textes : présentation et applications industrielles. ILN'96, Nantes.

L. Dini, 1995 : Unaccusative behaviors. Quaderni di Linguistica. 9/95.

R. Evans, G. Gazdar, D. Weir, 1995 : Encoding Lexicalized TAG in a non-monotonic inheritance hierarchy. ACL'95.

R. Frank, 1992 : Syntactic locality and Tree Adjoining Grammar: Grammatical, Acquisition and Processing Perpectives. Ph.D. thesis. Univ. of Pennsylvania.

R. Kasper, B. Kiefer, K. Netter, K. Vijay-Shanker, 1995 : Compilation of HPSG to TAG. ACL'95.

I. Mel'cuk, 1988 : *Dependency Syntax: Theory and Practice*. State Univ. Press NY, Albany (NY).

P. Monachesi, 1996 : A grammar of Italian clitics. Ph. D. thesis. Univ. of Tilburg.

L. Renzi, 1988 : *Grande grammatica di consultazione* (3 vol.) Il Mulino, Bologna.

J. Rogers, K. Vijay-Shanker, 1994 : Obtaining trees from their descriptions : an application to Tree Adjoining Grammars. Computational Intelligence, vol. 10, # 4.

Y. Schabes, A. Joshi, A. Abeillé, 1988 : Parsing strategies with lexicalized grammars : Tree adjoining grammars. COLING'88.

K. Vijay-Shanker, Y. Schabes, 1992 : Structure sharing in Lexicalized TAG. COLING'92.