# LEXICAL AND SYNTACTIC RULES IN A TREE ADJOINING GRAMMAR

Anne Abeillé*
LADL and UFRL
University of Paris 7-Jussieu
abeille@zeta.ibp.fr

## ABSTRACT

Taking examples from English and French idioms, this paper shows that not only constituent structures rules but also most syntactic rules (such as topicalization, wh-question, pronominalization ...) are subject to lexical constraints (on top of syntactic, and possibly semantic, ones). We show that such puzzling phenomena are naturally handled in a 'lexicalized' formalism such as Tree Adjoining Grammar. The extended domain of locality of TAGs also allows one to 'lexicalize' syntactic rules while defining them at the level of constituent structures.

## 1 INTRODUCTION TO 'LEXICALIZED' GRAMMARS

### 1.1 Lexicalizing Phrase Structure rules

In most current linguistic theories the information put in the lexicon has been increased in both amount and complexity. Viewing constituent structures as projected from the lexicon for example avoids the often noted redundancy between Phrase Structure rules and subcategorization frames. Lexical constraints on the well-formedness of linguistic outputs has also simplified the previous transformational machinery.

Collapsing phrase-structure rules into the lexicon is the overt purpose of 'lexicalized' grammars as defined by Schabes, Abeillé, Joshi 1988 : a 'lexicalized' grammar consists of a finite set of elementary structures, each of which is systematically associated with one (or more) lexical item serving as 'head'. These structures are combined with one another with one or more combining operation(s). These structures specify extended domains of locality (as compared to CFGs) over which lexical constraints can be stated. The 'grammar' consists of a lexicon where each lexical item is associated with a finite number of structures for which that item is the 'head'.
We here assume familiarity with Tree Adjoining Grammars, which are naturally 'lexicalized' according to this definition[2]. Each elementary tree is constrained to have at least one terminal at its frontier which serves as 'head' (or 'anchor'). Sentences of a Tag language are derived from the composition of an S-rooted initial tree with other elementary trees by two operations: substitution (the same operation used by context free grammars) or adjunction, which is more powerful.

Schabes, Abeillé, Joshi 1988 show that context free grammars cannot in general be lexicalized (using substitution only as the combining operation). They also show that lexicalized grammars are interesting from a computational point of view since lexicalization simplifies parsing techniques, because the parser uses only a relevant subset of the entire grammar: in a first stage, the parser selects a set of elementary structures associated with the lexical items in the input sentence, and in a second stage the sentence is parsed with respect to this set. As shown by Schabes, Joshi 1989, a parser's performances are thus improved.
We show here that such 'lexicalization' should be extended to other components of the grammar as well, thus challenging the usual distinction between 'lexical' and 'syntactic' rules. Further parsing simplification is therefore expected.

### 1.2 'Lexicalizing' lexical rules

As has often been noticed, rules (or transitivity alternations) such as passive, particle hopping, middle, dative-shift ... are subject to lexical idiosyncrasies. There are of course syntactic and semantic constraints governing such phenomena, but lexical ones seem to be at stake to.

If one considers double objects constructions, passivation of the second NP is regularly ruled out on syntactic grounds. Passivation of the first NP, on the other hand is subject to lexical restrictions as the example of 'cost', opposed to 'envy' or 'spare', shows:

They envy John his new car.
John is envied his new car.
The mistake cost Mary a chance to win.
?* Mary was cost a chance to win.
The judge kindly spared John the ordeal.
John was kindly spared the ordeal.

One might argue that such differences may be due to some semantic constraints, but even verbs with similar meaning may exhibit striking differences. For example, in French, 'regarder' in

* The author wants to thank Yves Schabes, Aravind Joshi, Maurice Gross, Sharon Cote and Tilman Becker for fruitful discussions, and Robert Giannasi and Beatrice Santorini for their help.

2 Categorial grammars are also 'lexicalized'.

its figurative reading (to concern) and 'concerner', which is a true synonym in this context, behave differently:

Cette affaire regarde Jean
* Jean est regardé par cette affaire
Cette affaire concerne Jean
Jean est concerné par cette affaire (M. Gross 1975)

It also seems a lexical phenomenon that "change" but not "transform" allows for ergative alternation in English:

The witch changed/transformed John into a wolf
John changed into a wolf
* John transformed into a wolf (G. Lakoff 1970)

To take another example, dative shift (or there-insertion) is often thought of as applying to a semantically restricted set of verbs (eg verbs of communication or of change of possession, for dative), but this does not predict the difference between 'tell' that allows for it, and 'announce' or 'explain' which do not[3]:

John told his ideas to Mary
John told Mary his ideas
John explained his ideas to Mary
* John explained Mary his ideas

Lexicalist frameworks such as GPSG, which handles such phenomena by metarules (defined on 'lexical' PS rules), or LFG, which defines them at the f-structure level (ie between 'lexical forms') could capture such restrictions. D. Flickinger 1987 handles them explicitly with a hierarchical lexicon in HPSG, considering such rules to hold between two word classes (verbs here) and to apply by default unless they are explicitly blocked in the lexicon.

But all these representations rely on a clear-cut distinction between lexical and syntactic rules and it is not clear how they could be extended to the latter.

## 2 LEXICAL CONSTRAINTS ON SYNTACTIC RULES

The distinction between 'syntactic' rules[4] that do not usually change argument structure nor meaning of the sentence and are supposed to apply regularly on syntactic structures, and 'lexical' rules that alter argument structure, may change the meaning of the predicate and may exhibit some lexical idiosyncrasies, usually overlooks the fact that both are subject to lexical constraints.

There has often been discussions about whether certain rules, (eg passive or extraposition) should be considered of one kind or the other. But it has seldom been realized, to the best of our knowledge, how often 'syntactic' rules are prevented to apply on what seems purely lexical grounds5.

Our discussion crucially relies on idiomatic or semi-idiomatic constructions. We believe that a sizable grammar of natural language, as well as any realistic natural language application, cannot ignore them, since their frequency is quite high in real texts (M. Gross 1989). We first present examples of such lexical constraints on topicalization, pronominalization and wh-question for both English and French idioms. We then show that similar constraints can be found in non idiomatic sentences.

### 2.1 Flexibility of idiomatic constructions

Idioms are usually divided into two sets (eg J. Bresnan 1982, T. Wasow et al. 1982): 'fixed' ones (not subject to any syntactic rule) and flexible ones (presumably subject to all). However, there is quite a continuum between both.

Let us take two French idioms usually considered as "fixed": 'casser la croûte' (to have a bite) and 'demander la lune' (to ask for the impossible). It is true that passivation or wh-question do not apply to either. But pronominalization for the former, cleft-extraction (c'est que) for the latter do[6]:

Paul a cassé la croûte (Paul had a bite)
# Quelle croûte casse-t-il ?
# C'est une petite croûte qu'il a cassée.

---

3 To dismiss 'announce' or 'explain' on the mere basis of their latin origin would not do, since 'offer', which comes from latin as well, does exhibit dative shift.

4 We use the term 'rule' for conveniency. It does not matter for our purpose, whether these phenomena are captured by derivation rules as such or by constraints on the well-formedness of output structures.

5 An interesting exception being Kaplan and Zaenen 1989's proposal that wh-movement and topicalization be constrained at the f-structure level, ie by LFG's 'lexical forms'.

6 # marks that the sentence is not possible with the desired idiomatic interpretation. There may be some variations among speakers about acceptability judgements on such sentences (and on some of the following ones). Such variability is indeed a property of lexical phenomena.

? Paul est en train de casser une petite croûte et j'en casserais bien une aussi. (Paul is having a little bite, I wouldn't mind having one too)

Jeanne demande la lune
# Quelle lune demande-t-elle ?
C'est la lune qu'elle demande !
# Jeanne demande la lune et Paule la demande aussi. (Jeanne is asking for the moon and I'm asking for it too)

These idioms are thus not completely fixed (as opposed to idioms such as 'casser sa pipe' or 'kick the bucket'), and some grammatical function must be assigned to their frozen NPs. But the differences among them are somewhat unexpected: 'casser la croûte' (where the noun can be modified and take several determiners) does not allow for more rules than 'demander la lune' (where the frozen NP is completely fixed).
If one now takes an idiom usually considered as flexible, 'briser la glace' (to break the ice), which does passivize, we notice the same distribution as with 'casser la croûte':

Paul a brisé la glace
# Quelle glace a-t-il brisée ?
# C'est la glace qu'il brise
?? Jean a brisé la glace hier et c'est à moi de la briser aujourd'hui. (Paul broke the ice yesterday and I have to break it today)

Passive is allowed but not wh-question, nor cleft extraction. It is difficult to dismiss such phenomena as rare exceptions. Looking at numerous idioms shows that one combination of such rules is not more frequent than the other. It is also difficult to find a clear semantic principle at work here.

Similar restrictions seem to be at work in English. If one takes some English idioms usually considered as 'flexible' (or even not idiomatic at all): NP0 give hell/the boot to NP1. The main verb 'give' seems to behave syntactically and semantically as in non idiomatic constructions: Dative shift applies and we have the regular semantic alternation : NP1 get hell/the boot (from NP0), with identical meaning. But it is not the case that all expecte rules apply: passive is blocked, pronominalization on the object too:

# Hell was given to Mary (by John)
# The boot was given to Mary (by John)
# Alice gave hell to Paul yesterday and she is giving it to Oscar now.
# Oscar gave the boot to Mary, and he will give it to Bob too.

Syntactic rules may also apply differently to distinct 'flexible' idioms. It is easy to find idioms which do passivize but don't allow for pronominalization or topicalization in the same way:

They hit the bull's eye.
The bull's eye, they hit.
? John hit the bull's eye and Paul hit it too.
They buried the hatchet.
?? The hatchet, they buried.
# John buried the hatchet and Paul buried it/one too.

For relativation also, there might be similar differences:

The strings that Chris pulled helped hime get the job (Wasow et al. 1982)
# The bull's eye that John hit helped him get the job.
# The hatchet that he buried helped him get the job.

Distinguishing between fixed and flexible idioms is thus not sufficient. Because different rules apply to them differently, without a clear hierarchy (contrary to Fraser 1970), one should distinguish as many different types of flexibility as there are possible combinations of such rules. Similarly, if one wants to follow T. Wasow et al. 1982 's suggestion that some kind of compositional semantics should be held responsible for the syntactic flexibility of idioms, as many degrees of compositionality should be defined as there are combinations of syntactic properties. Direct encoding of the latter is thus preferable, and such a semantic 'detour' does not seem to help.

This does not mean that no regularities could be found for idioms' syntax but that they have to be investigated at a more lexical level.

**2.2 Some lexical constraints on non idiomatic constructions**

Going back to non idiomatic constructions, it seems that their syntactic properties may be subject to similar lexical idiosyncrasies.
If one considers double objects constructions, It seems a lexical phenomenon that wh-question on the second NP is allowed with 'give' or 'spare', and not with 'envy' or 'cost', and that topicalization is allowed with 'spare' only:

They envy John his new car
* What/ * Which car do they envy John ?
* This brand new car, everyone envies John

The mistake cost Mary a chance to win
* What/ *Which chance did the mistake cost Mary ?
* This unique chance, the mistake cost Mary

The judge spared John the ordeal
What / Which ordeal did the judge spare John ?
This ordeal, the judge kindly spared John

If one now considers the first NP, topicalization applies differently to:
* Mary, the mistake cost a chance to win
? John, you have always envied his extraordinary luck
John, the judge kindly spared the ordeal

In French, as noted by M. Gross 1969, properties usually thought of as applying to all 'direct objects'(passivation, Que-question and Le-cliticization) may apply in fact unpredictingly. Although the objects of a verb like 'aimer' (love) take objects undergoing the three of them, the object of 'valoir'(be worth) only allows for Que-question and Le-cliticiization, that of 'coûter' (cost) only for Que-question and that of 'passer' (spend (time)) only allows for Le-cliticization:

Ce livre vaut cents francs.
(This book is worth 100 francs)
Ce livre les vaut.
Que vaut ce livre ?
Ce livre coûte cent francs.
(This book costs 100 francs)
* Ce livre les coûte.
Que coûte ce livre ?
Il a passé la nuit à travailler. (He spent the night working)
Il l'a passée à travailler.
*Qu' a-t-il passé à travailler ?[7]

These differences are all the more surprising that 'coûter' and 'valoir' are otherwise very close verbs (same subcategorization frames, same selectional restrictions).
Looking for some generalization principles with which to predict such restrictions should be pursued, but it seems that they will be of a lexical kind.

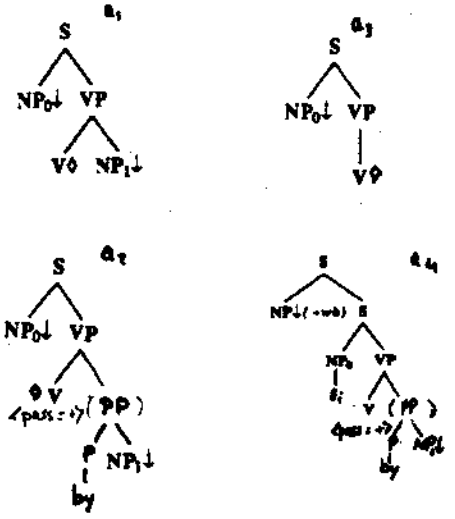
## 3 LEXICALIZED RULES IN A TREE ADJOINING GRAMMAR

### 3.1 Tree Families

Each elementary tree in a Tag is lexicalized in the sense that it is headed by (at least) one lexical item. The category of a word in the lexicon is the name of the tree it selects. We only consider here sentential trees for the sake of simplicity.

What lexical heads select is in fact a set of such elementary trees called a "Tree Family" (Abeillé 1988, Abeillé et al. 1990), each tree corresponding to a certain constituent sructure (initial trees for wh-questions, auxiliary trees for relative clauses...). This is the level at which syntactic generalizations can be stated, since each elementary tree may bear specific constraints independantly of any lexical items[8]. A Tree Family consists in fact of all the constituent structures trees which are possibly allowed for a given predicate[9].

Examples of trees in the n0Vn1 Family (verbs taking two NP arguments) are the following10:




7 ? Quelle nuit a-t-il passée à travailler ? would be better.

8 Further subdividing these Tree Families, similarly to M. Gross 1975's verb tables for French, and to D. Flickinger 1987's word classes for English, will help reduce the number of features, and thus the amount of seemingly idiosyncratic information, associated with each verb. However, as noted by both authors, lexical idiosyncrasies will never be eliminated altogether.

9 Tree Family names (n0V, n0Vn1...) are somewhat similar to 'lexical forms' in LFG in the sense that they capture both the predicate argument structure and the associated grammatical functions (which we note by indices: 0 for subject,1 for first object...). Notice that the Tree Family name does not change when lexical rules apply.

10↓marks a substitution node,◊marks the head. We use here standard TAG trees for commodity of exposition, although recent independent linguistic work suggests to slightly modify them, challenging for example the distinction between VP and V levels (see Abeillé, in preparation).

Each tree is identified by a Tree family name associated to a feature bundle corresponding to the rules it involves. For example, a2, a3 and a4 are respectively marked[11]:

a1 (n0Vn1)
passive = -
Wh-0 = -
Wh-1 = -
erg = -

a2 (n0Vn1)
passive = +
Wh-1 = -
Wh-0 = -

a3 (n0Vn1)
passive = -
Wh-1 = -
erg = +

a4 (n0Vn1)
passive = +
Wh-1 = +
Wh-0 = -

A given tree can belong to several tree families at the same time, which helps factorizing the grammar in a parsing perspective. For example, a3 can also be considered as belonging to the n0V Family (for verbs with one NP argument) with a different feature bundle : passive = -; Wh-0 = -. The lexical items heading the tree constrains its interpretation, eg 'sleep' will interpret a3 as n0V, while 'bake' or 'walk' interpret it as n0Vn1.

Lexical constraints on syntactic and lexical rules are handled by having the head select its own subset of trees in its tree family.

For example, 'resemble' selects only active trees; 'rumored' only passive ones, and 'love' select both[12]:

[love],V : n0Vn1 [erg = -]
[resemble],V: n0Vn1 [passive = -; erg = -]
[rumored],V: n0Vn1 [passive = +]

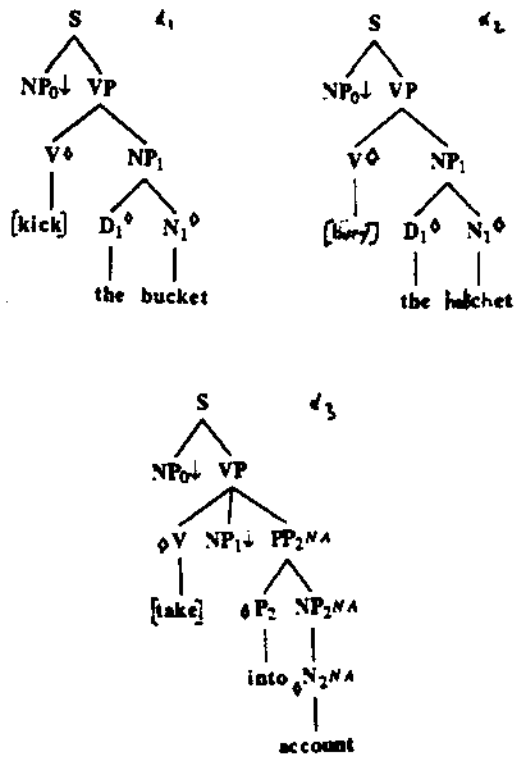
[donate],V; to,P : n0VPn1 [dative = -;erg = -]
[give],V;to,P: n0VPn1 [erg = -]
[spare],V:n0Vn1n2

These features work as follows: when nothing is said about a feature, it means that the predicate selects trees with the feature being plus of minus; when a feature is marked plus, it means that only trees with this feature plus are selected (ie that the corresponding rule is 'forced' to apply).

Such 'lexicalization' of syntactic rules applies similarly in idiomatic and non idiomatic constructions.

### 3.2 Idioms in a Lexicalized Tree Adjoining Grammar

Tags seem a natural framework to represent structures which at the same time are semantically non compositional and should be assigned regular syntactic structures (Abeillé and Schabes 1989, 1990). Idioms are thus made fall into the same grammar as non idiomatic constructions. The only specificity of idioms is that they are selected by a multicomponent head (called 'anchor') and may select elementary trees which are more extended than non idiomatic constructions. Here are some examples of elementary trees for 'kick the bucket', 'bury the hatchet' and 'take NP into account':

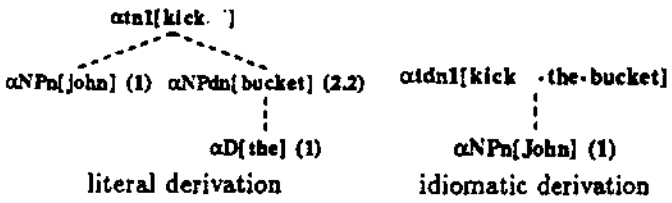



The lexical anchors are respectively 'kick', 'the' and 'bucket' for α1, 'bury', 'the' and 'hatchet' for α2, and 'take', 'into' and 'account' for α3. The idiomatic interpretation of sentences such as 'John kicked the bucket', as opposed to their

11 One might explicitly define materules, or links between such trees: a passive rule for example, changes the feature passive of the tree and intervert the features bearing on N0 and N1. Work is being curretly done along this line with T. Becker, Y. Schabes and K. Vijay Shanker.

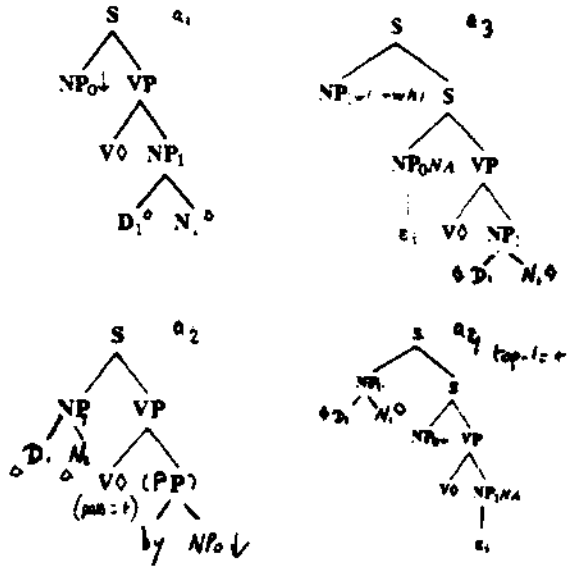
12 We note with square [] the set of inflected forms of a lexical item. For example, [give] = give, gives, gave, giving, given. We use a restriction principle to rule out erg = + whenever passive = + (or dative = +), and vice versa, so the ergative feature does not have to appear in the lexicon for 'rumored'.

literal readings, is straighforwardly based on their distinct derivation trees[13]:



literal derivation idiomatic derivation

Idiomatic and non idiomatic elementary trees are gathered into tree families according to the same principles. Here are some examples of the trees belonging to the Family of idioms with a frozen object (n0VDN1):



Similarly, idioms bear syntactic features constraining the elementary trees of the Tree Family they select. In the n0VDN1 Tree Family, for example, 'kick the bucket' selects only a1, and the trees corresponding to wh-movement on N0; 'bury the hatchet' selects also the trees for passive (and possibly topicalization on N1).

[bury],V;the,D;hatchet,N: n0VDN1 [Wh-N1 = -]
[kick],V;the,D;bucket,N: n0VDN1 [passive = -; Wh-N1 = -; Top-N1 = -]

There are some idioms which exist only in the passive form, or in the question form, and the corresponding trees are directly selected. In French, "être pris par le temps" (to be very busy) lacks its active counterpart (* Le temps prend Jean), and "Quelle mouche a piqué NP ?" (What's eating NP ?) lacks its non interrogative counterpart, although it allows for passive : "Par quelle mouche a-t-il été piqué ?" (M. Gross 1989).

[prendre],V;le,D;temps,N: DN0Vn1[passive = +]
[piquer],V;mouche,N : N0Vn1 [Wh-N0 = +]

Notice that the tree familiy name tells not only about the argument structure but also about the head being multicomponent or not (all head elements are noted with capital letter). Usually, no part of a multicomponent head can be omitted, and trees that are possible for this argument structure but in which all head elements could not be inserted will be ruled out. For example, what-questions (noted Wh-i) are generally disallowed with frozen nominals (and thus not noted for each lexical entry), whereas questions with wh-determiners (noted Wh-Ni) are not:

John took a trip to Spain
# What did John take ?
? Which trip to Spain did John take ? (Abeillé et al. 1990)

In fact, as has been noted by M. Gross 1989 for French, Wh-Ni questions seem to be ruled out when the determiner of the argument is completely fixed, as the following contrasts show:
John spilled the/those beans
John buried the/#this/#a hatchet
Which bean(s) did John spill ?
# Which hatchet did John bury ?

This generalization which can be captured since the Tree family names will be different (with D for frozen determiners, and d for not frozen ones):

[spill],V; [bean],N: n0VdN1
[bury],V; the,D; hatchet,N: n0VDN1

The trees for the Wh-N questions will thus belong only to the corresponding 'd' Families, and not to the 'D' ones.

## CONCLUSION

It has been shown that taking idiomatic or semi-idiomatic constructions into account in a French or English grammar forces one to define some lexical constraints on syntactic rules such as wh-question, pronominalization and topicalization. Such a lexical treatment has been exemplified using Lexicalized Tree Adjoining grammars. An interesting point about TAGs is that, due to their extended domain of locality, they enable one to consider as 'lexical' syntactic rules bearing on

13 The derived trees are the same (modulo the syntactic features explained above).

constituent structures, and not only rules changing the syntactic category of a predicate (as D. Dowty 1978) or rules changing the argument structure of a predicate (as in T. Wasow 1977 or D. Flickinger 1987).

## REFERENCES


Abeillé A., 1988. "Parsing French with Tree Adjoining grammar", Coling'88, Budapest.

Abeillé A., Schabes Y., 1989. "Parsing idioms with Lexicalized Tags", Proceedings of the European ACL meeting, Manchester.

Abeillé A., Schabes Y., 1990. "Non compositional discontinuous constituents in Lexicalized TAG", Proceedings of the international workshop on discontinuous constituency, Tilburg.

Abeillé A., K. Bishop, S. Cote, Y. Schabes, 1990. A lexicalized Tree Adjoining Grammar for English, Technical Report, CIS Dpt, University of Pennsylvania, Philadelphia.

Bresnan J., 1982. "Passive in lexical theory", in Bresnan (ed), *The Mental Representation of Grammatical Relations*, MIT Press.

Dowty D., 1978. "Governed transformations as lexical rules in a Montague grammar", *Linguistic Inquiry*, 9:3.

Flickinger, D. 1987. Lexical rules in the hierarchical lexicon, PhD Dissertation, Stanford University.

Gross M., 1969. "Remarques sur la notion d'objet direct en Français", *Langue française*, n°3, Paris.

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

Gross M., 1989. "Les expressions figées en Français", Technical Report, LADL, University Paris 7, Paris.

Kaplan R., Zaenen A., 1989. "Long distance dependencies, Constituent structure and Functional uncertainty", in Baltin & Kroch (eds), *Alternative Conceptions of Phrase Structures*, Chicago Press.

G. Lakoff, 1970. *Irregularity in Syntax*, Holt, Rinehart and Winston, New York.

Schabes Y., Abeillé A., Joshi A., 1988. "Parsing strategies with 'lexicalized' grammars", Proceedings of COLING'88, Budapest.

Wasow T., 1977. "Transformations and the lexicon", in P. Culicover et al. (eds), *Formal syntax*, Academic press, New York.

Wasow T., Sag I., Nunberg G., 1982. "Idioms: an interim report", Proceedings of the XIIIth international Congress of Linguists, Tokyo.