Parsing Idioms in Lexicalized TAGs *

Anne Abeillé and Yves Schabes
Laboratoire Automatique Documentaire et Linguistique
University Paris 7, 2 place Jussieu, 75005 Paris France
and Department of Computer and Information Science
University of Pennsylvania, Philadelphia PA 19104-6389 USA
abeille/schabes@linc.cis.upenn.edu

ABSTRACT

We show how idioms can be parsed in lexicalized TAGs. We rely on extensive studies of frozen phrases pursued at L.A.D.L.[1] that show that idioms are pervasive in natural language and obey, generally speaking, the same morphological and syntactical patterns as 'free' structures. By idiom we mean a structure in which some items are lexically frozen and have a semantics that is not compositional. We thus consider idioms of different syntactic categories : NP, S, adverbials, compound prepositions... in both English and French.

In lexicalized TAGs, the same grammar is used for idioms as for 'free' sentences. We assign them regular syntactic structures while representing them semantically as one non-compositional entry. Syntactic transformations and insertion of modifiers may thus apply to them as to any 'free' structures. Unlike previous approaches, their variability becomes the general case and their being totally frozen the exception. Idioms are generally represented by extended elementary trees with 'heads' made out of several items (that need not be contiguous) with one of the items serving as an index. When an idiomatic tree is selected by this index, lexical items are attached to some nodes in the tree. Idiomatic trees are selected by a single head node however the head value imposes lexical values on other nodes in the tree. This operation of attaching the head item of an idiom and its lexical parts is called **lexical attachment**. The resulting tree has the lexical items corresponding to the pieces of the idiom already attached to it.

We generalize the parsing strategy defined for lexicalized TAG to the case of 'heads' made out of several items. We propose to parse idioms in two steps which are merged in the two steps parsing strategy that is defined for 'free' sentences. The first step performed during the lexical pass selects trees corresponding to the literal and idiomatic interpretation. However it is not always the case that the idiomatic trees are selected as possible candidates. We require that all basic pieces building the minimal idiomatic expression must be present in the input string (with possibly some order constraints). This condition is a necessary condition for the idiomatic reading but of course it is not sufficient. The second step performs the syntax analysis as in the usual case. During the second step, idiomatic reading might be rejected. Idioms are thus parsed as any 'free' sentences. Except during the selection process, idioms do not require any special parsing mechanism. We are also able to account for cases of ambiguity between idiomatic and literal interpretations.

Factoring recursion from dependencies in TAGs allows discontinuous constituents to be parsed in an elegant way. We also show how regular 'transformations' are taken into account by the parser.

Topics: **Parsing, Idioms**.

1 Introduction to Tree Adjoining Grammars

Tree Adjoining Grammars (TAGs) were introduced by Joshi et al. 1975 and Joshi 1985 as a formalism for linguistic description. Their linguistic relevance was shown by Kroch and Joshi 1985 and Abeillé 1988. A lexicalized version of the formalism was presented in Schabes, Abeillé and Joshi 1988 that makes them attractive for writing computational grammars. They were proved to be

*This work is partially supported (for the second author) by ARO grant DAA29-84-9-007, DARPA grant N0014-85-K0018, NSF grants MCS-82-191169 and DCR-84-10413. We have benefitted immensely from our discussions with Aravind Joshi, Maurice Gross and Mitch Marcus. We want also to thank Kathleen Bishop, and Sharon Cote.

[1] Laboratoire d'Automatique Documentaire et Linguistique, University of Paris 7.

parsable in polynomial time (worst case) by Vijay Shanker and Joshi 1985 and an Earley-type parser was presented by Schabes and Joshi 1988.

The basic component of a TAG is a finite set of elementary trees that have two types: initial trees or auxiliary trees (See Figure 1). Both are minimal (but complete) linguistic structures and have at least one terminal at their frontier (that is their 'head'). Auxiliary trees are also constrained to have exactly one leaf node labeled with a non-terminal of the same category as their root node.

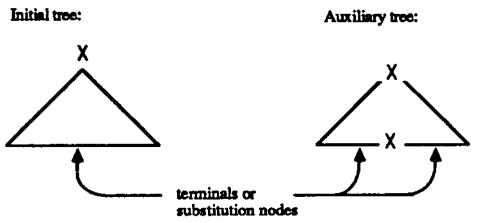


Figure 1: Schematic initial and auxiliary trees

Sentences of the language of a TAG are derived from the composition of an S-rooted initial tree with elementary trees by two operations: substitution or adjunction.

Substitution inserts an initial tree (or a tree derived from an initial tree) at a leaf node bearing the same label in an elementary tree (See Figure 2).[2] It is the operation used by CFGs.

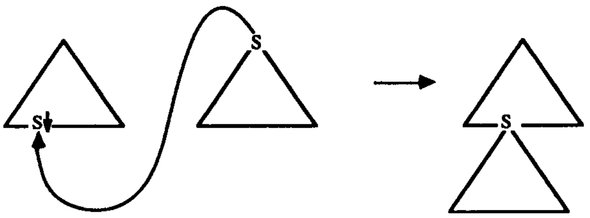


Figure 2: Mechanism of substitution

Adjunction is a more powerful operation: it inserts an auxiliary tree at one of the corresponding node of an elementary tree (See Figure 3).[3]

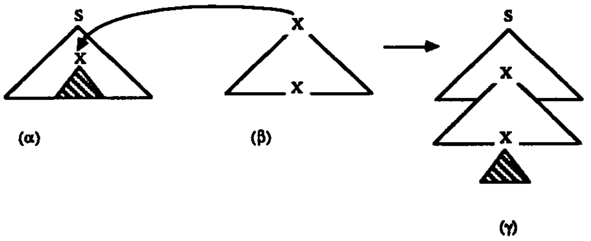


Figure 3: Adjoining

TAGs are more powerful than CFGs but only mildly so (Joshi 1983). Most of the linguistic advantages of the formalism come from the fact that it factors recursion from dependencies. Kroch and Joshi 1985 show how unbounded dependencies can be 'localized' by having filler and gap as part of the same elementary tree and having insertion of matrix clauses provided by recursive adjunctions. Another interesting property of the formalism is its extended domain of locality, as compared to that of usual phrase structure rules in CFG. This was used by Abeillé 1988 to account for the properties of 'light' verb (often called 'support' verb for Romance languages) constructions with only one basic structure (instead of the double analysis or reanalysis usually proposed).

We now define by an example the notion of derivation in a TAG.

Take for example the derived tree in Figure 4.

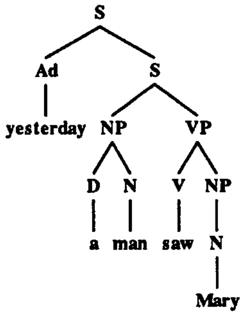


Figure 4: Derived tree for: *yesterday a man saw Mary*

It has been built with the elementary trees in Figure 5.

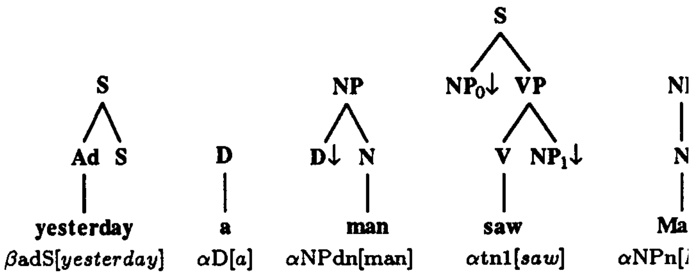


Figure 5: Some elementary trees

Unlike CFGs, from the tree obtained by deriva-

[2] ↓ is the mark for substitution.

[3] At each node of an elementary tree, there is a feature structure associated with it (Vijayshanker and Joshi, 1988). Adjunction constraints can be defined in terms of feature structures and the success or failure of unification.

tion (called the *derived tree*) it is not always possible to know how it was constructed. The *derivation tree* is an object that specifies uniquely how a derived tree was constructed.

The root of the derivation tree is labeled by an S-type initial tree. All other nodes in the derivation tree are labeled by auxiliary trees in the case of adjunction or initial trees in the case of substitution. A tree address is associated with each node (except the root node) in the derivation tree. This tree address is the address of the node in the parent tree to which the adjunction or substitution has been performed. We use the following convention: trees that are adjoined to their parent tree are linked by an unbroken line to their parent, and trees that are substituted are linked by dashed lines.

The derivation tree in Figure 6 specifies how the derived tree was obtained:

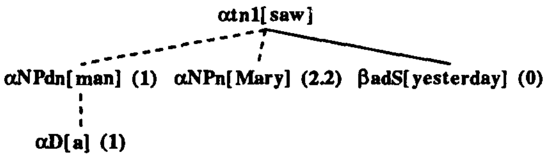


Figure 6: Derivation tree for *Yesterday a man saw Mary*

αD[a] is substituted in the tree αNPdn[man] at node of address 1, αNPdn[man] is substituted in the tree αtnl[*saw*] at address 1, αNPn[*Mary*] is substituted in the tree αtnl[*saw*] at node $2 \cdot 2$ and the tree βadS[*yesterday*] is adjoined in the tree αtnl[*saw*] at node 0.

In a 'lexicalized' TAG, the 'category' of each word in the lexicon is in fact the tree structure(s) it selects.[4] Elementary trees that can be linked by a syntactic or a lexical rule are gathered in a Tree Family, that is selected as a whole by the head of the structure. A novel parsing strategy follows (Schabes, Abeillé, Joshi 1988). In a first step, the parser scans the input string and selects the different tree structures associated with the lexical items of the string by looking up the lexicon. In a second step, these structures are combined together to produce a sentence. Thus the parser uses only a subset of the entire (lexicalized) grammar.

2 Linguistic Properties of Idioms

Idioms have been at stake in many linguistic discussions since the early transformational grammars, but no exhaustive work based on extensive listings of idioms have been pursued before Gross 1982. We rely on L.A.D.L.'s work for French that studied 8000 frozen sentences, 20, 000 frozen nouns and 6000 frozen adverbs. For English, we made use of Freckelton's thesis (1984) that listed more than 3000 sentential idioms. They show that, for a given structure, idiomatic phrases are usually more numerous in the language than 'free' ones. As is well known, idioms are made of the same lexicon and consist of the same sequences of categories as 'free' structures. An interesting exception is the case of 'words' existing only as part of an idiomatic phrase, such as *escampette* in *prendre la poudre d'escampette* (to leave furtively) or *umbrage* in *to take umbrage at NP.*

The specificity of idioms is their **semantic non-compositionality**. The meaning of *casser sa pipe* (to die), cannot be derived from that of *casser* (to break) and that of *pipe* (pipe). They behave semantically as one predicate, and for example the whole VP *casser sa pipe* selects the subject of the sentence and all possible modifiers. We therefore consider an idiom as **one entity in the lexicon.** It would not make sense to have its parts listed in the lexicon as regular categories and to have special rules to limit their distribution to this unique context. If they are already listed in the lexicon, these existing entries are considered as mere homonyms. Furthermore, usually idioms are **ambiguous between literal and idiomatic readings**.

Idioms do not appear necessarily as continuous strings in texts. As shown by M. Gross for French and P. Freckelton for English, more than 15% of sentential idioms are made up of **unbounded arguments**, (e.g. *NP_0 prendre NP_1 en compte*, *NP_0 take NP_1 into account*, *Butter would not melt in NP's mouth*). Discontinuities can also come from the **regular application of syntactic rules**. For example, interposition of adverbs between verb and object in compound V-NP phrases, and interposition of modals or auxiliaries between subject and verb in compound NP-V phrases are very general (Laporte 1988).

As shown by Gazdar et al. 1985 for English, and Gross 1982 for French, most sentential idioms are **not completely frozen and 'transformations' apply to them** much more regularly

[4] The nodes of the tree structures have feature structures associated with them, see footnote 3.

than is usually thought. Freckelton 1984's listings of idiomatic sentences exhibit passivization for about 50% of the idioms comprised of a verb (different from *be* and *have*) and a frozen direct argument. Looking at a representative sample of 2000 idiomatic sentences with frozen objects (from Gross's listings at LADL) yields similar results for passivization and relativization of the frozen argument for French. This is usually considered a problem for parsing, since the order in which the frozen elements of an idiom appear might thus vary.

Recognizing idioms is thus dependent on the whole syntactic analysis and it is not realistic to reanalyze them as simple categories in a preprocessing step.

3 Representing Idioms in Lexicalized TAGs

We represent idioms with the same elementary trees as 'free' structures. The values of the arguments of trees that correspond to a literal expression are introduced via syntactic categories and semantic features. However, the values of arguments of trees that correspond to an idiomatic expression are not only introduced via syntactic categories and semantic features but also directly specified.

3.1 Extended Elementary Trees

Some idioms select the same elementary tree structures as 'free' sentences. For example, a sentential idiom with a frozen subject *il faut* S_1 selects the same tree family as any verb taking a sentential complement (ex: NP_0 *dit* S_1), except that *il* is directly attached in subject position, whereas a 'free' NP is inserted in NP_0 in the case of 'dit' (See Figure 7).

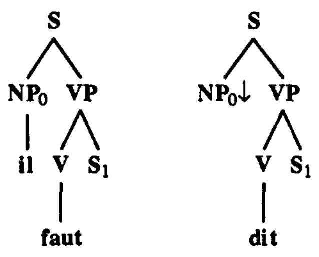


Figure 7: trees for *il faut* and *dit*

Usually idioms require elementary trees that are more expanded. Take now as another example the sentential idiom NP_0 *kicked the bucket*. The corresponding tree must be expanded up to the D_1 and N_1 level. *the* (resp. *bucket*) is directly attached to the D_1 (resp. N_1) node (See Figure 8).

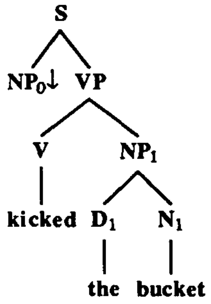


Figure 8: Tree for NP_0 *kicked the bucket*

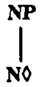
3.2 Multicomponent Heads

In the lexicon, idiomatic trees are represented by specifying the elements of the idiom. An idiom as NP_0 *kicked the bucket* is indexed by a 'head' (*kicked*) which specifies the other pieces of the idiom. Although the idiom is indexed by one item, the pieces are considered as its multicomponent heads.[5]

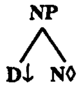
We have, among others, the following entries in the lexicon:[6]

kicked	, V	: Tnl (transitive verb)	(
kicked	, V	: Tdnl[D_1 = the, N_1 = bucket] (idiom)	(
the	, D	: αD	(
bucket	, N	: αNPdn	(
John	, N	: αNP	(

The trees $\alpha NPdn$ and αNPn are:[7]

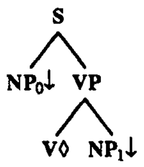


(αNPn)

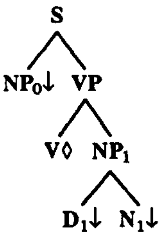


(αNPdn)

Among other trees, the tree αtnl is in the family *Tnl* and the tree αtdnl is in the family Tdnl:



(αtnl)



(αtdnl)

[5] The choice of the item under which the idiom is indexec is most of the time arbitrary.

[6] The lexical entries are simplified to just illustrate hov idiom are handled.

[7] ◊ marks the node under which the head is attached.

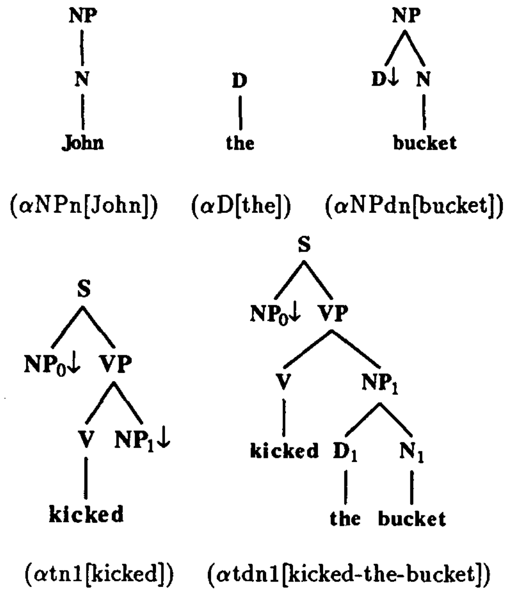


Figure 9: Trees selected for the input
John kicked the bucket

Suppose that the input sentence is *John kicked the bucket.* The first entry for kicked (a) specifies that kicked can be attached under the *V* node in the tree αtdnl (See the tree αtnl[kicked] in Figure 9). However the second entry for kicked (b) specifies that *kicked* can be attached under the *V* node and that *the* must be attached under the node labeled by D_1 and that *bucket* must be attached under the node labeled N_1 in the tree αtnl (See the tree αtdnl[kicked-the-bucket] in Figure 9).

In the first pass, the trees in Figure 9 are be selected (among others).

Some idioms allow some lexical variation, usually between a more familiar and a regular use of the same idiom, for example in French NP_0 *perdre la tête* and NP_0 *perdre la boule* (to get mad). This is represented by allowing disjunction on the string that gets directly attached at a certain position in the idiomatic tree. NP_0 *perdre la tête/boule* will thus be one entry in the lexicon, and we do not have to specify that *tête* and *boule* are synonymous (and restrict this synonymy to hold only for this context).

3.3 Selection of Idiomatic Trees

We now explain how the first pass of the parser is modified to select the appropriate possible candidates for idiomatic readings. Take the previous example, *John kicked the bucket.* The verb *kicked* will select the tree αtdnl[kicked-the-bucket] for an idiomatic reading. However, the values of the determiner and the noun of the object noun phrase are imposed to be respectively *the* and *bucket.* The determiner and the noun are attached to the tree αtdnl[kicked-the-bucket], however the tree αtdnl[kicked-the-bucket] is selected if the words *kicked, the* and *bucket* appear in the input string at position compatible with the tree αtdnl[kicked-the-bucket]. Therefore they must respectively appear in the input string at some position i, j and k such that $i < j < k$. If it is not the case, the tree αtdnl[kicked-the-bucket] is not selected. This process is called **lexical attachment.**

For example the word *kicked* in the following sentences will select the idiomatic tree αtdnl[kicked-the-bucket]:

John kicked the bucket *(s1)*
John kicked the proverbial bucket *(s2)*
John kicked the man who was
carrying the bucket *(s3)*

The parser will accept sentences *s1* and *s2* as idiomatic reading but not the sentence *s3* since the tree αtdnl[kicked-the-bucket] will fail in the parse. In the following sentence the word *kicked* will not select the idiomatic tree αtdnl[kicked-the-bucket]:

John kicked Mark *(s4)*
John kicked a bucket *(s5)*
John who was carrying a bucket
kicked the child *(s6)*
What did John kick? *(s7)*

This test cuts down the number of idiomatic trees that are given to the parser as possible candidates. Thus a lot of idioms are ruled out before starting the syntactic analysis because we know all the lexical items at the end of the first pass. This is important because a given item (e.g. a verb) can be the head of a large number of idioms (Gross 82 has listed more than 50 of them for the verb *manger,* and *prendre* or *avoir* yield thousands of them). However, as sentence *s3* illustrates, the test is not sufficient.

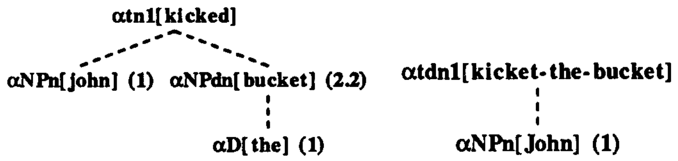
What TAGs allow us to do is to define multicomponent heads for idiomatic structures without requiring their being contiguous in the input string. The formalism also allows us to access directly the different elements of the compound without flattening the structure. As opposed to CFGs, for example, direct dependencies can be expressed between arguments that are at different levels of depth in the tree without having to pass features across local domains. For example, in NP_0 *vider DET sac* (to express all of one's secret thoughts), the determiner of the object *sac* has to be a possessive and agree in person with the subject : *je vide mon sac, tu vides ton sac...*

In NP_0 *dire DET quatre verités a* NP_2 (to tell someone what he really is), the determiner of the object *verités* has to be a possessive and agree in person with the second object NP_2 : *je te dis tes quatre verités, je lui dis ses quatre verités.*

4 Literal and Idiomatic Readings

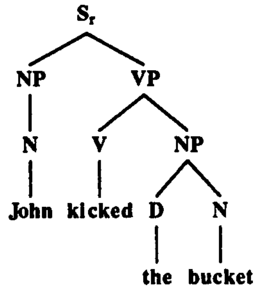
Our representation expresses correctly that idioms are semantically non-compositional. Trees obtained by lexical attachment of several lexical items act as one syntactic unit and also one semantic unit.

For example, the sentence *John kicked the bucket* can be parsed in two different ways. One derivation is built with the trees: αtnl[kicked] (transitive verb), αNPn[John], αD[the] and αNPn[bucket] . It corresponds to the literal interpretation; the other derivation is built with the trees: αtdnl[kicked-the-bucket] (idiomatic tree) and αNPn[John] (John):



literal derivation idiomatic derivation

However, both derivations have the same derived tree:



The meaning of *kicked the bucket* in its idiomatic reading cannot be derived from that of *kicked* and *the bucket*. However, by allowing arguments to be inserted by substitution or adjunction (in for example αtdnl[kicked-the-bucket]), we represent the fact that NP_0 *kicked the bucket* acts as a syntactic and semantic unit expecting one argument NP_0. Similarly, NP_0 *kicked* NP_1 in αtnl[kicked] acts as a syntactic and semantic unit expecting two arguments NP_0 and NP_1. This fact is reflected in the two derivation trees of *John kicked the bucket.*

However, the sentential idiom 'il faut S_1', is not parsed as ambiguous, since *faut* has only one entry (that is idiomatic) in the lexicon. When a certain item does not exist except in a specific idiom, for example *umbrage* in English, the corresponding idiom *to take umbrage of NP* will not be parsed as ambiguous. The same holds when a item selects a construction only in an idiomatic expression. *Aller*, for example, takes an obligatory *PP* (or adverbial) argument in its non-idiomatic sense. Thus the idiom:

aller son train (to follow one's way)

is not parsed as ambiguous since there is no free NP_0 *aller* NP_1 structure in the lexicon.

We also have ambiguities for compound nominals such as *carte bleue*, meaning either *credit card* (idiomatic) or *blue card* (literal), and for compound adverbials like *on a dime*: *John stopped on a dime* will mean either that he stopped in a controlled way or on a 10 cent coin.

Structures for literal and idiomatic readings are both selected by the parser in the first step. Since syntax and semantics are processed at the same time, the sentence is analyzed as ambiguous between literal and idiomatic interpretations. The derived trees are the same but the derivation trees are different. For example, the adjective *bleue* selects an auxiliary tree that is adjoined to *carte* in the literal derivation tree, whereas it is directly attached in a complex initial tree in the case of idiomatic interpretation.

All frozen elements of the idiom are directly attached in the corresponding elementary trees and do not have to exist in the lexicon. They are thus distinguished from 'free' arguments that select their own trees (and their own semantics) to be substituted in a standard sentential tree. Therefore we distinguish two kinds of semantic operations: substitution (or adjunction) corresponds to a compositional semantics; direct attachment on the other hand, makes different items behave as one semantic unit.

One should notice that non-idiomatic readings are not necessarily literal readings. Since feature structures are used for selectional restrictions on arguments, metaphoric readings can be taken into account (Bishop, Cote and Abeillé 1989).

We are able to handle different kinds of semantic non-compositionality, and we do not treat as idiomatic all cases of non-literal readings.

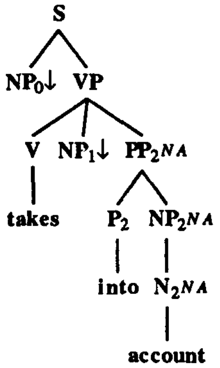


Figure 10: Tree for NP_0 *takes* NP_1 *into account*

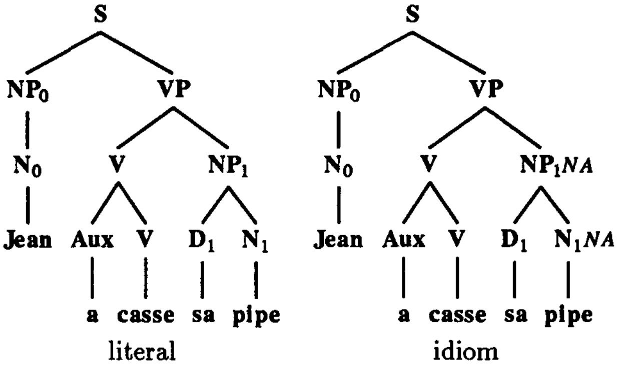


Figure 11: *Jean a cassé sa pipe*

5 Recognizing Discontinuous Idioms

Parsing flexible idioms has received only partial solutions so far (Stock 1987, Laporte 1988). Since TAGs factor recursion from dependencies, discontinuities are captured straightforwardly without special devices (as opposed to Johnson 1985 or Bunt et al. 1987). We distinguish two kinds of discontinuities: discontinuities that come from internal structures and discontinuities that come from the insertion of modifiers.

5.1 Internal Discontinuities

Some idioms are internally discontinuous. Take for example the idioms NP_0 *prendre* NP_1 *en compte* and NP_0 *takes* NP_1 *into account* (see Figure 10).[8]

The discontinuity is handled simply by arguments (here NP_0 and NP_1) to be substituted (or adjoined in some cases) as any free sentences. The internal structures of arguments can be unbounded.

5.2 Recursive Insertions of Modifiers

Some adjunctions of modifiers may be ruled out in idioms or some new ones may be valid only in idioms. If the sentence is possibly ambiguous between idiomatic and literal reading, the adjunction of such modifiers force the literal interpretation. For example, in NP_0 *casser sa pipe* (to die) , the NP_1 node in the idiomatic tree bears a null adjunction constraint (NA). The sentence *Il a cassé sa pipe en bois* (he broke his wooden pipe) is then parsed as non-idiomatic. This NA constraint will be the only difference between the two derived trees (See Figure 11): *Jean a cassé sa pipe* (literal) and *Jean a cassé sa pipe* (idiomatic).

But most idioms allow modifiers to be inserted in them. Each modifier can be unbounded (e.g. with embedded adjunct clauses) and their insertion is recursive. We treat these insertion by adjunction of modifiers in the idiomatic tree. However constraint of adjunction and feature structure constraints filter out partially or totally the insertion of modifiers at each node of an idiomatic tree. In a TAG, the internal structure of idioms is specified in terms of a tree, and we can get a unified representation for such compound adverbials as *à la limite* and *à l'extreme limite* (if there is no other way) or such complex determiners as *a bunch of* (or *la majorité de NP*) and *a whole bunch of NP* (resp. *la grande majorité de NP*) that will not have to be listed as separate entries in the lexicon. The adjective whole (resp. grande) adjoins to the noun *bunch* (resp. *majorité*), as to any noun. Take *a bunch of NP*. The adjective *whole* adjoins to the noun *bunch* as to any noun (See Figure 12) and builds *a whole bunch of.*

In order to have a modifier with the right features adjoining at a certain node in the idiom, we associate some features with the head of the idiom (as for heads of 'free' structures) but also with elements of the idiom that are directly attached. Unification equations, such as those constraining agreement, are the same for trees selected by idioms and trees selected by 'free' structures. Thus only *grande* that is feminine singular, and not *grand* for example, can adjoin to *majorité* that is feminine singular. In *il falloir NP*, the frozen subject *il* is marked 3rd person singular, and only an auxiliary like va (that is 3rd person singular) and not vont (3rd person plural) will be allowed

[8] *NA* expresses the fact that the node has null adjunction constraint

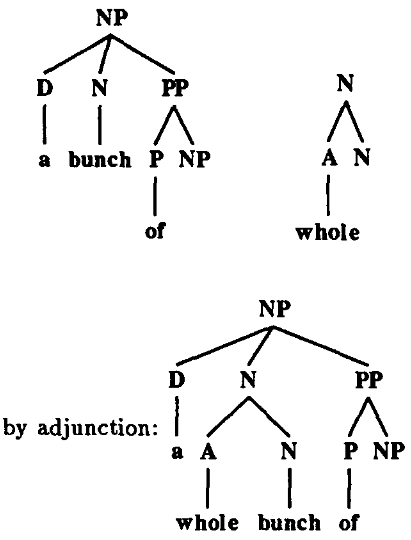


Figure 12: Trees for *a whole bunch of*

to adjoin to the VP: *il va falloir S_1* and not *il vont falloir S_1*.

As another example, an idiom such as *la moutarde monte au nez de NP* (NP looses his temper) can be represented as contiguous in the elementary tree. Adjunction takes place at any internal node without breaking the semantic unity of the idiom. For example, an adjunct clause headed by *aussitôt* can adjoin between the frozen subject and the rest of the the idiom in *la moutarde monter au nez de NP_2 : la moutarde, aussitôt que Marie entra, monta au nez de Max* (Max, as soon as Marie got in, lost his temper). Similarly, auxiliaries adjoin between frozen subjects and verbs as they do to 'free' VPs: *There might have been a box on the table* is parsed as being derived from the idiom : *there be NP_1 P NP_2*.

It should be noted that when a modifier adjoins to an interior node of an idiom, there is a semantic composition between the semantics of the modifier and that of the idiom as a whole, no matter at which interior node the adjunction takes place. For example, in *John kicked the proverbial bucket* semantic composition happens between the 3 units *John*, *kick-the-bucket*, and *proverbial*.[9] Semantic composition will be done the same way if an adjunct clause were adjoined into the *VP*. In *John kicked the bucket, as the proverb says*, composition will happen between *John*, *kick-the-bucket*, and the adjunct clause considered as one predicate *as-proverb-say*:

Therefore parsing flexible idioms is reduced to the general parsing of TAGs (Schabes and Joshi 1988).

6 Tree Families and Application of 'Transformations' to Idioms

As in the case of predicates in lexicalized TAGs, sentential idioms are represented as selecting a set of elementary trees and not only one tree. These tree families gather all elementary trees that are possible syntactic realizations of a given argument structure. The family for transitive verbs, for example, is comprised of trees for wh-question on the subject, wh-question on the object, relativization on the subject, relativization on the object, and so on. In the first pass, the parser loads all the trees in the tree family corresponding to an item in the input string (unless certain trees in that family do not match with the feature of the head in the input string).

The same tree families are used with idioms. However some trees in a family might be ruled out by an idiom if it does not satisfy one of the three following requirements.

First, the tree must have slots in which the pieces of the idiom can be attached.[10] If one distinguishes syntactic rules that keep the lexical value of an argument in a sentence (e.g. topicalization, cleft extraction, relativization...), and syntactic rules that do not (deleting the node for that argument, or replacing it by a pronoun or a wh-element; e.g.: wh-question, pronominalization), it can be shown that usually only the former applies to frozen elements of an idiom. If you take the idiom *bruler un feu* (to run a (red) light), relativization and cleft extraction, but not wh-question, are possible on the noun *feu*, with the idiomatic reading:

Le feu que Jean a brulé.
C'est un feu que Jean a brulé.
* *Que brule Jean ?*

Second, if all the pieces of an idiom can be attached in a tree, the order imposed by the tree must match with the order in which the pieces appear in the input string. Thus, if *enfant* appears before *attendre* in the input string, the hypothesis for an idiomatic reading will be made but only the trees corresponding to relativization, cleft ex-

[9] This is the case of a modifier where adjoining is valid only for the idiom.

[10] This requirement is independent of the input string.

traction, topicalization in which *enfant* is required to appear before *attendre* will be selected. But if the string *enfant* is not present at all in the input string, the idiomatic reading will not be hypothesized, and trees corresponding to *qui attend-elle* will never be selected as part of the family of the idiom *attendre un enfant.*

Third, the features of the heads of an idiom must unify with those imposed on the tree (as for 'free' sentences). For example, it has to be specified that *bucket* in *to kick the bucket* does not undergo relativization nor passivization, whereas *tabs* in *to keep tabs on NP* does. It is well known that even for 'free' sentences application of the passive, for example, has somehow to be specified for each transitive verbs since there are lexical idiosyncrasies.[11] The semantics of the passive *tabs were kept on NP by NP* is exactly the same as that of the active *NP keep tabs on NP*, since different trees in the same tree families are considered as (semantically) synonymous.

7 Conclusion

We have shown how idioms can be processed in lexicalized TAGs. We can access simultaneously frozen elements at different levels of depths where CFGs would either have to flatten the idiomatic structure (and lose the possibility of regular insertion of modifiers) or to use specific devices to check the presence of an idiom. We can also put sentential idioms in the same grammar as free sentences. The two pass parsing strategy we use combining with an operation of direct attachment of lexical items in idiomatic trees, enables us to cut down the number of idiomatic trees that the parser takes as possible candidates. We easily get possibly idiomatic and literal reading for a given sentence. The only distinctive property of idioms is the non-compositional semantics of their frozen constituents. The extended domain of locality of TAGs allows the two problems of internal discontinuity and of unbounded interpositions to be handled in a nice way.

References

Abeillé, Anne, 1988. Parsing French with Tree Adjoining Grammar: some Linguistic Accounts. In *Proceedings of the 12th International Conference on Computational Linguistics (Coling'88).* Budapest.

Bishop, Kathleen M.; Cote, Sharon; and Abeillé, Anne, 1989. *A Lexicalized Tree Adjoining Grammar for English.* Technical Report, Department of Computer and Information Science, University of Pennsylvania.

Bunt, et al., 1987. Discontinuous Constituents in Trees, Rules and Parsing. In *Proceedings of European Chapter of the ACL'87.* Copenhagen.

Freckelton, P., 1984. *Une Etude Comparative des Expressions Idiomatiques de l'Anglais et du Français.* PhD thesis, Thèse de troisième cycle, University Paris 7.

Gazdar, G.; Klein, E.; Pullum, G. K.; and Sag, I. A., 1985. *Generalized Phrase Structure Grammars.* Blackwell Publishing, Oxford. Also published by Harvard University Press, Cambridge, MA.

Gross, Maurice, 1982. Classification des phrases figées en Français. *Revue Québécoise de Linguistique* 11(2).

Johnson, M., 1985. Parsing with discontinuous elements. In *Proceedings of the 23rd ACL meeting.* Chicago.

Joshi, Aravind K., 1985. How Much Context-Sensitivity is Necessary for Characterizing Structural Descriptions—Tree Adjoining Grammars. In Dowty, D.; Karttunen, L.; and Zwicky, A. (editors), *Natural Language Processing—Theoretical, Computational and Psychological Perspectives.* Cambridge University Press, New York. Originally presented in a Workshop on Natural Language Parsing at Ohio State University, Columbus, Ohio, May 1983.

Joshi, A. K.; Levy, L. S.; and Takahashi, M., 1975. Tree Adjunct Grammars. *J. Comput. Syst. Sci.* 10(1).

Kroch, A. and Joshi, A. K., 1985. *Linguistic Relevance of Tree Adjoining Grammars.* Technical Report MS-CIS-85-18, Department of Computer and Information Science, University of Pennsylvania.

Laporte, E., 1988. Reconnaissance des expressions figées lors de l'analyse automatique. *Langages* . Larousse, Paris.

Schabes, Yves and Joshi, Aravind K., 1988. An Earley-Type Parsing Algorithm for Tree Adjoining Grammars. In 26th *Meeting of the Association for Computational Linguistics.* Buffalo.

Schabes, Yves; Abeillé, Anne; and Joshi, Aravind K., 1988. Parsing Strategies with 'Lexicalized' Grammars: Application to Tree Adjoining Grammars. In *Proceedings of the 12th International Conference on Computational Linguistics.*

Stock, O., 1987. Getting Idioms in a Lexicon Based Parser's Head. In *Proceedings of ACL'87.* Stanford.

Vijay-Shanker, K. and Joshi, A. K., 1985. Some Computational Properties of Tree Adjoining Grammars. In 23rd *Meeting of the Association for Computational Linguistics,* pages 82–93.

Vijay-Shanker, K. and Joshi, A.K., 1988. Feature Structure Based Tree Adjoining Grammars. In *Proceedings of the 12th International Conference on Computational Linguistics (Coling'88).* Budapest.

[11] Unless one thinks that some regularity might show up if one distinguishes different kinds of direct complements with thematic roles.