lemmatization9903/Revised 990530/991226

LEMMATIZATION OF ENGLISH VERBS IN COMPOUND TENSES

Maurice Gross

Laboratoire d'Automatique Documentaire et Linguistique[1]

University Paris 7

In general, lemmatization is performed on verbs conjugated by means of suffixes, that is, on simple verbs. In English, we have lemmas and paradigms such as:

to work: *work, works, worked, working*
to eat: *eat, eats, ate, eaten, eating*

But, there is no reason why forms composed of an auxiliary verb and a non finite verb such as *is working* or *has eaten* should not be part of these paradigms and lemmatized accordingly; after all, they are full-fledged conjugated forms. From the point of view of parsing, there is a difficulty in recognizing compound tenses, because inserts may occur between their parts:

Jo is today working on an essay
Bob has not much eaten

Hence, inserts have to be recognized in order to bring together the parts of a compound verb.

1. Inserts

Inserts are of various types, ranging from simple adverbs to complex combinations of adverbial phrases; some of these phrases can be sentential, in which case, their length is unbounded and their analysis requires the full power of a sentence parser. Nonetheless, it is possible to construct detailed grammars for most adverbial phrases and a study of corpora by C. Fairon 1999 has shown that the number of compound verbs which are 'disrupted' by inserts is small and that moreover long inserts occur in texts quite rarely.

[1] UMR N°7546 du CNRS.

The negation *not* has a special status as an insert: it only occurs between auxiliaries and verbs. *Not* interferes in various ways with the auxiliary system. Firstly, it is merged into a simple form *cannot* with the modal *can* and into many contracted forms *(isn't, shouldn't,* etc.). Secondly, it is introduced with most verbs by means of the auxiliary *do. Do* has itself no compound tenses and is thus limited to the forms *do, does* and *did.* Thirdly, with some auxiliary verbs (e.g. *to be, to have*) and with *to dare* and *to need, not* is introduced without *do,* it must then have a special treatment. As a consequence, we lemmatize negative verbs, such as *do not V* as *V*s in the negative form, hence we treat negative verbs as compounds. *Never* has some of the properties of *not,* but is more of an adverbial, only requiring the auxiliary *do* in sentences with subject inversions such as:

Never did Bob accept the situation
= *Bob never accepted the situation*

The negative word *neither* is tagged Conjunction (*CONJ*) in the electronic dictionary system DELA, hence, not recognized as an adverbial. But it can interrupt a compound verbal sequence in the same way adverbials do, as a consequence, we introduced *neither* in the graphs of inserts, and by continuity, we added *either.*

Resources

To parse adverbials, the following resources are available:

- in the electronic dictionary of simple inflected words DELAF, adverbs (e.g. *again, furiously*) are marked with the symbol *ADV*;
- frozen adverbs (e.g. *here and there, from time to time*), have been represented in a lexicon-grammar (M. Gross 1991), they are used by the parsing procedure with the same tag *ADV*;
- various inserts, such as time adverbials and some sentential inserts have been described in terms of local grammars. Again, the parser treats them like the other *ADV* forms.

In our local grammars, we use three types of inserts, depending on the presence or not of the negations *not* and *never*, these inserts are noted **Insert**, **InsNot** and **Ins** (*cf.* figure 1). When all these occurrences of adverbials are parsed, practically all compound forms of verbs found with inserts in corpora can be recognized.

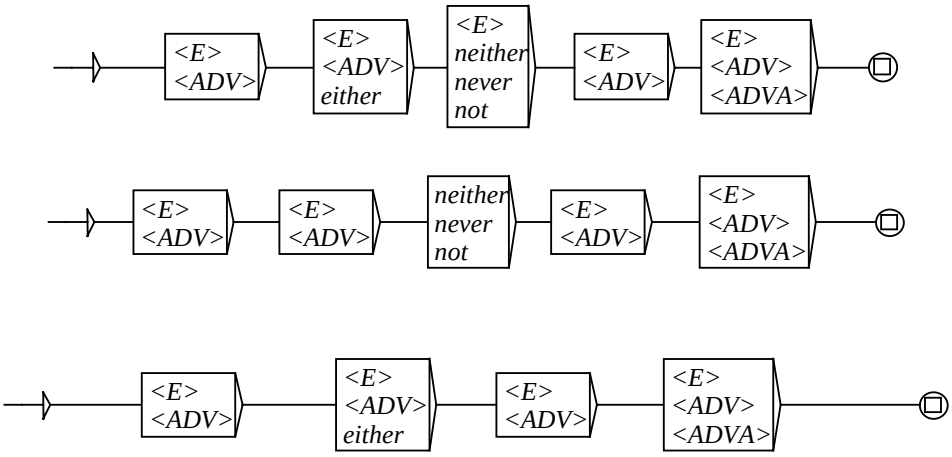


Figure 1**. Insert**, **InsNot** and **Ins**

2. Auxiliary verbs

On a crude intuitive basis, one can consider auxiliary verbs *Vaux* as verbs that add some meaning to the meaning of a main verb, or rather, to the meaning of a given subject-verb complex, noted N_0 *V*. The following examples where the main verb is *sleep* (with subject *Bob*) contain such auxiliary verbs:

Bob (is + ought + begins + wants) to sleep
Bob (is + went on + thought of) sleeping

More generally, our lemmatization process recognizes patterns of the following form: [2]

$(Vaux^0)^n\ V^0$, with n > 0 is the number of auxiliaries of the sequence and where V^0 is the lemmatized verb, the superscript ^{0} refers to the subscript of N_0, our notation for subject noun phrases.

The length of an auxiliary sequence, namely the number *n* of *Vaux*s is not limited, but all *Vaux*s and *V* must share the same subject N_0, hence, the result of lemmatization for the two following sentences is as marked in bold characters:

[2] This notation ignores governing prepositions and governed moods, these constraints are fully taken into account in the various graphs of the grammar (e.g. figures 2, 3, 4).

*People **have attempted to go** to the beach*
*People **have recommended** to go to the beach*

Moreover, there is no reason to limit the analysis to morphologically simple auxiliaries: other forms consisting of adjectives built with *to be,* of nouns built with *to be, to have* or other support verbs and of frozen forms are auxiliaries from the same semantic point of view:

Bob (is unable + has a right + found a way) to sleep
Bob (is on the verge of + has trouble + came close to) sleeping

As a consequence, defining the notion of auxiliary verb on a formal basis is a complex process. From a syntactic point of view, our examples present sharp differences, hence, we classified them according to elementary grammatical categories, namely, categories generally found in textbooks. Although categories of auxiliary verbs are described in all kinds of grammars, from high school textbooks to ambitious academic studies, constructing a full list for them is not an easy task in the absence of coherent definitions. But even when operational definitions are given, that is, syntactic definitions, constructing a list of lexical items is an exercise that has never been attempted. One can safely predict that due to the variety of interests competing on the market of linguistic theories, no agreement is possible today. We nonetheless propose a concrete classification of these verbs (i.e. lists), largely based on the various descriptions available in current grammars.

By definition, an auxiliary verb governs another verb in either one of the three forms: past participle, infinitive or *-ing* form. We have subdivided auxiliary verbs *Vaux* into five categories:

- tense auxiliaries,
- passive auxiliaries,
- aspectual verbs (noted *VAsp,* e.g. *to begin,* figure 2),
- modality verbs (noted *VMod,* e.g. *to attempt,* figure 3), different from modal verbs which are considered as the tense auxiliaries,
- verbs with sentential complements (noted *VS,* e.g. *to hope,* figure 4).

We consider the first three categories as reasonably complete, but the limit between *VMod*s and *VS*s is difficult to assess. We have only listed a limited number of verbs of the last category, *VS,* they are the most numerous (in the thousands), their lists should thus be substantially extended. The INTEX system

can use these categories to tag verbs (M. Silberztein 1993).

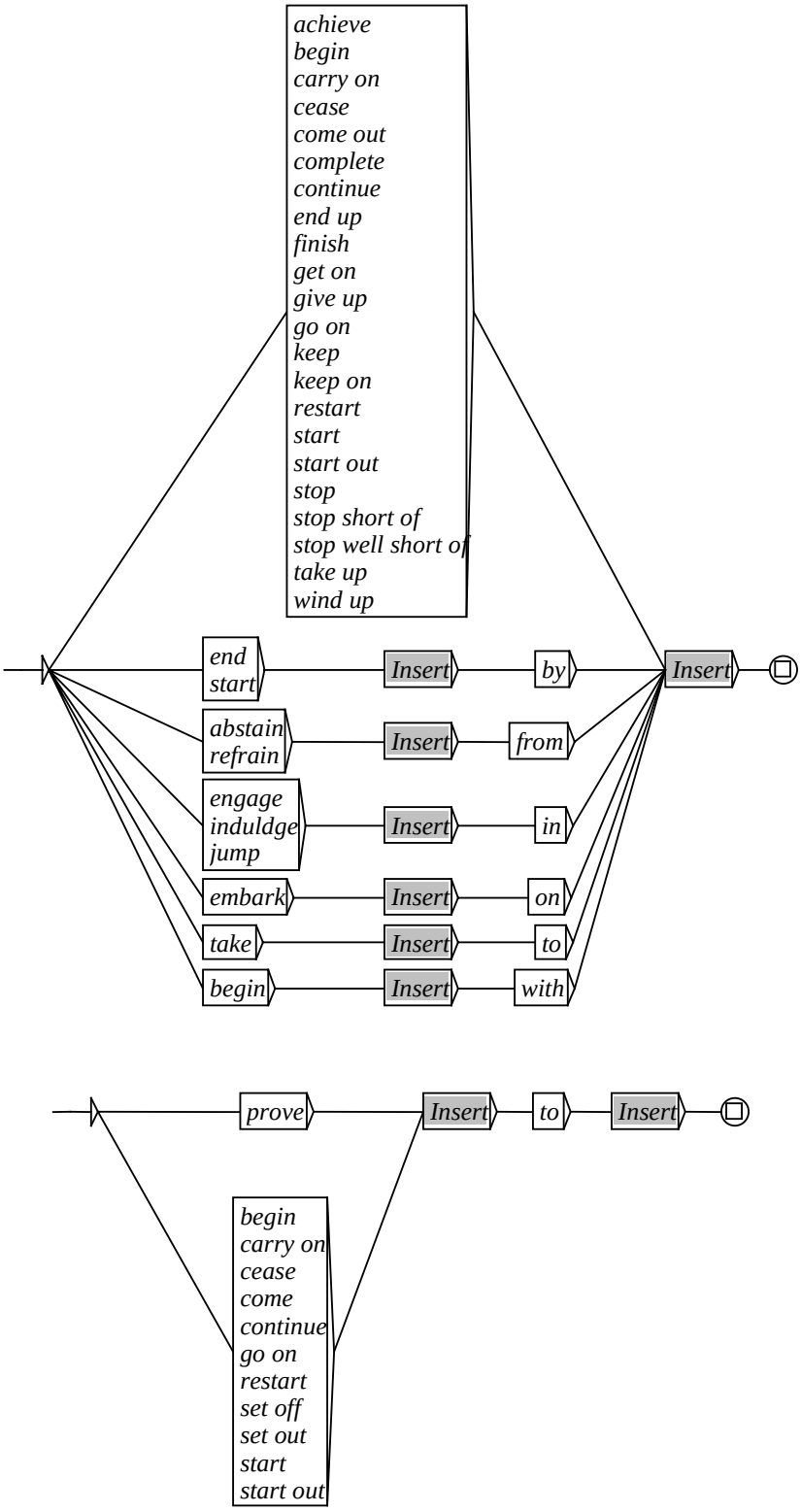
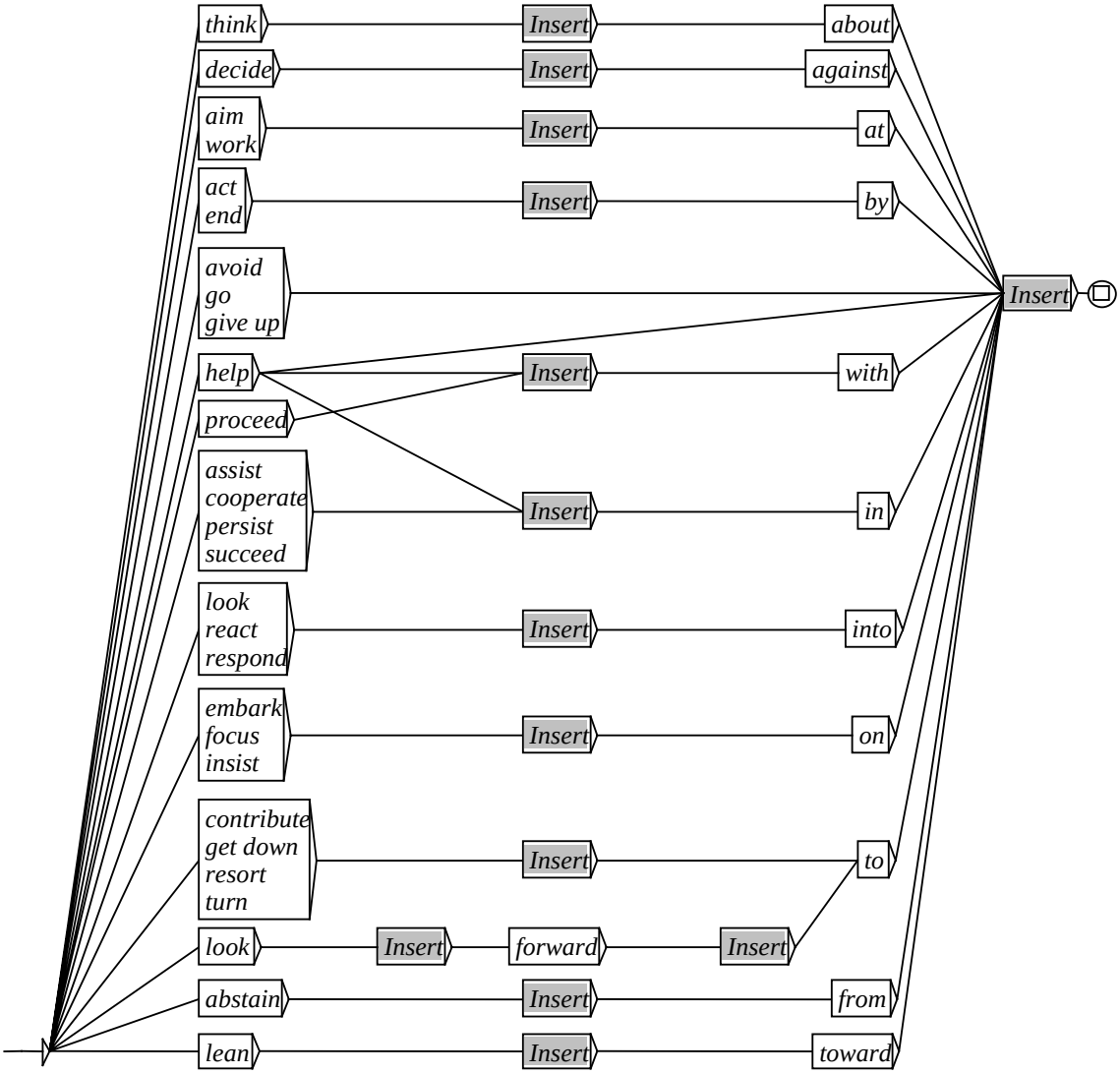


Figure 2. **VAspPrepVing** and **VAspToV**



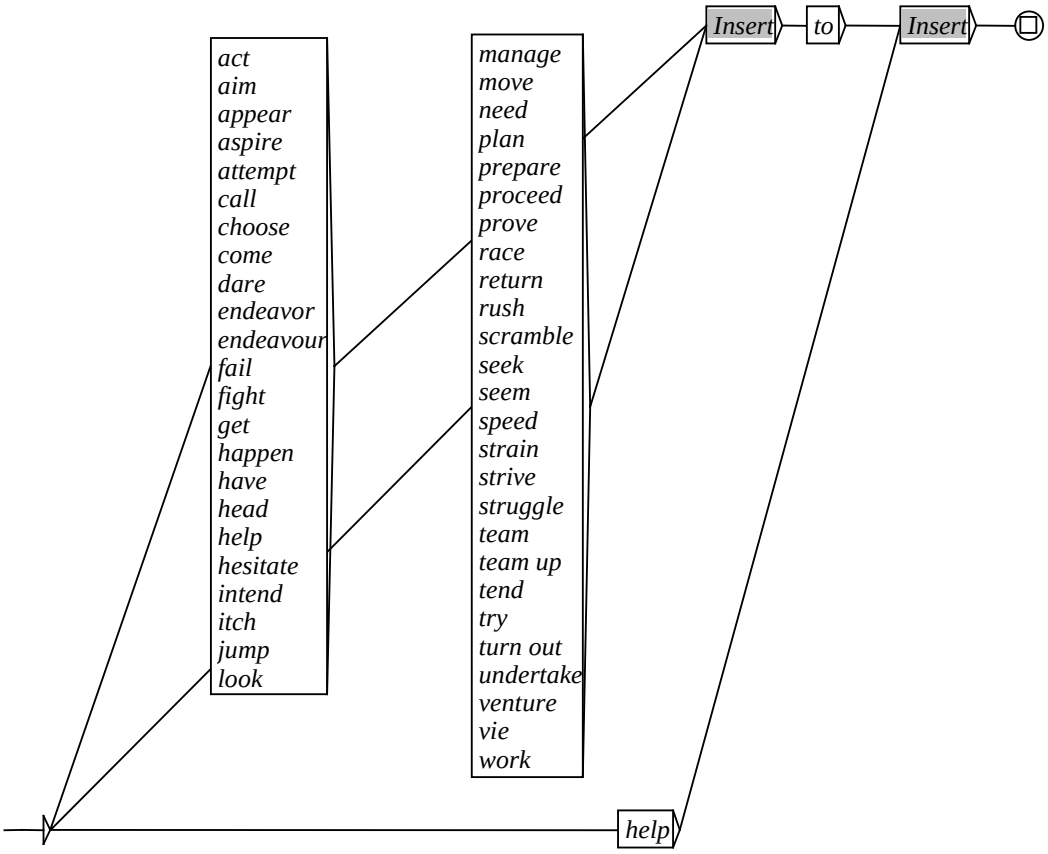


Figure 3. **VModPrepVing** and **VModToV**

2.1. ***Tense auxiliaries*** are rather well defined, from a formal point of view. Criteria commonly used to delimit this set are:

- lack of autonomous constructions, that is, constructions where the auxiliary verb would be a main verb,
- negation not constructed with the auxiliary *do*, (*be*, *have*, *will*, *dare*, *need*),
- defective tenses (*e.g.* the modal verbs, *used to*),
- transparency to the semantic selection of subjects. 'Real' auxiliaries occur, independently of distributional constraints between subject and verb. The following series of examples presents a variety of subject-verb constraints:

> *It began raining*, *Jo began reading*, *The pipe began leaking*
> *That Jo keeps protesting begins to annoy Bob*

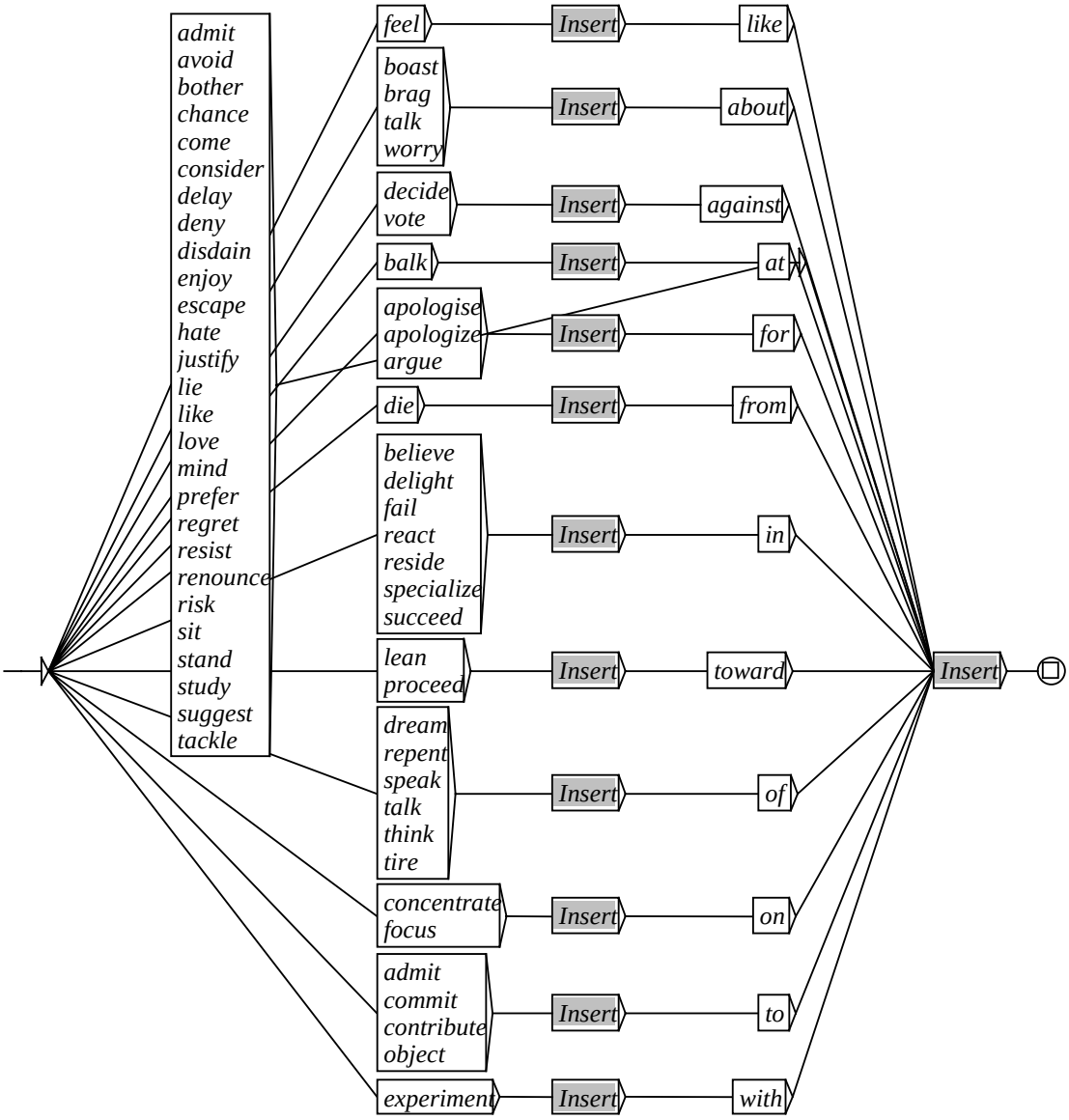
The presence of *to begin* does not interfere with the distributional constraints. In the sentence:

Jo attempted to read

the subject of *Vaux* =: *to attempt* has to be **human**, conflicting with **non-human** subjects, hence:

?**It attempted to rain* ?**The pipe attempted to leak*

As a consequence, *to attempt* should be less of an auxiliary than *to begin*.



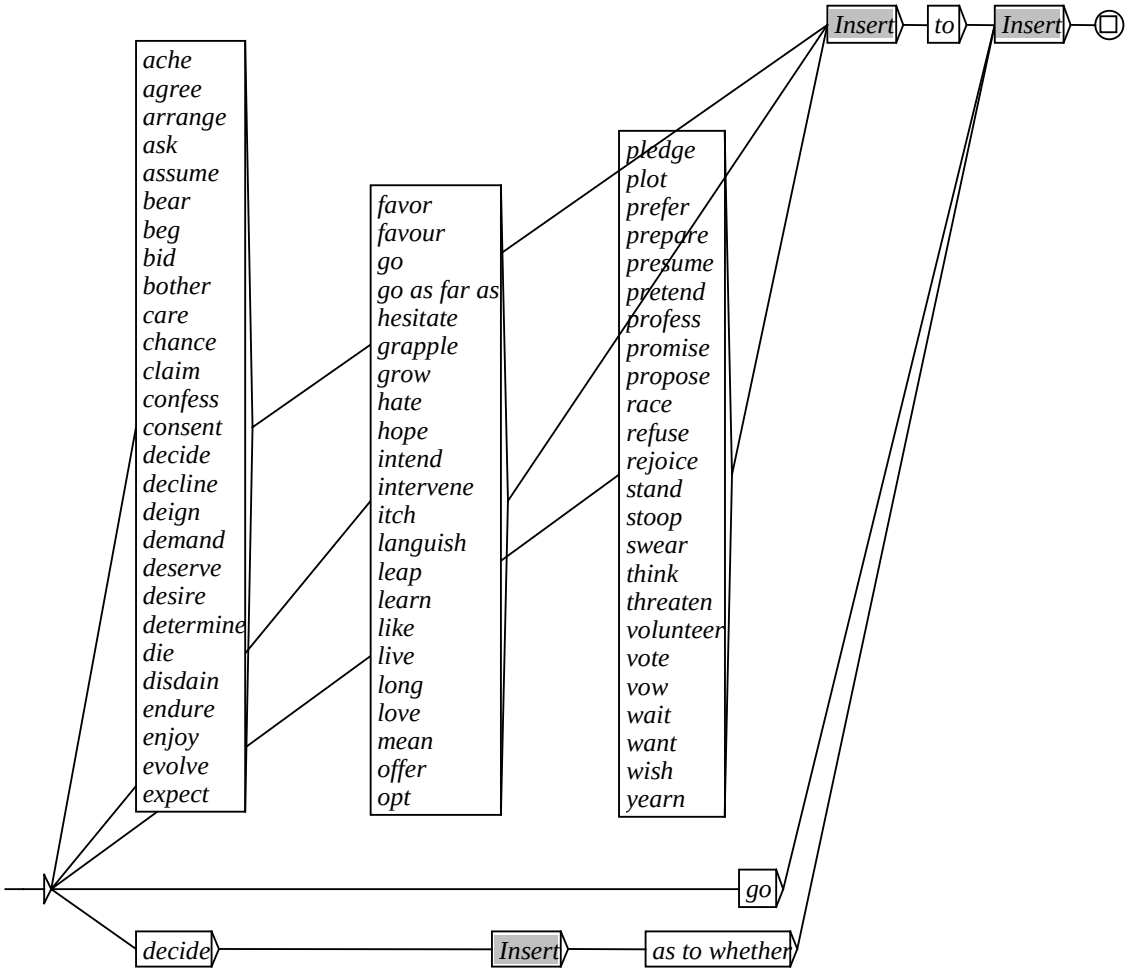


Figure 4. **VSPrepVing** and **VSToV**

The verbs *have to* and *need to* have been treated both as modality verbs *VMod* and as strict auxiliaries when they carry the negation *not* without *do*, these last forms are represented in the graph called **TenseNot (**figure 6). The auxiliary use of *to get to* has also be placed in the class *VMod*. Applying these criteria leads to the following list and to the local grammar called **Tense (**figure 5):[3]

- past tense auxiliaries: *have* and *have had* govern past participles; in the progressive forms, *be* and *have been* govern present participles;
- modal verbs: *can, could, may, might, will, would, shall, should, ought to, used to, be to*. These verbs have restricted conjugations :

[3] The names of the local grammars involved in the lemmatization process will appear in bold characters (*cf.* § 4).

- some have no infinitive form (*can, will, ought to, used to*),
- some cannot be conjugated or are highly defective:

Bob (is + was) to sell his car
* *Bob will be to sell his car*

- *to get* followed by gerund could be limited to a small set of verbs: *get (going + moving)*, in which case, it would be better described as entering idiomatic forms. Independently, *to get* is a variant of *to be*, when followed by adjectives and participles and similar to *to have* in sentences such as *Bob has (E + had) to sell his car*.

Simple tenses apply more or less regularly to auxiliary verbs. Tensed auxiliary verbs are all listed in the graphs, including forms contracted with subject pronouns (figures 5 and 6). Some contractions are ambiguous, for example *I'd = I had* or = *I would*, but some contexts disambiguate them. Contractions with subject noun phrases are observed: *My cousin's left*, *The best part's left*, they are locally ambiguous with possessive case; eliminating the ambiguity requires a deeper analysis of sentences. Roughly the same forms are described with negations in the separate graph **TenseNot**.[4]

2.2. ***Passive auxiliaries***. The graphs **Tense** and **TenseNot** are for active sentences. Passive sentences, which contain the auxiliary *to be* combined with past participles of transitive verbs are treated separately, they are described in several graphs stemming from the initial graph **BeTVen (**figure 7). The auxiliary *to be* has extensions: *become, get, grow, remain, stay,* some of these auxiliaries have aspectual meanings, accordingly, they are not accepted by all verbs, thus, many additional constraints will have to be introduced among verb combinations. Notice that lemmatization is not a goal in itself, it is a first step of the general operation of sentence parsing. Hence, lemmatization of a passive construction must be followed by an operation which links the passive fomr to its active form, such an operation is treated in a component of the grammar different from the one we present here.

[4] In this graph, we must separate inserts that may contain a negation (**Insert**), inserts that must contain a negation (**InsNot**) and those that may not contain a negation (**Ins**).

Figure 5. Tense

Figure 6. TenseNot

Sentences that have an auxiliary of the form *be Adjective*, *be N* or *be Prep N* are not fundamentally different from passive sentences, they can be treated together, they are called into the local grammar **BeTVen** by the subgraph **BeComp(**figure 8). Verbs like *to appear, to look, to seem* have forms similar to the variants of *be,* although they will be analyzed as reductions of sentences containing *to be* (*She seemed satisfied = She seemed to be satisfied*). Examples of these more general verbs have been included in various graphs.

The graph **AdjBeToV (**figure 9), a subgraph of **BeComp**, contains adjectival constructions of the type:

(1) *Bob is easy to please*

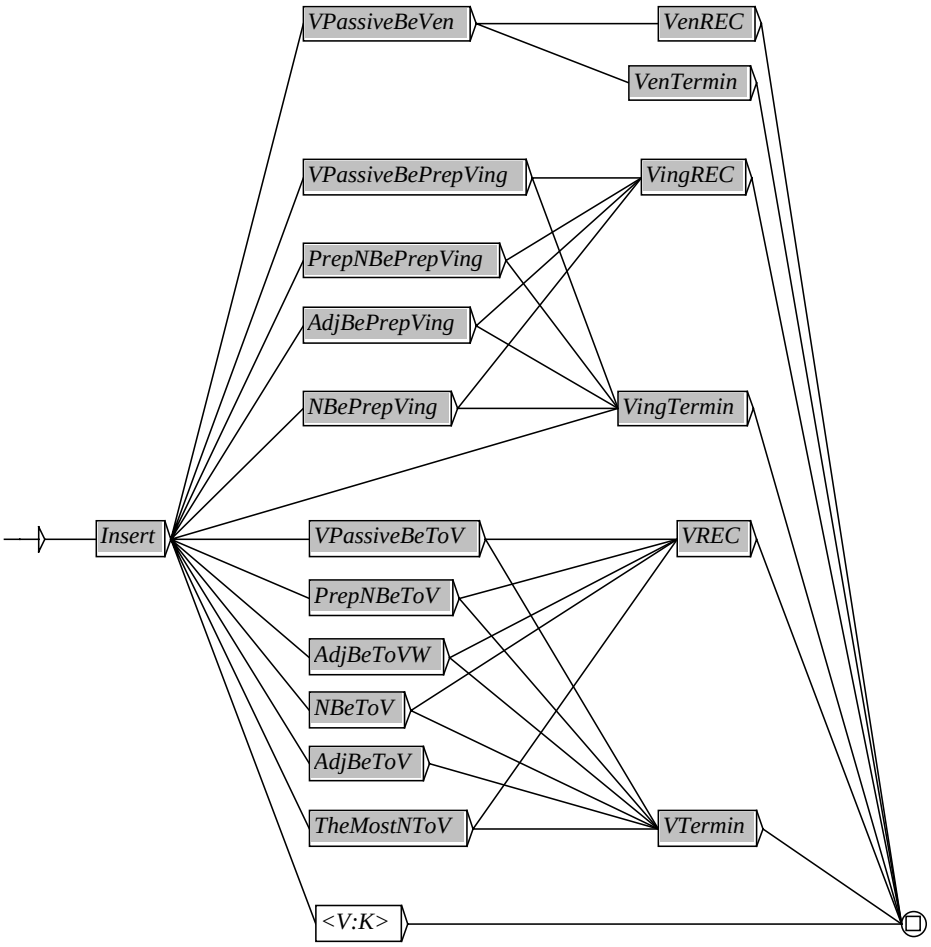
whereas the graph **AdjBeToVW (**figure 10), also subgraph of **BeComp**, corresponds to constructions such as:

(2) *Bob is eager to please (E + Jo)*

The names of the two graphs differ by the presence or not of *W*, the symbol for a complement sequence, they reflect the syntactic structures. Example (1) is close to Passive forms in the sense that the subject of *is easy* is the object of *please*. A plausible analysis of (1) is by reduction of the form *Bob is easy to be pleased* obtained by insertion of the modal sequence *is easy* in the Passive form *Bob is pleased*, hence, we feel justified to submit these constructions to the lemmatization procedure. In (2), the complement sequence *W* is unaffected.

Figure 10. AdjBeToVW

Figure 7. BeTVen



aBeComp.grf

Figure 8. BeComp
Figure 9. AdjBeToV

The lemmatized form of a Passive construction will not be a simple verb *V*, but a form *be Ven* (past participle). We chose this convention for two reasons:

- we call subject of a verb (or of a sentence) the noun phrase N_0 that agrees in number with the verb. In the case of Passive forms, the finite verb is the auxiliary *to be*, it agrees with a noun phrase which is an object of the verb to be lemmatized. Without a full analysis of sentences, which requires recognition of noun phrases, it is not be possible to locate the deep subject, hence, we deal here only with superficial subjects,
- the lemmatized form *to be* is similar to forms *to be Adjective* and *to be Prep* N_1, which, in general, are not to be lemmatized further.

Interrogative and imperative forms make use of specific auxiliaries: *do* and *let*, as in:

(Does + Did) Bob sleep well?
Let Bob read the book!

along with forms without auxiliaries :

Is Bob sleeping? Has Bob already left?
Read the book!

Interrogative forms involve an inversion of the subject. In imperative forms, let is followed by the subject of the main verb. The subjects, here *Bob*, can be noun phrase of any complexity, but no grammar of these constructions is available. However, a list of pronouns is easy to establish and can then be used in local grammars limited to the parsing of auxiliary sequences with subject pronouns for questions and object pronouns for imperative sentences. Since such pronouns are frequent in many texts, these restricted grammars might be useful in some applications.

We have built local grammars for interrogative forms with auxiliary *do*, but we limited subjects to a few pronominal forms (graphs **PronSubject,** figure 13), such as *they* in:

Did (they + the people who arrived before) read the book?

and *them* in imperative forms with auxiliary *let* :

Let (them + the people who arrived before) read the book!

We also parse forms with permuted subject pronouns with auxiliaries *be, have, will,* etc. in interrogative forms and in some other contexts. We did not include old forms such as *thou (dost + hast + goest + shalt + shouldest),* but they could be easily included in the grammar, if needed. Questions are described in figures 11, 12 and 13, imperative forms appear in figures 5 and 6.

Figure 11. Question
Figure 12. QuestionNot
Figure 13. PronSubject

2.3. ***Aspectual verbs*** have been distinguished, because they modify the time complements of the main verb. In the sentences that follow, acceptabilities are reversed, when an aspectual verb such as *to begin* is introduced:

Bob read both books in six hours
? *Bob read both books at 6 p.m.*

? *Bob began to read both books in six hours*
Bob began to read both books at 6 p.m.

The compatibility between the verb and the two adverbials of **date** and **duration** can be described in a natural way if one considers *to begin to read* as a verbal unit distinct from *to read*. Independently, we can argue that aspectual verbs have no autonomous constructions. Consider the sentences:

(1) *Bob began a reading of both books*
(2) *Bob began both books*

and Harris' nominalization relation of the base sentence:

(3)a *Bob read both books* = (3)b *Bob made a reading of both books*

The corresponding sentences with aspectual auxiliary both have the form N_0 V

N_1 (*i.e.* subject-verb-object), they are:

(3)Aa *Bob began to read both books*
= (3)Ab *Bob began (E + to make) a reading of both books*

(1) is then a reduction of (3)Ab. Now, (2) also seems to be an autonomous sentence, however, it is interpreted with an underlying verb, such as in :

Bob began to (read + write +sell + etc.*) both books*

that is, (2) is a further contraction of (1) or of the equivalent verbal sentence.

2.4. ***Modality verbs***

Z.S. Harris 1964 has syntactically defined a set of verbs or operators noted *U* that cover our set of aspectual and modality verbs. In principle, they do not have autonomous constructions. Sentences containing *U* verbs are of the form N_0 U^0 *(Prep)* V^0 *W* and <u>cannot</u> be derived from more complex sentences, as it is the case for *to want* for example:

Bob wants him to sell his car
= *Bob wants to sell his car* when *Bob = him*

Although the last sentence has the form of a sentence with auxiliary verb *U, to want* is not a *U* but a more complex operator (cf. *VS,* § 2.5). As already said, modality verbs, noted *Vmod* are sometimes difficult to separate from the verbs *VS* with full sentential complements (compare the lists of figures 3 and 4).[5] Examples of *U*s are:

Bob (attempted + failed +tried) to sell his car

Let us recall that we call auxiliaries the whole set of *U*s, including complex verbal phrases such as *be able to, be in the position of, have the ability to.*

The two verbs *come* and *go* behave somewhat like auxiliaries in sentences such as:

He came and talked to her, She goes and hit him

[5] Formal criteria of classification of these verbs are found in P.S. Rosenbaum 1967 and M. Gross 1968. A resulting classification of about 2500 verbs has been published in M. Gross 1975.

we included these constructions in our grammars (figure 14).

Figure 14. ComeAndV

2.5. ***Verbs with full sentential complements***
The more complex operators have sentential complements not directly relevant to our study:

Bob loved it that Jo solved the problem
Bob want us to leave
Bob insisted on Jo's leaving

But some of these constructions can be transformed and then become similar to aspectual or modality verbs, such as:

Bob wants to leave
Bob insisted on leaving

Notice that *to want* becomes auxiliary-like, whereas *to recommend* is not in *Bob recommended to leave*, since *Bob* is not the subject of *to leave*.

Let us sum up the various cases of auxiliary verbs that have been considered: Aspectual and modality verbs and verbs with sentential complements govern verbs in the infinitive or in the gerund form. The different formal types are:

- simple verbs: *to help, to need* and verbs with a preposition: *insist on Ving W, tend to V W*:[6]

Bob (accepted + needed + tried + wanted) to leave
Bob (insisted on + thought of) leaving

- adjectives constructed with *be*: *be able to V W, be aware of Ving W,*
- prepositional phrases constructed with *be*:

be on the verge of V^0*-ing W, (Bob is on the verge of leaving)*
be on $Poss^0$ *way to* V^0 *W, (Jo is on her way to give a lecture)*

[6] The symbol *W* corresponds to a variable ranging over the sequence of complements of the sentence.

Poss^0 is a possessive adjective that must refer to N_0, the subject;
- noun phrases whose support verbs is *to have* **(**figures 15 and 16):

Bob has the ability to remove scars
Bob has a procedure for removing scars

identical forms are nominalizations of verbs:

Bob has an urgent need to leave

Figure 15. NHaveToV
Figure 16. NHavePrepVing

Other nominalizations with other support verbs should also be included in the grammar;

- passive forms. Numerous verbs with sentential complements, when passivized, have forms that behave exactly like adjectival auxiliaries:

N_0 *V* N_1 *to* V^1*-ing W =*: *The impact sent the car spinning*
N_0 *V* N_1 *to* V^1 *W =*: *Bob ordered Jo to leave early*

N_1 *be Ven to* V^1 *W =*: *Joe was ordered to leave early*
N_1 *be Ven* V^1*-ing W =*: *The car was sent spinning*

Hence, we consider that they have to be lemmatized in the same way as the proper auxiliaries. We have listed a certain number of these forms in the figures 17 and 18.

Figure 17. VPassiveBePrepVing
Figure 18. VPassiveBeToV

- idiomatic forms of auxiliaries can be combinations of simple verbs with particles or can be more complex, involving nouns. They may belong to either of the categories we defined. We have included other examples in various graphs, other examples are found in figure 19.

N_0 *go on* V^0*-ing* W
N_0 *turn out to* V^0 W
N_0 *get the green light to* V^0 W
Jo broached the idea of selling her car

Figure 19. VingSCompPrepVing

There also exists idiomatic forms which do not enter the auxiliary pattern:

N_0 V N_1 *(the green light)*$_2$ *to* V^1 W =:

but whose passive are relevant:

N_0 *be Ven the green light to* V^0 W =:
Bob was given the green light to sell his car

such a form is similar to:

N_0 *be Ven to* V^0 W =: *I was able to sell his car*

where the form *be Adjective* is a modality similar to *can* in:

N_0 *Vaux*^{0} V^0 W =: *I could sell his car*

Our reason to include verbs with full sentential complements in the lemmatization process is to show how such lexical items can be parsed and interpreted within the local grammars of auxiliaries.

3. Combinations of verbs

Except for the first level of tense auxiliaries (left-most part) found in **Tense** (cf. figure 5), **TenseNot** (cf. figure 6) and **BeVen** (cf. figure 7), *Vaux*s combine in many different ways, with no clear pattern of restrictions emerging from the observations. For example, aspectual (*VAsp*) and modality verbs (*VMod*) can combine in different orders:

Bob began insisting on leaving early
Bob insists on beginning to work early

In principle, the content of a sentential complement (e.g. *that S*) is largely independent from the governing verb, hence, the verb found after a *VS* is not restricted. There are indeed restrictions, but no general rule seems to exist, which means that restrictions have to be handled verb by verb (M. Gross 1975). As a first approximation, we let the three categories *VAsp, VMod* and *VS* combine freely and recursively, that is, without restricting the length of the lemmatized sequences. The loops that represent these combinations are found in the graphs **VREC**, **VenREC**, **VingREC**, and **VTREC**. Since recursions are determined by government, there are four basic types of loops:

- a verb in the infinitive form that governs a verb in the infinitive form can loop on itself,
- a verb in the gerund form that governs a verb in the gerund form can loop on itself,
- a verb in the infinitive form may govern a verb in the gerund form, in turn this gerund verb may govern a verb in the infinitive form, this sequence of two verbs can loop on itself,
- a verb in the gerund form may govern a verb in the infinitive form, in turn this infinitive verb may govern a verb in the gerund form, this sequence of two verbs can loop on itself.

In figure 20, we give the example of **VREC**.

Figure 20. VREC

Families of graphs have been defined accordingly:

- for example, the graphs **VTXxxVToV**, begin with any simple form of verb (i.e. *VT*) that governs infinitive forms (i.e. *to V* forms), **Xxx** stands for one of the types: **Asp**, **Mod** or **S**,
- the graphs **VTXxxVPrepVing**, begin with any simple form of verb that governs gerund forms through a specific preposition,
- past participles are noted in the same manner, (i.e. the graphs **VenXxxVPrepVing** correspond to sequences beginning with a past participle and they governing the gerund), for gerund and infinitive forms, we use the same notation (i.e. **VingXxxVPrepVing**, **VXxxVPrepVing**).

Since we have three classes and four patterns of tenses, we had to construct the twelve corresponding graphs for the various verbs. In figure 21, we give the example of **VingAspVPrepVing**.

Figure 21. VingAspVPrepVing

Recursions must end in terminal forms. Simple verb forms <V:K>, <V:W>, <V:G>, are terminals, we added *be*-Adjective and *have-Noun* constructions ending in simple verbs to the terminal graphs **Vtermin**, **VenTermin**, **VingTermin**.

The recursive combinations are described by means of the classes of verbs, but in order to make abstraction of the lexical differences (i.e. classes of simple verbs, *be*-Adjective, *have-Noun* constructions, frozen expressions, etc.), we use higher level graphs that retain only the government constraints:

- **VenVPrepVing** (figure 22) corresponds to utterances that begin with a past participle and that govern gerund verbs,
- **VingVToV** corresponds to utterances that begin with a verb in the gerund and that govern infinitive verbs,
- since there are five patterns of tenses and two governing types, we built ten other graphs on the same basis and noted on the same principles.

Figure 22. VenVPrepVing

4. Ambiguities

4.1. Systematic ambiguities

1) A sequence of words such as:

plan to leave at night

is recognized with the modality verb *to plan*, auxiliary of *to leave*, as in *I plan to leave*. However, *plan* is also a noun that governs the same infinitive verb phrases, as in:

He presented (a plan)$_{N}$ *to eliminate pesticides*

Another example is:

The states use the Civil Justice Reform (Act to require)$_{V}$

where parentheses indicate the 'wrong' analysis. This ambiguity can only be

resolved by exploring the context of the ambiguous sequence. To this aim, we have introduced examples of local grammars which describe noun phrases whose head nouns govern verbs in the same way as verbs do. In the case of verb/noun homographs, taking into account left contexts resolves some ambiguities, for example, when a determiner such as a or *the* is present. We outlined several graphs for syntactic categories of nouns constructed with the support verbs *to be, there be* and *to have* (figures 15, 16 and 23 to 28). If an ambiguous sequence such as *need to leave* occurs as part of the utterance: *had an urgent need to leave*, it will be analyzed as a noun phrase whose recognition involves the graph **NhaveToV** (figure 15).

Figure 23. NBeToV
Figure 24. NBeToVW
Figure 25. NBePrepVing
Figure 26. ThereBeToV
Figure 27. ThereBePrepVing
Figure 28. TheMostNToV

In the same way, we have extended the description of left contexts of nouns phrases to prepositions, that is:

- to prepositional noun phrases with complex prepositions, such as *in an attempt to increase its influence* (figure 30) and *in view of increasing its influence* (figure 31). Some ambiguities are also resolved by these two local grammars,
- to noun phrases with simple prepositions, but in this case, we have restricted the verbal forms to complex ones, that is with at least two combined verbs. This restriction is achieved by using the graphs **VREC** and **VingREC** inside the subgraphs **PrepV** (figure 32) and **PrepVing** (figure 33). These four graphs are grouped into the graph of higher level **PrepVW** (figure 29), which is placed in **VAUX** (figure 34), the complete grammar of verbs preceded by auxiliary sequences.

Figure 29. PrepVW
Figure 30. InOrderToV
Figure 31. InViewOfVing
Figure 32. PrepV
Figure 33. PrepVing
Figure 34. VAUX

We could also have represented left contexts of verbs, which, in general, are subject noun phrases, often subject pronouns. Adjoining pronouns to the graphs of conjugated auxiliaries is a natural extension, since we already introduced many of them that are contracted with the verb (e.g. *I'm, we're*). But in this presentation, our goal is restricted, and we only suggest several generalizations which can be implemented at a further stage.[7]

2) Some of the aspectual and modality verbs have nominal complements parallel to their verbal ones, that is, they enter both structures:

N_0 V Prep V^0 W (=: Bob began to read the text)
N_0 V Prep N_1 (=: Bob began the reading of the text) (*Prep* is zero here)

This situation may generate analyses such as: *(help)*$_{\text{Vaux}}$*(low)*$_{\text{V}}$, where the noun or adjective *low* is tagged *Verb* in the sentence:[8]

They help low income families

The following examples are of a similar type:

These (are daunting)$_{\text{V}}$ *goals* (model: *They are flying planes*)
(helping) $_{\text{Vaux}}$ *(even)*$_{\text{V}}$ *young children*
(is encouraging) news

3) Some ambiguities may result from a particular choice of grammatical codes in the dictionary. For example, verb particles, such as *about, up, down, around, off,* are coded *Adverb*. As a consequence, our local grammars will analyze a sequence such as:

All this is about giving Bob a chance

with *about Adverb*, hence *is giving* with be parsed as the progressive form of *to give*. By using the code *Prt* for *Particle, be about* becomes a verbal unit which does not belong to the set of auxiliary verbs, the difficulty is thus avoided.

[7] For a more systematic approach to the resolution of ambiguities by means of local grammars, see E. Laporte 1995.
[8] The 'tag' is a solution provided by the dictionary.

4.2. Accidental ambiguities

1) *Homographs*. In the sentences:

Why do people behave so?
(When + Where) do (right + wrong) people behave so?
They do go to school

the nouns or adjectives *people*, *right*, *wrong*, *school* are also verbs; when preceded by *do*, the program recognizes verbal complexes. The same ambiguity occurs with inversions of the subject other than questions, as in:

Only in good times do people want to read

Some types of homography involve words which are, on the one hand, participles or adjectives and on the other, verbal forms:

(are dead) ends
(Head Start) funding (*Head Start* is a proper name).
It is (cost)$_{\mathrm{V}}$ *effective* (model: *is gone*)
a lot of (work) $_{\mathrm{Vaux}}$ *(left)*$_{\mathrm{ADV}}$ *to (do)*$_{\mathrm{V}}$
doesn't (clean)$_{\mathrm{ADV}}$ *(up)*$_{\mathrm{V}}$ *his room*

2) *My friends (at first)*$_{\mathrm{ADV}}$ *(glance)*$_{\mathrm{V}}$ *seemed lost*
My friends (at first glance)$_{\mathrm{ADV}}$ *(seemed)*$_{\mathrm{V}}$ *lost*

In this example, the longest match principle resolves the ambiguity (i.e. the adverb *at first* is a prefix of the longer adverb *at first glance).*

3) The utterances *go to bed* and *come to power* can be analyzed either as phrasal verbs where *bed* and *power* are nouns or with the auxiliaries *go to* and *come to* preceding the verbs *to bed* and *to power.*

4) The auxiliary *be to* in the sentence *Our son is to arrive at noon* is ambiguous, in the sense that its sentence structure can be confused with sentences of the type:

(1) *Our goal is to arrive at noon*

Resolving this ambiguity requires a detailed analysis of the sentence, and in

particular, a list of the nouns such as *goal* which allow construction (1). We can make human nouns a special case, since they cannot be found in the subject position of (1). We could introduce human subject pronouns for the auxiliary verb *to be to.* Such pronouns are easy to recognize and are numerous in various texts. Then, when a human pronoun is found, the ambiguity (i.e. the wrong analysis) would disappear. The general case requires a grammar of full human noun phrases, which is not available.

5. The local grammars

Throughout this presentation,[9] we have mentioned the use of graphs or local grammars to describe constrained sequences of words and/or categories. These local grammars have the form of finite-state grammars or transducers, given in the form of graphs, they contain words and grammatical categories (M. Gross 1998). We make more precise the notations of these devices.

5.1. Notations

<*V*> corresponds to the simple inflected forms of any verb, hence, <**eat**> corresponds to the forms *eat*, *eats*, *ate*, *eaten*, *eating*. <*V:K*> is 'verb in the preterit tense', <*V:W*> is 'verb in the infinitive', <*V:G*> is 'verb in the gerund'.[10] <*A*> is adjective, <*ADV*> is adverb, <*N*> is noun, <*N:s*> is a singular noun, <*N:p*> is a plural noun.

5.2 Graphs

The design of our finite-state graphs is not standard, it had to be adapted to the description of linguistic phenomena, where many parallel edges are observed to be equivalent. Such sets of edges are represented by boxes: each line in a box corresponds to one edge of a standard graph. Words and categories appear in boxes. Names of graphs appear in shaded boxes as subgraphs of other graphs. Chains of graphs embedded into each other can be relatively long, their depth can easily be 10 before reaching a final state, moreover, they are allowed to form loops, that is, to describe recursive structures of verbs governing verbs of the same type.

We list the various graphs and recall their function hereafter :

Tense is a basic grammar, in the sense it describes the finite tenses of the

[9] J. Senellart 199 has written similar grammars for French (M. Gross and J. Senellart 1998).

[10] At some future stage of development of parsers, this notation should cover the compound tenses.

auxiliaries of § 2.1. Paths start with the tensed forms and lead to verbs of the three categories of governing verbs: *VAsp*, *VMod*, *VS* and to the other verbs *V*;
TenseNot differs from **Tense** by the presence of the negation *not* (and accessorily *never*);
BeTVen describes the finite tenses of the Passive auxiliary *to be*.

The tensed auxiliaries of **Tense**, **TenseNot** and **BeTVen** precede three categories of verbal complexes: **VedREC**, **VenREC**, **VingREC**, **VREC**:

- VenREC contains structures that all begin with a past participle,
- VingREC contains structures that all begin with a present participle,
- VREC contains structures that all begin with an infinitive form simple: <V:W> or compound: *have* <*V:K*> or *be* <*V:K*>,
- VTREC appears in the graph **VAUX**, at the same level as **Tense**, it contains structures that all begin with a governing verb in a simple form, namely, <*VAsp*>, <*VMod*>, <*VS*>.

These five graphs are the only recursive ones. Recursivity results from the phenomenon of government: an auxiliary verb (in the semantic sense, cf. § 2) can govern a verb that governs another, etc. We considered two types of government:

- verbs govern verbs in the gerund form (i.e. *Ving* or more formally <*V:G*>),
- verbs govern verbs in the infinitive form (i.e. *V* or more formally <*V:W*>).

Let us recapitulate the components of the local grammar used to lemmatize compound tenses. The graph **VAUX**, is the highest level graph. It sums up the various structures involved:

(i) structures considered as compound verbs, they include:
- tensed verbs in graphs **Tense**, **TenseNot**, **BeTVen**, **VTREC**,
- passive forms without the auxiliary *to be*: **VPassiveToV**, **VPassivePrepVing**, **VpassiveVen**;

(ii) questions, with negation or not: graphs **Question** and **QuestionNot**;

(iii) structures whose left context contribute to disambiguation:
- prepositional phrases, which involve a left context that contributes to disambiguate verbs (4 graphs contained in **PrepVW**),
- noun phrases, whose support verbs are *be* and *have*: **NBeToV**,

NBePrepVing,, **NHaveToV**, **NHavePrepVing**,
- a special form of noun phrase: **TheMostNToV,**
- examples of sentences with the verb *to be*: **NBeToVW** and *There be*: **ThereBeToV**, **ThereBePrepVing.**

The number of graphs constructed in view of lemmatizing compound verbs is 66, to which must be added 17 graphs of noun phrases that disambiguate some of the verbal forms.

Experience has shown that even without any semantic restrictions on combinations of verbs, we found practically no error of analysis. We checked our grammars on Henry James' novel *The Portrait of a Lady* and we found about 13,000 examples of compound verbs, and on a recent sample of the *International Herald Tribune* (1997) in which we recognized 4,000 examples, we found very few errors (in the order of magnitude of 1 or 2 per 1000), some omissions did occur, but they were due to the state of incompleteness of the lists of generalized auxiliaries, which is something to be expected. In figure 35, we give a sample of text with auxiliaries sequences as they are marked by the INTEX parser using the grammar VAUX and in figure 36 we have a sample of a concordance of these sequences.

Tense auxiliaries only modify verbs, whereas aspectual and modality verbs may have nominal complement equivalent to verbal ones *(begin the reading* vs *begin to read)*, hence, aspectual and modality verbs may syntactically behave like ordinary verbs and have to be lemmatized accordingly. Ultimately, if we consider their nominal complements as belonging to a nominal voice of the verb, verbal nouns will have to be lemmatized as well.

The United States plans to stick to a go-slow approach on expanding NATO membership to East European countries despite sharp divisions among American officials and protests from East European leaders, administration officials say.

Little more than a week before a NATO summit meeting in Brussels, all sides acknowledge that the decision whether to let Poland, the Czech Republic, Hungary and other nations into the North Atlantic Treaty Organization is one of the most important in the history of the alliance.

The issue has become even more pressing for East European governments in recent weeks as their fears of resurgent Russian nationalism have been heightened by Vladimir V. Zhirinovsky's aggressive statements.

The prevailing view in Washington, reaffirmed since the Russian vote, is that the West should move slowly to avoid aggravating Moscow's traditional fears of encirclement and strengthening Russians opposed to reform.

On the other side are administration officials who believe that democratic gains in Eastern Europe must be consolidated and that those countries must be protected from past predators by inclusion in the alliance.

Providing new details about the debate that embroiled the Clinton administration, officials said that Secretary of State Warren M. Christopher had initially favored expanding NATO's membership to the East.

But he was persuaded to reverse course after the intervention of Strobe Talbott, the journalist-turned-policymaker who was named Mr. Christopher's deputy last week.

On the weekend before a critical cabinet-level meeting in October, Mr. Talbott, who has been ambassador-at-large to the former Soviet republics, typed a memo on his home computer arguing against NATO expansion and sent it to Mr. Christopher.

Within days, Mr. Christopher and Defense Secretary Les Aspin were flying to Europe to explain the change of plan.

Under the go-slow approach, which will be formally presented at the summit meeting on Jan. 10 and 11, the United States and its allies will endorse the principle that NATO's membership should eventually be enlarged.

But NATO will not ease expansion of the alliance by outlining a clear set of standards for admitting new members. Instead, East European countries and former Soviet republics are being invited to take part in a program of military training and exercises that will allow them to associate themselves with the alliance without offering formal membership or the security guarantees that come with it.

The text has been automatically segmented into sentences. The parts underlined are those recobnized by the grammar we described.

Figure 35. Marked text

sits to the chapel, I periodically felt	a strong urge to walk over to one of the wooden benches
cellent example of how America can play	a vital role in moderating tensions between Ukraine and
ko never lost his love for shooting. As	a way of sparking conversation, he took potshots at hat
ork Times The technicians have found	a way to bring back the musical past. Memorable perform
Clinton will, therefore, probably find	a way to renew China's most-favored- nation trading sta
Higuita, 27, was detained on June 4 and	accused of receiving a $64,000 payoff for helping negot
rawn angry letters from former children	accustomed to feeding on demand. A proofreader was guts
nbeaten home record. Papin was sent off	after appearing to elbow an opponent in the 37th minute
"See you later, folks," he said Monday	after being asked once if he planned to resign and twic
the IRA lays down its arms. Britain	agreed not to stand in the way if the people of Norther
"if the dominant powers in each region	agreed to exercise authority in accordance with an inte
the future of the province, if the IRA	agreed to put down its arms after 25 years and abandon
en the United States and China, Beijing	agreed to remove three-quarters of its non-tariff trade
een as a secondary target for tourists,	aimed at attracting one-day visits. But it has prove
rs and abandon its campaign of violence	aimed at driving the British from Northern Ireland.
ced to hard labor for seven years. I	am also advocating the death penalty for those who garb
s into contact with a human voice. I	am also starting a campaign to stop people from using t
reams of stardom in the movie world. "I	am beginning to think about my future beyond the world
. Continuing with the crime issue, I	am calling for the prohibition of hardened criminals ap
alked with the Knicks. "And as far as I	am concerned," Wilson added, "Isiah Thomas is welcome h
to the disk, at least not this week. I	am correcting this error lest my works be immortalized
ek winter break from the circuit. "I	am definitely giving myself the whole year," said the 1
ched the microphone as he confessed, "I	am frightened by the way I feel, maybe you are too / I
Bettis then talked about improving. "I	am going to try to get a little quicker laterally, and
g their personal finances better. "I	am hearing a lot of people give themselves permission n
y the way I feel, maybe you are too / I	am losing faith in everything and everyone but you." By
(even if it does bounce). I thought: I	am not going to play silly defensive cricket shots. I a
rison without parole for 10 years. I	am not talking about personal voice-mail boxes, but tho
to pay tomorrow." She added, "What I	am really dreaming of is a small inn, a cozy place with
ng with another electronic system. I	am recommending a five-day waiting period for the autho
e:International Herald Tribune I	am tired of hearing of the carnage that afflicts Americ
ut her mother, Haller Jones, did. "I	am tired of those dang reporters," Mrs. Jones said, sta
"I do not really complain because I	am used to it," he said as he stamped his feet in the c
and was, after all, a grown man. "I	am very saddened that this has happened, that he was ev
ame, is nursing a sore left foot. "I	am worried," said North Carolina coach Dean Smith. "The
ublished in Polish, contain accounts of	an abortive attempt to poison the Czar, on the occasion
d. Mr. Reynolds's comments were seen as	an attempt to offer further inducement to the IRA to en
eptance through their commodification -	another way of describing their Americanization. Jul
be "settled fairly," but he warned that	any effort to press his country to make broader concess
incorporated in Malaysia but ruled out	any plans of selling its equity to local shareholders.
nfrontation with Moscow would eliminate	any Russian temptation to fill the vacuum. Mr. Clint
d the Palestine Liberation Organization	appear to be frozen amid growing rancor and recriminati
s to be accelerating and interest rates	appear to favor the dollar. The Federal Reserve Board i
sketchy, but most guerrilla fatalities	appear to have been inflicted while the rebels were ret
ai army officials said. The fighting	appeared to be increasing in intensity, and both sides
y raids, fled for the hills Tuesday and	appeared to be pulling out of their remaining stronghol
ning set. Although he won the set, Cash	appeared to be uncomfortable during the remainder of th
acy. But even though the transaction	appears far from going through, analysts saw in Federat
e Israeli-Palestinian dispute, however,	appears now to be overtaken by the anger toward Mr. Ara
closest neighbors. U.S. involvement	appears restricted to trying to buy back the Stinger gr
s from Egypt to Tajikistan, where Islam	appears to be a popular alternative but governments are

Concordances present the sequences parsed by the grammar **VAUX** in alphabetic order. This sample comprises compound verbs as well as noun phrases whose head noun has been disambiguated (i.e. cannot be a verb, *urge* and *plan* for example).

Figure 36. Concordance

References

Fairon, Cédrick 1999. To appear, *Lingvisticae Investigationes*, Amsterdam-Philadelphia: John Benjamins.

Gross, Maurice 1968. *Grammaire transformationnelle du français. 1- Syntaxe du verbe*, Paris: Larousse, 186 p.

Gross, Maurice 1975. *Méthodes en syntaxe*, Paris: Hermann, 412 p.

Gross, Maurice 1990. *Grammaire transformationnelle du français. 3-Syntaxe de l'adverbe*, Paris: ASSTRIL, 670 p.

Gross, Maurice 1997. The Construction of Local Grammars, Roche, Emmanuel & Yves Schabes, eds. 1997. *Finite State Language Processing*, Cambridge, Mass.: The MIT Press, pp. 329-352.

Gross, Maurice & Jean Senellart 1998. Nouvelles bases statistiques pour les mots du français, *Proceedings des JADT*, Université de Nice.

Harris, Zellig S. 1964. The Elementary Transformations, Philadelphia: University of Pennsylvania, TDAP No 54. Reprinted in *Papers in Structural and Transformational Linguistics*, 1970, Dordrecht: Reidel, pp. 482-532.

Laporte, Eric 1995. Levée d'ambiguïtés par grammaires locales, in J. Labelle & C. Leclère eds: *Lexiques-grammaires comparés en français*, LIS 17, Amsterdam-Philadelphia: John Benjamins, pp. 97-114.

Rosenbaum, Peter S. 1967. *The Grammar of English Predicate Constructions*, Cambridge, Mass.: The MIT Press.

Senellart, Jean 1999. Reconnaissance automatique des entrées du lexique-grammaire des phrases figées, *Travaux de linguistique*, Bruxelles: Duculot, pp. 109-125.

Silberztein, Max, 1993. *Dictionnaires électroniques et analyse automatique de textes*. Paris: Masson, 233 p.

SUMMARY

We generalize the process of lemmatization of verbs to their compound tenses.Usually, lemmatization is limited on verbs conjugated by means of suffixes ; tense auxiliaries and modal verbs (e.g. *I have left, I am leaving, I could leave*) are ignored. We have constructed a set of 83 finite-state grammars which parse auxiliary verbs and thus recognizes the 'head verb', that is, the lemma.

We generalize the notion of auxiliary verb to verbs with sentential complements which have transformed constructions (e.g. *I want to go*) that can be parsed in exactly the same way as tense auxiliaries or modal verbs.

Ambiguities arise, in particular because adverbial inserts occurr inside the compound verbs,. We show how local grammars describing nominal contexts can be used to reduce the degree of ambiguity.

Address of the author :

LADL
Université Paris 7
2, place Jussieu
F-75251 Paris CEDEX 05
mgross@ladl.jussieu.fr