The essential feature of a lexicon-grammar is that the elementary unit of computation and storage is the simple sentence subject-verb-complement(s). This type of representation is obviously needed for verbs: limiting a verb to its shape has no meaning other than typographic, since a verb cannot be separated from its subject and essential complement(s). We have shown (M. Gross 1976) that given a verb, or equivalently a simple sentence, the set of syntactic properties that describes its variations is unique; in general, no other verb has an identical syntactic paradigm. As a consequence, the properties of each verbal construction must be represented in a lexicon-grammar. The lexicon has no significance taken as an isolated component and the grammar composed, viewed as independent of the lexicon, will have to be limited to certain complex sentences.

Since be-Adjectives forms are close to verbs, their description is quite similar, that is, they are considered as sentences.

We have applied lexicon-grammar representation not only to the two obvious predicative parts of speech, verb and adjective, but to nouns and adverbs as well. In the same way as one adjoins the verb to be to adjectives, we have systematically introduced support verbs (Vsup) for nouns and adverbs, as in the following examples (Z.S. Harris 1976, M. Gross 1982, 1986):

- Vsup: to be Prep: The text is in contradiction with the law
- Vsup: to have This text has certain importance for Bob
- Vsup: to occur, etc. Accidents occur at random The accident (was, happened, occurred, took place) late at night

Support verbs are frequent in technical texts, and may have stylistic variants, as in this last example:

Grammatical elements such as determiners, prepositions and conjunctions, do not belong to the lexicon-grammar in the same sense as the four major parts of speech do, since they are parts of structures or rules. For example, prepositions appear in the columns of the lexicon-grammar.

An early representation of verbs in a lexicon-grammar of about 12,000 verbs is given in Figure 1. Each row of the matrix is an entry whose main construction is defined by a type or class code. In Figure 1, the code Vg corresponds to the class of constructions subject-verb-direct sentential complement, noted:

1. Vg V que P

Each column is a syntactic property, and corresponds to a structure into which V may enter, roughly a syntactic transform of the main structure. For example, in columns we have placed the Passive forms, Extraposed and nominal forms. Thus, the related structures are semantically close. "..." sign at the intersection of a row and a column indicates that the entry in that cell is accepted in the structure associated to the column, "..." sign corresponds to inacceptability. The process of accumulation that led to the formalized lexicon-grammar of 12,000 French verbs has run into what seemed to be a minor problem of representation of words: the difference between simple and compound words. On the one hand, there are simple words such as the verb keep and complex (idiomatic) forms such as keep in mind. Both forms play the same syntactic and semantic role in sentences such as:

Bob kept in mind that Max has moved to Tampa
Bob keeps in mind that Max has moved to Tampa

Indeed compound terms raise a problem of representation. The unit of representation in a linear lexicon is roughly the word as defined by its written form, that is, a sequence of letters separated from neighboring sequences by boundary blanks. As a consequence, compound words cannot be directly put into a dictionary the way simple words are. An identification procedure is needed for their occurrences in texts, and this procedure will make use of the various simple parts of the compound utterance. Hence, the formal linguistic properties of compound terms will determine both the procedure of identification in texts and the type of storage they require.
We thus have to discuss the main types of compounds and to single out those properties that bear on automatic parsing and dictionary lookup.

1. Compound adverbs

We call adverb any circumstantial complement, including sentential phrases, as in the following examples:

1) The show took place at night during a busy night the night Bob missed his plane

By compound adverbs, or frozen or idiomatic adverbs, we mean adverbs that can be separated into several words, with some or all of their words frozen, that is, semantically and/or syntactically noncompositional. In (1), at night is a compound adverb, the lack of compositionality is apparent from lexical restrictions such as:

*at day, *at afternoon, *at evening

and by the impossibility of inserting material that is a priori plausible, syntactically and semantically:

*at (coming, present) night
*at (cold, dark) night

during the (coming, present) night during a (cold, dark) night

5. Notice that nightly can also be considered as a frozen compound, though not constituted of words but of a word and a suffix. Again, lack of compositionality stems from the observation that daily, weekly, monthly, yearly, etc. which are compounds of the same formal type have a regular formation, in the sense that their interpretation is homogeneous. Thus, nightly is an isolated case, as opposed to an open series of identical forms with a different interpretation.

The two other adverbs of (1) are free forms. Thus, the determiners Det and modifiers (Adj and Rel clause) of:

during Det Adj night Rel clause

can vary freely (within semantic constraints). In the same way, the event associated with the sentence S in the form:

the night (E, that) S

can be expressed by a large variety of unconstrained forms.

Frozen or compound adverbs constitute the simplest case of compound forms because they do not allow variations of their components. As mentioned above, in at night no adjective is authorized. Moreover, one cannot insert a determiner: *at (a, this) night, the plural is forbidden: *at nights and no relative clause can be appended: *at night (that, which) was agreed on.

Such observations are general, and apply to many adverbs of varied form and lexical content:

It rained cats and dogs
*many cats and dogs
*big cats and dogs
*cat and dog

Table 6: Verbs with Sentential Complements
(From M. Gross 1975)
Consequently, these compound adverbs could be identified by a simple recognition procedure, as they do not require any lemmatization or syntactic analysis to be reduced to a dictionary form, as is the case with verb forms for example.

A lexical study of compound adverbs has been performed in French and a systematic inventory has been compiled from various dictionaries. Phrasal texts have been examined as well. It is interesting to note that whereas in current dictionaries there are about 1,500 one word adverbs, most of them in current dictionaries. There are about 1,500 one word adverbs, most of them in modern dictionaries. We have found over 5,000 compound adverbs.

These compound adverbs have been classified according to their syntactic shape. The syntactic forms are described at the elementary level of sequences of parts of speech. We use symbols with obvious interpretations such as Prep, Det, Adj, N, V, Conj (for conjunction) and W for a variable ranging over verb components, etc. We write:

\[
\begin{align*}
\text{Prep} & \quad \text{Det} & \quad \text{N} & \quad \text{Adj} & \quad \text{Conj} & \quad \text{W} & \quad \text{S} \\
\end{align*}
\]

Figure 2 shows the classes that have been defined on this basis, together with examples and the number of items in each class:

| Class | Structure | Examples | Items |
|-------|-----------|----------|-------|
| PADV  | Adv       | soudain  | 328   |
| PC    | Prép C    | en bref  | 460   |
| pdu   | Prép Det C| contre toute oriente | 570 |
| PAC   | Prép Adj C| de sa belle mort | 440 |
| PCA   | Prép C Adj | à gorge déployée | 400 |
| PCDC  | Prép C de C| en désespoir de cause | 350 |
| PCDN  | Prép C de N| au moyen de N | 330 |
| PCPN  | Prép C prép N| par rapport à | 30 |
| PCPC  | Prép C, prép C| avec pieds à la tête | 140 |
| PCNF  | Prép C Conj C| en tout et pour tout | 170 |
| PV    | Prép V W | à dire vrai | 150 |
| PF    | P (phrase figée) | Dire vrai le sain | 230 |
| PICO  | (Adj) comme C, | comme ses pieds | 200 |
| PVCO  | (V) comme C, | comme un cheveu sur la soupe | 210 |
| PVNC  | (V) comme Prép C, | comme dans le buché | 30 |
| PIC   | Conj C, | et tout le tremblement | 100 |

**TOTAL:** 840

Table 2

The examples discussed so far are entirely frozen. Hence, as a practical matter, they can be located in a text by using the search function available for strings in any text editor system. There are however more complex examples that require deeper analysis. Consider for example the idiomatic adverb in the sentence:

**Max proposed solutions from the top of his hat**

It is largely frozen: no other determiner is allowed, no adjectives can be appended to either noun, etc., but the person of the possessive adjective, Max, may vary. This possessive adjective must refer to the subject of the sentence, and varies accordingly:

*Max proposed ideas from the top of his hat*
*My older brother proposed ideas from the top of his hat*
*Bob and Max proposed ideas from the top of their hats*

In this case, the recognition procedure is no longer a simple string matching operation, since a variable slot must be dealt with inside the fixed string. More general matching rules are required here. Once this compound adverb has been identified in a text to be processed, it can be given an interpretation, for example in terms of a simple adverb such as leisurely or lightly and the referential information carried by Pass can then be ignored. However, one can easily construct particular discourses where the obligatory coreference relation involved will disambiguate some analysis. Thus, not only the variation of Pass must be accounted for at the lexical level, but its referential information has to be kept for possible use in a parser.

Other compound adverbs offer different degrees of variation. There are cases where one part of the adverb is frozen and another part is entirely free:

**Max organized a party in honor of Bob**
**Max hid the car at the far end of the parking lot**

The parts in honor, of the far end are frozen. For example, they do not allow modifiers. The parts of N are free, for we observe variations such as:

**Max organized a party in his honor**
**Max hid the car at the far end, I think, of the parking lot**

Consider the adverbs:

**for the sake of raising things**
**for the sake of Bob**
**for God's sake**

We call the combination for--sake frozen, since the noun sake does not occur elsewhere than in adverbial phrases with the preposition for: it cannot be the subject or object of any verb. On the other hand, the modifiers of sake are quite varied and regular from the point of view of the syntax of noun modifiers.

There are also cases of seemingly free adverbs which require an ad hoc treatment. For example, dates such as:

**Monday March 13, 1968 at 9 p.m.**

are described in a natural way by a finite automaton.

Technical or specialized families of adverbs come close to being frozen adverbs:

(2) They elected Bob on the (first, second) ballot
(3) Max ate his noodles in a bowl

The special semantic relations that hold between the adverbial complement and the rest of the sentence are limited. There are few verbs such as to eat which combine with in a bowl and which have the non locative interpretation of (3). The usual interpretation is that found in:

**Max ate his noodles**

6. PRELOG rules are particularly well adapted to recognizing such frozen forms (P. Sabatier 1988).

7. There are nonetheless restrictions on them:

*for a heavenly sake*
Max put his noodles in a bowl

Endowing frozen adverbs into a lexicon-grammar raises many new questions. The bulk of adverbs can be described by means of the following type of derivation (Z.S. Harris 1976):

\[
\begin{align*}
Bob & \text{ left; The Bob moved occurred at 9} \\
& = \text{Bob left, this occurred at 9} \\
& = \text{Bob left at 9}
\end{align*}
\]

and support verbs play a crucial role here. However, there are cases where no general support verb is found and where adverbs have to be considered as a part of the elementary sentence. Consider the adverb in:

\[
Bob \text{ sang at the top of his voice}
\]

It is syntactically and semantically analogous to free adverbs such as noisily, powerfully. For these two free adverbs, a derivational source involving the adjective is available:

The way Bob sang was (noisily, powerfully)

This is not the case for at the top of his voice which is practically limited to modifying the verbs of saying. Moreover, the obligatory coreference link of his leads to a representation where this adverb is not analyzed. Thus two semantically similar types of adverbs have to be represented quite differently in the lexicon-grammar.

All the situations just exemplified with adverbs are quite common, and are also encountered with nouns, adjectives and verbs. The paradox is: these cannot modify adjectives and nouns as in

\[
\text{Such nouns can become quite complex in various technical fields.}
\]

In general, compound nouns allow variations of determiners and modifiers, but many situations are encountered:

- the moon is a frozen combination, -- definite article-noun -- which behaves like a proper name, because of its unicity of reference. It cannot be modified by adjectives without losing its reference: the (big, yellow) moon;

- crude oil takes restricted determiners. Since it is a mass noun, there are difficulties in accepting its plural. It can be modified by adjectives and nouns as in (cheap, high quality) crude oil, but these cannot modify oil: *crude, (cheap, high quality) oil;

- stroke of luck has unrestricted determiners and modifiers, but no insertion is allowed immediately before or near to it, in particular luck cannot be modified: *stroke of good luck*.

8. *Stroke of bad luck* would be a different compound word, whose relation to *stroke of luck* is only etymological.

- board of governors can be modified in several ways: board and governors take separate determiners and modifiers; the powerful boards of the twelve governors of my bank. Such a compound noun comes close to being a free form. It is the limited number of second nouns such as director, governor or regent that suggests we are dealing with a compound noun. Also, the meaning of these phrases is non-compositional in the sense that they have a legal or institutional meaning that their components do not have clearly.

The variations of form we have enumerated can be partly handled by attaching a finite automaton to a given entry, and this automaton will describe the main grammatical changes allowed. The adjunction of free relative clauses to compound nouns may require a different treatment.

The kinds of variation of compound nouns are so numerous that determining whether a given nominal construction is a compound noun or not almost requires a original demonstration. Thus, automatizing the construction of a lexicon is an activity that will present severe limitations.

Determining the support verbs for compound nouns does not seem to raise other problems than those encountered with simple nouns.

**Remark**

Compound nouns raise other questions in some languages:

- in German, where no blanks occur between components, segmentation is a problem;

- in French (G. Gross 1965), where the spelling of the plural is in general not standardized, extra variations have to be expected.

**Compound modifiers**

Adjectives, noun complements and relative clauses can be complex and yet apply to free nouns. From the point of view developed here, that is, the representation in terms of sequences of grammatical categories allowing for efficient matching procedures with texts, they do not differ from adverbs and nouns.

**Examples are:**

- The table is as clean as a new pin.

- The book is up to date.

- Bob is the world's (least, worst) teacher.

- They discussed it, on a take it or leave it basis.

**3. Compound verbs**

Compound verbs or frozen sentences as we have termed them (M. Gross 1982), can be described as sequences of categories. We write $N_i$ for variable noun phrases and $C_j$ for frozen noun phrases. For subjects: $i = 0$, for complements: $i = 1$. Examples are:

\[
\begin{align*}
(1) & \quad N_0 \text{ V } C_1 =: \text{Bob hit the jackpot!} \\
(2) & \quad N_0 \text{ V } N_1 \text{ Prep } C_2 =: \text{Bob took your project into account} \\
(3) & \quad N_0 \text{ V } C_1 \text{ Prep } C_2 =: \text{Bob took the bull by the horns} \\
(4) & \quad N_0 \text{ Prep } C_0 \text{ C_1} =: \text{Bob's dream came true}
\end{align*}
\]

We outlined in 1 the description of a lexicon-grammar of French verbs and the reasons why compound verbs had to be separated from simple ones.

Systematic search through dictionaries (monolingual, bilingual, and specialized) has yielded close to 20,000 compound verbs belonging to the same level of language as the 12,000 simple verbs. A syntactic classification has been built for them (Figure 3).

**Compound verbs** are the most complex forms that have to be entered into a lexicon. The compounds discussed previously were simple.

9. There are however a limited number of frozen discourses such as:

\[
\text{It was for all the world as it S} \]

which need an extra level of complexity (L. Danlos 1985).
because by and large they were topologically commm, that is, either
t heir parts could not be separated by any extraneous linguistic material
or else the inserted material could be easily described (i.e. by means
of a finite automaton).

| Tables | Structures | Examples | Final
|--------|------------|---------|-------|
| CI     | N, V C    | Il a roup le coche | 700   |
| CAN    | N, V (C à de N) | Cela a été la langue de Max (bui) | 500   |
| CDN    | N, V (C de N) | Il a bu le rappel de ses murs | 500   |
| CP1    | N, V Prép C | Il chante dans les bihagies | 1300  |
| CPN    | N, V Prép C C de N | Il abonde dans le sens de Max | 250   |
| CP1N   | N, V C Poly N | Il a déchargé sa balle sur Max | 1750  |
| CPN2   | N, V N, Prép C | Il est passé Max par les sages | 1350  |
| CIP2   | N, V C, Prép C | Il met de l'eau dans son vin | 800   |
| C5     | Que V P Prép C | Que Max reste infirme en sa fauve | 150   |
| Clb    | N, V Que P Prép C | Il a pris du bon côté que Max reste | 500   |
| C7     | N, V À à Que P | Il a dit pour à que Max reste | 150   |
| Clt    | N, V À à Que P | Il se mord les dents de ce qu'il est | 500   |
| CADV   | N, V Adv | Cela ne place pas loin | 200   |
| CN     | N, V N | Il est pur et sans baiser d'adresse | 500   |
| CO     | C, V W | La montagne monte au nex de Max | 1300  |
| AI     | N, avoir C | Il a eu le mot de la fin | 150   |
| A1PN   | N, avoir C, Prép N | Il a barre sur Max | 100   |
| ANP2   | N, avoir N, Prép C | Il a Max en loqueteur | 100   |
| A12    | N, avoir Adj | Il a la vue basée | 100   |
| A1P2   | N, avoir C, Prép C | Il a mal aux cheveux | 250   |
| E01    | C, de N, Prép Adj | La friche de Max est fleurie | 350   |
| E0P1   | C, être Prép C | Les lieux sont du côté de Max | 200   |
| TOTAL  |            |         | 412 800 |

Frozen Verbs
(M. Gross 1982)
Table 3

In the case of compound verbs, the various parts of each utterance
remain syntactically independent. Thus, the verbs of (1)-4) can take
any tensed form, as in:

At that time, Bob will be hitting the jackpot

Sentential inserts can separate a verb from its complements:

Bob hit, it seems to me, the jackpot

in example (2), the direct complement N is free and general; hence,
sentential structures can separate the verb from its second
(frozen) complement:

Bob looked the fact that Jo was absent yesterday into account

Notice that parts of compound verbs may be recognized directly, for
example the jackpot, or into account; but these parts may be
ambiguous, whereas the full utterances can rarely be confused with free
forms.10

4. Some conclusions

How to organize the lexicon of compound utterances is an open
question. From a computational point of view, many solutions are
available for the lookup of a compound term:

(i) In classical algorithms in which left-to-right analysis is
essential, the compound term could be viewed as an extension of the
first major element met while scanning the sentence. For example, the
adjective long is the first such element of the compound in the long run.
Among many other possibilities, the program, passing on the word long
would test the occurrence of the in and in the to the left of long, and the occurrence of run to the
right. Notice that the left-to-right constraint has to be somewhat
relaxed in order to test both left and right contexts of long.

(ii) In a futuristic view of parsing involving parallel computing, one
might envision several levels of lexicon. At the first level, long
on the one hand and run on the other, would be two sets of
constructions whose interaction would contain the compound in the
long run; the latter can then be searched for in the input text. For
compound verbs, one would have to synthesize a matching utterance,
rather than simply looking it up. Such a procedure can always be
implemented sequentially.

In all cases, the representation of utterances which we have used,
namely the sequences of syntactic categories, allows for the separation
of the lexicon of compound terms into classes for which direct access
can be provided. In this way, dictionary lookups can be sped up.11

REMARK

In favor of left-to-right analysis one could point to the fact that
complex forms can occur in abbreviated and that abbreviations are mostly
right truncations. In such situations the remaining part (the "leftmost"
part) of the truncated term must carry the information that describes
the right context in order to allow reconstruction of the reduced part.
There are however examples where abbreviations are carried out on the
left part of a term, (e.g. a programming language = a
language).

Preliminary figures have shown that compound terms form the
essential part of a lexicon-grammar. It is also interesting to observe
that they force both the linguist and the computer specialist to adopt a
much more abstract view of language.

- semantically, by definition, compound utterances cannot be decomposed
into simple utterances; in other terms, meaning is not compositional
for compounds. Hence, in a certain sense, one has to recognize that meaning
has not much to do with words;

- syntactically, it has become a rather general habit to attach
properties to individual words. In the case of compounds this mode of
representation is no longer possible. Why privilege one part of a
compound with marks rather than some other part? For example, there
is no reason to attach the Passive marking to the verb rather than to
one of the complements of the utterance to put the cart before the
horse. Lexicon-grammar representations eliminate such questions by
defocalizing the syntactic information and by attaching it to the full
sentence. In this sense, compound expressions provide a powerful
motivation for representing lexical and syntactic phenomena in the form
of a lexicon-grammar.

10. As a matter of fact, when an utterance is taken to be ambiguous,
with one analysis as a frozen form and the other as a free form,
ignoring competing free forms altogether is a good parsing strategy.

11. The same use of sequences of syntactic categories is found in a
strong grammar (Z.S. Harris 1961), which has proven to be quite
efficient in syntactic recognition (N. Sager 1981, M. Salkoff 1973,
1979).
REFERENCES

Boeas, Jean-Paul., Guillet, Alain. and Lecîbre, Christian. 1976a. La structure des phrases simples en français. I Constructions intransitives, Geneva: Droz, 377p.

Boeas, Jean-Paul., Guillet, Alain. and Lecîbre, Christian. 1976b. La structure des phrases simples en français. III Classes de constructions transitives, Rapport de recherches No 6, Paris: University Paris 7, L.A.D.L., 143p.

Boeas, Jean-Paul., Guillet, Alain. and Lecîbre, Christian. 1987. La structure des phrases simples en français. II Classes de constructions locatives, Paris: Cantilène.

Danlos, Laurence. 1985. Génération automatique de textes en langues naturelles, Paris: Masson, 239p.

Gross, Gaston. 1985. Le langage électronique des mots composés du français, Rapport ATP CNRS, Paris: University Paris XIII.

Gross, Gaston; Vivès Robert, eds. 1986. Syntaxe des noms, Langue française 69, Paris: Larousse, 129p.

Gross, Maurice 1975. Méthodes en syntaxe, Paris: Hermann, 414p.

Gross, Maurice 1981. Les bases empiriques de la notion de prédicat sémantique, Langages 63, Paris: Larousse, pp.7-52.

Gross, Maurice 1982. Une classification des phrases figées de français, Revue québécoise de linguistique, Vol. 11, No 2, Montréal : Presses de l'Université du Québec à Montréal, pp.151-195.

Gross, Maurice 1986. Grammaire transformationnelle du français. III Syntaxe de l'adverbe, Paris : Cantilène.

Gross, Maurice; Tremblay, Diane 1985. Étude du contenu d'une banque terminologique, Rapport de recherche du LADL, Paris: MIDIST.

Harris, Zellig S. 1961. String Analysis of Sentence Structure, Papers on Formal Linguistics, The Hague: Mouton.

Harris, Zellig S. 1964. The Elementary Transformations, Transformations and Discourse Analysis Papers 54, in Harris, Zellig S. 1970, Papers in Structural and Transformational Linguistics, Dordrecht: Reidel, pp.482-532.

Harris, Zellig S. 1976. Notes du cours de syntaxe, Paris : Le Seuil, 237p.

Harris, Zellig S. 1982. A Grammar of English on Mathematical Principles, New York: Wiley Interscience, 426p.

Sabatier, Paul 1980. Dialogue en français avec un ordinateur, Doctoral thesis, Marseille: Groupe d'Intelligence artificielle.

Sager, Niels 1981. Natural Language Information Processing. A Computer Grammar of English and Its Applications, Reading: Addison-Wesley, xvi-599p.

Salkoff, Morris 1973. Une grammaire en chaîne du français. Analyse distributionnelle, Paris: Dunod, xv-199p.

Salkoff, Morris 1979. Analyse syntaxique du français. Grammaire en chaîne, Amsterdam: John Benjamins B.V., 334p.