Lexicon-Grammar
Maurice Gross

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

*Chronometric 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 table or class code. In figure 1, the code 6 corresponds to the class of constructions subject-verb-direct sentential complement, noted:

\[
\begin{array}{c}
\text{(S)} \quad \text{Prep:} \\
\text{Bob knows that Max has moved to Tampa}
\end{array}
\]

Each column is a syntactic property, and corresponds to a structure into which a verb may enter, roughly a syntactic transform of the main structure. For example, in columns we have placed the Passive forms, Extramed and adnominal forms. Thus, the related structures are semantically close. "+" sign at the intersection of a row and a column indicates that the entry in the row is accepted in the structure associated to the column, while "-" sign corresponds to insufficiency. The process of accumulation has led to the formalized lexicon-grammar of 12,000 French verbs which has run into what seemed to be at first 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 know and complex (idiomatic) forms such as keep in mind. Both forms play the same syntactic and semantic role in sentences such as:

\[
\begin{align*}
\text{Bob} & \text{ knows that Max has moved to Tampa} \\
\text{Bob keeps in mind that Max has moved to Tampa}
\end{align*}
\]

but the lexical content (one word vs three) requires different identification procedures (simple dictionary lookup vs a certain amount of syntactic analysis).

The representation of figure 1 treats two forms such as *to know (someone, something)* and *to keep (someone, something) in mind* in the same way, thus emphasizing the semantic equivalence between simple and compound verbs.

Not 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 nightly 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. Note that words or roots are often considered as units in most attempts to devise semantic representations.

TABLE 6 Verbs with Sentential Complements
(From M. Gross 1975)

Figure 1

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 context:

it rained cats and dogs
*many cats and dogs
*big cats and dogs
*cat and dog
Consequently, these compound adverbs could be identified by a simple recognition procedure, for 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. Running texts have been examined as well. It is interesting to note that whereas in current dictionaries there are about 1,500 one word adverbs, in general, six word adverbs (e.g. for example, consequently, consequently, consequently). Consequently, these compound adverbs could be identified by a simple recognition procedure, for they require no lemmatization or syntactic analysis to be reduced to a dictionary form, as is the case with verb forms for example.

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 \( P \), \( D \), \( A \), \( N \), \( V \), \( C \) (for conjunction) and \( W \) for a variable ranging over verb components, etc. We write:

\[
\begin{align*}
\text{Prep} & \quad \text{N} & \text{=: at night} \\
\text{Prep} & \quad \text{Det} & \quad \text{N} & \text{=: in the end} \\
\text{Prep} & \quad \text{Det} & \quad \text{Adj} & \quad \text{K} & \text{=: in the long run} \\
\text{Prep} & \quad \text{Det} & \quad \text{N} & \quad \text{of} & \quad \text{Det} & \quad \text{N} & \text{=: in every sense of the word at the point of a gun} \\
\text{Prep} & \quad \text{Det} & \quad \text{Adj} & \quad \text{V} & \quad \text{W} & \text{=: to begin with} \\
\text{S} & \text{=: all things being equal}
\end{align*}
\]

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:

\[
\text{Max proposed solutions from the top of his hat}
\]

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, \( \text{Max's} \), may vary. This possessive adjective must refer to the subject of the sentence, and varies accordingly:

*Max proposed ideas from the top of your hat*

*My sister proposed ideas from the top of his hat*

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 leniency or lightly and the referential information carried by \( \text{Pass} \) can then be ignored. However, one can easily construct particular discourses where the obligatory coreference relation involved will disambiguate some analyses. Thus, not only the variation of \( \text{Pass} \) must be accounted for at the lexical level, but its referential information has to be kept for possible use in a parser.

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

<table>
<thead>
<tr>
<th>Table</th>
<th>Structure</th>
<th>Examples</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>PADV</td>
<td>Adv</td>
<td>seulain</td>
<td>328</td>
</tr>
<tr>
<td>PC</td>
<td>P &amp; C</td>
<td>en bref</td>
<td>460</td>
</tr>
<tr>
<td>PDEC</td>
<td>P &amp; D C</td>
<td>contre toute étrangère</td>
<td>570</td>
</tr>
<tr>
<td>PAC</td>
<td>P &amp; A C</td>
<td>de sa belle mère</td>
<td>440</td>
</tr>
<tr>
<td>PCA</td>
<td>P &amp; C A D</td>
<td>à gorge déployée</td>
<td>400</td>
</tr>
<tr>
<td>PCDC</td>
<td>P &amp; D C</td>
<td>en désespoir de cause</td>
<td>350</td>
</tr>
<tr>
<td>PCDN</td>
<td>P &amp; C N</td>
<td>au moyen de N</td>
<td>330</td>
</tr>
<tr>
<td>PCNP</td>
<td>P &amp; C prép N</td>
<td>par rapport à N</td>
<td>90</td>
</tr>
<tr>
<td>PCPC</td>
<td>P &amp; C prép C</td>
<td>des pieds à la tête</td>
<td>140</td>
</tr>
<tr>
<td>PCNP</td>
<td>P &amp; C conj C</td>
<td>en tout et pour tout</td>
<td>170</td>
</tr>
<tr>
<td>PV</td>
<td>P &amp; V W</td>
<td>à dire vrai</td>
<td>150</td>
</tr>
<tr>
<td>PF</td>
<td>P (phrase figée)</td>
<td>Dieu seul le sait</td>
<td>239</td>
</tr>
<tr>
<td>PICO</td>
<td>(A) comme C</td>
<td>comme ses pieds</td>
<td>200</td>
</tr>
<tr>
<td>PICO</td>
<td>(V) comme C</td>
<td>comme un cheveu sur la soupe</td>
<td>210</td>
</tr>
<tr>
<td>PICO</td>
<td>(V) comme prép C</td>
<td>comme dans le bureau</td>
<td>30</td>
</tr>
<tr>
<td>PIC</td>
<td>conj C</td>
<td>et tout le tremblement</td>
<td>100</td>
</tr>
<tr>
<td>TOTAL</td>
<td>3,419</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Frozen Adverbs (N. Gross 1976)

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:

\[
\text{Max proposed solutions from the top of his hat}
\]

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, \( \text{Max's} \), may vary. This possessive adjective must refer to the subject of the sentence, and varies accordingly:

*Max proposed ideas from the top of your hat*

*My sister proposed ideas from the top of his hat*

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 leniency or lightly and the referential information carried by \( \text{Pass} \) can then be ignored. However, one can easily construct particular discourses where the obligatory coreference relation involved will disambiguate some analyses. Thus, not only the variation of \( \text{Pass} \) 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, at 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 adverbials:

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:

6. PROLOG 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.

Enclosing 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):

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

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 part of the elementary sentence. Consider the adverb in:

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

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:

\[
\text{The way Bob sang was (noisy, powerful)}
\]

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 then: whereas they lead to can only be solved by introducing a complex level of semantic equivalence for the entries of the lexicon-grammar.

2. Compound nouns

Compound nouns form the bulk of the lexicon of languages. Language creativity is largely associated with the growth of technical vocabularies which consist mainly of technical nouns. Compound nouns number in the millions for European languages. They are usually built from the vocabulary of simple words by means of grammatical rules which may involve grammatical words. By definition, their meaning is recompositional. The compound nouns can be described in terms of the sequence of their grammatical categories, in the same way as for adverbs (M. Gross, D. Tremblay 1995). We have for example:

\[
\begin{align*}
\text{Det N} & \Rightarrow \text{the moon} \\
\text{Adj N} & \Rightarrow \text{crude oil, real estate} \\
\text{N of N} & \Rightarrow \text{stroke of luck, board of governors, regents} \\
\text{Det N of Det N} & \Rightarrow \text{the talk of the town} \\
\text{N N} & \Rightarrow \text{feel lula, color IV}
\end{align*}
\]

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:

-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 board 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 noncompositional 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, automatising 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 1985), 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:

\[
\begin{align*}
\text{The table is as clean as a new pin} \\
\text{The book is up to date} \\
\text{Bob is the world's (best, worst) teacher} \\
\text{They discussed it, on a take it or leave it basis}
\end{align*}
\]

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 \( N_j \) for variable noun phrases and \( C_k \) for frozen noun phrases. For subjects: \( i = 0 \), for complements: \( i = 1, 2 \). Examples are:

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

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 30,000 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\(^8\). The compounds discussed previously were simple.

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

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

which need an extra level of complexity (L. Danlos 1985).
because by and large they were topologically commutative, that is, either
their 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).

<table>
<thead>
<tr>
<th>Table</th>
<th>Structures</th>
<th>Examples</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>N0, V C1</td>
<td>Il a coupé le coche</td>
<td>2400</td>
</tr>
<tr>
<td>C1N</td>
<td>N0, V (C a de N)</td>
<td>Cela a été la langue de Max (but)</td>
<td>500</td>
</tr>
<tr>
<td>C1DN</td>
<td>N0, V (C de N)</td>
<td>Il but le rappel de ses amis</td>
<td>500</td>
</tr>
<tr>
<td>C1P</td>
<td>N0, V Prép C1</td>
<td>Il change dans les pigments</td>
<td>1300</td>
</tr>
<tr>
<td>C1PN</td>
<td>N0, V Prép C de N)</td>
<td>Il abonde dans le sens de Max</td>
<td>250</td>
</tr>
<tr>
<td>C1PN2</td>
<td>N0, V prép C1, Poby N0</td>
<td>Il a déchargé sa bûche sur Max</td>
<td>1750</td>
</tr>
<tr>
<td>C1PN2</td>
<td>N0, V prép C1, Poby N0</td>
<td>Il est passé Max par les uvaux</td>
<td>1350</td>
</tr>
<tr>
<td>C1PN2</td>
<td>N0, V C1, Prép C1</td>
<td>Il met de la peau dans son jeu</td>
<td>800</td>
</tr>
<tr>
<td>C5</td>
<td>Que P V Prép C4</td>
<td>Que Max reste malade de sa faim</td>
<td>150</td>
</tr>
<tr>
<td>C6</td>
<td>N0, V Qu P Prép C5</td>
<td>Il a pris du bon côté que Max reste</td>
<td>300</td>
</tr>
<tr>
<td>C7</td>
<td>N0, V C1 à ce Qu P</td>
<td>Il a dit oui à ce que Max reste</td>
<td>150</td>
</tr>
<tr>
<td>CII</td>
<td>N0, V C2 à ce Qu P</td>
<td>Il se moule dans le rôle de ce que Max reste</td>
<td>300</td>
</tr>
<tr>
<td>CADY</td>
<td>N0, V Adv</td>
<td>Celui ne place pas loin</td>
<td>200</td>
</tr>
<tr>
<td>CX</td>
<td>N0, V X</td>
<td>Il est parti sans baiser d'adresse</td>
<td>300</td>
</tr>
<tr>
<td>CO</td>
<td>C1 X W</td>
<td>La monture monte au nez de Max</td>
<td>1300</td>
</tr>
<tr>
<td>AI</td>
<td>N1 avoir C1</td>
<td>Il a eu le mot de la fin</td>
<td>150</td>
</tr>
<tr>
<td>A1P1</td>
<td>N2 avoir C1, Prép N1</td>
<td>Il a barre sur Max</td>
<td>160</td>
</tr>
<tr>
<td>A1P2</td>
<td>N0, N1 Prép C1</td>
<td>Il a vu le loyer</td>
<td>160</td>
</tr>
<tr>
<td>AI2</td>
<td>N1 avoir Adj</td>
<td>Il a la vue basse</td>
<td>100</td>
</tr>
<tr>
<td>AI2P1</td>
<td>N0, N1 avoir C Prép C1</td>
<td>Il a mal aux cheveux</td>
<td>230</td>
</tr>
<tr>
<td>E01</td>
<td>C1 de N autre Adj</td>
<td>La trêve de Max est fleurie</td>
<td>350</td>
</tr>
<tr>
<td>E0P1</td>
<td>C2 être Prép C1</td>
<td>Les reins sont du côté de Max</td>
<td>200</td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
<td></td>
<td>347 800</td>
</tr>
</tbody>
</table>

**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 N1 is free and general, hence, sentential structures can separate the verb from its second (frozen) complement:

Bob took 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.

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 noun element met while scanning the sentence. For example, the adjective long is the first such element of the compound adverb in the long run. Among many other possibilities, the program, passing on the word long would test the occurrence of the and in the to 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 intersection 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 simulated 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 forms into classes for which direct access can be provided. In this way, dictionary lookup can be sped up.11

**Remark**

In favor of left-to-right analysis one could point to the fact that complex forms can often be abbreviated and that abbreviations are mostly right truncations. In such situations the remaining part (the suffix 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 or 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 either 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 found 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 proved to be quite efficient in syntactic recognition (N. Sager 1981, M. Salkoff 1973, 1979).
REFERENCES


