Methods and Tactics in the Construction of a Lexicon-grammar*

Maurice Gross

University Paris 7

By building a lexicon and a grammar for a language, we mean attempting to reach a significant coverage of the syntactic forms and the semantic units that are contained in the language.

We distinguish our programme from those of traditional grammar and commercial lexicography, two activities which also aim at obtaining adequate coverage for languages. The distinction is imposed by the theoretical framework adopted for the description: the transformational theory of Zellig S. Harris, which, along with most modern theories, requires complete formalization of the data. These modern frameworks substantially modify the notion of coverage, since many variations of forms and numerous details neglected in most traditional approaches have to be entirely spelled out.

No reasonably complete formalized grammars and lexicons exist so far, for any language, although a strong demand is developping in the community of Artificial Intelligence and Computational Linguistics.

We shall describe our experience based on the work performed by a team of about ten linguists who have been working on the construction of a French lexicon-grammar over a period of about fifteen years, and the results that have been obtained. First, we discuss the methods and working principles that have shaped up in the course of this research enterprise.

1. Selecting a vocabulary

We do not attempt, at least as a first step, to collect the largest possible vocabulary. Since we want to construct a formalized grammar of the language, we are committed to use combinatorial methods to determine which sequences of words are accepted and/or grammatical. Thus the judgments

The Linguistic Society of Korea (ed.), 1988,
Linguistics in the Morning Calm 2, Seoul: Hanshin Publishing Co.

of acceptability which are needed in the course of this activity can only be carried out on constructions in which the meaning of all words are known.

For example, in order to study French verbs, we started with lists whose size was about 12,000 verbs. These are lists that have been compiled by various authors of conjugation manuals (e.g. Bescherelle, Larousse). Retaining only verbs that could reasonably well be submitted to linguistic manipulations led us to a working list of about 6,000 verbs.[1]

This obvious requirement eliminates many words obsolete or too technical to be known by all linguists. This requirement is also important for the reason that it prevents interferences between accidental knowledge of the vocabulary by an individual and the general scientific requirement of reproducibility of the acceptability judgments.

Indeed, reproducibility of judgments is essential (Cf. 2). In our team, no single information has been retained, unless its validity was corroborated by the independent judgments of several linguists: at least two, preferably three or four. All information resulting from acceptability judgments have to be made explicit and preferably published, in order to be exposed to critical evaluation by other linguists.

Since our major criterion for retaining lexical entries is mastery of their use, thus enabling linguists to perform experiments on them, various lexical factors have been ignored. For example, literary, technical and slang entries may be retained if they are tractable, hence independently of their belonging to different levels of discourse. As a matter of fact, there is not much choice but to ignore such distinctions, when there is no reliable classifying procedure for the vocabulary based on them. In case a classification based on levels of language could be effectively devised, for example by means of statistics on an extremely large corpus, the lexical items already described could always receive marks or features which do not seem to have any bearing on the syntactic properties currently being studied.

2. Acceptability

Evaluating the acceptability of a string of words is considered as performing a syntactic experiment. Hence, the requirement of reproducibility is as crucial as in any scientific experiment.

An acceptability judgment has many components, cultural and psychological. Since we want to retain the sole syntactic component, procedures have to be devised to eliminate the influence of extra-syntactic factors. The famous example:

(1) *Green ideas sleep furiously*

can be used to illustrate some of these procedures. There has been many

discussions about the acceptability status of (1). We consider it as syntactically acceptable, although no interpretation can be immediately attached to it. The main reason is that (1) can be easily read and pronounced. This is so because an intonation pattern can be placed on it. One can reasonably suppose that this intonation pattern is determined by syntactic clues such as grammatical categories, a supposition confirmed by the fact that a string such as:

(2) *Dallable pertussions brated dossally*

can receive the same intonation pattern. String (2) is composed of artificial word forms that do not belong to the English vocabulary, but their morphological structure imitates that of English words since they have endings determining their syntactic category. Hence, both (1) and (2) can be analyzed as:

(3) *Adj N V Adv*

and without having to make any hypothesis about the possibility that this sequence of categories receives a superstructure (e.g. a tree), a unique intonation pattern can be given to it.

A further observation is that a string such as (1) can be easily memorized.[2] This observation does not apply to the strings:

(4) *Sleep ideas green furiously*

Brated pertussions dallable dossally

which cannot be easily pronounced, that is, without break points in the intonation, nor memorized. The reason appears to be that no intonation pattern can be put on the sequence of categories:

V N Adj Adv

Memorization is certainly facilitated by the use of a standard intonation pattern.

Such examples have been devised (by N. Chomsky 1957 and L. Tesnière 1959) in order to show that syntactic structures exist independently of semantic interpretations. For us, they are important in a different respect, they show that it is possible to avoid interferences between the components of a language by devising experimental procedures that will render much more specific acceptability judgments that tend to be too global, because in general they merge intuitions of varied origins.

Thus we insist on our general position which asserts that carrying an acceptability judgment on a string is not at all emitting an opinion based on a cursory reading of the string, but consists in devising an experiment aiming at the demonstration of some hypothesis. Hence, careful building of examples, with appropriate choice of words should be made; in order to

avoid ambiguities that too often bring parasitic intuitions. In general, sentences that constitute actual written or spoken discourses do not have such particular properties. Thus, examples have to be constructed with a particular testing aim in view. This can only be done by professional linguists. As a consequence, the procedure that consists in interviewing native speakers in order to verify whether a given string of words is acceptable or not has to be limited to professional circles. One could also argue that if such a question is to be asked, it is because the data are unstable. The question then becomes a study about the distribution of the acceptabilities, and as such belongs to the field of socio-linguistics rather than to structural syntax.

We insist on the fact that acceptability judgments be reserved to professional linguists, mainly because evaluating a given sentence may take on a large variety of points of view, thus eliciting a large spectrum of commentaries belonging to semantics, style, intonation or expressivity. Often, dubious examples can be interpreted in terms of plays on words, jokes, incongruities or vague autobiographic memories.

Besides this precaution, the fact that recourse to intuition may not be entirely reliable imposes that acceptabilities be determined with great care. The following guidelines should be taken into account:

(1) When an acceptability judgment is not convincing, TRY EXAMPLES THAT INVOLVE WORDS SYNONYMOUS (perhaps antonymous) to those of the original examples. Examples can often be varied within a given range of structures, and since the uncontrolled component acceptabilities may depend on choice of words, shifting to a similar lexical item may lead to sharper judgments. Conversally, analogical arguments can be used in the following situation: Suppose one has arrived at a given set of similar verbs, for example semantically similar verbs (e.g. verbs of saying). Suppose now that one has to carry out an acceptability judgment on one of the verbs of this class. (e.g. verify whether *to gargle* accepts the structure *I gargled to Max that he should leave*). Suppose moreover that judgments on this form are not conclusive. In this case, one should check all verbs of the class, and concentrate on those which provide clear data (e.g. *to tell*). Then, the existence of a class leads to accept dubious cases. A negative judgment within the class should then be extremely sharp in order to be recorded as such.

(2) LOOKING FOR ATTESTATIONS. By chance, the example under test may be given in a dictionary (often in a slightly different form). Dictionaries are intended to represent usage, that is, current texts. Examples have been chosen in a representative way, but may not correspond to the precise syntactic property under discussion. Searching a corpus for significant examples may constitute a verification procedure. However it should

be reminded that as of to day, computerized searching procedures have not been developped to the point where they could be used in the evaluation of acceptabilities.[3]

(3) LET THE THEORY DECIDE. In most cases, acceptability judgments are performed in order to verify a hypothesis or a rule of grammar or a theory. Since grammars are characteristic devices which separate strings into two types: grammatical or not without intermediary cases, terms such as "dubious" or "intermediate" acceptabilities have no status with respect to such rules. A systematic attitude could consist in accepting all dubious forms that are determined as grammatical by the rules involved.

Remarks

(i) Since theoretical factors may take part in the evaluation of acceptabilities, the restriction to professional circles is further justified.
(ii) The set of sentence forms marked as accepted corresponds to sentences of varied acceptabilities, but variations are not represented. Hence the relations between actual sentences and their representation in a lexicon-grammar is quite complex. In other words, the French language represented is a theoretical brand of French, its relation to reality cannot be expected to be simple.

(4) ACCEPT ALL DUBIOUS FORMS. A third reason for obeying this principle is empirical: Experience has shown that intuition underestimates acceptabilities. Many dubious forms can be improved by adding adjectives or adverbs, that is without altering the initial structure. However, one should be cautious when one lengthens sentences: long and complex sentences tend to be many times ambiguous, hence they offer more opportunities of accepting them, possibly unjustified ones.

Notice that an acceptability judgment must be carried out on a full sentence. Judgments made on phrases may be valid in simple cases, such as:

a blue cup of tea

We know immediately that this phrase is well formed from grammatical information we have internalized. But notice that a certain amount of uncertainty is present when one deals with the phrase *a cup of tea* which can be either interpreted as a container or as an amount of tea. Besides these trivial examples, we encounter phrases such as (Z. S. Harris 1976):

a quick cup of tea

and here, it is clear that no judgment can be carried out outside a sentence because of the difference:

**I broke a quick cup of tea*
I drank a quick cup of tea

The empirical requirement of restricting judgments to sentences coincides with a theoretical position essential to dealing with lexicon-grammars:[4]

THE UNIT OF MEANING IS THE ELEMENTARY SENTENCE, AND NOT THE WORD.

By elementary sentence, we mean (for European languages at least), the structures subject-verb-essential complements.

This postulate is obviously necessary in the case of verbs. From a syntactic point of view, a given morphemic verb has no status. Subjects and objects are entirely determined and are unique for each verb. The same remark carries over to *be Adjective* forms. For nouns, it is less obvious so, and we will discuss this point below in 4. (Cf. M. Gross 1981). For adverbs, the treatment given by Z. S. Harris 1976, provides a general solution. As for the other categories of words, most of them are grammatical constants belonging either to basic structures (of verbs, adjectives, nouns, adverbs) or to the transformations they undergo.

3. Separating entries underlying a word

It is well known that many word forms are ambiguous, and dictionaries provide a good picture of their different meanings, at least in examples such as the word *bear* which is a noun on the one hand, and on the other a verb which has nothing to do with the noun.

However, there are many situations where distinctions are not made and where close examination by means of syntactic tools reveals differences not recorded in dictionaries. Whereas lexicographers proceed largely on an intuitive basis when they choose examples to reflect the meanings of words, the construction of a large formalized lexicon cannot rely on meaning intuitions. Consequently, formal syntactic criteria, which guarantee reproducibility of judgments, will constitute our basis for the separation of the meanings of words.

As an illustration of this activity, consider the verb *to daydream* in the constructions:

(1) *Bob daydreamed about Lea's not arriving in time*
(2) *Bob daydreamed that Lea did not arrive in time*

These two constructions of the verb form *daydream* being different, one has to decide whether one is dealing with one or two verbs. In the hypothesis of one verb, one must describe the synonymy relation (1) = (2) by a

syntactic process. One could use the basic unattested form:

(3) **Bob daydreamed about that Lea did not arrive in time*

and derive from it:
—(1), by a nominalization rule applied to the sentential complement *that S*,
—(2), by deleting *about*.
Although these analyses have been successfully argued for in other contexts, there is a difficulty which is not accounted by this solution. Consider the following discourses:

(4) *Lea did not arrived in time, Bob daydreamed about it*
(5) *Lea did not arrived in time, Bob daydreamed it*

In (4), the pronoun *it* must refer to the sentence of the first member, which can be represented by the relation:

(6) *it = Lea did not arrive in time*

In (5), the meaning of *it* can be either (6) or:

(7) *it = Lea arrived in time*

Short of accounting for such a difference by syntactic means, we have to conclude that there are two independent entries for the verb *to daydream*, in the sentences (1) and (2).

Intuitions involving what had been informally called metaphoric meanings of a word may signal a situation where several lexical entries are involved. The first question that has to be asked then is whether one is dealing with an isolated case or with an open family.

A terminology such as: *Sentence (phrase or word)* (A) *is a metaphor of sentence (phrase or word)* (B), implies a relation between (A) and (B). The terminology 'metaphor' implies the existence of at least two meanings attached to (A) and (B). A common example of this situation is the pair proper and figurative meanings. Questions about the nature of the relation between the two meanings have to be raised. First consider isolated cases, for example:

The stock market exploded yesterday
The Dow Jones exploded yesterday

which is considered as a figurative meaning, whereas:

The bomb exploded yesterday

corresponds to proper meaning. Consider now a close synonym of *to explode, to burst*. The sentence:

The stock market burst yesterday

is not accepted. A similar example is the following, in which associated to the proper meaning:

The bomb burst out

we find a figurative meaning of *to burst*:

Lea (burst, broke) into tears

which is called a metaphor. We do not accept:

Lea blew up into tears

which is parallel to:

The bomb blew up into a hundred pieces

We have here examples where the relation proper-figurative is highly restricted: only one verb allows it, close synonyms are excluded from it. Sometimes, unexpected verbs can enter into the pair, while synonyms are excluded (e.g. *Lea melted into a flood of tears* is accepted).

The discussion about lexical extensions of metaphors as in our previous example constitutes a proof that the metaphor is an isolated case. Situations can vary further, as the study of the nouns dependent on the verbs can show. We find two main cases:

—the nouns are free, within distributional limitations, as in the case of the subjects of:

The (costs + market + price) exploded yesterday

in this case we consider that this meaning of *to explode* is an entry independent of the entry with the proper meaning. Notice that there exist syntactic arguments for separating the two uses into separate entries. We have the following difference in nominalization:

There was a sudden explosion of the stock market, and all the prices soared
There was a sudden explosion of the bomb, and the room collapsed

The bomb made a tremendous explosion
?**The stock market made a tremendous explosion*

—the metaphor seems limited to the combination of particular words, as with:

Lea burst into tears

The combination *burst-tear* is particular in this range of meaning, thus the

sentence has to be considered as frozen.

We have encountered several thousands such situations and we have not detected any regularity in the formation of these types of metaphors. It is often easy to reconstruct the origins of these metaphors. But in such a process, there is often more imagination, that is folk etymology, than serious diachronic work.

Thus, whether free or frozen, the mechanism of formation of new entries, which starts from an existing utterance is viewed as a historical process, hence not belonging to the syntactic description of the language. We do not deny that we have here one or several important creative processes at work, but we consider that the devices involved in the institutionalization and the linguistic freezing of new utterances into the language is of a nature that has not much to do with the combinatorial processes used for the description of the lexicon-grammar. We think that the fact initial forms and new forms coexist in the language has misled many authors into looking for a synchronic relation between such forms, the motivation being that often the new form shares syntactic properties with the initial form. This parallelism of forms is explainable by any diachronic process, but what can never be explained is the fact that new syntactic differences appear most of the time between the initial and the new form. Namely, the difference in syntactic behaviors can never be correlated to any syntactic or lexical property of the initial form. As far as we can tell from the numerous cases studied, the acquisition of a syntactic paradigm by a new metaphor is entirely accidental. This is rather surprising in a sense: since syntactic restrictions are largely unconscious, one should have thought that mechanical processes (e. g. redundancy rules) would attribute predictable properties to utterances. However, we know that the semantic and conscious steps that have been taken to create a metaphor or an idiom are entirely anecdotal; specialized dictionaries do record many entertaining sources that do not seem to share any semantic regularity, but for syntax a parallel anecdotal situation is harder to accept.

Within present frameworks, the only possibility would be that there exists an interpretive device acting between meaning and syntactic properties. But the device cannot be the traditional relation considered by grammarians:

—first, it would have to be independent of the meaning of individual words (because of idioms),

—second, the nature of the syntax involved would be quite mysterious, the meaning cannot be linked to known syntactic properties, since examination of several thousand cases of verbal entries has failed to reveal any correlation.

One can of course always speculate about a biological (neurological?)

relation between meaning and form, but it would have to be of a purely computational nature, and at that point even significant questions are quite difficult to raise.

Remark

Notice that in a frozen form such as:

Bob took exception to your remark

which cannot be called a metaphor, or such as the idiom:

Bob carried coal to Newcastle

it is difficult to say that we are dealing with the nouns *exception, coal* or *Newcastle,* or else one has to say that each of these nouns has a special meaning (aside from the ordinary meaning(s) described in dictionaries) and that the special meaning is limited to the corresponding preceding sentences. In such situations, it is clear that the nouns cannot be separated from the sentences to which they belong, and that it is preferable to say that one is dealing with unanalysable units. The same is true for complex expressions other than sentences, for example .with compound nouns or adjectives such as:

red tape, magnetic field, odd number, tax free

4. The lexicon-grammar of French

We have applied the methodology we exposed to the French language, and we now summarize the stage of the construction of a lexicon-grammar of French that has been reached today.

First, in order to provide a quantitative background, we recall that the current commercial dictionaries of French contain numbers of simple words in the following orders of magnitude:

—30.000 nouns,
—12,000 verbs,
—12,000 adjectives,
—1,500 adverb.

With this way of counting, a French verb such as *voler* will be counted as one morphological unit, although in our lexicon-grammar it will have at least the two entries that mean *to fly* and *to steal.*

Work on the lexicon-grammar starts from this type of elementary material.

In the case of verbs, which is the part of the lexicon-grammar which has been the most deeply investigated, among the 12,000 verbs, linguists re-

tained only about 6,000, a subset that could be safely submitted to acceptability judgments:

N_0 V	500
N_0 V N_1	5 000
N_0 V *à* N_1	500
N_0 V *de* N_1	300
N_0 V *Prep* N_1	1 200
N_0 V N_1 N_2	110
N_0 V N_1 *à* N_2	1 200
N_0 V N_1 *de* N_2	1 500
N_0 V N_1 *Prep* N_2	1 500
N_0 V *à* N_1 *de* N_2	4
N_0 V $Prep_1$ N_1 $Prep_2$ N_2	130
Total	11 944

The notations are transparent: the N_i's are noun phrases: N_0 is the subject, N_1 and N_2 the complements. *Prep* holds for prepositions other than *à* and *de*, described here are direct and indirect objects, along with some locative and instrumental complements which have a special relation to the verb.

TABLE 1

Separating the various meanings and constructions of these verbs into independent entries has led to a set of almost 12,000 verbal entries. The main syntactic types are given in table 1, together with the number of verbs corresponding to each type. The actual classification is more refined than table 1 shows, and the sizes of the classes are better balanced. About 50 classes of verbs have been defined by means of syntactic criteria, each contains a few hundred verbs.

More precisely, a class of verbal entries is defined in terms of one or two syntactic structures into which all of its verbs may enter. Moreover, to each class is attached a description of other sentence forms that individual verbs may enter or not. Classes are thus outlayed as matrices (or tables) whose rows contain verbs and columns sentence forms. At the intersection of a row and a column a "+" sign appears when the corresponding verb and structure match, a "—" sign otherwise (cf. table 2).

On the whole, the sentence forms that appear in the columns of a given class (or table) are transforms of the simple declarative sentence which defines the class. Hence, one finds for example passive forms, extraposed and pronominal forms and the possibility to accept a sentential comple-

Subjet N_{hum}	Subjet N_{nc}		Auxiliary *avoir*	Auxiliary *etre*	N_0 *est Upp*	N_0 *U*	Indirects Complements	*que p*	*que Psubj Q*	Infinitives $T_p = T_c$	Infinitives T_c = "Hpast"	Infinitives T_c = " present"	Infinitives T_c = "future"	Infinitives V_c = *devoir*	Infinitives V_c = *pouvoir*	Infinitives V_c = *savoir*	*ppv*	N_0 *U Prep* N_1	N_{hum}	N_{-hum}	*ppv*	*dans N*	N_0 UN_1	N_{hum}	N_{-hum}
+	+	manquer	+	−	−	−	de	−	−	+	−	−	−	−	+	+	−	−	−	−	−	−	−	−	−
+	+	ne manquer Neg	+	−	−	−	de	−	−	+	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−
+	+	menacer	+	−	−	−	de	−	−	+	−	−	−	−	+	+	−	−	−	−	−	−	−	−	−
+	+	se mettre	−	+	−	−	a	−	−	+	−	−	−	−	+	+	−	−	−	−	−	−	−	−	−
+	−	negliger	+	−	−	−	de	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
+	−	omettre	+	−	−	−	de	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
+	−	oser	+	−	−	+	0	−	−	+	−	−	−	−	−	+	+	−	−	−	−	−	+	−	+
+	−	oublier	+	−	−	−	de	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
+	−	passer	+	−	−	−	pour	−	−	−	+	+	−	+	+	+	−	+	+	+	−	−	−	−	−
+	−	perseverer	+	−	−	+	a	−	−	+	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−
+	+	persister	+	−	−	+	a	−	−	+	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−
+	+	pouvoir	+	−	−	−	o	−	−	−	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−
+	−	se prendre	−	+	−	−	a	−	−	+	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−
+	+	recommencer	+	−	−	+	a	−	−	+	−	−	−	−	+	+	−	−	−	−	−	−	+	−	+
+	+	risquer	+	−	−	+	de	−	−	−	+	+	+	+	+	+	−	−	−	−	−	−	+	−	+
−	−	ne savoir	−	−	−	−	o	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
+	+	stopper	+	−	−	+	de	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	+	+	+
+	+	tarder	+	−	−	+	a	−	−	+	−	−	−	−	−	+	−	−	−	−	−	+	−	−	−
+	−	tenter	+	−	−	−	de	−	−	+	−	−	−	−	−	+	−	−	−	−	−	−	+	−	+
+	−	terminer	+	−	−	−	de	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	+	−	+
+	−	trouver	+	−	−	−	a	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
+	+	vernir	−	−	−	−	de	−	−	+	−	−	−	+	+	−	−	−	−	−	−	−	−	−	−
+	+	venir	−	+	−	−	a	−	−	+	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−

From M. Gross 1975

TABLE 2

ment along with the associated constructions with reduced infinitive predicates (M. Gross 1968, J. Rosenbaum 1967).

One also finds represented for all noun phrases the basic categories "human" and "non human", since they often bear on syntactic processes, such as nominalizations.

In the process of enumerating the constructions and meanings of verbs, we had to raise problems of representation for verbal idiomatic expressions. We first classified idioms according to their syntactic function. Thus, *to kick the bucket* is considered as a verb since it is more or less equivalent to the simple intransitive verb *to die*. We call such frozen utterances compound verbs and we extend their description to utterances such as *to take into account, to bear in mind, God recalled Nhum to Him, to pass the bucket, to diagonalize a matrix,* which can be qualified differently, namely as frozen, semi-frozen, metaphoric or technical, for example. Second, in order to enter such utterances in a lexicon-grammar, one has to describe them according to their content. Let us consider again our examples:
—*to kick the bucket* has a free human subject and a frozen object. We use for it the representation:

N_0 *V the* C_1
where $V =$: *kick*, $C_1 =$: *bucket*;

—*to diagonalize a matrix* is roughly of the same type, but the object accepts varied determiners and modifiers, as in:

You should diagonalize the matrix you just obtained
hence the notation:

N_0 *V Det* C_1

—*to take into account* and *to bear in mind* both have a free subject, a free direct objet and a frozen second complement, hence the notation:

N_0 *V* N_1 *Prep* $C_2 =$:
Bob took your advice into account
—among other types, we find:

N_0 *V* C_1 *Prep* $C_2 =$: *Bob took the bull by the horns*
C_0 *V* N_1 *Prep* $C_2 =$: *God recalled Bob to Him.*

—in the sentence:

God recalled Bob to Him

the subject and the second complement are frozen together; the direct object is free, the sentence can be passivized;

607

Tables	Structures	Examples	Numbers
C1	N_0 V C_1	Il a loupé le coche	1700
CPOS1	N_0 V Poss C_1	Il a passé son tour	600
CDET1	N_0 V Det C_1	Il agite un drapeau blanc	1000
CAN	N_0V (C à de N)$_1$	Cela a délié la langue de Max (lui)	600
CDN	N_0V (C de N)$_1$	Il bat le rappel de ses amis	500
CP1	N_0V Prép C_1	Il charrie dans les bégonias	1700
CPN	N_0V Prép (C de N)$_1$	Il abonde dans le sens de Max	300
C1PN	N_0V C_1 Prép N_2	Il a déchargé sa bile sur Max	1950
CNP2	N_0 V N_1 Prép C_2	Ils ont passé Max par les armes	1550
C1P2	N_0 V C_1 Prép C_2	Il met de l'eau dans son vin	1000
C5	Que P V Prép C_1	Que Max reste milite en sa faveur	200
C6	N_0V Qu P Prép C_2	Il a pris du bon côté que Max reste	400
C7	N_0V C_1 à ce Qu P	Il a dit non à ce que Max reste	150
C8	N_0V C_1 de ce Qu P	Il se mord les doigts de ce qu'il est resté	300
CADV	N_0V Adv	Cela ne pisse pas loin	300
CV	N_0V V W	Il est parti sans laisser d'adresse	400
CO	C_0 V W	La moutarde monte au nez de Max	1500
COE	V W	Vive le vin	1000
A1	N_0 avoir C_1	Il a eu le mot de la fin	250
A1PN	N_0 avoir C_1 Prép N_2	Il a barre sur Max	100
ANP2	N_0 avoir N_1 Prép C_2	Il a Max en horreur	100
A12	N_0 avoir C_1 Adj$_1$	Il a la vue basse	100
A1P2	N_0 avoir C_1 Prép C_2	Il a mal aux cheveux	250
EO1	C_0 de N être Adj	La barbe de Max est fleurie	450
EOP1	C_0 être Prép C_1	Les rieurs sont du côté de Max	300
		TOTAL	16,600

Structures and numbers of frozen sentences

TABLE 3

Frozen Expressions

Samples of tables C1, CPOS1, CDET1

Nhum	N-hum		N_0V	Passive		Npc	Other determiners	
+	−	CONNAITRE	+	−	LE	−	−	COUP
+	−	CONNAITRE	−	−	POSS-Ø	−	−	DOULEUR
+	−	CONNAITRE	−	+	LE	−	−	TRUC
+	−	NE CONNAITRE PAS	−	−	POSS-Ø	−	−	BONHEUR
+	−	NE CONNAITRE QUE	−	−	—	−	−	CA
+	−	CONSERVER	−	−	POSS-Ø	−	−	CHEMISE
+	−	SE CONTEMPLER	−	−	LE	+	−	NOMBRIL
+	−	COUPER	−	+	LE	−	−	CORDON OMBILICAL
+	+	DEBLOQUER	−	+	DET	−	−	SITUATION
+	−	DETENIR	+	−	LA	−	−	VERITE
−	+	DISTILLER	−	+	LE	−	+	VENIN
+	+	DOMINER	−	+	LE	−	−	LOT
+	−	DRESSER	−	+	POSS-Ø	−	+	BATTERIES
+	−	ENDOSSER	−	+	LE	−	−	HARNOIS
+	+	ENFONCER	−	+	LE	−	−	CLOU
+	−	ETRE . N PAS	−	−	UNE	−	−	LUMIERE
+	−	ETRE . N PAS	−	−	—	−	−	MANCHOT
−	+	ETRE . N PAS	−	−	LA	−	−	MORT
+	−	ETRE . S DIT	−	−	—	−	−	TOUT
+	−	FAIRE	−	−	UN	−	−	BRIN DE TOILETTE
+	−	FRIRE	−	−	—	−	−	GRISE MINE
+	−	FAIRE	−	−	—	−	−	HARA-KIRI
−	+	FAIRE	−	−	—	−	−	JURISPRUDENCE
+	−	FAIRE	−	+	UNE	−	+	MINUTE DE SILENCE
−	+	FAIRE	−	−	—	−	+	NOMBRE
+	−	FAIRE	−	+	DET	−	−	OPERATION PORTE OUVERTE
+	−	FAIRE	−	−	DU	−	−	QUARANTE CINQ FILLETTE
+	−	FAIRE	−	−	—	−	−	TAPIS
+	−	FAIRE	−	−	—	−	−	TINTIN
−	+	FERMER	−	+	POSS-Ø	−	−	PORTES
+	−	FLETRIR	−	+	DET	−	−	CRIME
+	−	FORCER	−	−	LA	−	+	CHANCE
+	−	FORMER	−	+	LE	−	−	CARRE
+	−	FORMER	−	+	DET	−	−	NUMERO
+	−	FORMER	−	+	DET	−	−	NUMERO DE TELEPHONE
+	−	FORMER	−	+	LES	−	−	RANGS
+	+	FRANCHIR	−	+	DET	−	−	CAP

TABLE 4

Table Adj-n 4

609

	suffix to remove	suffix to add	(N_0 = Nhum) être d'un Adj-n Modif	(N_0 = N-hum) être d'un Adj-n Modif	(N_0 = Qu P) être d'un Adj-n Modif	(N_0 = Nhum) avoir Dét Adj-n	(N_0 = Nhum) avoir Dét Adj-n	(N_0 = Qu P) avoir Dét Adj-n	N_0 avoir un Adj-n	N_0 avoir du Adj-n	N_0 avoir des Adj-n	N_0 avoir un Adj-n Modif
RECRUDESCENT	—	CE	−	−	−	−	+	−	−	−	−	+
REDONDANT	—	CE	−	+	+	−	+	+	+	−	−	+
RÉEL	—	ITÉ	−	+	−	−	+	−	−	+	+	+
RÉGULIER	—	ITÉ	−	+	+	+	+	−	+	−	+	+
RENTABLE	—	ITÉ	−	+	+	−	+	+	−	+	−	+
RÉPUTÉ	É	ATION	+	+	−	+	+	−	−	−	−	+
RESPECTABLE	—	ITÉ	+	+	+	+	+	+	−	−	−	+
RÉTROACTIF	—	ITÉ	−	+	−	−	++		−	−	−	+
RICHE	—	ESSE	+	+	−	+	+	−	−	−	−	+
RIGIDE	—	ITÉ	−	+	−	−	+	−	−	−	−	+
RUGUEUX	—	ITÉ	−	+	−	−	+	−	−	−	−	+
RUSTIQUE	—	ITÉ	−	+	+	−	+	+	−	−	−	+
SAIN	—	TÉ	+	−	−	+	−	−	−	+	−	+
SALE	—	TÉ	+	+	−	+	+	−	−	+	−	+
SALUBRE	—	ITÉ	−	+	−	−	+	−	−	−	−	+
SAVOUREUX	EUX	—	−	+	−	−	+	−	−	+	−	+
SENESCENT	—	CE	−	+	+	−	+	+	−	+	−	+
SEXUEL	—	ITÉ	+	−	−	+	−	−	−	−	−	+
SIMULTANÉ	É	ITÉ	−	+	−	−	+	−	−	−	−	+
SOLIDE	—	ITÉ	+	+	−	+	+	−	−	−	−	+
SOLUBLE	—	ITÉ	−	+	−	−	+	−	−	−	−	+
SOLVABLE	—	ITÉ	+	+	−	+	−	−	−	−	−	+
SOMPTUEUX	—	ITÉ	−	+	+	+	+	−	−	+	−	+
SOUDAIN	—	TÉ	−	+	−	−	+	−	−	+	−	+
SOUPLE	—	ESSE	+	+	−	+	+	−	−	+	−	+

(From A. Meunier 1977)

TABLE 5

The present stage of the classification of compound verbs is given in table 3. About 20,000 compound verbs have been compiled and described in about 30 classes. Each of these classes receives a syntactic description in the form of tables, exactly as for simple verbs, we give a sample in table 4.

The representation which has been given to nouns is similar to that of verbs. It is based on the following observation made by Z.S. Harris. Consider the pair:

Bob walked in the park
=*Bob took a walk in the park*

Harris sees it as a relation of nominalization, that is, as a transformation operating between two sentences. The verb *to take* has been introduced as a grammatical constant, we call it a support verb. Support verbs are limited, the main ones being *to be, to have,* as for example in the pairs:

This text contradicts your statement
=*This text is in contradiction with your statement*

Bob dreams
=*Bob has a dream*

But the crucial point here is that support verbs occur independently of nominalization relations, as in:

Bob is in charge of the report
Bob had a nightmare

where the nouns *charge* and *nightmare* do not correspond to verbs.

Large classes of nouns (so far about 10,000) can then be described as entries of the lexicon-grammar of French. Since they occur in sentence structures, they can be described by the same type of matrices that were used for verbs (tables 2 and 4), but the entry appearing in a row is the noun instead of the verb (table 5).

Support verbs have a syntactic behavior different from that of other simple verbs. An intuitive difference between ordinary verbs and support verbs can be stated in terms of noun distributions:

—with ordinary verbs, the nouns which can be attached in subject and object positions are semantically predictable, often they receive intentional definitions, that is an intuitive general paraphrase. As a consequence, when two nouns are approximate synonyms (or antonyms), if one of them belongs to the distribution of an ordinary verb, the other will too. For example, if we consider the direct object of *to read,* one can immediately pre-

dict that all "nouns of texts" will be accepted in its distribution: if one accepts as a normal activity *to read a dictionary*, one will also accept *to read an encyclopedia* as a normal activity. Notice that if we set aside the following distributions: (i) human nouns, (ii) unrestricted positions, (iii) adverbs, (iv) abstract nouns, then the situation described for *to read* is not so often encountered in the lexicon;

—with support verbs, supported nouns are not semantically predictable. Whereas this is easily understandable with support verbs such as *to be, to have* and *to make*, there are many support verbs harder to detect because they are identical to ordinary verbs but they lack this distributional property. Consider the objects of *to sign* in the construction:

Max signed an agreement with Bob

In the position of *agreement*, we find nouns such as *treaty, pact, contract,* but *conflict,* an antonym, and *tacit deal,* a reasonable synonym of *agreement* are not accepted with *to sign:*

Max has an agreement with Bob
Max has a tacit deal with Bob
**Max signed a (conflict + tacit deal) with Bob*

This is one reason among others that leads to the classification of *to sign* as a support verb.

In the same way that one can say ordinary verbs semantically select their nouns, one can consider that nouns select support verbs. From the various examples studied, it appears that the supports, whatever their lexical and syntactic form, are limited to certain aspectual and modal notions.

Remark

Supported nouns can select nouns as ordinary verbs do. This is obvious in the case of nominalizations of ordinary verbs, when the resulting noun in general retains the distributions of the verb. Hence in a support (or operator) structure, we can find both types of noun-verb relations: support and selectional.

Adverbs can be entered in a lexicon-grammar in a way similar to that of nouns. There is a special set of support verbs for them, such as *to occur, to take place,* in:

The accident took place (in the park + at four)
Such explosions occur (frequently + from time to time)

About 5.000 such adverbs have been classified in a way they can be entered in the lexicon-grammar.

The coverage of French entries is far from complete, numerous com-

pound words remain to be detected, before they can be incorporated into a lexicon-grammar. Estimates for compound nouns range over 200,000 for the style of language which has been adopted for simple words (cf. 1), and the description of compound verbs, adverbs and adjectives is not complete either.

Besides these empirical questions, they are theoretical points left open, we just mention a few obvious ones:

—What is the status of "concrete" nouns (e.g. *girl, grocer, pineapple, arm chair*) in a lexicon-grammar? While support verbs are relatively easy to determine for "abstract" nouns, the question is open for most "concrete" nouns;

—How are combined supported nouns with ordinary verbs? Namely, what is the analysis of sentences such as:

> *The contradiction between this text and your statement surprises me*
> *We should overcome the contradiction between this text and your statement*

in terms of the entries of the lexicon-grammar?

—How are introduced adverbs into a sentence? Namely what is the decomposition into elementary sentences of:

> *Such pipes explode (frequently + from time to time)*

Whereas this question has been raised in other theoretical contexts, Z. S. Harris hypotheses on support verbs constrain the range of possible analyses.

Although these questions can be seen as limitations, we think that the framework of lexicon-grammar theory is a promising way of describing languages in a realistic way. Confirmation of this point of view comes independently from the descriptions of other languages, some close to French: English (P. Freckleton 1985, M. Salkoff 1983), Italian (A. Elia, M. Martinelli, E. d'Agostino 1961), Portuguese (E. Macedo 1981, E. Ranchhod 1983), and Spanish (B. Lamiroy 1983, C. Subirats 1987), others quite distant: Arabic (M. Chad 1981), Korean (Hong Chai-song 1985, Han Jung-kill 1981, Pak Hyong-ik 1987), Malagasy (R. -B. Rabenilaina 1985). Also, the level of the description allows a potential for applications of lexicon-grammar to fields as different as language pedagogy and computer science.

Notes

* UA 819 of the CNRS. This study has been partly financed by the contract: "Programme de Recherches Coordonnées en Informatique Linguistique" of the Ministry of Research and Higher Education.

1. Later on, once the description of verbs had progressed, the list could be

augmented, by using arguments of analogy, which make predictions about words whose functioning is not entirely mastered by the linguists.

2. This is also true for (2) to a lesser extent.
3. The amounts of texts that would be necessary are not yet available, nor is automatic lemmatization.
4. We do not give yet any technical meaning to the notion essential complement. Our position is that the corresponding traditional problem of separating objects from circumstantial complements is not fully solved (Cf. M. Gross 1988; A. Guillet, C. Leclère, J.-P. Boons 1988 for proposals on locative complements).

REFERENCES

d'Agostino, Emilio. 1983. *Lessico e sintassi delle costruzioni locative,* Naples: Liguori, 182p.

Boons, Jean-Paul, Alain Guillet; Christian Leclère. 1976. *La structure des phrases simples en français. Constructions intransitives,* Genève: Droz, 377p.

Chomsky, Noam. 1957. *Syntactic Structures,* The Hague: Mouton.

Chad, Mohamed. 1981. *Les constructions intransitives de l'arabe classique,* Paris: LADL, Université Paris 7, Doctoral thesis, 250p.

Elia, Annibale. 1979. *Lexique des constructions complétives en italien. Verbes à un complément,* Paris: LADL, Université Paris 7, Doctoral thesis, 1984. *Le verbe italien. Les complétives dans les phrases à un complément,* Bari: Schena-Nizet, 303p.

Elia, Annibale, Maurizio Martinelli, Emilio d'Agostino 1981. *Lessico e structure sintattiche. Introduzione alla sintassi del verbo Italiano,* Naples: Liguori, 430p.

Freckleton, Peter. 1985. *Une comparaison des expressions de l'anglais et du français,* Paris: Université Paris 7, LADL, Doctoral thesis.

Gross, Maurice 1968. *Grammaire transformationnelle du français.* **1.** *Syntaxe du verbe,* Paris: Cantilène, 185p., reprinted. 1986.

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 1988. *Grammaire transformationnelle du français.* **3.** *Syntaxe de l'adverbe.* Paris: Cantilène.

Guillet, Alain, Christian Leclère, Jean-Paul Boons 1988. *Lexique-grammaire des constructions transitives à complèment locatif,* to appear.

Han, Jung-kill 1981. *Analyse des constructions causatives en coréen contemporain,* Paris: Université Paris 7, LADL, Doctoral thesis.

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

Harris, Zellig 1984. *A Grammar of English on Mathematical Principles,* New York: Wiley-Interscience, 429p.

Hong, Chai-song. 1985. *Syntaxe des verbes de mouvement en coréen contemporain,* Amsterdam: J. Benjamins B.V., 324p.

Lamiroy, Béatrice, 1983. *Les verbes de mouvement en français et en espagnol. Etude de syntaxe comparée de leurs infinitives.* Amsterdam: J. Benjamins B.V.,

337p.

Macedo Oliveira, Maria Elisa de. 1979. *Syntaxe des verbes psychologiques du portugais,* Paris: Université Paris 7, LADL, Doctoral thesis, 1984. In *Textos de linguistica* 7, Lisbonne: INIC, Centre de linguistique de l'Université de Lisbonne, 198p.

Meunier, Annie 1977. Sur les bases syntaxiques de la morphologie dérivationnelle, *Lingvisticae Investigationes* 1:2. Amsterdam: J. Benjamins B.V., pp. 287–332.

Pak, Hyong-ik 1987. Verbes datifs en coréen, *Lingvisticae Investigationes* Amsterdam: J. Benjamins B.V.

Ranchhod, Elisabete 1983. On the Support Verbs *ser* and *estar* in Portuguese, *Linguisticae Investigationes* VII:2. Ansterdam: J. Benjamins B.V., pp. 317–353.

Rabenilaina, Roger-Bruno 1985. *Lexique-grammaire du malgache. Constructions transitives et intransitives,* Paris: University Paris 7, LADL, Doctoral thesis

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

Salkoff, Morris 1983. Bees are swarming in the garden: A systematic synchronic study of productivity, *Language* 59 No. 2. Baltimore: The Waverly Press, pp. 288–346.

Subirats-Rüggeberg, Carlos 1987. *Sentential Complementation in Spanish,* Amsterdam: J. Benjamins B.V., 291p.

Tesnières, Lucien 1959. *Eléments de syntaxe structurale,* Paris: Klincksieck, 671p.