

ACA.DOC 910201

Linguistic Representations and Text Analysis

Maurice Gross_

Understanding a text, whether by a human being or by a computer, implies that units of meanings be identified in the text and that rules composing these units and the corresponding meaning units provide the complete meaning of the text. Such a statement raises many fundamental questions we shall not be concerned with (e.g. What is meaning?). We will limit ourselves to lexical and grammatical procedures that lead to the recognition of patterns of words on which the process of understanding is based.

First, we will illustrate the patterns of words to be detected by analyzing a short text in English (figure 1_). Already a large variety of grammatical combinations of words will be encountered. Then, we will discuss the implications of these observations for the construction of an explicit system of understanding. We will illustrate the shape and the size of this system mostly through data obtained for the French language.

Figure 1

1. Lexical analysis of simple words

Among the most obvious units of meaning are the simple words, they are defined as sequences of characters limited by consecutive spaces.

However, attributing meaning to a simple word runs into two fundamental problems:

- in many instances simple words are ambiguous, that is, they have several meanings, as recorded in ordinary dictionaries,
- often, they have no meaning at all by themselves, either because they are grammatical words used to combine words (e.g. of, and, to be, etc.) or else because they are part of compound words which carry meanings only as a whole (e.g. the idiom red herring, and the noun tunnel in the technical term of quantum physics tunnel effect).

These two situations do not exclude each other, and they both require analysis of the context of the individual words whose interpretation is sought. For example, the word show can be either a noun or a verb, and each of these two grammatical forms has several meanings. The text of figure 1 contains such meanings:

- in the sequences for showing that, showed that by using, the verb has a basic construction which authorizes sentential complements, as in the normalized form:

N0 V N1 =: The authors showed that their solution was coherent

The meaning of this construction_, roughly that of to prove, must be distinguished from the meaning found in the sentence:

N0 V N1 to N2 =: The authors showed their book to Max

roughly that of to exhibit, but which has a different structure: two complements N1 and to N2, instead of one and where the direct complement is 'concrete';

- in the sequences:

crystals that show icosahedral symmetry
which shows an overall five-fold symmetry

The verb has practically no meaning, it has only a grammatical function that we call support verb (Z.S. Harris 1964, M. Gross 1981). It is approximately synonymous with to have, a more general support verb;

- and a priori, there are other meanings, as illustrated by the following examples:

The authors showed us into the conference room
Max is showing off
Results are showing up

We are now in a position to define more precisely the problem of the formal analysis of a text. Texts are available as sequences of simple words (defined themselves as sequences of characters on a given alphabet). Simple words are described in dictionaries. An automaton acting very much as a beginner student of Greek or Latin, consults a dictionary which provides a whole range of solutions for the interpretation of a given word. In the example of the word show, we have already listed 7 interpretations. In principle, texts are not ambiguous, at least with respect to these interpretations. Hence, 6 irrelevant interpretations have to be eliminated, which can only be done by exploration of the context of the word.

We just presented one class of problems of the analysis of the text, namely the recognition of the lexical units of a texts. Another question consists in providing the organization of these words into sentences, a problem to be generalized to the organization of the texts into autonomous discourses. We will now describe an example of syntactic analysis, clearly distinguished from lexical analysis by the fact that the grammar rules involved are largely independent of the words to which they apply. We will then discuss a more elaborate type of lexical analysis, and we will then see that many rules apply only to interdependent lists of words, revealing the complex structure of the lexicon of a language.

2. Syntactic analysis

Consider the sentence (extracted from our text):

(1) The addition ... can be made to happen quickly and uniquely and in a way that is consistent with Penrose tiling

The rules of English grammar define the sequence:

(1a) The addition can be made to happen

as a well-formed sentence containing the main verb. It is clear that (1a) is the grammatical 'backbone' of (1), it has a structure noted $S =: N0 V$ (i.e. subject-verb, more precisely: subject-verbal complex). Other rules state that the adjunction of an adverb (noted Adv) to such a sentence results in a sentence. We can write the equation:

(R1) $S =: S \text{ Adv}$ and apply it to our example: $(1) = (1a) \text{ Adv}$

This (recursive_) notation reflects the fact that any number of adverbs can be added to a sentence S. But adverbs can also be built by means of the conjunction and, hence the rule (i.e. equation):

(R2) $\text{Adv} =: \text{Adv and Adv}$

and it is clear that this rule applies twice in sentence (1), yielding the following analysis where phrases are delimited by parentheses marked with the grammatical symbols of the rules:

(1S) (The addition ... can be made to happen)S (((quickly)Adv and (uniquely)Adv and (in a way that is consistent with Penrose tiling)Adv)Adv

The deepest level of parentheses, for example those attached to quickly and uniquely, is the result of a dictionary look-up for these words_. The other levels are obtained by the application of the grammar rules (R1) and (R2), which here indicate the way a complex adverb is constituted from simpler adverbial shapes.

There are many other rules in the grammar, corresponding to the many other sentence shapes. Among others, we will have:

(R3) $S \text{ and } S$

a rule stating that a sentence can be formed by conjoining two other sentences. As we are going to see, this particular rule has consequences for the analysis of (1). Let us now mimic a mechanical process of analysis for (1). To do so, we scan (1) from left to right. By definition of the problem, we know where the beginning of the sentence is (it is marked by a period, followed by a space, followed by a capital letter). Now, in order to locate the end of the sentence, let us attempt to define precisely the whole adverb (i.e. the outer level of adverb parentheses in (1S)).

We have analyzed (1a) intuitively as a sentence, the application of rule (R1) forces us to do the same for the following two other subsequences of (1):

- (1b) The addition ... can be made to happen quickly
- (1c) The addition ... can be made to happen quickly and uniquely
- (1b) has the shape S Adv, that is, (1a) quickly
- (1c) has the shape S Adv, that is, (1a) quickly and uniquely

We can paraphrase this analysis in the following way: we intend to analyze (1) as a full sentence S. Our grammar is composed of the three rules (R1), (R2) and (R3), a priori this grammar proposes two global competing structures for a sentence such as (1): S Adv and S and S. We already analyzed the adverbial structure (S Adv) in (1S), we now have to delimit all the sequences that are determined as Ss by the grammar in order to check for the possible presence of the structure S and S. The beginning of the sentence is the left-most word The, and an end for S is a priori possible after quickly, or after uniquely, or at the period.

(i) Consider the hypothesis 'end of S after quickly'. In order to be validated, it must be followed by the structure and S, then (1) would have the global form:

(1b) and S

But when we examine the rest of the sentence:

(2) and uniquely and in a way that is consistent with Penrose tiling

we verify that this sequence of words is not a sequence and S, hence the hypothesis must be rejected.

(ii) Let us finally consider the hypothesis 'end of S after uniquely', the rest of (1) is:

(3) and in a way that is consistent with Penrose tiling

this sequence has been analyzed in (1S) as a conjoined adverbial complement of the form:

(and in a way that S)Adv

However, if we examine the sequence (3) more closely, we do find another possibility of analysis, with a sentential rest of the form and S:

- the sequence in a way is by itself an adverb, as in: in a way, Bob is wrong,
- the word that is a subordinating conjunction in the previous analysis, but the dictionary also tells us that it can be a pronoun_, similar to this.

Let us now combine these two possibilities to produce the following variant of sentence (1):

(4) The addition ... can be made to happen quickly and uniquely, and in a way, that (= this) is consistent with Penrose tiling

(4S) ((The addition ... can be made to happen)S ((quickly)Adv and (uniquely)Adv)S

and

((in a way)Adv, (that (= this) is consistent with Penrose tiling)Adv)Adv)S)S

To reach this analysis, the only modifications we made_ are the two commas delimiting and in a way. These commas induces a substantial change in intonation and in meaning for sentence (1). But the use of commas in mechanical syntactic analysis is far from reliable, as a consequence, the analysis we have just arrived at forces us to consider that our initial sentence (1) is twice ambiguous, with the second reading (4).

—

3. Lexical analysis of complex forms

We mentioned the existence of complex sequences of words which function as simple words. In general, they can be tagged by the usual names of parts of speech; examples are:

- the compound nouns red herring and tunnel effect, already mentioned,
- from time to time, now and then which are complex, compound, frozen or idiomatic adverbs, there is no fixed terminology for qualifying such constructs,
- as soon as, inasmuch as, are complex conjunctions, etc.

The intuition lying behind the notion of complex words can be termed semantic noncompositionality, in other words, the meaning of the sequence cannot be obtained by composing the meanings of the component words. This notion is also relevant to adjectives and verbs:

- solid blue and well to do are complex adjectives,
- to take the bull by the horns and Bob's dream came true are complex verbs (or equivalently, complex elementary sentences).

Practically all of the examples we gave were idiomatic, hence their semantic noncompositionality was fairly obvious. There are however many examples where it may not seem so. A compound such as cruise missile has the meaning of missile, however, the word cruise cannot by itself indicate the supplement of meaning which corresponds to the special guiding system of this type of missile. Most complex technical terms are made of simple words that evoke parts of the meaning of the whole, but the complete definition lies outside the range of meaning of each word. The text we examine has for its main theme such a term: perfect Penrose tiling structure, whose meaning is given by a mathematical definition which cannot be deduced from the words.

Such complex terms are quite numerous in languages that handle science and technology. A new technical problem is associated with them. Today, practically all texts (books, newspaper, journals, commercial mail, etc.) are produced by means of computers. Hence, in principle, archives can now be stored in computer form. Computer programmes could search the texts of such archives for specific information. But information given in a linguistic form, that is in terms of words,

always presents the difficulties of interpretation discussed for words: ambiguity and compositionality.

—

Let us return to our text, and study the occurrences of the technical term perfect Penrose tiling structure. We observe the following occurrences:

perfect	Penrose	tiling structure
perfect	Penrose	tiling
	Penrose	tiling
	Penrose	structure

namely, we observe the full name and variable abbreviations. One can safely predict that the following forms will also occur in texts dealing with the same theme:

	Penrose	tiling structure
		tiling structure
perfect		tiling structure

There is however a difficulty in drawing such a list: whereas it is clear that the list of terms found in the text refer to one given object, this is less clear with the last three constructions; in fact, the use of a set of abbreviations is determined by a stylistic choice that may vary with the subject of each paper and within a given domain of knowledge. More generally, given a long term as in our example, two types of abbreviations have to be distinguished:

- a set of institutional abbreviations, that is, short forms used instead of the long form by the community of specialists of the domain,
- short forms used by individual authors in specific papers; there, forms may be different from the consensual abbreviations, a problem similar to the search of an antecedent for a pronoun.

As a consequence, different treatments apply to both situations:

- institutional abbreviations are listed a priori, that is, they are recorded in a dictionary,
- other abbreviations are to be detected during the analysis of a particular text.

The graph of figure 2 is a dictionary entry structured in order to make explicit the equivalence of the possible forms. The formalism of finite automata has been applied to it (M. Gross, D. Perrin 1989). More exactly, figure 2 is a directed acyclic graph that reads as follows: the nodes of the graph are called states, the leftmost state is the initial state, circled states are final states. Arrows are labelled by simple words, the empty (zero) word is noted E. An utterance is characterized by a path between an

initial and a final state.

	E			E
I	perfect	Penrose	tiling	structure F

Representation of families of strings by finite automata

This automaton represents the four strings found in the text and the string Penrose tiling structure in addition. The symbol E represents the null string.

Figure 2

The representation by finite automata of families of strings that are semantically equivalent is well adapted to noun phrases, and particularly to phrases corresponding to concrete or technical notions. It could also be used to represent families of strings belonging to other grammatical categories. We list in figure 3 complex units found in the text:

–

COMPLEX LEXICAL UNITS

Complex nouns of both shapes:

Adj N:

physical sciences
local rule
icosahedral symmetry
experimental study
growing cluster

N Prep N:

laws of crystallography

solution to a problem

Other shapes of noun phrases:

five-fold symmetry
local rules of interaction

Complex adverbs:

in part
one by one
that is,
in a way

Complex adjectives: three-dimensional

Figure 3

The formal variations of noun phrases representing technical terms are limited. We discussed their abbreviations, but other variations are possible for such terms:

- morphological variations (singular, plural, case),
- adjunctions of determiners (definite or indefinite articles, quantifiers, etc.),
- adjunctions of modifiers (adjectives, noun complements, relative clauses, etc.).

Adjunctions can only occur to the left or to the right of the sequence of words representing the term. This is not the case for sentences which can vary greatly in shape and which can be combined in quite complex ways.

4. Sentential components

4.1 Frozen sentences

We will also consider the following complex verbs (i.e. frozen sentences written in their normal form) and we will return to the problem of retrieving their information content from the text:

N0 receive an award
N0 fill (space) nonperiodically
N0 grow (a cluster + a tiling)
There exist N1

4.2 Government

A text contains many other elements that contribute to its meaning. We proceed to analyze our text in order to show that these other elements must also be represented, but in a different way, namely in terms of sentences, not of phrases, as

already suggested by the preceding frozen sentences.

As a first general step, we will pay attention to the grammatical phenomenon called government, that is to situations where a word belonging to one of the four major categories, Noun, Verb, Adjective, Adverb determines the use of a grammatical word (Preposition or Conjunction) which in turn introduces some complement.

We list in figure 4 the combinations found in the text.

—

GOVERNMENT

of Prepositions and Conjunctions by the four major categories.

Verbs:

showing that	N0 show that S
showed that	
be grown by	N0 grow N1
been taken as	N0 take N1 as N2
believed that	N0 believe that S
adding ...to	N0 add N1 to N2

Nouns:

award for	N0 be an award for N1
a solution to	N0 be a solution to N1
the problem of whether	Whether S or S is a problem
a model for	N0 is a model for N1
the connection between	There is a connection between N1 and N2
the addition of ... to	N0 make the addition of N1 to N2
the link between	There be a link between N1 and N2

Adjectives:

consistent with	N0 be consistent with N1
-----------------	--------------------------

Adverbs:

in contradiction with	N0 be in contradiction with N1
according to	N0 (occur + happen) according to N1
in a way that	N0 (occur + happen) in a way that S

In the left part of the table, we have the 'binary' combinations, and in the right part, a corresponding elementary sentence shape in a normal form which, in a minimal way, makes explicit the meaning of the relations determined by government. For nouns, we give a full sentence with a support verb.

Figure 4

The notion government can be extended to combinations of verbs, as in the following examples found in the text:

- can fill
- could not be grown
- came to be
- can be made to happen

4.3 Transformations

Z.S. Harris 1952 proposed a model for describing sentence variations, a model based on the notion of transformation. Transformations between sentences are equivalence relations that leave invariant the basic meaning of the sentence: rules such as [Passive], introduction of Modals and Negation are transformations written as in the following examples, again taken from the text:

- N0 grew a Penrose tiling
- [Passive] = A Penrose tiling was grown
- [Modal i.] = A Penrose tiling could be grown
- [Negation i.] = A Penrose tiling could not be grown

- N0 believed (that S)1
- [Passive] = (That S)1 was believed
- [Extrapolation] = It was believed (that S)1

The transformation of Relativization combines two sentences into (1):

- N0 adds N1 to a cluster. This cluster is growing
- = N0 adds N1 to a cluster that is growing
- = N0 adds N1 to a growing cluster

It relates the elementary sentence N0 grow a cluster to the noun phrase a growing cluster.

4.4 Sentences with support verbs.

Transformations with support verbs introduce an equivalence relation called nominalization (and noted [Nomin]) between sentences constructed with a noun and sentences built around a verb, as in:

- (1) N0 (relates + links) N1 (and + with) N2
- [Nomin] = N0 (makes + establishes) a (relation + link) of N1 with N2
- [Sym] = N0 (makes + establishes) a (relation + link) between N1 and N2

(1) [Passive] = N1 is (related + linked) with N2
[Sym] = There is a (relation + link) between N1 and N2

N0 added N1 to N2
[Nomin] = N0 made the addition of N1 to N2
= There was an addition of N1 to N2
= An addition of N1 to N2 happened
[Causative i.] = N0 made to happen an addition of N1 to N2
[Passive] = An addition of N1 to N2 was made to happen

N0 contradicts N1
[Nomin] = N0 (is + enters) in contradiction with N1
[Sym] = There is a contradiction between N0 and N1

The possibility for a verb or a noun to enter into a given syntactic form, that is to undergo a transformation, cannot be predicted from its meaning or from other properties. For example, the nouns relation, link and contradiction are observed in the same symmetrical construction (noted [Sym]), but link, contrary to these two other nouns, is not accepted in the construction with support verbs (be + enter) in:

*N1 (is + enters) in link with N2

This restriction is hard to attribute to the fact that to link does not have the transitive construction of to contradict:

N1 (connects + links) N2

In fact we can observe a transformation such as:

N0 communicates with N1
[Nomin] = N0 (is + enters) in communication with N1

which does not apply to an identical structure containing to relate:

N1 relates to N2

This situation is quite general, we will present in 5 the solution adopted to represent such lexical dependencies.

The sample of sentential phenomena and descriptions we have given was arbitrary in the sense that they were observed in a randomly selected and rather short text. However, the broad types of facts we have collected are quite general and they occur in most texts. Adding new texts would provide many new particular situations attached to specific lexical items but no basically new phenomena; however, the need to classify these detailed facts would become urgent.

We have performed such a classification for French and we have discovered that a rather small number of classificational features were powerful enough to accommodate a large number of lexical items, which at first sight seemed to enter

into an endless variety of structures. In order to reach such a stage, it is essential to distinguish two types of linguistic elements:

- (i) terms in the form of noun phrases: a technical term such as perfect Penrose tiling structure is a typical example. Such forms are to be described in dictionaries, possibly dictionaries of automata,
- (ii) words or compounds that must be described within sentences. They are described in lexicon-grammars.

These two components of a language are not independent. In fact, an electronic lexicon of simple words (called DELAS) has been constructed (B. Courtois 1990), it contains about 80.000 entries, which can be automatically inflected_ and used as keys to enter from texts into the lexicon-grammar and into the dictionaries of compound words. The main dictionaries of compound words which have been constructed so far for French include:

- compound nouns (G. Gross 1988, M. Silberztein, 1989; cf. figure 5),
- compound adverbs (M. Gross, 1990) which can also be described with their supporting verb, (cf. figure 6),
- compound conjunctions (M. Piot, 1978).

This sample of compound nouns illustrates their representation in an electronic dictionary. The shape of the nouns of this class is: N à (Det) N, that is, a noun N, followed by the preposition à, possibly by a determiner Det and a second noun N. The signs '+' and '-' indicate authorized variations: feminine and plural. Numerical codes (e.g. N1, N21) are inflection codes describing the endings corresponding to the se variations. An article and marks of gender and number (e.g. ms for masculine singular) are attached to each compound noun.

Figure 5

Compound adverbs are described according to their syntactic shape. The above sample (table PCDN) correponds to structures Prep (Det) C de N, where the first noun C is frozen, it is followed by the preposition de and by a free noun phrase N. an example is: Bob a agi sur la recommandation de Guy, (Bob acted on Guy's recommendation). In each column, syntactic properties appear. For example, the '+' sign in the leftmost column: N =: de V0 W indicates that an infinite complement is accepted: Bob a agi sur la recommandation de faire cela.

Figure 6

5. Lexicon-grammars

The theory of lexicon-grammar is founded on the following axiom:

The linguistic unit of meaning is the sentence.

As a consequence of this axiom, words are not units of meaning, a statement that needs to be justified:

- that simple words are not units of meaning is obvious for compound words. Since by definition compound words have no compositional meaning, the simple words used to form them cannot be said to carry meaning. It turns out that compound nouns are much more numerous than simple nouns in the lexicon of any language. The technical vocabulary (up to several millions of terms) is constituted of compound nouns;
- frozen (e.g. idiomatic) sentences also are more numerous than ordinary sentences, the simple words that constitute them cannot be said to have a meaning of their own.

These quantitative observations have been confirmed during the study of French, Italian, Spanish, English and Portuguese.

That sentences are elementary units of meaning is clear in the case of verbs: verbs cannot be considered without their subject and possible objects_. The same is true for to be Adjective forms, and also for predicative nouns and adverbs, although in a less obvious way (cf. figure 4). Converging observations led to this theoretical position:

- more syntactic properties of sentences than usually thought depend on the main verb. For example, determiners which are mostly represented as locally constrained by their noun are often selected by the verb to which their noun is attached:

Bob wants some beer
*Bob loves some beer

Bob hunts the wildgoose
*Bob hunts a wildgoose

The following sentence presents a constraint of number between its subject and its adverbial complement:

The soldier crossed the river during one hour

it has the interpretation of multiple crossings by the subject in the singular, whereas the sentence with a plural subject:

The soldiers crossed the river during one hour

Modern transformational grammar has systematized this approach. Sentences are related when they share an invariant of meaning which can roughly be seen as their lexical content. By this token, the Passive form is related to the Active (i.e. declarative) one by the pairing of the two structures, that is, the rule or relation:

- (1) N0 V N1 =: Bob criticized the report
 (2) N1 be V-ed by N0 =: The report was criticized by Bob

The invariant of meaning or lexical content is here the triple of words {Bob, criticize, report}. The Active-Passive relation is a synonymy relation, but other types of sentences share the same invariant and are thus related to (1) and (2):

- (3) He criticized it
 (4) The report cannot be criticized by him
 (5) It is uncriticizable

Transformational relations are equivalence relations, they define equivalence classes of sentences (e.g. (1)-(5)). Moreover, the relations between sentences are stated in combinatorial terms, that is by rules of transformations which all have the following formal features, which include:

- permutations of noun phrases (e.g. Passivization),
- operations of deletion and insertion of words involving mostly grammatical words, namely a fixed set of words such as: Prepositions, Conjunctions, Support verbs (e.g. to be, to have), etc.

Hence, the Passive relation involves permutation of the two noun phrases N0 and N1 and insertion of the preposition by and of the support verb to be governing the verbal suffix -ed. Pronominalization relations such as:

Ni = he, Prep Ni = him, Ni = it, Ni = who, etc.,

and contraction rules such as:

by him = 'zero', can be V-ed = V-able, not V-suffix = un-V-suffix

are combinatorial operations too, they are entirely explicit and do not involve particular intuitions of meaning in order to be applied, that is intuitions of meaning that could be difficult to attach to the words of the lexicon of the language.

More precisely, the study of elementary sentences of French (i.e. subject-verb-objects) has shown that they all enter into one of the three general structures:

N0 V
 N0 V N1
 N0 V N1 N2

where N0 is the subject, N1 and N2 are two possible object (or essential) complements which may be preceded by a preposition noted Prep. In French, the main prepositions are "zero" à and de. A limitation to two complements has been

observed in the course of a detailed study of about 12 000 verbs or elementary structures_. Complex sentences can be described as obtained from two simple ones by rules of compositions called binary transformations (Z.S. Harris 1952), as in the example discussed in 2.

A classification of these 12 000 elementary structures has been constructed (J.-P. Boons, A. Guillet, C. Leclère 1976a, 1976b, ; M. Gross 1975 ; A. Guillet, C. Leclère, 1991). It is based on the following features:

- the nature of the prepositions of N1 and N2: "zero", à, de, and a few others;
- the content of the Nis:
 - nominal (ordinary nouns),
 - sentential, that is, of the form que S or not (without excluding nouns),
 - frozen, namely constituting a compound with the verb: C0 V W or V Ci.

Such characters define syntactic tables (about 100) in which are represented the equivalence classes defined by the transformational rules or relations (see tables below). Hence, each table has a structural definition, namely, all of its verbs enter into one of the syntactic forms we have defined, a declarative form. Columns in the tables (over 500) correspond to equivalent syntactic forms. A verb given in a row may enter (it then has a '+' mark) or not (it has a '-' mark) into the forms represented in the columns. As a consequence, the '+' marks in a row define the content of the equivalence class of the verbal entry.

We already discussed elementary structures found in the text of figure 1. Further examples of the link between the text and the syntactic tables are:

N0 receive N1 from N2, N0 =: Georges V Onoda and David P. Di Vincenzo,
N1 =: Outstanding Innovation awards

but in the text, the complement from N2 is elliptical (N2 is a division of IBM), in a column, the property N0 V N1 allows the possibility of indicating the absence of Prep N2 by marking it "+". In the same sentence, the adverbial complement for showing ... is not considered as essential, it is analyzed in terms of the declarative form:

N0 show N1, N1 =: that S

The elliptical subject N0 of showing is the same as the subject of the sentence N0 receive N1. The sentence S embedded in the complement N1 of showing is the passive form of:

(purely local rules)0 can grow (a perfect Penrose tiling)1

Government phenomena in sentences, as discussed in 3, are thus described in syntactic tables. The table in figure 7 is extracted from a description of verbs with sentential complements.

The class 6 corresponds to structures N0 V (Qu S)1, that is verbs with one sentential complement. The key element of the table is a simple verb.

Figure 7

Similar elementary structures containing a frozen element (cf. figure 3) are described in the table of figure 8 (C6).

The global structure of the entries of the table C6 is N0 V (Qu S)1 Prep C2. Such a table has two entries the verb V and the noun C2 of the complement.

Figure 8

Nouns are described in a similar way: by means of sentences with a support verb. The entries of tables are nouns, associated to a sentence with a minimal support verb: to be, to have, to get, to put, etc. (cf. figure 4). Columns are the same as for the other two types of elementary sentences.

There is however an important difference of structure for these tables: Support verbs govern specific prepositions, namely complements of varied forms, in the same way as ordinary verbs do. But now, given a minimal support verb and its supported noun, one often observes that a set of equivalent support verbs can be substituted for the minimal verb, keeping the meaning invariant. As a consequence, each equivalent support verb must be described syntactically by means of specific columns. Hence, to each column containing an equivalent support verb, one must attach columns describing the properties of its constructions, namely, a syntactic table representing its transformations. Let us consider the example of the class of symmetrical nouns (Nsym) that are defined by the following structures:

N0 have Nsym with N1 = N0 and N1 have Nsym
=: Bob has an agreement with Jo = Bob and Jo have an agreement

Some equivalent support verbs are: There is, to be in and to sign:

There is Nsym between N0 and N1 =: There is an agreement between Bob and Jo
N0 be in Nsym with N1 =: Bob is in complete agreement with Jo
N0 sign Nsym with N1 =: Bob signed an agreement with Jo

It is clear that each of these structures has specific properties (i.e. columns) which must be attached to each equivalent support verb. For example, to sign is the only verb here that has a Passive form:

An agreement has been signed by Bob with Jo

Hence, to each support structure one must attach a set of properties that practically constitutes an autonomous syntactic table.

6. Results and conclusions

Three groups of sentences (ordinary, frozen and those with support verbs) can be clearly distinguished in European languages, we have based their presentation on our experience on French, but we have verified that this classification is more general. This organization has been applied to other languages than French, samples of similar syntactic tables are given in the annex for English (M. Salkoff 1983), Italian (A. Elia 1984) and Spanish (C. Subirats 1986).

This formalization of the descriptions, including their practical presentation, led to a number of results of a quantitative nature:

1. The systematic description of French verbs (simple sentences) has shown that no two verbs have the same set of syntactic properties, as a consequence, verbs have to be described individually and not in terms of intensional classes.
2. The proportion in the lexicon of idiomatic sentences, of metaphoric and technical sentences that have non compositional meanings, is very high. All these sentences or sentence types have anecdotal origins. The consequence is that they must be described individually, that is without reference to other classes of lexical combinations or of interpretation rules.
3. The large number of verb-complement combinations that cannot be qualified in terms of semantic (i.e. selectional) restrictions leads to the notion of support verb. Their variety also implies detailed individual descriptions of nouns.

This method of construction of equivalence classes for elementary sentences could be applied today to a whole language, leading to a coverage of structures so complete that computer analysis of syntactic forms would become possible for texts. Texts would then be reduced to sets of elementary units of meaning (Z.S. Harris 1982), allowing the tremendous variety of the expression of information to be reduced to a more tractable number of standardized forms.

ANNEX

Samples of syntactic tables:

- for English (M. Salkoff 1983, figure 9), the table is defined by verbs entering into the

two related forms:

Bees are swarming in the garden = The garden is swarming with bees

Figure 9

- for Italian (A. Elia 1984, figure 10) and for Spanish (C. Subirats 1986, figure 11), the structures are the same as the defining structure of the French table 6 (cf. figure 7).

Figure 10

Figure 11

—

REFERENCES

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

Boons, Jean-Paul, Alain Guillet & Christian Leclère 1976b. La structure des phrases simples en français, II Constructions transitives, Paris: Rapport de recherches du LADL, No 6, 85p., tables & index, 58 p.

Chomsky, Noam 1965. Aspects of the Theory of Syntax, Cambridge, Mass.: The MIT Press, 251 p.

Chomsky, Noam & Marcel-Paul Schützenberger 1963. The Algebraic Theory of Context-Free Languages, Computer Programming and Formal Systems, Brafford and Hirschberg eds., Amsterdam: North Holland Pub. Co.

Courtois, Blandine 1990. Un système de dictionnaires électroniques pour les mots simples du français, Langue française No 87, Paris: Larousse, pp. 11-22.

Elia, Annibale 1984. Le verbe italien. Les complétives dans les phrases à un complément, Fasano di Puglia: Schena-Nizet, 305 p.

Gross, Gaston 1988. La classification des noms composés du français, Langages No 90, Paris: Larousse, pp. 57-72.

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

Gross, Maurice 1979. On the Failure of Generative Grammar, Language, Vol.55, No4, Baltimore: The Waverly Press, pp. 859-885.

Gross, Maurice 1981. Les bases empiriques de la notion de prédicat sémantique, Formes syntaxiques et prédicats sémantiques, A. Guillet et C. Leclère eds., Langages, No 63, Paris: Larousse, pp.7-52.

Gross, Maurice 1982. Une classification des phrases figées du français, Revue québécoise de linguistique, Vol. 11, No 2, Montreal: Presses de l'Université du Québec à Montréal, pp.151-185.

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

Gross, Maurice & Dominique Perrin eds. 1989 Electronic Dictionaries and Automata in Computational Linguistics, Berlin: Springer Verlag, 110 p.

Guillet, Alain & Christian Leclère 1991. La structure des phrases simples en français. Verbes à complément direct et complément locatif, Geneva: Droz.

Harris, Zellig S. 1952. Discourse Analysis, Language 28, Baltimore: The Waverly Press, pp. 1-30.

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

Harris, Zellig S. 1982. A Grammar of English on Mathematical Principles, New York: John Wiley & Sons Inc., 429 p.

Piot, Mireille 1978. Etudes transformationnelles de quelques classes de conjonction de subordination en français, thèse de l'Université Paris VIII-Vincennes, Paris: LADL, 455 p.

Salkoff, Morris 1983. Bees are swarming in the garden, Language, Vol. 59, No 2, Baltimore: The Waverly Press, pp. 288-346.

Silberztein, Max 1990. Le dictionnaire électronique des mots composés, Langue française No 87, Paris: Larousse, 71-83.

Subirats-Rüggeberg, Carlos 1987. Sentential Complementation in Spanish, A lexico-

grammatical study of three classes of verbs, *Linguisticae Investigationes Supplementa*, No 14, Amsterdam-Philadelphia: J. Benjamins Pub. Co., 290 p.

Vasseux, Philippe 1979. Le système LEXSYN de gestion des données lexicosyntaxiques du LADL, Rapport de recherche du LADL, Paris: LADL.

_. Université Paris 7, Laboratoire d'Automatique Documentaire et Linguistique, Institut Blaise Pascal, CNRS. This work was done partly during a stay at the IBM Yorktown Heights Research Laboratory. I thank E. Black, R. Dougherty, F. Jelinek, C. Leacock and M. Salkoff for their comments and help.

_. This text is taken from an IBM leaflet about research news.

_. Notations are the following: we write N0 for the subject, V for the verb, N1 for the first complement, N2 for the second, etc. We will define other categories in a similar intuitive way.

_. In a purely formal way, the equation we have written can be solved by successive approximations, which provides a non-commutative power series which is a representation of the language characterized by the equation (N. Chomsky and M.-P. Schützenberger 1963).

_. We did not tag all the words in this way, in order to keep simple the form (1S).

_. And also a demonstrative adjective, a hypothesis to be rejected, since it is followed by the verb form is.

_. As a matter of fact, we also used a new rule: S = Adv S, a stylistic variant of rule (R1). Also, the parentheses around in a way are obtained by looking up a dictionary of compound adverbs.

_. The lexicon of inflected forms (DELAF) contains over 600.000 forms. Both lexicons contain a phonemic transcription (E. Laporte 1988).

_. Hence, there are two notions of verbs which should not be confused:

- morphological verbs, that are simple words, a notion relevant to the morphological level of description (conjugation, derivational morphology),
- syntactic verbs, that is, elementary sentences.

Our use of the term verb should be clear from the context.

_. The small number of structures N0 V N1 N2 N3 that have been found can be seen as exceptions on several grounds (i.e. frozen or support structures).

_. Studies on Arabic, Chinese, German, Korean and Madagascan have also been performed.

_. In order to maintain the lexicon-grammar and to use it in applications such as automatic syntactic analysis, it is necessary to access the data base in a convenient way. An index of the simple words contained in the tables can be produced automatically, these words are used as keys to enter the tables (P. Vasseux 1978).

_. Choosing as a representative of equivalence classes either the elementary declarative sentences (Z.S. Harris 1952) or abstract structures (N. Chomsky 1965) is a minor theoretical difference (M. Gross 1979).