

## CHAPTER 3

# Consequences of the metalanguage being included in the language

Maurice Gross

Laboratoire de Linguistique Informatique, NRS-Université Paris Nord

On several occasions, Z. S. Harris stated that the metalanguage of grammar was part of the language. At first sight, this statement is disturbing, but when understood in respect to Harris's practice of grammar construction, it has far-reaching consequences. In principle, the metalanguage of a scientific field is made of concepts and of statements involving these concepts: the laws of the field. In quantum physics for example, concepts are elementary particles, Planck's constant, etc., and statements are Heisenberg's uncertainty relations, etc. In syntax the concepts are essentially the grammatical categories of words (i.e. the parts of speech), and statements are the rules that assemble the words and/or categories into higher units such as phrases and sentences. Modern structural linguists, such as Leonard Bloomfield,<sup>1</sup> set out to formalize the metalanguage, and this activity has become the main trend, whether in generative syntax or in the various logical systems that aim at representing meaning. Meanwhile, the corresponding descriptive work has all but disappeared, at least for languages such as English that should be the main empirical background for theories. Formalization results in a set of abstract symbols and well-defined formal rules, which, in an obvious way, have not much to do with the units of natural language.

Inclusion of the metalanguage in the language can be seen as a methodological principle or as an empirical discovery. We will discuss various aspects of this statement by presenting different examples. We are convinced that the principle has deep consequences for linguistics, but that it may take time and research efforts to measure its full impact.

---

1. E.g. Bloomfield (1933).

For Harris, grammar is the formalized description of a given language, say English.

As in any scientific activity, the metalanguage is constructed by the specialists of the field who agree on an object to describe, that is, on facts to be accounted for. Then abstract entities are defined and refined in order to improve the understanding of facts. Consensus among specialists is reached through experiments, but facts and experiments must be *reproducible*. It goes without saying that research programmes should be common to the linguistic community, whether involved in particular language descriptions or in comparing and abstracting descriptions across languages.

Elements of the metalanguage of grammar have been deeply engrained by education among people. Examples are:

- The categories of words such as *verb, noun, adjective, preposition, affixes*; more abstract units are the *phrases: noun phrases, verb phrases*, etc. and *grammatical functions* such as *subject* or *object*.
- The rules of grammar, such as agreement rules, pronominalization rules, etc.

All of these concepts have been refined into subcategories according to descriptive needs and according to the main application of grammar, which is the teaching of first and second languages.

Most of these concepts are part of a cultural heritage, dating at least to Greek and Roman civilization. Until recently, they have been thought to be universal and have been exported as such by Christian missionaries who used them to describe the languages of Africa, America, Asia, and Oceania. Although specialists have often argued that the Greco-Roman categories are irrelevant to most of these exotic languages, the educational systems of most colonized countries are stuck with this grammatical framework which has been transmitted from generation to generation with remarkable stability.

In fact, the relevance of the Greek-Roman metalanguage even to European languages is far from obvious, but has almost never been questioned. Categories of words have been demonstrated to be useful, for example in the formulation of agreement rules. Confirmation of their value and generality dates back only the nineteenth century, when dictionaries with substantial coverage of the words of a language were built and categories assigned to each word.

## 1. Sentences

The category *sentence* has a special status as the main object of grammar: a grammar of a language must describe all sentences of a language. Sentences are defined on an intuitive basis. An intuition of grammatical acceptability has been developed independently of meaning and has become the empirical basis of syntactic studies. We recall Chomsky's emblematic example; the string of words:

- (1) Colorless green ideas sleep furiously

This has no meaning, but it is grammatical (i.e. it is easy to pronounce, even analyzable in terms of phrases and grammatical functions). The string composed of the same words:

- (2) Sleep colorless furiously ideas green

has no meaning either, but in addition no grammatical structure: it cannot be pronounced with a sentence intonation, no grammatical relation can be seen between words. Let us consider another example:

- (3) Where fell on the floor?

This is perceived as a sentence (interrogative<sup>2</sup>), but the very similar string:

- (4) Where fell on the floor?

is not felt to be a sentence. The string:

- (5) the book that fell on the floor

may be recognized as well-formed according to rules we have learned, but is not felt to be a sentence. Experiences shared by linguists have demonstrated the reproducibility of the intuition of sentence acceptability and its limits. Let us consider examples of the current metalanguage and evaluate them in the framework of grammar construction.

---

## 2. Another term of the metalanguage.

## 2. The predicate

The term *predicate* has numerous interpretations by various authors. The introduction of the linguistic notion can be attributed to Aristotle. It was used in a parallel way in logic and grammar in the Middle Ages, it has recently become a technical term in mathematical logic, and has a wide variety of uses in linguistics. There is no use of the term in ordinary language; it clearly belongs to the respective metalanguages of the mentioned fields. The linguistic predicate appears to have the following descriptive use, based on a notion of sentence quite different from the modern one: as mentioned before, sentences are used to provide information by uttering statements about 'things', which can be concrete objects or abstract entities. Hence, a sentence is made of two components: the 'thing' or subject of the statement, and what is said about the 'thing', which is called a predicate. It does not take long to find examples of sentences that are accepted according to the modern definition and for which the analysis in terms of subject and predicate is irrelevant, for example sentences such as *It is six o'clock* or symmetrical sentences such as *Jo is married to Bob* where it is hard to distinguish the role of *Bob* from that of *Jo*. Nonetheless, over centuries, grammarians have attempted to justify in formal terms the two notions: subject and predicate. The notion 'grammatical subject' is a rather operational notion (M. Gross 1999), it has a definition based on agreement rules between the 'thing' of the sentence and the verb, sometimes also called the predicate. Grammarians keep trying to match the notion of grammatical subject with the semantic notion, defining more and more abstract levels of description to meet the ever-growing number of difficulties. For example, since the two sentences *Bob loves Indian literature* and *Indian literature imparts Bob* are more or less synonymous, they should have the same subjects and predicates. In order to arrive at this result, it is necessary to invent an abstract decomposition or representation of both sentences that will satisfy the requirement, something like:

(6) Indian literature causes Bob to be in a state of love

and then assert whatever is desirable: *Bob* is the subject or *Indian literature* is the subject, or both are subjects. Despite centuries of failure, this notion is so firmly established that such analyses have seemed reasonable. Harris demonstrated exceptional intellectual courage in abandoning the notion and adopting for the description of sentences the general schema:

$N_0 V W$ 

where  $N_0$  is the grammatical subject,  $V$  the verb and  $W$  the sequence of the complements.<sup>3</sup> This seemingly trivial description has in fact deep empirical consequences:

- It eliminates endless and useless discussions that involve the notion of predicate.
- The representation of sentences is based on the widely-recognized fact that the content of  $W$  depends on each verb, and thus, will have to be described case by case.
- An obvious invariant appears for sentences: the sequence  $N_0 V$ .

One thus states that all sentences contain a grammatical subject and a verb. The empirical adequacy of this statement has to be discussed. Many languages do not have an obvious agreement rule that adapts a suffix form of the verb to certain changes in the subject; no such agreement phenomenon is observed with complements. Is such a notion of subject appropriate, say, to Chinese, Japanese, or Korean? Also, in Indo-European languages, there are utterances which are clearly recognized as sentences, but which have no grammatical subject or which have no verb. What is their status? Grammarians have given answers in various cases: imperative sentences have a zeroed grammatical subject which can be reconstructed, some impersonal subjects (i.e. *it*) have been 'regularized' under various proposals such as the Extraposition transformation:

(7) *It seems dangerous to act today = To act today seems dangerous*

There are also utterances such as:

(8) *Good night! Merry Christmas!*

(9) *No point going there.*

to which one naturally attaches the notion of sentence but which linguists have often been reluctant to analyze by ellipsis of a verbal unit: *have* in the imperative sentences:

(10) *Have (a good night + a merry Christmas).*

---

3. Harris used the symbol  $\Omega$  for  $W$ .

or *there is* in impersonal sentences:

- (11) *There is no point going there.*

There are also fully idiomatic utterances, such as:

- (12) *The hell with N!*

which are felt to be sentences but resist analyses other than etymological. Such examples are exceptions to the general statement that all sentences have a subject and a verb. But to claim that they are exceptions, two conditions have to be verified:

- It should not be possible to analyze the utterances in question according to the schema  $N_o V W$ .
- Their number should be small.

The first condition is not too difficult to check. However, depending on the willingness to adopt (or reject) zeroing rules as a tool for reconstructing regular sources, the outcome of the count of exceptions may change substantially. The second condition is much harder to check. It requires quantitative data of various sorts:

- The number of sentences which do have the regular shape  $N_c V W$  should be known.
- An enumeration procedure for the exceptions should be provided.

Precise quantitative data of both kinds have been obtained for French and for a few other languages. They suggest the construction of syntactic tables for elementary sentences, that is, sentences made of a subject, a verb, and its essential complements, if any.

Approximations are easy to obtain: there are about 15,000 verbs in French that are morphologically simple, that is, made of a single word. In English and in other Romance languages, the number is about the same. This figure takes into account the various meanings of homographic or ambiguous verbs such as *drive* in examples (13-15):

- (13) *Bob drives a Ford.*

- (14) *What are you driving at?*

- (15) *Bob drove Jo to her school.*

Here, we count three verbs *to drive* (there are others). But the figure 15,000

does not include idiomatic or frozen forms such as:

(16) *Bob drove away.*

(17) *The noise drove him crazy.*

which have to be counted separately, since they are composed of two units: *drive* and *away*, *drive* and *crazy*.

In French, we have described more than more than 30,000 frozen sentences, more than twice as many as the number of free sentences. Other important corrections have to be made to these numbers. For example, one has to enumerate  $N_0$  *be Adjective* sentences (e.g. *Jo is tall*), of which there are about 10,000 in French, and others such as  $N_0$  *be Prep*  $N_1$  (e.g. *Jo is in trouble*) for which there is no descriptive tradition (i.e. no name has been given to them); of these there are over 7,000 in French.

Only at this point does the enumeration of utterances that meet the intuitive test of sentencehood yet cannot be analyzed according to the schema  $N_0 V W$  become a meaningful enterprise. In French, we allowed zeroing operations of the type given above; under these conditions, about 1000 examples of sentences without a subject and/or a verb were found.

Applying the term 'exception' to them is an interesting issue. We have to balance 1000 unanalyzable forms against more than 50,000 regular schemata. Two per cent may appear a reasonable figure for qualifying an event as rare, or it may not. The term exception is in the eye of the linguist, who may vary his point of view. Let us comment on this situation. The sentence forms involving the different verbs *to drive* can be represented by the following schemata:

(18)  $N_0$  *drive*  $N_1$

(19)  $N_0$  *drive*  $N_1$  *to*  $N_2$

(20)  $N_0$  *drive* *at*  $N_1$

Here, the  $N_i$  s are variable noun phrases whose content is semantically constrained by the verb:

- $N_0$  is human in all three cases.
- $N_1$  is a vehicle in (18), is human in (19), and is difficult to name in (20).
- $N_2$  is a place.

Schemata can be seen as notational variants for functions of several variables such as:

(21) *drive* ( $N_0$ ,  $N_1$ ), *drive* ( $N_0$ ,  $N_1$ ,  $N_2$ )

At this point, formalization takes on a mathematical character. However, the indices attached to the arguments (i.e. noun phrases) of the verbs are also used to define the syntactic transformations the sentences undergo. For example, passive forms are:

(22)  $N_1$  *be driven by*  $N_0$

(23)  $N_1$  *be driven to*  $N_2$  *by*  $N_0$

For the frozen sentences, the schemata are:

(24)  $N_0$  *drive away*

(25)  $N_0$  *drive*  $N_1$  *crazy*

where the noun phrases  $N_0$  and  $N_1$  are human in (24) and (25);  $N_0$  in (25) is either human agentive or unrestricted causative, that is roughly, sentential. But many examples are such that their frozen parts must be indexed too, because frozen parts may undergo the same transformations as free parts. For example, we have:

(26)  $N_0$  *made up his mind* = *His mind was made up*

It then becomes much less natural and much less convenient to use the formal notation of functions:

(27) *drive away* ( $N_0$ )

(28) *drive crazy* ( $N_0$ ,  $N_1$ )

Actually, frozen forms have always been regarded as exceptions. Grammars mention them only briefly and sometimes not at all. However, we have observed a large number of these utterances — as a matter of fact, we observed many more frozen forms than free ones. We are thus entitled to claim that frozenness is a very general phenomenon when it comes to the constitution of sentences. Free variables are not exceptions, since their number is of the same order of magnitude, but they are on a par with frozen items.

Returning to the sentences that do not respect the schema  $N_0$  V W, we have observed that practically all of them were frozen in some sense, hence they are not exceptional from the point of view of sentence formation. Their exceptional character has to be looked for elsewhere, which complicates the situation.

Such a discussion carries us away from the initial question: the interest of



the subject–predicate dichotomy. In fact, from the very beginning, it appears that a cut between the invariant part of the sentence  $N_0 V$  and its variable part  $W$  is more meaningful. But in the end, neither cut appears to be significant: the schema  $N_0 V W$  reflects the general structure of (Indo-European) sentences in a very precise way. Eliminating the notion of predicate leaves us with an improved metalanguage, that is, more operational: the notion of verb is that of a category defined in extension:

- (i) Verbs are easy to recognize by their endings and a list of them is easily established.
- (ii) Extensions of verbs by prefixation and compounding is productive, and it may be hard to determine which verbs will accept such prefixes as *re-*, *co-*, *un-*, etc. Also, appearance of compounds of the form *to chain-smoke*, *to code-name an operation*, *to radiocarbon-date bones* is not predictable.

Nonetheless, a clear picture of the set of verbs of a language can be reached and used to classify words and processes, leaving a residue that only then can be seen as made of exceptions.

### 3. Metalinguistic sentences for morphology

Consider the following sentences:

(29) (*The word + A word such as*) *arrival* (*takes + has + contains*) *two r's*.

(30) *Arrival* (*does not have + never contains*) *a y*.

From a distributional point of view, subjects contain essentially one variable which ranges over the list of English words. In the same way, the main variable of the complements in (29) and (30) ranges over the English alphabet. The determiner may vary, but within a narrow range, as in:

(31) *English* *has a capital e*.

Modifying adjectives may be introduced that correspond to some comment made by the utterer of the sentence, as in:

(32) *Oxygen* *contains a nasty y*.<sup>4</sup>

---

4. Nasty, because oxide does not.

To be semantically (logically) correct, sentences (29), (31), and (32) must present an identity relation between the spelling of the subject and the letters of the complement. In (30), the relation is different. There are many other analogous sentences that describe the shapes of words:

(33) *Few English words (have + are of) length two.*

(34) *Many English words contain an e.*

(35) *Oxygen begins with an o and ends in n.*

Even without introducing the metalinguistic terms 'prefix', 'ending', or 'suffix', such sentences allow a detailed morphological description of written English. By replacing letters with sounds in the preceding examples, the description involves phonemes, that is, it becomes morphophonological. In other words, the sentences constitute the metalanguage of morphology, extensible by the introduction of terms that are more technical. Notice that a device similar to that of morphology is used for expressing intensity in:

(36) *He is stupid with a capital S.*

As Harris (1991: 123-144) observed, all these metalinguistic sentences belong to English. At this point, one may argue that they constitute a special subset of English sentences and should not be considered as common English — in other words, that this metalanguage is outside of the language. It seems, however, difficult to sustain such an argument, since many other families of sentences with a similar specialized focus can be easily distinguished. Consider, for example, sentences dealing with costs or prices:

(37) *This book costs ten dollars.*

(38) *I (paid + spent) ten dollars for this book, etc.*

The direct complements of *to cost* and *to pay* have a highly specialized distribution. The following sentences that correspond to length measurements or to weight or time are similar:

(39) *This river is 100 meters (deep + long + wide).*

(40) *The book weighs five pounds.*

(41) *I spent six hours and 23 minutes reading this book.*

It is hard to say that we are dealing here with technical sublanguages, since

many of these sentences have non-numerical variants that are well-rooted in ordinary language and that are not essentially different, either from a syntactic or a semantic point of view. This is the case for the sentences:

(42) *This river is very (deep + long + wide).*

(43) *The book is heavy.*

(44) *I spent a lot of time reading the book.*

Historically, numerical utterances have become available and have been made precise in a gradual way, following scientific and technical progress. In many cases, they have been allowed to occupy the same syntactic positions as informal utterances also used to express quantities. If we exclude these sentences from the languages, not much will be left to be considered as ordinary, non-technical language. There can be little doubt, then, that the metalanguage of grammar is a part of its subject matter, language itself.

## References

- Bloomfield, Leonard. 1933. *Language*. New York: Henry Holt & Company.  
Gross, Maurice. 1999. "Sur la définition d'auxiliaire du verbe". *Langages* 135: 8–21.  
Harris, Zellig S. 1991. *A Theory of Language and Information: A mathematical approach*. Oxford and New York: Clarendon Press.