1. INTRODUCTION:

General review of relations between formal grammar theory, natural linguistics and psycholinguistics.

1.1 Origin and basic problems of formal grammar theory

This chapter is introductory to the following three. Its aim is to give an historical outline of the mutual inspiration that we have seen in the last fifteen or so years between formal grammar theory, natural language theory and psycholinguistics. In the following three chapters we will discuss some recent characteristic examples of this interaction.

Fifteen to twenty years are long enough to have almost forgotten how formal grammar theory came into existence. The origin of this theory comes from the study of natural language. A description of natural language is traditionally called a grammar. It specifies construction of sentences, relations between linguistic units, etc.. Formal grammar theory started from the need to give a formal mathematical basis for such descriptions. Initially the creation of these new formal systems was largely the work of Noam Chomsky. His aim was not so much to refine linguistic descriptions, but to construct a formal basis for the discussion of the foundations of linguistics. "What should be the form of a linguistic theory?", "What sort of problems can be expressed by way of different formal means, and what do we take to be a solution?": these were the main issues to be tackled. In short, formal grammars were developed as mathematical models for linguistic structure.

The first developments only concerned the syntax of natural languages, not their semantics. The most successful application of formal grammar theory have been up to now in the area of syntax. All our discussion will therefore be largely limited to syntactic issues.

The first and most obvious use of formal grammar theory in linguistics was to create a variety of more or less restrictive grammars, and to compare their generative power to the empirical requirements of linguistic data. Let us

call this the problem of "generative power". In this chapter we will discuss and criticize some historical highlights in the approach to this problem. In the next chapter some important recent results will be discussed.

The explicit use of formal grammars in linguistics also created a more general and more philosophical problem. It is one thing to formalize a linguistic theory, it is quite another thing to formulate the relation between such a theory on the one hand, and empirical linguistic data on the other hand. The problem here consists in clarifying what, exactly, is the empirical domain of the linguistic theory, and what is the empirical interpretation of the elements and relations that figure in the theory. We will call this problem the interpretation problem, after Bar-Hillel. In this chapter we will only make some general points relating to this issue. The third chapter, however, will be devoted to a formal psycholinguistic analysis of the interpretation problem.

The linguistic origin of formal grammar theory, finally, also led to the early development of theories of grammatical inference. There were two reasons for this. Firstly, a main theme in structural linguistics had for a long time been the development of so-called "discovery procedures", i.e. methods to detect structures in linguistic data. Secondly, probably under the influence of the psychologist George Miller, Chomsky had realized the fundamental problem of language acquisition. The description of a language is one thing, but the causation of linguistic structures is another more fundamental issue. Only a solution of this latter problem will give linguistic theory an explanatory dimension. Efforts to write formal systems which are able to infer a grammar from a data corpus can be found as early as 1957. Since then, inference theory has had a considerable development. In the last chapter we will be concerned with some relations between recent inference theory and psycholinguistic models of language acquisition.

1.2 Observational adequacy of regular and context-free grammars.

Let us now return to the early developments of formal

grammar theory. We will first very quickly review the variety of grammars that Chomsky developed in the second half of the fifties. Then we will discuss some problems relating to the linguistic adequacy of regular and contex-free grammars.

According to Chomsky, a grammar is defined as a system

$$G = \langle V_N, V_T, P, S \rangle ,$$

where V_N is a finite nonempty set, the nonterminal vocabulary, whose elements are called category symbols or auxiliary variables;

V_T is a finite nonempty set, the terminal vocabulary whose elements are usually called"words" or "morphemes";

S is an element of V_N (the start symbol).

Given a set E of symbols, we denote by E^* the set of all strings of finite length which can be obtained by concatenation of symbols in E; by E^+ we shall denote the set $E^* - \{\Lambda\}$, where Λ is the null string (of zero length).

Now P (the set of production rules of the grammar) is a finite set of rules of the form $\alpha \longrightarrow \beta$, with $\alpha \in V^+$ and $\beta \in V^*$, where $V = V_T \cup V_N$.

We shall say that a string $\gamma \in V^+$ directly produces a string $\delta \in V^*$ (in symbols $\gamma \Rightarrow \delta$) if $\gamma = \varphi\,\eta\,\psi$, $\delta = \varphi\,\theta\,\psi$, for some $\eta, \theta, \varphi, \psi \in V^*$, and $\eta \longrightarrow \theta$ is in P. Finally, we say that $\gamma \in V^+$ derives (directly or not) a string $\delta \in V^*$ (in symbols $\gamma \overset{*}{\Rightarrow} \delta$) if either $\gamma = \delta$, or there exist strings $\gamma_0, \gamma_1, \ldots, \gamma_n$, for some finite, n, such that

$$\gamma_0 = \gamma, \quad \gamma_i \Rightarrow \gamma_{i+1}, \text{ for } i = 0, \ldots, n-1,$$

and $\gamma_n = \delta$.

Now the language L_G generated by a grammar G as above is defined as the set

$$L_G = \{ \alpha \mid \alpha \in V_T^*, S \overset{*}{\Rightarrow} \alpha \}$$

The variety of grammars that Chomsky defined came about by putting more and more restrictive conditions on the format of production rules.

These are:

(0) no restriction: type 0 grammars.

(1) for all rules $\alpha \longrightarrow \beta$ of P, the length of β should be not less than the length of α : context-sensitive grammars (type 1)

(2) for all rules $\alpha \longrightarrow \beta$ of P, we must have $\alpha \in V_N, \beta \neq \Lambda$: context-free grammars (type 2).

(3) for all rules $\alpha \longrightarrow \beta$ of P, we must have $\alpha \in V_N$, and either $\beta \in V_T$, or β equal to the concatenation of an element of V_T and one of V_N, in that order: regular grammars (type 3).

A language is called type-i if it can be generated by a type-i grammar.

There is a strict inclusion relation among the classes of languages defined above: if C_i is the class of languages of type i, then $C_{i+1} \subsetneq C_i$. In particular there are not regular (i.e. type 3). These are exactly the languages that are called "self-embedding". A context-free language is self-embedding if all grammars generating it are self-embedding. A context-free grammar is self-embedding if there is a $B \in V_N$ such that $B \overset{*}{\Longrightarrow} \alpha B \gamma$, where α and γ are non-empty strings.

Chomsky (1956, 1957) rejected regular languages as adequate models for natural languages. The argument used by Chomsky to conclude that natural languages are at least non-regular had an enormous influence on the development of modern linguistics; this justifies a rather detailed discussion of it. It is also the case that the argumentation, as given in Syntactic Structures (1957), is not completely balanced (the same is true, to a lesser degree, of Chomsky's treatment of the question in 1956). A consequence of this has been that the same sort of evidence is incorrectly used for the rejection of other types of grammars, and erroneous