Generative Capacity

Emmon Bach and Philip Miller

Final version

GENERATIVE CAPACITY was introduced by Chomsky (1963) in the context of the theory of formal grammars and automata (→ *Finite State Grammars and Languages,* → *Context Free Grammars and Languages,* → *Mildly Context Sensitive Grammars and Languages,* → *Automata Theory*). A language is defined as a set of strings over some vocabulary (e.g., a set of sentences over a vocabulary of words). A formal grammar defines a language, in the sense that it provides an abstract mechanism which generates the strings belonging to the language. (We will henceforth use the terminology of grammars; completely parallel formulations are possible in terms of languages accepted by automata within automata theory, or in terms of languages defined by constraint satisfaction within a declarative formalism.) The W[EAK] G[ENERATIVE] C[APACITY] of a grammar is then simply defined as the language generated by the grammar, and the WGC of a class of grammars is defined as the set of languages generated by the grammars in the class. For example, the WGC of a given C[ontext] F[ree] G[rammar] is the C[ontext] F[ree] L[anguage] which it generates, whereas the WGC of the class of CFGs is the set of CFLs. If a theory of grammar is defined as a specification of a set of grammars, then the WGC of that theory is the WGC of the set of grammars thus specified.

If a grammar generates strings while at the same time assigning them a structural description of some kind (e.g. a CFG associates each sentence it generates with a tree), then one can define the S[TRONG] G[ENERATIVE] C[APACITY] of the grammar as the set of structural descriptions generated by the grammar. Likewise, the SGC of a class of grammars (or theory) is the set of sets of structural descriptions generated by the grammars in the class: for example, the SGC of the class of CFGs is the set of sets of trees generated by CFGs.

Questions of weak and strong generative capacity can be empirically relevant in theoretical linguistics. Suppose a general linguistic theory T is defined in a way that is precise enough for us to give a formal characterization of the class of grammars projected by T. It then becomes possible to investigate whether T is too restrictive or too powerful in WGC or SGC for natural languages, and to compare various theories in this respect. This kind of research tries to specify as narrowly as possible some class of formal grammars that is just powerful enough to characterize all and only the possible human languages. Part of the early work in generative grammar was devoted to showing that certain types of formalisms (for example, Finite State Grammars) are intrinsically incapable of assigning some types of structural descriptions (for example, recursive self-embedding) that are present within natural languages. Such formalisms are thus insufficiently powerful in SGC to provide an adequate analysis of natural language phenomena. Other early arguments focussed on the inadequacy of CFGs in WGC, but they have been shown to be flawed (see Pullum and Gazdar 1982). Since then new results (see Culy 1985, Shieber 1985, Miller 1991) have indicated that natural languages are → Mildly Context Sensitive.

Issues of SGC are in principle very relevant to theoretical linguistics, because linguists are usually more interested in the appropriateness of the structural descriptions than in the simple enumeration of well-formed sentences. However there has been very little study of SGC because the concept, as initially defined, did not allow meaningful comparison, neither within a single theory (it was quickly pointed out, for instance, that two CFGs are strongly equivalent only if they are identical), nor between theories, since the structural descriptions are then usually not of the same type, and are consequently not comparable (see Kuroda 1976).

During the late 80s and the 90s, there was renewed interest in SGC especially for → mildly contexts sensitive formalisms. Moreover, two new approaches to SGC have recently been developed in order to overcome the problem of incommensurability and allow comparison between different formalisms. Rogers (1998) proposes a logical analysis of the SGC of CFGs which allows him to show that the languages licensed by particular theories within the → Government and Binding framework are strongly context-free. Miller (1999) proposes that SGC should be understood as a model-theoretic semantics for linguistic formalism, mapping structural descriptions from different theories into interpretation domains, which are specifically set up to represent aspects of linguistic structure in a theory-neutral way.

References Chomsky, Noam. 1963. Formal properties of grammar. In Luce, R.D., R.R. Bush and E. Galanter (eds), Handbook of Mathematical Psychology, vol.II. New York: Wiley, pp. 323-418.

Chomsky, Noam. 1965. Aspects of the Theory of Syntax. Cambridge, MA: MIT Press.

Culy, Christopher. 1985. The complexity of the vocabulary of Bambara. Linguistics and Philosophy 8.3, 345-353.

Kuroda, S.-Y. 1976. A topological study of phrase structure languages. Information and Control 30, 307-379.

Miller, Philip H. 1991. Scandinavian extraction phenomena revisited: weak and strong generative capacity. Linguistics and Philosophy 14.1, 101-113.

Miller, Philip H. 1999. Strong Generative Capacity: The Semantics of Linguistic Formalism. Stanford: CSLI Publications

Pullum, Geoffrey K. and Gazdar, Gerald. 1982. Natural languages and context free languages. Linguistics and Philosophy 4, 471-504.

Rogers, James 1998. A Descriptive Approach to Language-Theoretic Complexity. Stanford: CSLI Publications.

Savitch, Walter J., et al., eds. 1987. The Formal Complextiy of Natural Language. Dordrecht: Reidel

Shieber, Stuart M. 1985. Evidence against the context-freeness of natural language. Linguistics and Philosophy 8.3., 333-345.

Index generative capacity, weak generative capacity, strong generative capacity