ON THE STRUCTURE OF CONTEXT-SENSITIVE GRAMMARS[†]

by

Ronald V. Book

6-71

Center for Research in Computing Technology

Harvard University
Cambridge, Massachusetts 02138

ON THE STRUCTURE OF CONTEXT-SENSITIVE GRAMMARS[†]

by

Ronald V. Book

6-71

Center for Research in Computing Technology

Harvard University

Cambridge, Massachusetts 02138

[†] This paper was written while the author was visiting the Department of System Science, University of California, Los Angeles. It was supported in part by the National Science Foundation under Grant No. GJ-803, and by the National Aeronautics and Space Administration under Grant No. NGR 22-007-176.

Introduction.

It is well known that the family of context-sensitive grammars generate languages which are not context-free and that it is undecidable whether a context-sensitive grammar generates a context-free language. However the mechanism by which the use of context allows a non-context-free language to be generated is not well understood--in fact, the question itself is vague: what does context do for you? In this paper we attempt to make this question more precise by surveying some of the results which speak to two more specific questions: (i) what constraints can be placed on the form of the rules of context-sensitive grammars without restricting the weak generating capacity?; and (ii) what (nontrivial) constraints can be placed on the form of the rules of context-sensitive grammars such that only context-free languages will be generated?

Section 1.

In this section we review some basic definitions and facts about context-sensitive grammars and languages.

A grammar $G = (V, \Sigma, R, X)$ is context-sensitive if each rule is of the form $\alpha Z\beta \to \alpha\gamma\beta$ where $\alpha, \beta, \gamma \in V^*$, $\gamma \neq e$, and $Z \in V - \Sigma$. A language is context-sensitive if it is generated by some context-sensitive grammar. Let us recall a few facts about context-sensitive grammars and languages.

A language is context-sensitive if and only if it is accepted by a nondeterministic linear bounded automaton (a nondeterministic Turing acceptor which uses an amount of tape proportional to the length of the input). Hence every context-sensitive language has a primitive recursive characteristic function but there are sets with primitive recursive characteristic functions which are not context-sensitive. It is not known whether every context-sensitive language is accepted by a deterministic linear bounded automaton.

The family of context-sensitive languages possesses many of the positive closure properties of the context-free languages; for example, this family is a principal AFL closed under e-free substitution. However, this family is not closed under arbitrary homomorphic mappings; in fact the image of the family of context-sensitive languages under arbitrary homomorphic mappings is the family of recursively enumerable sets. In addition, this family is closed under intersection and it is an open question whether it is closed under complementation.

The form of the rules of a context-sensitive grammar places no restrictions on either left or right context. It is clear that no generative capacity is lost if one requires that the context be free of terminal symbols. Similarly, no generative capacity is gained (or lost) if more than one symbol is transformed by a rule, that is, if the non-context-free rules do not transform terminal symbols and obey the restriction that length is preserved, i.e., $\rho \to \theta$ where $|\rho| \leq |\theta|$. In this case the grammar is called monotonic.

Perhaps the best known normal form for context-sensitive grammars is that introduced by Kuroda. [10]. A context-sensitive grammar $G = (V, \Sigma, R, X)$ is in Kuroda Normal Form if each rule in R is of one of the following forms:

$$\left.\begin{array}{l} Z \to YY' \\ ZY \to Z'Y \\ YZ \to YZ' \\ Z \to Z' \\ Z \to a \end{array}\right\} \qquad Z, Z', Y, Y' \in V - \Sigma \text{ , } a \in \Sigma \text{ .}$$

Given any context-sensitive grammar G_1, one can effectively construct a context-sensitive grammar G_2 in Kuroda Normal Form such that $L(G_2) = L(G_1)$.

The restriction on the form of the rules of a context-sensitive grammar prevents the empty word from being generated. Often the restriction is altered so that a context-sensitive (or extended context-sensitive) language is allowed to contain the empty word, usually by adding some ad hoc rule which generates only the empty word. In general, if erasing (i.e., length-decreasing) rules are present, then a non-context-sensitive language may be generated. Ginsburg and Greibach [5] have shown that limited use of erasing rules in derivations preserve the property of generating only a context-sensitive language--in this case, "limited use" means applying no more than a number proportional to the length of the terminal string generated. In general, the upper bound for such "limited use" is not known [1].

Section 2.

In this section we consider some of the constraints which can be placed on the form of the rules of context-sensitive grammars without forcing the language generated to be only context-free. This is done in order to gain some insight into the question, "what does context do for you?"

If one is constructing a context-sensitive grammar to generate some given non-context-free language, then one often proceeds as if context can be used to "store and transmit information." Thus one builds rules so that in derivations "messages" (or "pulses") are transmitted along a string. Sometimes this effect is achieved by building a grammar which imitates the action of some linear bounded automaton; hence, the action of the read-write head must be imitated as it travels back and forth across the tape.

The "ability to send messages" has not been formalized in such way as to explain "what context does for you," although some properties of the structure of derivations have been studied [1, 6, 7]. However this notion does provide an intuitive "handle" for studying some questions and for gaining perspective on some results on context-sensitive grammars and languages. (For a different approach to some of these questions, see [3] particularly Section 3.)

Haines [8] has established an important result on the structure of context-sensitive grammars. A left context-sensitive grammar is a context-sensitive grammar $G = (V, \Sigma, R, X)$ such that every non-context-free rule is of the form

$$\alpha Z \rightarrow \alpha\gamma \text{ where } \alpha, \gamma \in V^{*}, Z \in V - \Sigma \text{ and } \gamma \neq e.$$

(Thus, every non-context-free rule has only left context.) Haines has shown that each context-sensitive language is generated by a left context-sensitive grammar; in addition, the context need not be more than a single symbol, i.e., $|\alpha| = 1$.

Interpreting this result in terms of "message-sending," we see that it is enough to "send messages" only in one direction--in this case, to the right--so that one need only "guess" that the "message was received." (Of course one has the same result for right context-sensitive grammars since the theory is symmetric.) This result may be interpreted as a new normal form theorem, that is, it is enough to consider just those grammars with context-free rules and left context-sensitive rules.

From the results of Kuroda [10] (in particular, Kuroda Normal Form), it is easy to see that one loses no generative capacity if one considers only those grammars such that in any derivation only context-free rules are applied first and then only length-preserving context-sensitive rules are applied. One might well ask whether the same result is obtained if one relaxes the requirement that the non-context-free be length-preserving. It is shown in [1] that this is not the case.

If one considers those context-sensitive grammars such that every rule is either length-increasing or generates only a terminal symbol, then the family $\mathscr{L}$ of languages generated has the following properties:

(i) every e-free context-free language is in $\mathscr{L}$;

(ii) for every recursively enumerable set L, there exists a language L_0 in $\mathscr{L}$ and a homomorphism h such that $h[L_0] = L$ - hence, $\mathscr{L}$ contains languages which are not context-free;

(iii) $\mathscr{L}$ is a proper subfamily of the context-sensitive languages--in particular, $wcw^Rcw \mid w \varepsilon \{a,b\}^*\}$ is not in $\mathscr{L}$.

Note that (ii) implies that a grammar of this form must have some "message-sending capacity" but the proof of (ii) given in [1] depends on the grammar also generating some "padding" which is eventually erased by the homomorphism. On the other hand, (iii) shows that the power to "send messages" is quite restricted.

One effect of context is the ability to achieve a permutation of symbols, that is, using context-sensitive rules one can obtain $AB \overset{*}{\Longrightarrow} BA$. One might well ask if context-sensitive rules allow anything more than the composition of permutations along with the type of substitution allowed by context-free rules. Sillars [13] has shown that this is not the case. He did this by investigating "permutation grammars," grammars which have only permutation rules in addition to the context-free rules.

A monotonic grammar $G = (V, \Sigma, R, X)$ is a <u>permutation grammar</u> if each non-context-free rule in R is of the form $Y_1 \dots Y_m \to Y_{\pi(1)} \dots Y_{\pi(m)}$ where each $Y_j \in V-\Sigma$ and where π is some permutation of $\{1,\dots, m\}$ with no fixed points. Let Π_R be the set of "permutation" rules of R and let $G_r = (V, \Sigma, R-\Pi_R, X)$, so that G_r is a context-free grammar. It can be shown that L(G) is infinite if and only if $L(G_r)$ is infinite, so that L(G) is infinite if and only if it has an infinite context-free subset. Thus the permutation grammars cannot generate all the context-sensitive languages; for example, $\{a^n b^n c^n \mid n \geq 1\}$ has no infinite context-free subset. On the other hand, it is clear that the permutation grammars do generate non-context-free languages; for example, it is easy to construct a permutation grammar to generate $\{w \in \{a,b,c\}^* \mid w$ has an equal number of a's, b's, and c's$\}$.

Note that a permutation grammar has only length-preserving rules in addition to its context-free rules. However it is not clear how to analyze such grammars in terms of "message-sending." In fact, these grammars may be best analyzed by considering them to be weak extensions of context-free grammars.

It should be noted that there are other types of grammars which extend the context-free languages but generate only a proper subfamily of the context-sensitive languages. For example, there are matrix grammars, indexed grammars, programmed grammars, time-variant grammars, and others. In these cases either the type of rule is different from the definition of a context-sensitive rule or the way that the rules are applied is regulated in some fashion. It is not clear whether the study of these grammars would throw light on the questions considered in this paper.

Section 3.

In this section we consider certain restrictions on the form of the rules of a context-sensitive grammar which force the language generated to be context-free. In each case it is decidible whether the rules of a grammar satisfy the restriction.

Recall that a context-free language may be represented in various ways: by a context-free grammar; by a nondeterministic pushdown store acceptor; by a formula $h_1(h_2^{-1}(D) \cap R)$, where h_1 and h_2 are homomorphisms, D is a Dyck set, and R is a regular set. However every grammar has a particular context-free language associated with it, the language obtained by left-to-right derivations.

In an arbitrary rewriting system or grammar G = (V, Σ, R, X), a derivation $\psi_0 \Rightarrow \psi_1 \Rightarrow \cdots \Rightarrow \psi_n$ is a left-to-right derivation if for each i=1,..., n, there are strings $\alpha \in \Sigma^*$, $\beta \in V^*$, and a rule $\rho \to \theta \in R$ such that $\psi_{i-1} = \alpha\rho\beta$ and $\psi = \alpha\theta\beta$ --that is, at each step the rewriting rule is applied at the leftmost possible position in the string. It has been shown by Evey [4] and by Matthews [11] that for any rewriting system or grammar, the set of terminal strings obtained by left-to-right derivations from the initial symbol is a context-free language. If one relaxes the "leftmost" condition to allow the rewriting rule to be applied within some fixed distance of the leftmost possible position in the string, then once again one obtains a context-free language. Clearly one can substitute right-to-left for left-to-right since the general theory of rewriting systems is symmetric.

In an arbitrary rewriting system or grammar G = (V, Σ, R, X), a derivation $\psi_0 \Rightarrow \psi_1 \Rightarrow \cdots \Rightarrow \psi_n$ is a two-way derivation if for each i-1,..., n, there are strings α, β and a rule $\rho \to \theta \in R$ such that $\psi_{i-1} = \alpha\rho\beta$, $\psi_i = \alpha\theta\beta$, and either $\alpha \in \Sigma^*$ or $\beta \in \Sigma^*$ --that is, at each step the rewriting rule is applied at either the leftmost or the right most possible position in the string. It has been shown by Matthews [12] that for any rewriting system or grammar the set of terminal strings obtained by two-way derivations from the initial symbol is a context-free language. (A short and elegant proof of this result was given by Ginsburg and Greibach [5].) Again one can allow the rewriting rule to be applied within some fixed distance of the ends and still generate only a context-free language.

These results provide a useful tool for showing that a given grammar generates a context-free grammar. Thus, if one can show that for some given grammar the set of left-to-right (or right-to-left or two-way) derivations yields the entire language generated by that grammar, then one can conclude that the grammar generates only a context-free language. This method of attack is illustrated by the following result.

Let $G = (V, \Sigma, R, X)$ be any context-sensitive grammar such that each non-context-free rule is of the form $\alpha Z\beta \rightarrow \alpha\gamma\beta$ where $\alpha \varepsilon \Sigma^*$, $Z \varepsilon V - \Sigma$, β , $\gamma \varepsilon V^*$, and $|\alpha| \geq |\beta|$ (i.e., the left context is a terminal string and is at least as long as the right context). It is shown in [2] that L(G) must be context-free. This is done by showing that L(G) may be obtained from derivations such that at each step the rewriting rule is applied within some fixed distance of the leftmost possible position in the string--in this case the fixed distance is $m(m+1)/2$ where $m = \max\{ |\alpha| \mid \alpha Z\beta \rightarrow \alpha\gamma\beta \; \varepsilon \; R \}$.

In this case the terminal context requirement allows the strings to be generated while "sending messages" only a bounded distance, in fact, a distance bounded by $m(m + 1)/2$.

It should be noted that the requirement that the terminal left context be at least as long as the right context cannot be relaxed [2]--that is, if it is relaxed, then it is possible to generate non-context-free languages.

There are several other results whose known proofs do not depend on reduction to the yield of left-to-right derivations. In some cases the restrictions on the form of the rules may be viewed as creating "barriers" which keep "messages" from being "passed."

Let $G = (V,\Sigma,R, X)$ be a monotonic grammar. Suppose $<$ is a partial order on V with the property that for each rule $Z_1 \ldots Z_p \rightarrow Y_1 \ldots Y_q$ in R, there exists Y_j in $\{Y_1,\ldots,Y_q\}$ such that for $i = 1,\ldots, p$, $Z_i < Y_j$. In this case $L(G)$ is context-free [9].

Let $G = (V, \Sigma, R, X)$ be a monotonic grammar. Suppose that every non-context-free rule in R is of the form $\rho \rightarrow \theta$ where $\rho \in (V - \Sigma)^*$ and $\theta \in V^*\Sigma V^*$ (so that every application of a non-context-free rule generates a new terminal symbol which cannot be used as part of context). In this case $L(G)$ is context-free [5]. This result strongly suggests the notion of "barrier": terminal symbols cannot be used as part of context and a new terminal symbol is generated whenever context is used -- thus "messages" cannot be "transmitted" more than some bounded distance.

Let $G = (V, \Sigma, R, X)$ be a context-sensitive grammar. Suppose that each non-context-free rule in R is of the form $\alpha Z\beta \rightarrow \alpha\gamma\beta$ where $\alpha, \beta \in \Sigma^*$, $Z \in V-\Sigma$, $\gamma \in V^*$. In this case $L(G)$ is context-free [2].

In each case cited above the known proofs depend on the representation of a context-free language as $h_2^{-1}((D) \cap R)$ or as $h_1(D \cap R)$ where h_1 and h_2 are homomorphisms, D is a Dyck set and R is a regular set. In each case the proof suggests the idea of a "barrier" but there is no formalization of this notion.

A careful analysis of the proofs of the results cited above suggests that a context-sensitive grammar generates a non-context-free language only if an infinite number of strings in the language are generated by derivations which transform symbols arbitrarily far from either the leftmost or rightmost nonterminal symbol, and that arbitrarily many steps of the derivation "interact" in some way. What is needed is a formalization of the notion of "message-sending" and of "barrier" in order to prove the appropriate necessary conditions for a grammar to generate a non-context-free language. Such results would provide a reasonable answer to the question: What does context do for you?

References

1. R. Book, Time-bounded grammars and their languages, J. Computer and System Sciences 5(1971), to appear.

2. R. Book, On the effect of terminal context on context-sensitive grammars, Math. Linguistics and Automatic Translation, NSF-18 (1967), Computation Laboratory, Harvard University, Cambridge, Mass.

3. N. Chomsky, Formal properties of grammars, in Luce, Bush, Galanter (Eds.), Handbook of Math. Psychology, Vol. II, John Wiley, New York, 1963.

4. R. Evey, "The Theory and Application of Pushdown Store Machines," Ph.D. Dissertation Harvard Univ., 1963. Also appears as Math. Linguistics and Automatic Translation NSF-10 (1963), Computation Lab., Harvard Univ., Cambridge, Mass.

5. S. Ginsburg and S.A. Greibach, Mappings which preserve context-sensitive languages, Info. and Control 9 (1966), 563-582.

6. A. Gladkii, On the complexity of derivations in phrase structures grammars, Algebri i Logika Sem. 3 (1964), 29-44.

7. T. Griffiths, Some remarks on derivations in general rewriting systems, Info. and Control 12 (1968), 27-54.

8. L. Haines, Representation theorems for context sensitive languages, submitted for publication.

9. T. Hibbard, "Scan Limited Automata and Context Limited Grammars," Ph.D. Dissertation Univ. of California, Los Angeles, 1966.

10. S.Y. Kuroda, Classes of languages and linear-bounded automata, Info. and Control 7 (1964), 207-223.

11. G.H. Matthews, A note on asymmetry in phrase structure grammars, Info. and Control 7 (1964), 360-365.

12. G.H. Matthews, Two-way languages, Info. and Control 10 (1967), 111-119.

13. W. Sillars, "Formal Properties of Essentially Context-Dependent Languages," Ph.D. Dissertation, Penn. State Univ., 1968.