CATEGORIAL AND NON-CATEGORIAL LANGUAGES

Joyce Friedman
Ramarathnam Venkatesan

Computer Science Department
Boston University
111 Cummington Street
Boston, Massachusetts 02215 USA

ABSTRACT

We study the formal and linguistic properties of a class of parenthesis-free categorial grammars derived from those of Ades and Steedman by varying the set of reduction rules. We characterize the reduction rules capable of generating context-sensitive languages as those having a partial combination rule and a combination rule in the reverse direction. We show that any categorial language is a permutation of some context-free language. thus inheriting properties dependent on symbol counting only. We compare some of their properties with other contemporary formalisms.

INTRODUCTION

Categorial grammars have recently been the topic of renewed interest, stemming in part from their use as the underlying formalism in Montague grammar. While the original categorial grammars were early shown to be equivalent to context-free grammars, [1,2,3] modifications to the formalism have led to systems both more and less powerful than context-free grammars.

Motivated by linguistic considerations, Ades and Steedman[4] introduced categorial grammars with some additional cancellation rules. Full cancellation rules correspond to application of functions to arguments. Their partial cancellation rules correspond to functional composition. The new backward combination rule is motivated by the need to treat preposed elements. They also modified the formalism by making category symbols parenthesis-free, treating them in general as governed by a convention of association to the left, but violating this convention in certain of the rules.

This treatment of categorial grammar suggests a family of categorial systems, differing in the set of cancellation rules that are allowed. Earlier, we began a study of the mathematical properties of that family of systems,[5] showing that some members are fully equivalent to context-free grammars, while others yield only a subset of the context-free languages, or a superset of them.

In this paper we continue with these investigations. We characterize the rule systems that can obtain context-sensitive languages, and compare the sets of categorial languages with the context-free languages. Finally, we discuss the linguistic relevance of these results, and compare categorial grammars with TAG systems in this regard.

PRELIMINARIES

A *categorial grammar* under a set R of reduction rules is a quadruple $CG_R(VT, VA, S, F)$, whose elements are defined as follows: VT is a finite set of morphemes. VA is a finite set of atomic category symbols. $S \in VA$ is a distinguished element of VA. To define F, we must first define CA, the set of category symbols. CA is given by: i) if $A \in VA$, then $A \in CA$; ii) if $X \in CA$ and $A \in VA$, then $X/A \in CA$; and iii) nothing else is in CA. F is the lexicon, a function from VT to 2^{CA} such that for every $a \in VT$, $F(a)$ is finite. We often write CG_R to denote a categorial grammar with rule set R, when the elements of the quadruple are known.

Notation: Morphemes are denoted by a, b; morpheme strings by u, v, w. The symbols S, A, B, C denote atomic category symbols, and U, V, X, Y denote arbitrary (complex) category symbols. Complex category symbols whose left-most symbol is S (symbols "headed" by S) are denoted by X_S, Y_S. Strings of category symbols are denoted by x, y.

The language of a categorial grammar is determined in part by the set R of reduction rules. This set can include any subset of the following five rules. In each statement, $A \in VA$, and $U/A, A/U, A/V, V/A \in CA$.

(1) (F Rule) The string of category symbols $U/A\ A$ can be replaced by U. We write: $U/A\ A \rightarrow U$;

(2) (FP Rule) The string $U/A\ A/V$ can be replaced by U/V. We write: $U/A\ A/V \rightarrow U/V$;

(3) (B Rule) The string $A\ U/A$ can be replaced by U. We write: $A\ U/A \rightarrow U$;

(4) (B_S Rule) Same as B rule, except that U is headed by S.

(5) (BP Rule) The string $A/U\ V/A$ can be replaced by V/U. We write: $A/U\ V/A \rightarrow V/U$.

If $XY \rightarrow Z$ by the F-rule, XY is called an *F-redex*. Similarly, for the other four rules. Any one of them may simply be called a *redex*.

The reduction relation determined by a subset of these rules is denoted by $=>$ and defined by: if $X\ Y \rightarrow Z$ by one of the rules of R, then for any α, β in CA*, $\alpha X Y \beta\ > \alpha Z \beta$. The reflexive and transitive closure of the relation $->$ is $=>^*$. A morpheme string $w = w_1 w_2 \cdots w_n$ is *accepted* by $CG_R(VT, VA, S, F)$ if there is a category string $x = X_1 X_2 \cdots X_n$ such that $X_i \in F(w_i)$ for each $i = 1,2, \cdots n$, and $x =>^* S$. The *language* $L(CG_R)$ accepted by $CG_R(VT, VA, S, F)$ is the set of all morpheme strings that are accepted.

I. NON-CONTEXT-FREE CATEGORIAL LANGUAGES

In this section we present a characterization theorem for the categorial systems that generate only context-free languages.

First, we introduce a lexicon F_{EQ} that we will show has the property that for any choice R of metarules any string in $L(CG_R)$ has equal numbers of a, b, and c. We define the lexicon F_{EQ} as $F_{EQ}(a) = \{A\}$, $F_{EQ}(b) = \{B\}$, $F_{EQ}(c) = \{C/A/C/B, C/D\}$, $F_{EQ}(d) = \{D\}$, $F_{EQ}(e) = \{S/A/C/B\}$.

We will also make use of two languages on the alphabet $\{a, b, c, d, e\}$ $L_1 = \{a^n db^n e c^n \mid n \geq 1\}$, and $L_{EQ} = \{w \mid \#a = \#b = \#c \geq 1, \#d = \#e = 1\}$.

A lemma shows that with any set R of rules the lexicon F_{EQ} yields a subset of L_{EQ}.

Lemma 1 Let G be any categorial grammar, $CG_R(VT, VA, S, F_{EQ})$, where $VT = \{a, b, c, d, e\}$, $VA = \{S, A, B, C, D\}$, with $R \subseteq \{F, FP, B, BP\}$. Then $L(G) \subseteq L_{EQ}$.

Proof Let $x = X_1X_2...X_n \Rightarrow^* S$. Let $w = w_1...w_n$ be a corresponding morpheme string. To differentiate between the occurrence of a symbol as a head and otherwise, write $C/A/C/B = CA^{-1}C^{-1}B^{-1}$, $S/A/C/B = SA^{-1}C^{-1}B^{-1}$ and $C/D = CD^{-1}$. For any rule system R, a redex is two adjacent categories, the tail of one matching the head of the other, and is reduced to a single category after cancelling the matching symbols. Since all occurrences of A must cancel to yield a reduction to S, $\#A = \#A^{-1}$. This holds for all atomic categories except S, for which $\#S = \#S^{-1}+1$. This lexicon has the property that any derivable category symbol, either has exactly one S and is S-headed or does not have an occurrence of S. Hence in x, $\#S = 1$, i.e., w has exactly one e. Let the number of occurrences in x of $C/A/C/B$ and C/D be p and q respectively. It follows that $\#C = p+q$, $\#C^{-1} = p+1$. Hence $q = 1$ and w has exactly one d. Each occurrence of $C/A/C/B$ introduces one A^{-1} and B^{-1}. Since w has one e, $\#A^{-1} = \#B^{-1} = p+1$. Hence $\#A = \#B = p+1$. Since for each A, B and C in x there must be exactly one a, b and c, $\#a = \#b = \#c$. □

We show next that in the restricted case where R contains only the two rules FP and B_S, the language L_1 is obtained.

Lemma 2 Let CG_R be the categorial grammar with lexicon F_{EQ} and rule set $R = \{FP, B_S\}$. Then $L(CG_R) = L_1$.

Proof Any $x \in L_1$ has a unique parse of the form $(B_S FP)^n B_S B_S{}^n$, and hence $L_1 \subseteq L(CG_R)$. Conversely, any x having a parse must have exactly one e. Further, all b's and c's can appear only on the left and right of e respectively. Any derivable category having an A has the form $S/(A/)^n U$ where U does not have any A. Thus all A's appear consecutively on the left of the e. For the rightmost c, $F(c) = C/D$. A d must be in between a's and b's. By lemma 1, $\#(a) = \#(b) = \#(c)$. Thus $x = a^n db^n ec^n$, for some n. Hence $L_1 = L(CG_R)$. □

The next lemma shows that no language intermediate to L_1 and L_{EQ} can be context-free. It really does not involve categorial grammar at all.

Lemma 3 If $L_1 \subseteq L \subseteq L_{EQ}$, then L is not context-free.

Proof Suppose L is context-free. Since L contains L_1, it has arbitrarily long strings of the form $a^n b\ db^n e\ c^n$. Let k and K be pumping lemma constants. Choose $n > \max(K, k)$. This string, if pumped, yields a string not in L_{EQ}, hence we have a contradiction. □

Corollary Let $\{FP, B_S\} \subseteq R$. Then there is a non-context-free language $L(CG_R)$.

Proof Use the lexicon F_{EQ}. Then by lemma 1 $L(CG_R) \subseteq L_{EQ}$. But $\{FP, B_S\} \subseteq R$, so $L_1 \subseteq L(CG_R)$. □

The following theorem summarizes the results by characterizing the rule sets that can be used to generate context sensitive languages.

Main Theorem A categorial system with rule set R can generate a context-sensitive language if and only if R contains a partial combination rule and a combination rule in the reverse direction.

Proof The "if" part follows for $\{FP, B_S\}$ by lemmas 1, 2, and 3. It follows for $\{BP, F\}$ by symmetry. For the "only if" part, first note that any unidirectional system (system with rules that are all forward, or all backward) can generate only context-free languages.[5] The only remaining cases are $\{F, B\}$ and $\{FP, BP\}$. The first generates only context free languages.[5] The second generates only the empty language, since no atomic symbol can be derived using only these two rules.

II. CATEGORIAL LANGUAGES ARE PERMUTATIONS OF CONTEXT-FREE LANGUAGES

Let $VT = \{a_1, a_2, \cdots, a_k\}$. A Parikh mapping[6] Ψ is a mapping from morpheme strings to vectors such that $\Psi(w) = (\#a_1, \#a_2, \ldots, \#a_k)$. u is a permutation of v iff $\Psi(u) = \Psi(v)$. Let $\Psi(L) = \{\Psi(w) \mid w \in L\}$. A language L is a permutation of $\bar{L}$ iff $\Psi(L) = \Psi(\bar{L})$. We define a rotation as follows. In the parse tree for $u \in L$, at any node corresponding to a B redex or BP-redex exchange its left and right subtrees, obtaining an F-redex or an FP-redex. Let v the resulting terminal string. We say that u has been transformed into v by rotation.

We now obtain results that are helpful in showing that certain languages cannot be generated by categorial grammars. First we show that, every categorial language is a permutation of a context-free language. This will enable us to show that properties of context-free languages that depend only on the symbol counts must also hold of categorial languages.

Theorem Let R ⊆ {F, FP, B, BP}. Then there exists a L_{CF} such that $\Psi(L(CG_R)) = \Psi(L_{CF})$, where L_{CF} is context free.

Proof Let $x \in L(CG_R)$. In its parse tree at each node corresponding to a B-redex or a BP-redex perform a rotation, so that it becomes a F-redex or a FP-redex. Since the transformed string y is obtained by rearranging the parse tree, $\Psi(x) = \Psi(y)$. Also y derivable using $R_1 = \{FP, F\}$ only. Hence the set of such y obtained as a permutation of some x is the same as $L(CG_{R_1})$, which is context free,[5] i.e., $L(CG_{R_1}) = L_{CF}$. □

Corollary For any R $\subseteq$ {F, FP, B, BP}, $L(CG_R)$ is semilinear , Parikh bounded and has the linear growth property.

Semilinearity follows from Parikh's Lemma and linear growth from the pumping lemma for context-free languages. Parikh boundedness follows from the fact that any context-free language is Parikh bounded.[6] □

Proposition Any one-symbol categorial grammar is regular.

Note that if L is a semilinear subset of nonnegative integers, $\{a^n \mid n \in L\}$ is regular.

III. NON-CATEGORIAL LANGUAGES

We now exhibit some non-categorial languages and compare categorial languages with others. From the corollary of the previous section we have the following results.

Theorem Categorial languages are properly contained in the context-sensitive languages.

Proof The languages $\{a^{h(n)} \mid n \geq 0\}$, where $h(n)=n^2$ or $h(n)=2^n$ which do not have linear growth rate, are not generated by any CG_R. These are context sensitive. Also $\{a^m b^n \mid$ either $m > n$, or m is prime and $n \leq m$ and m is prime $\}$ is not semilinear [7] and hence not categorial.

It is interesting to note that lexical functional grammar can generate the first two languages mentioned above[8] and indexed languages can generate $\{a^n b^{n^2} a^n \mid n \geq 1\}$.

Linguistic Properties

We now look at some languages that exhibit cross-serial dependencies.

Let G_3 be the CG_R with $R = \{FP, B_S\}$, $VT = \{a, b, c, d\}$, and with the lexicon $F(d) = \{S/S_1\}$, $F(c) = \{S_1\}$, $F(a) = \{S_1/A/S_1, A\}$, $F(b) = \{S_1/B/S_1, B\}$. Then $L_3 = L(G_3) = \{wcdw \mid w \in \{a, b\}^*\}$. The reasoning is similar to that of lemma 1. First $\#c = \#d = 1$, from $\#S = 1$. Since we have B_S rule, c occurs on the left of d and all occurrences of a and b on the left of c get assigned A and B respectively. Similarly all a and b on the right of c, get assigned to the complex category as defined by F. It follows that all symbols to the right of d get combined by FP rule and those on the left by B_S rule. Hence a symbol occurring n symbols to the right of d must be matched by an occurrence n symbols to the right of the left-most symbol.

For any k, let $G_4(k)$ be the CG_R with $R = \{FP, B_S\}$ again, $VT = \{a_i, b_i \mid 1 \leq i \leq k\} \cup \{c_i \mid 1 \leq i < k\} \cup \{d, e\}$, and the lexicon $F(b_i) = \{S_i/A_i/S_i\}$, $F(a_i) = \{A_i\}, 1 \leq i \leq k$, $F(c_i) = \{S_i/S_{i+1}\}, 1 \leq i < k$, $F(d) = \{S_k\}$, $F(e) = \{S/S_1\}$. Then $L(G_4(k)) = \{a_1^{n_1} a_2^{n_2} \cdots a_k^{n_k} d e b_1^{n_1} c_1 \cdots c_{k-1} b_k^{n_k}\}$ for any k. Note that $\#A_i = \#A_i^{-1}$. This implies $\#b_i = \#a_i$. The rest of the argument parallels that for L_3 above . Thus {FP, B_S } has the power to express unbounded cross-serial dependencies.

Now we can compare with Tree Adjoining Grammars (TAG).[8] A TAG without local constraints cannot generate L_3. A TAG with local constraints can generate this, but it cannot generate $L_6 = \{a^m b^n c^m d^n \mid m, n \geq 1\}$. $L_4(2)$ can be transformed into L_6 by the homomorphism erasing c_1, d and e. TAG languages are closed under homomorphisms and thus the categorial language $L_4(2)$ is not a TAG language. TAG languages exhibit only limited cross serial dependencies. Thus, though TAG Languages and CG languages share some properties like linear growth, semilinearity, generation of all context-free languages, limited context sensitive power, and Parikh boundedness, they are different in their generative capacities.

Acknowledgements We would like to thank Weiguo Wang and Dawei Dai for helpful discussions.

References

1. Yehoshua Bar-Hillel, "On syntactical categories," *Journal of Symbolic Logic*, vol. 15 , pp. 1-16 , 1950. Reprinted in Bar-Hillel (1964), pp. 19-37.
2. Haim Gaifman, *Information and Control*, vol. 8, pp. 304-337, 1965.
3. Yehoshua Bar-Hillel, *Language and Information*, Addison-Wesley, Reading, Mass., 1964 .
4. Anthony E. Ades and Mark J. Steedman, "On the order of words," *Linguistics and Philosophy*, vol. 4, pp. 517-558, 1982.
5. Joyce Friedman, Dawei Dai, and Weiguo Wang, "Weak Generative Capacity of Parenthesis-free Categorial Grammars," Technical Report #86-1, Dept. of Computer Science, Boston University, 1986.
6. Meera Blattner and Michel Latteux, "Parikh-Bounded Languages," in *Automata, Languages and Programming, LNCS 115*, ed. Shimon Even and Oded Kariv, Springer-Verlag, 1981.
7. Harry R. Lewis and Christos H. Papadimitriou, *Elements of the Theory of Computation*, Prentice-Hall, 1981.
8. Aravind K. Joshi, "Factoring recursion and dependencies: an aspect of tree adjoining grammars and a comparison of some formal properties of TAGs, GPSGs, PLGs and LFGs," *21st Ann. Meeting of the Assn. for Comp. Linguistics*, 1983.