Proceedings of the First International Symposium: Category Theory Applied to Computation and Control, published by the Mathematics Department and the Department of Computer and Information Science, University of Massachusetts at Amherst, 1974.

AN ALGEBRAIC FORMULATION OF THE CHOMSKY HIERARCHY

Mitchell Wand
Computer Science Department

Indiana University
Bloomington, Indiana 47401, U.S.A.

In the classic paper [2], Chomsky introduced a hierarchy of types of grammars: regular, context-free, context-sensitive, phrase-structure. This was an ad hoc list. More recently, the results of [1] and [8] have suggested that there is a deep algebraic relationship between the regular and context-free sets, and that a similar relationship holds between the context-free and indexed languages [6]. Thus the indexed languages may be taken as the natural next step in the hierarchy.* In this paper we define a sequence of operations E_k on theories such that for a certain theory T, the constants in $E_k(T)$ for $k = 0, 1, 2$ are precisely the regular, context-free, and indexed sets. This proves the conjecture of [9, p. 59] and [10, pp. 182-183].

---

*This view is held even by authors not normally associated with algebraic techniques, e.g. [4].

1. Definitions

We will use the properties of Cartesian closed categories, and the results of [5] in particular. Let $\underline{CL}$ denote the Cartesian closed category of small complete lattices with morphisms continuous over nonvoid directed chains. Let the usual isomorphism $\underline{CL}(L \times M, N) \to \underline{CL}(L, N^M)$ be denoted φ.

Let 1 be the singleton set, as usual. Define sets T_k of strings over the alphabet $\{0, 1, x, \to\}$ as follows:

(i) $0, 1 \in T_0$

(ii) if $t, u \in T_k$ then $*tu \in T_k$

(iii) if $t \in T_k$, then $t \in T_{k+1}$

(iv) if $t, u \in T_k$, then $\to tu \in T_{k+1}$

(v) nothing else.

Note $\cup\{T_k \mid k \in \omega\}$ is the set of objects of the free Cartesian closed category generated by 1, in which 0 is the terminal object and $*$ and $\to$ have their usual meanings. Let $\underline{I}_k$ denote the full subcategory with objects T_k.

Definition: A k-theory is a functor $F: \underline{I}_k \to \underline{CL}$ preserving product and exponentiation.

Any algebraic theory has a unique extension to a k-theory $T_k(X)$, which may be built "from below" i.e., without considering objects of higher type. We often identify k-theories with their images in $\underline{CL}$.

Let $Y \in \underline{CL}(M^M, M)$ be the morphism sending each continuous function $M \to M$ to its least fixed point. Let $\underline{A}$ be a subcategory of $\underline{CL}$, closed under finite products. We say $\underline{A}$ is iteration-closed iff for every $g \in A(L \times M, M)$, the morphism $Y.\varphi(g)$ belongs to $\underline{A}(L, M)$. If $\underline{A}$ is a subcategory of $\underline{CL}$ with finite products, then $\underline{A}$ has an iteration-closure $\mu(\underline{A})$.

Definition: A k-system is the iteration-closure of a k-theory

Thus every theory X has a unique extension to a k-system $\mu(T_k(X))$. We call this system $E_k(X)$.

2. Results

Theorem 1. Let A be any subcategory of CL with finite products. Then every morphism in $\mu(A)$ is of the form $E.Y.\varphi(\Pi g_i)$, where E is a projection, Y is the fixed point morphism, and each g_i is a morphism in A.

This is a straightforward application of the normal form theorem for μclones [10].

Theorem 2. (Chomsky theorem) Let $k>0$, let $L \in CL$ and for $n \geqslant 0$ let Ω_n be a set of morphisms $L^n \to L$. Let T be the theory generated by $\Omega = \coprod \Omega_n$. Then every morphism in $E_k(T)$ is of the form $E.Y.\varphi(\Pi g_i)$, where E and Y are as before, and each g_i is of the form $t_i.E_i$ where E_i is a product of projections and t_i is either

(i) $\varphi(u)$ where $u \in \Omega$

(ii) $comp \in CL([x \to y]*[y \to z], [x \to z])$

(iii) $eval \in CL([x \to y]*x, y)$

(iv) $\underline{\varphi} \in CL([x*y \to z], [x \to [y \to z]])$

Sketch of proof. We first apply theorem 1. We then show that, without loss of generality, the g_i may be taken to be of the form $E_i.\varphi(u_i)$. We then apply the cut-elimination lemma of [9] to the u_i. This sufficiently simplifies the u_i so that a case analysis becomes feasible. We may then transform each u_i into the proper form.

Let V be a countable set (of terminal symbols), and let V* be the set of strings over V. Let $L = P(V^*)$. Let $\Omega_0 = \{\{e\}\}$ (the empty string), $\Omega_1 = \{\lambda S[aS] \mid a \in V\}$ and let $\Omega_2 = \{\lambda ST[SUT]\}$. Let T be

the theory generated by Ω, and $S_k = E_k(T)$. Then $S_k(0, L)$ is a class of languages for each k.

Theorem 3. For $k = 0, 1, 2$, $S_k(0, L)$ is the class of regular, context-free, and indexed languages.

Sketch of proof. For each class $\mathcal{L}_k$, showing $\mathcal{L}_k \subseteq S_k(0, L)$ is tedious but routine. In the other direction $k = 0$ was shown in [10]. For $k = 1, 2$, we apply Theorem 2 in a weakened form, treating unions specially. Then no nontrivial multi-place functions develop, and we arrive at the Chomsky normal form theorem and Fischer's OI normal form theorem [3], respectively.

3. Conclusions and Open Problems

Our results confirm the suggestion that the (revised) Chomsky hierarchy is of algebraic, rather than ad hoc, origin. They also suggest some new problems. Of particular interest are the characterization of $S_\omega(0, L)$ and the question of strict hierarchy problem (i.e., is $S_k(0, L) \neq S_{k+1}(0, L)$ for all k ?). In addition, the algebraic properties of E_k have not been studied. These become particularly relevant when L is the lattice of flow diagrams [7]: in this case we construct classes of flow diagrams which seem related to the theory of program schemes.

(Added in proof). Systems similar to S_k were studied by Turner (Doctoral Dissertation, University of London, 1973). When our base lattice T(1) is a power set, our systems become a special case of equational sets over a many-typed algebra, studied by Maibaum ("a Generalized Approach to Formal Languages", JCSS 8 (1974), 409-439).

REFERENCES

1. Brainerd, W.S., Tree Generating Regular Systems, Information and Control 14 (1969) 217-231.

2. Chomsky, N., On certain formal properties of grammars, Information and Control 2: 2 (1959), 137-167.

3. Fischer, M.J., Grammars with Macro-Like Productions, Proc 9th IEEE Conf Sw & Auto Th (1968), 131-142.

4. Greibach, S.A., Full AFL and Nested Iterated Substitution, Information and Control 16 (1970), 7-35.

5. Lambek J., Deductive Systems & Categories II, Category Theory, Homology Theory & Their Applications I (P. Hilton, ed.) Berlin: Springer-Verlag, Lecture Notes in Mathematics, Vol. 86, (1969), 76-122.

6. Rounds, W.C., Mappings and Grammars on Trees, Math Sys Th 4 (1970), 257-287.

7. Scott, D. The Lattice of Flow Diagrams, Oxford U. Comp. Lab. Rep. PRG-3 (1970).

8. Thatcher, J.W., Characterizing Derivation Trees of Context-Free Grammars through a Generalization of Finite Automata Theory, J Comp & Sys Sci 1 (1967), 317-322.

9. Wand, M. An Usual Application of Program-Proving Proc 5th ACM Symp on Th of Computing (Austin, 1973), 59-66.

10. Wand, M. Algebraic Foundations of Formal Language Theory, MIT Project MAC TR-108, Cambridge, Mass., 1973.