MÉMOIRES DE LA S. M. F.

WOLFGANG THOMAS

An application of the Ehrenfeucht-Fraisse game in formal language theory

Mémoires de la S. M. F. 2^e série, tome 16 (1984), p. 11-21

<http://www.numdam.org/item?id=MSMF_1984_2_16__11_0>

Société Mathématique de France
2° série, mémoire n° 16, 1984, p. 11-21

AN APPLICATION OF THE EHRENFEUCHT-FRAISSÉ GAME IN FORMAL LANGUAGE THEORY

Wolfgang Thomas

Abstract A version of the Ehrenfreucht-Fraissé game is used to obtain a new proof of a hierarchy result in formal language theory: It is shown that the concatenation hierarchy ("dot-depth hierarchy") of star-free languages is strict.

Résumé Une version du jeu de Ehrenfeucht-Fraissé est appliquée pour obtenir une nouvelle preuve d'un théorême dans la théorie des langages formels: On montre que la hiérarchie de concaténation ("dot-depth hierarchy") des langages sans étoile est stricte.

1. Introduction.

The present paper is concerned with a connection between formal language theory and model theory. We study a hierarchy of formal languages (namely, the dot-depth hierarchy of star-free regular languages) using logical notions such as quantifier complexity of first-order sentences. In this context we apply a form of the Ehrenfeucht-Fraissé game which serves to establish the elementary equivalence between structures with respect to sentences of certain prefix types.

The class of star-free regular languages is of a very basic nature: It consists of all languages (= word-sets) over a given alphabet A which can be obtained from the finite languages by finitely many applications of boolean operations and the concatenation product. (For technical reasons we consider only nonempty words over A , i.e.

0037-9484/84 03 11 11/$ 3.10/

languages $L \subset A^+$; in particular, the complement operation is applied w.r.t. A^+.) General references on the star-free regular languages are McNaughton-Papert (1971), Chapter IX of Eilenberg (1976), or Pin (1984b).

A natural classification of the star-free regular languages is obtained by counting the "levels of concatenation" which are necessary to build up such a language: For a fixed alphabet A, let

$$\mathcal{B}_0 = \{L \subset A^+ \mid L \text{ finite or cofinite}\},$$

$$\mathcal{B}_{k+1} = \{L \subset A^+ \mid L \text{ is a boolean combination of languages of the form } L_1 \cdot \ldots \cdot L_n (n \geq 1) \text{ with } L_1, \ldots, L_n \in \mathcal{B}_k\}.$$

The language classes $\mathcal{B}_0, \mathcal{B}_1, \ldots$ form the so-called dot-depth hierarchy (or: Brzozowski hierarchy), introduced by Cohen/Brzozowski (1971). In the framework of semigroup theory, Brzozowski/Knast (1978) showed that the hierarchy is infinite (i.e. that $\mathcal{B}_k \supsetneq \mathcal{B}_{k-1}$ for $k \geq 1$). The aim of the present paper is to give a new proof of this result, based on a logical characterization of the hierarchy that was obtained in Thomas (1982). The present proof does not rely on semigroup-theory; instead, an intuitively appealing model-theoretic technique is applied: the Ehrenfeucht-Fraissé game.

Let us first state the mentioned characterization result, taking $A = \{a,b\}$. One identifies any word $w \in A^+$, say of length n, with a "word model"

$$w = (\{1,\ldots,n\}, <, \min, \max, S, P, Q_a, Q_b)$$

where the domain $\{1,\ldots,n\}$ represents the set of positions of letters in the word w, ordered by $<$, where min and max are the first and the last position, i.e. $\min = 1$ and $\max = n$, S and P are the successor and predecessor function on $\{1,\ldots,n\}$ with the convention that $S(\max) = \max$ and $P(\min) = \min$, and Q_a, Q_b are unary predicates over $\{1,\ldots,n\}$ containing the positions with letter a, b respectively. (Sometimes it is convenient to assume that the position-sets of two words u,v are disjoint; then one takes any two nonoverlapping segments of the integers as the position-sets of u and v.) Let L be the first-order language with equality and nonlogical symbols $<, \min, \max, S, P, Q_a, Q_b$. Then the satisfaction of an L-sentence φ in a word w

(written: $w \models \varphi$) can be defined in a natural way, and we say that $L \subset A^+$ is defined by the L-sentence φ if $L = \{w \in A^+ \mid w \models \varphi\}$.

For example, the language $L = (ab)^+$ is defined by

$$Q_a\min \wedge Q_b\max \wedge \forall y(y < \max \rightarrow (Q_a y \leftrightarrow Q_b S(y))) .$$

As usual, a Σ_k-formula is a formula in prenex normal form with a prefix consisting of k alternating blocks of quantifiers, beginning with a block of existential quantifiers. A $B(\Sigma_k)$-formula is a boolean combination of Σ_k-formulas.

1.1 Theorem. (Thomas (1982)). Let $k > 0$. A language $L \subset A^+$ belongs to $\mathcal{B}_k$ iff L is defined by a $B(\Sigma_k)$-sentence of L.

For the formalization of properties of words the symbols min,max,S,P are convenient. But of course they are definable in the restricted first-order language L_0 with the nonlogical constants $<, Q_a, Q_b$ alone. Indeed, we have:

1.2 Lemma. Let $k > 0$. If $L \subset A^+$ is defined by a $B(\Sigma_k)$-sentence of L, then L is defined by a $B(\Sigma_{k+1})$-sentence of L_0.

Proof. The quantifier-free kernel of a Σ_k-formula φ of L can be expressed both by a Σ_2- and a Π_2-formula of L_0. For example, $Q_a S(\min)$ is expressible in the following two ways:

(+) $\exists y(y = S(\min) \wedge Q_a y)$, $\forall y(y = S(\min) \rightarrow Q_a y)$

where $y = S(\min)$ is rewritten as a Π_1-formula of L_0 using

$$x = \min \leftrightarrow \forall z(x = z \vee x < z), \quad x = \max \leftrightarrow \forall z(z = x \vee z < x)$$
$$S(x) = y \leftrightarrow (x = \max \wedge x = y) \vee (x < y \wedge \forall z \neg (x < z \wedge z < y)).$$

Hence we obtain a Σ_{k+1}-sentence of L_0 which is equivalent (in all word-models) to φ by applying one of the two definitions in (+), depending on the case whether the innermost quantifier-block of φ is existential or universal.

We mention without proof that (for $k > 0$) the $B(\Sigma_k)$-sentences of L_0 define exactly those languages $L \subset A^+$ which occur on the k-th level of another hierarchy of star-free regular languages, introduced by

Straubing (1981). For details concerning the Straubing hierarchy and its relation to the Brzozowski hierarchy cf. Pin (1984a,b). The proof to be given below also shows that the Straubing hierarchy is infinite.

2. The Example Languages

In order to show that $\mathcal{B}_k \supsetneq \mathcal{B}_{k-1}$ for $k \geq 1$, we introduce "example languages" L_k, L_k^+, L_k^- over $A = \{a,b\}$.

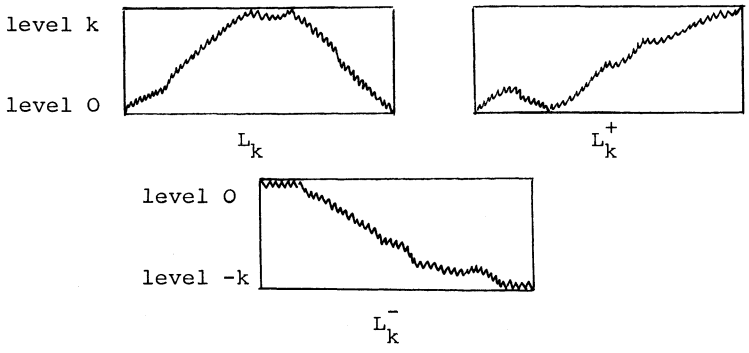
Let $|w|_a$ (resp. $|w|_b$) denote the number of occurrences of the letter a (resp. b) in w, and define the weight $\|w\|$ of a word w by

$$\|w\| = |w|_a - |w|_b \ .$$

In the sequel we write $v \sqsubseteq w$ if the word v is an initial segment (left factor) of w. Let

$$L_k = \{w \in A^+ \mid \|w\| = 0,\ \forall v \sqsubseteq w\ 0 \leq \|v\| \leq k,\ \exists v \sqsubseteq w\ \|v\| = k\} \ ,$$
$$L_k^+ = \{w \in A^+ \mid \|w\| = k,\ \forall v \sqsubseteq w\ 0 \leq \|v\| \leq k\} \ ,$$
$$L_k^- = \{w \in A^+ \mid \|w\| = -k,\ \forall v \sqsubseteq w\ -k \leq \|v\| \leq 0\} \ .$$

To obtain a more intuitive picture of these languages, it is useful to represent the letter a by the stroke ↗ and b by ↘. Then the word abababa, for example, is represented by ∧∧∧. Thus L_k contains all words whose "graph" has the following properties: It ends on the same level where it starts ("level 0"), it is confined to level 0 and the next k levels, and it assumes the k-th level at least once. Similarly for L_k^+, L_k^-. The "typical shape" of words in L_k, L_k^+, L_k^- is indicated in the following diagrams:



L_k L_k^+ L_k^-