JOHANNES ARZ

Syntactic congruences and syntactic algebras

SYNTACTIC CONGRUENCES
AND SYNTACTIC ALGEBRAS (*)

by Johannes ARZ ([1])

Communicated by J. BERSTEL

Résumé. — On prouve quelques propositions sur les congruences syntactiques et les algèbres syntactiques. On définit la notion de la séparabilité des langages. Les résultats obtenus nous donnent des relations entre les algèbres syntactiques, introduites par C. Reutenauer, et les algèbres des monoïdes syntactiques, introduites par E. Shamir et G. Hotz.

Enfin, en analysant la construction d'une version non-déterministique d'un langage, on trouve une relation supplémentaire entre les algèbres syntactiques et les monoïdes syntactiques.

Abstract. — This paper gives new results about syntactic congruences and algebras. Especially, the notion of separability makes clear the connection between the syntactic algebra, defined by C. Reutenauer, and the algebra of the syntactic monoid, defined by E. Shamir and G. Hotz.

Further, the construction of a nondeterministic version of any language gives another relation between syntactic algebras and syntactic monoids.

INTRODUCTION

It has been introduced by E. Shamir [7] a monoid algebra $R[M]$ over a ring R, where M is the set of all so far as possible cancelled expressions of nested parentheses. He proved, that all context free languages are of the form $L = v^{-1}(K_\gamma)$ for a homomorphism v and a coset K_γ in $R[M]$.

G. Hotz has heighted this results by dealing with $\mathbb{B}[M_D]$, where $\mathbb{B}$ is the Boolean semiring and M_D the syntactic monoid of the Dyck language. His main idea was the use of zero divisors computing in his algebra. He called $\mathbb{B}[M_D]$ the syntactic algebra [3, 4].

Later, C. Reutenauer has defined an algebra over a field K by introducing a congruence, induced by a given power series, on $K[X]$, the set of polynomials over K, which he called syntactic algebra, too. In this way, he advanced the theory of formal power series using high sophisticated algebraic technics [5].

In this paper, we compare both concepts by regarding them as algebras over a semiring R and dealing with the characteristic power series of a language.

We call $A_L = R[M_L]$ the syntactic algebra and $\mathscr{A}_S = R[X]$ modulo the congruence induced by char (L) the characteristic algebra. Our results are:

— $\mathscr{A}_S$ is a quotient of A_L, for any semiring R and language L;

— we define the property of a language to be (syntactically) separable and prove: L separabel $\Rightarrow \mathscr{A}_S \cong A_L$ for any R;
the converse is true for $R = \mathbb{B}$;

— $\mathscr{A}_S$ is the syntactic monoid of a language $\tilde{L}$ related to some submonoid, where $\tilde{L}$ is the nondeterministic version of L.

— $\mathscr{A}_S$ is a submonoid of the syntactic monoid of $\tilde{L}$.

1. NOTATIONS AND DEFINITIONS

Let X be a finite alphabet and X^* the free monoid generated by X with neutral element 1. For each language $L \subset X^*$, the syntactic monoid M_L is the quotient of X^* modulo the syntactic congruence $\sim_L$ which is defined (*see* [2]):
for all w, $w' \in X^*$:

$$w \sim w' \quad \Leftrightarrow \quad \text{for all contexts } u, v \in X^* \text{ holds} \quad (uwv \in L \Leftrightarrow uw'v \in L).$$

The monoid homomorphism $\sigma : X^* \to M_L : = X^*/\sim_L$, $w \to \bar{w}$, is called the syntactic morphism.

We define $\overline{0} : = \{ w \in X^*/\text{for all } u, v \in X^* : uwv \notin L \}$.

A generalisation of this concept is the restriction of X^* to any submonoid M of X^*. The syntactic monoid of a language L related to some submonoid $M \subset X^*$ is (*see* [2]):

$$\text{Syn}(M, L) : = \{ [x]/x \in M \}$$

and:

$$y \in [x] \quad \Leftrightarrow \quad \text{for all contexts } u, v \in M : (uxv \in L \Leftrightarrow uyv \in L).$$

By definition: $\text{Syn}(X^*, L) = M_L$.

It is easy to see that $\text{Syn}(M, L)$ divides $\text{Syn}(X^*, L)$, that is: There exists a submonoid $\tilde{S} \subset \text{Syn}(X^*, L)$ and a monoid epimorphism $\tilde{S} \to \text{Syn}(M, L)$.

Let R denotes any semiring. For a monoid M, $R[M]$ is the set of all polynomials in M with coefficients in R:

$$R[M] : = \{ p = \sum a_w w/w \in M, a_w \in R, a_w = 0 \text{ for allmost all } w \}$$

With addition:

$$\sum a_w\, w + \sum b_w\, w = \sum (a_w + b_w)\, w$$

and multiplication:

$$\sum a_w\, w \cdot \sum b_w\, w = \sum_{w} \left(\sum_{w=uv} a_u\, b_v \right) w,$$

$R[M]$ is a monoid algebra.

In this monoid algebra we identify the zero in R and the zero in the monoid (if it exists): $0_R = 0_M$, that is, we always deal with the contracted algebra of a monoid (*see* [1]).

We shall often use the following theorem:

LEMMA 0: *Let* A, B, C *be monoids (monoid algebras). Let* $f : A \to B$, $g : A \to C$ *monoid epimorphism (monoid algebra epimorphism):*

1) *There exists a canonical monoid (algebra) epimorphism* $h : B \to C$, *such that* $h \circ f = g$, *iff* $f(a) = f(a')$ *implies* $g(a) = g(a')$ *for all* $a, a' \in A$;

2) h *is an isomorphism iff* $f(a) = f(a') \Leftrightarrow g(a) = g(a')$.

Proof: Define $h : B \to C$ with $h(b) : = g(a)$ if $f(a) = b$.

2. SYNTACTIC CONGRUENCES

Let X be a finite alphabet and $L \subset X^*$.

DEFINITION 1: We write $R\langle X \rangle$ instead of $R[X^*]$ and call $R\langle X \rangle$ the *algebra of polynomials* over R.

A series $S = \sum s_w\, w$, where $s_w \in R$ and $w \in X^*$, is called a formal power series in X with coefficients in R.

The characteristic power series of a formal language $L \subset X^*$ is $S = \sum s_w\, w$, where $s_w = 1$ iff $w \in L$ and $s_w = 0$ elsewhere. It is sometimes denoted by $\underline{L}$.

We define $(S, .) : R\langle X \rangle \to R$:

$$p \mapsto (S, p) = \left(\sum s_w\, w, \sum a_w\, w \right) = \sum s_w\, a_w \in R.$$

This sum is finite, because p is a polynomial.

DEFINITION 2: Let S be a formal power series on X. We define a congruence on $R\langle X \rangle$ by:

$$p \equiv q \quad \Leftrightarrow \quad (S, rpr') = (S, rqr') \quad \text{for all } r, r' \in R\langle X \rangle.$$

We give some properties and show, that $\equiv$ is really a congruence satisfying the two operations in $R \langle X \rangle$:

LEMMA 1: *For all p, $q \in R \langle X \rangle$, we have:*

1) $p \equiv q$, $p' \equiv q' \Rightarrow p + p' \equiv q + q'$, $p \cdot p' \equiv q \cdot q'$;

2) $p \equiv q \Leftrightarrow$ *for all u, $v \in X^*$: $(S, upv) = (S, uqv)$;*

3) *for all w, $w' \in X^*$: $w \equiv w' \Leftrightarrow w \sim w'$.*

Proof: 1) is easy to see;

2) follows because of 1) and the linearity of $(S, .)$;

3) $w \sim w' \Leftrightarrow$ for all u, $v \in X^*$: $uwv \in L \Leftrightarrow uw'v \in L$;

$\qquad\qquad \Leftrightarrow$ for all u, $v \in X^*$: $(S, uwv) = 1 \Leftrightarrow (S, uw'v) = 1$;

$\qquad\qquad \Leftrightarrow$ for all u, $v \in X^*$: $(S, uwv) = (S, uw'v)$;

$\qquad\qquad \Leftrightarrow w \equiv w'$.

DEFINITION 3: We define $\mathscr{A}_S : = R \langle X \rangle / \equiv$ and call $\mathscr{A}_S$, remembering that it depends on the characteristic power series of a language, the *characteristic algebra of L* and the algebra homomorphism $\chi : R \langle X \rangle \to \mathscr{A}_S$ the characteristic homomorphism of L.

The syntactic morphism $\sigma : X^* \to M_L$ extends to an algebra homomorphism (also called) $\sigma : R \langle X \rangle \to R [M_L]$, defined by linear extension:

$$\sigma : \quad p = \sum a_w w \mapsto \bar{p} = \sum a_w \sigma(w) = \sum a_w \bar{w}.$$

DEFINITION 4: We write A_L instead of $R [M_L]$ and call:

$$A_L : = \left\{ \sum a_{\bar{w}} \bar{w} / \bar{w} \in M_L, a_{\bar{w}} \in R, a_{\bar{w}} = 0 \text{ for allmost all } w \right\},$$

the *syntactic algebra of L*.

An obvious question is: Is there a relation between the characteristic algebra and the syntactic algebra of a language? A first answer is:

LEMMA 2: *$\mathscr{A}_S$ is a quotient of A_L. There exists a monoid algebra epimorphism $\mathscr{C} : A_L \to \mathscr{A}_S$.*

Proof: Because A_L is a free R monoid, we may define $\mathscr{C}$ on the basis $\bar{w} \in M_L$:

$$\bar{w} \mapsto \chi(w) \qquad \text{if} \quad w \in \sigma^{-1}(\bar{w}).$$

Using Lemma 1,3 we see that $\mathscr{C}$ is a well defined algebra epimorphism.

3. SYNTACTICAL SEPARABILITY

Let $L \subset X^*$ and S the characteristic power series of L.

Let $R = \mathbb{B}$, the Boolean semiring with elements 0 and 1.

DEFINITION 5: A language L is (*syntactically*) *separable* iff for all $w \in X^*$, $w \notin \overline{0}$, exists a context $u, v \in X^*$ such that: (a) $uwv \in L$; (b) for all $w' \in X^* : uw'v \in L \Rightarrow w \sim w'$.

In other words, L is separable, if for each class $\overline{w} \neq \overline{0}$ a special context exists that only takes the class $\overline{w}$ into the language and no other class.

Some examples may help for a better understanding.

Examples:

1) the Goldstine language is separable:
$$X = \{a, b\}; \qquad L = \{abaaba^3 b \dots a^n b / n \geq 1\};$$

2) $X = \{a\};$ $L_1 = a^+ = \{a^i / i \geq 1\}$ is not separable; $L_2 = \{a^{2i} / i \geq 0\}$ is separable;

3) $X = \{a, b, c\};$ $L = \{aa, ba, ac, bc\}$ is not separable;

4) $L = D$, the Dyck language is not separable;

5) let $\mathscr{C}$ a morphism on a group, then $\mathscr{C}^{-1}(1)$ is separable.

Proof: Hints: In 1) observe that, for instance, choosing $w = aba$ and $w' = abaabaa$, $w \sim w'$ and $wab \in L$ and $w'ab \in L$, but $waba^3 b \in L$ and $w'aba^3 b \notin L$. In 4) choose $w =))(($ and $w' =)($, for example.

LEMMA 3: *For all* $p = w_1 + \dots + w_n$, $q = w'_1 + \dots + w'_m \in \mathbb{B}\langle X \rangle$: $p \equiv q \Leftrightarrow$ *for all* $u,$ $v \in X^*$ *the following hold:* $uw_i v \in L$ *for some* $1 \leq i \leq n \Leftrightarrow uw'_j v \in L$ *for some* $1 \leq j \leq m.$

Proof: Follows from the fact that:
$$(S, upv) = 1 \quad \Leftrightarrow \quad uw_i v \in L \text{ for some } 1 \leq i \leq n.$$

Now we give the main result in this section.

THEOREM 1: L *separable* $\Leftrightarrow A_L \cong \mathscr{A}_S;$ *that is, the algebra homomorphism* $\mathscr{C}$ *defined in Lemma 2 is bijective.*

Proof: " $\Rightarrow$ " Using Lemma 2 it remains to show:
$$\chi(p) = \chi(q) \quad \Rightarrow \quad \sigma(p) = \sigma(q) \text{ for all } p, q \in \mathbb{B}\langle X \rangle.$$

Let $p = \sum a_i w_i,$ $q = \sum b_j w'_j;$ $a_i, b_j \in \mathbb{B};$ $a_i, b_j \neq 0.$

We may assume that $w_i, w'_j \notin \overline{0}$ for each $i, j.$

Let us fix some $w_i.$

Claim: there exists $j \in \{1, \dots, m\} : w_i \sim w'_j.$

Proof of the claim: From the definition of separability exists a special context u, $v \in X^*$ such that $uw_i v \in L$ and from $\chi(p) = \chi(q)$ and Lemma 3 exists $j \in \{1, \ldots, m\}$: $uw'_j v \in L$. It follows $w_i \sim w'_j$.

We have shown: To each summand w_i in p exists a summand w'_j in q such that $\bar{w}_i = \bar{w}'_j$ and vice versa.

Thus $\sigma(p) = \sigma(q)$.

" $\Leftarrow$ " L being not separable means: there exists $w \in X^*$ such that for all u, $v \in X^*$: $uwv \notin L$ or there exists $w' \in X^*$ with the property $uw'v \in L$ and $w' \not\sim w$.

It follows $\bar{w} + \bar{w}' \neq \bar{w}'$, but for all u, $v \in X^*$:

$$(S, u(w + w')v) = (S, uwv) + (S, uw'v) = (S, uw'v).$$

REMARK: Let us assume for a moment, that $\mathbb{B}$ is any semiring R. It can be proved by similar arguments that L separable is a sufficient condition for $A_L \cong \mathscr{A}_S$, but in contrary not a necessary condition: choose $R = \mathbb{Q}$ and $L = a^+$ in example 3) and compute kernel of $\sigma =$ kernel of $\chi =$

$$\{ \sum x_w w = \sum x_i a^i \in \mathbb{Q} \langle a \rangle / x_0 = 0 \text{ and } \sum_{i \geq 1} x_i = 0 \}$$

and therefore $A_L \cong \mathscr{A}_S$.

4. CHARACTERISTIC ALGEBRAS AND SYNTACTIC MONOIDS

Again, let X be a finite alphabet, $L \subset X^*$ and $S = \sum_{w \in L} s_w w$. All homomorphisms considered are monoid homomorphisms.

Let $\mathbb{B}$ be the basisring of $\mathscr{A}_S$.

DEFINITION 6: Let $\tilde{X} = X \cup \{[,], +, 0\}$; $X_0 = X \cup \{0\}$:

$$L = \{ [x_{11} + \ldots + x_{1m_1}][x_{21} + \ldots + x_{2m_2}] \ldots [x_{n1} + \ldots + x_{nm_n}]/x_{ij} \in X_0^*,$$

$$\text{it exists } i_1, i_2, \ldots, i_n : x_{1i_1} x_{2i_2} \ldots x_{ni_n} \in L \}.$$

We call $\tilde{L}$ the nondeterministic version of L (*see* [6]).

Example: $L = D$, the Dyck language; $\tilde{L}$ the Greibach language.

We want to show, that $\mathscr{A}_S$ the characteristic algebra of L is isomorphic to the syntactic monoid of L related to some submonoid of $\tilde{X}^*$. But there is a small problem. There is always a zero in $\mathscr{A}_S$, but is there one in the syntactic monoid? The answer is yes, because we had forced it by adjoining the special symbol "0" to the alphabet. It is easy to see, that a word " $\ldots [0] \ldots$ " represents the zero in the syntactic monoid.

Let $\mathcal{M}$ be the free submonoid of $\tilde{X}^*$ with basis:

$$\{[x]/x \in \tilde{X}^* \setminus \{[\ ,\ ]\}, \ x = x_1 + \ldots + x_n, \ x_i \in X_0^*,$$

where some x_i may be $1\}$.

Let $\tilde{S} := \text{Syn}(\mathcal{M}, \tilde{L})$ the syntactic of $\tilde{L}$ related to $\mathcal{M}$: $\tilde{S} = \mathcal{M}/\underline{\underline{\bot}}$, where for $w, w' \in \mathcal{M}$:

$$w \underline{\underline{\bot}} w' \iff \text{for all } u, v \in \mathcal{M}: \quad uwv \in \tilde{L} \iff uw'v \in \tilde{L}.$$

We need the following fact:

LEMMA 4: *For all $w, w' \in \{[x_1 + \ldots + x_n]/x_i \in X_0^*\}$:*

$$w \underline{\underline{\bot}} w' \iff \text{for all } u, v \in [X^*] \cup \{1\}: \quad uwv \in \tilde{L} \iff uw'v \in \tilde{L}.$$

Proof: Is easy to see.
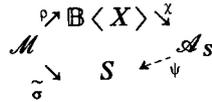
Let $\tilde{\sigma} : \mathcal{M} \to \tilde{S}$.

There exists a homomorphism $\rho : \mathcal{M} \to \mathbb{B}\langle X \rangle$, defined on the basis of $\mathcal{M}$ by:

$$[x] = [x_1 + \ldots + x_n] \mapsto (\hat{x}_1 + \ldots + \hat{x}_n) \in \mathbb{B}\langle X \rangle,$$

where $\hat{x}_i = 0_{\mathbb{B}\langle X \rangle}$ if 0 is a factor in x_i and $\hat{x}_i = x_i$ otherwise.

It is easy to see, that ρ is a monoid epimorphism.

Let us recall our situation (all mappings are considered as monoid homomorphisms):

$$\mathcal{M} \xrightarrow{\rho} \mathbb{B}\langle X \rangle \xrightarrow{\chi} \quad \mathcal{A}_S$$
$$\mathcal{M} \searrow_{\tilde{\sigma}} \quad S \xleftarrow{\psi}$$

We want to show that there exists a monoid isomorphism $\psi : \mathcal{A}_S \to \tilde{S}$.

Claim: for x, y in the basis of $\mathcal{M}$:

$$\chi\rho(x) = \chi\rho(y) \iff \tilde{\sigma}(x) = \tilde{\sigma}(y).$$

Proof of the claim: Let:

$$x = [x_1 + \ldots + x_n]; \quad y = [y_1 + \ldots + y_m]; \quad x_i, y_i \in X_0^*.$$

Now $\chi\rho(x) = \chi\rho(y)$
$\iff$ for all $u, v \in X^*$: ($ux_i v \in L$ for some $i \iff uy_j v \in L$ for some j);
$\iff$ for all $u, v \in [X^*] \cup \{1\}$: ($uxv \in \tilde{L} \iff uyv \in \tilde{L}$);
$\iff$ for all $u, v \in \mathcal{M}$: ($uxv \in \tilde{L} \iff uyv \in \tilde{L}$);
$\iff \tilde{\sigma}(x) = \tilde{\sigma}(y)$.

We have proved:

THEOREM 2: $\mathscr{A}_S \cong \tilde{S}$; there exists a monoid isomorphism between the characteristic algebra of a language and the syntactic monoid of the nondeterministic version of the language related so some submonoid.

In general, $\tilde{S} = \mathrm{Syn}(\mathscr{M}, \tilde{L})$ divides $\mathrm{Syn}(\tilde{X}^*, \tilde{L})$, the syntactic monoid. But we can prove still more.

LEMMA 5: $\mathrm{Syn}(\mathscr{M}, \tilde{L})$ is a submonoid of $\mathrm{Syn}(\tilde{X}^*, \tilde{L})$.

Proof: Claim: for all w, $w' \in \mathscr{M}$: $w \stackrel{m}{=} w' \Leftrightarrow w \sim w'$.

Proof of the claim: Observe that for u, $v \in \tilde{X}^* \backslash \mathscr{M}$: $uwv \notin \tilde{L}$. We have shown:

$$\mathrm{Syn}(\mathscr{M}, \tilde{L}) = \{ \bar{x} \in \mathrm{Syn}(\tilde{X}^*, \tilde{L}) / \text{it exists } y \in \mathscr{M} : x \sim y \}.$$

THEOREM 2: $\mathscr{A}_S \subset \mathrm{Syn}(\tilde{X}^*, \tilde{L})$; the characteristic algebra of a language is a submonoid of the syntactic algebra of the nondeterministic version of the language.

Let L be a contextfree language. Then $\tilde{L}$, the nondeterministic version of L, is also contextfree.

COROLLARY: For all contextfree languages $L \subset X^*$ exist a contextfree language $\tilde{L} \subset (X \cup \{ [\ , \ ], \ +, \ 0 \})^*$ such that:

1) $\mathscr{A}_L$ is the syntactic monoid of $\tilde{L}$ related to a submonoid $\mathscr{M} \subset (X \cup \{ [\ , \ ], \ +, \ 0 \})^*$.

2) $\mathscr{A}_L$ is a submonoid of the syntactic monoid of $\tilde{L}$.

REFERENCES

1. A. H. CLIFFORD and G. B. PRESTON, *The Algebraic Theory of Semigroups*, Amer. Math. Soc. Providence R. I., 1967.
2. S. EILENBERG, *Automata, Languages and Machines*, Academic Press, New York, Vol. A et B, 1974.
3. G. HOTZ, *Der Satz von Chomsky-Schützenberger und die schwerste kontextfreie Sprache von S. Greibach*, Société Mathématique de France, Astérisque, Vol. 38-39, 1976.
4. G. HOTZ, *k-dimensionale syntaktische Monoide*, EIK, Vol. 17, H. 8/9, 1981, pp. 435-449.
5. Ch. REUTENAUER, *Séries formelles et algèbres syntactiques*, J. Algebra, Vol. 66, 1980, pp. 448-483.
6. I. H. SUDBOROUGH, *A Note on Tape Bounded Complexity Classes and Linear Context-free Languages*, J.A.C.M., Vol. 22, 1975, pp. 499-500.
7. E. SHAMIR, *A Representation Theorem for Algebraic and Context-free Power Series in Noncommuting Variables*, Inform. and Control, Vol. 11, 1967, pp. 239-254.