A NEW CHARACTERIZATION OF THE REGULAR LANGUAGES

Christophe Reutenauer

LITP

Institut de Programmation

4 place Jussieu

75005 Paris France

Résumé On définit les notions de langage faiblement commutatif et de langage périodique. On montre qu'un langage est rationnel si et seulement s'il est périodique et faiblement commutatif.

Zusammenfassung Es werden zwei Eigenschaften der Sprachen definiert: schwache Kommutativität und Periodizität. Es wird gezeigt, dass eine Sprache regulär ist dann und nur dann, wenn sie schwach kommutativ und periodisch ist.

Summary We define weakly commutative languages and periodic languages. We show that a language is regular if and only if it is simultaneously weakly commutative and periodic.

1. INTRODUCTION

We define first a property of formal languages, which we call weak commutativity and which is related to a classical mathematical object: the alternating sum. Let us recall that a language is said to be commutative if it is saturated for the Parikh mapping. Our first result states the perhaps surprising fact: every regular language is weakly commutative. More surprising is the fact that we can state a converse. For that we say that a language is periodic if its syntactic monoid is periodic (that is, each cyclic submonoid is finite). This property is clearly related to the "pumping lemma" property. It should be noticed that weak commutativity and periodicity are orthogonal properties: you can find, for each combination of these properties or their negations, a language

satisfying it. And also that a periodic language is in general not regular; the problem of finding conditions in which a periodic language is regular could be called the Burnside problem for languages. Our main result is:

A language is regular if and only if it is periodic and weakly commutative.

2. WEAK COMMUTATIVITY

Let us recall that a language $L \subset A^*$ is called commutative if for each pair of words that have same commutative image, these words are simultaneously inside or outside L .

This can be formally described as:

$$\forall u, x_1, x_2, v \in A^* : ux_1x_2v \in L \iff ux_2x_1v \in L$$

We define now weak commutativity. A language $L \subset A^*$ is weakly commutative if there exists an integer $n \geq 2$ such that L satisfies the following property

(S_n) For any words $u, v, x_1, x_2, \dots, x_n$ the two subsets of $\mathfrak{S}_n$ (=the group of permutations of $\{1, 2, \dots, n\}$)

$$\{\sigma \in \mathfrak{S}_n \mid \sigma \text{ even} \ \& \ ux_{\sigma(1)}x_{\sigma(2)}\dots x_{\sigma(n)}v \in L\}$$

$$\text{and } \{\sigma \in \mathfrak{S}_n \mid \sigma \text{ odd} \ \& \ ux_{\sigma(1)}x_{\sigma(2)}\dots x_{\sigma(n)}v \in L\}$$

have the same cardinality.

One sees immediately that a language and its complement are simultaneously weakly commutative or not. Further (S_2) is exactly commutativity and (S_n) implies (S_{n+1}). What is a little bit surprising, but easy to prove, is that (S_{2n+1}) implies (S_{2n}): to show it take $x_{2n+1} = 1$ the empty word. Hence, only the (S_n) with n even are interesting.

We come now to a first result

Proposition 1 Every regular language is weakly commutative.

Before proving it, let us recall the definition of the syntactic monoid of a language.

Given a language $L \subset A^*$ one shows that in the set of all congruences of A^* (a congruence is an equivalence relation that is compatible with the multiplication of words) for which L is union of classes, there exists a biggest element: it is the syntactic congruence of L , denoted by $\sim_L$ or $\sim$ when no confusion arises. It may be defined by the relation

$$(1) \quad x \sim y \quad \text{iff for any words } u \text{ and } v \quad uxv \in L \Leftrightarrow uyv \in L$$

(see (1) chapter 3 §10 or (2) chapter 6 §1)

The syntactic monoid of L is the quotient $A^*/\sim$.

By Kleene's theorem, a language is regular if and only if it's syntactic monoid is finite.

Proof of proposition 1 Let M be the syntactic monoid of a given regular language

$L \subset A^*$ and $\rho: A^* \rightarrow M$ the canonical morphism of monoids. Let $\mathbb{Q}[M]$ denote the algebra of M over the field $\mathbb{Q}$ of rational numbers. Each element of $\mathbb{Q}[M]$ is a linear combination over $\mathbb{Q}$ of elements of M . Let $n = |M| + 1$ (where $|M|$ denotes the cardinality of M which is finite). Let $y_1, y_2, \dots, y_n$ be new letters and define the following noncommutative polynomial, called the standard polynomial of order n and denoted S_n

$$(2) \quad S_n(y_1, y_2, \dots, y_n) = \sum_{\sigma \in \mathfrak{S}_n} \varepsilon_\sigma y_{\sigma(1)} y_{\sigma(2)} \cdots y_{\sigma(n)}$$

where ε_σ is the signature of the permutation σ . Then for any $p_1, p_2, \dots, p_n$ in $\mathbb{Q}[M]$ we have $S_n(p_1, p_2, \dots, p_n) = 0$. Indeed, the left-hand side is an alternating n -linear mapping of the n -tuple $(p_1, p_2, \dots, p_n)$ and because $\mathbb{Q}[M]$ has dimension $< n$, this mapping classically vanishes.

Let $M' = \rho(L)$. Then $x \in L$ iff $\rho(x) \in M'$ because L is union of classes of $\sim$.

Let $\varphi: M \rightarrow \mathbb{Q}$ be the characteristic function of M' : $\varphi(m) = 1$ if $m \in M'$ and $= 0$ if not. Then for any word x $x \in L \Leftrightarrow \varphi \circ \rho(x) = 1$ and $x \notin L \Leftrightarrow \varphi \circ \rho(x) = 0$

Now for any words $u, v, x_1, x_2, \dots, x_n$ one has $\rho(u) S_n(\rho(x_1), \dots, \rho(x_n)) \rho(v) = 0$

$$\text{hence} \quad \sum_{\sigma \text{ even}} \varphi \circ \rho(u x_{\sigma(1)} \cdots x_{\sigma(n)} v) = \sum_{\sigma \text{ odd}} \varphi \circ \rho(u x_{\sigma(1)} \cdots x_{\sigma(n)} v)$$

and the proposition follows.

3. PERIODICITY

A monoid M is called periodic if each cyclic submonoid is finite (one also says torsion instead of periodic). It is equivalent to: For any $m \in M$ there exist integers k and $p \geq 0$, $k \neq p$ such that $m^k = m^p$. We say that a language is periodic if its syntactic monoid is periodic. By the above and (1) it is equivalent to

(3) For any word x there exist integers k and p , $k \neq p$ such that for any words u and v $ux^k v \in L \Rightarrow ux^p v \in L$.

Now the following proposition is trivially true.

Proposition 2 Every regular language is periodic.

4. THE MAIN RESULT

It is the converse of the two preceding propositions.

Theorem Every periodic and weakly commutative language is regular.

Before proving this result, let us give some examples showing that periodicity and weak commutativity are orthogonal properties (as said in the introduction). The language of square-free words is easily seen to be periodic, but is not regular (when the alphabet has at least 3 letters) hence not weakly commutative. Any commutative nonregular language is of course weakly commutative but not periodic. The languages of palindromes is neither periodic nor weakly commutative (when the alphabet has at least 2 letters): it can be shown directly, but also follows from (3) proposition 1.3.3.

For the proof of the theorem we need some results on formal power series. A formal power series is a mapping $S: A^* \rightarrow \mathbb{Q}$. The image of a word w by S will be denoted (S, w) . Let $\mathbb{Q}\langle A \rangle$ denote the algebra of noncommutative polynomials on A over $\mathbb{Q}$. Because A^* is a basis of the vector space $\mathbb{Q}\langle A \rangle$, each formal power series S defi-

nes on $\mathbb{Q}\langle A \rangle$ a linear form extending the mapping $S: A^* \rightarrow \mathbb{Q}$; the image by S of a polynomial P will be denoted by (S, P) . With this convention, the kernel of S is as usually the subspace of $\mathbb{Q}\langle A \rangle$: $\text{Ker}(S) = \{ P \in \mathbb{Q}\langle A \rangle \mid (S, P) = 0 \}$

The syntactic ideal of S is the biggest (two-sided) ideal of $\mathbb{Q}\langle A \rangle$ contained in $\text{Ker}(S)$. It exists because any sum of ideals is again an ideal. We denote it $\mathcal{I}_S$ or simply $\mathcal{I}$. It is not difficult to show that

$$(4) \quad \mathcal{I} = \{ P \in \mathbb{Q}\langle A \rangle \mid \forall u, v \in A^* \quad (S, uv) = 0 \}$$

(see (3) 1.2). The syntactic algebra of S is the quotient $\mathbb{Q}\langle A \rangle / \mathcal{I}$ that

we denote by $\mathcal{M}_S$ or $\mathcal{M}$. As a consequence of the Kleene-Schützenberger theorem

(see (1) th.5.1 chapter 7 or (5) th.2.3 chapter 2) a formal power series is rational if and only if it's syntactic algebra is finite dimensional (see (3) th.2.1.2)

If L is a language, it's characteristic series, still denoted L , is defined by

$(L, w) = 1$ if w is in L and $(L, w) = 0$ if not. We shall say syntactic algebra of

L instead of syntactic algebra of it's characteristic series, and so on...

Now, by a theorem due to Schützenberger, a language is regular if and only if it's characteristic series is rational (see (5) th.5.2 chapter 2).

Proof of the theorem Let $L \subset A^*$ be a periodic and weakly commutative language. Let

$\mathcal{M}$ be it's syntactic algebra and $\mu: \mathbb{Q}\langle A \rangle \rightarrow \mathcal{M}$ the canonical algebra homomorphism.

$\mathcal{M}$ is generated as an algebra by the finite set $\mu(A)$. Moreover, let x be a word;

then by (3) there exist two integers k and p such that for any words u and v

$ux^k v$ is in L iff $ux^p v$ is in L . This means $(L, u(x^k - x^p)v) = 0$ and implies

that $x^k - x^p$ is in the syntactic ideal of L by (4). In other words

$$\mu(x)^k - \mu(x)^p = 0$$

Furthermore, let L verify (S_n) and consider the n -linear mapping from $\mathcal{M}^n$ into $\mathcal{M}$

defined by $(p_1, \dots, p_n) \mapsto S_n(p_1, \dots, p_n)$ where the polynomial is defined by

(2). We show that this mapping is zero; it suffices to do so when $p_1, \dots, p_n$ are elements

of a generating system of the vector space $\mathcal{M}$. Now $\mu(A^*)$ is such a system

and we have for any words $u, v, x_1, \dots, x_n$ $(L, u S_n(x_1, \dots, x_n) v) = 0$ what is

just another way to write that L satisfies (S_n) . By (4) this means that

$S_n(x_1, \dots, x_n)$ is in the syntactic ideal of L hence $S_n(\mu(x_1), \dots, \mu(x_n)) = 0$

and $S_n = 0$ in $\mathcal{M}$. This shows that $\mathcal{M}$ satisfies the standard identity $S_n = 0$ which is a special case of a polynomial identity, see (4). Now we may apply the following theorem of Shirshov (6) (whose application is the main point of the present proof): Let $\mathcal{M}$ be an algebra generated by a finite set A and satisfying a polynomial identity. If each element of the multiplicative semigroup generated by A is algebraic then $\mathcal{M}$ is finite dimensional (for a proof see (4) §2 chapter 4). Hence $\mathcal{M}$ is finite dimensional, L is rational as a series and L is a regular language.

The proof makes clear that a language satisfies (S_n) if and only if its syntactic algebra satisfies the standard identity $S_n = 0$.

The theorem may be improved: instead of (S_n) one may assume the weaker property $(S_{n,a})$ for a given integer $a \geq 2$:

$(S_{n,a})$ For any words $u, v, x_1, \dots, x_n$ the same two finite subsets of $\mathcal{M}_n$ as in (S_n) have the same cardinality modulo a .

The proof still works by replacing $\mathbb{Q}$ by the ring $\mathbb{Z}/a\mathbb{Z}$ because the necessary results on formal power series are still true for any ring in place of $\mathbb{Q}$ (see (3) th.1.2 and (7) th.2.3).

As a conclusion, remark that the theorem is a special case of a characterization of the rationality of formal power series, which is directly related to Kurosh's problem in ring theory (see (3) 2.3).

REFERENCES

- (1) S.Eilenberg: Automata, languages and machines vol.A Acad. Press (1974)
- (2) G.Lallement: Semigroups and combinatorial applications John Wiley (1979)
- (3) C.Reutenauer: Séries formelles et algèbres syntactiques J.Algebra 66 (1980)
448-483.
- (4) L.H.Rowen: Polynomial identities in ring theory Acad. Press (1980)
- (5) A.Salomaa and M.Soittola: Automata-theoretic aspects of formal power series
Springer (1978)
- (6) A.I.Shirshov: On rings with identity relations Mat. Sb. N. S. 43 (85) (1957)
277-283 (russian)
- (7) E.D.Sontag: On some questions of rationality and decidability J.Comput.System
Sci. 11 (1975) 375-385