INFORMATION AND CONTROL 57, 205–213 (1983)

Some Applications of a Theorem of Shirshov to Language Theory*

ANTONIO RESTIVO

Universitá di Palermo, Palermo, Italy

AND

CHRISTOPHE REUTENAUER

CNRS, Paris, France

Some applications of a theorem of Shirshov to language theory are given: characterization of regular languages, characterization of bounded languages, and a sufficient condition for a language to be Parikh-bounded.

1. INTRODUCTION

In 1957, A. I. Shirshov solved affirmatively the famous problem of Kurosch (which is the analogue for algebras of the Burnside problem for groups), in the general case of algebras with polynomial identities. The heart of the proof is a combinatorial result which states, roughly speaking, that each long word either contains some power of a word or has some permutation property.

First we use this theorem of Shirshov to give a characterization of regularity: let us say that a language L has the transposition property if for some integer m, in each word $w = ux_1 \cdots x_m v$, it is possible to transpose two consecutive blocks of x's, obtaining a word w' such that $w \in L$ iff $w' \in L$. Together with periodicity, which is a kind of pumping property (related to the Burnside problem), the transposition property characterizes regularity (Theorem 3.2). This theorem has some analogy with a theorem of Ehrenfeucht, Rozenberg, and Parikh (1981), which characterizes regularity by some cancellation property. Also, the transposition property has some connection with the weak commutativity of Reutenauer (1981).

Our next result (Theorem 4.1) characterizes the boundedness property of languages: a language is bounded iff for some integer n it does not contain n-

* This work was done while the second author was a visiting professor at the University of Palermo, supported by the CNR.

0019-9958/83 $3.00

643/57/2-3-9

divided words. For the proof, we need also Shirshov's theorem. As a corollary (Corollary 4.4), we obtain, with the use of Restivo (1977) and Boasson and Restivo (1977), a nice property of regular and context-free languages.

Latteux and Leguy (1979 have introduced the following concept: a language is Parikh-bounded if it contains some bounded language having the same commutative image. We give a sufficient condition for this property (Theorem 5.1) and obtain, as a corollary, the fact that all supports of rational power series are Parikh-bounded languages (Corollary 5.2).

2. A Theorem of Shirshov

Let A be a totally ordered and finite alphabet. In the free monoid A^* generated by A, words of equal length are ordered laxicographically (from the left to the right); the order is denoted by $\leqslant$ and $u < v$ means that $u \leqslant v$ and $u \neq v$.

Let w be a word in A^*. An *n-division* of w is a factorization

$$w = ux_1 \cdots x_n v$$

such that for any permutation σ of $\{1,..., n\}$, $\sigma \neq \mathrm{id}$, one has

$$w < ux_{\sigma(1)} \cdots x_{\sigma(n)} v.$$

We say that a word is *n-divided* if it admits at least one n-division.

We say that a word w *contains a pth power* of a word x if x is nonempty and if w may be written $w = ux^p v$ for some words u and v.

The following theorem is due to Shirshov (1957). A proof may be found in Lothaire (1983) or Rowen (1980).

Theorem 2.1 (Shirshov). *For any integers k, p, $n \geqslant 1$ such that $p \geqslant 2n$, there exists an integer $N(k, p, n)$ such that each word of length at least $N(k, p, n)$ on an alphabet of cardinality k either is n-divided or contains a pth power of a word of length at most $n - 1$.*

3. A Characterization of Regularity

We say that a language $L \subset A^*$ has the *transposition property* if there exists an integer m such that for each words $w, u, x_1,..., x_m, v$ in A^* verifying $w = ux_1 \cdots x_m v$, there exist i, j, k, $1 \leqslant i < j < k \leqslant m$, such that

$$w \in L \Leftrightarrow ux_1 \cdots x_{i-1} x_j \cdots x_{k-1} x_i \cdots x_{j-1} x_k \cdots x_m v \in L \qquad (3.1)$$

(the word of the right member is obtained by interchanging in w the consecutive blocks $x_i \cdots x_{j-1}$ and $x_j \cdots x_{k-1}$).

Remark 3.1. There is a formal analogy between the transposition property and the cancellation property of Ehrenfeucht *et al.* (1981), although there is no evident mathematical relation between them. The same remark applies to the weak commutativity of Reutenauer (1981).

Note that each regular language L has the transposition property: indeed, let $m =$ twice the number of states of some finite deterministic automaton recognizing L; let q_0 be the initial state and $w = ux_1 \cdots x_m v$. Then in the sequence of $m+1$ states

$$q_0 u, q_0 u x_1, q_0 u x_1 x_2, \ldots, q_0 u x_1 \cdots x_m$$

there is one state, say q, which appears at least three times. This implies that one can interchange the two corresponding blocks in w, obtaining a word w' such that

$$w \in L \Leftrightarrow w' \in L.$$

Hence L has the transposition property.

Recall that the *syntactic congruence* of a language L is the congruence of A^* defined by: $x \sim y$ if and only if for any words u and v

$$uxv \in L \Leftrightarrow uyv \in L$$

($x \sim y$ means exactly that x and y have the same *contexts* in L).

The syntactic monoid of L is the quotient monoid $A^*/\sim$; see Eilenberg (1974). A monoid is *periodic* if any element of it is periodic, i.e., generates a finite submonoid. We call a language periodic if its syntactic monoid is periodic.

Note that for any finite cyclic monoid generated by an element x, there exists an integer $p \geqslant 1$ such that $x^{2p} = x^p$. Hence a language is periodic if and only if for each word x, there exists an integer $p \geqslant 1$ verifying, for any words u and v,

$$ux^p v \in L \Leftrightarrow ux^{2p} v \in L. \tag{3.2}$$

Note also that each regular language is periodic because by Kleene's theorem, its syntactic monoid is finite. We come now to the converse.

THEOREM 3.2. *If a periodic language has the transposition property, then it is regular.*

Proof. (i) We use a particular case of Ramsey's theorem: For each set X, denote by $X[3]$ the set of subsets of X of cardinality 3. Then: for each

$m \geqslant 1$, there exists an integer $n(m)$ such that for each set X, $\mathrm{card}(X) \geqslant n(m)$, and each partition $X[3] = I \cup J$, there is some subset Y of X, $\mathrm{card}(Y) = m$, such that

$$Y[3] \subset I \qquad \text{or} \qquad Y[3] \subset J. \tag{3.3}$$

See Harrison (1978, Theorem 1.7.1).

(ii) Note that if L is a periodic language and W a finite set of words, then it is possible to find p such that (3.2) holds for all $x \in W$; moreover p may be chosen arbitrarily large. Denote by $\mathscr{L}_{m,p}$ the set of languages on the given finite alphabet A which have the transposition property for m and which are periodic, with the property that all words x of length at most $n(m)$ verify (3.2).

By the previous remark, each periodic language having the transposition property is in some $\mathscr{L}_{m,p}$ with $p \geqslant n(m)$.

(iii) Let $\mathscr{L} = \mathscr{L}_{m,p}$ with $p \geqslant n(m)$. It will suffice to show that $\mathscr{L}$ is finite: indeed, $L \in \mathscr{L}$ implies $a^{-1}L = \{w / aw \in L\} \in \mathscr{L}$ for each letter a and one applies Nerode's criterion (see Eilenberg, 1974, Theorem III.8.1).

Let $n = n(m)$ and $N = N(k, 2p, n)$ defined as in Theorem 2.1, with $k = \mathrm{card}(A)$. Then each word of length at least N is either n-divided or contains a $(2p)$th power of some word of length at most $n - 1$.

(iv) Let $L, L' \in \mathscr{L}$ such that for each word w,

$$|w| < N: w \in L \Leftrightarrow w \in L'.$$

We show that this implies $L = L'$ (hence $\mathscr{L}$ is finite). For this, order $A^*: u \prec v$ means either that $|u| < |v|$ or that $|u| = |v|$ and $u > v$ (lexicographic order). We show by induction on this order that for each word w, $w \in L$ iff $w \in L'$. This is true if $|w| < N$.

Let $|w| \geqslant N$. Suppose w contains a $(2p)$th power of a word x, $|x| \leqslant n - 1$: $w = ux^{2p}v$. Then because $L, L' \in \mathscr{L}_{m,p}$, one has by (3.2) and induction:

$$w \in L \Leftrightarrow ux^p v \in L \Leftrightarrow ux^p v \in L' \Leftrightarrow w \in L'.$$

Suppose now that w contains no such $(2p)$th power: then w admits an n-division

$$w = ux_1 \cdots x_n v. \tag{3.4}$$

Let $X = \{1, 2, \ldots, n\}$ and define a subset I of $X[3]$ by: for $1 \leqslant i < j < k \leqslant n$, $\{i, j, k\} \subset I$ iff

$$ux_1 \cdots x_{i-1} x_j \cdots x_{k-1} x_i \cdots x_{j-1} x_k \cdots x_n v \in L.$$

Note that, because (3.4) is an n-division and by induction, I remains unchanged if L is replaced in the above definition by L'.

Let $J = X[3] \backslash I$. Then by Ramsey's theorem, there exists $Y \subset X$, $\mathrm{card}(Y) = m$, such that (3.3) holds.

Let $w = u' y_1 \cdots y_m v'$ be the subfactorization of (3.4) corresponding to Y. Because L has the transposition property, there exist i, j, k with $1 \leqslant i < j < k \leqslant m$ and such that

$$w \in L \Leftrightarrow u' y_1 \cdots y_{i-1} y_j \cdots y_{k-1} y_i \cdots y_{j-1} y_k \cdots y_m v' \in L.$$

Hence if $w \in L$, then there is some $\{i, j, k\}$ in $I \cap Y[3]$. By (3.3) this implies that $Y[3] \subset I$. Conversely if $Y[3] \subset I$, then by the transposition property of L one has $w \in L$. Hence $w \in L \Leftrightarrow Y[3] \subset I$.

A previous remark and the transposition property of L' also imply $w \in L' \Leftrightarrow Y[3] \subset I$. Thus $w \in L \Leftrightarrow w \in L'$ and the theorem is proved. ∎

Remark 3.1. The transposition property defines an infinite hierarchy, as is easily seen in the following example: let $A = \{a_1, ..., a_m\}$; then the singleton-language $\{a_1 \cdots a_m\}$ has the transposition property for $m + 1$ but not for m.

PROBLEM. Modify the transposition property in the following way: if $w = u x_1 \cdots x_m v$ then there exists some permutation σ of $\{1, ..., m\}$, $\sigma \neq \mathrm{id}$, such that

$$w \in L \Leftrightarrow u x_{\sigma(1)} \cdots x_{\sigma(m)} v \in L.$$

Is the theorem still true with this weaker property?

4. BOUNDED LANGUAGES

Recall that a language is *bounded* if for some words $u_1, ..., u_q$, it is contained in $u_1^* \cdots u_q^*$.

THEOREM 4.1. *A language is bounded if and only if for some integer n it contains no n-divided word.*

Proof. We show first that for each bounded language L, there exists n such that L contains no n-divided word. It suffices to do so for $L = u_1^* \cdots u_q^*$.

Let $n = q \max\{2|u_i| + 1,\ 1 \leqslant i \leqslant q\}$. Suppose that $w \in L$ is n-divided; then w may be written

$$w = u_1^{n_1} \cdots u_q^{n_q} = u x_1 \cdots x_n v.$$

Hence, for some i, $1 \leqslant i \leqslant q$, and some j, k, $1 \leqslant j < k \leqslant n$, one has $u_i^{n_i} = u' x_{j+1} \cdots x_k v'$ and $k - j \geqslant 2|u_i| + 1$. This implies that for some words u_i',

u_i'' and for some integers k_1, k_2, k_3, $k_1 < k_2 < k_3$, one has $u_i = u_i' u_i''$ and $x_{k_1+1} \cdots x_{k_2}$, $x_{k_2+1} \cdots x_{k_3} \in (u_i'' u_i')^*$. But then $x_{k_1+1} \cdots x_{k_2}$, $x_{k_2+1} \cdots x_{k_3}$ commute and thus $ux_1 \cdots x_n v$ is not an n-division of w. Hence L contains no n-divided word.

For the converse fix an integer $n \geqslant 1$ and let $N = N(k, 2n, n)$ with $k = \text{card}(A)$.

LEMMA 4.2. *Let w be a word of length at least N which is not n-divided. Then it may be written*

$$w = ux^p v$$

with $|u| < N + n$, $0 < |x| < n$, $p \geqslant 2n$, and either v is empty or $Fv \neq Fx$ (where Fv denotes the first letter of v).

Proof. Let $w = w'w''$ with $|w'| = N$. Then by Shirshov's theorem, we have $w' = sy^{2n}t$ for some words s, t, y such that $|s| < N$ and $0 < |y| < n$. Hence $w = sy^p r$ with $p \geqslant 2n$ and y is not a prefix of r. Let y' be the longest prefix common to y and r: then $y = y'y''$, $y'' \neq 1$, $r = y'v$, where either v is empty or $Fv \neq Fy''$. Put $u = sy'$, $x = y''y'$. Then $w = sy^p r = s(y'y'')^p y'v = ux^p v$ with $v = 1$ or $Fv \neq Fx$, because $Fx = Fy''$. Because $|x| = |y|$ and $|u| = |s| + |y'| < N + n$, the lemma is proved. ∎

LEMMA 4.3. *Let w be a word which is not n-divided. Then it admits a factorization*

$$w = u_0 x_1^{p_1} u_1 x_2^{p_2} \cdots x_q^{p_q} u_q \tag{1}$$

with $|u_i| < N + n$, $0 < |x_i| < n$, $p_i \geqslant 2n$ for $1 \leqslant i \leqslant q$, $Fx_i \neq F(u_i x_{i+1})$ if $1 \leqslant i \leqslant q - 1$ and either $u_q = 1$ or $Fx_q \neq Fu_q$.

Proof. If $|w| < N$ the lemma holds true for w. Suppose $|w| \geqslant N$; then by Lemma 4.2, $w = u_0 x_i^p v$, where $|u_0| < N + n$, $0 < |x_i| < n$, $p \geqslant 2n$, and $v = 1$ or $Fv \neq Fx$. By induction, the lemma is true for v. This implies that it is true for w. ∎

Now we can prove the theorem.

Let $Q = 1 + kln$ where $k = \text{card}(A)$ and $l = \text{Card}\{x \in A^*,\ 0 < |x| < n\}$. We show that if w is as in Lemma 4.3, then $q \leqslant Q$.

Suppose the contrary: then there exist $i_1, \ldots, i_n$ such that $1 \leqslant i_1 < \cdots < i_n \leqslant q$ and that $x_{i_1} = \cdots = x_{i_n}$ $(=x)$, $Fu_{i_1} = \cdots = Fu_{i_n}$ $(=b)$. Let $a = Fx$; then $a \neq b$.

Now w may be written

$$w = v_0 x^n b v_1 x^n b \cdots x^n b v_n.$$

Suppose $a < b$: then w admits the n-division

$$w = v_0(x^n b v_1 x)(x^{n-1} b v_2 x^2) \cdots (x b v_n).$$

Suppose $a > b$: then w admits the n-division

$$w = v_0 x^{n-1}(x b v_1 x^{n-1})(x^2 b v_2 x^{n-2}) \cdots (x^n b v_n).$$

Hence, in both cases, w is n-divided: contradiction. This shows that each word w which is not n-divided admits a factorization of the form (1) with $q \leqslant Q$. Hence the set of all these words is a bounded language. ∎

COROLLARY 4.4. *Let L be a regular language. The two following conditions are equivalent*:

(i) *For some p, L contains arbitrary long words without pth power.*

(ii) *For each n, L contains an n-divided word.*

Proof. The second condition is equivalent to: L is not bounded (Theorem 4.1). But so is the first, by Theorem 2 of Restivo (1977). ∎

Remark 4.5. From Restivo (1977) and Shirshov's theorem, it follows directly that any regular language without any n-divided word is bounded. The same is true for any context-free language, by Boasson and Restivo (1977). This raises the question whether the language

$$L_n = \{w, w \text{ is not } n\text{-divided}\}$$

is regular or context-free. For $A = \{a, b\}$, $a < b$ one has

$$L_2 = b^* a^*.$$

Moreover it is easy to show that

$$L_n \cap a(ba^*)^{n-1} = \{aba^{i_1} \cdots ba^{i_{n-1}}, \exists j, i_j \leqslant i_{j+1}\},$$

which is not regular, but context-free. Hence L_n is not regular. It remains open if L_n is context-free or not.

COROLLARY 4.5. *Let L be a context-free language. The two following conditions are equivalent*:

(i) *For some p, there are arbitrarily long words without pth power which are factors of words of L.*

(ii) *For each n, L contains an n-divided word.*

Proof. As for Corollary 4.4, but using Boasson and Restivo (1977). ∎

5. Parikh-Bounded Languages

Following Blattner and Latteux (1981) and Latteux and Leguy (1979), we say that a language L is *Parikh-bounded* if it contains some bounded language L' such that $p(L) = p(L')$, where $p: A^* \to \mathbb{N}^k$ is the Parikh-mapping and $k = \text{card}(A)$. In these papers it is shown that each context-free language is Parikh-bounded.

Theorem 5.1. *Let L be a language and $n \geqslant 1$ such that for any $w = ux_1 \cdots x_n v$ in L, there is some permutation σ of $\{1,...,n\}$, $\sigma \neq \text{id}$, such that $ux_{\sigma(1)} \cdots x_{\sigma(n)} v$ is still in L. Then L is Parikh-bounded.*

Corollary 5.2. *If L is the support of some rational power series, then L is Parikh-bounded.*

Recall that a language L is the *support* of some rational power series exactly when there exist a monoid homomorphism $\mu: A^* \to K^{n \times n}$ (the multiplicative monoid of n by n matrices over a field K) and a linear mapping $\varphi: K^{n \times n} \to K$ such that

$$L = \{w \in A^*, \varphi(\mu w) \neq 0\} \tag{5.1}$$

See Salomaa and Soittola (1978) for this and more about supports; especially each regular language is a support, but the converse is not true.

Proof of the theorem. Let L_n be as in Remark 4.5. Let $L' = L \cap L_n$. Then L' is bounded (Theorem 4.1) and $p(L') \subset p(L)$. It remains to show that $p(L) \subset p(L')$.

Let $w \in L$. Then either $w \in L_n$, hence $w \in L'$ and $p(w) \in p(L')$, or $w \notin L_n$: then w is n-divided,

$$w = ux_1 \cdots x_n v.$$

By hypothesis there is some permutation σ of $\{1,...,n\}$ such that $w' = ux_{\sigma(1)} \cdots x_{\sigma(n)} v$ is still in L. Then $|w'| = |w|$ and $w' > w$: hence by induction $p(w) = p(w') \in p(L')$. Thus $p(L) \subset p(L')$. ∎

Proof of the corollary. It suffices to show that L, as defined by (5.1), satisfies the hypothesis of the theorem. By the theorem of Amitsur–Levitzki (see Rowen, 1980, Theorem 1.4.1), for any matrices $m_1,..., m_{2n}$ in $K^{n \times n}$, one has

$$\sum_{\sigma \in \mathfrak{S}_{2n}} (-1)^\sigma m_{\sigma(1)} \cdots m_{\sigma(2n)} = 0$$

where $(-1)^\sigma$ is the signature of the permutation σ.

Let $w = ux_1 \cdots x_{2n} v \in L$. Then

$$\sum_{\sigma} (-1)^{\sigma} \mu(ux_{\sigma(1)} \cdots x_{\sigma(2n)} v) = 0.$$

Apply φ to this equality. Because $\varphi(\mu(ux_1 \cdots x_{2n} v)) \neq 0$, there is some σ such that $\varphi(\mu(ux_{\sigma(1)} \cdots x_{\sigma(2n)} v)) \neq 0$, hence $ux_{\sigma(1)} \cdots x_{\sigma(2n)} v \in L$. ∎

Remark 1. Corollary 5.2 gives a new proof for the fact that each regular language is Parikh-bounded. Unfortunately this proof does not work for context-free languages, because they do not satisfy in general the hypothesis of Theorem 5.1 (for example, the set of palindrome words).

2. The bounded language $L' \subset L$ of Corollary 5.2 may effectively be constructed. Indeed, by the proof of Theorem 4.1, there exists an effective bounded regular set L'_n containing L_n. Then $L'' = L'_n \cap L$ is the support of some rational power series which may effectively be given (see Salomaa and Soittola, 1978, Theorem 2.4.5).

RECEIVED: July 11, 1983

REFERENCES

BOASSON, L., AND RESTIVO, A. (1977), Une caracterisation des languages algebriques bornes, *RAIRO Inform.* **11**, 203–205.

BLATTNER, M., AND LATTEUX, M. (1981), Parikh-bounded languages, *in* "Lecture Notes in Computer Science No. 115," pp. 316–323, Springer-Verlag, New York/Berlin.

EHRENFEUCHT, A., PARIKH, R., AND ROZENBERG, G. (1981), Pumping lemmas for regular sets, *Siam J. Comput.* **10**, 536–541.

EILENBERG, S. (1974), Automata, Languages and Machines," Vol. A, Academic Press, New York.

HARRISON, M. (1978), "Introduction to Formal Language Theory," Addison–Wesley, Reading, Mass.

LATTEUX, M. AND LEGUY, J. (1979), Une propriete de la famille GRE," pp. 255–261, Foundations of Computer Science, Akademie-Verlag, Berlin.

LOTHAIRE, M. (1983), "Combinatorics on Words," Addison–Wesley, Reading, Mass.

RESTIVO, A. (1977), Mots sans repetitions et languages rationnels bornes, *RAIRO Inform. Theor.* **11**, 197–202.

REUTENAUER, C. (1981), A new characterization of the regular languages, *in* "Lecture Notes in Computer Science No. 115," pp. 177–183, Springer-Verlag, New York/Berlin.

ROWEN, L. H. (1980), Polynomial Identities in Ring Theory," Academic Press, New York.

SALOMAA, A., AND SOITTOLA, M. (1978), "Automata-Theoretic Aspects of Formal Power Series," Springer-Verlag, New York/Berlin.

SHIRSHOV, A. I. (1957), On rings with identity relations, *Mat. Sb.* **43**, 277–283. [In russian]