RAIRO-Inf. Theor. Appl. **40** (2006) 295-301
DOI: 10.1051/ita:2006014

ON THE SIMPLEST CENTRALIZER OF A LANGUAGE*

Paolo Massazza[1] and Petri Salmela[2]

Abstract. Given a finite alphabet Σ and a language $L \subseteq \Sigma^+$, the centralizer of L is defined as the maximal language commuting with it. We prove that if the primitive root of the smallest word of L (with respect to a lexicographic order) is prefix distinguishable in L then the centralizer of L is as simple as possible, that is, the submonoid $L^\star$. This lets us obtain a simple proof of a known result concerning the centralizer of nonperiodic three-word languages.

Mathematics Subject Classification. 68Q70, 68R15.

1. Introduction

Language equations, that is, equations with languages as variables, have been considered with particular interest since the work of Chomsky and Schützenberger [4]. For example, in [2] special systems of equations (called left language equations) have been studied in their relation with boolean automata and sequential networks.

The equation $XL = LX$ (known as the commutation equation) has been deeply investigated since 1971, when Conway raised a problem concerning the commutation with rational languages [5]. More precisely, his question was about the centralizer of a rational language, that is, the maximal solution of the equation $XL = LX$: is it true that the centralizer of any rational language is rational?

Several interesting results concerning the commutation of languages have been presented since then. In particular, in the case of codes, in [14] it has been shown

Keywords and phrases. Commutation equation, centralizer, lexicographic order.

* *Partially supported by: (1) Project M.I.U.R. PRIN 2005-2007:* Automata and formal languages: mathematical and application driven studies*; (2) Academy of Finland under the grant 203354.*

[1] Dipartimento di Informatica e Comunicazione, Università degli Studi dell'Insubria, via Mazzini 5, 21100 Varese, Italy; paolo.massazza@uninsubria.it

[2] Department of Mathematics and TUCS, University of Turku, 20014 Turku, Finland; pesasa@utu.fi

that if a code L and a circular code X commute (see [1] for definitions), then $L = X^k$ for a suitable integer k. Moreover, for each prefix code L (no word is a prefix of another) its centralizer is always $\rho(L)^\star$, where $\rho(L)$ is the primitive root of L. This shows that Conway's Problem has a positive answer for rational prefix codes.

Partial answers to Conway's Problem have been given in the last years. In [3] the commutation with a two-word language L has been studied, showing that the centralizer is A^+, where either $A = L$ (if L consists of two noncommuting words) or $A = \{t\}$ and t is a primitive word (if L consists of two commuting words $x = t^r$, $y = t^s$). Later on, a similar result has been given for three-word languages (see [7, 9]).

Conway's Problem has been solved for particular classes of codes. More precisely, in [8] it has been proved that the centralizer of any rational code is rational and that the centralizer of any finite code is finitely generated.

More recently, a definitive and negative answer to the original question raised by Conway has been found. In fact, a finite language with a centralizer which is not recursively enumerable is shown in [10].

In this note we deal with the problem of identifying a suitable class of languages for which the commutation equation $XL = LX$ has a trivial maximal solution, that is, the submonoid $L^\star$. As proved in [10], the centralizer of a very simple language (like a finite language) can be very complex, therefore it is quite natural to look for suitable conditions under which the centralizer is as simple as possible.

According to this aim, we present a sufficient condition under which the centralizer of L is as simple as possible. More precisely, we prove that if the primitive root of the smallest word of L (with respect to a lexicographic order) is not a prefix of other words in L then the centralizer of L is $L^\star$, that is, the submonoid generated by L. Moreover, from this result we can easily obtain some corollaries that let us immediately prove that the centralizer of a nonperiodic three-word language $L \subseteq \Sigma^+$ is $L^\star$ (see [7, 9]).

2. Preliminaries

Let Σ be a finite alphabet. A language L on Σ is a subset of the free monoid generated by Σ, $L \subseteq \Sigma^\star$. Given a word $w \in \Sigma^\star$ we denote its length by $|w|$. The word of length 0 is the *empty* word ϵ. For two words $x, w \in \Sigma^\star$, we say that x is a *prefix* of w if and only if $w = xy$ for a suitable $y \in \Sigma^\star$. In this case we write $x \leq w$. If $y \neq \epsilon$ we say that x is a *proper* prefix of w and write $x < w$. Analogously, a word y is a *suffix* of w if and only if $w = xy$ for a suitable word x. We use the notation $x = wy^{-1}$ to denote the operation of erasing a suffix y from w. This also applies to languages by setting $Lx^{-1} = \{wx^{-1} \mid w \in L\}$.

Given a word w and an integer e, $0 \leq e \leq |w|$, we denote by $\mathrm{pref}_e(w)$ the prefix of w having length e, that is, the string $x \in \Sigma^e$ such that $w = xy$ for a suitable $y \in \Sigma^{|w|-e}$. Similarly, we indicate by $\mathrm{suf}_e(w)$ the suffix of w of lenght e.

Two words x, y are said to be *prefix incomparable* if $x \not\leq y$ and $y \not\leq x$. A word $w \in L$ is said *prefix distinguishable* in L if and only if for any $y \in L \setminus \{w\}$, w and y are prefix incomparable. We point out that prefix distinguishable words are also called left singular by some authors (see, for instance [9,14]). We recall here one of the oldest results in combinatorics on words regarding the commutation of words (see, for example [11]).

Theorem 2.1. *Let $u, v \in \Sigma^\star$. The following properties are equivalent*

(1) *$uv = vu$;*
(2) *there exist $t \in \Sigma^\star$ and $r, s \in \mathbb{N}$ such that $u = t^r, v = t^s$.*

A word w is called *primitive* if $w = u^r$ implies $u = w$ and $r = 1$. We say that u is a *root* of w if there is $r \in \mathbb{N}$ such that $u^r = w$. By Theorem 2.1, it is immediate to prove that every word admits exactly one *primitive root*, that is, a root which is primitive. Thus, a word $w \in L$ is said *root prefix distinguishable* in L if its primitive root ρ is prefix distinguishable in $(L \setminus \{w\}) \cup \{\rho\}$. Note that every root prefix distinguishable word in L is also prefix distinguishable in L. So, a simple but useful result concerning prefix distinguishable words is given in the following lemma.

Lemma 2.2. *Let $w \in L$ be prefix distinguishable in L. Then for each $Y \subseteq \Sigma^\star$ the condition $wy \in LY$ implies $y \in Y$.*

Proof. Let u be the prefix of wy which belongs to L. Note that if $u < w$ or $w < u$ then w would not be prefix distinguishable in L. Thus, $u = w$ and $y \in Y$. □

Given $L \subseteq \Sigma^\star$, we say that R is a *root* of L if there is $e \in \mathbb{N}$ such that $L = R^e$. A root which is not a proper power of another language is called *minimal*. Given a language L, if there is only one minimal root R of L, we say that R is the *primitive* root of L and we denote it as $\rho(L)$. So, L is primitive if $L = \rho(L)$.

We recall that prefix codes admit primitive root (since prefix codes form a free semigroup, see [1, 13]), while Conjecture 2 in [14] (any code has a primitive root) is still open.

We say that $X \subseteq \Sigma^\star$ commutes with L if and only if $XL = LX$. Note that if S and R commute with L then $S \cup R$ commutes with L as well. So, we define the *centralizer* of $L \subseteq \Sigma^\star$ as the largest subset of $\Sigma^\star$ which commutes with L, that is, the maximal solution of the equation $XL = LX$. We indicate by $\mathcal{C}(L)$ the centralizer of L; note that if A commutes with L then $A \subseteq \mathcal{C}(L)$, that is, $\mathcal{C}(L) = \bigcup_{Y \mid YL = LY} Y$. In particular, since for any L we have $L^\star L = LL^\star$, the submonoid generated by L is always contained in $\mathcal{C}(L)$. Moreover, if ϵ belongs to L then $\mathcal{C}(L)$ is $\Sigma^\star$. Henceforth, we are interested in the centralizer of languages which do not contain ϵ.

Given $L \subseteq \Sigma^\star$, we say that L is *branching* if we can find two words $v, w \in L$ such that $\mathrm{pref}_1(v) \neq \mathrm{pref}_1(w)$. Finally, L is said *periodic* if $L \subseteq u^\star$ for some $u \in \Sigma^\star$. If $L \subseteq \Sigma^+$ is not branching then we have $L = aL_1$ for suitable $a \in \Sigma$, $L_1 \subseteq \Sigma^\star$. In this case, we define the *circular shift* of L as the language $L^{\hookrightarrow} = L_1 a$. We recall here Theorem 2 in [6].

Theorem 2.3. *Let $L \subseteq \Sigma^+$ be a nonperiodic language. Then there is a branching language $\hat{L} \subseteq \Sigma^+$ such that $\mathcal{C}(L) = L^\star$ if and only if $\mathcal{C}(\hat{L}) = \hat{L}^\star$.*

The following lemma illustrates the relation between the centralizer of a language which is not branching and the centralizer of its circular shift. It can be proved as shown in the proof of Theorem 2 in [6].

Lemma 2.4. *Let $L \subseteq \Sigma^+$ be nonperiodic and nonbranching, $L = aL_1$. Then*

$$\mathcal{C}(L) = (a\mathcal{C}(L^{\hookrightarrow}))a^{-1}.$$

Let $\prec$ be a linear order on Σ. We can extend the relation $\prec$ in order to define a lexicographic order on $\Sigma^\star$. The *pure lexicographic* order $\leq_{\text{lex}}$ is defined as follows: given $x, y \in \Sigma^\star$, we write $x \leq_{\text{lex}} y$ if and only if either $x \leq y$ or there exist $\alpha, u, v \in \Sigma^\star$ and $\sigma, \tau \in \Sigma$ such that $x = \alpha\sigma u$ and $y = \alpha\tau v$ with $\sigma \prec \tau$. We write $x <_{\text{lex}} y$ if $x \leq_{\text{lex}} y$ and $x \neq y$. We denote by $\min_{\text{lex}}(L)$ the smallest word of L with respect to $\leq_{\text{lex}}$, that is, the word $x \in L$ such that $x \leq_{\text{lex}} y$ for all $y \in L$.

3. The equation $XL = LX$

Let L be a language with at least two words and such that $u = \min_{\text{lex}}(L)$ is root prefix distinguishable in L. Note that if u is primitive then it is root prefix distinguishable in L if and only if it is prefix distinguishable in L. Thus, the following theorem generalizes a result in [12] which states that $\mathcal{C}(L) = L^\star$ if $\min_{\text{lex}}(L)$ is primitive and prefix distinguishable in L.

Theorem 3.1. *Let $L \subseteq \Sigma^+$ be a language such that $\sharp L > 1$ and $u = \min_{lex}(L)$ is root prefix distinguishable. Then $\mathcal{C}(L) = L^\star$.*

Proof. We show that if $\mathcal{C}(L) \neq L^\star$ then $\mathcal{C}(L)L \neq L\mathcal{C}(L)$ since we can find a word in $\mathcal{C}(L)L$ which has not prefixes in L.

Let $y \in \mathcal{C}(L) \setminus L^\star$, $d = |u|$ and $e = |y|$. Clearly, $uy = y_1\alpha_1$ for suitable $y_1 \in \mathcal{C}(L)$, $\alpha_1 \in L$. Note that if $y_1 \in L^\star$ then $uy \in LL^\star$ and so, by Lemma 2.2, y would belong to $L^\star$. Then, it is immediate to see that for any $n \geq e + d$ there are $\alpha_1, \alpha_2, \ldots, \alpha_n \in L$ such that

$$u^n y = u^{n-1}y_1\alpha_1 = u^{n-2}y_2\alpha_2\alpha_1 = \ldots = y_n\alpha_n \ldots \alpha_2\alpha_1$$

with

- $y_n = u^m v \in \mathcal{C}(L) \setminus L^\star$ for some $0 \leq m \leq n-2$;
- $u = vw$ $(v, w \neq \epsilon)$;
- $z = w(vw)^{n-m-1}y = \alpha_n \cdots \alpha_2\alpha_1 \in L^\star$.

Now, consider the word $y_n u = u^m vu \in \mathcal{C}(L)L = L\mathcal{C}(L)$. By Lemma 2.2, we first get $u^{m-1}vu \in \mathcal{C}(L)$ and then, after m steps,

$$\hat{y} = vu^m = v(vw)^m \in \mathcal{C}(L).$$

Moreover, we have $\hat{y}u = \hat{\alpha}\tilde{y}$ for suitable $\hat{\alpha} \in L$ and $\tilde{y} \in \mathcal{C}(L)$, with $u \leq_{\text{lex}} \hat{\alpha}$ and $\hat{\alpha} \leq v(vw)^{m+1}$. Note that v is a common prefix of u and $\hat{\alpha}$ since $\hat{\alpha} \not< u$. So, let $k = |w|$ and observe that

$$w \leq_{\text{lex}} \text{pref}_k(vw),$$

otherwise $\hat{\alpha} <_{\text{lex}} u$. By considering the previously defined word $z \in L^\star$, we have

$$\text{pref}_k(vw) \leq_{\text{lex}} \text{pref}_k(z) = w$$

and then $w = \text{pref}_k(vw)$. This means that w is both a suffix and a prefix of u, that is, $vw = wx$ with $x = \text{suf}_{d-k}(vw)$.

Now, note that z can not have a prefix which is lexicographically smaller than $u = wx$ and so, since wv is a prefix of z (with $|v| = |x|$), we get $x \leq_{\text{lex}} v$. Moreover, by Lemma 2.2, from $\hat{y}u = v(vw)^{m+1} = v(wx)^{m+1} = \hat{\alpha}\tilde{y} \in L\mathcal{C}(L)$ we obtain $\tilde{y} = x(wx)^m \in \mathcal{C}(L)$. Then, the word $x(wx)^{m+1} \in \mathcal{C}(L)L$ has a prefix in L which is not smaller than v, that is, $v \leq_{\text{lex}} x$. So, we have $v = x$ and $u = vw = wv$. Thus, by Theorem 2.1, we can find a primitive word ρ and $p, q \in \mathbb{N}$ such that

$$v = \rho^p, \quad w = \rho^q, \quad u = \rho^{p+q}.$$

Note that ρ is the primitive root of u since the primitive root of a word is unique.

Finally, we consider the word $y_n = \rho^{(p+q)m+p}$. By recalling that u is root prefix distinguishable in L and $\sharp L > 1$, for any $\beta \in L \setminus \{u\}$ we have $\rho \not< \beta$ and we can write the equalities

$$\begin{aligned} y_n\beta &= uz_1 \\ z_1\beta &= uz_2 \\ &\vdots \\ z_{m-1}\beta &= uz_m \end{aligned}$$

with $z_m = \rho^p\beta^m \in \mathcal{C}(L)$. Therefore, the word $z_m\beta = \rho^p\beta^{m+1}$ belongs to $\mathcal{C}(L)L$ and for any $\gamma \in L$ we have $\gamma \not\leq z_m\beta$ (since the word ρ is prefix distinguishable in $(L \setminus \{\rho^{p+q}\}) \cup \{\rho\}$). This is a contradiction since $\mathcal{C}(L)L = L\mathcal{C}(L)$. □

An immediate consequence of the previous theorem is given by

Corollary 3.2. *Let $L \subseteq \Sigma^+$, $\sharp L > 1$, be a language such that there is $w \in L$ with*

$$\mathit{pref}_1(w) \neq \mathit{pref}_1(y)$$

for any $y \in L \setminus \{w\}$. Then $\mathcal{C}(L) = L^\star$.

Proof. It is immediate to see that w is root prefix distinguishable in L. Then, by choosing a suitable order on Σ such that $\text{pref}_1(w) = \min(\Sigma)$, it turns out that w is the smallest word of L. Then, the result follows from Theorem 3.1. □

We point out that Corollary 3.2 leads to a simple proof of a known result about the centralizer of a three-word language [9]. In fact we can state:

Corollary 3.3. *Let $L \subseteq \Sigma^+$ be a three-word language which is not periodic. Then*

$$\mathcal{C}(L) = L^\star.$$

Proof. If L is branching then, since $\sharp L = 3$, there is $v \in L$ such that $\mathrm{pref}_1(v) \neq \mathrm{pref}_1(y)$ for any $y \in L \setminus \{v\}$. So, the result follows from Corollary 3.2. Otherwise, we determine the branching language $\hat{L}$ associated with L by Theorem 2.3. Since $\hat{L}$ can be chosen such that $\sharp\hat{L} = \sharp L = 3$, we have $\mathcal{C}(\hat{L}) = \hat{L}^\star$ and then $\mathcal{C}(L) = L^\star$. □

4. Conclusions and open problems

We conclude by observing that the conditions given in this paper are not necessary. We first consider a trivial example.

Example 4.1. Let us consider the three-word code $L = \{a, aba, ababa\}$ which does not satisfy the conditions of Theorem 3.1 or Corollary 3.2. Nevertheless, Corollary 3.3 states that its centralizer is $\{a, aba, ababa\}^\star$. In fact, L is not branching and we can consider its circular shift, $L^{\hookrightarrow} = \{a, baa, babaa\}$. Since $L^{\hookrightarrow}$ satisfies the conditions of Corollary 3.2, we get $\mathcal{C}(L^{\hookrightarrow}) = \{a, baa, babaa\}^\star$. Then, by Lemma 2.4, we have $\mathcal{C}(L) = a\{a, baa, babaa\}^\star a^{-1} = \{a, aba, ababa\}^\star$.

A more interesting case consists of the following:

Example 4.2. Let $L = \{a^iba^i, b^iab^i \mid i > 0\}$. This is a primitive language which is also a prefix code, so we have $\mathcal{C}(L) = L^\star$. Note that Theorem 3.1 can not be applied since $\min_{\mathrm{lex}}(L)$ is not defined (independently of the order $\prec$ on $\{a, b\}$, for any fixed $w \in L$ we can find $y \in L$ with $y <_{\mathrm{lex}} w$). Observe that L is branching and that circular shift can not be used.

In particular, note that circular shift can be successfully applied to all three-word languages, while the same assertion is not true for four-word languages.

Example 4.3. Let $L = \{aaa, bbb, ab, ba\}$. L is a prefix code and primitive, so $\mathcal{C}(L) = L^\star$. Note that either $\min_{\mathrm{lex}}(L) = aaa$ or $\min_{\mathrm{lex}}(L) = bbb$, depending on whether $a \prec b$ or $b \prec a$. In both cases $\min_{\mathrm{lex}}(L)$ is not root prefix distinguishable. Moreover, circular shift can not be used (L is branching).

So, while the problem of characterizing the class of languages L with $\mathcal{C}(L) = L^\star$ is still open, it is quite natural to look for other sufficient conditions, possibly weaker than those we have presented here.

In particular, the language $L = \{aa, ab, ba, bb\}$ shows that something more than the existence of a root prefix distinguishable word is needed. In fact, ab is root prefix distinguishable in L, but $\mathcal{C}(L) = \rho(L)^\star$ with $\rho(L) = \{a, b\}$. So, the minimality with respect to a lexicographic order might be replaced with some weaker condition.

References

[1] J. Berstel and D. Perrin, *Theory of codes*. Academic Press, New York (1985).

[2] J.A. Brzozowski and E. Leiss, On equations for regular languages, finite automata, and sequential networks. *Theor. Comp. Sci.* **10** (1980) 19–35.

[3] C. Choffrut, J. Karhumäki and N. Ollinger, The commutation of finite sets: a challenging problem. *Theor. Comp. Sci.* **273** (2002) 69–79.

[4] N. Chomsky and M.P. Schützenberger, The algebraic theory of context-free languages. *Computer Programming and Formal Systems*, edited by P. Braffort and D. Hirschberg. North-Holland, Amsterdam (1963) 118–161.

[5] J.H. Conway, *Regular Algebra and Finite Machines*. Chapman & Hall, London (1971).

[6] J. Karhumäki and I. Petre, The branching point approach to Conway's problem, in *Formal and Natural Computing*, edited by W. Brauer, H. Ehrig, J. Karhumäki, A. Salomaa. *Lect. Notes Comput. Sci.* **2300** (2002) 69–76.

[7] J. Karhumäki and I. Petre, Conway's problem for three-word sets. *Theor. Comp. Sci.* **289** (2002) 705–725.

[8] J. Karhumäki, M. Latteux and I. Petre, Commutation with codes. *Theor. Comp. Sci.* **340** (2005) 322–333.

[9] J. Karhumäki, M. Latteux and I. Petre, Commutation with ternary sets of words. *Theory Comput. Syst.* **38** (2005) 161–169.

[10] M. Kunc, The power of commuting with finite sets of words, in *Proc. of STACS 2005*. *Lect. Notes Comput. Sci.* **3404** (2005) 569–580.

[11] R.C. Lyndon and M.P. Schützenberger, The equation $a^m = b^n c^p$ in a free group. *Michigan Math. J.* **9** (1962) 289–298.

[12] P. Massazza, On the equation $XL = LX$, in *Proc. of WORDS 2005*, Publications du Laboratoire de Combinatoire et d'Informatique Mathématique, Montréal **36** (2005) 315–322.

[13] D. Perrin, Codes conjugués. *Inform. Control* **20** (1972) 222–231.

[14] B. Ratoandromanana, Codes et motifs. *RAIRO-Inf. Theor. Appl.* **23** (1989) 425–444.