Regular Languages are Church-Rosser Congruential

Volker Diekert *
Manfred Kufleitner *
Klaus Reinhardt †
Tobias Walter *

* Institut für Formale Methoden der Informatik
University of Stuttgart, Germany

† Wilhelm-Schickard-Institut für Informatik
University of Tübingen, Germany

Abstract. This paper proves a long standing conjecture in formal language theory. It shows that all regular languages are Church-Rosser congruential. The class of Church-Rosser congruential languages was introduced by McNaughton, Narendran, and Otto in 1988. A language L is Church-Rosser congruential, if there exists a finite confluent, and length-reducing semi-Thue system S such that L is a finite union of congruence classes modulo S. It was known that there are deterministic linear context-free languages which are not Church-Rosser congruential, but on the other hand it was strongly believed that all regular language are of this form. Actually, this paper proves a more general result. [1]

Keywords. String rewriting; Church-Rosser system; regular language; finite monoid; finite semigroup; local divisor.

[1] The research on this paper was initiated during the program *Automata Theory and Applications* at the Institute for Mathematical Sciences, National University of Singapore in September 2011. The second author was supported by the German Research Foundation (DFG) under grant DI 435/5-1.

1 Introduction

It has been a long standing conjecture in formal language theory that all regular languages are Church-Rosser congruential. The class of Church-Rosser congruential languages was introduced by McNaughton, Narendran, and Otto in 1988 [8]. A language L is Church-Rosser congruential, if there exists a finite confluent, and length-reducing semi-Thue system S such that L is a finite union of congruence classes modulo S. One of the main motivations to consider this class of languages is that the membership problem for L can be solved in linear time; this is done by computing normal forms using the system S, followed by a table look-up. For this it is not necessary that the quotient monoid A^*/S is finite, it is enough that L is a finite union of congruence classes modulo S. It is not hard to see that $\{a^n b^n \mid n \in \mathbb{N}\}$ is Church-Rosser congruential, but $\{a^m b^n \mid m, n \in \mathbb{N} \text{ and } m \geq n\}$ is not. This led the authors of [8] to the more technical notion of Church-Rosser languages; this class of languages captures all deterministic context-free languages. For more results about Church-Rosser languages see e.g. [2, 9, 14, 15].

From the very beginning it was strongly believed that all regular languages are Church-Rosser congruential in the pure sense. However, after some significant initial progress [9, 10, 11, 12, 13] there was some stagnation.

Before 2011 the most advanced result was the one announced in 2003 by Reinhardt and Thérien [13]. According to this manuscript the conjecture is true for all regular languages where the syntactic monoid is a group. However, the manuscript has never been published as a refereed paper and there are some flaws in its presentation. The main problem with [13] has however been quite different for us. The statement is too weak to be useful in the induction for the general case. So, instead of being able to use [13] as a black box, we shall prove a more general result in the setting of weight-reducing systems. This part about group languages is a cornerstone in our approach.

The other ingredient to our paper has been established only very recently. Knowing that the result is true if the the syntactic monoid is a group, we started looking at aperiodic monoids. Aperiodic monoids correspond to star-free languages and the first two authors together with Weil proved that all star-free languages are Church-Rosser congruential [5]. Our proof became possible by *loading the induction hypothesis*. This means we proved a much stronger statement. We showed that for every star-free language $L \subseteq A^*$ there exists a finite confluent semi-Thue system $S \subseteq A^* \times A^*$ such that the quotient monoid A^*/S is finite (and aperiodic), L is a union of congruence classes modulo S, and moreover all right-hand sides of rules appear as scattered subwords in the corresponding left-hand side. We called the last property *subword-reducing*, and it is obvious that every subword-reducing system is length-reducing.

We have little hope that such a strong result could be true in general. Indeed here we step back from subword-reducing to weight-reducing systems.

We prove in Theorem 6 the following result: Let $L \subseteq A^*$ be a regular language and $\|a\| \in \mathbb{N} \setminus \{0\}$ be a positive weight for every letter $a \in A$ (e.g., $\|a\| = |a| = 1$). Then we can construct for the given weight a finite, confluent and weight-reducing semi-Thue system $S \subseteq A^* \times A^*$ such that the quotient monoid A^*/S is finite and recognizes L. In particular, L is a finite union of congruence classes modulo S.

Note that this gives us another characterization for the class of regular languages. By Corollary 7 we see that a language $L \subseteq A^*$ is regular if and only if L is recognized by a finite Church-Rosser system S with finite index. As a consequence, a long standing conjecture about regular languages has been solved positively.

2 Preliminaries

Words and languages Throughout this paper, A is a finite alphabet. An element of A is called a *letter*. The set A^* is the free monoid generated by A. It consists of all finite sequences of letter from A. The elements of A^* are called *words*. The empty word is denoted by 1. The *length* of a word u is denoted by $|u|$. We have $|u| = n$ for $u = a_1 \cdots a_n$ where $a_i \in A$. The empty word has length 0, and it is the only word with this property. The set of word of length at most n is denoted by $A^{\leq n}$, and the set of all nonempty words is A^+. We generalize the length of a word by introducing weights. A *weighted alphabet* $(A, \|\cdot\|)$ consists of an alphabet A equipped with a weight function $\|\cdot\| : A \to \mathbb{N} \setminus \{0\}$. The *weight* of a letter $a \in A$ is $\|a\|$ and the *weight* $\|u\|$ of a word $u = a_1 \cdots a_n$ with $a_i \in A$ is $\|a_1\| + \cdots + \|a_n\|$. The weight of the empty word is 0. The length is the special weight with $\|a\| = 1$ for all $a \in A$. A word u is a *factor* of a word v if there exist $p, q \in A^*$ such that $puq = v$, and u is a *proper factor* of v if $pq \neq 1$. The word u is a *prefix* of v if $uq = v$ for some $q \in A^*$, and it is a *suffix* of v if $pu = v$ for some $p \in A^*$. We say that u is a factor (resp. prefix, resp. suffix) of v^+ if there exists $n \in \mathbb{N}$ such that u is a factor (resp. prefix, resp. suffix) of v^n. Two words $u, v \in A^*$ are *conjugate* if there exist $p, q \in A^*$ such that $u = pq$ and $v = qp$. An integer $m > 0$ is a *period* of a word $u = a_1 \cdots a_n$ with $a_i \in A$ if $a_i = a_{i+m}$ for all $1 \leq i \leq n - m$. A word $u \in A^+$ is *primitive* if there exists no $v \in A^+$ such that $u = v^n$ for some integer $n > 1$. It is a standard fact that a word u is not primitive if and only $u^2 = puq$ for some $p, q \in A^+$. This follows immediately from the result from combinatorics on words that $xy = yx$ if and only if x and y are powers of a common root; see e.g. [7, Section 1.3].

A monoid M *recognizes* a language $L \subseteq A^*$ if there exists a homomorphism $\varphi : A^* \to M$ such that $L = \varphi^{-1}\varphi(L)$. A language $L \subseteq A^*$ is *regular* if it is recognized by a finite monoid. There are various other and well-known characterizations of regular languages; e.g., regular expressions, finite automata or monadic second order logic. Regular languages L can be classified in terms of structural properties of the monoids recognizing L. In particular, we consider group languages; these are languages recognized by finite groups.

Semi-Thue systems A *semi-Thue system* over A is a subset $S \subseteq A^* \times A^*$. In this paper, all semi-Thue systems are finite. The elements of S are called *rules*. We frequently write $\ell \to r$ for rules (ℓ, r). A system S is called *length-reducing* if we have $|\ell| > |r|$ for all rules $\ell \to r$ in S. It is called *weight-reducing* with respect to some weighted alphabet $(A, \|\cdot\|)$, if $\|\ell\| > \|r\|$ for all rules $\ell \to r$ in S. Every system S defines the rewriting relation $\underset{S}{\Longrightarrow} \subseteq A^* \times A^*$ by setting $u \underset{S}{\Longrightarrow} v$ if there exist $p, q, \ell, r \in A^*$ such that $u = p\ell q$, $v = prq$, and $\ell \to r$ is in S.

By $\overset{*}{\underset{S}{\Longrightarrow}}$ we mean the reflexive and transitive closure of $\underset{S}{\Longrightarrow}$. By $\overset{*}{\underset{S}{\Longleftrightarrow}}$ we mean the symmetric, reflexive, and transitive closure of $\underset{S}{\Longrightarrow}$. We also write $u \overset{*}{\underset{S}{\Longleftarrow}} v$ whenever $v \overset{*}{\underset{S}{\Longrightarrow}} u$. The system S is *confluent* if for all $u \overset{*}{\underset{S}{\Longleftrightarrow}} v$ there is some w such that $u \overset{*}{\underset{S}{\Longrightarrow}} w \overset{*}{\underset{S}{\Longleftarrow}} v$. It is *locally confluent* if for all $v \underset{S}{\Longleftarrow} u \underset{S}{\Longrightarrow} v'$ there exists w such that $v \overset{*}{\underset{S}{\Longrightarrow}} w \overset{*}{\underset{S}{\Longleftarrow}} v'$. If S is locally confluent and weight-reducing for some weight, then S is confluent; see e.g. [1, 6]. Note that $u \underset{S}{\Longrightarrow} v$ implies that $\|u\| > \|v\|$ for weight-reducing systems. The relation $\overset{*}{\underset{S}{\Longleftrightarrow}} \subseteq A^* \times A^*$ is a congruence, hence the congruence classes $[u]_S = \{v \in A^* \mid u \overset{*}{\underset{S}{\Longleftrightarrow}} v\}$ form a monoid which is denoted by A^*/S. The size of A^*/S is called the *index of* S. A finite semi-Thue system S can be viewed as a finite set of defining relations. Hence, A^*/S becomes a finitely presented monoid. By $\mathrm{IRR}_S(A^*)$ we denote the set of irreducible words in A^*, i.e., the set of words where no left-hand side occurs as a factor.

Whenever the weighted alphabet $(A, \|\cdot\|)$ is fixed, a finite semi-Thue system $S \subseteq A^* \times A^*$ is called a *weighted Church-Rosser system* if it is finite, weight-reducing for $(A, \|\cdot\|)$, and confluent. Hence, a finite semi-Thue system S is a weighted Church-Rosser system if and only if (1) we have $\|\ell\| > \|r\|$ for all rules $\ell \to r$ in S and (2) every congruence class has exactly one irreducible element. In particular, for weighted Church-Rosser systems S, there is a one-to-one correspondence between A^*/S and $\mathrm{IRR}_S(A^*)$. A *Church-Rosser system* is a finite, length-reducing, and confluent semi-Thue system. In particular, every Church-Rosser system is a weighted Church-Rosser system. A language $L \subseteq A^*$ is called a *Church-Rosser congruential language* if there is a finite Church-Rosser system S such that L can be written as a finite union of congruence classes $[u]_S$.

Definition 1. *Let $\varphi : A^* \to M$ be a homomorphism and let S be a semi-Thue* *system. We say that* φ factorizes through S *if for all* $u, v \in A^*$ *we have:*

$$u \underset{S}{\overset{*}{\Longleftrightarrow}} v \quad \textit{implies} \quad \varphi(u) = \varphi(v).$$

Note that if S is a semi-Thue system and $\varphi : A^* \to M$ factorizes through S, then the following diagram commutes:

$$\begin{tikzcd} & A^*/S \arrow[d, "\psi"] \\ A^* \arrow[ur, "\pi"] \arrow[r, "\varphi"] & M \end{tikzcd}$$

Here, $\pi(u) = [u]_S$ is the canonical homomorphism and $\psi([u]_S) = \varphi(u)$.

3 Finite Groups

Our main result is that every homomorphism $\varphi : A^* \to M$ to finite monoid M factorizes through a Church-Rosser system S. Our proof of this theorem distinguishes whether or not M is a group. Thus, we first prove this result for groups. Before we turn to the general case, we show that for some particular groups, proving the claim is easy. The techniques developed here will also be used when proving the result for arbitrary finite groups.

3.1 Groups without proper cyclic quotient groups

The aim of this section is to show that finding a Church-Rosser system is very easy for many cases. This list includes systems of all finite (non-cyclic) simple groups, but it goes far beyond this. Let $\varphi : A^* \to G$ be a homomorphism to a finite group, where $(A, \|\cdot\|)$ is a weighted alphabet. This defines a regular language $L_G = \{w \in A^* \mid \varphi(w) = 1\}$. Let us assume that the greatest common divisor $\gcd \{\|w\| \mid w \in L_G\}$ is equal to one; e.g. $\{6, 10, 15\} \subseteq \{\|w\| \mid w \in L_G\}$. Then there are two words $u, v \in L_G$ such that $\|u\| - \|v\| = 1$. Now we can use these words to find a constant d such that all $g \in G$ have a representing word v_g with the exact weight $\|v_g\| = d$. To see this, start with some arbitrary set of representing words v_g. We multiply words v_g with smaller weight with u and words v_g higher weights with v until all weights are equal.

The final step is to define the following weight-reducing system

$$S_G = \left\{ w \to v_{\varphi(w)} \;\middle|\; w \in A^* \text{ and } d < \|w\| \leq d + \max \{\|a\| \mid a \in A\} \right\}.$$

Confluence of S_G is trivial; and every language recognized by φ is also recognized by the canonical homomorphism $A^* \to A^*/S_G$.

Now assume that we are not so lucky, i.e., $\gcd\{\|w\| \mid w \in L_G\} > 1$. This means there is a prime number p such that p divides $\|w\|$ for all $w \in L_G$. Then, the homomorphism of A^* to $\mathbb{Z}/p\mathbb{Z}$ defined by $a \mapsto \|a\| \bmod p$ factorizes through φ and $\mathbb{Z}/p\mathbb{Z}$ becomes a quotient group of G. This can never happen if G is simple and non-cyclic, because a simple group does not have any proper quotient group. But there are many other cases where a natural homomorphism $A^* \to G$ for some weighted alphabet $(A, \|\cdot\|)$ satisfies the property $\gcd\{\|w\| \mid w \in L_G\} = 1$ although G has a non-trivial cyclic quotient group. Just consider the length function and a presentation by standard generators for dihedral groups D_{2n} or the permutation groups $\mathcal{S}_n$ where n is odd.

For example, let $G = D_6 = \mathcal{S}_3$ be the permutation group of a triangle. Then G is generated by elements τ and ρ with defining relations

$$\tau^2 = \rho^3 = 1 \text{ and } \tau\rho\tau = \rho^2.$$

The following six words of length 3 represent all six group elements:

$$1 = \rho^3,\ \rho = \rho\tau^2,\ \rho^2 = \tau\rho\tau,\ \tau = \tau^3,\ \tau\rho = \rho^2\tau,\ \tau\rho^2.$$

The corresponding monoid $\{\rho, \tau\}^* / S_G$ has 15 elements.

It is much harder to find a Church-Rosser system for the homomorphism $\varphi : \{a, b, c\}^* \to \mathbb{Z}/3\mathbb{Z}$ where $\varphi(a) = \varphi(b) = \varphi(c) = 1 \bmod 3$. In some sense this phenomenon suggests that finite cyclic groups or more general commutative groups are the obstacle to find a simple construction for Church-Rosser systems.

3.2 The general case for group languages

In this section, we consider arbitrary groups. We start with some simple properties of Church-Rosser systems. Then, in Theorem 5, we state and prove that group languages are Church-Rosser congruential.

Lemma 2. *Let $(A, \|\cdot\|)$ be a weighted alphabet, let $d \in \mathbb{N}$, and let $S \subseteq A^* \times A^*$ be a weighted Church-Rosser system such that $\mathrm{IRR}_S(A^*)$ is finite. Then*

$$S_d = \left\{ u\ell v \to urv \;\middle|\; u, v \in A^d \text{ and } \ell \to r \in S \right\}$$

is a weighted Church-Rosser system satisfying:

1. *$\mathrm{IRR}_{S_d}(A^*)$ is finite.*
2. *All words of length at most $2d$ are irreducible with respect to S_d.*
3. *The mapping $[u]_{S_d} \mapsto [u]_S$ for $u \in A^*$ is well-defined and yields a surjective homomorphism from A^*/S_d onto A^*/S.*

Proof. First, one shows that local confluence of S transfers to local confluence of S_d. For "1" and "2" note that $\mathrm{IRR}_{S_d}(A^*) = A^{\leq 2d} \cup A^d \cdot \mathrm{IRR}_S(A^*) \cdot A^d$. The remaining proof is straightforward and therefore left to the reader. $\square$

Lemma 3. *Let $(A, \|\cdot\|)$ be a weighted alphabet and let $\Delta \subseteq A^+$ such that all words in Δ have length at most t. Then, for every $n \geq 1$, the set of rules*

$$T = \left\{\delta^{t+n} \to \delta^t \;\middle|\; \delta \in \Delta,\ \delta \text{ is primitive}\right\}$$

yields a weighted Church-Rosser system.

Proof. Every rule in T is weight-reducing. Thus it suffices to show that T is locally confluent. Let $\delta, \tilde{\delta} \in \Delta$ be primitive with $|\delta| \geq |\tilde{\delta}|$ and suppose $x\delta^{t+n} = \tilde{\delta}^{t+n}y$. If δ^{t+n} is a suffix of $\tilde{\delta}^t y$, then $\tilde{\delta}^{t+n}$ is a prefix of $x\delta^t$; and the two T-rules $\delta^{t+n} \to \delta^t$ and $\tilde{\delta}^{t+n} \to \tilde{\delta}^t$ can be applied independently of one another. Thus we can assume $|\delta^{t+n}| > |\tilde{\delta}^t y|$. In particular, $\tilde{\delta}^t$ is a factor of δ^+. Note that $|\tilde{\delta}^t| \geq |\delta|$. Thus $|\tilde{\delta}|$ is a period of δ.

Let us first consider the case $|\delta| > |\tilde{\delta}|$. Since δ is primitive, $|\tilde{\delta}|$ cannot be a divisor of $|\delta|$. In particular, we have $|\tilde{\delta}| \geq 2$. Suppose $|\tilde{\delta}| = 2$. Then $\delta = (ab)^m a$ for $a, b \in A$ and some $m \geq 1$. We conclude that the suffix $a\delta$ or the prefix δa of δ^2 is a factor of $\tilde{\delta}^+$. Since both words $a\delta$ and δa have a factor aa and $|\tilde{\delta}| = 2$, this contradicts $\tilde{\delta}$ being primitive. Therefore, we can assume $|\tilde{\delta}| \geq 3$ and hence, $|\tilde{\delta}^t| \geq |\delta^3|$. It follows that δ^2 is a factor of $\tilde{\delta}^+$ and $|\tilde{\delta}|$ is a period of δ^2. By shifting the prefix δ of δ^2 by this period, we can write $\delta^2 = p\delta q$ with $p, q \in A^+$ and $|p| = |\tilde{\delta}|$. We conclude that δ is not primitive, which is a contradiction.

Let now $|\delta| = |\tilde{\delta}|$. In this case, the words δ and $\tilde{\delta}$ are conjugate. Therefore, applying one of the rules $\delta^{t+n} \to \delta^t$ and $\tilde{\delta}^{t+n} \to \tilde{\delta}^t$ yields the same word. $\square$

Lemma 4. *Let $\Delta \subseteq A^+$ be a set of words such that all words in Δ have length at most n. If $u \in A^{>2n}$ is not a factor of some δ^+ for $\delta \in \Delta$, then there is a proper factor v of u which is also not a factor of some δ^+ for $\delta \in \Delta$.*

Proof. Assume that such a factor v of u does not exist. Let $u = awb$ for $a, b \in A$. Then aw is a factor of δ^+ and wb is a factor of δ'^+ for some $\delta, \delta' \in \Delta$. Let $p = |\delta|$ and $q = |\delta'|$. Now, p is a period of aw and q is a period of wb. Thus p and q are both periods of w. Since $|w| \geq 2n - 1 \geq p + q - \gcd(p, q)$, we see that $\gcd(p, q)$ is also a period of w by the Periodicity Lemma of Fine and Wilf [7, Section 1.3]. The $(p+1)$-th letter in aw is a. Going in steps $\gcd(p, q)$ to the left or to the right in w, we see that the $(q+1)$-th letter in aw is a. Thus awb is a factor of δ'^+, which is a contraction. $\square$

We are now ready to prove the main result of this section: Group languages are Church-Rosser congruential. An outline of the proof is as follows. By induction on the size of the alphabet, we show that every homomorphism $\varphi : A^* \to G$ factorizes through a weighted Church-Rosser system S with finite index. Remove some letter c from the alphabet A. This leads to a system R for the remaining letters B. Lemma 2 allows to assume that certain words are irreducible. Then we consider $K = \mathrm{IRR}_R(B^*)c$ which is a prefix code in A^*. We consider K as a new alphabet. Essentially, it is this situation where weighted alphabets come into play because we can choose the weight of K such that it is compatible with the weight over the alphabet A. Over K, we introduce two sets of rules T_Δ and T_Ω. The T_Δ-rules reduce long repetitions of short words Δ, and the T_Ω-rules have the form $\omega\, u\, \omega \to \omega\, v_g\, \omega$. Here, Ω is some finite set of markers and $\omega \in \Omega$ is such a marker. The word v_g is a normal form for the group element g. The T_Ω-rules reduce long words without long repetitions of short words. Then we show that T_Δ and T_Ω are confluent and that their union has finite index over K^*. Here, the confluence of the T_Δ-rules is Lemma 3. The confluence of the T_Ω-rules relies on several combinatorial properties of the normal forms v_g and the markers Ω. Using Lemma 4, we see that all sufficiently long words are reducible. Since by construction all rules in $T = T_\Delta \cup T_\Omega$ are weight-reducing, the system T is a weighted Church-Rosser system over K^* with finite index such $\varphi : K^* \to G$ factorizes through T. Since $K \subseteq A^*$, we can translate the rules $\ell \to r$ in T over K^* to rules $c\ell \to cr$ over A^*. This leads to the set of T'-rules over A^*. The letter c at the beginning of the T'-rules is require to shield from R-rules. Finally, we show that $S = R \cup T'$ is the desired system over A^*.

Theorem 5. *Let $(A, \|\cdot\|)$ be a weighted alphabet and let $\varphi : A^* \to G$ be a homomorphism to a finite group G. Then there exists a weighted Church-Rosser system S with finite index such that φ factorizes through S.*

Proof. In the following n denotes the exponent of G; this is the least positive integer n such that $g^n = 1$ for all $g \in G$. The proof is by induction on the size of the alphabet A. If $A = \{c\}$, then we set $S = \{c^n \to 1\}$. Let now $A = \{a_0, \ldots, a_s, c\}$ and let a_0 have minimal weight. We set $B = A \setminus \{c\}$. Let

$$\gamma_i = a_{i \bmod s}^{n+\lfloor i/s \rfloor}\, c.$$

Since A and $\{a_0 c, \ldots, a_s c, c\}$ generate the same subgroups of G and since every element $a_j c \in G$ occurs infinitely often as some γ_i, there exists $m > 0$ such that for every $g \in G$ there exists a word

$$v_g = \gamma_0^{n_0} \cdots \gamma_m^{n_m} \gamma_0$$

with $n_i > 0$ satisfying $\varphi(v_g) = g$ and $\|v_g\| - \|v_h\| < n\, \|a_0\|$ for all $g, h \in G$. The latter property relies on $\|\gamma_0\| + \|a_0\| = \|\gamma_s\|$ and pumping with γ_0^n and γ_s^n which both map to the neutral element of G: Assume $\|v_g\| - \|v_h\| \geq n \|a_0\|$ for some $g, h \in G$. Then we do the following. All v_g with maximal weight are multiplied by γ_0^n on the left, and for all other words v_h the exponent n_s of γ_s is replaced by $n_s + n$. After that, the maximal difference $\|v_g\| - \|v_h\|$ has decreased at least by 1 (and at most by $n \|a_0\|$). We can iterate this procedure until the weights of all v_g differ less than $n \|a_0\|$. Let

$$\Gamma = \{\gamma_0, \ldots, \gamma_m\}$$

be the generators of the v_g. By induction there exists a weighted Church-Rosser system R for the restriction $\varphi : B^* \to G$ satisfying the statement of the theorem. By Lemma 2, we can assume $\Gamma \subseteq \mathrm{IRR}_R(B^*)\, c$. Thus $v_g \in \mathrm{IRR}_R(A^*)$ for all $g \in G$. Let

$$K = \mathrm{IRR}_R(B^*)\, c.$$

The set K is a prefix code in A^*. We consider K as an extended alphabet and its elements as extended letters. The weight $\|u\|$ of $u \in K$ is its weight as a word over A. Each γ_i is a letter in K. The homomorphism $\varphi : A^* \to G$ can be interpreted as a homomorphism $\varphi : K^* \to G$; it is induced by $u \mapsto \varphi(u)$ for $u \in K$. The length lexicographic order on B^* induces a linear order $\leq$ on $\mathrm{IRR}_R(B^*)$ and hence also on K. Here, we assume $a_0 < \cdots < a_s$. The words v_g can be read as words over the weighted alphabet $(K, \|\cdot\|)$ satisfying the following five properties: First, v_g starts with the extended letter γ_0. Second, the last two extended letters of v_g are $\gamma_m \gamma_0$. Third, all extended letters in v_g are in non-decreasing order from left to right with respect to $\leq$, with the sole exception of the last letter γ_0 which is smaller than its predecessor γ_m. The fourth property is that all extended letters in v_g have a weight greater than $n \|a_0\|$. And the last important property is that all differences $\|v_g\| - \|v_h\|$ are smaller than $n \|a_0\|$. Let

$$\Delta = \left\{ \delta \in K^+ \;\middle|\; \delta \in K \text{ or } \|\delta\| \leq n \|a_0\| \right\}.$$

Note that Δ is closed under conjugation, i.e., if $uv \in \Delta$ for $u, v \in K^*$, then $vu \in \Delta$. We can think of Δ as the set of all "short" words. Choose $t \geq n$ such that all normal forms v_g have no factor δ^{t+n} for $\delta \in \Delta$ and such that $\|c^t\| \geq \|u\|$ for all $u \in K^{2n}$. Note that $c \in \Delta$ has the smallest weight among all words in Δ.

The first set of rules over the extended alphabet K deals with long repetitions of short words: The Δ-rules are

$$T_\Delta = \left\{ \delta^{t+n} \to \delta^t \;\middle|\; \delta \in \Delta \text{ and } \delta \text{ is primitive} \right\}.$$

Let $F \subseteq K^*$ contain all words which are a factor of some δ^+ for $\delta \in \Delta$ and let $J \subseteq K^+$ be minimal such that $K^* J K^* = K^* \setminus F$. By Lemma 4, we have $J \subseteq K^{2n}$. In particular, J is finite. Since J and Δ are disjoint, all words in J have a weight greater than $n\,\|a_0\|$. Let Ω contain all $\omega \in J$ such that $\omega \in \Gamma K^*$ implies $\omega = \gamma\gamma'$ for some $\gamma > \gamma'$, i.e.,

$$\Omega = J \cap \{\omega \in K^* \mid \omega \notin \Gamma K^* \text{ or } \omega = \gamma\gamma' \text{ for some } \gamma > \gamma'\}\,.$$

As we will see below, every sufficiently long word without long Δ-repetitions contains a factor $\omega \in \Omega$.

Claim 1. *There exists a bound $t' \in \mathbb{N}$ such that every word $u \in K^*$ with $\|u\| \geq t'$ contains a factor $\omega \in \Omega$ or a factor of the form δ^{t+n} for $\delta \in \Delta$.*

Proof of Claim 1. Let $t'' = (t+n+2) \cdot \max\{\|v\| \in \mathbb{N} \mid v \in K\}$. First, suppose $u \in K^* \setminus K^*\Gamma K^*$ and $\|u\| \geq t''$. If u is a factor of δ^+, then δ^{n+d} is a factor of u since $\|\delta\| \leq \max\{\|v\| \in \mathbb{N} \mid v \in K\}$. Thus we can assume $u \in K^* \setminus F$. By definition of J, the word u contains a factor $\omega \in J$. We have $\omega \in \Omega$ because u (and thus ω) has no factor in Γ.

If $u \in K^* b\gamma K^*$ for $b \in K \setminus \Gamma$ and $\gamma \in \Gamma$, then u contains a factor $\omega = b\gamma \in \Omega$. Similarly, if $u \in K^*\gamma\gamma' K^*$ for $\gamma, \gamma' \in \Gamma$ and $\gamma > \gamma'$, then u contains a factor $\omega = \gamma\gamma' \in \Omega$. Thus, if $u \in K^*\Gamma K^*$, then we can assume $u = \gamma_{i_1} \cdots \gamma_{i_k} u'$ with

- $\gamma_{i_j} \in \Gamma$ and $\gamma_{i_1} \leq \cdots \leq \gamma_{i_k}$, and
- $u' \notin K^*\Gamma K^*$ and $\|u'\| < t''$.

We set $t' = (t+n-1) \cdot |\Gamma| \cdot \max\{\|v\| \in \mathbb{N} \mid v \in \Gamma\} + 1 + t''$. If $\|u\| \geq t'$, then $k \geq (t+n-1) \cdot |\Gamma| + 1$. By the pigeon hole principle, there exists $\gamma \in \{\gamma_{i_1}, \ldots, \gamma_{i_k}\} \subseteq \Delta$ such that γ^{t+n} is a factor of u. This completes the proof of Claim 1. $\diamond$

Since Δ is closed under factors, u contains no factor of the form δ^{t+n} for $\delta \in \Delta$ if and only if $u \in \mathrm{IRR}_{T_\Delta}(K^*)$. In particular, it is no restriction to only allow primitive words from Δ in the rules T_Δ. Every sufficiently long word u' can be written as $u' = u_1 \cdots u_k$ with $\|u_i\| \geq t'$ and k sufficiently large. Thus, by repeatedly applying Claim 1, there exists a non-negative integer d_Ω such that every word $u' \in \mathrm{IRR}_{T_\Delta}(K^*)$ with $\|u'\| \geq t_\Omega$ contains two occurrences of the same $\omega \in \Omega$ which are far apart. More precisely, u' has a factor $\omega\, u\, \omega$ with $\|u\| > \|v_g\|$ for all $g \in G$.

This suggests rules of the form $\omega\, u\, \omega \to \omega\, v_{\varphi(u)}\, \omega$; but in order to ensure confluence we have to limit their use. For this purpose, we equip Ω with a linear order $\preceq$ such that $\gamma_m\gamma_0$ is the smallest element, and every element in $\Omega \cap K^+\gamma_0$ is smaller than all elements in $\Omega \setminus K^+\gamma_0$. By making t_Ω bigger, we can assume that every word u' with $\|u'\| \geq t_\Omega$ contains a factor $\omega\, u\, \omega$ such that

- $\|u\| > \|v_g\|$ for all $g \in G$, and
- for every factor $\omega' \in \Omega$ of $\omega\, u\, \omega$ we have $\omega' \preceq \omega$.

The following claim is one of the main reasons for using the above definition of the normal forms v_g, and also for excluding all words $\omega \in \Gamma K^*$ in the definition of Ω except for $\omega = \gamma\gamma' \in \Gamma^2$ with $\gamma > \gamma'$.

Claim 2. *Let $\omega, \omega' \in \Omega$ and $g \in G$. If $\omega\, v_g\, \omega \in K^*\omega' K^*$, then $\omega' \preceq \omega$.*

Proof of Claim 2. All normal forms v_g have $\gamma_m\gamma_0$ as a suffix. In addition, the word $\gamma_m\gamma_0$ is the only element in Ω which is a factor of some v_g for $g \in G$. The reason is that all other letters in v_g are in non-decreasing order whereas all $\gamma\gamma' \in \Omega$ are in decreasing order. In particular, if $\gamma_m\gamma_0\, v_g\, \gamma_m\gamma_0 \in K^*\omega' K^*$ for $\omega' \in \Omega$, then $\omega' = \gamma_m\gamma_0$, i.e., $\gamma_m\gamma_0$ is the only factor of $\gamma_m\gamma_0\, v_g\, \gamma_m\gamma_0$ which is in Ω.

Let now $\omega = b\gamma_0$ for $b \in K \setminus \{\gamma_0\}$. Note that $\omega \in \Omega$ and that all elements in $\Omega \cap K^+\gamma_0$ have this form. Then the set of factors of $\omega v_g \omega$ which are in Ω is $\{\gamma_m\gamma_0, \omega\}$. Since $\gamma_m\gamma_0$ is the smallest element with respect to $\preceq$, each of them satisfies the claim.

Next, suppose $\omega \in K^+b$ for $b \in K \setminus \{\gamma_0\}$. Then the set of factors of $\omega v_g \omega$ which are in Ω is $\{\gamma_m\gamma_0, b\gamma_0, \omega\}$. Since every element ending with γ_0 is smaller than any other element in Ω, the claim also holds in this case. This completes the proof of Claim 2. $\diamond$

We are now ready to define the second set of rules over the extended alphabet K. They are reducing long words without long repetitions of words in Δ. We set

$$T'_\Omega = \left\{ \omega\, u\, \omega \to \omega\, v_{\varphi(u)}\, \omega \;\middle|\; \begin{array}{l} \left\| v_{\varphi(u)} \right\| < \|u\| \le t_\Omega \text{ and} \\ \omega\, u\, \omega \text{ has no factor } \omega' \in \Omega \text{ with } \omega \prec \omega' \end{array} \right\}.$$

Whenever there is a shorter rule in $T'_\Omega \cup T_\Delta$ then we want to give preference to this shorter rule. Thus the Ω-rules are

$$T_\Omega = \left\{ \ell \to r \in T'_\Omega \;\middle|\; \begin{array}{l} \text{there is no rule } \ell' \to r' \in T'_\Omega \cup T_\Delta \\ \text{such that } \ell' \text{ is a proper factor of } \ell \end{array} \right\}.$$

Let now

$$T = T_\Delta \cup T_\Omega\,.$$

Claim 3. *The system T is locally confluent over K^*.*

Proof of Claim 3. The system T_Δ is confluent by Lemma 3. Suppose we can apply two rules $\ell \to r \in T_\Omega$ and $\ell' \to r' \in T_\Delta$. Then ℓ' is not a factor of ℓ. Let $\ell = \omega u \omega$. Since ω is not a factor of ℓ', it is possible to first apply $\ell \to r$ and then apply $\ell' \to r'$. Moreover, by choice of d we have $\|\omega\| \le \|r'\|$. Thus we also can first apply $\ell' \to r'$ and then $\ell \to r$.

If $u \in \mathrm{IRR}_{T_\Delta}(K^*)$ and $u \underset{T_\Omega}{\Longrightarrow} v$, then $v \in \mathrm{IRR}_{T_\Delta}(K^*)$ by definition of the normal forms v_g and the set Ω. Thus, it remains to show that T_Ω is locally confluent on $\mathrm{IRR}_{T_\Delta}(K^*)$. By minimality of J, no $\omega \in \Omega$ is a proper factor of another word $\omega' \in \Omega$. Let $\omega u \omega \to r$ and $\omega' u' \omega' \to r'$ be two Ω-rules with $\omega \neq \omega'$. By construction of T'_Ω, the left sides of both rules can overlap at most $\min\{|\omega|, |\omega'|\} - 1$ positions. Thus the two rules can always be applied independently of one another.

Let now $\omega u \omega \to \omega v_g \omega$ and $\omega u' \omega \to \omega v_h \omega$ be two Ω-rules. By construction of T_Ω, neither is $\omega u' \omega$ a proper factor of $\omega u \omega$ nor vice versa. If $x\omega = \omega y$ for some $x, y \in K^+$ with $\|x\| \leq n\,\|a_0\|$, then $x \in \Delta$ and ω is a prefix of x^+ which contradicts the definition of $J \subseteq K^* \setminus F$. Therefore, whenever $x\omega = \omega y$ for $x, y \in K^+$ then $\|x\| > n\,\|a_0\|$ and $\|y\| > n\,\|a_0\|$. Suppose now $x\omega u\omega = \omega u'\omega y = \omega u''\omega$ for $x, y \in K^+$. If $|x| \geq |\omega u|$, then the two rules can be applied independently of one another. Thus let $|x| < |\omega u|$. As seen before, we have $\|x\| > n\,\|a_0\|$ and $\|y\| > n\,\|a_0\|$. We will show

$$x\,\omega\,v_g\,\omega \;\underset{T_\Omega}{\overset{*}{\Longrightarrow}}\; \omega\,v_{\varphi(u'')}\,\omega \;\underset{T_\Omega}{\overset{*}{\Longleftarrow}}\; \omega\,v_h\,\omega\,y.$$

If $x\,\omega\,v_g\,\omega \in K^*\omega' K^*$ or $\omega\,v_h\,\omega\,y \in K^*\omega' K^*$, then by Claim 2 we have $\omega' \preceq \omega$. We can write $x\omega = \omega x'$. Since $\|x'\| = \|x\| > n\,\|a_0\|$, we have $\|x'v_g\| > n\,\|a_0\| + \|v_g\| > \|v_{g'}\|$ for every $g' \in G$. This relies on the fact that the weights all normal forms $v_{g'}$ differ less than $n\,\|a_0\|$. This shows that the weight of $x'v_g$ is sufficiently high. If $\|x'v_g\| > t_\Omega$, then by Claim 1 we have $x'v_g \underset{T_\Omega}{\overset{*}{\Longrightarrow}} x''$ such that $\|v_{g'}\| < \|x''\| \leq t_\Omega$ for every $g' \in G$. Therefore, without loss of generality we can assume that the weight of $x'v_g$ is not too high, i.e., $\|x'v_g\| \leq t_\Omega$. Since $\varphi(x'v_g) = \varphi(u'')$, we have $x\omega v_g\omega \underset{T_\Omega}{\overset{*}{\Longrightarrow}} \omega v_{\varphi(u'')}\omega$. Similarly, $\omega v_h \omega y \underset{T_\Omega}{\overset{*}{\Longrightarrow}} \omega v_{\varphi(u'')}\omega$. This completes the proof of Claim 3. $\diamond$

Since all rules in T are weight-reducing, local confluence implies confluence. Moreover, all rules $\ell \to r$ in T satisfy $\varphi(\ell) = \varphi(r)$. We conclude that T is a weighted Church-Rosser system such that K^*/T is finite and $\varphi : K^* \to G$ factorizes through T. Remember that every element in K^* can be read as a sequence of elements in A^*. Thus every $u \in K^*$ can be interpreted as a word $u \in A^*$. We use this interpretation in order to apply the rules in T to words in A^*; but in order to not destroy K-letters when applying rules in R, we have to guard the first K-letter of every T-rule by appending the letter c. This leads to the system

$$T' = \{c\ell \to cr \in A^* \times A^* \mid \ell \to r \in T\}.$$

Combining the rules R over the alphabet B with the T'-rules yields

$$S = R \cup T'.$$

Since left sides of R-rules and of T'-rules can not overlap, the system S is confluent. By definition, each S-rule is weight-reducing. This means that S is a weighted Church-Rosser system. We have

$$\mathrm{IRR}_S(A^*) \;=\; \mathrm{IRR}_R(B^*) \,\cup\, \mathrm{IRR}_R(B^*) \cdot \mathrm{IRR}_{T'}\Big(c\big(\mathrm{IRR}_R(B^*)c\big)^*\Big) \cdot \mathrm{IRR}_R(B^*).$$

Therefore $\mathrm{IRR}_S(A^*)$ and A^*/S are finite. Since all rules $\ell \to r$ in S satisfy $\varphi(\ell) = \varphi(r)$, the homomorphism φ factorizes through S. $\square$

4 Arbitrary Finite Monoids

This section contains the main result of this paper. We show that every homomorphism $\varphi : A^* \to M$ to finite monoid factorizes through a weighted Church-Rosser system S with finite index. The proof relies on Theorem 5 and on a construction called local divisors.

4.1 Local divisors

The notion of *local divisor* has turned out to be a rather powerful tool when using inductive proofs for finite monoids, see e.g. [3, 4, 5]. The same is true in this paper. The definition of a local divisor is as follows: Let M be a monoid and let $c \in M$. We equip $cM \cap Mc$ with a monoid structure by introducing a new multiplication $\circ$ as follows:

$$xc \circ cy = xcy.$$

It is straightforward to see that $\circ$ is well-defined and $(cM \cap Mc, \circ)$ is a monoid with neutral element c.

The following observation is crucial. If $1 \in cM \cap Mc$, then c is a unit. Thus if the monoid M is finite and c is not a unit, then $|cM \cap Mc| < |M|$. The set $M' = \{x \mid cx \in Mc\}$ is a submonoid of M, and $c\cdot : M' \to cM \cap Mc : x \mapsto cx$ is a surjective homomorphism. Since $(cM \cap Mc, \circ)$ is the homomorphic image of a submonoid, it is a divisor of M. We therefore call $(cM \cap Mc, \circ)$ the *local divisor* of M at c.

4.2 The main result

We are now ready to prove our main result: Every homomorphism $\varphi : A^* \to M$ to a finite monoid factorizes through a weighted Church-Rosser system S with finite index. The proof uses induction on the size of M and the size of A. If $\varphi(A^*)$ is a group, then we apply Theorem 5; and if $\varphi(A^*)$ is not a group, then we find a letter $c \in A$ such that c is not a unit. Thus in this case we can use local divisors.

Theorem 6. *Let $(A, \|\cdot\|)$ be a weighted alphabet and let $\varphi : A^* \to M$ be a homomorphism to a finite monoid M. Then there exists a weighted Church-Rosser system S of finite index such that φ factorizes through S.*

Proof. The proof is by induction on $(|M|, |A|)$ with lexicographic order. If $\varphi(A^*)$ is a group, then the claim follows by Theorem 5. If $\varphi(A^*)$ is not a group, then there exists $c \in A$ such that $\varphi(c)$ is not a unit. Let $B = A \setminus \{c\}$. By induction on the size of the alphabet there exists a weighted Church-Rosser system R for the restriction $\varphi : B^* \to M$ satisfying the statement of the theorem. Let

$$K = \mathrm{IRR}_R(B^*)c.$$

We consider the prefix code K as a weighted alphabet. The weight of a letter $uc \in K$ is the weight $\|uc\|$ when read as a word over the weighted alphabet $(A, \|\cdot\|)$. Let $M_c = \varphi(c)M \cap M\varphi(c)$ be the local divisor of M at $\varphi(c)$. We let $\psi : K^* \to M_c$ be the homomorphism induced by $\psi(uc) = \varphi(cuc)$ for $uc \in K$. By induction on the size of the monoid there exists a weighted Church-Rosser system $T \subseteq K^* \times K^*$ for ψ satisfying the statement of the theorem. Suppose $\psi(\ell) = \psi(r)$ for $\ell, r \in K^*$ and let $\ell = u_1c \cdots u_jc$ and $r = v_1c \cdots v_kc$ with $u_i, v_i \in \mathrm{IRR}_R(B^*)$. Then

$$\begin{aligned}\varphi(c\ell) &= \varphi(cu_1c) \circ \cdots \circ \varphi(cu_jc)\\ &= \psi(u_1c) \circ \cdots \circ \psi(u_jc)\\ &= \psi(\ell) = \psi(r) = \varphi(cr).\end{aligned}$$

This means that every T-rule $\ell \to r$ yields an φ-invariant rule $c\ell \to cr$. Thus we can transform the system $T \subseteq K^* \times K^*$ for ψ into a system $T' \subseteq A^* \times A^*$ for φ by

$$T' = \{c\ell \to cr \in A^* \times A^* \mid \ell \to r \in T\}.$$

Since T is confluent and weight-reducing over K^*, the system T' is confluent and weight-reducing over A^*. Combining R and T' leads to

$$S = R \cup T'.$$

The left sides of a rule in R and a rule in T' cannot overlap. Therefore, S is a weighted Church-Rosser system such that φ factorizes through A^*/S. Suppose that every word in $\mathrm{IRR}_T(K^*)$ has length at most k. Here, the length is over the extended alphabet K. Similarly, let every word in $\mathrm{IRR}_R(B^*)$ have length at most m. Then

$$\mathrm{IRR}_S(A^*) \subseteq \{u_0cu_1 \cdots cu_{k'+1} \mid u_i \in \mathrm{IRR}_R(B^*),\ k' \leq k\}$$

and every word in $\mathrm{IRR}_S(A^*)$ has length at most $(k+2)m$. In particular $\mathrm{IRR}_S(A^*)$ and A^*/S are finite. □

The following corollary is a straightforward translation of the result in Theorem 6 about homomorphisms to a statement about regular languages.

Corollary 7. *A language $L \subseteq A^*$ is regular if and only if there exists a Church-Rosser system S of finite index such that $L = \bigcup_{u \in L}[u]_S$.*

Proof. If L is regular, then there exists a homomorphism $\varphi : A^* \to M$ recognizing L. By Theorem 6 there exists a finite Church-Rosser system S of finite index such that φ factorizes through S. The latter property implies $\varphi^{-1}(x) = \bigcup_{u \in \varphi^{-1}(x)}[u]_S$ for every $x \in M$. Thus $L = \bigcup_{x \in \varphi(L)} \varphi^{-1}(x) = \bigcup_{u \in L}[u]_S$. The converse is trivial. □

In particular, we see that all regular languages are Church-Rosser congruential.

References

[1] R. Book and F. Otto. *String-Rewriting Systems.* Springer-Verlag, 1993.

[2] G. Buntrock and F. Otto. Growing context-sensitive languages and Church-Rosser languages. *Information and Computation*, 141:1–36, 1998.

[3] V. Diekert and P. Gastin. First-order definable languages. In *Logic and Automata: History and Perspectives*, Texts in Logic and Games, pages 261–306. Amsterdam University Press, 2008.

[4] V. Diekert, M. Kufleitner, and B. Steinberg. The Krohn-Rhodes Theorem and Local Divisors. *ArXiv e-prints*, Nov. 2011.

[5] V. Diekert, M. Kufleitner, and P. Weil. Star-free languages are Church-Rosser congruential. *Theoretical Computer Science*, 2012. DOI: 10.1016/j.tcs.2012.01.028.

[6] M. Jantzen. *Confluent String Rewriting*, volume 14 of *EATCS Monographs on Theoretical Computer Science.* Springer-Verlag, 1988.

[7] M. Lothaire. *Combinatorics on Words*, volume 17 of *Encyclopedia of Mathematics and its Applications.* Addison-Wesley, Reading, MA, 1983. Reprinted by *Cambridge University Press*, 1997.

[8] R. McNaughton, P. Narendran, and F. Otto. Church-Rosser Thue systems and formal languages. *J. ACM*, 35(2):324–344, 1988.

[9] P. Narendran. *Church-Rosser and related Thue systems.* Doctoral dissertation, Dept. of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY, USA, 1984.

[10] G. Niemann. *Church-Rosser Languages and Related Classes.* Kassel University Press, 2002. PhD thesis.

[11] G. Niemann and F. Otto. The Church-Rosser languages are the deterministic variants of the growing conext-sensitive languages. *Inf. Comput.*, 197:1–21, 2005.

[12] G. Niemann and J. Waldmann. Some regular languages that are Church-Rosser congruential. In *Revised Papers from the 5th International Conference on Developments in Language Theory*, DLT '01, pages 330–339, London, UK, 2002. Springer-Verlag.

[13] K. Reinhardt and D. Thérien. Some more regular languages that are Church Rosser congruential. In *13. Theorietag, Automaten und Formale Sprachen, Herrsching, Germany*, pages 97–103, 2003.

[14] J. R. Woinowski. *Church-Rosser Languages and Their Application to Parsing Problems.* PhD thesis, TU Darmstadt, 2001.

[15] J. R. Woinowski. The context-splittable normal form for Church-Rosser language systems. *Inf. Comput.*, 183:245–274, 2003.