Context-Free Languages and Primitive Words

Pál Dömösi
Masami Ito

World Scientific

Context-Free Languages and Primitive Words

Pál Dömösi (Nyíregyháza College, Hungary)
Masami Ito (Kyoto Sangyo University, Japan)

World Scientific

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

Published by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

Library of Congress Cataloging-in-Publication Data
Dömösi, Pál, author.
Context-free languages and primitive words / by Pál Dömösi (Nyíregyháza College, Hungary), Masami Ito (Kyoto Sangyo University, Japan).
pages cm
Includes bibliographical references and index.
ISBN 978-9814271660 (hardcover : alk. paper)
1. Computer science--Philosophy. 2. Combinatorial analysis. 3.Word (Linguistics) 4. Context (Linguistics) 5. Coding theory. 6. Number theory. I. Ito, Masami, 1941– author. II. Title.
QA76.167.D66 2014
003'.54--dc23

2014024517

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

Printed by FuIsland Offset Printing (S) Pte Ltd Singapore

Context-Free Languages and Primitive Words

In Memory of Our Late Good Friend and Teacher
Professor István Peák

Preface

This monograph is an attempt to give an overview of the theory of context-free languages and also the most important results on combinatorics of words in relation to primitive words.

Combinatorial properties of words play an important role in mathematics and theoretical computer science. One of the well-known open problems is related to the language of primitive words. A word is called *primitive* if it is not a repetition of another word. (Thus the empty word is non-primitive.)

We conjectured that the language Q of all primitive words over a non-singleton alphabet is not context-free (P. Dömösi, S. Horváth, M. Ito [1991]). The problem seems to be simple but we have not yet found the solution.

Apart from the conditions of the Wise lemma (D. S. Wise [1976]), Q has all the well-known iteration conditions of context-free languages (P. Dömösi, S. Horváth, M. Ito, L. Kászonyi, M. Katsura [1992,1993]).[1] Another test of context-freeness is the so-called Interchange lemma (W. Ogden, R. J. Ross, K. Winklmann [1982]). It is also proved that Q fulfils the conditions of this test (S. Horváth [1995]). Therefore, Q resists almost all well-known tests of context-freeness.

It is also well-known that the intersection of a regular and a context-free language is again a context-free language. Therefore, if we find a regular language R such that $R \cap Q$ is not context-free then we can show that Q is not context-free. By some results of L. Kászonyi and M. Katsura [1996, 1997, 1999a, 1999b], this approach also seems to be hopeless.

Perhaps an appropriate homomorphic characterization of languages (see N. Chomsky and M. P. Schützenberger [1963], R. J. Stanley [1965], S. Hirose

[1] Note that the applicability problem of the Wise lemma is equivalent to the original problem.

and M. Yoneda [1985], P. Dömösi and S. Okawa [2003]) could help to prove our conjecture about the context-freeness of Q. Another possible direction of research to prove or disprove our conjecture is to follow the approach to formal language theory by means of Kolmogorov complexity started by M. Li and P. Vitányi [1995], and also O. Glier [2003].

The monograph is almost completely self-contained in the sense that no further sources are necessary for the proofs of the results. No prerequisite knowledge on formal languages and combinatorics of words is necessary. In very rare cases some additional statements are mentioned without proof. Several recent developments are discussed. In addition, a number of well-known classical results with new, alternative proofs are shown.

The authors are grateful to Kyoto Sangyo University, Nyíregyháza College, Debrecen University, the Japan Society for the Promotion of Science (JSPS), the Hungarian Academy of Sciences (HAS), and the Hungarian Foundation of Science and Technology (TéT Foundation) for their constant support during the development of this monograph.

Special thanks to Francine Blanchet-Sadri, Szilárd Zsolt Fazekas, László Kászonyi, Yoshiyuki Kunimochi, Peter Leupold, Gerhard Lischke, and Jeffrey Shallit for their useful comments concerning the manuscript. The authors are also very grateful to Andrea Pákozdy and Attila Gilányi for their careful linguistic revision. We are especially grateful to the staff at World Scientific and especially, Ms. Tan Rok Ting, for their encouragement and help in bringing about this monograph. In addition, the first author is grateful to his wife, Éva Tünde Rápolti, who always supported him in his scientific activity.

Pál Dömösi
Professor Emeritus
Nyíregyháza College, Hungary
and
Masami Ito
Professor Emeritus
Kyoto Sangyo University, Japan

March, 2014

Contents

Chapter 1

Preliminaries

1.1 Background

We start with a discussion of some set-theoretic notation. The set S consisting of all the elements that have the *property* P is written $S = \{s \mid s \text{ has the property } P\}$.[1] If s is an element of S, we write $s \in S$. The opposite case is expressed by $s \notin S$. If $s \in S$ implies that $s \in T$, then S is a *subset* of T and we write $S \subseteq T$. The *set of all subsets* of S is called the *power set* of S and it is denoted 2^S. The *set difference* $S \setminus T$ is $\{s \mid s \in S \text{ and } s \notin T\}$. Two sets S and T are *equal*, in symbols $S = T$, if $S \subseteq T$ and $T \subseteq S$. Moreover, S is a *proper subset* of T, denoted $S \subsetneq T$, if $S \subseteq T$ and $S \neq T$. The set containing no elements, the *empty* or *void* set, is denoted $\emptyset$. The *intersection* of S and T is the set consisting of all the elements in both S and T and we write $S \cap T = \{s \mid s \in S \text{ and } s \in T\}$. The *union* of S and T is the set consisting of the elements in either S or T. In symbols, $S \cup T = \{s \mid s \in S \text{ or } s \in T\}$. The set operations naturally extend to families of sets $\{S_i \mid i \in I\}$ where I is referred to as an *index set*:[2]

$$\bigcup_{i \in I} S_i = \{s \mid s \in S_i \text{ for some } i \in I\},$$

$$\bigcap_{i \in I} S_i = \{s \mid s \in S_i \text{ for all } i \in I\}.$$

If I is a finite (nonempty) set then we also say that $\bigcup_{i \in I} S_i$ is a *finite union* and $\bigcap_{i \in I} S_i$ is a *finite intersection* of sets $S_i, i \in I$, respectively. Two sets

[1] This way of specifying sets suffices for the purposes of this monograph and will not lead us into any foundational difficulties. To avoid ambiguity, sometimes we also use the form $S = \{s : s \text{ has the property } P\}$ instead of $S = \{s \mid s \text{ has the property } P\}$.

[2] An index set may be empty, but in this monograph we will consider only nonempty index sets.

are *disjoint* if $S \cap T = \emptyset$ and a family of sets $\{S_i \mid i \in I\}$ is *disjoint* if the sets are *pairwise disjoint*: $S_i \cap S_j \neq \emptyset$ implies $i = j$ for all $i, j \in I$. The cardinality of a set S is denoted by $|S|$. The set S is called *finite* if it has finitely many elements. Thus $|S|$ denotes the number of elements for a finite set S. In particular, if $|S| = 1$, then S is called a *singleton.*

Let S and T be sets. A *function* f of S into T, written $f : S \to T$, assigns to every element $s \in S$ an element $t \in T$, written $f(s) = t$.[3] Then t is the *image* of s, and s is an *inverse image* or *pre-image* or *counter image* of t under f. S is called the *source* and T is the *target* of f. If the source and the target coincide, then we also say that f is a *transformation* and is said to *transform* the elements of S. We put $f^{-1}(t) = \{s \mid f(s) = t, s \in S\}$ for every $t \in T$. We will also use the notation $f(S') = \{f(s) \mid s \in S'\}$ and $f^{-1}(T') = \bigcup_{t \in T'} f^{-1}(t)$ for any $S' \subseteq S, T' \subseteq T$. The function f is sometimes called a *map* or *mapping* from S to T. The set $f(S) = \{f(s) \mid s \in S\}$ is called the *image* of $f : S \to T$. The *rank* of f is the cardinality of its image. If $|f(S)| = 1$, then f is a *constant function,* or in short, a *constant.* If $f(S) = T$ then f is an *onto* or *surjective* function. If f is surjective, we may also write $f : S \twoheadrightarrow T$. The function f is *one-to-one* or *injective* if for every $s_1, s_2 \in S$, $s_1 \neq s_2$ implies that $f(s_1) \neq f(s_2)$. If f is injective, we sometimes write $f : S \hookrightarrow T$. If f is surjective and injective then it is called *bijective.* A bijective transformation is a *permutation* and is said to *permute* the elements of S. A *partial function,* or in other words, a *partially-defined function* from S to T is a function $f : S' \to T$, where S' is a subset of S.

Given a pair of sets S and Y, a *multi-valued function* f or a *multiple-valued function* of S into Y is a partially-defined function of S into 2^Y.[4]

Let $f : A \to B, g : C \to D$ be functions with $C \subseteq A$ and $g(c) = f(c)$ for each $c \in C$. Then we say that f is an *extension* of g (to A) and that g is a restriction of f (to C), and sometimes we write $g = f|_C$. The *(right) composite* or *(right) product* fg of functions $f : S \to T, g : T \to U$ is the function $h : S \to U$ with $h(s) = g(f(s))$ for all $s \in S$. For any transformation $f : S \to S$ and positive integer k we define the k-th power f^k of f as a transformation $f^k : S \to S$ having $f^k(s) = f(s), s \in S$ if $k = 1$ and $f^k(s) = f(f^{k-1}(s)), s \in S$ if $k > 1$.[5]

[3] Considering such an $f : S \to T$, sometimes we say that f is *well-defined.*

[4] In more precise terms, a multi-valued function may not be a function at all, at least not in the conventional sense.

[5] The definition $f^0(s) = s, s \in S$ is also allowed. In this monograph we consider f^k with $k > 0$.

Throughout this monograph, $\Im$ is the set of complex numbers, $\Re$ is the set of real numbers, $\mathbb{Q}$ is the set of rational numbers, $\mathbb{N}$ denotes the set of positive integers, $\mathbb{N}_0$ denotes the set of non-negative integers, and for integers $k, n\,(n \geq 2)$, $k \bmod n$ denotes the least positive integer k' such that n divides $k - k'$. (In particular, $0 \mod n = n$.) In addition, if n is a positive integer which divides $x - y$ for some pair of integers x, y, then we write $x \equiv y \pmod{n}$. Moreover, let us remark that for a real number x, $\lfloor x \rfloor$ and also $[x]$ denote the greatest integer which is smaller or equal to x (called *integer part* or *integer part* of x), and $\lceil x \rceil$ denotes the smallest integer which is greater than or equal to x. By a *strict divisor* of a positive integer n we mean any divisor $s > 1$ of n (including n itself). Finally, given a list $c_1, \ldots, c_n$ of integers, let $\mathrm{lcm}(c_1, \ldots, c_n)$ denote the least common multiple, and let $\gcd(c_1, \ldots, c_n)$ denote the greatest common divisor of $c_1, \ldots, c_n$.

The *Cartesian product* of a finite sequence of sets $S_1, \ldots, S_n$ is the set $S_1 \times \cdots \times S_n = \{(s_1, \ldots, s_n) \mid s_1 \in s_1, \ldots, s_n \in S_n\}$. If $S_1 = S_2 = \cdots = S_n$ then we call it *Cartesian power*. It is also defined for a not necessarily finite family $\{S_i \mid i \in I\}$ of sets as the set of all functions $\varphi : I \to \bigcup_{i \in I} S_i$ such that, for every $i \in I$, $\varphi(i)$ is in S_i. For this concept we use the notation $\prod_{i \in I} S_i$. (For a finite index set I, it is more convenient to think of the elements of a Cartesian product as a set of n-tuples as defined above.)

Let $\mathbb{N}_0(= \{0, 1, 2, \ldots\})$ be the set of non-negative integers, n be a positive integer, and $\mathbb{N}_0^n$ be the Cartesian power of $\mathbb{N}_0$ with n items. Let $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ with $x, y \in \mathbb{N}_0^n$. Define $x + y = (x_1 + y_1, \ldots, x_n + y_n)$ and for $m \geq 0$, define $mx = (mx_1, \ldots, mx_n)$.

A set either of the form $F = \emptyset$ or

$$F = \{p_0 + \sum_{i=1}^{r} k_i p_i \mid k_i \geq 0\},$$

where $p_0, \ldots, p_r$ are elements of $\mathbb{N}_0^n$, is said to be a *linear subset of* $\mathbb{N}_0^n$ or, in short, a *linear set*. p_0 is called the *constant* of F and $F_P = \{p_1, \ldots, p_r\}$ is the *set of periods* of F. A *semi-linear set* is a finite union of linear sets.

A subset H of $\mathbb{N}_0^n$ is said to be *stratified* if the following two conditions are satisfied:

(i) for every $(x_1, \ldots, x_n) \in H$, $|\{x_i \mid x_i > 0, i \in \{1, \ldots, n\}| \leq 2$;
(ii) there are no integers i, j, k, ℓ and $(x_1, \ldots, x_n), (x'_1, \ldots, x'_n) \in H$, such that $1 \leq i < j < k < \ell \leq n$ and $x_i x'_j x_k x'_\ell \neq 0$.

Thus, condition (ii) asserts that there are no two elements x and x' of H such that the indices of two non-zero components of x interlace with the indices of two non-zero components of x'.

The following two obvious facts are occasionally used in Section 6.2.

Fact 1.1.1. Given positive integers n, m and the set $\mathbb{N}_0$ of non-negative integers, let H be a stratified subset of $\mathbb{N}_0^n$. Then $H \times \{0\}^m$ and $\{0\}^m \times H$ are stratified subsets of $\mathbb{N}_0^{n+m}$. □

Fact 1.1.2. Given positive integers n, m and the set $\mathbb{N}_0$ of non-negative integers, let $H \subseteq \mathbb{N}_0^n$ be stratified. Moreover, let $1 \leq i_1 < \cdots < i_m \leq n$. If $f : \mathbb{N}_0^n \to \mathbb{N}_0^m$ is a function defined by $f((z_1, \ldots, z_n)) = (z_{i_1}, \ldots, z_{i_m})$, then $f(H)$ is stratified. □

A *relation* between a set S and a set T is a subset ρ of $S \times T$. For $(s,t) \in \rho$ we write $s \; \rho \; t$. Thus $\rho = \{(s,t) \mid s \; \rho \; t\}$.

A relation ρ between S and S itself is simply called a relation on S. It is called *reflexive* if, for all $s \in S$, $s \; \rho \; s$; *symmetric* if, for every $s, t \in S$, $s \; \rho \; t$ implies $t \; \rho \; s$; *antisymmetric* if, for every $s, t \in S$, $s \; \rho \; t$ and $t \; \rho \; s$ imply $s = t$; and *transitive* if $s \; \rho \; t$ and $t \; \rho \; u$ imply $s \; \rho \; u$ for every $s, t, u \in S$. A relation ρ on S is an *equivalence relation* on S if ρ is reflexive, symmetric, and transitive. If ρ is an equivalence relation on S, then for every $s \in S$, the set $s/\rho = \{t \mid s \; \rho \; t\}$ is the *equivalence class* of s under ρ. This notation is extended to an arbitrary subset S' of S by $S'/\rho = \{s'/\rho \mid s' \in S'\}$. A *partition* π on S is a collection of disjoint subsets of S whose set union is S. Then, in symbols, $\pi = \{S_i \mid i \in I\}$ such that $S_i \cap S_j = \emptyset$ for $i \neq j, i, j \in I$, and $\bigcup_{i \in I} S_i = S$. Sometimes we refer to the elements of π as *blocks*. For every $s \in S$, $\pi(s)$ will denote the block containing the element s. It is clear that if ρ is an equivalence relation on S then S/ρ is a partition of S and that every partition of S can be given this way.

Given a set S and a positive integer n, the mapping $S^n \to S$ is called an *n-ary operation* on S. We also define the concept of 0-ary operation on S as a fixed constant c of S, which is also written in the form $c : S^0 \to S$ sometimes. An n-ary operation on S, with $n \geq 0$, is also simply called an *operation* on S.

Given a nonempty set S, let A be a set of operations on A. We say that the equivalence relation ϱ_A on S is a *congruence relation with respect to A* if, for every $f : S^n \to S, a_1, a_1', \ldots, a_n, a_n' \in A$, with $f \in A$ and $a_i \varrho_A a_i'$, $i = 1, \ldots, n$, $f(a_1, \ldots, a_n) \varrho_A f(a_1', \ldots, a_n')$. We note that 0-ary operations have no arguments and thus, considering the set $T \subseteq A$ of 0-ary operations on S, the congruence relations on A and $A \setminus T$ coincide. In addition, if A consists of 0-ary operations, then every equivalence relation on S is a