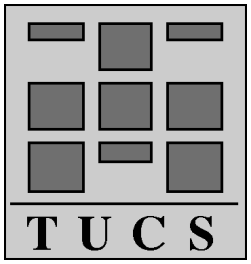
Commutation Problems on Sets of Words and Formal Power Series

Ion Petre



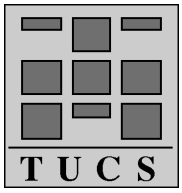
Turku Centre for Computer Science

TUCS Dissertations

No 38, June 2002

Commutation Problems on Sets of Words and Formal Power Series

Ion Petre
Turku Centre for Computer Science and
Department of Mathematics, University of Turku
Turku 20520, Finland
ion.petre@cs.utu.fi



Turku Centre for Computer Science
TUCS Dissertations No 38
June 2002
ISBN 951-29-2270-3
ISSN 1239-1905

Commutation Problems on Sets of Words and Formal Power Series

Ion Petre

To be presented, with the permission of the Faculty of Mathematics and Natural Sciences of the University of Turku, for public criticism in the Auditorium of the Computer Science Department on May 27th, 2002, at 12 noon.

University of Turku
Department of Mathematics
2002

Commutation Problems on Sets of Words and Formal Power Series

Ion Petre

Supervisor

Professor Juhani Karhumäki

Department of Mathematics
University of Turku
FIN-20014 Turku, Finland

Opponent

Professor Werner Kuich

Institut für Algebra und Diskrete Mathematik
Technische Universität Wien
Wiedner Hauptstrasse 8-10
A-1040 Wien, Austria

Reviewers

Professor Michel Latteux

Laboratoire d'Informatique Fondamentalle de Lille
Université des Sciences et Technologies de Lille
59655 Villeneuve d'Ascq, Lille, France

Professor Sheng Yu

Department of Computer Science
University of Western Ontario
Middlesex College, London, Ontario
N6A 5B7 Canada

ISBN 951-29-2270-3
ISSN 1239-1883
Painosalama Oy

To my family

Abstract

We study in this thesis several problems related to commutation on sets of words and on formal power series. We investigate the notion of semilinearity for formal power series in commuting variables, introducing two families of series - the semilinear and the bounded series - both natural generalizations of the semilinear languages, and we study their behaviour under rational operations, morphisms, Hadamard product, and difference. Turning to commutation on sets of words, we then study the notions of centralizer of a language - the largest set commuting with a language -, of root and of primitive root of a set of words. We answer a question raised by Conway more than thirty years ago - asking whether or not the centralizer of any rational language is rational - in the case of periodic, binary, and ternary sets of words, as well as for rational ω-codes, the most general results on this problem. We also prove that any code has a unique primitive root and that two codes commute if and only if they have the same primitive root, thus solving two conjectures of Ratoandromanana, 1989. Moreover, we prove that the commutation with an ω-code X can be characterized similarly as in free monoids: a language commutes with X if and only if it is a union of powers of the primitive root of X.

Keywords: commutation, sets of words, formal power series, semilinearity, closure properties, Combinatorics on Words, centralizer, Conway's problem, BTC-property, root, primitive root, motif, code, ternary language.

Acknowledgments

My PhD studies in Turku have been supervised by Professor Juhani Karhumäki. I wish to express my gratitude for his guidance, advice, and support throughout these years. I want to thank him for unveiling to me the beauty of academic research and for setting the highest levels of quality for my work. I also wish to thank him for encouraging me to explore some other fields of research, independently of the scope of this thesis.

It is my pleasure to thank Professor Werner Kuich (Technical University of Wien), Professor Michel Latteux (University of Lille), and Professor Sheng Yu (University of Western Ontario) for reviewing my thesis and for their useful comments and suggestions. It is an honor for me that Professor Werner Kuich accepted to act as an opponent for the public debate of my thesis.

I have been very fortunate to work in this period with Professor Grzegorz Rozenberg (Leiden University). I would like to thank him for introducing me to the beauty of Biocomputing and for his constant support. His influence on me goes far beyond our work and I am very grateful for this "magical touch". I have also greatly benefited from working close to Professor Arto Salomaa, learning from him and having his valuable advice on many issues.

I would like to thank my good friend Dr. Lucian Ilie (University of Western Ontario) for his support, especially in the early stages of my stay in Turku, as well as for the countless rewarding discussions we had on most various matters. I also wish to thank Professor Tero Harju (University of Turku) for our joint work and for the many things I have learned from him. Special thanks are due to Professor George Păun (Romanian Academy) and Professor Alexandru Mateescu (University of Bucharest) for our valuable discussions, and to Professor Adrian Atanasiu (University of Bucharest) for his advice during my first academic steps.

I have benefited from the excellent working conditions and the financial support provided by Turku Centre for Computer Science, a top research centre in Finland. I thank all the TUCS staff for this and especially its chairman, Professor Ralph-Johan Back.

Finally, I wish to thank my family for always believing in me, for their warm encouragement, and ever-lasting support. Special thanks to my wife, Luigia, for her love and devotion. This thesis is dedicated to them.

Turku
May 2002 Ion Petre

Contents

Chapter 1

Introduction

The commutation is an essential concept in mathematics, playing a decisive role in a huge number of principles and results we constantly deal with. Even in noncommutative mathematical structures, one encounters interesting - and often challenging - problems concerning the commutation of some elements of that structure, or some commutative substructures.

In Automata Theory, one deals in general with non-commutative structures, such as the free monoid of all finite words over a given alphabet. The commutation, however, gives rise to some challenging problems, often connected to concepts and results from algebra, geometry, or analysis. We investigate in this thesis several such problems, mainly related to semilinearity and commutation of sets of words.

The semilinearity is an important notion in mathematics, especially in algebra and geometry. It is also an important notion in theoretical computer science, mostly because it gives a simple characterization for the rational subsets of any commutative monoid - this result is often referred to as Eilenberg-Schützenberger theorem, [22]. Perhaps even better known is this notion through the celebrated theorem of Parikh, [56], proving that any context-free language is letter-equivalent with a rational language. Equivalently, Parikh's result proves that in a free commutative monoid, the context-free languages coincide with the rational languages and also, with the semilinear ones. Our investigations on semilinearity originated from the question of whether or not Parikh's theorem holds for multisets, i.e., if it holds when the multiplicities are taken into consideration. We gave a negative answer to this question: there are context-free grammars for which no automaton accepts the same language with multiplicities. It turned out, however, from this study that the semilinearity for formal power series (in commuting variables) has been considered in literature only in passing, and that this notion was not even properly defined, nor was it clear how to define it in a natural way. This was the starting point for a systematic study of the semilinearity for formal power series. We extended this notion in

two different ways, introducing the semilinear and the bounded power series in commuting variables, comparing their behaviour under elementary operations - rational operations, Hadamard product, morphisms, and difference - to the behaviour of other known families of series, both in commuting and noncommuting variables. As it turned out, some of the properties of the semilinear languages can be extended to series, and some cannot, or at least, they do not hold for both families of series; this gives a new insight on the notion of semilinearity. The results of this investigation are presented in Chapter 3, following a Chapter 2 of preliminaries.

Turning our attention to the commutation itself, i.e., to the equation $XY = YX$, a number of old open problems related to words and sets of words came up. It is an elementary result of combinatorics on words that two words commute if and only if they are powers of a same word. Moreover, the same result holds also for multisets of words: two formal power series with coefficients in a commutative field commute if and only if they are combinations of another series - Cohn's theorem, [16]. On the other hand, almost nothing was known on the commutation of sets of words.

The centralizer $\mathcal{C}(X)$ of a set of words X is the largest set of words commuting with it, i.e., $X\mathcal{C}(X) = \mathcal{C}(X)X$. The following question has been proposed by Conway, [18], more than thirty years ago: is it true that the centralizer of any rational language is rational ? The answer to this question is still unknown up-to-date; as a matter of fact, much weaker question than Conway's problem are also open, such as: is it true that the centralizer of any finite language is a recursively enumerable language ? We develop in Chapter 4 two strategies to attack Conway's problem: the equational method and the branching point approach. We give Conway's question an affirmative answer in the case of periodic, binary, and ternary sets of words, as well as for finite biprefix codes, results presented in Chapter 4. We return to this problem in Chapter 6, where we prove that the centralizer of any rational ω-code is rational, the most general result known on Conway's problem.

Closely related to commutation are the notions of root and primitive root. In general, X is said to be a root of Y if $Y = X^n$, for some $n \geq 1$; X is primitive if it has no roots other than itself. For words it is well-known, see [11], or [47], that any nonempty word has a unique primitive root; for sets of words on the other hand, such a property does not hold in general, despite the fact that the notion of root has been extended in two different ways to sets of words, to the notions of root and premotif, see [73] and [3], respectively. For codes, however, we prove that these two notions coincide. Moreover, we prove that any code has a unique primitive root and that two codes commute if and only if they have the same primitive root, thus solving two conjectures stated in 1989 by Ratoandromanana, [66]. To prove these results, we develop a new strategy, the so-called multiplicity approach, transferring the problems to formal power series, solving them in

this framework using algebraic tools, and then transferring the results back to sets of words. We present this strategy and its outcome in Chapter 5.

The multiplicity approach and the connection to formal power series is further refined in Chapter 6 to study the commutation of two sets of words. While it is clear that characterizing the commutation of two languages similarly as for words, polynomials, and formal power series is not possible in general, it has been conjectured by Ratoandromanana [66], 1989, that the situation is different for codes: for any code X, if a language Y commutes with X, then $Y = \cup_{i \in I} \rho(X)^i$, where $\rho(X)$ is the primitive root of X and I is a set of nonnegative integers. Using the multiplicity approach and Cohn's theorem, we give a solution to this conjecture in the case of ω-codes. In particular, this leads to solutions also in the case of prefix, elementary, binary, and ternary codes. As a matter of fact, we also give independent proofs for this conjecture in the case of binary and ternary sets of words, based on the branching point approach and on combinatorial properties of finite and infinite words.

At the end of each of these chapters, we discuss directions for further research and give some open problems. The last chapter of the thesis is dedicated to conclusions.

The material of this thesis comes mainly from our papers [59], [60], [61], and [62], and our joint papers [29], [37], [38], [39], and [40]. However, some new results are introduced and some of the proofs are improved.

Chapter 2

Preliminaries

This preliminary chapter briefly surveys the main notions and concepts used throughout the thesis for the purpose of serving as a quick reference whenever needed. More specific definitions and results are given in the appropriate places in the thesis, where they are employed. We recall in the following some elements of algebra, combinatorics of words, formal languages, codes, and formal power series.

2.1 Elements of algebra

A *semigroup* consists of a set S and a binary associative operation $\circ$ on S. We denote this semigroup by $(S, \circ)$, or just S when $\circ$ is clearly understood.

A *monoid* consists of a set M, a binary associative operation $\circ$ on M, and a *neutral* element $\lambda \in M$ such that $a \circ \lambda = \lambda \circ a = a$, for all $a \in M$. We denote this monoid by $(M, \circ, \lambda)$, or just M, whenever $\circ$ and λ are understood. The monoid $(M, \circ, \lambda)$ is *commutative* if the operation $\circ$ is commutative, i.e., $a \circ b = b \circ a$, for all $a, b \in M$. E.g., denoting by $\mathbb{N}$ the set of all nonnegative integers, $(\mathbb{N}, \circ, 0)$ is a commutative monoid. A subset N of M is a *submonoid* of M if it is closed under the operation $\circ$ and $\lambda \in N$. Thus, $(N, \circ, \lambda)$ is a monoid. A *morphism* h of a monoid $(M, \circ, \lambda)$ into a monoid $(M', \circ', \lambda')$ is a mapping $h : M \to M'$ such that $h(\lambda) = \lambda'$, and $h(a \circ b) = h(a) \circ' h(b)$, for all $a, b \in M$. The *direct product* of the monoids M and M' is the monoid $M \otimes M' = (M \times M', \circ'', (\lambda, \lambda'))$, where $M \times M'$ denotes the *cartesian product* of the sets M and M', and $(a, b) \circ'' (c, d) = (a \circ c, b \circ' d)$.

A *semiring* consists of a set R, two binary operations $+$ and $\cdot$ on M, and two constants $0, 1 \in R$ such that

- (i) $(R, +, 0)$ is a commutative monoid,
- (ii) $(R, \cdot, 1)$ is a monoid,
- (iii) $(a + b) \cdot c = a \cdot c + b \cdot c,\ c \cdot (a + b) = c \cdot a + c \cdot b$, for all $a, b, c \in R$ *(distributivity)*,
- (iv) $0 \cdot a = a \cdot 0 = 0$, for all $a \in R$.

We denote this semiring as $(R, +, \cdot, 0, 1)$, or simply R, whenever the operations and the constants are understood. The semiring R is called *commutative* if $a \cdot b = b \cdot a$, for all $a, b \in R$. Note that by definition, the $+$ operation is commutative for all semirings. R is called *idempotent* if $1 + 1 = 1$. Thus, in such a semiring R, $a + a = a$, for all $a \in R$. E.g., $(\mathbb{N}, +, \cdot, 0, 1)$ is a commutative semiring. If $B = \{0, 1\}$, then $\mathbb{B} = (B, +, \cdot, 0, 1)$ is the *Boolean semiring*, where $1 + 1 = 1$. Thus, the Boolean semiring is idempotent.

A subset S of R is a *subsemiring* of R if it is closed under sum and product, and $0, 1 \in S$. Thus, $(S, +, \cdot, 0, 1)$ is a semiring.

A *morphism* h of a semiring $(R, +, \cdot, 0, 1)$ into another semiring $(S, \oplus, \odot, 0', 1')$ is a mapping $h : R \to S$ such that $h(0) = 0'$, $h(1) = 1'$, $h(a + b) = h(a) \oplus h(b)$, and $h(a \cdot b) = h(a) \odot h(b)$, for all $a, b \in M$.

A semiring R is *naturally ordered* if the relation $\leq$ defined as

$$a \leq b \text{ iff there is } c \text{ such that } a + c = b,$$

is a partial order on the set R. The relation $\leq$ is called the *natural order* of R.

A commutative semiring $(R, +, \cdot, 0, 1)$ is called *complete* if for any family $(a_i)_{i \in I}$ of elements of R, it is possible to define sums $\sum_{i \in I} a_i$ such that the following conditions are satisfied:

(i) $\sum_{i \in \emptyset} a_i = 0$, $\sum_{i \in \{n\}} a_i = a_n$, $\sum_{i \in \{m,n\}} a_i = a_m + a_n$, for $m \neq n$,

(ii) $\sum_{j \in J} \left(\sum_{i \in I_j} a_i \right) = \sum_{i \in I} a_i$, if $I = \cup_{j \in J} I_j$ and $I_j \cap I_{j'} = \emptyset$, for all $j \neq j'$,

(iii) $\sum_{i \in I} (c \cdot a_i) = c \cdot \left(\sum_{i \in I} a_i \right)$, $\sum_{i \in I} (a_i \cdot c) = \left(\sum_{i \in I} a_i \right) \cdot c$.

Intuitively, the semiring R is complete if it is possible to define "infinite" sums as extensions of the finite ones - condition (i), such that they are commutative, associative - condition (ii), and distributive - condition (iii). The semiring R is called *ω-continuous* if the following conditions are satisfied:

(i) R is complete,

(ii) R is naturally ordered, and

(iii) for any sequence $(a_i)_{i \in \mathbb{N}}$ of elements of R and any $c \in R$, if $\sum_{0 \leq i \leq n} a_i \leq c$, for all $n \in \mathbb{N}$, then $\sum_{i \in \mathbb{N}} a_i \leq c$.

For more details and related notions and results, we refer to [9], [10], [15], [21], [33], [43], and [49].

2.2 Words and languages

We introduce in this section some basic notions and results on combinatorics on words, formal languages, and codes.

2.2.1 Words

An *alphabet* is a finite nonempty set, its elements being called *letters*. A (finite) *word* over an alphabet Σ is a finite sequence of letters of Σ. In particular, the *empty word* consists of zero letters and it is denoted by 1. For instance, 1, a, ab, $babb$ are all words over the alphabet $\{a, b\}$.

Let Σ be an alphabet. The set of all words over Σ is denoted by Σ^*, while the set of all nonempty words over Σ is $\Sigma^+ = \Sigma^* \setminus \{1\}$. The *concatenation* (or *catenation*, or *product*) of two words u and v over Σ is the word denoted by uv obtained by writing v at the end of u. For instance, the concatenation of a and bb is the word abb. The sets Σ^* and Σ^+ are the *free monoid* and the *free semigroup*, respectively, generated by Σ with respect to the catenation product (see Section 2.2.3 for a general definition of a free monoid).

For a word $u \in \Sigma^*$, its *n-th power* is denoted by u^n and it is defined as follows:

$$u^0 = 1, \quad u^n = u^{n-1}u, \text{ for all } n \geq 1.$$

The *length* of a word $u \in \Sigma^*$ is the number of letters of which u consists. Thus, if $u = a_1 a_2 \dots a_n$, with $a_i \in \Sigma$, $1 \leq i \leq n$, then its length $|u|$ is n. In particular, the length of the empty word is zero. Thus, the length application $|\cdot|$ is a monoid morphism $|\cdot| : (\Sigma^*, \cdot) \to (\mathbb{N}, +)$. For a letter $a \in \Sigma$, the number of occurrences of a in u is denoted by $|u|_a$. Consequently, $|u| = \sum_{a \in \Sigma} |u|_a$. For a word $u = a_1 a_2 \dots a_n$, $a_i \in \Sigma$, $1 \leq i \leq n$, its *reverse* is the word $\overline{u} = a_n \dots a_2 a_1$.

There exist several important partial order relations on words, defined in the following. For two words u, v, we say that u is a *prefix* of v and denote $u \leq_p v$, or simply $u \leq v$, if $v = uw$, for some word w. The set of prefixes of v is denoted by $\mathrm{Pref}(v)$. We say that u is a *suffix* of v and denote $u \leq_s v$, if $v = wu$, for a word w. The set of suffixes of v is denoted by $\mathrm{Suf}(v)$. The word u is said to be a *factor* of v, denoted by $u \leq_f v$, if $v = w_1 u w_2$, for some words w_1, w_2. Also, u is said to be a *subword* (or a *scattered subword*) of v if $u = u_1 u_2 \dots u_n$ and $v = w_0 u_1 w_1 u_2 \dots, u_n w_n$, for some $n \geq 1$ and some words $w_0, w_1, \dots, w_n$, $u_1, u_2, \dots u_n$. The word u is a *proper* prefix (suffix, factor, subword) of v if u is a prefix (suffix, factor, subword, resp.) and moreover, $u \neq 1$ and $u \neq v$. For some other interesting relations on words we refer to [31].

A word v is a *root* of a word u if $u = v^n$, for some $n \geq 1$. Moreover, v is a *nontrivial root* of u if $v \neq 1$ and $v \neq u$. A nonempty word u is *primitive* if it is its only root; in other words, if $u = v^n$, then necessarily, $n = 1$ and $v = u$. For instance, aba is a primitive word, while $abab$ is not, having ab as a nontrivial root.

The following result characterizes the commutation of two words in terms of roots.

Theorem 2.2.1. *Let $u, v \in \Sigma^*$. The following conditions are equivalent: (i) u and v commute; (ii) u and v satisfy a nontrivial relation; (iii) u and v have a common root.*

As a consequence of this result, it follows that any nonempty word has a unique primitive root.

Corollary 2.2.2. *For each $u \in \Sigma^+$, there exists a unique primitive root $\rho(u)$ of u.*

For a word u, $\rho(u)$ is called the *primitive root* of u. In particular, it follows that two words u and v commute if and only if $\rho(u) = \rho(v)$.

The following simple property of primitive words is often useful in the sequel.

Lemma 2.2.3. *Let ρ be a primitive root. Then ρ has exactly two occurrences in ρ^2, as a prefix and a suffix, i.e., if u and v are words such that $\rho^2 = u\rho v$, then necessarily, $u = 1$ or $v = 1$.*

Two words u and v are called *conjugates* if there are some words x and y such that $u = xy$ and $v = yx$. Thus, v can be obtained from u through some cyclic shifts. The following result characterizes the conjugacy of words, both in terms of word equations and in terms of solutions of word equations.

Theorem 2.2.4. *Let $u, v \in \Sigma^+$. The following conditions are equivalent: (i) u and v are conjugates; (ii) there exists a word z such that $uz = zv$; (iii) there exist words z, p, and q such that $u = pq$, $v = qp$, and $z \in (pq)^*p$; (iv) the primitive roots of u and v are conjugates.*

A *right-infinite* word, or simply, an *infinite* word u over the alphabet Σ is an infinite - to the right - sequence of letters from Σ:

$$u = a_0 a_1 a_2 \ldots,$$

with $a_n \in \Sigma$, for all $n \geq 0$. We denote the set of all right-infinite words over Σ by Σ^ω. A *left-infinite* word v over Σ is an infinite - to the left - sequence of letters from Σ:

$$v = \ldots a_{-3} a_{-2} a_{-1},$$

with $a_{-n} \in \Sigma$, for all $n \geq 1$. We denote the set of all left-infinite words over Σ by ${}^\omega\Sigma$.

For a finite nonempty word $u \in \Sigma^+$, we denote $u^\omega = uuu\ldots \in \Sigma^\omega$ and ${}^\omega u = \ldots uuu \in {}^\omega\Sigma$.

For other notions and results on Combinatorics on Words we refer to [11], [47], and [48].

2.2.2 Languages

A *language* over the alphabet Σ is a set of words over Σ, i.e., a subset of Σ^*. The empty language is denoted by $\emptyset$. The operation of union, intersection, and difference are defined in the usual way for sets. We denote the union of two languages L and R by $L + R$, preserving for the other two operations the usual notations for sets: $L \cap R$ for intersection and $L \setminus R$ for difference.

The *product* of two languages L and R is the language $LR = \{uv \mid u \in L, v \in R\}$. This product is called *unambiguous* if for any $w \in LR$, there is exactly one pair $(u, v) \in L \times R$ such that $w = uv$. The *n-th power* of a language L is then defined as follows: $L^0 = \{1\}$, $L^n = L^{n-1}L$, for all $n \geq 1$. The *catenation closure*, or the *Kleene star* of a language L is the language $L^* = \cup_{n \geq 0} L^n$.

A *binary (ternary)* set is a set of cardinal two (three, resp.) A language L is *periodic* if there is a finite word u such that $L \subseteq u^*$. For any finite language L, $|L|$ denotes its cardinality. Also, $\overline{L}$ denotes its *reverse*, i.e., $\overline{L} = \{\overline{u} \mid u \in L\}$.

The operations of sum, product, and star are called *rational operations*. The smallest family of languages containing all finite languages and being closed under rational operations is called the family of *rational* languages. Equivalently, a language is *rational* if it can be obtained from some finite languages by applying a finite number of rational operations. E.g., if $\Sigma = \{a, b\}$, then Σ^* is a rational language, as well as $\Sigma^* \setminus \{a\}$. Indeed, $\Sigma^* \setminus \{a\} = \{1, b\} + \{a, b\}^2\{a, b\}^*$.

A *deterministic finite automaton* - DFA, in short - is a 5-tuple $\mathcal{A} = (Q, \Sigma, \delta, q_0, F)$, where Σ is an alphabet, Q is a finite set of states, $q_0 \in Q$ is the *initial* state, $F \subseteq Q$ is the set of *final* states, and $\delta : Q \times \Sigma \to Q$ is the *transition* mapping. The mapping δ is extended to $\overline{\delta} : Q \times \Sigma^* \to Q$ as follows:

$$\overline{\delta}(q, 1) = q \quad \text{and} \quad \overline{\delta}(q, wa) = \delta(\overline{\delta}(q, w), a),$$

for all $q \in Q$, $w \in \Sigma^*$, $a \in \Sigma$. The language $\mathcal{L}(\mathcal{A})$ *accepted* by the automaton $\mathcal{A}$ is $\mathcal{L}(\mathcal{A}) = \{u \mid \overline{\delta}(q_0, u) \in F\}$. A language is called *regular* if it is accepted by some DFA.

The following result, stating the equivalence between the rationality and the regularity of a language is due to Kleene, [41].

Theorem 2.2.5 (Kleene's Theorem, [41]). *A language is regular if and only if it is rational.*

A *context-free grammar* is a 4-tuple $G = (N, \Sigma, S, P)$ where N and Σ are two disjoint alphabets, of *nonterminals* and of *terminals*, respectively, $S \in N$ is the *start symbol* and $P \subseteq N \times (N \cup \Sigma)^*$ is a finite set. The elements (A, α) of P are called *productions* and are customarily written as rewriting rules in the form $A \to \alpha$. For any $\beta, \gamma \in (N \cup \Sigma)^*$, we say that $\beta A \gamma$ *derives directly* $\beta\alpha\gamma$ in G, and we write $\beta A\gamma \Rightarrow_G \beta\alpha\gamma$, if $A \to \alpha$ is a production

of G. The reflexive and transitive closure of the relation $\Rightarrow_G$ is denoted by $\Rightarrow_G^*$ and it is called *the derivation relation*. The language *generated* by the grammar G is $\mathcal{L}(G) = \{u \in \Sigma^* \mid S \Rightarrow_G^* u\}$. A language is called *context-free*, or *algebraic*, if it is generated by a context-free grammar.

For details on rational and context-free languages, we refer to [2], [24], [51], [69], [70], and [75].

A set $L \subseteq \mathbb{N}^k$, $k \geq 1$, is called *linear* if there are some $u_0, u_1, \ldots, u_i \in \mathbb{N}^k$ such that

$$L = \{u_0 + m_1 u_1 + \ldots + m_i u_i \mid m_1, \ldots, m_i \in \mathbb{N}\}.$$

Equivalently, L is linear if $L = u_0 + F^*$, where $F \subseteq \mathbb{N}^k$ is a finite set and the iteration F^* is done with respect to addition, i.e., in the additive monoid $(\mathbb{N}^k, +)$. A *semilinear* subset of $\mathbb{N}^k$ is a finite union of linear sets.

For an alphabet $\Sigma = \{a_1, \ldots, a_n\}$, the *Parikh mapping* is $\pi : \Sigma^* \to \mathbb{N}^k$, $\pi(u) = (|u|_{a_1}, \ldots, |u|_{a_k})$. Two languages over Σ are called *letter-equivalent* if their images through the Parikh mapping coincide.

The following celebrated result is due to Parikh, [56].

Theorem 2.2.6 (Parikh's Theorem, [56]). *Each context-free language is letter-equivalent with a regular language.*

Corollary 2.2.7. *For any context-free language L, $\pi(L)$ is a semilinear language. Moreover, for any semilinear set S, there is a context-free language L such that $S = \pi(L)$.*

For a definition of Turing machines we refer to [74], as well as to [52], [54], [68], and [70]. A language $L \subseteq \Sigma^*$ is *recursively-enumerable* if there exists a Turing machine $\mathcal{M}$ accepting L. The language L is *recursive* if there exists a Turing machine $\mathcal{M}$ accepting L and halting on all input words.

If $\mathcal{M}$ is an always halting Turing machine, then we implicitly have an algorithm to decide whether or not an input w is in $\mathcal{L}(\mathcal{M})$. We say in this case that the membership problem for $\mathcal{L}(\mathcal{M})$ is *decidable*. On the other hand, for an arbitrary Turing machine, we only have a procedure to confirm that $w \in \mathcal{L}(\mathcal{M})$ if this is the case, while giving no information if $w \notin \mathcal{L}(\mathcal{M})$. In general, we say that the membership to a language L is decidable if L is recursive.

A function is called *effectively computable* if there exists an algorithm that produces the values of the function correctly for each possible argument.

When considering a problem with a "yes/no" answer, we can identify the problem with the set of its "yes" instances. For instance, the set of "yes" instances of the problem "Is u primitive ?" is the set of primitive words. Thus, instead of deciding the problem itself, we may decide the membership to the set of "yes" instances. Accordingly, we say that a problem is *decidable* if the set of its "yes" instances is a recursive language. We refer to [68] for a detailed discussion on decidability aspects.

2.2.3 Codes

Let $(M, \cdot)$ be a monoid and $X \subseteq M$. We say that M is *generated* by X if $M = X^*$, i.e., for any $\alpha \in M$, there exist $x_1, \ldots, x_n \in X$, $n \geq 0$, such that $\alpha = x_1 \ldots x_n$. In this case, we also say that X is a set of *generators* of M.

We say that $X \subseteq M$ is *independent* in M if whenever $x_1 x_2 \ldots x_m = y_1 y_2 \ldots y_n$, with $x_i, y_j \in X$, $1 \leq i \leq m$, $1 \leq j \leq n$, then $m = n$ and $x_i = y_i$, for all $1 \leq i \leq m$.

A monoid M is *free* if it has an independent set of generators X. We say in this case that M is *freely generated* by X and that X is a *base* of M.

Let Σ be an alphabet. A submonoid M of Σ^* is *stable* if and only if for any word $w \in \Sigma^*$, one has $w \in M$ whenever there exists $x, y \in M$ such that $xw, wy \in M$.

The following result is often referred to as *Schützenberger criterium.*

Theorem 2.2.8. *Let M be a submonoid of Σ^*. Then M is stable if and only if M is free.*

Let $X \subseteq \Sigma^*$ and $u \in \Sigma^*$. An *X-factorization* of u is a finite sequence $x_1, \ldots, x_m$, $m \geq 0$ of words of X such that $u = x_1 \ldots x_m$. An X-factorization of an infinite (left-infinite) word v is a sequence $(x_n)_{n \geq 1}$ such that $v = x_1 x_2 \ldots$ ($v = \ldots x_2 x_1$, resp.).

A set $X \subseteq \Sigma^+$ is a *code* if any word $u \in \Sigma^*$ has at most one X-factorization. This condition can in fact be expressed in several different ways, as shown in the next result.

Theorem 2.2.9. *Let Σ be an alphabet and $X \subseteq \Sigma^+$. The following conditions are equivalent:*

(i) X is a code;

(ii) for any words $x_1, \ldots, x_m, y_1, \ldots, y_n \in X$, $m, n \geq 1$, if $x_1 \ldots x_m = y_1 \ldots y_n$, then $m = n$ and $x_i = y_i$, for all $1 \leq i \leq m$;

(iii) X^* is freely generated by X;

(iv) X^* is stable in Σ^*

The name of "code" is motivated by the next result. Intuitively, if the letters of a source alphabet Δ are put into a 1-to-1 correspondence with the words of a code X over a target alphabet Σ, then a message $\alpha \in \Delta^*$ is encoded into $\beta \in \Sigma^*$ by replacing each letter of α with its correspondent in X. The fact that X is a code ensures the unique decipherability of any such message.

Lemma 2.2.10. *Let Σ and Δ be two alphabets. A set $X \subseteq \Sigma^*$ is a code if and only if any morphism $\phi : \Delta^* \to \Sigma^*$ induced by a bijection from Δ onto X is injective.*

The simplest families of codes are those of prefix, suffix, and biprefix codes. A language $X \subseteq \Sigma^+$ is a *prefix* (*suffix*) code if no word of X is a proper prefix (suffix, resp.) of another word in X. X is a *biprefix* if no word of X is a proper prefix or a proper suffix of another word in X, i.e., X is both a prefix and a suffix code.

A language X over Σ is an *ω-code (reversed ω-code)* if any infinite (left-infinite, resp.) word over Σ has at most one X-factorization. In particular, X is a reversed ω-code if and only if $\overline{X}$ is an ω-code. As it is easy to see ([7]), any ω-code and any reversed ω-code is a code, and any prefix, suffix, and biprefix is an ω-code.

The following two results characterize a binary and a ternary code, respectively, in terms of ω-codes. Note that neither of the results holds for codes with at least four words.

Lemma 2.2.11. *A binary set is a code if and only if it is an ω-code.*

Lemma 2.2.12 ([34],[35]). *A ternary set is a code if and only if it is an ω-code or it is a reversed ω-code.*

A *bounded-decoding delay code* $L \subseteq \Sigma^+$ is a language for which there is an integer $d \geq 0$ such that for any $x, x', y_1, \ldots, y_d \in L$ and $z \in L^*$, if $xy_1 \ldots y_d$ is a prefix of $x'z$, then $x = x'$. The smallest integer d with this property is called the *decoding delay of the code L*.

A finite set of words X is called *simplifiable* if there is a finite language Y such that $X \subseteq Y^*$ and $|Y| < |X|$. Otherwise, X is called an *elementary* set. As it is well-known, see [11], [47], [48], any elementary set is an ω-code.

The next result summarizes the relations among the above defined families of codes.

Theorem 2.2.13. *(i) Any ω-code is a code.*

(ii) Any code with bounded decoding delay is an ω-code. Conversely, any finite ω-code has bounded decoding delay.

(iii) Any prefix and biprefix code is a code with decoding delay zero. Any suffix code is an ω-code.

(iv) Any elementary code is an ω-code.

We refer to [7], [47], [48], [73] for the above results, as well as for deeper notions and results on the theory of codes.

2.3 Formal power series

Let (M, λ) be a monoid and $(R, +, \cdot, 0, 1)$ a semiring. Mappings r of M into R are called *formal power series over M* with *coefficients* in R. The

values of r are denoted (r, w), for all $w \in M$; r is written as a *formal sum*:

$$r = \sum_{w \in M} (r, w)w.$$

The values (r, w) are called the *coefficients* of the series r. The set of all formal power series defined above is denoted by $R\langle\langle M \rangle\rangle$.

For an alphabet $\Sigma = \{a_1, \ldots, a_n\}$, we denote by $\Sigma^{\oplus}$ the direct product $a_1^* \otimes \ldots \otimes a_n^*$.

Most often, we will be interested in series over Σ^* with coefficients in a semiring R. Then we say that we deal with series in *noncommuting* variables in Σ. The series over $\Sigma^{\oplus}$ are called formal power series with *commuting* variables in Σ. As a matter of fact, the family $R\langle\langle \Sigma^{\oplus} \rangle\rangle$ is often denoted in Analysis and Combinatorics as $R[[\Sigma]]$. However, we adopt in this thesis the notation $R\langle\langle \Sigma^{\oplus} \rangle\rangle$ to underline the fact that we are not interested in summing up the series, or in convergence aspects, but rather in various operations defined for series. For this reason, our power series are called *formal*.

The subset of M defined by $\{w \mid (r, w) \neq 0\}$ is called the *support* of r and it is denoted by $\mathrm{supp}(r)$. Thus, for any series $r \in R\langle\langle M \rangle\rangle$, we associate a subset of M, $\mathrm{supp}(r) \subseteq M$. Conversely, for any $L \subseteq M$, the *characteristic series* of L, $\mathrm{char}(L)$, is the series having coefficients 1 for all $w \in L$ and 0 for $w \in M \setminus L$:

$$\mathrm{char}(L) = \sum_{w \in L} w.$$

The subset of $R\langle\langle M \rangle\rangle$ consisting of all series with *finite support* is denoted by $R\langle M \rangle$ and its elements are called *polynomials*. We denote by 0 the series having all its coefficients equal to 0. For $a \in R \setminus \{0\}$, we denote by $a \cdot \lambda$, or simply a, the series having all coefficients 0, except for that of λ, which is equal to a. In general, for an arbitrary $w \in M$, we denote by aw, or simply w for $a = 1$, the series having all coefficients equal to 0, except for that of w, which is equal to a. These series aw are elements of $R\langle M \rangle$ and are called *monomials*. A series r is called *quasiregular*, or *proper*, if $(r, \lambda) = 0$.

We define in the following some algebraic operations on $R\langle\langle M \rangle\rangle$. The *sum* of two series $r, s \in R\langle\langle M \rangle\rangle$ is a series $r + s \in R\langle\langle M \rangle\rangle$ defined by

$$r + s = \sum_{w \in M} \left((r, w) + (s, w)\right) w.$$

For any $a \in R$, the *multiplication* of r by the *scalar* a is the series $ar \in R\langle\langle M \rangle\rangle$ defined by

$$ar = \sum_{w \in M} \left(a \cdot (r, w)\right) w.$$

Assume that the monoid M has the property that any $w \in M$ has finitely many factorizations $w = w_1 w_2$, with $w_1, w_2 \in M$; in particular, the monoids

Σ^* and $\Sigma^\oplus$ have this property. The *Cauchy product*, or simply the product, of two series $r, s \in R\langle\langle M\rangle\rangle$ is the series $rs \in R\langle\langle M\rangle\rangle$ defined by

$$rs = \sum_{w \in M} \left(\sum_{w_1 w_2 = w} (r, w_1)(s, w_2) \right) w.$$

Consequently, $(R\langle\langle M\rangle\rangle, +, \cdot, 0, 1)$ is a semiring and so is $R\langle M\rangle$. Moreover, if R is a ring, then both $R\langle\langle M\rangle\rangle$ and $R\langle M\rangle$ are also rings.

The *Hadamard product* of r and s is the series $r \odot s$ defined by

$$r \odot s = \sum_{w \in M} ((r, w)(s, w))\, w.$$

Let M_1 and M_2 be two monoids, $r \in R\langle\langle M_1\rangle\rangle$ and $s \in R\langle\langle M_2\rangle\rangle$. The *direct product* of r and s is the series $r \times s \in R\langle\langle M_1 \otimes M_2\rangle\rangle$ defined by

$$r \times s = \sum_{(w_1, w_2) \in M_1 \times M_2} ((r, w_1) \cdot (s, w_2))\, (w_1, w_2).$$

Let h be a monoid morphism from M_1 into M_2. Then h can be extended to a semiring morphism $h : R\langle\langle M_1\rangle\rangle \to R\langle\langle M_2\rangle\rangle$ as follows:

$$h(r) = \sum_{w \in M_1} (r, w) h(w),$$

under the assumption that the right-hand side is well defined. This is the case if, e.g., h is injective. The reverse mapping h^{-1} can also be extended to a mapping on power series $h^{-1} : R\langle\langle M_2\rangle\rangle \to R\langle\langle M_1\rangle\rangle$ as follows:

$$(h^{-1}s, u) = (s, h(u)).$$

h^{-1} is called an *inverse morphism*.

Let M be a monoid, R_1 and R_2 two semirings and $H : R_1 \to R_2$ a semiring morphism from R_1 into R_2. Then H can be extended to a semiring morphism $H : R_1\langle\langle M\rangle\rangle \to R_2\langle\langle M\rangle\rangle$ as follows:

$$H(r) = \sum_{w \in M} H((r, w))w.$$

Consider now, for the rest of this section, some finite alphabets $\Sigma_1, \ldots, \Sigma_n$, $n \geq 1$, and let $M = \Sigma_1^* \otimes \ldots \otimes \Sigma_n^*$. We will often call such a monoid M simply a *product* monoid, without explicitly naming the alphabets Σ_i. If $n = 1$, then $M = \Sigma_1^*$ and if $\Sigma_i = \{a_i\}$, for all $1 \leq i \leq n$, then $M = \Sigma^\oplus$, with $\Sigma = \{a_1, \ldots, a_n\}$. For $w = (w_1, \ldots, w_n) \in M$, the *length* of w is $|w| = |w_1| + \ldots + |w_n|$.

A sequence $r_1, r_2, \ldots$, of series of $R\langle\langle M\rangle\rangle$ *converges* to the *limit* r, denoted $\lim_{n\to\infty} r_n = r$, if for any $n \geq 1$, there is $k \geq 1$ such that

$$(r_l, w) = (r, w),$$

for all $w \in M$ with $|w| \leq n$, for all $l \geq k$.

Let $r \in R\langle\langle M\rangle\rangle$ be a quasiregular series. Then the sequence $r, r^2, r^3, \ldots$ converges to the limit 0. Consequently, the limit $\lim_{n\to\infty} \sum_{k=1}^{n} r^k$ exists. This limit is called the *quasi-inversion* of r and it is denoted by r^+. The series r^+ is the only series satisfying the property

$$r + rr^+ = r + r^+r = r^+.$$

We also denote $r^* = 1 + r^+$.

The operations of sum, product, and quasi-inversion are called *rational operations*. The family of *R-rational* series over M, denoted by $R^{Rat}\langle\langle M\rangle\rangle$, is the smallest subsemiring of $R\langle\langle M\rangle\rangle$, containing all polynomials and being closed under rational operations. Thus, every R-rational series can be obtained from some polynomials by a finite number of applications of some rational operations. If $M = \Sigma^*$ and $|\Sigma| = 1$, then the R-rational series are also called R-rational *sequences*.

Let $(S_n)_{n\geq 0}$ be an increasing sequence of sets of R-rational power series - with respect to inclusion - defined as follows:

(i) $S_0 = R\langle M\rangle$;

(ii) S_n is closed under sum and product, for any $n \geq 0$;

(iii) if $s \in S_n$ is a proper series, then $s^+ \in S_{n+1}$, for all $n \geq 0$.

Clearly, $\cup_{n\geq 0} S_n$ is the set of all R-rational power series. For any R-rational series r, the minimal n such that $r \in S_n$ is called the *star-height* of r. We refer for details to [20] and [67], and for results on the star-height of rational languages to [26].

A series $r \in R\langle\langle M\rangle\rangle$ is called *R-recognizable* if

$$r = (r, \lambda)\lambda + \sum_{w\neq\lambda} p(\mu w)w,$$

where $\mu : M \to R^{n\times n}$, $n \geq 1$, is a representation, and $p : R^{n\times n} \to R$ is a mapping such that, for any matrix $X = (x_{ij})_{1\leq i,j\leq n}$, the value $p(X)$ is a linear combination of the entries of X:

$$p(X) = \sum_{1\leq i,j\leq n} p_{ij}x_{ij},$$

with $p_{ij} \in R$. We denote the family of R-recognizable series by $R^{Rec}\langle\langle M\rangle\rangle$.

Let R be a commutative semiring, M a product monoid, $M = \Sigma_1^* \otimes \ldots \otimes \Sigma_n^*$, and $Z = \{z_1, \ldots, z_m\}$ an alphabet with $Z \cap \cup_{i=1}^n \Sigma_i = \emptyset$. An *algebraic system* with variables in Z is a set of equations of the form

$$z_i = p_i, \quad i = 1, \ldots, m, \tag{2.1}$$

where $p_i \in R\langle(M \cup Z)^*\rangle$. The system (2.1) is called *proper* if $(p_i, \lambda) = 0$ and $(p_i, z_j) = 0$, for all $1 \leq i, j \leq m$. Let $r_1, \ldots, r_m \in R\langle\langle M\rangle\rangle$ and let $h : R\langle(M \cup Z)^*\rangle \to R\langle\langle M\rangle\rangle$ be the unique semiring morphism such that

$$\begin{array}{ll} h(z_i) = r_i, & \text{for all } 1 \leq i \leq m, \\ h(aw) = aw, & \text{for all } w \in M, a \in R. \end{array}$$

Then $(r_1, \ldots, r_m)$ is called a *solution* of the system (2.1) if $h(z_i) = h(p_i)$, for all $1 \leq i \leq m$.

The algebraic systems can be viewed as generalizations of the context-free grammars, and so, the solutions of the algebraic systems are extensions of the context-free languages. Note however that an algebraic system may have several distinct solutions. We say that a solution σ of (2.1) is a *least solution* if $\sigma \leq \tau$ for all solutions τ of (2.1), where $\leq$ is the natural order on $R\langle\langle M\rangle\rangle$. Clearly, a least solution, if it exists, is unique. The following theorem shows that in some conditions, any algebraic system has a least solution and it gives a recurrent procedure to compute it.

Theorem 2.3.1 ([43],[45],[71]). *If R is an ω-continuous semiring, then the least solution of any algebraic system* (2.1) *exists in R^n and it is approximated by the sequence $(p^n(0) \mid n \in \mathbb{N})$.*

A series $r \in R\langle\langle M\rangle\rangle$ is called *R-algebraic* if it is a component of the least solution of an R-algebraic system. The family of R-algebraic series is denoted by $R^{Alg}\langle\langle M\rangle\rangle$.

For further details on formal power series, as well as for connections between the context-free languages and the algebraic series, we refer to [8], [21], [43], [45], [69], and [71].

Chapter 3

Families of semilinear formal power series

The semilinearity is a central notion in the theory of Formal Languages that until now has not been considered *per se* for formal power series. We discuss in Chapter 3 this notion. We define two subfamilies of rational series, both natural generalizations of the family of semilinear sets, and we clarify their relations with some well-known classes of formal power series. It turns out that the behavior of the semilinear power series under rational operations, morphisms, Hadamard product, and difference is in some respects similar to the behavior of the semilinear languages under the corresponding operations. In particular, the comparison between the behaviors of the two families of semilinear series gives a new insight on the structural details that are essential in various classical results on semilinearity. Some directions for further research are discussed at the end of this chapter.

The results of this chapter are based on our papers [59], [60], [61], [62].

3.1 Notions of semilinearity

The notion of semilinearity was originally introduced by Eilenberg and Schützenberger ([22]) as an equivalent form of the rationality in commutative monoids. For a commutative monoid $(M, \cdot)$, a subset X of M is called *linear* if

$$X = aB^*, \tag{3.1}$$

for some $a \in M$ and a finite subset B of M. Equivalently, X is called linear if

$$X = \{ab_1^{n_1} \dots b_r^{n_r} \mid n_1, \dots, n_r \in \mathbb{N}\}, \tag{3.2}$$

for some $a, b_1, \dots, b_r \in M$.

A finite union of linear sets is called a *semilinear* set. We denote the family of semilinear subsets of M by $\mathrm{SLin}(M)$. We also denote the family of

rational subsets of M by $\mathrm{Rat}(M)$. As it is straightforward to prove, see [22], it turns out that $\mathrm{Rat}(M) = \mathrm{SLin}(M)$, thus obtaining a simple representation for all rational subsets of a commutative monoid M.

As a matter of fact, quite often in literature one defines as *linear*, a set of the form (3.2), rather than a set defined by (3.1). For the classical theory of languages, this is irrelevant, as both (3.1) and (3.2) define the same family of linear sets. On the contrary, for power series, one can define two different notions of semilinearity, both equally natural generalizations of the classical notion, by considering either (3.1), or (3.2) as the definition of the linear sets. In naming these two notions of semilinearity, we choose to follow the original definition of semilinearity given by Eilenberg and Schützenberger in [22]. Thus, we define in the following a family of *semilinear series* similar to (3.1) above, and a different family of *bounded series* similar to (3.2).

Definition 1. *Let M be a commutative monoid and K a semiring. A formal power series $r \in K\langle\langle M\rangle\rangle$ is called a linear series if r is a monomial or*

$$r = pq^*, \tag{3.3}$$

for some monomial $p \in K\langle M\rangle$, and a proper polynomial $q \in K\langle M\rangle$. We denote the family of linear series by $K^{Lin}\langle\langle M\rangle\rangle$.

A semilinear series is a finite sum of linear series. We denote the family of semilinear series by $K^{SLin}\langle\langle M\rangle\rangle$.

We note that the semilinear formal power series are natural generalizations of the semilinear sets. Indeed, if we consider semilinear series with coefficients in the Boolean semiring and we identify the subsets of M with their characteristic series, then we obtain the old notion of a semilinear set.

If K is a commutative semiring then, for $p = a_0 m_0$ and $q = a_1 m_1 + \ldots + a_n m_n$, $a_i \in K, m_i \in M$, then the linear series $r = pq^*$ has the following form:

$$r = \sum_{i_1,\ldots,i_n \geq 0} \binom{i_1 + \ldots + i_n}{i_1, \ldots, i_n} a_0 a_1^{i_1} \ldots a_n^{i_n} \cdot m_0 m_1^{i_1} \ldots m_n^{i_n}, \tag{3.4}$$

where $\binom{i_1+\ldots+i_n}{i_1,\ldots,i_n}$ is the *multinomial coefficient*, $\binom{i_1+\ldots+i_n}{i_1,\ldots,i_n} = \frac{(i_1+\ldots+i_n)!}{(i_1)!\cdot\ldots\cdot(i_n)!}$

Definition 2. *A series $r \in K\langle\langle M\rangle\rangle$ is called a linear bounded series if r is a monomial or*

$$r = p_0 p_1^* \ldots p_m^*,$$

for some monomials $p_0, p_1, \ldots, p_m$, $m \geq 0$, with p_i proper monomials, for all $1 \leq i \leq m$.

A bounded series is a finite sum of linear bounded series. We denote the family of bounded series by $K^{Bound}\langle\langle M\rangle\rangle$.

The term *bounded* was introduced in [24] for subsets of $w_1^* \dots w_r^*$ for some words $w_1, \dots, w_r$; the adjective "bounded" was referring to the existence of a finite number of words w_i capable of representing in the form $w_1^{n_1} \dots w_r^{n_r}$, each word of that subset. Thus it is natural to use it here as well. In general, if $p_i = a_i m_i$, for all $0 \leq i \leq n$, then

$$p_0 p_1^* \dots p_n^* = \sum_{i_1,\dots,i_n \geq 0} a_0 a_1^{i_1} \dots a_n^{i_n} m_0 m_1^{i_1} \dots m_n^{i_n}, \tag{3.5}$$

similarly to the form (3.2) of a linear set. Note that the bounded series also generalize the family of semilinear sets. Indeed, if the coefficients are taken in the Boolean semiring, the family of bounded series coincides with the family of characteristic series of semilinear sets.

In general, although they have similar forms, the family of semilinear series and the family of bounded series are incomparable. We will prove this result in Section 3.5. The following sections are dedicated to studying their behavior under rational operations, Hadamard product, and difference.

3.2 Examples

We give here some examples of semilinear and bounded power series.

Example 3.2.1. The series $r = \sum_{n\geq 0} F_{n+1} a^n$, with $(F_n)_{n\geq 0}$ being the Fibonacci sequence, is $\mathbb{N}$-semilinear: $r = (a + a^2)^*$. Also, the series $s = \sum_{m,n\geq 0} \binom{m+n}{m} a^m b^n$ is $\mathbb{N}$-semilinear: $s = (a + b)^*$.

Example 3.2.2. The series $r' = \sum_{n\geq 0} n a^n$ is $\mathbb{N}$-bounded: $r = aa^*a^*$. Also, the series $s' = \sum_{m,n\geq 0} a^m b^n$ is $\mathbb{N}$-bounded: $s = a^*b^*$. Note that $\mathrm{supp}(r') = \mathrm{supp}(r)$ and $\mathrm{supp}(s') = \mathrm{supp}(s)$, where r and s are the series of Example 3.2.1.

3.3 Closure under rational operations

It is straightforward to prove that both families of semilinear and bounded power series contain all polynomial and are closed under sum. Moreover, the bounded power series are also closed under product. We prove in this section that although they are defined similarly, their behaviour under rational operations can be different, depending on the semiring of coefficients.

It follows directly from the definition that both the semilinear and the bounded series are rational power series of star-height at most one.

Theorem 3.3.1. *For any commutative monoid M and any semiring K, if $r \in K^{SLin}\langle\langle M \rangle\rangle \cup K^{Bound}\langle\langle M \rangle\rangle$, then r has star-height at most one.*

As it is well-known, see [22], the semilinear subsets of a commutative monoid are closed under rational operations and therefore, they coincide with the rational subsets of that monoid. The next theorem shows that a similar result holds also for series with coefficients in an idempotent semiring. Consequently, for such a semiring, the notions of rationality, semilinearity, and bounded coincide.

Theorem 3.3.2. *If K is an idempotent commutative semiring and M is a commutative monoid, then*

$$K^{Rat}\langle\langle M\rangle\rangle = K^{SLin}\langle\langle M\rangle\rangle = K^{Bound}\langle\langle M\rangle\rangle.$$

Proof. Since K is an idempotent commutative semiring, it can be easily proved that both $K^{SLin}\langle\langle M\rangle\rangle$ and $K^{Bound}\langle\langle M\rangle\rangle$ are closed under the rational operations and thus, they coincide with $K^{Rat}\langle\langle M\rangle\rangle$. The first equality of this theorem was proved in [45] in the particular case when $K = \mathbb{B}$ and $M = \Sigma^{\oplus}$, for a finite alphabet Σ. □

If K is a commutative ring, a stronger result holds, see [45]. In this case, the notions of rationality, semilinearity, and bounded coincide with the notion of linearity. In other words, any rational series has in this case a linear representation.

Theorem 3.3.3 ([45]). *If K is a commutative ring, then*

$$K^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle \;=\; K^{Lin}\langle\langle\Sigma^{\oplus}\rangle\rangle.$$

Example 3.3.1. Let $\mathbb{Z}$ be the set of integers and let $r = a^+$, $s = b^*$ be two $\mathbb{Z}$-rational (and linear) series. Then rs and r^* are $\mathbb{Z}$-linear: $rs = a(a+b-ab)^*$ and $r^* = (1-a)(2a)^*$.

If we consider series with coefficients in $\mathbb{N}$, the above results do not hold anymore: in this case, the family of rational series is strictly larger than those of semilinear series and bounded series. As a matter of fact, in this case, none of our families of series is closed under (all) rational operations. The $\mathbb{N}$-semilinear series are only closed under sum, but not under either product, or quasi-inversion, and the $\mathbb{N}$-bounded series are closed under sum and product, but not under quasi-inversion. We prove these results in the following.

Lemma 3.3.4. *For any alphabet Σ, $\mathbb{N}^{SLin}\langle\langle\Sigma^{\oplus}\rangle\rangle$ is not closed under either product, or quasi-inversion.*

Proof. Consider the semilinear series a^* and aa^*, and assume that their product

$$r = aa^*a^* = \sum_{n\geq 0} na^n,$$

is a semilinear series: $r = \sum_{i=1}^{q} r_i$, with r_i linear series.

If any of the series r_i has more than two generators, i.e., it is of the form

$$r_i = p_0 a^{k_0} (p_1 a^{k_1} + p_2 a^{k_2} + \ldots + p_m a^{k_m})^*,$$

for some $m \geq 2$, with $p_i \in \mathbb{N} \setminus \{0\}$, for all $0 \leq i \leq m$, then denoting $x_n = (r_i, a^{n+k_0})$, we obtain for large enough n that

$$\begin{aligned} x_n &= p_1 x_{n-k_1} + p_2 x_{n-k_2} + \ldots + p_m x_{n-k_m} \geq \\ &\geq x_{n-k_1} + x_{n-k_2} + \ldots + x_{n-k_m} \geq x_{n-k_1} + x_{n-k_2}. \end{aligned}$$

In particular,

$$\begin{aligned} x_{nk_1+nk_2} &\geq x_{(n-1)k_1+nk_2} + x_{nk_1+(n-1)k_2} \geq \\ &\geq x_{(n-2)k_1+nk_2} + 2x_{(n-1)k_1+(n-1)k_2} + x_{nk_1+(n-2)k_2} \geq \\ &\geq \ldots \geq \sum_{i=0}^{n} \binom{n}{i} x_{(n-i)k_1+ik_2} \geq \sum_{i=0}^{n} \binom{n}{i} p_1^{n-i} p_2^i = \\ &= (p_1 + p_2)^n \geq 2^n, \end{aligned}$$

for all large enough n. On the other hand, since r_i is a component of r, we have:

$$x_{nk_1+nk_2} = (r_i, a^{nk_1+nk_2+k_0}) \leq (r, a^{nk_1+nk_2+k_0}) = n(k_1 + k_2) + k_0,$$

and so, $2^n \leq n(k_1 + k_2) + k_0$, for all large enough n, which is absurd.

Hence, all the components of r should have only one generator:

$$r_i = p_{0i} a^{k_{0i}} (p_i a^{k_i})^*, \text{ for all } 1 \leq i \leq q.$$

If $p_i > 1$, then we obtain again exponential multiplicities. Thus, all p_i's must be equal to 1 and so, r must have bounded multiplicities. This is a contradiction, since $(r, a^n) = n$, for all $n \geq 0$. Consequently, the $\mathbb{N}$-semilinear series are not closed under product.

Also, the $\mathbb{N}$-rational series $(2a + a^2 a^*)^*$ is not semilinear. Indeed, it is proved in [4] that its star-height is two (see [4] for a detailed discussion). Thus, since $2a + a^2 a^*$ is clearly $\mathbb{N}$-semilinear, it follows by Theorem3.3.1 that the $\mathbb{N}$-semilinear series are not closed under quasi-inversion. □

Corollary 3.3.5. *For any alphabet Σ, the family of $\mathbb{N}$-rational series over Σ is strictly larger than the family of $\mathbb{N}$-semilinear series over Σ:*

$$\mathbb{N}^{SLin}\langle\langle \Sigma^{\oplus} \rangle\rangle \subset \mathbb{N}^{Rat}\langle\langle \Sigma^{\oplus} \rangle\rangle.$$

Lemma 3.3.6. *For any alphabet Σ, $\mathbb{N}^{Bound}\langle\langle \Sigma^{\oplus} \rangle\rangle$ is not closed under quasi-inversion.*

Proof. The series $2a + a^2a^*$ is clearly $\mathbb{N}$-bounded. On the other hand, $(2a + a^2a^*)^*$ is a rational series of star-height two, see [4], and thus, by Theorem 3.3.1, it is not bounded. □

Corollary 3.3.7. *For any alphabet Σ, the family of $\mathbb{N}$-rational series over Σ is strictly larger than the family of $\mathbb{N}$-bounded series over Σ:*

$$\mathbb{N}^{Bound}\langle\langle\Sigma^{\oplus}\rangle\rangle \subset \mathbb{N}^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle.$$

Having clarified the strict inclusion of the $\mathbb{N}$-semilinear and $\mathbb{N}$-bounded series in the family of $\mathbb{N}$-rational series, it remains to discuss the relationship between the two families of series. We have partially answered this question in the proof of Lemma 3.3.4: aa^*a^* is a bounded series, but not a semilinear one. A complete answer to this question is postponed for Sections 3.4 and 3.5 of this chapter.

3.4 A hierarchy of formal power series in commuting variables

We clarify in this section the relationships among the families of recognizable, rational, and algebraic series and also, the newly introduced families of semilinear and bounded power series. Moreover, we prove an extension of Mezei's theorem to the family of recognizable series over product monoid, an instrumental result in our considerations.

As usual in this chapter, we restrict ourselves to series in $\mathbb{N}\langle\langle\Sigma^{\oplus}\rangle\rangle$, where Σ is a finite alphabet. Corollary 3.4.8 and Figure 3.1 completely describe the hierarchy of the above mentioned families of series. To that aim, several relationships are considered separately in the following subsections.

3.4.1 Rational and algebraic series

We first consider the relationship between rational and algebraic series in commuting variables. It is well-known that ignoring the multiplicities, the notions of rational and algebraic coincide, as proved by Parikh's Theorem (see [56] and also Section 3.7 of this chapter). In general, the result does not hold for power series, as shown in the next theorem. We prove this result using two different algebraic methods.

Theorem 3.4.1. *For any alphabet Σ,*

$$\mathbb{N}^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle \subset \mathbb{N}^{Alg}\langle\langle\Sigma^{\oplus}\rangle\rangle.$$

First proof of Theorem 3.4.1. As it is well-known, the algebraic series are closed under rational operations. We refer to [45] and [71].

Let us assume that the $\mathbb{N}$-rational and $\mathbb{N}$-algebraic power series in commuting variables coincide. Then, the algebraic series r given by the algebraic equation

$$z = a + z^2 \tag{3.6}$$

also has a rational form.

Clearly, $supp(r) = a^+$. Denoting $x_n = (r, a^n)$ for all $n \geq 1$, we obtain for $n \geq 2$ that $x_n = (r, a^n) = (r^2, a^n) = \sum_{i=1}^{n-1} x_i x_{n-i}$. On the other hand, any rational series over a^* is generated by a rational generator function, see [71]. Consequently, a sequence $(x_n)_{n\geq 0}$ is rational if and only if for large values of n a recursion formula

$$x_n = q_1 x_{n-1} + \ldots + q_l x_{n-l}$$

holds for some $q_i \in \mathbb{N}$. Let n_0 be a positive integer such that for $n \geq n_0$, this recursion formula holds, and consider $n = 2m + 1$, with $m \geq max(l, n_0)$. Then we have:

$$\begin{aligned}
x_n =& \sum_{k=1}^{n-1} x_k x_{n-k} = \sum_{k=1}^{m} 2x_k x_{2m+1-k} = \\
=& \sum_{k=1}^{m} 2x_k (q_1 x_{2m-k} + q_2 x_{2m-k-1} + \ldots + q_l x_{2m-k-l+1}) = \\
=& q_1 \sum_{k=1}^{m} 2x_k x_{2m-k} + q_2 \sum_{k=1}^{m} 2x_k x_{2m-k-1} + \ldots + q_l \sum_{k=1}^{m} 2x_k x_{2m-k-l+1} = \\
=& q_1 (2x_1 x_{2m-1} + 2x_2 x_{2m-2} + \ldots + 2x_{m-1} x_{m+1} + 2x_m^2) + \\
& + q_2 (2x_1 x_{2m-2} + 2x_2 x_{2m-3} + \ldots + 2x_{m-1} x_m + 2x_m x_{m-1}) + \\
& \ldots \\
& + q_l (2x_1 x_{2m-l} + 2x_2 x_{2m-l-1} + \ldots + 2x_{[\frac{2m-l+1}{2}]} x_{2m-l+1-[\frac{2m-l+1}{2}]} + \\
& + \ldots + 2x_m x_{m-l+1}) = \\
=& q_1 (x_{2m} + x_m^2) + q_2 (x_{2m-1} + 2x_m x_{m-1}) + \ldots + q_l (x_{2m-l+1} + S) = \\
=& (q_1 x_{n-1} + q_2 x_{n-2} + \ldots + q_l x_{n-l}) + T,
\end{aligned}$$

where S and T are sums of positive integers. Thus, $x_n = x_n + T$, which is impossible. □

Second proof of Theorem 3.4.1. Consider the algebraic series r given by the algebraic equation (3.6) or, equivalently, generated by the grammar

$$S \to SS \mid a.$$

Denote by $f(z)$ the *structure generating function* of the above grammar, see [50], cf. also [43], [44], and [71]:

$$f(z) = \sum_{n\geq 1} (r, a^n) z^n.$$

Assuming that the series r is rational, we obtain by [71] that the function f is rational, in the sense that $f(z) = \frac{p(z)}{q(z)}$, for some polynomials p and q.

On the other hand, as shown in [14], see also [43], [44], and [45], one can compute the structure generating function of the above grammar by solving the quadratic equation (3.6) and taking the solution which vanishes at the origin. This yields the function

$$f(z) = \frac{1 - \sqrt{1 - 4z}}{2},$$

for $z < 1/4$, which is not rational. This is a contradiction, and so, r is not a rational series. □

3.4.2 Recognizable and rational series

Kleene's foundational theorem, [41], on the equivalence of rationality and regularity in formal languages has been extended in many ways. E.g., Schützenberger proved in [72] that the rational power series coincide with the recognizable power series in noncommuting variables. It turned out however, that in general, the recognizable and rational series do not coincide, see, e.g., [55] and [71]. In particular, Droste and Gastin considered in [19] the recognizable series over trace monoids, i.e., series in partially commuting variables. They characterized in the same paper the recognizable series as being the *m-c rational series*, when the semiring of coefficients is commutative, thus extending to series Ochmański's result on recognizable languages over trace monoids, see [55]. A free commutative monoid $\Sigma^{\oplus}$ is obviously a trace monoid. In this case, the recognizable series are included, by [19], in those rational series for which the star operation is restricted to series having the support included in a^*, for some $a \in \Sigma$. This proves that the recognizable series are strictly included in the rational series; e.g., $(ab)^*$ is an $\mathbb{N}$-rational series which is not recognizable by the above characterization.

We give in this section a new proof that the recognizable power series in commuting variables are strictly included in the rational power series. To this aim, we extend to formal power series a theorem due to Mezei, characterizing the recognizable languages over a product monoid.

Theorem 3.4.2 (Mezei's Theorem, [6], [21]). *A language L over a product monoid $M_1 \times M_2$ is recognizable if and only if it is a finite union of languages $L_1 \times L_2$, with L_i recognizable language over M_i, $i = 1, 2$.*

This result has been extended in [71] to recognizable series over $X_1^* \times X_2^*$, with X_1, X_2 alphabets, and we prove it here in general for series over any product monoid. If we consider series with coefficients in the Boolean semiring, we obtain the original theorem of Mezei.

The *Kronecker product* of two matrices $A \in K^{n_1 \times n_1}, B \in K^{n_2 \times n_2}$, denoted as $A \otimes B \in K^{n_1 n_2 \times n_1 n_2}$, is defined as follows:

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \dots & a_{1n_1}B \\ \dots & \dots & \dots & \dots \\ a_{n_1 1}B & a_{n_1 2}B & \dots & a_{n_1 n_1}B \end{pmatrix}.$$

The following basic property of the Kronecker product is well-known, see [45]:

$$(A \otimes C)(B \otimes D) = (AB) \otimes (CD),$$

for any matrices $A, B \in K^{n_1 \times n_1}$, $C, D \in K^{n_2 \times n_2}$.

Theorem 3.4.3 (Mezei's Theorem for formal power series). *If M_1 and M_2 are two product monoids and K is a commutative semiring, then a series $r \in K\langle\langle M_1 \otimes M_2 \rangle\rangle$ is recognizable if and only if r is a finite sum of series of the form $r_1 \times r_2$, with $r_i \in K^{Rec}\langle\langle M_i \rangle\rangle$, $i = 1, 2$.*

Proof. Consider two recognizable series $r_1 \in K^{Rec}\langle\langle M_1 \rangle\rangle$, $r_2 \in K^{Rec}\langle\langle M_2 \rangle\rangle$, with the representations $\mu_1 \in K^{m \times m}$, resp. $\mu_2 \in K^{n \times n}$:

$$r_1 = (r_1, 1)1 + \sum_{w_1 \neq 1} (\mu_1, w_1)_{1,m} w_1,$$
$$r_2 = (r_2, 1)1 + \sum_{w_2 \neq 1} (\mu_2, w_2)_{1,n} w_2.$$

Let $M = M_1 \times M_2$. We define the mapping $\mu : M \to K^{mn \times mn}$ as follows:

$$\mu(w_1, w_2) = (\mu_1 w_1) \otimes (\mu_2 w_2),$$

for all $w_1 \in M_1$, $w_2 \in M_2$.

Consider $w', w'' \in M$, $w' = (w'_1, w'_2), w'' = (w''_1, w''_2)$. Then we have:

$$\begin{aligned} \mu w' w'' =& \mu(w'_1 w''_1, w'_2 w''_2) = \mu_1 w'_1 w''_1 \otimes \mu_2 w'_2 w''_2 = (\mu_1 w'_1 \cdot \mu_1 w''_1) \otimes \\ & \otimes (\mu_2 w'_2 \cdot \mu_2 w''_2) = (\mu_1 w'_1 \otimes \mu_2 w'_2) \cdot (\mu_1 w''_1 \otimes \mu_2 w''_2) = \\ =& \mu w' \cdot \mu w''. \end{aligned}$$

Thus, μ is a representation. Then,

$$\begin{aligned} r_1 \times r_2 &= \sum_{\substack{w_1 \in M_1 \\ w_2 \in M_2}} (r_1, w_1)(r_2, w_2)(w_1, w_2) = \\ &= (r_2, 1) \sum_{w_1 \in M_1} (r_1, w_1)(w_1, 1) + (r_1, 1) \sum_{w_2 \in M_2} (r_2, w_2)(1, w_2) + \\ &\quad + \sum_{\substack{w_1 \in M_1 \setminus \{1\} \\ w_2 \in M_2 \setminus \{1\}}} (r_1, w_1)(r_2, w_2)(w_1, w_2) = \\ &= 2(r_1, 1)(r_2, 1)(1, 1) + (r_1, 1) \sum_{w_1 \in M_1 \setminus \{1\}} (\mu_1 w_1)_{1,m}(w_1, 1) + \\ &\quad + (r_1, 1) \sum_{w_2 \in M_2 \setminus \{1\}} (\mu_2 w_2)_{1,n}(1, w_2) + \sum_{\substack{w_1 \in M \setminus \{1\} \\ w_2 \in M_2 \setminus \{1\}}} (\mu w)_{1,mn}(w_1, w_2), \end{aligned}$$

and so, since the family of recognizable series is closed under sum, it follows that $r_1 \times r_2 \in K^{Rec}\langle\langle M \rangle\rangle$.

For the reverse direction, consider now a series $r \in K^{Rec}\langle\langle M \rangle\rangle$, $r = (r, 1)1 + \sum_w (\mu w)_{1,n} w$, where $\mu : M \to K^{n \times n}$ is a representation for r. We define two representations μ_1 and μ_2 on M_1 and M_2, respectively, $\mu_i : M_i \to K^{n \times n}$:

$$\mu_1 w_1 = \mu(w_1, 1), \qquad \mu_2 w_2 = \mu(1, w_2),$$

for any $w_1 \in M_1$, $w_2 \in M_2$. Then, for $w = (w_1, w_2) \in M$, we have:

$$\mu w = \mu((w_1, 1) \cdot (1, w_2)) = \mu(w_1, 1) \cdot \mu(1, w_2) = \mu_1 w_1 \cdot \mu_2 w_2,$$

and so, for $w = (w_1, w_2)$, $w_1, w_2 \neq 1$, $(r, w) = \sum_{k=1}^{n} (\mu_1 w_1)_{1,k} (\mu_2 w_2)_{k,n}$. Consequently,

$$\begin{aligned} r =& (r, 1)(1, 1) + \sum_{w_1 \in M_1 \setminus \{1\}} (\mu_1 w_1)_{1,n}(w_1, 1) + \sum_{w_2 \in M_2 \setminus \{1\}} (\mu_2 w_2)_{1,n}(1, w_2) + \\ &+ \sum_{\substack{w_1 \in M_1 \setminus \{1\} \\ w_2 \in M_2 \setminus \{1\}}} \sum_{k=1}^{n} (\mu_1 w_1)_{1,k} (\mu_2 w_2)_{k,n}(w_1, w_2), \end{aligned}$$

a decomposition into a finite sum of products of recognizable series over M_1 and M_2. □

Using Mezei's theorem, we can describe now the relationship between the $\mathbb{N}$-recognizable and the $\mathbb{N}$-rational formal power series over $\Sigma^{\oplus}$.

Theorem 3.4.4. *(i) If $|\Sigma| = 1$, then $\mathbb{N}^{Rec}\langle\langle \Sigma^{\oplus} \rangle\rangle = \mathbb{N}^{Rat}\langle\langle \Sigma^{\oplus} \rangle\rangle$.*

(ii) If $|\Sigma| \geq 2$, then $\mathbb{N}^{Rec}\langle\langle \Sigma^{\oplus} \rangle\rangle \subset \mathbb{N}^{Rat}\langle\langle \Sigma^{\oplus} \rangle\rangle$.

Proof. The first part of the theorem follows from Schützenberger's theorem, [72]; indeed, if $|\Sigma| = 1$, then $\Sigma^{\oplus} = \Sigma^*$. Consider then that $|\Sigma| \geq 2$. We prove first that $K^{Rec}\langle\langle\Sigma^{\oplus}\rangle\rangle \subseteq K^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle$, for any commutative semiring K. To this aim, consider $r \in K^{Rec}\langle\langle\Sigma^{\oplus}\rangle\rangle$, with $r = (r,1)1 + \sum_{|w|>0}(\mu w)_{1,n}w$. If we denote $A_i = \mu a_i \in K^{n\times n}$ for any $a_i \in \Sigma$, then we have:

$$\begin{aligned} r =&(r,1)1 + (A_1)_{1,n}a_1 + \ldots + (A_m)_{1,n}a_m + (A_1^2)_{1,n}a_1^2 + \ldots + \\ &+ (A_1A_m)_{1,n}a_1a_m + \ldots + (A_1^{k_1}\ldots A_m^{k_m})_{1,n}a_1^{k_1}\ldots a_m^{k_m} + \ldots \end{aligned}$$

Let R be a series over $\Sigma^{\oplus}$ with coefficients matrices in $K^{n\times n}$, given by $R = (A_1a_1)^*\ldots(A_ma_m)^*$. Then $(R, a_1^{k_1}\ldots a_m^{k_m}) = A_1^{k_1}\ldots A_m^{k_m}$ and so, r is the $(1,n)$ entry of R. In order to prove that r is a rational series it is enough to prove that R has all its entries in $K^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle$.

Clearly, if $A, B \in K^{n\times n}\langle\langle\Sigma^{\oplus}\rangle\rangle$ have all the entries in $K^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle$ then so has $A + B$ and AB. We prove that the series $Aa + A^2a^2 + \ldots$ also has all its entries in $K^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle$. Clearly,

$$(Aa + A^2a^2 + \ldots)_{ij} = \sum_{n\geq 1}(\mu a^n)_{ij}a^n,$$

and so, it is a K-recognizable series over a^*. Thus, by the representation theorem of Schützenberger, [72], it follows that $(Aa + A^2a^2 + \ldots)_{ij}$ is also a K-rational series. Consequently, $K^{Rec}\langle\langle\Sigma^{\oplus}\rangle\rangle \subseteq K^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle$.

Consider now the $\mathbb{N}$-rational series r over $a^*\otimes b^*$, $r = (ab)^* = \sum_{k\geq 0}a^kb^k$, and assume it is also an $\mathbb{N}$-recognizable series. Then, Mezei's theorem implies that

$$r = r_1 \times s_1 + \ldots + r_n \times s_n,$$

for some $\mathbb{N}$-recognizable series r_i over a^* and some $\mathbb{N}$-recognizable series s_i over b^*, for all $1 \leq i \leq k$. Clearly, we can assume without loss of generality that all series r_i, s_i are non-zero. Since r has infinite support, at least one of the series $r_i \times s_i$ has infinite support as well. We can assume without loss of generality that $r_1 \times s_1$ is that one. Thus, either r_1, or s_1 has infinite support. To make a choice, assume that r_1 has infinite support and let $n \geq 0$ be such that $(s_1, b^n) > 0$. Then, since r_1 has infinite support, there is $m > n$ such that $(r_1, a^m) > 0$. It follows then that $(r, a^mb^n) \geq (r_1 \times s_1, a^mb^n) = (r_1, a^m)(s_1, b^n) > 0$, for some $m > n$, which is impossible. Thus, $(ab)^* \in \mathbb{N}^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle \setminus \mathbb{N}^{Rec}\langle\langle\Sigma^{\oplus}\rangle\rangle$. □

3.4.3 Semilinear and recognizable series

We consider next the family of $\mathbb{N}$-semilinear power series. As it is well-known, see [22], the semilinear subsets of a commutative monoid coincide with the rational ones. In turn, it is a simple consequence of a result in [19],

that the recognizable subsets of $\Sigma^{\oplus}$ are strictly included in the rationals and so, also in the semilinears. For formal power series in commuting variables, the situation is completely different: the $\mathbb{N}$-semilinear series are strictly included in the $\mathbb{N}$-rational series (see Section 3.3) and moreover, they are incomparable with the $\mathbb{N}$-recognizable series. We prove this in the next theorem.

Theorem 3.4.5. *Any semilinear power series is rational. Moreover, for any alphabet Σ, the following holds:*

(i) If $|\Sigma| = 1$, then $\mathbb{N}^{SLin}\langle\langle\Sigma^{\oplus}\rangle\rangle \subset \mathbb{N}^{Rec}\langle\langle\Sigma^{\oplus}\rangle\rangle$.

(ii) If $|\Sigma| \geq 2$, then $\mathbb{N}^{SLin}\langle\langle\Sigma^{\oplus}\rangle\rangle$ and $\mathbb{N}^{Rec}\langle\langle\Sigma^{\oplus}\rangle\rangle$ are incomparable.

Proof. (i) If $|\Sigma| = 1$, then $\Sigma^{\oplus} = \Sigma^*$, and so, the rational and the recognizable power series over $\Sigma^{\oplus}$ coincide. Clearly, any semilinear series is a rational, and thus, also recognizable. However, as shown in Lemma 3.3.4, the series aa^*a^* is recognizable, but not semilinear.
(ii) Consider the series $a^*b^* \in \mathbb{N}\langle\langle a^* \otimes b^*\rangle\rangle$. It follows from Mezei's theorem that this is a recognizable series, since it can be decomposed as $a^* \times b^*$. Assume it is also a semilinear series:

$$a^*b^* = \sum_{i,j\geq 0} a^i b^j = \sum_{k=1}^{m} v_{k0}(v_{k1} + \ldots + v_{kn_k})^*,$$

where $v_{ki} = l_{ki}w_{ki}$, with $l_{ki} \in \mathbb{N} \setminus \{0\}$ and $w_{ki} \in a^*b^*$, for all $1 \leq k \leq m$ and $1 \leq i \leq n_k$. If in this sum there is a term with $n_k \geq 2$ then the series has some binomial coefficients larger than 1. Moreover, all coefficients must be equal to 1. Thus, a semilinear form of a^*b^* must be

$$a^*b^* = \sum_{k=1}^{m} w_{k0}w_{k1}^*,$$

for some $w_{ki} \in a^*b^*$, $i = 0, 1$, $1 \leq k \leq m$. Observe now that the geometrical interpretation of a^*b^* is the discrete plane $\mathbb{N} \times \mathbb{N}$ while that of uv^* is a discrete line. However, the discrete plane cannot be covered by a finite union of lines. Consequently, a^*b^* is not a $\mathbb{N}$-semilinear series.

On the other hand, observe that, as proved in Theorem 3.4.4, $(ab)^*$ is not a $\mathbb{N}$-recognizable series, while clearly being a semilinear one. □

3.4.4 Bounded and recognizable series

The relationship between the family of $\mathbb{N}$-bounded power series and the family of $\mathbb{N}$-recognizable power series is described in the next result.

Theorem 3.4.6. *Any bounded power series is rational. Moreover, for any finite alphabet Σ, the following hold:*

(i) if $|\Sigma| = 1$, *then* $\mathbb{N}^{Bound}\langle\langle\Sigma^{\oplus}\rangle\rangle \subset \mathbb{N}^{Rec}\langle\langle\Sigma^{\oplus}\rangle\rangle$;

(ii) if $|\Sigma| \geq 2$, *then* $\mathbb{N}^{Bound}\langle\langle\Sigma^{\oplus}\rangle\rangle$ *and* $\mathbb{N}^{Rec}\langle\langle\Sigma^{\oplus}\rangle\rangle$ *are incomparable.*

Proof. (i) Clearly, any bounded power series is rational, and thus also recognizable, for a one-letter alphabet. However, $(2a + a^2a^*)^*$ is a recognizable series, but not a bounded one. As a matter of fact, this series has star-height two, [4].

(ii) We proved in Theorem 3.4.4 that $(ab)^*$ is not a $\mathbb{N}$-recognizable series, while being a bounded one. On the other hand, as shown above, there are $\mathbb{N}$-recognizable, which are not $\mathbb{N}$-bounded. □

3.4.5 Semilinear and bounded series

To complete our discussion, it remains to settle the relationship between the family of semilinear and bounded power series. We prove that these two families of series are incomparable: aa^*a^* is a bounded, but not semilinear formal power series, and $(a + a^2)^*$ is a semilinear, but not bounded series.

Theorem 3.4.7. *For any alphabet* Σ, $\mathbb{N}^{Bound}\langle\langle\Sigma^{\oplus}\rangle\rangle$ *and* $\mathbb{N}^{SLin}\langle\langle\Sigma^{\oplus}\rangle\rangle$ *are incomparable.*

Proof. We already proved that aa^*a^* is a bounded, not semilinear series. In turn, using closure properties of semilinear and bounded series under Hadamard product, we prove in Theorem 3.5.6 that $(a+a^2)^*$ is a semilinear, not bounded series. □

The results of this section are summarized in the next corollary and illustrated in Figure 3.1, where an arrow from the family A to the family B means that A is strictly included in B, and a dashed line between A and B means that A and B are incomparable.

Corollary 3.4.8. *(i) If* $|\Sigma| = 1$, *then*

$$\mathbb{N}^{Bound}\langle\langle\Sigma^{\oplus}\rangle\rangle, \mathbb{N}^{SLin}\langle\langle\Sigma^{\oplus}\rangle\rangle \subset \mathbb{N}^{Rec}\langle\langle\Sigma^{\oplus}\rangle\rangle = \mathbb{N}^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle \subset \mathbb{N}^{Alg}\langle\langle\Sigma^{\oplus}\rangle\rangle,$$

and the families of $\mathbb{N}$*-bounded and* $\mathbb{N}$*-semilinear series over* $\Sigma^{\oplus}$ *are incomparable.*

(ii) If $|\Sigma| \geq 2$, *then*

$$\mathbb{N}^{Bound}\langle\langle\Sigma^{\oplus}\rangle\rangle, \mathbb{N}^{SLin}\langle\langle\Sigma^{\oplus}\rangle\rangle, \mathbb{N}^{Rec}\langle\langle\Sigma^{\oplus}\rangle\rangle \subset \mathbb{N}^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle \subset \mathbb{N}^{Alg}\langle\langle\Sigma^{\oplus}\rangle\rangle,$$

and the families of $\mathbb{N}$*-bounded,* $\mathbb{N}$*-semilinear, and* $\mathbb{N}$*-recognizable series over* $\Sigma^{\oplus}$ *are incomparable.*

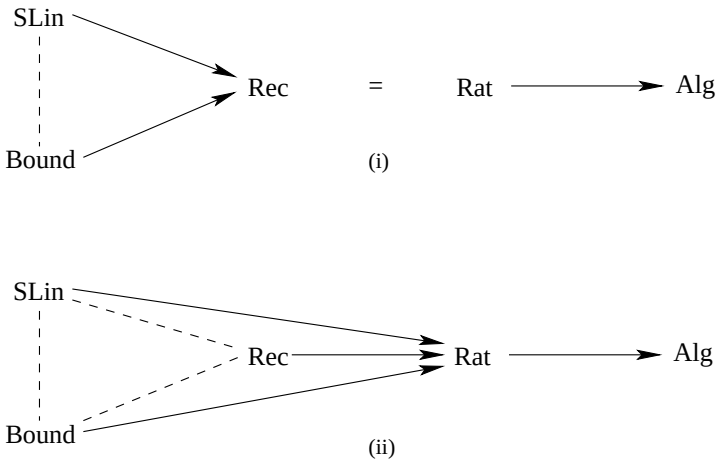


Figure 3.1: A hierarchy of families of power series over $\Sigma^{\oplus}$ for: (i) $|\Sigma| = 1$, (ii) $|\Sigma| \geq 2$.

3.5 Closure under Hadamard product

As it is well-known, see, e.g., [70], the rational languages over Σ are closed under intersection, the context-free languages are not, and the intersection of any rational language with a context-free language is context-free. Similar properties are known also for some of the corresponding families of formal power series; we recall these results in the following theorem.

Theorem 3.5.1 ([71]). *(i)* $\mathbb{N}^{Rec}\langle\langle\Sigma^{\oplus}\rangle\rangle$ *is closed under Hadamard product.*

(ii) For any $r \in \mathbb{N}^{Rec}\langle\langle\Sigma^{\oplus}\rangle\rangle$ *and* $s \in \mathbb{N}^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle$, $r \odot s \in \mathbb{N}^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle$.

(iii) $\mathbb{N}^{Alg}\langle\langle\Sigma^{\oplus}\rangle\rangle$ *is not closed under Hadamard product.*

(iv) For any $r \in \mathbb{N}^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle$ *and* $s \in \mathbb{N}^{Alg}\langle\langle\Sigma^{\oplus}\rangle\rangle$, $r \odot s \in \mathbb{N}^{Alg}\langle\langle\Sigma^{\oplus}\rangle\rangle$.

In turn, the closure properties under Hadamard product of the $\mathbb{N}$-rational series over $\Sigma^{\oplus}$ can be found only implicitly in the literature. We clarify these properties in the following result.

Theorem 3.5.2. *(i) If* $|\Sigma| = 1$, *then* $\mathbb{N}^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle$ *is closed under Hadamard product.*

(ii) If $|\Sigma| \geq 2$, *then* $\mathbb{N}^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle$ *is not closed under Hadamard product.*

Proof. (i) If $|\Sigma| = 1$, then $\mathbb{N}^{Rec}\langle\langle\Sigma^{\oplus}\rangle\rangle = \mathbb{N}^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle$, and the claim follows by Theorem 3.5.1.

(ii) Consider the $\mathbb{N}$-rational series

$$r = (a+b)^* = \sum_{i,j\geq 0} \binom{i+j}{i} a^i b^j$$

and consider its Hadamard square:

$$s = r \odot r = \sum_{i,j\geq 0} \binom{i+j}{i}^2 a^i b^j.$$

If the $\mathbb{N}$-rational series over $\Sigma^{\oplus}$ are closed under Hadamard product, then s is $\mathbb{N}$-rational.

Let h be the morphism $h : a^* \times b^* \to a^*, h(a) = h(b) = a$. Then $h(s)$,

$$h(s) = \sum_{i,j\geq 0} \binom{i+j}{j}^2 a^{i+j}$$

is $\mathbb{N}$-rational (see [71]). Let $x_n = (h(s), a^n) = \sum_{i=0}^{n} \binom{n}{i}^2 = \binom{2n}{n}$. It follows that $(x_n)_n$ is also a $\mathbb{Z}$-rational sequence and thus, it satisfies a linear recurrence as follows:

$$x_n = r_1 x_{n-1} + \ldots + r_k x_{n-k},$$

for all large enough n. Consequently,

$$\begin{aligned}(2n)(2n-1)\ldots(2n-2k+1) = r_1 n^2 (2n-2)\ldots(2n-2k+1) + \\ + r_2 n^2 (n-1)^2 (2n-4)\ldots(2n-2k+1) + \ldots + \\ + r_k n^2 (n-1)^2 \ldots (n-k+1)^2,\end{aligned}$$

for all large enough n. This is an equality of polynomials in n. But the polynomial on the left hand side has a non-zero monomial of degree 1 in n, and in turn, the polynomial on the right hand side has only monomials of degree larger than 2, which is impossible. □

Coming to semilinearity, it is known that for any commutative monoid, the intersection of two semilinear subsets is semilinear, see [22]. In particular, the intersection of two semilinear subsets of $\mathbb{N}^n$ is proved in [24] to be effectively computable.

Theorem 3.5.3 ([24]). *If X and Y are semilinear subsets of $\mathbb{N}^n$, then their intersection is also semilinear and effectively computable from X and Y.*

Although both the semilinear and the bounded series are natural generalizations of the semilinear sets, we prove in this section that they have different behaviors under Hadamard product: the $\mathbb{N}$-bounded series are closed

under Hadamard product, while the $\mathbb{N}$-semilinear series are not. As an interesting observation, we note that this provides a subfamily of rational series, that of bounded series, that is closed under Hadamard product, while the family of rational series is not closed under this operation.

We first consider the $\mathbb{N}$-semilinear series over $\Sigma^{\oplus}$. To prove that they are not closed under Hadamard product, we recall some basic results on the celebrated Fibonacci sequence.

The *Fibonacci sequence* $(F_n)_{n\geq 0}$ is recursively defined as

$$F_{n+2} = F_{n+1} + F_n,$$

for all $n \geq 0$, where $F_0 = 0$ and $F_1 = 1$. The explicit *Binet formula* for the n-th Fibonacci number is

$$F_n = \frac{\phi^n - (1-\phi)^n}{2\phi - 1},$$

where ϕ is the *golden ratio*, $\phi = (1+\sqrt{5})/2$. Of the many interesting known properties of the Fibonacci sequence, we will need in the sequel the following one:

$$F_{2n+1} = F_n^2 + F_{n+1}^2, \tag{3.7}$$

for all $n \geq 0$.

It is easy to see that the following two series generate the Fibonacci sequence:

$$a(a+a^2)^* = \sum_{n\geq 0} F_n a^n, \quad \text{and} \quad (a+a^2)^* = \sum_{n\geq 0} F_{n+1} a^n.$$

Also, we used several times in this chapter the series $(2a + a^2a^*)^*$, known to be a rational series of star-height two ([4]). It is straightforward to prove that

$$(2a + a^2a^*)^* = \sum_{n\geq 0} F_{2n+1} a^n. \tag{3.8}$$

We prove in the next theorem that the semilinear series are not closed under Hadamard product. Moreover, we also prove that the Hadamard product of two semilinear series is *always* rational if and only if we have a unary alphabet.

Theorem 3.5.4. *(i) The family of $\mathbb{N}$-semilinear series over $\Sigma^{\oplus}$ is not closed under Hadamard product.*

(ii) If $|\Sigma| = 1$, then the Hadamard product of any two $\mathbb{N}$-semilinear series over $\Sigma^{\oplus}$ is $\mathbb{N}$-rational.

(iii) If $|\Sigma| \geq 2$, then there are $\mathbb{N}$-semilinear series r and s over $\Sigma^{\oplus}$ such that $r \odot s$ is not $\mathbb{N}$-rational.

Proof. (i) Let $\Sigma = \{a\}$ and $r_1 = a(a+a^2)^*$, $r_2 = (a+a^2)^*$ be two $\mathbb{N}$-semilinear series over $\Sigma^{\oplus}$. If the $\mathbb{N}$-semilinear series are closed under Hadamard product, then $r_1 \odot r_1$ and $r_2 \odot r_2$ are semilinear. Thus,

$$r_1 \odot r_1 + r_2 \odot r_2 = \sum_{n\geq 0}(F_n^2 + F_{n+1}^2)a^n$$

is also semilinear. Note however that (3.7) and (3.8) imply that $r_1 \odot r_1 + r_2 \odot r_2 = (2a + a^2a^*)^*$; this is not a rational series of star-height one, and so, it is not semilinear, a contradiction.
(ii) For $|\Sigma| = 1$, the $\mathbb{N}$-rational series over $\Sigma^{\oplus}$ are closed under Hadamard product. The claim follows since any semilinear series is rational.
(iii) As shown in the proof of Theorem 3.5.2, the Hadamard square of $(a+b)^*$ is not $\mathbb{N}$-rational, although $(a+b)^*$ is semilinear. □

On the other hand, for $\mathbb{N}$-bounded series we can establish an extension of Ginsburg's result: this family is closed under Hadamard product and the result is effectively computable.

Theorem 3.5.5. *For any finite alphabet Σ and any $\mathbb{N}$-bounded series r_1, r_2 over $\Sigma^{\oplus}$, $r_1 \odot r_2$ is a bounded series and moreover, its bounded representation is effectively computable.*

Proof. The proof follows the ideas used in [24] for the closure of the semilinear subsets of $\mathbb{N}^n$ under intersection.

Obviously, $(\sum_{1\leq i\leq m} s_i) \odot (\sum_{1\leq j\leq n} t_j) = \sum_{\substack{1\leq i\leq m\\ 1\leq j\leq n}}(s_i \odot t_j)$ and so, it is enough to prove the claim for linear bounded series. Let r_1 and r_2 be two such series. Clearly, if at least one of them is a monomial, then $r_1 \odot r_2$ is a bounded series. Thus, without loss of generality, we can assume that r_1 and r_2 are of the form

$$r_1 = (\alpha_0 u_0)(\alpha_1 u_1)^* \dots (\alpha_m u_m)^*, \qquad r_2 = (\beta_0 v_0)(\beta_1 v_1)^* \dots (\beta_n v_n)^*,$$

for some $m, n \geq 0$, $\alpha_i, \beta_j \in \mathbb{N}\setminus\{0\}$, $u_i, v_j \in \Sigma^{\oplus}$, for all $0 \leq i \leq m$, $0 \leq j \leq n$. To simplify the notations, let

$$\alpha = (\alpha_1, \dots, \alpha_m), \quad \beta = (\beta_1, \dots, \beta_n)$$

be vectors with components from $\mathbb{N} \setminus \{0\}$ and

$$u = (u_1, \dots, u_m), \quad v = (v_1, \dots, v_n)$$

be vectors with components from $\Sigma^{\oplus}$. For a vector $\delta = (\delta_1, \dots, \delta_k)$ over $\mathbb{N}$, and a vector $x = (x_1, \dots, x_k)$ over either $\mathbb{N}$, or $\Sigma^{\oplus}$, $k \geq 1$, we denote

$$x^{\delta} = x_1^{\delta_1} \dots x_k^{\delta_k}.$$

If x is a vector over $\mathbb{N}$, then $x^\delta \in \mathbb{N}$ and if x is a vector over $\Sigma^\oplus$, then $x^\delta \in \Sigma^\oplus$.

For each $a \in \Sigma$, let $a_i = |u_i|_a$ and $a'_j = |v_j|_a$, for all $0 \leq i \leq m$, $0 \leq j \leq n$.

We first consider the support of $r \odot s$. Clearly, $\mathrm{supp}(r \odot s) = \mathrm{supp}(r) \cap \mathrm{supp}(s)$. Thus, to compute $\mathrm{supp}(r \odot s)$ we need to solve over $\mathbb{N}$ the following linear system of equations in $(k_1, \ldots, k_{m+n})$:

$$a_0 + a_1 k_1 + \ldots + a_m k_m = a'_0 + a'_1 k_{m+1} + \ldots + a'_n k_{m+n}, \qquad (3.9)$$

for all $a \in \Sigma$.

The general solution of this system is an *effectively* semilinear subset of $\mathbb{N}^{m+n}$ of the form $B + C^*$, where B and C are finite subsets of $\mathbb{N}^{m+n}$ and the iteration is done with respect to addition on $\mathbb{N}^{m+n}$, see [24].

For any vector $x = (x_1, \ldots, x_{m+n})$ over either $\Sigma^\oplus$ or $\mathbb{N}$, we denote $x' = (x_1, \ldots, x_m)$ and $x'' = (x_{m+1}, \ldots, x_{m+n})$.

It follows from (3.9) that for any $k \in B + C^*$, $u_0 u^{k'} = v_0 v^{k''}$. Moreover, $w \in \Sigma^\oplus$ is in the support of $r \odot s$ if and only if there is $k \in B + C^*$ such that $w = u_0 u^{k'}$. In this case,

$$(r \odot s, w) = \alpha_0 \alpha^{k'} \beta_0 \beta^{k''}.$$

Thus, $r \odot s$ has the following form:

$$r \odot s = \sum_{k \in B + C^*} \alpha_0 \alpha^{k'} \beta_0 \beta^{k''} u_0 u^{k'}.$$

Let $B = \{b_1, \ldots, b_p\}$ and $C = \{c_1, \ldots, c_q\}$, for some $p, q \geq 0$. Then

$$B + C^* = \{b_i + r_1 c_1 + \ldots + r_q c_q \mid 1 \leq i \leq p, r_1, \ldots, r_q \geq 0\}.$$

Thus, $w \in \mathrm{supp}(r \odot s)$ if and only if $w = u_0 u^{b'_i + r_1 c'_1 + \ldots + r_q c'_q}$, for some $1 \leq i \leq p$, $r_1, \ldots, r_q \geq 0$. Moreover, in this case,

$$\begin{aligned}(r \odot s, w) &= \alpha_0 \beta_0 \alpha^{b'_i + r_1 c'_1 + \ldots + r_q c'_q} \beta^{b''_i + r_1 c''_1 + \ldots + r_q c''_q} = \\ &= (\alpha_0 \alpha^{b'_i} \beta_0 \beta^{b''_i})(\alpha^{c'_i} \beta^{c''_i})^{r_1} \ldots (\alpha^{c'_q} \beta^{c''_q})^{r_q}.\end{aligned}$$

Denote $w_0^{(i)} = u_0 u^{b'_i}$, $w_j = u^{c'_j}$, all elements from $\Sigma^\oplus$, for all $1 \leq i \leq p$, $1 \leq j \leq q$. Denote also $\gamma_0^{(i)} = \alpha_0 \beta_0 \alpha^{b'_i} \beta^{b''_i}$, $\gamma_j = \alpha^{c'_j} \beta^{c''_j}$, elements from $\mathbb{N}$. Then, it follows that

$$\begin{aligned}r \odot s &= \sum_{w \in B + C^*} (r \odot s, w) w = \sum_{i=1}^{p} \sum_{r_1, \ldots, r_q \geq 0} \gamma_0^{(i)} w_0^{(i)} (\gamma_1 w_1)^{r_1} \ldots (\gamma_q w_q)^{r_q} = \\ &= (\gamma_0^{(1)} w_0^{(1)} + \ldots + \gamma_0^{(p)} w_0^{(p)})(\gamma_1 w_1)^* \ldots (\gamma_q w_q)^*.\end{aligned}$$

Thus, $r \odot s$ is a bounded series. Moreover, the system (3.9) can be effectively solved and so, $r \odot s$ is *effectively* bounded. □

Example 3.5.1. Let $r = a(3a^2)^*$ and $s = a^2(2a^3)^*$ be two bounded series. Clearly, $a^n \in \mathrm{supp}(r \odot s)$ if and only if $a^n = a(a^2)^i$ and $a^n = a^2(a^3)^j$, for some $i, j \geq 0$. Moreover, in this case, $(r \odot s, a^n) = 3^i \cdot 2^j$. Thus, we need to solve over $\mathbb{N}$ the following equation in i and j: $2i + 1 = 3j + 2$. Clearly, the solution of this equation is the set $\{(3k + 2, 2k + 1 \mid k \in \mathbb{N}\}$. Thus,

$$r \odot s = \sum_{k \geq 0} 3^{3k+2} \cdot 2^{2k+1} a(a^2)^{3k+2} = 18a^5(108a^6)^*.$$

Example 3.5.2. Consider the series $r = (a^2b^3)^*(3a^3b^2)^* \odot (2a^5b^6)^*$. Then $a^m b^n \in \mathrm{supp}(r)$ if and only if there are $k_1, k_2, k_3 \geq 0$ such that $m = 2k_1 + 3k_2 = 5k_3$ and $n = 3k_1 + 2k_2 = 6k_3$. In this case, $(r, a^m b^n) = 3^{k_2} \cdot 2^{k_3}$. Thus, we need to solve the Diophantine system

$$\begin{cases} 2k_1 + 3k_2 = 5k_3 \\ 3k_1 + 2k_2 = 6k_3. \end{cases}$$

The solution of this system is $\{(8n, 3n, 5n) \mid n \geq 0\}$ and so,

$$r = \sum_{n \geq 0} 3^{3n} \cdot 2^{5n} a^{25n} b^{30n} = (864a^{25}b^{30})^*.$$

Using the closure property of bounded series under Hadamard product, we can now establish the relationship between the $\mathbb{N}$-semilinear series and the $\mathbb{N}$-bounded series, announced already in Section 3.3.

Theorem 3.5.6. *For any alphabet Σ, the families of $\mathbb{N}$-semilinear and $\mathbb{N}$-bounded formal power series are incomparable.*

Proof. We have proved in the previous section that aa^*a^* is a bounded series, but not a semilinear one.

In turn, the series $r = (a + a^2)^*$ is a semilinear series. If this is a bounded series, then ar is also bounded and so, by Theorem 3.5.5, $r \odot r + (ar) \odot (ar)$ is bounded as well. However, we proved in Theorem 3.5.4(i) that this is a rational series of star-height two and so, r is not a bounded series. □

3.6 Closure under morphisms

We study in this section the behavior of the semilinear and bounded power series under (monoid) morphisms and inverse morphisms. The following theorem summarizes the known results on the closure properties of the recognizable, rational, and algebraic power series under these transformations.

Theorem 3.6.1 ([71]). *Let M_1 and M_2 be two monoids, K a semiring, and X, Y some finite alphabets.*

(i) If $h : M_1 \to M_2$ is a monoid morphism, then for any recognizable series $s \in K^{Rec}\langle\langle M_2\rangle\rangle$, $h^{-1}s \in K^{Rec}\langle\langle M_1\rangle\rangle$.

(ii) If $h : K\langle\langle M_1\rangle\rangle \to K\langle\langle M_2\rangle\rangle$ is a semiring morphism such that hw is a proper series for any $w \in M_1$, then for any rational series $r \in K^{Rat}\langle\langle M_1\rangle\rangle$, hr is well defined and $hr \in K^{Rat}\langle\langle M_2\rangle\rangle$.

(iii) If $h : X^* \to Y^*$ is a non-erasing morphism, then for any algebraic series $r \in K^{Alg}\langle\langle X^*\rangle\rangle$, hr is well defined and $hr \in K^{Alg}\langle\langle Y^*\rangle\rangle$.

The closure of the semilinear and bounded series under morphisms is proved in the next theorem.

Theorem 3.6.2. *Let Σ_1 and Σ_2 be some alphabets, $h : \Sigma_1^{\oplus} \to \Sigma_2^{\oplus}$ be a non-erasing morphism, and K a semiring.*

(i) If $r \in K^{SLin}\langle\langle \Sigma_1^{\oplus}\rangle\rangle$ then $h(r) \in K^{SLin}\langle\langle \Sigma_2^{\oplus}\rangle\rangle$.

(ii) If $r \in K^{Bound}\langle\langle \Sigma_1^{\oplus}\rangle\rangle$ then $h(r) \in K^{Bound}\langle\langle \Sigma_2^{\oplus}\rangle\rangle$.

Proof. Clearly, for a linear series $r = \alpha_0(\alpha_1 + \ldots \alpha_n)^*$, $h(r)$ is well defined and $h(r) = h(\alpha_0)(h(\alpha_1) + \ldots h(\alpha_n))^*$. Also, for a linear bounded series $s = \alpha_0\alpha_1^* \ldots \alpha_n^*$, $h(s)$ is well defined and $h(s) = h(\alpha_0)(h(\alpha_1))^* \ldots (h(\alpha_n))^*$. □

Regarding the closure under inverse morphisms, the situation appears once again to be different for semilinear and for bounded power series. It is an open question whether or not the semilinear series are closed under this transformation; on the other hand, we prove in the next theorem that the bounded series of $\mathbb{N}\langle\langle \Sigma^{\oplus}\rangle\rangle$ are closed under inverse morphisms. Our proof follows the ideas used in [24] to prove the closure of the semilinear subsets of $\mathbb{N}^n$ under inverse morphisms.

Theorem 3.6.3. *Let Σ and Δ be two disjoint alphabets. If $h : \Sigma^{\oplus} \to \Delta^{\oplus}$ is a monoid morphism, and r is a bounded series in $\mathbb{N}\langle\langle \Delta^{\oplus}\rangle\rangle$, then $h^{-1}(r)$ is a bounded series in $\mathbb{N}\langle\langle \Sigma^{\oplus}\rangle\rangle$.*

Proof. Let μ be the monoid morphism $\mu : \Sigma^{\oplus} \to \Sigma^{\oplus} \times \Delta^{\oplus}$, defined as $\mu(x) = (x, h(x))$. Assume $\Sigma = \{a_1, \ldots, a_n\}$ and consider the characteristic series of $\Sigma^{\oplus}$, the bounded series $\sigma = a_1^* \ldots a_n^*$. By Theorem 3.6.2, $\mu(\sigma)$ is a bounded series in $\mathbb{N}\langle\langle \Sigma^{\oplus} \times \Delta^{\oplus}\rangle\rangle$.

Let r be an arbitrary linear bounded series in $\mathbb{N}\langle\langle \Delta^{\oplus}\rangle\rangle$. Then $\sigma \times r$ is clearly a bounded series in $\mathbb{N}\langle\langle \Sigma^{\oplus} \times \Delta^{\oplus}\rangle\rangle$; as a matter of fact, $\sigma \times r = \sigma \cdot r$. Then, by Theorem 3.5.5, if $s = \mu(\sigma) \odot (\sigma \times r)$, then s is also a bounded series in $\mathbb{N}\langle\langle \Sigma^{\oplus} \times \Delta^{\oplus}\rangle\rangle$. Clearly,

$$\mu(\sigma) = \sum_{u \in \Sigma^{\oplus}} (u, h(u)) \quad \text{and} \quad \sigma \times r = \sum_{(u,v) \in \Sigma^{\oplus} \times \Delta^{\oplus}} (r, v) \cdot (u, v),$$

and so,

$$s = \sum_{u \in \Sigma^{\oplus}} (r, h(u)) \cdot (u, h(u)). \tag{3.10}$$

Let $\pi_1 : \Sigma^{\oplus} \times \Delta^{\oplus} \to \Sigma^{\oplus}$ be the canonical projection on the first component. It follows that

$$(\pi_1(s), u) = \sum_{v \in \Delta^{\oplus}} (s, (u, v)) = (s, (u, h(u))) = (r, h(u)) = (h^{-1}(r), u),$$

which implies that $h^{-1}(r) = \pi_1(s)$.

Note now that $s \in \mathbb{N}^{Bound}\langle\langle \Sigma^{\oplus} \times \Delta^{\oplus} \rangle\rangle$ and it is given by (3.10). Since the elements in the support of the bounded series s have the form $u \times h(u)$, s can be written as a finite sum of linear bounded series of the form $w_0 w_1^* \dots w_n^*$, with $w_i = x_i \cdot (u_i, h(u_i))$, $x_i \in \mathbb{N}$, $u_i \in \Sigma^{\oplus}$. Consequently,

$$h^{-1}(r) = \pi_1(s) = \sum x_0 u_0 (x_1 u_1)^* \dots (x_n u_n)^*,$$

which is a bounded series in $\mathbb{N}\langle\langle \Sigma^{\oplus} \rangle\rangle$.

Finally, for a bounded series $r = \sum_{i=1}^{m} r_i$, with r_i linear bounded series, $h^{-1}(r) = \sum_{i=1}^{m} h^{-1}(r_i)$, concluding the proof. □

3.7 Parikh's Theorem for formal power series

The notions of semilinear, rational, recognizable, and algebraic can be considered in various algebraic settings. We have been discussing them so far in this chapter in the monoid Σ^* of formal languages over the alphabet Σ, as well as in the semirings $\mathbb{N}\langle\langle \Sigma^* \rangle\rangle$ and $\mathbb{N}\langle\langle \Sigma^{\oplus} \rangle\rangle$ of formal power series in non-commuting, resp. commuting variables in Σ, with coefficients in $\mathbb{N}$. It comes as a natural problem to study the connections between the corresponding notions in these three different algebraic structures. We investigate the problem in this section. In particular, we prove that the famous theorem of Parikh stating that the above mentioned five notions coincide in $\Sigma^{\oplus}$, does not hold when the multiplicities are taken into consideration.

The following result summarizes the known results in Σ^* and $\mathbb{N}\langle\langle \Sigma^* \rangle\rangle$.

Theorem 3.7.1 ([70],[71]). *(i) For any finite alphabet Σ,*

$$\mathrm{Bound}(\Sigma^*) = \mathrm{SLin}(\Sigma^*), \qquad \mathrm{Rec}(\Sigma^*) = \mathrm{Rat}(\Sigma^*),$$

and $\mathrm{SLin}(\Sigma^*) \subseteq \mathrm{Rat}(\Sigma^*) \subseteq \mathrm{Alg}(\Sigma^*)$. Moreover, the inclusions are strict if and only if $|\Sigma| \geq 2$.

(ii) For any finite alphabet Σ,

$$\mathbb{N}^{Rec}\langle\langle \Sigma^* \rangle\rangle = \mathbb{N}^{Rat}\langle\langle \Sigma^* \rangle\rangle \subset \mathbb{N}^{Alg}\langle\langle \Sigma^* \rangle\rangle,$$

and $\mathbb{N}^{Bound}\langle\langle \Sigma^* \rangle\rangle$ and $\mathbb{N}^{SLin}\langle\langle \Sigma^* \rangle\rangle$ are incomparable families of rational series.

The equality $\mathrm{Rec}(\Sigma^*) = \mathrm{Rat}(\Sigma^*)$ is known in literature as Kleene's theorem, [41], and the equality $\mathbb{N}^{Rec}\langle\langle\Sigma^*\rangle\rangle = \mathbb{N}^{Rat}\langle\langle\Sigma^*\rangle\rangle$ is known as the representation theorem of Schützenberger, [71]. Note also that the family of $\mathbb{N}$-rational ($\mathbb{N}$-algebraic) series over Σ^* can be equivalently defined as the rational (context-free, resp.) languages over Σ with multiplicities - or ambiguities - given by a generating or accepting device (see [21], [70], [71] for details).

It is useful to observe that the Parikh mapping, [56], is the monoid morphism $\pi : \Sigma^* \to \Sigma^{\oplus}$ defined as $\pi(a) = a$, for all $a \in \Sigma$. Thus, for two words $u, v \in \Sigma^*$, $\pi(u) = \pi(v)$ if and only if $|u|_a = |v|_a$, for all $a \in \Sigma$.

Since π is a non-erasing monoid morphism, it can be naturally extended to a semiring morphism $\pi : \mathbb{N}\langle\langle\Sigma^*\rangle\rangle \to \mathbb{N}\langle\langle\Sigma^{\oplus}\rangle\rangle$. We prove in the following results that the bounded, semilinear, rational, and algebraic power series of $\mathbb{N}\langle\langle\Sigma^{\oplus}\rangle\rangle$ are the Parikh images of the corresponding families of $\mathbb{N}\langle\langle\Sigma^*\rangle\rangle$. It turns out however that this result does not hold for recognizable series.

Theorem 3.7.2. *For any $\mathbb{N}$-rational series r over Σ^*, πr is an $\mathbb{N}$-rational series over $\Sigma^{\oplus}$. Moreover,*

$$\pi : \mathbb{N}^{Rat}\langle\langle\Sigma^*\rangle\rangle \to \mathbb{N}^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle$$

is a surjective semiring morphism.

Proof. Let $r \in \mathbb{N}^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle$. Then, see [71], there is a proper linear system $Z = P + QZ$ such that

$$P = (P, 1)1, \quad Q = \sum_{\substack{|w|=1 \\ w\in\Sigma^*}} (Q, w)w$$

and r is the first component of its solution vector.

Then clearly, see also [43], πr is the first component of the system $Z' = \pi(P) + \pi(Q)Z'$. Since $\pi(P) = P$ and $\pi(Q) = Q$, πr is a rational series in $\mathbb{N}\langle\langle\Sigma^{\oplus}\rangle\rangle$.

To prove the surjectivity of π, consider $r \in \mathbb{N}^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle$. Then there exists P, Q such that

$$P = (P, 1), \quad Q = \sum_{\substack{|w|=1 \\ w\in\Sigma^{\oplus}}} (Q, w)w$$

and r is the first component of the solution for $Z = P + QZ$. Note however that the very same system can be considered in $\mathbb{N}\langle\langle\Sigma^*\rangle\rangle$; let Z' be its unique solution with rational components. Then,

$$\pi(Z') = \pi(P + QZ') = \pi(P) + \pi(Q)\pi(Z') = P + Q\pi(Z').$$

Since Q is proper, it follows, see [45], [71]), that the solution of this linear system is unique. Thus, $\pi(Z') = Z$, i.e., π is surjective. □

Corollary 3.7.3. *(i) If $|\Sigma| = 1$, then for any $\mathbb{N}$-recognizable series r over Σ^*, πr is an $\mathbb{N}$-recognizable series over $\Sigma^\oplus$. Moreover,*

$$\pi : \mathbb{N}^{Rec}\langle\langle\Sigma^*\rangle\rangle \to \mathbb{N}^{Rec}\langle\langle\Sigma^\oplus\rangle\rangle$$

is a surjective semiring morphism.

(ii) If $|\Sigma| \geq 2$, then there are series $r \in \mathbb{N}^{Rec}\langle\langle\Sigma^*\rangle\rangle$ such that $\pi r \notin \mathbb{N}^{Rec}\langle\langle\Sigma^\oplus\rangle\rangle$.

Proof. The result follows by Theorems 3.7.1, 3.7.2, and 3.4.4. Indeed, $\mathbb{N}^{Rec}\langle\langle\Sigma^\oplus\rangle\rangle \subset \mathbb{N}^{Rat}\langle\langle\Sigma^\oplus\rangle\rangle$, while $\mathbb{N}^{Rec}\langle\langle\Sigma^*\rangle\rangle = \mathbb{N}^{Rat}\langle\langle\Sigma^*\rangle\rangle$. E.g., $r = (ab)^*$ is an $\mathbb{N}$-recognizable series over Σ^*, such that πr is not recognizable over $\Sigma^\oplus$. Note however that πr is a rational series over $\Sigma^\oplus$. □

Theorem 3.7.4. *For any $\mathbb{N}$-algebraic series r over Σ^*, πr is an $\mathbb{N}$-algebraic series over $\Sigma^\oplus$. Moreover,*

$$\pi : \mathbb{N}^{Alg}\langle\langle\Sigma^*\rangle\rangle \to \mathbb{N}^{Alg}\langle\langle\Sigma^\oplus\rangle\rangle$$

is a surjective semiring morphism.

Proof. Let r be an $\mathbb{N}$-algebraic series over Σ^*, defined by the first component of the solution of the following system

$$z_i = p_i, \quad i = 1, \ldots, n.$$

Then, see [43], $\pi(r)$ is the first component of the solution of the following system:

$$z_i = \pi(p_i), \quad i = 1, \ldots, n$$

over $\Sigma^\oplus$, proving that $\pi(r)$ is an algebraic series over $\Sigma^\oplus$.

For the surjectivity of π we use the quadratic normal form of an algebraic series, see, e.g., [45]. For some proper series $r \in \mathbb{N}^{Alg}\langle\langle\Sigma^\oplus\rangle\rangle$, there is a system in the quadratic normal form such that r is the first component of the following system of equations:

$$z_i = p_i,\ i = 1, \ldots, n, \quad supp\ p_i \subseteq M_1 \cup Z^2,$$

where $M_1 = \{w \in \Sigma^\oplus \mid |w| = 1\} = \{a_1, \ldots, a_n\}$.

Clearly, the very same system can be also considered over Σ^*. Let r' be the first component of the solution of the system in this case. It follows that $\pi(r')$ and r verify the same proper algebraic system over $\Sigma^\oplus$. However, such a system has a unique proper solution, and so, $\pi(r') = r$. □

Theorem 3.7.5. *For any $\mathbb{N}$-semilinear series r over Σ^*, πr is an $\mathbb{N}$-semilinear series over $\Sigma^\oplus$. Moreover,*

$$\pi : \mathbb{N}^{SLin}\langle\langle\Sigma^*\rangle\rangle \to \mathbb{N}^{SLin}\langle\langle\Sigma^\oplus\rangle\rangle$$

is a surjective mapping.

Proof. Let $r = pq^*$ be an $\mathbb{N}$-linear series over Σ^*. Then $\pi r = \pi p(\pi q)^*$ is an $\mathbb{N}$-linear series over $\Sigma^\oplus$. Since π is a semiring morphism, for any series r_1, r_2, $\pi(r_1 + r_2) = \pi r_1 + \pi r_2$, concluding the claim. □

A similar argument holds also for bounded series.

Theorem 3.7.6. *For any $\mathbb{N}$-bounded series r over Σ^*, πr is an $\mathbb{N}$-bounded series over $\Sigma^\oplus$. Moreover,*

$$\pi : \mathbb{N}^{Bound}\langle\langle\Sigma^*\rangle\rangle \to \mathbb{N}^{Bound}\langle\langle\Sigma^\oplus\rangle\rangle$$

is a surjective mapping.

The celebrated theorem of Parikh, [56], states that for any context-free language, there is a rational (and in fact, even semilinear) language having the same Parikh image. Reformulated, Parikh's theorem states that the notions of bounded, semilinear, recognizable, rational, and algebraic coincide in $\Sigma^\oplus$. Moreover, the result holds also for formal power series, if the semiring of coefficients satisfies some specific properties, as we recall in the next theorem, see also [1], [30], and [63] for other approaches.

Theorem 3.7.7 ([42],[43]). *For any idempotent, commutative, ω-continuous semiring K, and any alphabet Σ,*

$$K^{Alg}\langle\langle\Sigma^\oplus\rangle\rangle = K^{Rat}\langle\langle\Sigma^\oplus\rangle\rangle = K^{SLin}\langle\langle\Sigma^\oplus\rangle\rangle.$$

Consequently, Parikh's theorem holds for K-series in commuting variables.

On the other hand, it follows from the results of this section and those of Section 3.4 that Parikh's theorem does not hold anymore when the *multiplicities* are taken into consideration. Indeed, the following hierarchy of formal power series is strict: $\mathbb{N}^{SLin}\langle\langle\Sigma^\oplus\rangle\rangle \subset \mathbb{N}^{Rat}\langle\langle\Sigma^\oplus\rangle\rangle \subset \mathbb{N}^{Alg}\langle\langle\Sigma^\oplus\rangle\rangle$. It follows from the proof of Theorem 3.4.1 that for the following simple context-free grammar,

$$S \to SS \mid a,$$

there is no rational language having the same Parikh image with multiplicities. In this grammar, a^n has the multiplicity (ambiguity) C_{n-1}, where C_n is *the n-th Catalan number*,

$$C_n = \frac{1}{n+1}\binom{2n}{n}.$$

Note though that the rational language aa^* has the same Parikh image (without multiplicities) as the context-free language generated by the above grammar. We thus proved the following result.

Theorem 3.7.8 (Parikh's theorem). *(i) For any alphabet Σ and any context-free set over Σ, there is a semilinear set having the same Parikh image.*

(ii) For any alphabet Σ, Parikh's theorem does not hold when the multiplicities are taken into consideration. We have the following strict hierarchy of formal power series:

$$\mathbb{N}^{SLin}\langle\langle\Sigma^{\oplus}\rangle\rangle \subset \mathbb{N}^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle \subset \mathbb{N}^{Alg}\langle\langle\Sigma^{\oplus}\rangle\rangle.$$

3.8 The difference operation on semilinear power series

The semilinear subsets of a commutative monoid are closed under (set) difference, see [22]. Thus, the semilinear subsets of $\mathbb{N}^k$, $k \geq 1$, are closed under (set) difference, and moreover, see [24], the result is effectively computable. The problem appears to be more difficult for formal power series. Thus, the $\mathbb{N}$-rational power series in both commuting and noncommuting variables are not closed under difference, see, e.g., [21] and [71]. However, an operation of *quasi-difference* $r \dot{-} s$ is defined as $(r \dot{-} s, x) = \max(0, (r, x) - (s, x))$ and it is proved that if r and s are rational series in noncommuting variables and s has bounded multiplicities, then the quasi-difference $r \dot{-} s$ is rational. Thus, if $r \geq s$ and s has bounded coefficients, then $r - s$ is rational.

Theorem 3.8.1 ([21]). *(i) For any finite alphabet Σ, the $\mathbb{N}$-rational power series over Σ^* are not closed under difference.*

(ii) For any finite alphabet Σ, if $r, s \in \mathbb{N}^{Rat}\langle\langle\Sigma^*\rangle\rangle$ and s has bounded coefficients, then $r \dot{-} s \in \mathbb{N}^{Rat}\langle\langle\Sigma^*\rangle\rangle$.

We prove in this section a result of the same type: if r and s are semilinear power series in commuting variables, s has bounded coefficients, and $r \geq s$, then $r - s$ is a rational series of star-height one, and thus, of an "almost semilinear" form. However, the commutation makes the problem completely different, and we prove our result by combinatorial means, avoiding the use of a deep algebraic tool as the *cross-section* theorem in [21].

3.8.1 Semilinear series with bounded coefficients

The main emphasis in this section will be on semilinear power series with bounded coefficients and on some of their properties. We say that a semilinear series $r \in \mathbb{N}\langle\langle\Sigma^{\oplus}\rangle\rangle$ has *bounded coefficients* if there is a constant $K \in \mathbb{N}$ such that $(r, u) \leq K$, for all $u \in \Sigma^{\oplus}$. Throughout this section, we will always consider series in $\mathbb{N}\langle\langle\Sigma^{\oplus}\rangle\rangle$.

We describe in the next theorem the general form of the semilinear series with bounded coefficients.

Theorem 3.8.2. *Let s be a semilinear series in $\mathbb{N}^{SLin}\langle\langle\Sigma^{\oplus}\rangle\rangle$. If s has bounded coefficients, then $s = m_1 u_1^* + \ldots + m_k u_k^* + p$, for some monomials $m_1, \ldots, m_k$, words $u_1, \ldots, u_k \in \Sigma^{\oplus}$, and a polynomial p.*

Proof. If s is a semilinear series, then it is of the form

$$s = m_1 p_1^* + \ldots + m_k p_k^* + p_{k+1},$$

for some monomials $m_1, \ldots, m_k$, and some proper polynomials $p_1, \ldots, p_{k+1}$. Consider the polynomial $p_1 = \alpha_1 u_1 + \ldots + \alpha_i u_i$, with $\alpha_1, \ldots, \alpha_i \in \mathbb{N} \setminus \{0\}$, and $u_1, \ldots, u_i \in \Sigma^{\oplus}$, $i \geq 1$. All the coefficients $\alpha_1, \ldots, \alpha_i$ must be equal to 1 since otherwise, p_1^* and s have exponential coefficients. Moreover, if p_1 is not a monomial, i.e. $i \geq 2$, then $(p_1^*, u_1^{j_1} u_2^{j_2}) \geq \binom{j_1+j_2}{j_1}$, and this is not bounded for $j_1, j_2 \geq 0$. Consequently, $p_1 = u_1$, for some $u_1 \in \Sigma^{\oplus}$. It follows similarly that $p_i = u_i$, for some $u_i \in \Sigma^{\oplus}$, for all $2 \leq i \leq k$. □

3.8.2 The base of a semilinear series

Computing the difference between a semilinear series and a semilinear series with bounded coefficients, we will be typically interested in knowing how words of the form vu^n, $n \in \mathbb{N}$, are generated by a linear series, for some $u, v \in \Sigma^{\oplus}$. The notion of *base* of a linear series turns out to be instrumental in this respect.

Definition 3. *Let $r = pq^*$ be a linear series, with p a monomial, and q a proper polynomial, $p = \alpha_0 u_0$, $q = \alpha_1 u_1 + \ldots + \alpha_m u_m$, for some $\alpha_i \in \mathbb{N} \setminus \{0\}$, $u_i \in \Sigma^{\oplus}$, for all $0 \leq i \leq m$. The base of the series r is the vector ω_r containing all the words appearing as monomials in q:*

$$\omega_r = (u_1, \ldots, u_m).$$

To any word $v \in \Sigma^{\oplus}$, we associate with respect to r a finite set of vectors from $\mathbb{N}^m$ in the following way: if

$$v = u_0 u_1^{n_1} \ldots u_m^{n_m},$$

for some $n_1, \ldots, n_m \in \mathbb{N}$, then the vector $t = (n_1, \ldots, n_m)$ is associated to v. In this case, we denote $v = u_0 \omega^t$ and say that t is a *representation of v with respect to the series r*. We denote by $\mathcal{R}_r(v)$ the set of all representations of v with respect to r.

Note that in general, a word v can have more than one representation with respect to a given linear series r, and that $\mathcal{R}_r(v) \neq \emptyset$ if and only if $(r, v) > 0$. Moreover, for any two words v_1, v_2, $\mathcal{R}_r(v_1) \cap \mathcal{R}_r(v_2) \neq \emptyset$ if and only if $v_1 = v_2$, if and only if $\mathcal{R}_r(v_1) = \mathcal{R}_r(v_2)$.

Example 3.8.1. For $r = (a + a^2)^*$, its base is $\omega = (a, a^2)$. Then, $\mathcal{R}_r(a^4) = \{(4, 0), (2, 1), (0, 2)\}$ and $\mathcal{R}_r(a^m b^n) = \emptyset$, for all $m \geq 0$, $n \geq 1$.

For two vectors $t, s \in \mathbb{N}^m$, $m \geq 1$, $t = (t_1, \ldots, t_m)$, $s = (s_1, \ldots, s_m)$, we say that $t \leq s$ if $t_i \leq s_i$, for all $1 \leq i \leq m$. The relation $\leq$ thus defined is an order relation on $\mathbb{N}^m$.

One important tool used often in the sequel is provided by the following decomposition result.

Lemma 3.8.3. *Let $r = pq^*$ be a linear series, for a monomial p and a proper polynomial q. For any $v \in \Sigma^\oplus$, if $(r, v) > 0$, then $r = vq^* + p'q^* + p''$, for some polynomials p' and p''.*

Proof. If $(r, v) > 0$, then $v = v_1v_2$, with $(p, v_1) > 0$, and $(q^*, v_2) > 0$, i.e., $(q^m, v_2) > 0$, for some $m \geq 0$. Then

$$\begin{aligned} r &= p(1 + q + \ldots + q^{m-1} + q^m q^*) = \\ &= p(1 + q + \ldots + q^{m-1}) + vq^* + (pq^m - v)q^*. \end{aligned}$$

□

The following representation theorem will be useful in the sequel.

Theorem 3.8.4. *Let $r = pq^*$ be a linear series, for a monomial p and a proper polynomial q, and let $u, v \in \Sigma^\oplus$ such that $(r, vu^n) > 0$ for some $n \geq 0$. Then r is of the form*

$$r = (vu^{n_1} + \ldots + vu^{n_k})q^* + p'q^* + p'',$$

for polynomials p', p'' and $n_1, \ldots, n_k \in \mathbb{N}$, such that for all $n \in \mathbb{N}$,

$$(r, vu^n) > 0 \quad \textit{if and only if} \quad (vu^{n_i}q^*, vu^n) > 0,$$

for some $1 \leq i \leq k$.

Proof. If $q = \alpha_1 u_1 + \ldots + \alpha_m u_m$, with $\alpha_i \in \mathbb{N} \setminus \{0\}$, $u_i \in \Sigma^\oplus$, then the base of r is the vector $\omega = (u_1, \ldots, u_m)$.

Let $\mathcal{R} = \bigcup_{n \geq 0} \mathcal{R}_r(vu^n)$. By König's Lemma, see, e.g., [24], there is in $\mathcal{R}$ only a finite set of minimal vectors with respect to $\leq$. Let $\{t_1, \ldots, t_k\}$ be this set, where $t_i = (t_{i1}, \ldots, t_{im})$, and let vu^{n_i} be the *unique* word such that $t_i \in \mathcal{R}_r(vu^{n_i})$, for all $1 \leq i \leq k$. We denote $t_{i0} = t_{i1} + \ldots + t_{ik}$, for all $1 \leq i \leq k$, and without loss of generality, we assume that $t_{i0} \leq t_{j0}$, for $i \leq j$. Then $(pq^{t_{10}}, vu^{n_1}) > 0$ and

$$r = p(1 + q + \ldots + q^{t_{10}-1} + q^{t_{10}}q^*) = p' + vu^{n_1}q^* + p''q^*,$$

where $p' = p(1 + q + \ldots + q^{t_{10}-1})$ and $p'' = q^{t_{10}} - vu^{n_1}$ are polynomials.

Observe now that $(p''q^*, vu^{n_i}) > 0$, for all $2 \leq i \leq k$. Indeed, if this is not true, then $(p' + vu^{n_1}q^*, vu^{n_i}) > 0$. Since $t_{10} \leq t_{i0}$, for all $2 \leq i \leq k$, we obtain that $(vu^{n_1}q^*, vu^{n_i}) > 0$ and so, $t_1 \leq \tau_i$, for all $\tau_i \in \mathcal{R}_r(vu^{n_i})$. This contradicts the fact that t_1 and t_i are incomparable for all $2 \leq i \leq k$.

Iterating the procedure, it follows that r is of the form

$$r = (vu^{n_1} + \ldots vu^{n_k})q^* + p'q^* + p''.$$

Let n be a positive integer such that $(r, vu^n) > 0$. Then, $\mathcal{R}_r(vu^n) \neq \emptyset$; let $t \in \mathcal{R}_r(vu^n) \subseteq \mathcal{R}$. Consequently, there is $1 \leq i \leq k$ such that $t_i \leq t$. It follows then from the definition of the representation that $(vu^{n_i}q^*, vu^n) > 0$, concluding the claim of the theorem. □

Using Theorem 3.8.4, we can prove now that for any semilinear series r and any word u, the set of powers of u included in the support of r is semilinear. We recall first a well-known number theoretical result.

Lemma 3.8.5. *Let $n_1, \ldots, n_k$ be positive integers, and let d be their greatest common divisor. For any large enough integer n, n can be written as a linear combination of $n_1 \ldots, n_k$ with coefficients in $\mathbb{N}$ if and only if n is a multiple of d.*

Theorem 3.8.6. *For any semilinear series r and any $u \in \Sigma^{\oplus}$, the set*

$$P_{r,u} = \{n \in \mathbb{N} \mid (r, u^n) > 0\}$$

is a semilinear set of nonnegative integers.

Proof. Clearly, $P_{r_1+r_2,u} = P_{r_1,u} \cup P_{r_2,u}$ and so, it is enough to prove the claim for a linear series. Assume thus that $r = pq^*$, for a monomial p and a proper polynomial q. By Theorem 3.8.4,

$$r = (u^{n_1} + \ldots + u^{n_k})q^* + p'q^* + p'',$$

for some $n_1, \ldots, n_k \in \mathbb{N}$ and polynomials p', p'' such that $(r, u^n) > 0$ if and only if $(u^{n_i}q^*, u^n) > 0$ for some $1 \leq i \leq k$. Thus, we can assume without loss of generality that $r = q^*$, for a proper polynomial q.

By Theorem 3.8.4, $r = (u^{n_1} + \ldots + u^{n_k})r + p'q^* + p''$, for some $n_1, \ldots, n_k \in \mathbb{N}$ and polynomials p', p'', such that for any positive integer n, $(r, u^n) > 0$ if and only if $(u^{n_i}r, u^n) > 0$, for some $1 \leq i \leq k$. Equivalently, for any nonnegative integer n, $(r, u^n) > 0$ if and only if n is a linear combination of $n_1, \ldots, n_k$. The claim follows now by Lemma 3.8.5. □

3.8.3 The difference operation

For $\mathbb{N}$-semilinear power series, it is an open and apparently difficult problem whether or not they are closed under difference. The smaller family of $\mathbb{N}$-linear series is not closed under this operation, as it can be proved by considering the difference $(2a)^* - a^*$. Note however, that this is a rational series of star-height one, namely $aa^*(2a)^*$ (as one can compute using the techniques developed in this subsection, see also [4]). We believe that the $\mathbb{N}$-semilinear series are not closed under difference either, even in the simpler case when one of the series has bounded coefficients. While this problem remains open, we prove in this section that for two $\mathbb{N}$-semilinear formal power series r and s, such that s has bounded coefficients and $r \geq s$, the difference $r - s$ is an $\mathbb{N}$-rational power series of star-height one, effectively computable from r and s. We first solve the problem in several simpler instances, to which we will reduce the general problem.

We need the following result, see [24].

Lemma 3.8.7 ([24]). *The family of semilinear sets of nonnegative integers is closed under set difference.*

Lemma 3.8.8. *Let* $r, s \in \mathbb{N}^{SLin}\langle\langle\Sigma^{\oplus}\rangle\rangle$,

$$r = pq^*, \qquad s = uv^*,$$

for a polynomial p, *a proper polynomial* q, *and words* $u, v \in \Sigma^{\oplus}$. *If* $r \geq s$, *then* $r - s$ *is a rational series. Moreover,* $r - s$ *is of the form*

$$p_1 + p_2 q^* + p_3 (u^d)^* + p_4 q^* (u^d)^*,$$

with p_1, p_2, p_3, p_4 *polynomials and* $d \in \mathbb{N}$. *Also,* $q^* \geq (u^d)^*$.

Proof. By Theorem 3.8.4, we can assume without loss of generality that $v = 1$ and $p = u^{l_1} + \ldots + u^{l_k}$, for some $k, l_1, \ldots, l_k \in \mathbb{N}$. Thus, it is enough to prove the claim for $r = (u^{l_1} + \ldots + u^{l_k})q^*$ and $s = u^*$.

Clearly, there must be a positive integer n such that $(q^*, u^n) > 0$; let d be the minimal such integer. If $r \geq s$, then $(r, u^i) > 0$, for all $0 \leq i < d$ and thus, we must have $\{0, 1, \ldots, d-1\} \subseteq \{l_1, \ldots, l_k\}$. Since $\sum_{i=0}^{d-1} u^i q^* \geq u^*$, it is enough to assume that $r = (1 + u + \ldots + u^{d-1})q^*$. Note also that $q^* \geq (u^d)^*$.

Since $u^* = (1 + u + \ldots + u^{d-1})(u^d)^*$, it follows that

$$r - s = \sum_{i=0}^{d-1} u^i (q^* - (u^d)^*).$$

It is enough to prove now that $q^* - (u^d)^*$ is a rational series.

By Lemma 3.8.3, if $q^* \geq (u^d)^*$, then $q^* = 1 + u^d q^* + p_1 q^* + p_2$, for some polynomials p_1 and p_2. Consequently,

$$\begin{aligned} q^* - (u^d)^* &= (1 + u^d q^* + p_1 q^* + p_2) - (1 + u^d (u^d)^*) = \\ &= u^d (q^* - (u^d)^*) + p_1 q^* + p_2 = p_1 q^* (u^d)^* + p_2 (u^d)^*, \end{aligned}$$

which is a rational series of the form specified in the lemma. □

The result proved in Lemma 3.8.8 may look somehow troubling because we need to compute the difference between one (semi)linear series and the sum of several linear series; however, Lemma 3.8.8 shows that after the first difference, the result is not necessarily a semilinear series and thus, it appears that we cannot have an iterative procedure. Lemma 3.8.10 shows that such an iterative procedure is in fact possible and the difficulty described above is solved in Lemma 3.8.9.

Lemma 3.8.9. *Let* $r = pq^*u^*$ *be a rational series, for a polynomial* p, *a proper polynomial* q, *and a word* $u \in \Sigma^{\oplus}$, *such that* $q^* \geq u^*$, *and let* $v \in \Sigma^{\oplus}$. *If* $(r, v) > 0$, *then also* $(pq^*, v) > 0$.

Proof. If $(r, v) > 0$, then $v = u_1 \cdot u_2 \cdot u^l$, for some $l \in \mathbb{N}$, $u_1, u_2 \in \Sigma^{\oplus}$, such that $(p, u_1) > 0$ and $(q^*, u_2) > 0$. But $q^* \geq u^*$, i.e. $(q^*, u^l) > 0$. Thus, by Lemma 3.8.3,

$$q^* = u^l q^* + p_1 q^* + p_2,$$

for some polynomials p_1, p_2. Consequently, $(q^*, u^l u_2) \geq (q^*, u_2) > 0$, and so, $(pq^*, v) > 0$. □

Lemma 3.8.10. *Let r be a semilinear series and $s = vu^*$ for some $u, v \in \Sigma^{\oplus}$. If $r \geq s$, then $r - s$ is a rational series. Moreover, $r - s$ is of the form*

$$r - s = p'_0 + \sum_{i=1}^{m} p'_i (q'_i)^* + \sum_{i=m+1}^{m+n} p'_i (q'_i)^* (u^{d_i})^*,$$

for some $m, n, d_{m+1}, \ldots, d_{m+n} \in \mathbb{N}$ and for some polynomials p'_0, p'_i, q'_i, $1 \leq i \leq m+n$. Also, $q'^*_i \geq (u^{d_i})^*$, for all $m+1 \leq i \leq m+n$.

Proof. Similarly as in the proof of Lemma 3.8.8, one can reduce the problem to the case when $v = 1$: $r - s = v(r' - u^*) + r''$, where r', r'' are semilinear series and $r' \geq u^*$. Assume thus that $v = 1$, and let $r = p_1 q_1^* + \ldots + p_m q_m^*$, for some polynomials p_i, q_i, $1 \leq i \leq m$.

By Theorem 3.8.6, the set $P_i = \{n \in \mathbb{N} \mid (p_i q_i^*, u^n) > 0\}$ is semilinear. But then, by Lemma 3.8.7, the sets

$$P'_i = P_i - \bigcup_{j=1}^{i-1} P_j$$

are also semilinear sets of nonnegative integers, for all $1 \leq i \leq m$, and moreover, since $r \geq u^*$,

$$\bigcup_{i=1}^{m} P'_i = \bigcup_{i=1}^{m} P_i = \mathbb{N}.$$

The difference $r - s$ can thus be written as

$$r - s = \sum_{i=1}^{m} (p_i q_i^* - \sum_{n \in P'_i} u^n).$$

Note that $\sum_{n \in P'_i} u^n$ is a semilinear series, as P'_i is a semilinear set of nonnegative integers, i.e., $\sum_{n \in P'_i} u^n = u^{l_{i1}} (u^{d_{i1}})^* + \ldots + u^{l_{ik_i}} (u^{d_{ik_i}})^* + p$, for some $l_{ij}, d_{ij} \in \mathbb{N}$, and a polynomial p. It is enough to prove the result for $p = 0$: one can decrease the polynomial p in the last step of the computation, using the decomposition given by Lemma 3.8.3.

Observe now that $p_i q_i^* \geq \sum_{n \in P'_i} u^n$ since $P'_i \subseteq P_i$. Thus, we reduced the problem to computing several differences of the form

$$pq^* - (u^{l_1} (u^{d_1})^* + \ldots + u^{l_k} (u^{d_k})^*).$$

Let $r_i = pq^* - (u^{l_1}(u^{d_1})^* + \ldots + u^{l_i}(u^{d_i})^*)$, for all $1 \leq i \leq k$.

By Lemma 3.8.8, $pq^* - u^{l_1}(u^{d_1})^* = p_1 + p_2q^* + p_3q^*(u^{t_1})^*$, for some polynomials p_1, p_2, p_3 and a nonnegative integer t_1, such that $q^* \geq (u^{t_1})^*$.

Since $r_1 \geq (u^{l_2}(u^{d_2})^* + \ldots + u^{l_k}(u^{d_k})^*)$, we have by Lemma 3.8.9 that also $p_1 + p_2q^* + p_3q^* \geq (u^{l_2}(u^{d_2})^* + \ldots + u^{l_k}(u^{d_k})^*)$. Moreover,

$$r_1 = p_1 + (p_2 + p_3)q^* + p_3u^{t_1}q^*(u^{t_1})^*,$$

and

$$r_2 = r_1 - u^{l_2}(u^{d_2})^* = (p_1 + (p_2 + p_3)q^* - u^{l_2}(u^{d_2})^*) + p_3u^{t_1}q^*(u^{t_1})^*.$$

Iterating the procedure, we obtain that indeed $r - s$ is a rational series of the form specified in the lemma. □

Using the above partial results, we can now solve the general problem.

Theorem 3.8.11. *Let r and s be semilinear series. If s has bounded coefficients and $r \geq s$, then $r - s$ is a rational series of star-height one.*

Proof. If s has bounded coefficients, then by Theorem 3.8.2, s is of the form

$$s = v_1u_1^* + \ldots + v_nu_n^* + p,$$

where $u_i, v_i \in \Sigma^{\oplus}$, for all $1 \leq i \leq m$, and p is a polynomial. we can assume without loss of generality that $p = 0$; indeed, for any rational series t of star-height one, $t - p$ is also a rational series of star-height one.

We prove the claim of the theorem by induction on n.

If $n = 1$, then $r - s$ is rational by Lemma 3.8.10. Assume now that the theorem holds for series s with up to $n - 1$ linear components, for some $n > 1$, and let us prove it for n. By Lemma 3.8.10, $r_1 = r - v_1u_1^*$ is of the form

$$r_1 = p_0 + \sum_{i=1}^{m} p_i(q_i)^* + \sum_{i=m+1}^{m+n} p_i(q_i)^*(u^{d_i})^*,$$

for some $m, n, d_{m+1}, \ldots, d_{m+n} \in \mathbb{N}$ and some polynomials p_0, p_i, q_i, $1 \leq i \leq m + n$, such that $q_i^* \geq (u^{d_i})^*$, for all $m + 1 \leq i \leq m + n$.

Let $r_1' = p_0 + \sum_{i=1}^{m+n} p_iq_i^*$. Since $r_1 \geq v_2u_2^* + \ldots + v_nu_n^*$, it follows by Lemma 3.8.9 that $r_1' \geq v_2u_2^* + \ldots + v_nu_n^*$. Thus, by the induction hypothesis, it follows that $r_1' - (v_2u_2^* + \ldots + v_nu_n^*)$ is a rational series of star-height one. Consequently,

$$\begin{aligned} r - s &= r_1 - (v_2u_2^* + \ldots + v_nu_n^*) = \\ &= r_1' - (v_2u_2^* + \ldots + v_nu_n^*) + \sum_{i=m+1}^{m+n} p_iu_1^{d_i}q_i^*(u_1^{d_i})^* \end{aligned}$$

is also rational of star-height one, proving the claim of the theorem. □

Example 3.8.2. Let $r = (a+b)^*$ and $s = a^* + bb^*$. Observe that both r and s are semilinear series, s has bounded coefficients, and $r \geq s$. Then, according to Theorem 3.8.11, the difference $r - s$ is computed as follows.

Let $s_1 = a^*$. We first compute $r - s_1$:

$$\begin{aligned} r - s_1 = (a+b)^* - a^* = 1 + a(a+b)^* + b(a+b)^* - 1 - aa^* = \\ = a\left((a+b)^* - a^*\right) + b(a+b)^* = b(a+b)^*a^*. \end{aligned}$$

The difference $r - s$ is then computed as follows:

$$\begin{aligned} r - s = b\left((a+b)^*a^* - b^*\right) = b[((a+b)^* - b^*) + a(a+b)^*a^*] = \\ = b[a(a+b)^*b^* + a(a+b)^*a^*] = ab(a+b)^*b^* + ab(a+b)^*a^*. \end{aligned}$$

3.9 Decompositions of semilinear series

In this section, we take a closer look at the family of semilinear power series with bounded coefficients. We prove that this family has strong closure properties, similar to the family of semilinear subsets of $\mathbb{N}^n$. We thus prove that the bounded multiplicities do not affect essentially the closure properties of the semilinear sets.

Any semilinear subset of $\mathbb{N}^n$ can be written as a finite union of linear sets, each of which has linearly independent periods, see [24] for details. Moreover, they are closed under the operations of complementation, intersection, and difference. While these properties are proved or believed to be false in general for semilinear power series, we prove that they hold for semilinear power series with bounded coefficients. These series can be decomposed in finite sums of linear series with disjoint supports, are closed under Hadamard product and difference. We prove these results in Section 3.9.

Theorem 3.9.1. *Any semilinear series with bounded coefficients can be written as a sum of linear series with disjoint supports.*

Proof. Consider first the decomposition of a "simple" semilinear series of the form $r + r'$, with r and r' linear series. Let $r = pw^*$, $r' = p'w'^*$, for $w, w' \in \Sigma^{\oplus}$ and p, p' monomials. Slightly more generally, we can assume that p and p' are in fact polynomials such that for any two monomials μ_1 and μ_2 of p, $\mu_1 w^*$ and $\mu_2 w^*$ have disjoint supports, and for any two monomials μ_1' and μ_2' of p', $\mu_1' w'^*$ and $\mu_2' w'^*$ have disjoint supports.

If the intersection of $\mathrm{supp}(r)$ and $\sup(r')$ is finite (or empty), then using Lemma 3.8.3, we can decompose r and r' as

$$r = r_1 + q, \qquad r' = r_1' + q',$$

for some polynomials q, q' and some semilinear series r_1, r_1' with disjoint supports. Thus, $r + r' = (q + q') + r_1 + r_1'$, and $(q + q')$, r_1, and r_1' have disjoint supports.

Assume now that the intersection of $\mathrm{supp}(r)$ and $\sup(r')$ is an infinite set of words. For the sake of simplicity of notations, we will assume that we have a two letter alphabet $\Sigma = \{a, b\}$. The general case can be treated in a similar way; we refer to the proof of Theorem 3.5.5, where a similar problem is treated in the general case.

Let

$$p = m_1 u_1^{l_1} + \ldots + m_k u_k^{l_k} \qquad \text{and} \qquad p' = m'_1 v_1^{l'_1} + \ldots + m'_{k'} v_{k'}^{l'_{k'}},$$

for $k, k', l_i, l'_j, m_i, m'_j \in \mathbb{N}$ and $u_i, v_j \in \Sigma^{\oplus}$, for all $1 \leq i \leq k$, $1 \leq j \leq k'$. Let also

$$u_i = a^{t_{i1}} b^{t_{i2}}, \quad v_j = a^{t'_{j1}} b^{t'_{j2}}, \quad w = a^{s_1} b^{s_2}, \quad w' = a^{s'_1} b^{s'_2},$$

with $s_1, s_2, s'_1, s'_2, t_{i1}, t_{i2}, t'_{j1}, t'_{j2} \in \mathbb{N}$, for all $1 \leq i \leq k$, $1 \leq j \leq k'$.

To compute the intersection of $\mathrm{supp}(r)$ and $\mathrm{supp}(r')$, one has to find the pairs $(m, n) \in \mathbb{N}^2$ such that $u_i^{l_i} w^m = v_j^{l'_j} w'^n$, for some $1 \leq i \leq k$, $1 \leq j \leq k'$. Thus, we need to solve the following Diophantine linear systems of two equations in $(m, n) \in \mathbb{N}^2$:

$$\begin{cases} t_{i1} + m \cdot s_1 = t'_{j1} + n \cdot s'_1, \\ t_{i2} + m \cdot s_2 = t'_{j2} + n \cdot s'_2, \end{cases} \tag{3.11}$$

for all $1 \leq i \leq k$, $1 \leq j \leq k'$. As it is well-known, see, e.g., [24], the solution of this system is either the empty set, or it is a union of linear subsets of $\mathbb{N}^2$ of the form

$$\{(l_{ij}^{(1)} + l_{ij}^{(2)} p, l_{ij}^{(3)} + l_{ij}^{(4)} p) \mid p \in \mathbb{N}\}. \tag{3.12}$$

It is important to observe now that only $l_{ij}^{(1)}$ and $l_{ij}^{(3)}$ depend on i and j, while $l_{ij}^{(2)}$ and $l_{ij}^{(4)}$ depend only on s_1, s_2, s'_1 and s'_2. Thus, $l_{ij}^{(2)} = \rho_2$, $l_{ij}^{(4)} = \rho_4$, for some $\rho_2, \rho_4 \in \mathbb{N}$.

Using the fact that for any $t \in \Sigma^{\oplus}$ and any $k \in \mathbb{N}$,

$$t^* = (1 + t + \ldots + t^{k-1})(t^k)^*,$$

it follows that

$$r = pw^* = (m_1 u_1^{l_1} + \ldots + m_k u_k^{l_k})(1 + w + \ldots + w^{\rho_2 - 1})(w^{\rho_2})^*,$$

and

$$r' = p'w'^* = (m'_1 v_1^{l'_1} + \ldots + m'_{k'} v_{k'}^{l'_{k'}})(1 + w' + \ldots + w'^{\rho_4 - 1})(w'^{\rho_4})^*.$$

Having the solutions (3.12) of the systems (3.11), it is easy now to isolate $\mathrm{supp}(r) \cap \mathrm{supp}(r')$ in both r and r'. Indeed, r and r' can be written as

$$r = p_1 (w^{\rho_2})^* + p_2 (w^{\rho_2})^*, \quad r' = p_3 (w'^{\rho_4})^* + p_4 (w'^{\rho_4})^*,$$

for some polynomials p_1, p_2, p_3, and p_4, such that $p_1(w^{\rho_2})^*, p_2(w^{\rho_2})^*$, and $p_3(w'^{\rho_4})^*$ have disjoint supports, while $p_2(w^{\rho_2})^*$ and $p_4(w'^{\rho_4})^*$ have the same support; in fact, by (3.11), $p_4(w'^{\rho_4})^* = p_5(w^{\rho_2})^*$, for a polynomial p_5. Thus,

$$r + r' = (p_1 + p_2 + p_5)(w^{\rho_2})^* + p_3(w'^{\rho_4})^*,$$

and $(p_1 + p_2 + p_5)(w^{\rho_2})^*$ and $p_3(w'^{\rho_4})^*$ have disjoint supports.

Consider now an arbitrary semilinear series r with bounded coefficients,

$$r = p_1 w_1^* + p_2 w_2^* + \ldots + p_n w_n^*,$$

where p_i is polynomial and $w_i \in \Sigma^{\oplus}$, for all $1 \leq i \leq n$. Let $r_i = p_1 w_1^* + \ldots + p_i w_i^*$, for all $2 \leq i \leq n$, and assume that for some i, we have decomposed r_i into a sum of linear series with disjoint supports,

$$r_i = q_1 u_1^* + \ldots + q_{k_i} u_{k_i}^* + q_0.$$

We add $p_{i+1} w_{i+1}^*$ to r_i as follows. As proved above,

$$q_1 u_1^* + p_{i+1} w_{i+1}^* = q_0' + q_1' u_1'^* + p_{i+1}' w_{i+1}'^*,$$

for some polynomials q_0', q_1', p_{i+1}' and words $u_1', w_{i+1}' \in \Sigma^{\oplus}$, such that q_0', $q_1' u_1'^*$, and $p_{i+1}' w_{i+1}'^*$ have disjoint supports, and moreover,

$$\mathrm{supp}(q_1' u_1'^*) \subseteq \mathrm{supp}(q_1 u_1^*), \quad \mathrm{supp}(p_{i+1}' w_{i+1}'^*) \subseteq \mathrm{supp}(p_{i+1} w_{i+1}^*).$$

We then add $p_{i+1}' w_{i+1}'^*$ to $q_2 u_2^*$:

$$q_2 u_2^* + p_{i+1}' w_{i+1}'^* = q_0'' + q_2' u_2'^* + p_{i+1}'' w_{i+1}''^*,$$

for some polynomials q_0'', q_2', p_{i+1}'', and words $u_2', w_{i+1}'' \in \Sigma^{\oplus}$, such that q_0'', $q_2' u_2'^*$, and $p_{i+1}'' w_{i+1}''^*$ have disjoint supports, and moreover,

$$\mathrm{supp}(q_2' u_2'^*) \subseteq \mathrm{supp}(q_2 u_2^*) \quad \text{and} \quad \mathrm{supp}(p_{i+1}'' w_{i+1}''^*) \subseteq \mathrm{supp}(p_{i+1}' w_{i+1}'^*).$$

Consequently, $q_{i+1}'' w_{i+1}''^*$ and $q_1' u_1'^*$ also have disjoint supports. We iterate the procedure up to n, concluding the claim. □

Example 3.9.1. Let $r_1 = a(a^2)^*$, $r_2 = a^3(a^4)^*$, and $r_3 = a^2(a^5)^*$, The semilinear series $r = r_1 + r_2 + r_3$ can be written as a sum of linear series with disjoint supports as follows. Let $s = r_1 + r_2$. To compute $\mathrm{supp}(r_1) \cap \mathrm{supp}(r_2)$, we consider the following equation in the nonnegative integers m and n:

$$2m + 1 = 4n + 3,$$

having the solution $\{(2p + 1, p) \mid p \in \mathbb{N}\}$. Thus,

$$r_1 + r_2 = a(1 + a^2)(a^4)^* + a^3(a^4)^* = a(a^4)^* + 2a^3(a^4)^*.$$

To add r_3 to s, we solve the following two equations in the nonnegative integers m and n:

$$4m + 1 = 5n + 2 \quad \text{and} \quad 4m + 3 = 5n + 2.$$

Their solutions are the sets $\{(5p + 4, 4p + 3) \mid p \in \mathbb{N}\}$ and $\{(5p + 1, 4p + 1) \mid p \in \mathbb{N}\}$, respectively. Thus,

$$\begin{aligned} r =& (a + 2a^3)(a^4)^* + a^2(a^5)^* \\ =& (a + 2a^3)(1 + a^4 + a^8 + a^{12} + a^{16})(a^{20})^* + a^2(1 + a^5 + a^{10} + a^{15})(a^{20})^* \\ =& (a + a^2 + 2a^3 + a^5 + 3a^7 + a^9 + 2a^{11} + a^{12} + a^{13} + 2a^{15} + 2a^{17})(a^{20})^*, \end{aligned}$$

a decomposition into a sum of linear series with disjoint supports.

In the proof of Theorem 3.9.1, we have proved essentially that for two linear power series with bounded coefficients, the intersection of their supports is semilinear. Using this argument, one can also conclude the closure under Hadamard product.

Theorem 3.9.2. *The family of $\mathbb{N}$-semilinear power series with bounded coefficients is closed under Hadamard product.*

Furthermore, it is easy to prove the closure under difference using the techniques in the proof of Theorem 3.8.11. Indeed, in the proof of Theorem 3.8.11, we reduce the problem to a difference of the form $r - v^*$. However, in our case, r has bounded coefficients and so, $r = pu^*$, for a polynomial p and $u \in \Sigma^{\oplus}$. Since $r \geq v^*$, there must be a minimal positive integer k such that $(u^*, v^k) > 0$. Moreover, p must contain the monomials $1, v, \ldots, v^{k-1}$. Thus, we reduce the problem to computing the difference

$$(1 + v + \ldots + v^{k-1})u^* - v^*.$$

Since $v^* = (1 + v + \ldots + v^{k-1})(v^k)^*$, we further reduce the problem to computing the difference $u^* - (v^k)^*$, where $u^i = v^k$, for some positive integer i. Thus,

$$\begin{aligned} u^* - (v^k)^* &= (1 + u + \ldots + u^{i-1}) + u^i u^* - 1 - v^k(v^k)^* = \\ &= (u + \ldots + u^{i-1}) + v^k(u^* - (v^k)^*) = \\ &= (u + \ldots + u^{i-1})(v^k)^*. \end{aligned}$$

We skip the other details here, as they are completely similar to the details of Theorem 3.8.11 and its auxiliary results proved in Section 3.8 of this chapter.

Theorem 3.9.3. *The family of $\mathbb{N}$-semilinear power series with bounded coefficients is closed under difference.*

Example 3.9.2. Let $r_1 = (a^2)^*$, $r_2 = a^3(a^6)^*$, $r = r_1 + r_2$, and $s = (a^3)^*$. To compute $r - s$, let $P_i = \{n \in \mathbb{N} \mid (r_i, a^n) \geq (s, (a^3)^n)\}$, with $i = 1, 2$, i.e., $P_1 = \{2n \mid n \in \mathbb{N}\}$ and $P_2 = \{2n + 1 \mid n \in \mathbb{N}\}$. Since $P_1 \cup P_2 = \mathbb{N}$ and $P_1 \cap P_2 = \emptyset$, it follows that

$$s = \sum_{n \geq 0} (a^3)^n = \sum_{n \in P_1} a^{3n} + \sum_{n \in P_2} a^{3n} = (a^6)^* + a^3(a^6)^*.$$

Let $s_1 = (a^6)^*$ and $s_2 = a^3(a^6)^* = r_2$.

Clearly, $r_1 \geq (a^6)^*$. In fact, $r_1 = (1 + a^2 + a^4)(a^6)^*$. Thus,

$$r - s = r_1 - s_1 = (a^2 + a^4)(a^6)^*.$$

Example 3.9.3. Let $r_1 = (a^2)^*$, $r_2 = a(a^4)^*$, $r_3 = a^3(a^4)^*$, $r = r_1 + r_2 + r_3$, and $s = (a^3)^*$. To compute $r - s$, denote as above $P_i = \{n \in \mathbb{N} \mid (r_i, a^n) \geq (s, (a^3)^n)\}$, for all $1 \leq i \leq 3$. As it is easy to see, it follows that $P_1 = \{2n \mid n \in \mathbb{N}\}$, $P_2 = \{4n + 3 \mid n \in \mathbb{N}\}$, and $P_3 = \{4n + 1 \mid n \in \mathbb{N}\}$. Thus, $P_1 \cup P_2 \cup P_3 = \mathbb{N}$ is a disjoint union, and so,

$$\begin{aligned} s = (a^3)^* &= \sum_{n \geq 0} (a^3)^n = \sum_{n \in P_1} a^{3n} + \sum_{n \in P_2} a^{3n} + \sum_{n \in P_3} a^{3n} = \\ &= (a^6)^* + a^9(a^{12})^* + a^3(a^{12})^*. \end{aligned}$$

Note that $r_1 \geq (a^6)^*$ and

$$r_1 - (a^6)^* = (1 + a^2 + a^4)(a^6)^* - (a^6)^* = (a^2 + a^4)(a^6)^*.$$

Similarly, $r_2 \geq a^9(a^{12})^*$ and

$$r_2 - a^9(a^{12})^* = a(1 + a^4 + a^8)(a^{12})^* - a^9(a^{12})^* = (a + a^5)(a^{12})^*.$$

Also, $r_3 \geq a^3(a^{12})^*$ and

$$r_3 - a^3(a^{12})^* = a^3(1 + a^4 + a^8)(a^{12})^* - a^3(a^{12})^* = (a^7 + a^{11})(a^{12})^*.$$

Thus,

$$r - s = (a^2 + a^4)(a^6)^* + (a + a^5 + a^7 + a^{11})(a^{12})^*.$$

3.10 Discussion

Starting from the notion of semilinearity, we have investigated in this chapter some aspects of the formal power series in commuting variables, with special stress on the series with coefficients in the semiring of nonnegative integers. We have clarified some basic properties of the recognizable, rational, and algebraic power series in commuting variables, otherwise only implicitly found in literature. We extended in two different ways the notion

of semilinearity to formal power series, introducing the families of semilinear and bounded power series; both these families are natural generalizations of the notion of semilinearity, as originated in [22]. We investigated the closure properties of these families of series under some elementary operations, comparing their behaviour with the behaviour of the recognizable, rational, and algebraic power series in commuting variables, as well as with the behaviour of the corresponding families of series in noncommuting variables and the corresponding families of formal languages. Thus, it turned out in several instances that classical results on semilinearity cannot be extended to formal power series, or at least, they cannot be extended to both semilinear and bounded power series. This gives a new insight on semilinearity.

There remain several interesting open problems concerning the semilinear formal power series. We describe some of them in the following.

Problem 3.10.1. Study the closure properties of the semilinear power series under inverse morphisms.

Problem 3.10.2. We have proved in Theorem 3.5.4 that there are semilinear series in two variables such that their Hadamard product is not even rational. On the other hand the Hadamard product of any two semilinear series in one variable is rational. Is it true that the Hadamard product of a *semilinear* and a *rational* power series in one variable is always a *rational* power series? Also, is it true that the Hadamard product of a *bounded* and a *rational* power series is always a *rational* power series?

Problem 3.10.3. Study the closure properties of the bounded power series under difference. Also, prove that the semilinear power series are not closed under difference.

Problem 3.10.4. It is easy to see that the rational series with bounded coefficients are in fact bounded power series. Establish whether or not the strong closure properties of the semilinear power series with bounded coefficients can be extended to rational power series with bounded coefficients.

Problem 3.10.5. It is known in the case of formal languages that the star-height is not bounded but it is decidable, see [26] and also [58]. On the other hand, no similar result is known for formal power series in noncommuting variables. However, it is known that any rational power series in one variable is of star-height at most two. Study if a similar result holds for formal power series in at least two *commuting* variables. In other words how many of the following inclusions are strict:

$$\mathbb{N}_0^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle \subseteq \mathbb{N}_1^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle \subseteq \mathbb{N}_2^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle \subseteq \ldots \subseteq \mathbb{N}^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle,$$

where $\mathbb{N}_i^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle$ is the set of the $\mathbb{N}$-rational series over $\Sigma^{\oplus}$, having star-height i ? We only know that the first two inclusions are strict for any alphabet Σ and that these are the only strict ones for $|\Sigma| = 1$.

Chapter 4

The notion of centralizer and Conway's Problem

The commutation $XY = YX$ is one of the most fundamental equations in any algebraic structure. Its solution is well-known in the case of free monoids: two words commute if and only if they are powers of a same word. On the other hand, almost nothing is known on solutions of language equations, except for some very special equations with two operations characterizing rational languages, see [21] and also [46] for some extensions. Nevertheless, there are several natural and apparently very difficult combinatorial problems arising from this area.

It was more than thirty years ago when Conway proposed such a problem, asking whether the largest set commuting with a given rational set is rational ([18]). The problem remained unanswered up-to-date; even worse, it is not known even if the centralizer of any rational set is recursively enumerable. Moreover, these questions are also unanswered for finite sets !

We discuss in this chapter the notion of centralizer of a set of words and prove that Conway's Problem has an affirmative answer for all sets of cardinal at most three, for periodic sets, as well as for finite biprefix sets. We return to Conway's Problem in Chapter 6. Using different techniques and some results on the commutation of two sets of words and the primitive roots of codes, we answer it affirmatively also in the case of rational ω-codes and rational reversed ω-codes.

The results of this chapter are based on our papers [37], [38], [39], and [40].

4.1 The notion of centralizer

As it is well-known, two words commute if and only if they are powers of a same word. Formally, $uv = vu$ if and only if $\rho(u) = \rho(v)$, i.e., $u = \rho^m$ and $v = \rho^n$, for some word ρ and some nonnegative integers m, n, where $\rho(x)$

denotes the primitive root of the word x. Thus, the largest set of words commuting with u is $\rho(u)^*$, where $\rho(u)$ denotes the primitive root of u.

We consider in Chapters 4 and 6 the corresponding problems for sets of words. For an alphabet Σ, let 2^{Σ^*} be the set of all languages over Σ. Throughout these chapters, we denote the union of two languages $L_1, L_2 \in 2^{\Sigma^*}$ by $L_1 + L_2$ and their concatenation by L_1L_2.

For a given language L, consider the set of all languages commuting with L,

$$\mathcal{COM}(L) = \{S \subseteq \Sigma^* \mid LS = SL\}.$$

It is easy to see that for any L, $\mathcal{COM}(L)$ is closed under product. Indeed, for any sets $L_1, L_2 \in \mathcal{COM}(L)$, we have $L_1L_2L = L_1LL_2 = LL_1L_2$ and so, $L_1L_2 \in \mathcal{COM}(L)$. Moreover, $\mathcal{COM}(L)$ is also closed under union. Indeed, for any sets $L_1, L_2 \in \mathcal{COM}(L)$, $(L_1 + L_2)S = L_1S + L_2S = SL_1 + SL_2 = S(L_1 + L_2)$. Consequently, $(\mathcal{COM}(L), +, \cdot)$ is a semiring with the empty set $\emptyset$ as its zero, and $\{1\}$ as its one.

Theorem 4.1.1. *For any sets of words L, $\mathcal{COM}(L)$ has a unique maximal element, with respect to inclusion. Moreover, this unique maximal element of $\mathcal{COM}(L)$ is the union of all sets in $\mathcal{COM}(L)$.*

Proof. Let

$$\mathcal{C}(L) = \cup_{S \in \mathcal{COM}(L)} S.$$

Then $\mathcal{C}(L)$ is a maximal element of $\mathcal{COM}(L)$. Moreover, for any two maximal elements L_1, L_2 of $\mathcal{COM}(L)$, since $L_1 + L_2 \in \mathcal{COM}(L)$, we obtain $L_1 = L_1 + L_2 = L_2$, and so, the uniqueness of the maximal element follows. □

Definition 4. *For a language L, its centralizer $\mathcal{C}(L)$ is the maximal set - with respect to inclusion - commuting with L.*

Example 4.1.1. Let $\Sigma = \{a, b\}$ be the alphabet.

(i) If $L_1 = \{a, ab\}$, then $\mathcal{C}(L_1) = \{a, ab\}^*$.

(ii) If $L_2 = \{aa, ab, ba, bb\}$, then $\mathcal{C}(L_2) = \{a, b\}^*$.

(iii) If $L_3 = \{aa, baa, bb, aab\}$, then $\mathcal{C}(L_3) = \{aa, b\}^*$. Also, in this case, we have the following strict inclusions: $L_3^* \subset \mathcal{C}(L_3) \subset \Sigma^*$.

Note that the notion of centralizer of a set is usually defined in *Algebra* with respect to element-wise commutation, see, e.g., [5], [9], [16], [23], [32]. Thus, for a *group* (alternatively, for a semigroup, a ring, or a semiring) R and a subset S of R, the *centralizer* of S in R is the set $\{x \in R \mid xs = sx, \forall s \in S\}$. In this respect, the centralizer of a word u in Σ^* is $\rho(u)^*$, for any $u \in \Sigma^+$. Also, by this definition, the centralizer of a language L in the semiring 2^{Σ^*} is the set of all languages commuting with L, i.e., the set we denoted

as $\mathcal{COM}(L)$. We investigate in Chapter 6 the problem of characterizing $\mathcal{COM}(L)$ for a given L.

As a matter of fact, the term normalizer is also used in the literature, see, e.g., [16]. For a *group* G and a subset H of G, the *normalizer* of H in G is the subset $\mathcal{N}_G(H) = \{g \in G \mid gH = Hg\}$. Thus, for a subgroup H of G, $\mathcal{N}_G(H)$ is the largest subgroup of G such that H is a normal subgroup of $\mathcal{N}_G(H)$. For singletons (e.g., for an element $L \in 2^{\Sigma^*}$, or in other words, for a set of words L), the notions of centralizer and normalizer as defined above coincide. We refer for more details to [9] and [23].

We consider in this thesis the centralizer of a set of words L defined as the largest set commuting with L.

For any set of words L, its centralizer $\mathcal{C}(L)$ is a monoid with respect to catenation.

Lemma 4.1.2. *For any $L \subseteq \Sigma^*$, the centralizer $\mathcal{C}(L)$ is a monoid.*

Proof. As we noted above, $\mathcal{COM}(L)$ is closed under product and so, in particular, $\mathcal{C}(L)\mathcal{C}(L) \in \mathcal{COM}(L)$. Thus, it follows by Theorem 4.1.1 that $\mathcal{C}(L)\mathcal{C}(L) \subseteq \mathcal{C}(L)$ and so, for any $x, y \in \mathcal{C}(L)$, $xy \in \mathcal{C}(L)$. Also, by definition, the empty word 1 is in $\mathcal{C}(L)$, for any L. □

As a matter of fact, one can define similarly as above a notion of a semigroup centralizer of a set L. Indeed, observe that for any set of words L, there is a unique maximal element of the set $\{S \subseteq \Sigma^+ \mid LS = SL\}$. We call this element the *semigroup centralizer* of L and we denote it as $\mathcal{C}_s(L)$. We refer to [13] for details.

Clearly, for any L, we have $\mathcal{C}_s(L) \cup \{1\} \subseteq \mathcal{C}(L)$. However, this is a strict inclusion in general: it is not true that for any L, $\mathcal{C}_s(L)$ and $\mathcal{C}(L)$ coincide modulo the empty word. The reason for this is that $\mathcal{C}(L) \setminus \{1\}$ does not commute with L for all sets of words L.

Example 4.1.2. (i) For $L = \{a, ab, ba, bb\}$,

$$\mathcal{C}(L) = \{a, b\}^* \quad \text{and} \quad \mathcal{C}_s(L) = \{a, b\}^+ \setminus \{b\}.$$

Note that $\mathcal{C}(L) = \mathcal{C}_s(L) \cup \{1, b\}$.

(ii) For $\Sigma = \{a, b\}$ and $L = \Sigma^* b \Sigma^*$, $\mathcal{C}(L) = \Sigma^*$ and $\mathcal{C}_s(L) = L^+$.

It is in fact an interesting open problem whether or not for any language (or at least for any rational language), the difference between its monoid and its semigroup centralizers is always a rational set. The relevance of this question comes from Conway's Problem, formulated bellow: it is not known whether or not Conway's Problem for monoid centralizers can be reduced to the same problem for semigroup centralizers and the other way around.

In this thesis, unless explicitly stated otherwise, we always consider monoid centralizers.

The problem of computing or studying the properties of the centralizer of a language L over the alphabet Σ is trivial if $1 \in L$. Indeed, in this case $\mathcal{C}(L) = \Sigma^*$, since $\mathcal{C}(L)\Sigma^* = \Sigma^*\mathcal{C}(L) = \Sigma^*$. Thus, we will always consider in the sequel languages L such that $1 \notin L$.

The best known problem concerning the centralizer of a language is the question raised by Conway more than thirty years ago, [18]. His question, still unanswered today, is whether or not the centralizer of any rational language is rational.

Conway's Problem [18]: *Is the centralizer of any rational language, rational as well ?*

Thirty years later, this problem is far from being answered. In fact, it is not even known whether or not the centralizer of a rational language is always recursive or recursively enumerable. Attempts to simplify Conway's problem have been made replacing the rational language by a finite language and replacing Conway's question by "Is the centralizer recursive or even recursively enumerable ?". The problem remains open even in this weaker form.

For context-free languages however, we know that the commutation of languages is a very difficult problem, as shown by the following result.

Theorem 4.1.3 ([27]). *Let C be a fixed two element code. It is undecidable for context-free languages L whether or not $CL = LC$.*

Several very different approaches have been taken to attack Conway's Problem: *combinatorial properties* of finite and infinite words ([13], [53], [66]), *equations on languages* ([37], [38]), *algebraic results* on the commutation of formal power series ([29], the *fixed point approach* ([18], [36]), and the *branching point approach* ([40]). We discuss in this chapter some of these approaches and their outcome up to date. We prove that Conway's Problem has an affirmative answer for periodic, binary, and ternary sets, as well as for finite biprefix sets.

4.2 A general property of the centralizer

In general, very little is known about the centralizer of a set of words. Even much weaker questions than Conway's question seem to be unanswered, namely it is not known whether the centralizer of a rational language is recursive or even recursively enumerable. What we can prove is that for any rational, and in fact, for any recursive language, the complement of its centralizer is always recursively enumerable.

Theorem 4.2.1. *For any recursive set, the complement of its centralizer is a recursively enumerable language.*

Proof. Let L be a recursive language and let $\mathcal{C}(L)$ be its centralizer. Our claim is that there is an algorithm such that given an input word x, the computation stops if and only if $x \notin \mathcal{C}(L)$.

Since $\mathcal{C}(L)$ is the maximal set commuting with L, an element y is not in $\mathcal{C}(L)$ if and only if there is a word $u \in L$ such that either one of the following conditions is satisfied:

(i) For all $v \in L$, if $yu = vz$, for some $z \in \Sigma^*$, then $z \notin \mathcal{C}(L)$.

(ii) For all $v \in L$, if $uy = zv$, for some $z \in \Sigma^*$, then $z \notin \mathcal{C}(L)$.

We set $\mathcal{L}_1 = \{x\}$ and in the n-th step of the algorithm, we test the words from $\mathcal{L}_n$ for their membership to $\mathcal{C}(L)$, in the following way: for each word $z \in \mathcal{L}_n$, we choose nondeterministically a word $u \in L$ (this is possible since L is recursive) and one of the conditions (i) or (ii) to be checked. Assuming that we chose (i), we consider the word zu, and for all words $v \in L$, such that there is a word z' with $zu = vz'$, we add z' to the set $\mathcal{L}_{n+1}$. If we chose (ii), then we are looking for words z' such that $uz = z'v$.

It is important to observe here that if none of the words in $\mathcal{L}_{n+1}$ is from $\mathcal{C}(L)$, then the same is true also for the words of $\mathcal{L}_n$, for any $n \geq 1$. Indeed, if we had a $z \in \mathcal{L}_n \cap \mathcal{C}(L)$, then for all $u \in \mathcal{C}(L)$ we would have $zu = v_1y_1$, and $uz = y_2v_2$, for some words $v_1, v_2 \in L$, and $y_1, y_2 \in \mathcal{C}(L)$, which implies that some words from $\mathcal{C}(L)$ should be in $\mathcal{L}_{n+1}$ as well.

If the list $\mathcal{L}_{n+1}$ remains empty then the algorithm stops: the initial word x is not in $\mathcal{C}(L)$. Otherwise we repeat the procedure with $\mathcal{L}_{n+1}$ instead of $\mathcal{L}_n$.

It is easy to conclude from the above that all the words for which there is a halting computation, are from the complement of $\mathcal{C}(L)$. For the reverse inclusion, let x be a word from the complement of $\mathcal{C}(L)$, and assume that our algorithm does not have any halting computation on the input x. Our claim is that there is $Z \supseteq \mathcal{C}(L) \cup \{x\}$ such that $ZL = LZ$. To begin with, let $Z = \mathcal{C}(L) \cup \{x\}$. If our algorithm does not halt on the input x, then for any $u \in L$, there are two words $v_1, v_2 \in L$ such that $xu = v_1y_1$ and $ux = y_2v_2$ and moreover, the algorithm does not halt on any of the input words y_1 and y_2 (indeed, if there is $u \in L$ such that the algorithm has halting computations for all y_1, y_2 as above, then it has a halting computation also for x: we just choose u in the first step of the algorithm). We add to Z the words y_1 and y_2 and we continue the same reasoning with y_1 and y_2 instead of x. The language Z obtained in this way clearly commutes with L. But $\mathcal{C}(L)$ is the maximal set commuting with L and so, $Z \subseteq \mathcal{C}(L)$. In particular, we obtain $x \in \mathcal{C}(L)$, which is a contradiction. □

The same result can be obtained also by defining $\mathcal{C}(L)$ as follows. Let

$X_0 = \mathrm{Pref}(L^*) \cap \mathrm{Suf}(L^*)$ and

$$X_{n+1} = X_n \setminus \left((L^{-1}(LX_n \triangle X_n L) \ \cup \ (LX_n \triangle X_n L)L^{-1})\right),$$

for all $n \geq 0$, where $\triangle$ denotes the symmetric difference of two sets: $R \triangle S = (R \setminus S) \cup (S \setminus R)$. Then $\mathcal{C}(L) = \cap_{n \geq 0} X_n$. Note also that $\mathcal{C}(L)$ is the maximal fixed point of the following mapping:

$$\phi(Y) = Y \setminus (L^{-1}(LY \triangle YL) \cup (LY \triangle YL)L^{-1}).$$

We refer to [36] for more details on this "fixed point approach".

A related approach is described in [18], where the centralizer $\mathcal{C}(L)$ is defined as follows. For $X_0 = \Sigma^*$, let X_1 be the maximal subset of X_0 such that $X_1 L \subseteq LX_0$. Then let X_2 be the maximal subset of X_1 such that $LX_2 \subseteq X_1 L$, and so on. Then $\mathcal{C}(L) = \cap_{n \geq 0} X_n$. We refer to [18] for details, as well as for an interesting conjecture on the maximal solutions of systems of semilinear inequalities. Conway's Problem is a particular instance of such a system.

4.3 The centralizer of periodic and binary sets of words

As it is well-known, two words commute if and only if they have the same primitive root, or equivalently, if and only if they are powers of another word. Based on this, it is not difficult to prove, see [53], that a set of words X commutes with a word u if and only if $X \subseteq \rho(u^*)$, where $\rho(u)$ denotes the primitive root of u. Consequently, for any word u, the centralizer of $\{u\}$ is $\rho(u)^*$.

If instead of a singleton, we consider a set of words, all powers of a same word, then the situation is not much different than that of a singleton, as we prove in Theorem 4.3.2.

We say that a set of words L is *periodic* if there is a word u such that $L \subseteq u^+$.

We use often in our considerations the following simple result.

Lemma 4.3.1. *For any language L, $1 \notin L$, and any $x \in \mathcal{C}(L)$, $xL^\omega \subseteq L^\omega$.*

Proof. Let $\alpha_n \in L$, for all $n \geq 1$. Let $x_1 = x$. Then for any $n \geq 1$, there is $\beta_n \in L$ and $x_{n+1} \in \mathcal{C}(L)$ such that $x_n \alpha_n = \beta_n x_{n+1}$. Thus,

$$x\alpha_1\alpha_2 \ldots \alpha_n = \beta_1\beta_2 \ldots \beta_n x_{n+1}.$$

Consequently,

$$x\alpha_1\alpha_2 \ldots \alpha_n \ldots = \beta_1\beta_2 \ldots \beta_n \ldots.$$

Indeed, the two infinite words have arbitrarily long common prefixes, as $1 \notin L$, and so, they coincide. Thus, $xL^\omega \subseteq L^\omega$. □

Theorem 4.3.2. *Let u be a primitive word and $L \subseteq u^+$. Then $\mathcal{C}(L) = u^*$ and moreover, for any set of words X, if $LX = XL$, then $X = \cup_{i\in I} u^i$, for some $I \subseteq \mathbb{N}$.*

Proof. Since $L\mathcal{C}(L) = \mathcal{C}(L)L$, for any word $x \in \mathcal{C}(L)$ and any $\alpha \in L$, $x\alpha^\omega \in L^\omega$. Thus, $xu^\omega = u^\omega$, and since u is primitive, $x = u^n$, for some $n \geq 0$. Due to the maximality of the centralizer, it follows that $\mathcal{C}(L) = u^*$. The second part of the claim follows from the maximality of $\mathcal{C}(L)$: any set commuting with L is a subset of $\mathcal{C}(L)$. □

We say that a set of words X is *branching* if not all words of X start with the same letter. In other words, there are two words of X starting with different letters.

Theorem 4.3.2 has the following corollary, instrumental in our considerations.

Corollary 4.3.3. *For any nonperiodic set of words L, $1 \notin L$, there is a branching set of words L' such that $\mathcal{C}(L)$ is rational if and only if $\mathcal{C}(L')$ is rational. Moreover, $\mathcal{C}(L) = L^*$ if and only if $\mathcal{C}(L') = L'^*$.*

Proof. If L is branching, then the claim is trivially true, with $L' = L$. Thus, let us assume that $L = aL_1$, for some letter a and some set of words L_1. Then, as $\mathcal{C}(L)L = L\mathcal{C}(L)$, it follows that $\mathcal{C}(L) = 1 + aX$, for some set of words X, and so, $aL_1 + aXaL_1 = aL_1 + aL_1aX$. Thus, $(Xa + 1)L_1a = L_1a(Xa + 1)$, i.e., $1 + Xa \subseteq \mathcal{C}(L_1a)$. The other inclusion can be proved similarly, and so, $\mathcal{C}(aL_1) = 1 + a(\mathcal{C}(L_1a)a^{-1})$. Note that $\mathcal{C}(aL_1)$ is rational if and only if $\mathcal{C}(L_1a)$ is rational and moreover, $\mathcal{C}(aL_1) = (aL_1)^*$ if and only if $\mathcal{C}(L_1a) = (L_1a)^*$.

If L_1a is not branching, then we repeat the reasoning with L_1a instead of L. Since L is not periodic, we find in a finite number of steps a branching set L' such that $\mathcal{C}(L)$ is rational if and only if $\mathcal{C}(L')$ is rational. Moreover, $\mathcal{C}(L) = L^*$ if and only if $\mathcal{C}(L') = L'^*$. □

We say that a set of words F is *binary* (*ternary*, resp.) if it consists of two (three, resp.) words only.

Let F be a binary set of words, $1 \notin F$, $F = \{u, v\}$. If $uv = vu$, then there is a primitive word t such that $u = t^m$ and $v = t^n$, $m, n \geq 1$. The centralizer $\mathcal{C}(F)$ in this case is given by Theorem 4.3.2: $\mathcal{C}(F) = t^*$. We prove in the next theorem that if $uv \neq vu$, then $\mathcal{C}(F) = F^*$ and we also characterize the commutation with F. This result was originally proved in [13] but we give here an elementary, somewhat simpler proof, based on branching arguments. This argument will be refined in Section 4.5, leading to the branching point approach, see [40].

Theorem 4.3.4 ([13]). *The centralizer of any binary set F over the alphabet Σ is rational. Moreover,*

(i) If $1 \in F$, *then* $\mathcal{C}(F) = \Sigma^*$.

(ii) If F *is periodic,* $F \subseteq u^+$, *for some primitive word* u, *then* $\mathcal{C}(F) = u^*$.

(iii) If F *is not periodic and* $1 \notin F$, *then* $\mathcal{C}(F) = F^*$.

Proof. The first case is trivial and it holds more generally for any set of words F. Also, Case (ii) is concluded in Theorem 4.3.2. For Case (iii), note that by Corollary 4.3.3, we can assume without loss of generality that F is branching. Assume thus that $F = \{au, bv\}$ and consider a set X commuting with F. Let $x \in X \setminus \{1\}$. Since $FX = XF$, then for any $\alpha \in F$, there is $\beta_\alpha \in F$ and $y_\alpha \in X$ such that $x\alpha = \beta_\alpha y_\alpha$. Observe now that β_α is uniquely determined by the first letter of x. Thus, in fact, there is $\beta_1 \in F$ such that $xau = \beta_1 y_1$ and $xbv = \beta_1 y_2$, with $y_1, y_2 \in X$. Consequently, β_1 is a prefix of x, since otherwise β should have both a and b on the same position. It follows that $x = \beta_1 x_1$ and so, $y_1 = x_1 au$ and $y_2 = x_1 bv$. Note that $|x_1| < |x|$ since $1 \notin F$.

If $x_1 \neq 1$, then using a similar argument for y_1 and y_2, there is $\beta_2 \in F$ (uniquely determined by the first letter of x_1) such that $x_1 auau = \beta_2 z_1$ and $x_1 bvbv = \beta_2 z_2$, for some $z_1, z_2 \in X$. Again, β_2 must be a prefix of x_1: $x_1 = \beta_2 x_2$ and so, $z_1 = x_2 auau$ and $z_2 = x_2 bvbv$, where $|x_2| < |x_1| < |x|$, $x = \beta_1\beta_2 x_2$. Since it is impossible to build an infinite decreasing sequence of positive integers, it follows by iterating the argument, that $x \in F^n$, for some $n \geq 1$. Thus, $X \subseteq F^*$. Since $\mathcal{C}(F)$ is the maximal set commuting with F, it follows that $\mathcal{C}(F) = F^*$. □

4.4 A solution to Conway's Problem for ternary sets

The case of ternary sets turns out to be very difficult for Conway's Problem, a boundary point in many respects. Firstly, no elementary solution is known for Conway's Problem in this case. Using involved arguments of equations on languages, we prove in this section that indeed, the centralizer of any ternary set is a rational set. Secondly, unlike in the case of periodic and binary sets, the exact form of the centralizer is not known. It has been conjectured that similarly as for binary sets, for any nonperiodic ternary set $F \subseteq \Sigma^+$, it holds that $\mathcal{C}(F) = F^*$, but this appears to be a challenging problem. In the case of an affirmative answer, we prove in Chapter 6 that the commutation with a ternary set can be characterized in similar terms as for periodic and binary sets of words: for any nonperiodic ternary set F, if $XF = FX$, then $X = \cup_{i \in I} F^i$, for some $I \subseteq \mathbb{N}$. This, however, does not hold for sets of cardinality at least four. E.g., for $F = \{a, ab, ba, bb\}$, the set $X = F \cup F^2 \cup \{bab, bbb\}$ commutes with F but it is not a union of powers of F, see [13].

For a finite language F we define two parameters:

$$l_F = \min_{u \in F} |u|, \quad L_F = \max_{u \in F} |u|,$$

where $|u|$ denotes the length of u.

Let Σ be a finite alphabet, and Ξ a finite set of unknowns in one-to-one correspondence with a set of nonempty words $X \subseteq \Sigma^*$, say $\xi_i \leftrightarrow x_i$, for some fixed enumeration of X. A (constant-free) *equation* over Σ with Ξ as the set of unknowns is a pair $(u, v) \in \Xi^\omega \times \Xi^\omega$, usually written as $u = v$. The subset X *satisfies* the equation $u = v$ if the morphism $h : \Xi^\omega \to \Sigma^\omega$, $h(\xi_i) = x_i$, for all $i \geq 0$, verifies $h(u) = h(v)$. These notions extend in a natural way to *systems of equations.*

The *dependence graph* of a system of equations S is the nondirected graph G, whose vertices are the elements of Ξ, and whose edges are the pairs $(\xi_i, \xi_j) \in \Xi \times \Xi$, with ξ_i and ξ_j appearing as the first letters of the left and right handsides of some equation of S, resp.

The following basic result on Combinatorics of Words, see [11], is very useful and efficient in our later considerations.

Lemma 4.4.1 ([11], Graph Lemma). *Let S be a system of word equations and let $X \subset \Sigma^+$ be a subset satisfying it. If the dependence graph of S has p connected components, then there exists a subset F of cardinal at most p such that $X \subseteq F^+$.*

Most often, we will use the following consequence of the Graph Lemma: a system S having a connected dependency graph has only periodic solutions. Note also that in Graph Lemma it is *essential* that all words of X are *nonempty.*

We say that the mapping $\phi : 2^{\Sigma^*} \to 2^{\Sigma^*}$ is *linear* if there are some languages A, $B_1, B_2, \ldots, B_n$, $C_1, C_2, \ldots, C_n$, $n \geq 0$, such that $\phi(L) = A + B_1LC_1 + B_2LC_2 + \ldots + B_nLC_n$, for any $L \in 2^{\Sigma^*}$. We say that a language equation $\phi(X) = \psi(X)$ with X as the only unknown is *linear*, if both ϕ and ψ are linear mappings.

We prove in the next result that the uniqueness of the centralizer is a general property of linear equations and their maximal solutions.

Lemma 4.4.2. *Any satisfiable linear language equation has a unique maximal solution. Moreover, the maximal solution is the union of all solutions of the equation.*

Proof. Any satisfiable linear equation has at least one maximal solution which is the union of all solutions of that equation. Due to the linearity, this is indeed a solution. To prove the uniqueness of the maximal solution, assume there is a satisfiable linear language equation having two maximal solutions X_1 and X_2. Since the equation is linear, $X_1 + X_2$ is also a solution of the equation. But $X_1, X_2 \subseteq X_1 \cup X_2$ and since both X_1 and X_2 are maximal, we must have $X_1 = X_1 \cup X_2 = X_2$. □

4.4.1 A decomposition result

We consider now Conway's Problem for ternary sets, i.e., sets with three words. Lemma 4.4.3 will be the main tool we will use in proving the rationality of the centralizer of any ternary set. We prove in this lemma that in the ternary case, all the nonempty words of the centralizer have as a prefix an element of the set. For the sake of completeness, we give two proofs of this result. The first one, the longest of them, is self-contained, it uses only elementary techniques of Combinatorics on Words, and it gives a deep insight on the nature of the problem. The second one is much shorter but it relies heavily on a deep result of [66] concerning the commutation with prefix codes.

Lemma 4.4.3. *Let F be a nonperiodic three-word set such that $1 \notin F$, and let $\mathcal{C}(F)$ be its centralizer. Then, all words of $\mathcal{C}(F) \setminus \{1\}$ have as a prefix a word from F.*

Proof. (First proof of Lemma 4.4.3) Let $F = \{u, v, w\}$ be a nonperiodic three-word set such that $1 \notin F$. Clearly, since $F\mathcal{C}(F) = \mathcal{C}(F)F$, all long enough words of $\mathcal{C}(F)$ have as a prefix a word from F. Let us consider the set of those words of $\mathcal{C}(F)$ which do not have as a prefix a word from F. The claim of the lemma is that this set, say X_0, contains only the empty word.

Obviously, $1 \in X_0$, so assume that there are nonempty words in X_0, and let x be a minimal such word, with respect to length. Let $r_0 = s_0 = t_0 = x$. Since $\mathcal{C}(F)F = F\mathcal{C}(F)$, there are $\alpha_n, \beta_n, \gamma_n \in F$, $r_n, s_n, t_n \in \mathcal{C}(F)$, such that

$$r_{n-1}u = \alpha_n r_n, \qquad s_{n-1}v = \beta_n s_n, \qquad t_{n-1}w = \gamma_n t_n,$$

for all $n \geq 1$. Consequently,

$$xu^n = \alpha_1 \dots \alpha_n r_n, \qquad xv^n = \beta_1 \dots \beta_n s_n, \qquad xw^n = \gamma_1 \dots \gamma_n t_n, \tag{4.1}$$

for all $n \geq 1$ and moreover,

$$\begin{aligned} xu^\omega &= \alpha_1\alpha_2 \dots \alpha_n \dots \\ xv^\omega &= \beta_1\beta_2 \dots \beta_n \dots \\ xw^\omega &= \gamma_1\gamma_2 \dots \gamma_n \dots \end{aligned} \tag{4.2}$$

Let us denote $A = \{\alpha_1, \beta_1, \gamma_1\}$.

If the cardinal of A is 3, that is to say, $A = F$, then applying Graph Lemma on (4.2), for the set of unknowns $\{x, u, v, w\}$, we conclude that F is periodic: a contradiction.

If, on the other hand, A is a singleton, e.g. $A = \{u\}$, then, as x is from X_0, we have that x is a proper prefix of u: $u = xt$, with $t \neq 1$. Hence, we conclude again by applying Graph Lemma on (4.2), for the set of unknowns $\{t, u, v, w\}$: F must be periodic, a contradiction.

Assume now that A has cardinal 2.

Claim 1. X_0 is totally ordered by the prefix relation.

Proof of Claim 1. Consider a word $x' \in X_0$, distinct from x. Then we obtain that

$$\begin{aligned} x'u^\omega &= \alpha'_1\alpha'_2\ldots\alpha'_n\ldots \\ x'v^\omega &= \beta'_1\beta'_2\ldots\beta'_n\ldots \\ x'w^\omega &= \gamma'_1\gamma'_2\ldots\gamma'_n\ldots, \end{aligned} \tag{4.3}$$

for some $\alpha'_i, \beta'_i, \gamma'_i \in F$. We can assume that $A' = \{\alpha'_1, \beta'_1, \gamma'_1\}$ has cardinal 2, since otherwise the problem is solved as above. Necessarily, the intersection $A \cap A'$ is nonempty, since $|A \cup A'| \leq 3$. Thus, for any $\delta \in A \cap A'$, we obtain from (4.2) and (4.3) that both x and x' are prefixes of δ, and thus, one is prefix of the other. The claim is thus proved.

Now we go back to the case when $|A| = 2$. To make a choice, let $A = \{u, v\}$ and $\gamma_1 = u$. If $\alpha_1 = u$ or $\beta_1 = u$, then we can conclude again by Graph Lemma, using the fact that $u = xt$, and the system (4.2). It remains the case when $\alpha_1 = \beta_1 = v$. By (4.1), we obtain that

$$xu = vy_1, \qquad xv = vy_2, \qquad xw = uy_3, \tag{4.4}$$

for some $y_1, y_2, y_3 \in \mathcal{C}(F)$. Note that x is a proper prefix of both u and v, and hence, $|x| < |u|$ and $|x| < |v|$.

Claim 2. If $|w| < |u|$ and $|w| < |v|$, then either $xw = wx$, or there are some integers $l, r \geq 1$ such that $w^l x = \delta w^{l-1}$ and $xw^r = w^{r-1}\delta'$, with $\delta, \delta' \in \{u, v\}$.

Proof of Claim 2. Let $x_1 = x$. Then clearly, there is $l \geq 1$ such that $wx_i = x_{i+1}w$, for all $1 \leq i \leq l-1$, and either $wx_l = x_k w$, for some $1 \leq k \leq l$, or $wx_l = y\delta$, for some $\delta \in \{u, v\}$ and $y \in \mathcal{C}(F)$.

In the former case we obtain that $x_k w^{l-k+1} = w^{l-k+1}x_k$, implying that $x_k w = wx_k$. In turn, this implies that $x_{k-1}, \ldots, x_1$ also commute with w; in particular, we obtain that $xw = wx$.

In the latter case, we obtain that $w^l x = y\delta w^{l-1}$, $\delta \in \{u, v\}$. Since $|x| < |\delta|$, it follows that $|y| < |w|$ and so, $|y| < l_F$. Thus, $y \in X_0$. Also note that, since $|w| < |u|$ and $|w| < |v|$, this implies that $|y| < |x|$. Since x is minimal in X_0 and X_0 is totally ordered by the prefix relation, it follows that $y = 1$. Consequently, $w^l x = \delta w^{l-1}$.

The second part of the claim can be proved using a symmetric argument.

We distinguish now two cases, depending on whether or not $y_2 \in X_0$.

Case 1. If $y_2 \notin X_0$ then, since $|y_2| = |x| < |u|, |v|$, we must have $w \leq y_2$. Consequently, $|w| < |u|$ and $|w| < |v|$. By Claim 2, we obtain that either $wx = xw$, or $w^n x = uw^{n-1}$, or $w^n x = vw^{n-1}$, for some $n \geq 1$. Adding

either of these relations to system (4.2), with $\alpha_1 = \beta_1 = v$ and $\gamma_1 = u$, we obtain by Graph Lemma on the set of unknowns $\{x, u, v, w\}$ that F must be periodic, a contradiction.

Case 2. If $y_2 \in X_0$, then, as $|x| = |y_2|$, we obtain by Claim 1 that $x = y_2$. Thus, $xv = vx$, i.e., $x = p^i$ and $v = p^j$, for some primitive word $p \in \Sigma^+$, and some positive integers $i < j$. Moreover, since $x \leq u$, u is of the form $u = p^i u'$, with $u' \neq 1$. The first and third equations of the system (4.2) can now be written as follows:

$$\begin{aligned} p^i p^i u' (p^i u')^\omega &= p^j \alpha_2 \alpha_3 \ldots, \\ p^i w^\omega &= p^i u' \gamma_2 \gamma_3 \ldots, \end{aligned} \tag{4.5}$$

with $\alpha_m, \gamma_m \in \{p^i u', p^j, w\}$, for all $m \geq 2$. It is straightforward to see that if $2i \neq j$, then Graph Lemma applied on (4.5) for the set of unknowns $\{p, u', w\}$ implies that F must be periodic; this is impossible. Consequently, $2i = j$, i.e., $v = x^2$, and so, by (4.4), $xu = x^2 y_1$ and $xw = uy_3$, i.e.,

$$u = xy_1, \tag{4.6}$$

$$v = x^2, \tag{4.7}$$

$$w = y_1 y_3. \tag{4.8}$$

In particular, we have that $wu^\omega = y_1 y_3 u^\omega$ and so, since $y_3 u^\omega \in F^\omega$,

$$wu^\omega = y_1 \nu_1 \nu_2 \ldots, \tag{4.9}$$

with $\nu_i \in F$, for all $i \geq 1$. Note also that by (4.6), $y_1 \neq 1$, as $x \neq u$.

Consider the word $y_1 v \in \mathcal{C}(F)F$. There are $\delta \in F$ and $z \in \mathcal{C}(F)$ such that $y_1 v = \delta z$.

If $\delta = u = xy_1$ or $\delta = v = x^2$, then $y_1 v u^\omega = \delta z u^\omega = xtu^\omega$, with $t = y_1 z$ or $t = xz$. In both cases, $t \in \mathcal{C}(F)$ and so, $tu^\omega \in F^\omega$, i.e.,

$$y_1 v u^\omega = x \mu_1 \mu_2 \ldots, \tag{4.10}$$

for some $\mu_i \in F$, for all $i \geq 1$. Applying Graph Lemma on (4.6), (4.7), (4.9), and (4.10), for the set of unknowns $\{u, v, w, x, y_1\}$, we obtain the periodicity of F. Indeed, the dependency graph of this system of equations is depicted in Figure 4.1.

Thus, $\delta = w$, i.e., $y_1 v = wz$, implying by (4.7) and (4.8) that

$$x^2 = y_3 z. \tag{4.11}$$

In this case, we prove the following claim.

Claim 3. For any $t \in \mathcal{C}(F) \setminus \{1\}$, if $|t| < |w|$, then $x \leq t$.

Proof of Claim 3. If $t \in X_0$, then by Claim 1 and the minimality of x, it follows that $x \leq t$. If $t \notin X_0$, then there is $\delta \in F$ such that $\delta \leq t$. However,

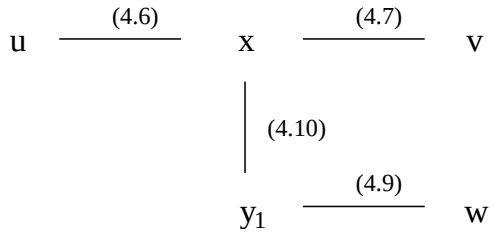


Figure 4.1: The dependency graph of the equations (4.6), (4.7), (4.9), and (4.10).

$|t| < |w|$, and so, $\delta = u$ or $\delta = v$. In both cases we have by (4.6) and (4.7) that $x \leq t$, proving the claim.

If $y_3 = 1$, then by (4.8), $y_1 = w$ and in this case $F = \{xw, x^2, w\}$. As ${}^{\omega}(xw)x \in {}^{\omega}F$, we obtain a nontrivial relation in x and w. Consequently, F must be periodic.

If $y_3 \neq 1$, then, as $y_1 \neq 1$, we obtain from (4.8) that $|y_1|, |y_3| < |w|$. Moreover, we obtain that $|w| > |x|$ as otherwise, y_1 and y_3 would be words from $X_0 \setminus \{1\}$ shorter than x. Consequently, $|x| < l_F$. Moreover, Claim 3 applied for $t = y_3$ gives that $x \leq y_3$ and in particular, $|y_3| \geq |x|$. Thus, we derive from (4.11) that $|z| \leq |x| < l_F$ and so, $z \in X_0$. Since x is minimal in $X_0 \setminus \{1\}$, either $z = 1$, or $z = x$. Thus, by (4.11), $y_3 \in \{x, x^2\}$ and by (4.8), $w \in \{y_1x, y_1x^2\}$.

Consequently, $F = \{xy_1, x^2, w\}$, where $w = y_1x$, or $w = y_1x^2$. In both cases, since $y_1w^{\omega} \in F^{\omega}$, we obtain a nontrivial relation in x and y_1 and therefore, F must be periodic: again a contradiction.

The conclusion is that $X_0 = \{1\}$, which was the claim of the lemma. □

Proof. (Second proof of Lemma 4.4.3) Let F be a nonperiodic three-word set $F = \{u, v, w\}$. By Corollary 4.3.3, we can assume without loss of generality that two words of F, say u and v, start with different letters.

Assume that there is a word $x \in \mathcal{C}(F) \setminus \{1\}$ such that no word of F is a prefix of x. Then, as $\mathcal{C}(F)F = F\mathcal{C}(F)$, there are some words $\alpha, \beta \in F$ and $y, z \in \mathcal{C}(F)$ such that $xu = \alpha y$ and $xv = \beta z$. It follows that x is a proper prefix of both α and β and so, α and β start with the same letter (as x). Moreover, they are prefix incomparable, since their common prefix is x and $|x| < |\alpha|$, $|x| < |\beta|$. Consequently, F is a prefix set. Thus, by [66], $\mathcal{C}(F) = F^*$ and so, $x \in F^+$. This is a contradiction. □

For a language L and a word $\alpha \in L$, we say that α is *suffix distinguishable* in L if for any $\beta \in L \setminus \{\alpha\}$, α and β are suffix incomparable.

For ternary sets, having a suffix distinguishable word is equivalent with having no word as a suffix of both the other two.

Lemma 4.4.4. *If F is a three-word language then F has a suffix distinguishable word if and only if no word of F is a suffix of both the other two.*

Proof. Clearly, if F has a suffix distinguishable word, then no word of F is a suffix of both the other two. For the reverse direction, let $F = \{u, v, w\}$. If no two words of F are suffix comparable, then the claim follows trivially. If there are two such words, say $u \leq_s v$, then we claim that w is suffix distinguishable in L. Indeed, if $u \leq_s w$, then u is a suffix of both v and w, and if $w \leq_s u$, then w is a suffix of both u and v. Thus, u and w are suffix incomparable. Since $u \leq_s v$, it follows that v and w are also suffix incomparable. □

Note though that in Lemma 4.4.4 it is essential to deal with ternary sets. E.g., the four-word set $\{a, b, ab, ba\}$ has no word as a suffix of all the other three, but nevertheless, it has no suffix distinguishable words.

We have proved in Corollary 4.3.3 that Conway's Problem can be reduced to branching sets of words. By symmetry, it follows also that for ternary sets, Conway's Problem can be reduced to those ternary sets of words having no word as a suffix of both the other two. Although this result, Lemma 4.4.5, is only a reformulation of Corollary 4.3.3, this is more convenient to us, as it allows us to work with prefix decompositions rather than suffix decomposition, i.e., in a left-to-right manner.

Lemma 4.4.5. *Let F be a nonperiodic ternary set. There is a ternary set G having a suffix distinguishable word such that $\mathcal{C}(F)$ is rational if and only if $\mathcal{C}(G)$ is rational. Moreover, $\mathcal{C}(F) = F^*$ if and only if $\mathcal{C}(G) = G^*$.*

4.4.2 The prefix decomposition

Let F be a finite language, $1 \notin F$. Clearly, the language F can be uniquely written in the form

$$F = u_1F_1 + u_2F_2 + \cdots + u_nF_n, \tag{4.12}$$

such that the following conditions are satisfied:

(i) $1 \in F_i$, for all $i = 1, \ldots, n$.

(ii) u_i and u_j are prefix incomparable, for al $i \neq j$.

We say that (4.12) is the *prefix decomposition* of F.

Example 4.4.1. The prefix-decomposition of $F = \{a, aa, b, bab\}$ is $F = a \cdot \{1, a\} + b \cdot \{1, ab\}$.

A ternary set F, $1 \notin F$, can only have one of the following three types of prefix decomposition:

I. $F = u_1 + u_2 + u_3$, with u_i, u_j prefix incomparable for all $1 \leq i < j \leq 3$.

II. $F = u_1 + u_1v + u_2$, with u_1 and u_2 prefix incomparable, and $v \neq 1$.

III. $F = u_1 + u_1v + u_1w$, with $v, w \neq 1$ and $v \neq w$.

We prove in the next three lemmata the rationality of the centralizer in the ternary case. We discuss separately each type of prefix decomposition.

4.4.3 Ternary sets with prefix decomposition of type I or II

The first case is solved in [66] in a more general setting (i.e., for all prefix sets), but for the sake of completeness, we give here an independent and shorter proof. Note that our proof relies on Lemma 4.4.3. The first proof of this lemma was independent of the results of [66].

Lemma 4.4.6 (Case I). *Let u_1, u_2, u_3 be three nonempty words, such that u_i and u_j are prefix incomparable for all $1 \leq i < j \leq 3$, and let $F = \{u_1, u_2, u_3\}$. If F has a suffix distinguishable word, then $\mathcal{C}(F) = F^*$.*

Proof. Let u_1 be the suffix distinguishable word of F.

By Lemma 4.4.3, $\mathcal{C}(F)$ is of the form

$$\mathcal{C}(F) = u_1X_1 + u_2X_2 + u_3X_3 + 1. \tag{4.13}$$

Thus, as $F\mathcal{C}(F) = \mathcal{C}(F)F$, it follows that $u_1\mathcal{C}(F) + u_2\mathcal{C}(F) + u_3\mathcal{C}(F) = u_1X_1F + u_2X_2F + u_3X_3F + F$, and so,

$$\mathcal{C}(F) = X_1F + 1, \quad \mathcal{C}(F) = X_2F + 1, \quad \mathcal{C}(F) = X_3F + 1. \tag{4.14}$$

Thus, $X_iF = X_jF$, for all $i \neq j$. We claim that $X_i = X_j$.

Let $x_i \in X_i$. Then, for $i \neq j$, $x_iu_1 \in X_jF$, i.e., $x_iu_1 = x_j\alpha$, for some $x_j \in X_j$ and $\alpha \in F$. As u_1 is suffix distinguishable in F, we must have $\alpha = u_1$ and so, $x_i = x_j \in X_j$. Consequently, $X_i \subseteq X_j$, for all $i \neq j$, proving the claim.

Consequently, by (4.13), $\mathcal{C}(F) = FX_1+1$. Then, by (4.14), it follows that $X_1F+1 = FX_1+1$, or equivalently, $FX_1 = X_1F$. Thus, $X_1 \subseteq \mathcal{C}(F)$ and so, $\mathcal{C}(F) \subseteq F\mathcal{C}(F)+1$. The other inclusion also holds and so, $\mathcal{C}(F) = F\mathcal{C}(F)+1$. In turn, this implies (see, e.g., [21]) that $\mathcal{C}(F) = F^*$. □

Our proof of Case II is similar to that of Case I.

Lemma 4.4.7 (Case II). *Let u_1, u_2, v be three nonempty words, such that u_1 and u_2 are prefix incomparable, and let $F = \{u_1, u_1v, u_2\}$. If F has a suffix distinguishable word, then $\mathcal{C}(F) = F^*$.*

Proof. Let δ_0 be the suffix distinguishable word of F.

If $F = u_1 + u_1 v + u_2$, then by Lemma 4.4.3, $\mathcal{C}(F)$ is of the form $\mathcal{C}(F) = u_1 X_1 + u_2 X_2 + 1$. Then, as

$$(u_1 + u_1 v + u_2)\mathcal{C}(F) = (u_1 X_1 + u_2 X_2 + 1)F, \tag{4.15}$$

we obtain that

$$(1 + v)\mathcal{C}(F) = X_1 F + (1 + v) \quad \text{and} \quad \mathcal{C}(F) = X_2 F + 1. \tag{4.16}$$

Thus,

$$(1 + v)X_2 F + 1 + v = X_1 F + 1 + v.$$

We claim that $X_1 = (1 + v)X_2$.

Let $x_2 \in (1 + v)X_2$. If $x_2\delta_0 \in 1 + v$, then $x_2\delta_0 = v$, i.e., $\delta_0 \in \mathrm{Suf}(u_1 v)$. Thus, $\delta_0 = u_1 v$ and so, $x_2 u_1 = 1$, which is impossible.

Hence, $x_2\delta_0 \in X_1 F$, i.e., $x_2\delta_0 = x_1\alpha$, for some $x_1 \in X_1$ and $\alpha \in F$. Since δ_0 is suffix distinguishable in F, we must have $\alpha = \delta_0$, and so, $x_2 \in X_1$. Consequently, $(1 + v)X_2 \subseteq X_1$.

Let $x_1 \in X_1$. If $x_1\delta_0 \in 1 + v$, then $x_1\delta_0 = v$, which is impossible as we proved above. Thus, $x_1\delta_0 \in (1 + v)X_2 F$, and as above we obtain that $x_1 \in (1 + v)X_2$, proving that $X_1 \subseteq (1 + v)X_2$, and concluding the claim.

We thus obtain that $X_1 = (1+v)X_2$ and so, $\mathcal{C}(F) = FX_2 + 1$. Moreover, by (4.16), $FX_2 + 1 = X_2 F + 1.$, i.e., $FX_2 = X_2 F$. Thus, $X_2 \subseteq \mathcal{C}(F)$, i.e., $\mathcal{C}(F) \subseteq F\mathcal{C}(F) + 1$. As the other inclusion also holds, it follows that $\mathcal{C}(F) = F\mathcal{C}(F) + 1$, implying that $\mathcal{C}(F) = F^*$. □

4.4.4 Ternary sets with prefix decomposition of type III

It turns out that Case III is much more difficult to settle than the first two cases. We prove here that $\mathcal{C}(F)$ is effectively rational for any set F having a type III prefix decomposition and a suffix distinguishable word. It remains as an open problem whether or not $\mathcal{C}(F) = F^*$ for all such sets F with $1 \notin F$, see also Section 4.6.

Lemma 4.4.8 (Case III). *Let F be a three-word set, $F = \{u, uu', uu''\}$, for some nonempty words u, u', u'', If F has a suffix distinguishable word, then $\mathcal{C}(F)$ is rational.*

Proof. Since F has a suffix distinguishable word δ_0, F is not periodic. The prefix decomposition of F is $F = u + u^m v + u^n w$, where $m, n \geq 1$, and u is not a prefix of either v, or w.

If both v and w are empty, then F is periodic. We can thus assume that at least one of them, say v, is nonempty. Moreover, if $w \neq 1$, we assume without loss of generality that $m \leq n$.

The centralizer $\mathcal{C}(F)$ is the maximal solution of the equation

$$FX = XF. \tag{4.17}$$

By Lemma 4.4.3, $\mathcal{C}(F)$ is of the form $\mathcal{C}(F) = uL_1 + 1$. Thus, by canceling the common prefix u, (4.17) can be refined to $(1 + u^{m-1}v + u^{n-1}w)uL_1 + (1 + u^{m-1}v + u^{n-1}w) = L_1F + (1 + u^{m-1}v + u^{n-1}w)$ and so, L_1 is a solution of the equation

$$(u+u^{m-1}vu+u^{n-1}wu)X_1+u^{m-1}v+u^{n-1}w = X_1F+u^{m-1}v+u^{n-1}w. \tag{4.18}$$

Note that for any solution Y_1 of (4.18), $uY_1 + 1$ is a solution of (4.17) and so, since $\mathcal{C}(F)$ is the maximal solution of (4.17), $uY_1 \subseteq \mathcal{C}(F)$ and $Y_1 \subseteq L_1$. Thus, L_1 is the maximal solution of (4.18). Moreover, $\mathcal{C}(F)$ is rational if and only if L_1 is rational.

If $m - 1 \geq 1$ and $n - 1 \geq 1$, then clearly, all long enough words of L_1 have u as a prefix. The short words are prefixes of u. Thus, L_1 is of the form $L_1 = uL_2 + L_{1,0}$, with $L_{1,0} \subseteq \mathrm{Pref}(u)$. Note also that all words of $L_{1,0}F$ start with u. It then follows from (4.18) that L_2 is a solution of the equation

$$\begin{aligned}(u + u^{m-2}vu^2 + u^{n-2}wu^2)X_2 + (1 + u^{m-2}vu + u^{n-2}wu)L_{1,0}+\\ + (u^{m-2}v + u^{n-2}w) = X_2F + u^{-1}(L_{1,0}F) + (u^{m-2}v + u^{n-2}w).\end{aligned} \tag{4.19}$$

Moreover, for any solution Y_2 of (4.19), $uY_2 + L_1$ is a solution of (4.18) and so, $uY_2 \subseteq L_1$ and $Y_2 \subseteq L_2$. Thus, L_2 is the maximal solution of (4.19). Also, L_1 is rational if and only if L_2 is rational.

Iterating the same argument, we derive the equation

$$\begin{aligned}(u + vu^m + u^{n-m}wu^m)X_m + \sum_{i=0}^{m-1}(vu^i + u^{n-m}wu^i)L_{i,0} + L_{m-1,0} =\\ = X_mF + \sum_{i=0}^{m-1} u^{-(m-i)}(L_{i,0}F),\end{aligned} \tag{4.20}$$

where $L_{0,0} = \{1\}$ and $L_{i,0} \subseteq \mathrm{Pref}(u)$, for all $1 \leq i \leq m - 1$. Our above statement of the maximality still holds, so that $\mathcal{C}(F)$ is rational if and only if the maximal solution L_m of (4.20) is rational.

Let us denote

$$G = u + vu^m + u^{n-m}wu^m,$$
$$A = \sum_{i=0}^{m-1} u^{-(m-i)}(L_{i,0}F), \quad B = \sum_{i=0}^{m-1}(vu^i + u^{n-m}wu^i)L_{i,0} + L_{m-1,0},$$

and observe that the set G has at least one word which is prefix incomparable to the other words of G. Thus, G is of the form I or II.

The proof of the lemma is reduced to proving that the maximal solution of the language equation

$$YF + A = GY + B. \tag{4.21}$$

is rational. Depending on the type of prefix decomposition of G, we distinguish two cases, discussed separately in the following.

Case 1. The prefix decomposition of G is of type I, i.e., $G = u_1 + u_2 + u_3$, with u_1, u_2, u_3 pairwise prefix incomparable. Recall also that $u \in G$, say $u_1 = u$.

Let

$$K = \mathrm{Pref}'(G) \cup B \cup \mathrm{Pref}'(B),$$

where Pref' denotes the set of proper prefixes: for any word u, $\mathrm{Pref}'(u) = \mathrm{Pref}(u) \setminus \{1, u\}$, and for any set L, $\mathrm{Pref}'(L) = \mathrm{Pref}(L) \setminus L$.

Observe that $u_j^{-1}K = \emptyset$, for all $1 \leq j \leq 3$. Indeed, $u_j^{-1}\,\mathrm{Pref}'(G) = \emptyset$, since all words of G are prefix incomparable. Also, since $L_{i,0} \subseteq \mathrm{Pref}(u)$, for all $0 \leq i \leq n-1$, one can readily see that $u_j^{-1}B = \emptyset$, and thus, also $u_j^{-1}\,\mathrm{Pref}'(B) = \emptyset$, for all $1 \leq j \leq 3$.

We construct a chain of language equations of the form

$$Y_nF + A_n = GY_n + B_n,$$

with

$$A_n \subseteq \{t \mid \text{ there is } k \geq 1, u^k t \in KF\} \quad \text{and} \quad B_n \subseteq K, \tag{4.22}$$

such that the maximal solution of the n-th equation is rational if and only if that of the $(n+1)$-st equation is so. Let R_n denote the maximal solution of the n-th equation of the chain.

The first equation in our chain is (4.21). Clearly, it is of the required form, as $u \in G$ and $B \subseteq K$.

Assume now that we have constructed the n-th equation in the chain, for some $n \geq 1$. Then R_n is of the form

$$R_n = u_1R_{n,1} + u_2R_{n,2} + u_3R_{n,3} + R_{n,0},$$

for some languages $R_{n,1}, R_{n,2}, R_{n,3} \subseteq \Sigma^*$ and

$$R_{n,0} \subseteq \{t \in \mathrm{Pref}'(B_n) \cup \mathrm{Pref}'(G) \mid \text{ for all } \alpha \in G, \alpha \not\leq t\} \subseteq K.$$

Then, from the n-th equation of the chain, we obtain that

$$R_{n,i}F + u_i^{-1}(R_{n,0}F) + u_i^{-1}A_n = R_n + u_i^{-1}B_n, \tag{4.23}$$

for all $1 \leq i \leq 3$. Clearly, as $B_n \subseteq K$, we have that $u_i^{-1}B_n = \emptyset$, for all $1 \leq i \leq 3$. Thus, $R_{n,i}F + u_i^{-1}(R_{n,0}F) + u_i^{-1}A_n = R_n$, for all $1 \leq i \leq 3$, implying that

$$R_{n,i}F + u_i^{-1}(R_{n,0}F) + u_i^{-1}A_n = R_{n,j}F + u_j^{-1}(R_{n,0}F) + u_j^{-1}A_n, \tag{4.24}$$

for all $1 \leq i, j \leq 3$.

Let $y_i \in R_{n,i}$. Since $y_i\delta_0 \in R_{n,i}F$, we have the following possibilities by (4.24):

(i) If $y_i\delta_0 \in R_{n,j}F$, then, as δ_0 is suffix distinguishable in F, we obtain $y_i \in R_{n,j}$.

(ii) If $y_i\delta_0 \in u_j^{-1}(R_{n,0}F)$, then $u_jy_i\delta_0 \in R_{n,0}F$. Since δ_0 is suffix distinguishable in F, $u_jy_i \in R_{n,0}$, which is impossible because $\alpha^{-1}R_{n,0} = \emptyset$, for all $\alpha \in G$.

(iii) If $y_i\delta_0 \in u_j^{-1}A_n$, then $u_jy_i\delta_0 \in A_n$, i.e., $u^ku_jy_i\delta_0 \in KF$, for some $k \geq 1$. Thus, $u^ku_jy_i \in K$, which is impossible since $u^{-1}K = \emptyset$.

Consequently, $R_{n,i} \subseteq R_{n,j}$, for all $i \neq j$, i.e., $R_{n,1} = R_{n,2} = R_{n,3}$. Thus, $R_n = GR_{n,1} + R_{n,0}$. Moreover, $R_{n,1}$ satisfies the relations (4.23), rewritten now as

$$R_{n,1}F + u_i^{-1}(R_{n,0}F) + u_i^{-1}A_n = GR_{n,1} + R_{n,0},$$

for all $1 \leq i \leq 3$.

Let $A_{n+1,i} = u_i^{-1}(R_{n,0}F) + u_i^{-1}A_n$, $B_{n+1,i} = R_{n,0}$, for all $1 \leq i \leq 3$. Also, let $A_{n+1} = GR_{n,1} \setminus R_{n,1}F$ and $B_{n+1} = R_{n,1}F \setminus GR_{n,1}$. Then $R_{n,1}$ is a solution of the equation

$$Y_{n+1}F + A_{n+1} = GY_{n+1} + B_{n+1}, \tag{4.25}$$

which is the $(n+1)$-st equation in our chain. Since $A_{n+1} \subseteq A_{n+1,i}$ and $B_{n+1} \subseteq B_{n+1,i}$, for all $1 \leq i \leq 3$, it is straightforward to prove that for any solution Y of (4.25), $GY + R_n$ is a solution of the n-th equation of the chain. Thus, since R_n is the maximal solution of this equation, it follows that $GY \subseteq R_n$ and so, $Y \subseteq R_{n,1}$. Consequently, $R_{n,1}$ is the maximal solution of the $(n+1)$-st equation in the chain. Clearly, R_n is rational if and only if $R_{n,1}$ is rational.

We have thus constructed a required chain of language equations. However, note that there are only a finite number of distinct equations in this chain. Indeed, by (4.22), the sets A_n and B_n contain only words of length up to $L_K + L_F$, for all $n \geq 1$ and so, there must be $m < n$ such that $A_m = A_n$ and $B_m = B_n$, i.e., the m-th and the n-th equations in the chain coincide. But this implies by Lemma 4.4.2 that $R_m = R_n$ and so, from the construction, we obtain that

$$R_m = G^{n-m}R_m + (G^{n-m-1}R_{n-1,0} + \ldots + GR_{m+1,0} + R_{m,0}).$$

As it is well-known (see, e.g., [21]), R_m is thus rational, implying the rationality of all R_p, $p \geq 1$. In particular we obtain that R_1 is rational and so, $\mathcal{C}(F)$ is rational.

Case 2. G has a type II prefix decomposition, i.e., it is of the form

$$G = u_1 + u_1t + u_2,$$

with u_1 and u_2 prefix incomparable.

Let

$$K = \mathrm{Pref}'(G) \cup B \cup \mathrm{Pref}'(B), \quad K' = \{\alpha \in K \mid u_1 \not\leq \alpha \text{ and } u_2 \not\leq \alpha\}.$$

Note that $u \in G$ and moreover, $u = u_1$, or $u = u_2$.

Claim. If $x\delta_0 \in u_1^{-1}K$, then $x \in \mathrm{Pref}'(t)$, where δ_0 is a suffix distinguishable word of F. Also, $u_2^{-1}K = \emptyset$.

Proof of the claim. The second part of the claim holds since u_1 and u_2 are prefix incomparable, and for any $0 \leq i \leq m-1$, the words in $L_{i,0}$ have lengths smaller than u (in fact, $L_{i,0} \subseteq \mathrm{Pref}(u)$).

For the first part of the claim, let x be a word such that $x\delta_0 \in u_1^{-1}K$. Thus, $u_1x\delta_0 \in K$. If $u_1x\delta_0 \in \mathrm{Pref}'(G)$, then necessarily $u_1x\delta_0 \in \mathrm{Pref}'(u_1t)$, i.e., $x \in \mathrm{Pref}'(t)$. Assume now that $u_1x\delta_0 \in B$.

If $u_1x\delta_0 \in L_{m-1,0}$, then we have in $L_{m-1,0}$ a word of length larger than $|u|$, contradicting $L_{m-1,0} \subseteq 1 + \mathrm{Pref}'(u)$.

If $u_1x\delta_0 = vu^ix_i$, for some $x_i \in L_{i,0}$, $0 \leq i \leq m-1$, then, as $|x_i| < |u| \leq |\delta_0|$, we have that $u_1x \leq vu^i$. Since G has a type II prefix decomposition, $vu^m = u_1t$ and so, $x \in \mathrm{Pref}'(t)$.

If $u_1x\delta_0 = u^{n-m}wu^ix_i$, for some $x_i \in L_{i,0}$, $0 \leq i \leq m-1$, then, as $|x_i| < |\delta_0|$, we have that $u_1x \leq u^{n-m}wu^i$, i.e., $u^{n-m}vu^m = u_1t$. Thus, $x \in \mathrm{Pref}'(t)$.

The claim is thus proved for B, and then clearly follows also for $\mathrm{Pref}'(B)$.

We construct in the following a chain of language equations of the form

$$Y_nF + A_n = GY_n + B_n, \tag{4.26}$$

with

$$\begin{aligned} &B_n \subseteq K, \text{ and} \\ &A_n \subseteq \{t \mid \exists k, l, k \leq m-1, l \leq n-1, \text{ such that } u_2^l u^k t \in K'F\}, \end{aligned} \tag{4.27}$$

such that the maximal solution of the n-th equation is rational if and only if that of the $(n+1)$-st equation is so. Let S_n denote the maximal solution of the n-th equation.

The first equation in our chain is (4.21). Clearly, it is of the required form, as $u \in G$, $B \subseteq K$, and $L_{i,0} \subseteq 1 + \mathrm{Pref}'(u)$, for all $0 \leq i \leq m-1$.

Assume now that we have constructed the n-th equation, for some $n \geq 1$. Then, it follows from (4.26) that S_n is of the form

$$S_n = u_1S_{n,1} + u_2S_{n,2} + S_{n,0},$$

for some languages $S_{n,1}, S_{n,2} \subseteq \Sigma^*$ and

$$S_{n,0} \subseteq \{s \in \mathrm{Pref}'(u_1, u_2) \cup \mathrm{Pref}'(B_n) \mid u_1 \not\leq s, u_2 \not\leq s\} \subseteq K.$$

Then, from (4.26), by canceling the common prefixes, we obtain that

$$\begin{aligned} S_{n,1}F + u_1^{-1}(S_{n,0}F) + u_1^{-1}A_n &= (1+t)S_n + u_1^{-1}B_n, \text{ and} \\ S_{n,2}F + u_2^{-1}(S_{n,0}F) + u_2^{-1}A_n &= S_n + u_2^{-1}B_n. \end{aligned} \tag{4.28}$$

Clearly, as $B_n \subseteq K$, we have by the claim that $u_2^{-1}B_n = \emptyset$. Thus,

$$\begin{aligned} (1+t)S_{n,2}F + (1+t)u_2^{-1}(S_{n,0}F) + (1+t)u_2^{-1}A_n + u_1^{-1}B_n \\ = S_{n,1}F + u_1^{-1}(S_{n,0}F) + u_1^{-1}A_n. \end{aligned} \tag{4.29}$$

Let $\alpha \in \{1, t\}$ and $y_2 \in S_{n,2}$. Since $\alpha y_2 \delta_0 \in (1+t)S_{n,2}F$, we have the following possibilities by (4.29):

(i) If $\alpha y_2 \delta_0 \in u_1^{-1}(S_{n,0}F)$, then $u_1 \alpha y_2 \delta_0 \in S_{n,0}F$. Since δ_0 is suffix distinguishable in F, we obtain that $u_1 \alpha y_2 \in S_{n,0}$, which is impossible.

(ii) If $\alpha y_2 \delta_0 \in u_1^{-1}A_n$, then $u_2^l u^k u_1 \alpha y_2 \delta_0 \in K'F$, for some $k \leq m-1$, $l \leq n-1$. Thus, $u_2^l u^k u_1 \alpha y_2 \in K'$, which is impossible.

(iii) If $\alpha y_2 \delta_0 \in S_{n,1}F$, then $\alpha y_2 \in S_{n,1}$, as δ_0 is suffix distinguishable in F.

Consequently, $(1+t)S_{n,2} \subseteq S_{n,1}$.

Consider now $y_1 \in S_{n,1}$. Since $y_1 \delta_0 \in S_{n,1}F$, we have the following four possible subcases by (4.29):

(iii.1) If $y_1 \delta_0 \in (1+t)S_{n,2}F$, then necessarily $y_1 \in (1+t)S_{n,2}$, as δ_0 is suffix distinguishable in F.

(iii.2) If $y_1 \delta_0 \in (1+t)u_2^{-1}A_n$, then either $u_2 y_1 \delta_0 \in A_n$, or $y_1 \delta_0 = tz$, for some $z \in u_2^{-1}A_n$.

In the former case, we obtain that $u_2^l u^k u_2 y_1 \delta_0 \in K'F$, for some $k \leq m-1$, $l \leq n-1$, and so, $u_2^l u^k u_2 y_1 \in K'$, a contradiction.

In the latter case, if $|y_1| \geq |t|$, then $y_1 = ty'$, for some $y' \in \Sigma^*$. Thus, $u_2 y' \delta_0 \in A_n$, and this was proved above to lead to contradiction. If $|y_1| < |t|$, then necessarily, $y_1 \leq t$.

(iii.3) If $y_1 \delta_0 \in (1+t)u_2^{-1}(S_{n,0}F)$, then either $u_2 y_1 \delta_0 \in S_{n,0}F$, or $y_1 \delta_0 = tz$, for some $z \in u_2^{-1}(S_{n,0}F)$.

In the former case, we obtain that $u_2 y_1 \in S_{n,0}$, a contradiction.

In the latter case, if $|y_1| \geq |t|$, then $y_1 = ty'$, for some $y' \in \Sigma^*$. Thus, $u_2 y' \delta_0 \in S_{n,0}F$, and this leads to a contradiction as in (iii.2). If $|y_1| < |t|$, then necessarily, $y_1 \leq t$.

(iii.4) If $y_1 \delta_0 \in u_1^{-1}B_n$, then, as $B_n \subseteq K$, we obtain by Claim 2 that $y_1 \leq t$.

Consequently, $S_{n,1} = (1+t)S_{n,2} + T_{n,0}$, for some language $T_{n,0} \subseteq 1 + \mathrm{Pref}'(t)$ and so,

$$S_n = GS_{n,2} + u_1 T_{n,0} + S_{n,0}.$$

Moreover, $S_{n,2}$ satisfies the relations (4.28), rewritten as

$$\begin{aligned}(1+t)S_{n,2}F + T_{n,0}F + u_1^{-1}(S_{n,0}F) + u_1^{-1}A_n &= \\ = (1+t)GS_{n,2} + (1+t)u_1T_{n,0} + (1+t)S_{n,0} + u_1^{-1}B_n,& \\ S_{n,2}F + u_2^{-1}(S_{n,0}F) + u_2^{-1}A_n = GS_{n,2} + u_1T_{n,0} + S_{n,0}.&\end{aligned}$$

Let $A'_{n+1} = u_2^{-1}(S_{n,0}F) + u_2^{-1}A_n$, $B'_{n+1} = u_1T_{n,0} + S_{n,0}$, $A''_{n+1} = T_{n,0}F + u_1^{-1}(S_{n,0}F) + u_1^{-1}A_n$, and $B''_{n+1} = (1+t)u_1T_{n,0} + (1+t)S_{n,0} + u_1^{-1}B_n$. Then

$$\begin{aligned}(1+t)S_{n,2}F + A''_{n+1} &= (1+t)GS_{n,2} + B''_{n+1},\\ S_{n,2}F + A'_{n+1} &= GS_{n,2} + B'_{n+1}.\end{aligned}$$

Denoting $A_{n+1} = GS_{n,2} \setminus S_{n,2}F$ and $B_{n+1} = S_{n,2}F \setminus GS_{n,2}$, it follows that $S_{n,2}$ is a solution of the equation

$$Y_{n+1}F + A_{n+1} = GY_{n+1} + B_{n+1}, \tag{4.30}$$

which is the $(n+1)$-st equation in our chain. Since $A_{n+1} \subseteq A'_{n+1}$, $B_{n+1} \subseteq B'_{n+1}$, $(1+t)A_{n+1} \subseteq (1+t)S_{n,2}F + A''_{n+1}$, and $(1+t)B_{n+1} \subseteq B''_{n+1}$, it is straightforward to prove that for any solution Y of (4.30), $GY + S_n$ is a solution of the n-th equation of the chain. Thus, since S_n is the maximal solution of the n-th equation of the chain, it follows that $GY \subseteq S_n$, and so, $Y \subseteq S_{n,2}$, for any solution Y of (4.30). Consequently, $S_{n,2}$ is the maximal solution of the $(n+1)$-st equation in the chain. Clearly, S_n is rational if and only if $S_{n,2}$ is rational.

We have thus constructed a required chain of language equations. Case 2 is now concluded similarly as in Case 1. □

We can conclude now Conway's Problem in the case of ternary sets.

Theorem 4.4.9. *The centralizer of any three-element set is rational.*

Proof. Let F be a three-word language. Clearly, if $1 \in F$, then $\mathcal{C}(F) = \Sigma^*$. Moreover, if F is periodic, then the claim follows by Theorem 4.3.2.

Assuming that F is not periodic and $1 \notin F$, by Lemma 4.4.5, we can assume without loss of generality that there is a word $\delta \in F$ suffix distinguishable in F. Then, F has a prefix decomposition of the form I, II, or III. The claim follows by Lemmata 4.4.6, 4.4.7, and 4.4.8, respectively. □

4.5 The branching point approach

Following ideas of [36], we investigate here a different approach to Conway's Problem, proposed in [40]. We proved in Corollary 4.3.3 that Conway's Problem can be reduced to branching sets of words, i.e., sets having words starting with different letters. In turn, for these sets of words, one only has

to establish the rationality of the set of the so-called *critical* points to obtain a solution to Conway's Problem. As an illustration of the approach, we give a simple, elementary solution to this problem for binary sets and for finite biprefixes. For binary sets, this gives a simpler solution than that in [13] and a simpler reformulation of the solution presented in Section 4.3. The result for finite biprefixes is obtained also in [66], through some involved combinatorial arguments, as well as in [29], using some algebraic result on the commutation of two formal power series. The proof based on the branching point approach on the other hand, is completely elementary, as well as self contained.

4.5.1 Branching points

Let X be a set of words and u a word. We say that u is a *branching point* of X if there are two distinct letters a and b such that both ua and ub are prefixes of some words in X^*. In other words, u can be extended in two different ways to words of X^*. We denote by $\mathcal{B}(X)$ the set of branching points of X. A branching point u of X is called *critical* if $u \notin X^*$.

We recall that a set of words L is called *branching* if L has words starting with different letters. We say that L is *marked* if no two words of L start with the same letter.

Example 4.5.1. (i) For any branching set of words X, $X \subseteq \mathcal{B}(X)$. Indeed, if $u, v \in X$ start with different letters, then for any $\alpha \in X$, we have $\alpha u, \alpha v \in X^*$.

(ii) Let $F = \{a, aba, bb\}$. Then a is a branching point of F: $aa, aba \in F^*$, while b is not a branching point. Indeed, ba is not a prefix of any word in F^*. Also, ab is a critical point: $aba, abb \in F^*$ and $ab \notin F^*$. Note that ab is not in $\mathcal{C}(F)$ since, by Lemma 4.4.7, $\mathcal{C}(F) = F^*$.

(iii) Let $F = \{aa, ab, ba, bb\}$. Then both a and b are critical points of F. Moreover, a and b are both in $\mathcal{C}(F)$. Indeed, $\mathcal{C}(F) = \{a, b\}^*$.

For a branching set of words L, the critical points are the only potential nontrivial elements of the centralizer $\mathcal{C}(L)$, i.e., elements outside L^*. Indeed, if L is branching, then $\mathcal{C}(L) \subseteq \mathcal{B}(L)$. We prove in the next result that for any rational language L, $\mathcal{B}(L)$ is rational, thus giving further support for a possible affirmative answer to Conway's Problem.

Theorem 4.5.1. *For any rational language R, the set of its branching points $\mathcal{B}(R)$ is rational.*

Proof. If R is rational, then R^* is rational and so is $\mathrm{Pref}(R^*)$, see [47]. Let $\mathcal{A}$ be a complete deterministic finite automaton accepting $\mathrm{Pref}(R^*)$, with δ its transition mapping, Q the set of states, F the set of final states, and q_0 its initial state. For each $q \in Q$ and each letter a, let $q_a = \delta(q, a)$.

We construct an automaton $\mathcal{A}'$ accepting $\mathcal{B}(A)$. Intuitively, to accept a word u, we walk in $\mathcal{A}$ with u and then we check whether or not both letters a and b lead to final states. Formally, let Q' be a set isomorphic to Q: $Q' = \{q' \mid q \in Q\}$, and let r, s be two new states, $r, s \notin Q \cup Q'$. The set of states of $\mathcal{A}'$ is $Q \times (Q' \cup \{r, s\}) \times (Q' \cup \{r, s\})$, the initial state is (q_0, r, r) and the transition mapping is defined as follows:

(i) $\delta'((q, r, r), x) = (\delta(q, x), r, r)$, for all letters x;

(ii) $\delta'((q, r, r), 1) = (q, s, s)$;

(iii) $\delta'((q, s, s), 1) = (q, q'_a, q'_b)$, for all letters $a \neq b$.

The set of final states of $\mathcal{A}'$ is $Q \times F' \times F'$, where $F' = \{q' \mid q \in F\}$. Consequently, $\mathcal{A}'$ is a so called generalized finite automaton and hence, it accepts a rational language. As it is easy to see, the language accepted by $\mathcal{A}'$ is $\mathcal{B}(R)$. Indeed, a word $u \in \mathcal{B}(R)$ if and only if $ua, ub \in \mathrm{Pref}(R^*)$. □

We proved in Corollary 4.3.3 that for any nonperiodic set of words L, there is a branching set of words L' such that $\mathcal{C}(L)$ is rational if and only if $\mathcal{C}(L')$ is rational. Moreover, $\mathcal{C}(L) = L^*$ if and only if $\mathcal{C}(L') = L'^*$. Consequently, Conway's Problem can be reduced to two types of sets: periodic sets and branching sets of words. The case of periodic sets is, however, easy to settle. As we proved in Theorem 4.3.2, for any periodic set $L \subseteq u^+$, with u primitive word, $\mathcal{C}(L) = u^*$.

Consider now the case of branching sets of words. Based on the above reduction, we give a simple proof for Conway's Problem in the case of binary sets. This result has been originally proved in [13] using somewhat more involved combinatorial arguments. Note also that we gave still another proof of this result in Theorem 4.3.4.

Theorem 4.5.2. *Let F be a nonperiodic binary set of words F, $1 \notin F$. Then $\mathcal{C}(F) = F^*$.*

Proof. By Corollary 4.3.3, we can assume without loss of generality that F is a branching set of words. Let $F = \{au, bv\}$, where a and b are distinct letters and u, v some words.

Assume that $F^* \neq \mathcal{C}(F)$ and let $x \in \mathcal{C}(F) \setminus F^*$. Since F is a prefix and $x \in \mathrm{Pref}(F^*)$, it follows that there are unique words $u_1, \ldots, u_m \in F$, $m \geq 0$ such that $x = u_1 \ldots u_m t$, for a word $t \in \mathrm{Pref}(F)$.

Observe now that x is a branching point. Indeed, since F is a branching set of words, all words of $\mathcal{C}(F)$ are branching points of F. Thus, t is also a branching point of F, and so, as $t \in \mathrm{Pref}(F)$, there are $\alpha, \beta \in \mathrm{Pref}(F)$ such that $ta \leq \alpha$ and $tb \leq \beta$. If $t \neq 1$ then, since F is marked, it follows that $\alpha = \beta$ and $a = b$, a contradiction. Thus, $t = 1$ and $x \in F^*$, again impossible. Consequently, $\mathcal{C}(F) = F^*$. □

4.5.2 A simple solution for finite biprefix sets

It is well-known, see [57], that the set of prefix codes is a free monoid. In particular, this implies that any prefix has a unique primitive root, similarly as words have. It is a consequence of a result of [66] characterizing the commutation with a prefix code that, for any prefix code X, we have $\mathcal{C}(X) = \rho(X)^*$, where $\rho(X)$ denotes the primitive root of X. This result was extended in [29] to ω-codes: any ω-code X has a unique primitive root $\rho(X)$ and $\mathcal{C}(X) = \rho(X)^*$. Also, as we prove in Chapter 6, if additionally, X is rational, then so is $\rho(X)$, and thus, also $\mathcal{C}(X)$. However, these results of [29] and [66] rely on some complex arguments: for prefix codes ([66]), one uses some involved combinatorial arguments, while for ω-codes ([29]), one relies on some results of Bergman and Cohn, [5], [16], [17], characterizing the commutation of two polynomials and of two formal power series, respectively.

Using the notion of branching point, we give here a simple, elementary solution for Conway's Problem in the case of finite biprefixes. We begin by proving that the centralizer of any biprefix set of words is necessarily of a very special form.

Theorem 4.5.3. *For any biprefix L, there is a set $T \subseteq \mathrm{Pref}(L)$ such that $\mathcal{C}(L) = L^* \sum_{t\in T} L^{k_t} t$, where $k_t \geq 0$, for all $t \in T$.*

Proof. By Corollary 4.3.3, we can assume without loss of generality that L is a branching set of words.

Let $x \in \mathcal{C}(L)$. By Lemma 4.3.1, $x \in \mathrm{Pref}(L^*)$ and thus, since L is a prefix, there are unique words $u_1, \ldots, u_k, t$, such that $x = u_1 \ldots u_k t$, with $u_i \in L$ and $t \in \mathrm{Pref}(L)$. Moreover, since L is branching, t is a branching point of L. Because L is a prefix code and $\mathcal{C}(L)L^k = L^k\mathcal{C}(L)$, it follows that $xL^k \subseteq L^k\mathcal{C}(L)$, and so $tL^k \subseteq \mathcal{C}(L)$. Similarly, since L is also a suffix, it follows that $L^k t \subseteq \mathcal{C}(L)$.

Let T be the set of all words t defined above, or more formally,

$$T = \{t \in (L^*)^{-1}\mathcal{C}(L) \mid u \not\leq t, \forall u \in L\}.$$

For any $t \in T$, let $k_t = \min\{k \geq 0 \mid L^k t \subseteq \mathcal{C}(L)\}$. We claim that

$$\mathcal{C}(L) = L^* \sum_{t\in T} \rho(L)^{k_t} t.$$

Clearly, by construction, $L^{k_t} t \subseteq \mathcal{C}(L)$, and so

$$L^* \sum_{t\in T} L^{k_t} t \subseteq \mathcal{C}(L),$$

since $\mathcal{C}(L)$ is closed under union and under multiplication by L.

For the reverse inclusion, let $x \in \mathcal{C}(L)$. Then, as shown above, there is $l \geq 1$ such that $x \in L^l t$, with $t \in T$. Since L is a biprefix, it follows as above that $L^l t \subseteq \mathcal{C}(L)$. Consequently, $l \geq k_t$ and so, $x \in L^* L^{k_t} t$, and the claim follows. □

Corollary 4.5.4. *The centralizer of any finite biprefix is rational.*

Proof. Let L be a finite biprefix. Using the notations in Theorem 4.5.3, it follows that T is finite and so, $\mathcal{C}(L)$ is rational. □

As a matter of fact, it is proved in [66] that the set T of Theorem 4.5.3 is always empty, for all prefix sets L. However, this result is proved in [66] using some lengthy, involved combinatorial arguments and we do not have at this point any simple proof of it. We extend the results of [66] and solve some conjectures of the same paper, using algebraic methods based on the commutation of two formal power series. We prove then that Conway's Problem has an affirmative answer for all rational ω-codes and rational reversed ω-codes. Moreover, in this case, and in particular for all prefix codes, we have $\mathcal{C}(X) = \rho(X)^*$.

4.5.3 Biprefix sets with at most one critical point

We consider next a simple case of finite biprefixes to further illustrate the branching point approach. Namely, we consider the case of finite biprefixes with at most one critical point.

Theorem 4.5.5. *Let L be a biprefix code. Then L has no critical points if and only if L is marked. Moreover, in this case, $\mathcal{C}(L) = L^*$.*

Proof. Let L be a biprefix code and assume that L has no critical points. If L is not marked, then there are two words u and v starting with the same letter. Thus, as L is a prefix, the longest common prefix of u and v is a critical point of L, a contradiction.

If L is marked, let us assume that L has a critical point x. Thus, $xa, xb \in \mathrm{Pref}(L^*)$, for distinct letters a and b, and $x \notin L^*$. Since L is a biprefix code, it follows that there are unique words $u_1, \ldots, u_m \in L$ and $t \in \mathrm{Pref}(L) \setminus L^*$ such that $x = u_1 \ldots u_m t$. Since $ta, tb \in \mathrm{Pref}(L^*)$, there are $\alpha, \beta \in L$ such that $ta \leq \alpha$ and $tb \leq \beta$. However, $t \neq 1$, and so, as L is marked, it follows that $\alpha = \beta$ and $a = b$, a contradiction.

If L is marked, then all points of $\mathcal{C}(L)$ are branching points. Since L has no critical points, it follows that $\mathcal{C}(L) = L^*$. □

Note that for any critical point u of a prefix code L, all words in $\rho(L)^* u$ are also critical points of L. We say that v is a *minimal* critical point of a code L if there is no critical point u of L such that $v \in \rho(L)^* u$.

Example 4.5.2. Let $F = \{aa, ab\}$. Then the set of critical points of F is F^*a. However, the only minimal critical point of F is a.

Theorem 4.5.6. *Let L be a biprefix with at most one minimal critical point. Then $\mathcal{C}(L) = L^*$.*

Proof. Similarly as in Corollary 4.3.3, we can assume without loss of generality that L is a branching set of words. If L has no critical point, then the claim follows by Theorem 4.5.5.

Assume that L has one minimal critical point. By Theorem 4.5.3, $\mathcal{C}(L) = L^*(1 + \sum_{t\in T} L^k t)$, for a set T of critical points of L. Since L has only one minimal critical point, $T \subseteq \{t\}$, with $t \in \mathrm{Pref}(L) \setminus \{1\}$. If $T = \emptyset$, then the claim follows. Assuming that $T = \{t\}$, we obtain $\mathcal{C}(L) = L^*(1 + L^k t)$, and so

$$L^*(1 + L^k t)L = LL^*(1 + L^k t). \tag{4.31}$$

We prove that L^*t commutes with L.

If $L^*L^k tL \cap LL^* \neq \emptyset$ then, as L is a prefix, it follows that $tL \cap L^* \neq \emptyset$. Since L is also a suffix, it follows that $t \in L^*$, a contradiction.

If $L^*L \cap LL^*L^k t \neq \emptyset$ then, as L is a prefix, it follows that $t \in L^*$, again a contradiction.

Consequently, it follows from (4.31) that $L^*L^k tL = LL^*L^k t$. Moreover, since L is a prefix, it follows that

$$L^*tL = LL^*t. \tag{4.32}$$

Since L has only one minimal critical point, there is a word $w \in L$ such that no other word of L starts with the same letter as w. This follows from the fact that L is branching and $|T| \leq 1$. Thus, t and w are prefix incomparable.

From (4.32), we obtain that $Lt \subseteq L^*tL$. In particular, $wt \in L^*tL$. Since t and w are prefix incomparable, and L is a prefix code, we obtain that $t \in L^*tL$, a contradiction.

Consequently, $\mathcal{C}(L) = L^*$, proving the claim. □

4.6 Discussion

We have investigated in this chapter the notion of centralizer of a set of words and the thirty years old problem of Conway, asking whether or not the centralizer of any rational language is rational. We believe that the answer to this question is affirmative, at least for finite sets, and we gave it a solution in this chapter for the special case of periodic sets, binary and ternary sets, as well as for finite biprefix codes. To prove these results, we have used three different approaches: combinatorial results on finite and infinite words, equations on languages, and the branching point approach.

Still another one is developed in the Chapter 6, in connection with the problem of characterizing the commutation of two sets of words. Using algebraic results on the commutation of two formal power series, we give an affirmative answer to Conway's Problem also in the case of rational ω-codes and rational reversed ω-codes, the most general known results on this problem, up to date.

We believe that Conway's Problem is a perfect example of a jewel of the Theory of Formal Languages: a natural question, defined in elementary terms, yet appearing to be amazingly difficult to solve. Despite its simple formulation, in terms of commutation and rational languages, we seem to lack the proper tools to solve it. The following observation may give a hint on the degree of difficulty of this problem, see also [36]. As it is easy to see, the commutation requires for any word z in the centralizer of X and any $\alpha \in X$, that there is $\beta \in X$ such that $\beta^{-1}(z\alpha) \in \mathcal{C}(X)$. This resembles a computation step in a *Post tag system*, see [64], [65], where one deletes a fixed length prefix of a word and appends some other word to it. It is known, see [54], that the Post tag systems have universal power of computation.

Many problems remain open in connection with Conway's Problem. The most striking one perhaps, is that, although the rationality of the centralizer is questioned in the original problem, it is not even known whether or not the centralizer of a rational set is always recursive. Furthermore, the question is open also for finite sets !

Problem 4.6.1. Is the centralizer of a rational (or finite) set of words always recursive ?

We have proved in Theorem 4.2.1 that for any recursive language, the complement of its centralizer is recursively enumerable. Thus, Problem 4.6.1 is equivalent with asking whether or not the centralizer of a rational (or finite) set of words is always recursively enumerable.

We have defined in this thesis the centralizer of a set of words X as the largest submonoid of Σ^* commuting with X. Alternatively, one can also define a notion of semigroup centralizer of X, as the largest subsemigroup of Σ^+ commuting with X, as done in [13], [39], or [40]. The following question comes then very natural and its answer is unknown up to date.

Problem 4.6.2. Is it true that for any rational set of words, the monoid and the semigroup centralizers coincide modulo a rational set ?

The relevance of this problem comes from Conway's Problem: we do not know yet whether or not the problem of the centralizer's rationality is essentially the same for monoid and for semigroup centralizers. We know however, that the monoid and the semigroup centralizers of any code (even a non-rational one) coincide modulo the empty word; this is a consequence of a result of [66].

Related to commutation and to the notion of centralizer, we have the following three decidability problems.

Problem 4.6.3. Decide whether or not, for a given rational language X, $\mathcal{COM}(X)$ contains a finite set. In other words, decide whether or not there is a finite set commuting with X.

Problem 4.6.4. Decide whether or not, for a given rational language X, $\mathcal{C}(X) = X^*$.

Problem 4.6.5. Decide, whether or not, for two given rational languages X and M, $\mathcal{C}(X) = M$.

The case of ternary sets is a particularly important one for Conway's Problem. We know that for any non-periodic binary set X, if $1 \notin X$, then $\mathcal{C}(X) = X^*$. On the other hand, this is not true for sets X with at least four words, as shown in Example 4.1.2, see also [13]. We conjecture that this property holds also for ternary sets.

Problem 4.6.6. Prove that for any nonperiodic ternary set of words X, if $1 \notin X$, then $\mathcal{C}(X) = X^*$.

We give in Chapter 6, Theorem 6.4.4, a solution to Problem 4.6.6 for the case of ternary codes. We also prove in Chapter 6 that a complete solution to Problem 4.6.6 implies that the commutation with ternary sets can be characterized similarly as in free monoids. We have been investigating several possible strategies for a complete solution and we mention here three of them.

We proved in Theorem 4.4.9 and the preceding three lemmata that the centralizer of any ternary set is rational. Moreover, we gave an algorithm how to compute the centralizer for any rational set and we proved that the centralizer of F always has a rational form, very close to F^*. One possibility to solve Problem 4.6.6 is to refine the language equations used in the proof of Lemma 4.4.8 and to conclude that $\mathcal{C}(F)$ is always F^*.

Another possible strategy to solve Problem 4.6.6 is to prove directly that $\mathcal{C}(F) \subseteq F^*$. Assuming the contrary, let x be a minimal word in $\mathcal{C}(F) \setminus F^*$. Using the fact that F is not a code (the case of codes is solved in Theorem 6.4.4) and the techniques of Lemma 4.4.3, one could be able to conclude using Graph Lemma that F must be periodic.

We also mention here the possibility of generalizing Lemma 4.4.3. We proved in that result that all the nonempty words in the centralizer start with a word in F. Proving that any word in $\mathcal{C}(F) \setminus (1 + F)$ starts with a word in F^2 and developing an induction step, clearly leads to a solution to Problem 4.6.6.

Based on the simple notions of branching and critical points, we have proposed the branching point approach to attack Conway's Problem. We

have demonstrated its usefulness by giving very simple solutions of the problem in the case of binary sets and finite biprefix sets. As a matter of fact, our result for biprefix sets, Corollary 4.5.4, can be easily extended to codes with bounded decoding delay in both directions, proving that also in this case, the centralizer has a simple, rational form.

We believe that the branching point approach can be used to derive also other results on Conway's problem, maybe even an affirmative answer for the case of all finite sets. To support this idea, let us denote by $T_{\mathcal{B}(L)}$ the tree of all words in $\mathcal{B}(L)$, for a rational language L. By Theorem 4.5.1, $\mathcal{B}(L)$ is rational, and so this tree - let us call it the *branching tree* of L - is of a regular type: it contains only a finite number of different (maximal) subtrees. For branching sets of words L, to which Conway's Problem can be reduced, all words in $\mathcal{C}(L)$ are branching and thus, they are nodes in $T_{\mathcal{B}(L)}$. Let $Z = \mathcal{C}(L) \setminus L^*$. Since $\mathcal{C}(L)$ is a monoid, for any $z \in Z$, $L^*z \cup z^* \cup zL^* \subseteq \mathcal{C}(L)$. Consequently, a single node of $T_{\mathcal{B}(L)}$ from Z determines many other nodes of $T_{\mathcal{B}(L)}$ to be in $\mathcal{C}(L)$. Thus, for a solution to Conway's Problem, an important question is: can one "saturate" $\mathcal{C}(L)$ within the nodes of $T_{\mathcal{B}(L)}$ in a finite number of steps of this type, at least for some types of (finite) sets X ? Intuitively, the regularity of $T_{\mathcal{B}(L)}$ supports this view.

Nevertheless, Conway's Problem remains open up to date. We conclude this chapter with its formulation.

Problem 4.6.7 (Conway's Problem). Is it true that for any rational language, its centralizer is rational ?

Chapter 5

Motifs and roots of sets of words

It is an elementary property of words that for each nonempty word u there is a unique minimal word v such that $u = v^m$, for some positive integer m; v is called the *primitive root* of u. Moreover, two words commute if and only if they have the same primitive root.

There are at least two different possibilities of defining a similar notion for sets of words and we will be investigating both of them in this chapter. The first one regards the root of a word as a pattern, or a *motif*, spanning the word. From this point of view, the notion of root extends to sets of words to the notion of *premotif*, as in Autebert et al. [3], 1989. A language R is said to be a premotif of a language L if $L = \bigcup_{i \in I} R^i$, for some $I \subseteq \mathbb{N}$. On the other hand, the root of a word can be viewed also as an arithmetic, or an algebraic notion, as in the case of numbers, polynomials, monoids, etc. We thus obtain the notion of *root* of a language, as in Shyr [73], 1983. A language R is said to be a root of a language L if $L = R^n$, for some $n \in \mathbb{N}$.

As it turns out, none of these notions ensure a desired uniqueness property for the premotif/root of all sets of words. Indeed, there are languages having several "minimal" premotifs and several "minimal" roots. The situation is different for codes. Not only that the two notions coincide in this case, but we prove that in fact, any code has a unique primitive root, as in the case of words, thus solving a conjecture dating back to 1989, due to Ratoandromanana, [66]. Moreover, using the *multiplicity approach*, we prove that two codes commute if and only if they have the same primitive root, solving another old open problem of [66]. These results are in fact somewhat surprising since the set of codes is not a free semigroup, see [7].

The result of this chapter are based, among other references, on our results in [29].

5.1 Motifs and premotifs

The notions of motif and premotif have been introduced by Autebert et al. [3], motivated by investigations on the commutation of two languages, and the BTC-property, discussed in Details in Chapter 6 of this thesis. Essentially, they generalize to sets of words the notions of root and primitive root of words.

Definition 5. *For two sets of words L and R, we say that R is a premotif of L if*

$$L = \bigcup_{i \in I} R^i,$$

for some $I \subseteq \mathbb{N}$.

Example 5.1.1. (i) For any language $L \subseteq a^*$, a is a premotif of L.

(ii) Both a and a^2 are premotifs of $(a^2)^*$.

(iii) a^*ba^* is a premotif of $a^*ba^*ba^*$.

It follows immediately from the definition that we can always consider premotifs R not containing the empty word. Indeed, if $1 \in R$ and $L = \cup_{i \in I} R^i$, then $R \setminus \{1\}$ is also a premotif of L: $L = \cup_{I \cup \{0\}} (R \setminus \{1\})^i$.

We then have the following notion of primitive language, see [3].

Definition 6. *A set of words $R \subseteq \Sigma^+$ is said to be primitive if for any set of words X, $R = \cup_{i \in I} X^i$ implies that $I = \{1\}$.*

Note that the requirement in the above definition of a primitive language is not equivalent with asking that the language has no premotifs other than itself. E.g., for any star language L, we have $L = L^n$, for all $n \geq 1$, and thus, L is not primitive. For languages $L \subseteq \Sigma^+$ however, the two conditions are equivalent. We refer to [3] for a proof of this result.

Lemma 5.1.1 ([3]). *A language $L \subseteq \Sigma^+$ is primitive if and only if it has no premotifs other than itself.*

The notion of motif of a language, the equivalent of the primitive root of a word, is defined as follows.

Definition 7. *A language $R \subseteq \Sigma^+$ is a motif of a language L if R is a primitive premotif of L.*

Example 5.1.2. (i) For any $L \subseteq a^*$, a is a motif of L. Thus, L is primitive if and only if $L = \{a\}$.

(ii) For any elementary code C, C is a motif of C^*.

Two desired properties for any root-like notion, in this case, the notion of motif, are clearly the existence and the uniqueness of that notion for any language. For the existence part, it is proved in [3] that indeed, any language has a motif. We sketch the proof in the following.

Let L be an arbitrary language over the alphabet Σ. We can assume without loss of generality that $1 \notin L$ since otherwise, any motif of $L \setminus \{1\}$ is a motif of L. Consider an arbitrary total order $\preceq$ on the alphabet Σ, i.e., for any two distinct letters a, b, either $a \preceq b$, or $b \preceq a$. We define the *alphabetical order* $\preceq$ of Σ^* as follows: for any two words $u \neq v$, we say that $u \preceq v$ if and only if

(i) $|u| < |v|$, or

(ii) $|u| = |v|$, $u = tau'$, and $v = tbv'$, for some words $t, u', v' \in \Sigma^*$ and some letters $a, b \in \Sigma$, such that $a \preceq b$.

Let $\Sigma^* = \{u_0, u_1, u_2, \ldots\}$ be the enumeration of Σ^* in the alphabetical order defined above. For any word $v \in \Sigma^+$, let $\mathrm{Inf}(v) = \{u \in \Sigma^+ \mid u \preceq v\}$.

We define a sequence of languages converging to a motif of L, as follows. Let u_{k_0} be the shortest word present in a premotif of L and let $M_0 = \{u_{k_0}\}$. Starting with u_{k_0}, we examine each word of Σ^* in the alphabetical order, determining whether or not that word is present in all premotifs of L. We add the word to the next language in the sequence if and only if this is the case. Formally, let

$$M_n = \begin{cases} M_{n-1}, & \text{if there is a premotif } P \text{ of } L \text{ such that} \\ & \quad P \cap \mathrm{Inf}(u_{k_0+n}) = M_{n-1} \\ M_{n-1} \cup \{u_{k_0+n}\}, & \text{otherwise.} \end{cases}$$

For each $n \geq 0$, let $I_n = \{i \in \mathbb{N} \mid M_n^i \subseteq L\}$. Denote $M = \cup_{n \geq 0} M_n$ and $I = \cap_{n \geq 0} I_n$. Then it is straightforward, though involved, to prove that M is a motif of L: $L = \cup_{i \in I} M^i$ and M has no premotifs other than itself. We refer for details to [3].

Theorem 5.1.2 ([3]). *Any language has a motif.*

Regarding the uniqueness of the motif for any set of words, it turns out that there are languages having several distinct motifs. Moreover, this is true even for finite languages, as shown in the next example.

Example 5.1.3. (i) For any alphabet Σ with $|\Sigma| \geq 2$, Σ^* has the following motifs:

- Σ is a motif of Σ^*: $\Sigma^* = \cup_{n \in \mathbb{N}} \Sigma^n$.

- for any $u \in \Sigma^*$, $\Sigma \cup \{u\}$ is a motif of Σ^*: $\Sigma^* = \cup_{n \in \mathbb{N}} (\Sigma \cup \{u\})^n$.

- if Prim is the set of all primitive words over Σ, then Prim is a motif of Σ^* as well: $\Sigma^* = \cup_{n\in\mathbb{N}} \text{Prim}^n$. Indeed, any nonempty word has a unique primitive root.

(ii, [3]) There are also finite sets of words having multiple motifs. E.g., let

$$F = a + a^2 + \{a^i b a^j \mid 0 \leq i, j \leq 4,\ i + j \leq 7\}.$$

Also, let

$$M_1 = a + \{a^i b a^j \mid 0 \leq i, j \leq 3\} \setminus \{aba^2\}, \quad \text{and}$$
$$M_2 = a + \{a^i b a^j \mid 0 \leq i, j \leq 3\} \setminus \{a^2 ba\}.$$

Both M_1 and M_2 are motifs of F, since $F = M_1 + M_1^2 = M_2 + M_2^2$.

(iii, [3]) There are also languages $L \subseteq \Sigma^+$ having an infinite number of motifs. E.g., let

$$L = \{b\} \cup \bigcup_{n\geq 4} \Sigma^n \quad \text{and} \quad P_k = \{b\} \cup \bigcup_{4\leq n\leq 15+k} \Sigma^n,$$

for all $k \geq 0$. Then P_k is a premotif of L, for all $k \geq 0$. Moreover, if k is not a multiple of 15, then P_k is a motif of L.

5.2 Roots and primitiveness

We consider next the notion of root of a language, introduced in [73] similarly as for words, numbers, polynomials, etc. We also consider a (different) notion of primitive language, defined similarly as for words.

Definition 8. *We say that a language R is a root of L if $L = R^n$, for some positive integer n. We say that a language L is primitive if for any root R of L, $L = R^n$, we have $n = 1$ and $L = R$.*

Example 5.2.1. (i) Any language consisting of primitive words only is a primitive language. In particular, the set of all primitive words, Prim, is a primitive language. However, the reverse is not true. E.g., $L = \{a^2, b^2\}$ is a primitive language.

(ii) For any alphabet Σ with $|\Sigma| \geq 2$ and any $u \in \Sigma^*$, $\Sigma \cup \{u\}$ is a primitive language.

(iii) Unlike for motifs, there are primitive languages containing the empty word. E.g., $\{1, a, b\}$ is a primitive language.

Note that the Definitions 6 and 8 introduce different notions of primitive language. E.g., for any alphabet Σ, $\Sigma \cup \Sigma^2$ is a primitive language by Definition 8, but not by Definition 6. On the other hand, any language

which is primitive by Definition 6 is primitive also in terms of Definition 8. We use however the same term of primitive, without risk of confusion. Most often, we consider the notion of primitive language in connection with that of root, and for this we do have different terminology: motifs and primitive roots.

Similarly as in the case of motifs, note that it is not enough to ask in the definition of a primitive language that it has no root other than itself. E.g., for any star language L, we have $L = L^n$, for all $n \geq 1$. However, this condition is true for all languages $L \subseteq \Sigma^+$.

Lemma 5.2.1. *A language $L \subseteq \Sigma^+$ is primitive if and only if L has no root other than itself.*

Proof. Let L be a primitive language and R a root of L, $L = R^n$. It then follows that $n = 1$, and so, $L = R$.

Conversely, let L be a language having no roots other than L itself. Let $n \geq 1$ be such that $L = L^n$. Denoting l the length of a minimal length word in L, it follows that $l = nl$, i.e., $l = 0$. Thus, $1 \in L$, contradicting the hypothesis that $L \subseteq \Sigma^+$. □

The following lemma gives a simple way to construct primitive languages.

Lemma 5.2.2. *Let $L \subseteq \Sigma^+$ be an arbitrary language. If $L \cap \Sigma \neq \emptyset$, then L is primitive.*

Proof. If L is not primitive, then $L = R^n$, for some language R and a positive integer $n \geq 2$. Let l be the length of a minimal word in R. Since $\Sigma \cap L \neq \emptyset$ and $1 \notin L$, it follows that the length of a minimal word in L is 1 and so, $1 = nl$. Consequently, $n = 1$, a contradiction. □

Corollary 5.2.3. *Let $L \subseteq \Sigma^+$ be an arbitrary language and $a \in \Sigma$. Then $L \cup \{a\}$ is primitive.*

Note that the above corollary is the generalization to sets of words of the following property of words: for any $u \in \Sigma^*$, there is $a \in \Sigma$ such that ua is primitive, see [47].

We prove in the next result that all languages have a primitive root. It is interesting to note that the proof of this result is very similar to the proof of the corresponding result for words, although the set of all languages over a given alphabet is not a free monoid.

Theorem 5.2.4. *Any nonempty language has a primitive root.*

Proof. Let L be a nonempty language. To begin with, assume that $1 \notin L$.

If L is primitive, then it is its own primitive root. Otherwise, let R_1 be a nontrivial root of L, $L = R_1^{n_1}$, $n_1 \geq 2$. If R_1 is not primitive, then let R_2 be a nontrivial root of R_1, $R_1 = R_2^{n_2}$, $n_2 \geq 2$. Thus, $L = R_2^{n_1 n_2}$. If R_2 is

not primitive, then we consider a root R_3 of R_2, etc. Observe now that this procedure cannot continue indefinitely, since for any root R of L, if $L = R^n$, then n is a divisor of the length of minimal words in L. Thus, there is $i \geq 1$ such that R_i obtained as above is a primitive language, and consequently, a primitive root of L.

Note that if $1 \in L$, then the above considerations do not necessarily yield a primitive root of L in a finite number of steps. Let L be a language such that $1 \in L$ and let $L' = L \setminus \{1\}$ and $B = L' \setminus (L'L'^{+})$. Clearly, for any root R of L, we have $B \subseteq R \subseteq L$.

If $L' = L'^{+}$, then $R = \{1\} \cup B(B^2)^*$ is a primitive root of L: $L = R^2$. Indeed, if $R = S^n$, $n \geq 2$, then necessarily, $\{1\} \cup B \subseteq S$ and so, $B^2 \subseteq R$, a contradiction.

If $L' \neq L'^{+}$, then there is $n_0 > 1$ such that $B^{n_0} \not\subseteq L$. Thus, for any root K of L, $L = K^t$, it follows that $t < n_0$ and one can find a primitive root of L in at most n_0 steps as above. □

It turns out that, similarly as for motifs, there are languages, even finite ones, having several primitive roots.

Example 5.2.2. (i,[12]) The set

$$X = \{a^n \mid 0 \leq n \leq 30, n \neq 1, 8, 11, 23\}$$

has two distinct primitive roots. Indeed, if

$$Y = \{a^n \mid n = 0, 2, 3, 7, 10, 12, 14, 15\} \quad \text{and}$$
$$Z = \{a^n \mid n = 0, 2, 3, 7, 12, 13, 14, 15\},$$

then $X = Y^2 = Z^2$ and both Y and Z are primitive languages.

(ii,[73]) There are languages $L \subseteq \Sigma^+$ having multiple primitive roots. E.g., the set $X = \{a^n \mid 2 \leq n \leq 10\}$ has two distinct primitive roots: $X = Y^2 = Z^2$, where $Y = \{a, a^2, a^4, a^5\}$ and $Z = \{a, a^2, a^3, a^4, a^5\}$.

(iii) Let $L = \bigcup_{n \geq 4} \Sigma^n$, for some alphabet Σ. Then L has an infinite number of primitive roots. Indeed, let $L_n = \Sigma^+ \setminus \Sigma^n$, for all $n \geq 3$ and observe that $L = L_n^4$. Also, by Lemma 5.2.2, L_n is primitive, since $\Sigma \subseteq L_n$, for all $n \geq 3$.

5.3 Decidability results

We consider in this section some natural decidability questions regarding the notions of motif and primitive root: is it decidable if a language is primitive, is it decidable if a language is a root/premotif of another one ?

We prove first that any rational language has a rational root. For this, we use the notion of *syntactical monoid*, see [25], and the techniques of [12]. Note that a similar result is proved in [28] using elementary techniques of finite automata.

We prove in fact that any root of a rational language can be enlarged to a rational root of that language. This is usually referred to as the *saturation method*. A general form of this result is given in [12], proving that the result holds with any rational operation instead that of exponentiation.

Lemma 5.3.1. *Let R be a rational language. If L is a root of R, $R = L^n$, with $n \geq 1$, then there is a rational root S of R such that $L \subseteq S$ and $R = S^n$.*

Proof. Let M be the syntactical monoid of R and $\phi : \Sigma^* \to M$ its syntactical morphism. Note that M is finite and $R = \phi^{-1}(\phi(R))$.

Let $F = \phi(L) \subseteq M$ and $\mathcal{M} = \{\phi(X) \mid X \subseteq \Sigma^*\}$. Also, let G be a maximal element of the set $\{Y \in \mathcal{M} \mid F \subseteq Y, F^n = Y^n\}$. We denote $S = \phi^{-1}(G) \subseteq \Sigma^*$.

Observe first that S is a rational language. Indeed, $\phi(S) = G$, and so, $S = \phi^{-1}(\phi(S))$. We prove that $L \subseteq S$ and $R = S^n$.

Clearly, $\phi(L) = F \subseteq G$ and so, $L \subseteq \phi^{-1}(\phi(L)) \subseteq \phi^{-1}(G) = S$. Thus, $R = L^n \subseteq S^n$.

To prove the reverse inclusion, note that

$$\phi(S^n) = \phi(S)^n = G^n = F^n = \phi(L)^n = \phi(L^n) = \phi(R).$$

Consequently, $S^n \subseteq \phi^{-1}(\phi(S^n)) = \phi^{-1}(\phi(R)) = R$ and so, it follows that $R = S^n$, i.e., S is a rational root of R. □

Using the techniques of Lemma 5.3.1, we prove now that it is decidable whether or not a rational language is primitive.

Theorem 5.3.2. *It is decidable whether or not a rational language R is primitive. Moreover, if $R \subseteq \Sigma^+$ is not primitive, then a primitive rational root of R can be effectively found.*

Proof. Let $\mathcal{M}$ be the syntactical monoid of R and ϕ its syntactical morphism, $\phi^{-1}(\phi(R)) = R$. As we proved in Lemma 5.3.1, to decide whether or not R admits a nontrivial root, one only needs to test the sets $S = \phi^{-1}(T)$, with $T \subseteq \mathcal{M}$. Since $\mathcal{M}$ is finite, there is a finite number of candidates S. We need however an upper limit on the exponent n such that $R = S^n$. Clearly, if $1 \notin R$, then n is a divisor of the length of a minimal word of R and thus, upper bounded. If $1 \in R$, then this argument does not hold anymore.

Let $p = |\mathcal{M}|$. We prove that if R is not primitive, then there is a rational root S of R, such that $R = S^n$, with $n \leq 2^p$.

Assume the contrary: the smallest exponent n such that $R = S^n$ is $n \geq 2^p + 1$. But then, since $\phi(S^i) \subseteq \mathcal{M}$, for all $1 \leq i \leq n$, it follows that there are r, t, $1 \leq r < t \leq n$ such that

$$\phi(S^r) = \phi(S^t).$$

Thus, $\phi(S^{r+i}) = \phi(S^{t+i})$, for all $i \geq 0$. In particular,

$$\phi(S^{n-t+r}) = \phi(S^n) = \phi(R),$$

and so, $S^{n-t+r} \subseteq \phi^{-1}(\phi(S^{n-t+r})) = \phi^{-1}(\phi(R)) = R = S^n$.

We claim that $1 \in S$. Indeed, if $1 \notin S$, then let u be a minimal length word in S. Then, since $S^{n-t+r} \subseteq S^n$, it follows that $(n-t+r)|u| \geq n|u|$, i.e., $|u| = 0$, a contradiction.

Consequently, since $R = S^n$ and $1 \in S$, it follows that $S^i \subseteq R$, and so, $\phi(S^i) \subseteq \phi(R)$, for all $0 \leq i \leq n$. Note also that $\phi(S^k) \in \{\phi(S^i) \mid 0 \leq i \leq t\}$, for all $k \geq 0$. Thus, $\phi(S^k) \subseteq \phi(R)$, for all $k \geq 0$. This implies that

$$S^k \subseteq \phi^{-1}(\phi(S^k)) \subseteq \phi^{-1}(\phi(R)) = R,$$

for all $k \geq 0$, i.e., $S^* \subseteq R$. But $R = S^n$, and so, $R = S^*$. In particular, we obtain that $R = R^n$, for all $n \geq 1$, contradicting our assumption.

Thus, to test whether or not R is primitive, we need to test the equality $R = S^n$, for all $S = \phi^{-1}(T)$, $T \subseteq \mathcal{M}$, and $n \leq 2^p$, proving the first claim of the theorem.

Assume now that $R \subseteq \Sigma^+$ is not primitive. In this case, the above procedure finds a rational root S of R, say $R = S^n$, for some $n \geq 2$. If S is not primitive, then we repeat the procedure with S instead of R and find a rational root T of S, $S = T^m$, $m \geq 2$. Thus, $R = T^{mn}$, and we continue the same procedure with T instead of S. Clearly, the procedure ends in a finite number of steps with a primitive rational root of R, since for any root X of R, if $R = X^p$, then p divides the length of the minimal words in R . □

Corollary 5.3.3. *It is decidable whether or not a rational language $R \subseteq \Sigma^+$ is a root of another rational language.*

Proof. If $S = R^n$, then n is a divisor of the length of a minimal length word of S. Thus, one only has a finite number of possible values for n to be tested. □

Similar results can be proved also for premotifs of rational languages. In this case, we choose to use techniques of finite automata, similar to those of [28].

Lemma 5.3.4. *Let R be a rational language. If L is a premotif of R, $R = \bigcup_{i \in I} L^i$, then there is a rational motif S of R such that $L \subseteq S$ and $R = \bigcup_{i \in I} S^i$.*

Proof. We can assume without loss of generality that $1 \notin R$. Indeed, any motif of $R \setminus \{1\}$ is a motif of R.

Let L be a premotif of R, $R = \cup_{i \in I} L^i$, where $I \subseteq \mathbb{N}$. Let $\mathcal{A}$ be the minimal automaton of R, with Q the set of states, q_0 its initial state, F the set of final states, and δ its transition function.

For each $n \geq 0$, we define the set of states

$$P_n = \delta(q_0, L^n),$$

and for each $n \geq 1$ we define the following sets of words:

$$L_n = \bigcap_{q \in P_{n-1}} \mathcal{A}(q, P_n) \quad \text{and} \quad S_n = \bigcap_{i=1}^{n} L_i,$$

where $\mathcal{A}(q, P_n)$ is the set of words accepted by $\mathcal{A}$ with the initial state q and the set of final states P_n.

Note that there is only a finite number of sets P_n, since $P_n \subseteq Q$ and moreover, $P_n \subseteq F$, for all $n \in I$. Consequently, the sequences $(L_n)_{n \geq 1}$ and $(S_n)_{n \geq 1}$ contain only a finite number of different terms.

Let $S = \bigcap_{i \in I} S_i$. We prove that S is rational, $L \subseteq S$, and $R = S^I$.

Since the sequence $(S_n)_{n \geq 1}$ has only a finite number of distinct terms, S is a finite intersection of rational languages and so, it is rational.

To prove that $L \subseteq S$, it is enough to prove that $L \subseteq L_n$, for all $n \geq 1$. For $n = 1$, this follows from the definition of L_1 and P_1. Let us assume that the claim holds up to $n - 1$, for some $n \geq 2$. Let $u \in L$ and let $q \in P_{n-1}$. There is a word $v \in L^{n-1}$ such that $\delta(q_0, v) = q$. Also, since $vu \in L^n$, it follows that $\delta(q_0, vu) \in P_n$, and so, $\delta(q, u) \in P_n$. Thus, for any $q \in P_{n-1}$, $\delta(q, u) \in P_n$. It follows then that $u \in \bigcap_{q \in P_{n-1}} \mathcal{A}(q, P_n) = L_n$, proving the second claim.

Finally to prove that $R = S^I$, we note that

$$R = L^I = \cup_{i \in I} L^i \subseteq \cup_{i \in I} S^i = S^I.$$

For the reverse inclusion, let $n \in I$ and let $u_1, \ldots, u_n \in S$. Then $u_j \in L_j$, for all $1 \leq j \leq n$, and so, there are states $q_k \in P_k$, for all $1 \leq k \leq n$ such that $\delta(q_0, u_1 u_2 \ldots u_k) = q_k$. In particular, $\delta(q_0, u_1 u_2 \ldots u_n) \in P_n \subseteq F$. Thus, $u_1 u_2 \ldots u_n \in R$ and so, $L^n \subseteq R$, for all $n \in I$.

Consequently, S is a rational premotif of R, $R = S^I$. Iterating the argument with S instead of R we obtain in a finite number of steps a rational motif of R. □

By Lemma 5.3.4, one can construct a rational motif of a rational language R, having a premotif of R. Some rather involved arguments are used in [3] to prove that in fact, one can always construct a rational motif of R.

Theorem 5.3.5 ([3]). *For any rational language $R \subseteq \Sigma^+$, one can construct a rational motif of R.*

Corollary 5.3.6. *It is decidable whether or not a rational language is primitive.*

Proof. Let R be a rational language. By Theorem 5.3.5, we can construct a rational motif S of R. Then R is primitive if and only if $R = S$, which can be easily decided. □

Note however that Theorems 5.3.2 and 5.3.5 do not hold for context-free languages, as shown in the next results. Our proofs follows the techniques used in [28] to prove that it is undecidable whether or not a context-free language is a square language.

Theorem 5.3.7. *It is undecidable whether or not a context-free language is primitive.*

Proof. Let L be an arbitrary context-free language over the alphabet Σ and let $K = aLaa\Sigma^*a \cup a\Sigma^*aa\Sigma^*a$, where a is a new letter, $a \notin \Sigma^*$. We prove that $L = \Sigma^*$ if and only if K is not primitive. Then the claim of the theorem follows since it is undecidable whether or not a context-free language equals the total language.

If $L = \Sigma^*$, then clearly, $K = (a\Sigma^*a)^2$.

Assume now that K is not primitive and let R be a non-trivial root of K, $K = R^n$, $n \in \mathbb{N}$. Since all words in K start with a and end with another a, the same must hold also for the words of R. Moreover, since $a^4 \in K$, it follows that $n = 2$. Consequently, $R = a\Sigma^*a$ and so, $L = \Sigma^*$. □

Theorem 5.3.8. *It is undecidable whether or not a context-free language admits a nontrivial premotif.*

Proof. Let L be an arbitrary context-free language over the alphabet Σ and let $K = aLaa\Sigma^*a \cup a\Sigma^*aa\Sigma^*a$, where a is a new letter, $a \notin \Sigma^*$. We prove that $L = \Sigma^*$ if and only if K is not primitive. Then the claim of the theorem follows as in the proof of Theorem 5.3.7.

If $L = \Sigma^*$, then clearly, $K = (a\Sigma^*a)^2$.

Assume now that K is not primitive and let R be a non-trivial root of K, $K = \cup_{i\in I} R^i$, $I \subseteq \mathbb{N}$, $I \neq \{1\}$. Since all words in K start with a and end with another a, the same must hold also for the words of R. Moreover, since all words of K have exactly four a-s, it follows that $I \subseteq \{1, 2\}$.

If $1 \in I$, then there are words in R having four a-s and so, $2 \notin I$. Thus, either $I = \{1\}$, or $I = \{2\}$. Since R is a non-trivial root of K, it follows that $I = \{2\}$, i.e., $K = R^2$. Consequently, $R = a\Sigma^*a$ and so, $L = \Sigma^*$. □

5.4 Two conjectures on roots and codes

As we discussed already, the two notions of root and premotif are different in general. In fact, any root of a language is also a premotif of it, but not the other way around. It turns out however that for codes, the two notions coincide. Consequently, we deal in this case with only one notion of root and one notion of primitive root of a language.

Lemma 5.4.1. *Let C be a code. A language R is a root of C if and only if R is a premotif of C.*

Proof. Clearly, if R is a root of C, then R is also a premotif of it. Reversely, let R be a premotif of C, $C = \bigcup_{i \in I} R^i$, $I \subseteq \mathbb{N}$. Since C is a code, it follows that I is a singleton, $I = \{n\}$, and so, $C = R^n$. □

It is well-known, see [7], that the set of codes is not a semigroup. E.g., $\{a, ab\}$ and $\{b, ab\}$ are codes but their product $\{ab, aab, abb, abab\}$ is not. For this reason, it is perhaps surprising that a central, elementary property of words (and of elements of any free semigroup) has been conjectured to hold also for codes. Indeed, it is a conjecture of Ratoandromanana [66] that any code has a unique primitive root.

Conjecture 1 ([66]). Any code has a unique primitive root.

The resemblance with the behavior of words goes further. It is still Ratoandromanana who conjectures in [66] that two codes commute if and only if they have the same primitive root.

Conjecture 2 ([66]). Two codes commute if and only if they have the same primitive root.

We discuss these conjectures in details in the next section.

5.5 Primitive roots of codes

We give in this section solutions to both conjectures above. As it turns out, these conjectures are in fact, equivalent. However, to give them a solution, we first give a partial solution to Conjecture 2, then prove Conjecture 1, from which Conjecture 2 follows in its general form. Some auxiliary results are needed first.

Lemma 5.5.1 ([66]). *Two codes X and Y commute if and only if there are some positive integers i and j such that $X^i = Y^j$. Moreover, XY and YX are non-ambiguous products.*

Lemma 5.5.2. *A language L is a code if and only if L^m is a code for some $m \geq 1$.*

Proof. Clearly, if L is a code, then L^m is also a code, for all $m \geq 1$, since any relation on L^m is a relation on L. Let us assume that there is a non-code L, such that L^m is a code, for some $m \geq 2$. Thus, there is a nontrivial relation $\alpha = \beta$ over L. Clearly, for any $x, y \in L^*$, $x\alpha\beta y = x\beta\alpha y$. Consider then words $x \in L^{m-1}$ and $y \in L^*$, such that $x\alpha\beta y, x\beta\alpha y \in L^{im}$, for some $i \geq 1$. Since L^m is a code, we obtain that α and β start with the same word from L, which is impossible since $\alpha = \beta$ is a nontrivial relation over L. □

Corollary 5.5.3. *If C is a code, then any root of C is also a code.*

We also recall here Cohn's theorem, [16], characterizing the commutation of two formal power series. We refer to Chapter 6 for a more detailed discussion on this result and on the multiplicity approach.

Theorem 5.5.4 (Cohn's Theorem, [16, 17]). *Let K be a commutative field and Σ a finite alphabet. Two formal power series $p, q \in K\langle\langle\Sigma^*\rangle\rangle$ commute if and only if there is a formal power series $r \in K\langle\langle\Sigma^*\rangle\rangle$ and two polynomials $p', q' \in K[[t]]$, such that $p = p'(r)$ and $q = q'(r)$, for a single variable t.*

We give in the next theorem a first characterization for the commutation of codes. Note that this does not solve yet Conjecture 2 but it provides an important tool for its proof.

Theorem 5.5.5. *Two codes X and Y commute if and only if they have a common root.*

Proof. Clearly, the condition of the theorem is sufficient for the commutation of two codes. To prove that it is also necessary, let X and Y be two codes such that $XY = YX$, and let r_X and r_Y be their characteristic formal power series, respectively. By Lemma 5.5.1, the products XY and YX are unambiguous, and so, the series products $r_X \cdot r_Y$ and $r_Y \cdot r_X$ are both unambiguous. Consequently, $r_X r_Y = r_Y r_X$. Thus, by Cohn's Theorem, it follows that there is a series r such that r_X and r_Y are combinations of r. If R is the support of r, this implies that $X = \cup_{i\in I} R^i$ and $Y = \cup_{j\in J} R^j$, for some $I, J \subseteq \mathbb{N}$. Since X and Y are codes, we obtain that both I and J are singletons, proving the claim. □

The next result shows that Conjectures 1 and 2 are equivalent.

Theorem 5.5.6. *Conjectures 1 and 2 are equivalent.*

Proof. If any code has a unique primitive root, then, by Theorem 5.5.5, Conjecture 2 holds. Reversely, assume that there is a code C having two primitive roots X and Y: $C = X^m = Y^n$. Then, by Corollary 5.5.3, both X and Y are codes, and by Lemma 5.5.1, $XY = YX$. Then, since Conjecture 2 holds and X, Y are primitive, it follows that $X = Y$, proving Conjecture 1. □

We can solve now Conjectures 1 and 2. Using Theorem 5.5.5, which was a partial solution to Conjecture 2, we solve next Conjecture 1. Conjecture 2 then follows by Theorem 5.5.6. In particular, note that our proof of Conjecture 1 for codes is highly similar to the proof of the corresponding result for words.

Theorem 5.5.7 (a solution to Conjecture 1). *Any code has a unique primitive root.*

Proof. Assume that there is a code C having two primitive roots X and Y. Thus, there are some positive integers m, n such that $C = X^m = Y^n$. By Corollary 5.5.3, both X and Y are codes, and by Lemma 5.5.1, $XY = YX$. Consequently, by Theorem 5.5.5, there is a code R such that $X = R^i$ and $Y = R^j$, for some positive integers i, j. However, since X and Y are primitive, we have that $i = j = 1$, i.e., $X = Y$. □

Definition 9. *For a code C, we denote by $\rho(C)$ its unique primitive root.*

Example 5.5.1. (i) For any alphabet Σ and any $n \geq 1$, the primitive root of Σ^n is Σ.

(ii) Theorem 5.5.7 is not valid in general for non-codes, as seen in Example 5.2.2. There are however, non-codes having unique primitive roots. E.g., for any nonperiodic binary or ternary set F, F^n has a unique primitive root, namely F, for all $n \geq 1$.

By Theorems 5.5.6 and 5.5.7, we also obtain a solution to Conjecture 2.

Theorem 5.5.8 (a solution to Conjecture 2). *Two codes commute if and only if they have the same primitive root.*

5.6 Discussion

The notion of primitive root of a word can be naturally extended in two different ways to sets of words, to the notion of premotif, as in Autebert et al. [3], 1989, and to the notion of root, as in Shyr [73], 1983. It turns out however, that for codes, these two notions coincide. Moreover, using the multiplicity approach, we prove that any code has a unique primitive root, just as words have. It also follows that two codes commute if and only if they have the same primitive root, another resemblance with the behaviour of words. Both these results were conjectured by Ratoandromanana, 1989, in [66].

It turns out that the notion of primitive root plays a central role in the commutation of languages, similarly as in the case of words. In the next chapter, we discuss in more details the multiplicity approach and establish an interesting connection between the notions of primitive root and centralizer of a language.

Chapter 6

The commutation of two sets of words

Characterizing the commutation of two sets of words is in general a very difficult problem. It is well-known that two words commute if and only if they are powers of another word, and in general, this property holds in any free monoid. The commutation of two multisets or of two finite multisets with coefficients in a commutative field can be described in similar terms: two (finite) multisets commute if and only if they are unions of powers of another (finite) multiset. However, nothing similar holds for the commutation of two arbitrary sets of words, unless one of them satisfies some special properties, e.g., it is a periodic or a binary set of words. Despite the fact that the set of codes is not a free monoid, it has been long conjectured in Ratoandromanana, 1989 [66], that such a characterization holds for the commutation with a code, as well as for the commutation of two codes. We solved completely the latter conjecture, regarding the commutation of two codes, in Chapter 5 of this thesis, and we investigate the former one in Chapter 6. We give here a partial solution to this problem, proving that it holds for all ω-codes and for all reversed ω-codes, the most general known results on this problem, up to date.

Two different approaches are used in this chapter. Using combinatorial properties of words, we characterize the commutation with periodic and binary sets of words, as well as the commutation with a ternary code. Using the multiplicity approach, already briefly mentioned in Chapter 5, we also characterize the commutation with ω-codes and reversed ω-codes, thus obtaining the most general result up to date on the commutation of two sets of words. As a matter of fact, all the other known results on this problem follow as simple consequences of our result.

We also establish a connection to Conway's problem, discussed in Chapter 4. We prove that Conway's question has an affirmative answer for rational ω-codes and for rational reversed ω-codes. Moreover, for such a code X,

we prove that the centralizer of X is $\rho(X)^*$, where $\rho(X)$ denotes the unique primitive root of X.

As we mentioned also in Chapter 4, the case of ternary sets is very interesting for the commutation, and a boundary point in many respects. Using the combinatorial approach, we characterize the commutation with ternary codes similarly as in free monoids and prove that extending this result to all ternary sets is equivalent with proving that $\mathcal{C}(F) = F^*$, for all ternary nonperiodic sets F. This, in turn, remains one of the most interesting open problems in the area. Some other open problems and research directions are discussed at the end of this chapter.

Among other references, the results of this chapter are based on our papers [29], [37], [38], and [39].

6.1 Characterizing the commutation

It is an elementary property of words that two words commute if and only if they have a common root, i.e., they are both powers of another word. As a matter of fact, this property holds in general for any free monoid, see, e.g., [73].

Theorem 6.1.1. *Let M be a free monoid. Two elements $u, v \in M$ commute if and only if there is $w \in M$ and $m, n \in \mathbb{N}$ such that $u = w^m$ and $v = w^n$.*

It has been long known that the commutation of two polynomials and the commutation of two formal power series in noncommuting variables, with coefficients in a field can be characterized similarly as for words. These results are due to Bergman, 1969 [5], and Cohn, 1962 [16], respectively.

Theorem 6.1.2 (Bergman's Theorem, [5]). *Let K be a commutative field and Σ a finite alphabet. Two polynomials $p, q \in K\langle\Sigma^*\rangle$ commute if and only if there are polynomials $r \in K\langle\Sigma^*\rangle$ and $p', q' \in K[t]$ such that $p = p'(r)$ and $q = q'(r)$, for a single variable t.*

Theorem 6.1.3 (Cohn's Theorem, [16, 17]). *Let K be a commutative field and Σ a finite alphabet. Two formal power series $p, q \in K\langle\langle\Sigma^*\rangle\rangle$ commute if and only if there are some formal power series $r \in K\langle\langle\Sigma^*\rangle\rangle$ and $p', q' \in K[[t]]$ such that $p = p'(r)$ and $q = q'(r)$, for a single variable t.*

Consequently, by Cohn's Theorem, the commutation of two multisets of words, i.e., formal power series, can be characterized similarly as in free monoids: two multisets commute if and only if they are unions of powers of another multiset. It is then tempting to conjecture that the commutation of two arbitrary sets of words can be characterized in similar terms: two languages commute if and only if they are unions of powers of a third language. However, this is not true, as shown in the next example.

Example 6.1.1 ([13]). The sets $X = \{a, a^3, b, ab, ba, aba\}$ and $Y = X \cup \{a^2\}$ commute, but they cannot be expressed as unions of powers of a same set.

As a matter of fact, it has been conjectured in [66] that the commutation of languages can be characterized as above, similarly as for words, if one of the languages is a code. Thus, Ratoandromanana, 1989 [66], conjectured that a code X commutes with a language L if and only if there is a language R such that $L = \cup_{i\in I} R^i$ and $X = R^m$, for some $I \subseteq \mathbb{N}$ and a positive integer m. We say in this case that X satisfies the *BTC-property*, i.e., the Bergman-type of commutation, in the sense that the commutation with X can be characterized as in Bergman's theorem.

BTC-Property: *For a given set $X \subseteq \Sigma^+$, we say that X satisfies the BTC-property if for any set Y commuting with X, there exists a set $V \subseteq \Sigma^+$ and sets I, J of nonnegative integer indices, such that*

$$X = \bigcup_{i\in I} V^i \quad \text{and} \quad Y = \bigcup_{j\in J} V^j.$$

Regarding the BTC-property, we have the following conjecture of [66].

Conjecture 3 ([66]). All codes satisfy the BTC-property. Equivalently, for any code X and any set Y commuting with X, there is a set $R \subseteq \Sigma^+$, a set $I \subseteq \mathbb{N}$, and $n \in \mathbb{N}$ such that $X = R^n$ and $Y = \cup_{i\in I} R^i$.

Conjecture 3 has been proved in the case of prefix codes in [66], using some involved combinatorial considerations on words and codes. Using a similar combinatorial approach and the results of Chapter 4, we give elementary proofs of Conjecture 3 in the case of binary and ternary codes, and we extend the result to periodic sets, the only case of noncodes for which the BTC-property is known to hold.

We also develop a new approach, the so called *multiplicity approach*, to this problem. The basic idea of this new strategy is to transfer relations on sets of words to relations on formal power series, solve them in this framework using strong algebraic tools, and then transfer the results back to sets of words. Based on the multiplicity approach, we prove that the BTC-property holds also for ω-codes and reversed ω-codes, the most general results on Conjecture 3 up to date. As a matter of fact, all the other known results on Conjecture 3 follow as simple consequences of our result.

6.2 Some examples

We have seen in Example 6.1.1 that the BTC-property does not hold in general for all sets of words. We give more examples in the following.

Example 6.2.1. (i,[3]) The sets

$$X = \{aa, ab, ba, bb, aaa\} \quad \text{and} \quad Y = \{a, b, aa, ab, ba, ba, aaa\}$$

commute, but they cannot be expressed in terms of another set of words. (ii,[13]) The sets

$$X = \{a, ab, ba, bb\} \quad \text{and} \quad Y = X \cup X^2 \cup \{bab, bbb\}$$

commute but they are not unions of powers of another set.

In particular, this example seems to be a boundary point, in the sense that no such example is known for sets with at most three words, and none is believed to exist. In fact, it is proved that the BTC-property holds for all sets with at most two words, see [13], [39], and [40]. The ternary case is a special one, as discussed also in Chapter 4. It is believed that the BTC-property holds for all ternary sets of words, but this has been proved only for ternary codes. We discuss this in details in Sections 6.4, 6.5, and 6.8.

Throughout this chapter, we always assume that the languages we consider do not contain the empty word. A similar approach was taken in Chapter 4, as well as in [3], [66], and [73]. The reason for this restriction is that there is little hope for a reasonable characterization of the commutation with languages L, with $1 \in L$. E.g., for any language L, even for non-recursively enumerable L, the sets a^* and a^*La^* commute. Also, for any language L, $1 + L$ and Σ^* always commute.

6.3 The commutation with periodic and binary sets of words

We recall that a set of words L is called *periodic* if $L \subseteq u^+$, for some primitive word u.

The BTC-property is known to hold for only one class of non-codes, namely for periodic sets, as we proved in Theorem 4.3.2. We recall this result in the following.

Theorem 4.3.2. *Let u be a primitive word and $L \subseteq u^+$. Then $\mathcal{C}(L) = u^*$. Moreover, for any set of words X, if $LX = XL$, then $X = \cup_{i \in I} u^i$, for some $I \subseteq \mathbb{N}$.*

The BTC-property holds also for all binary sets. In fact, the binary case is a rather simple one, sine any binary set of words is either periodic, or a code. To see this, let $F = \{u, v\}$ be a binary set. If $uv = vu$, then $\rho(u) = \rho(v)$, and so, $F = \{t^m, t^n\}$, for some $m, n \geq 1$. Thus, F is periodic, case which is covered by Theorem 4.3.2. If $uv \neq vu$, then F is a code, and by Theorem 4.3.4, $\mathcal{C}(F) = F^*$. Based on these considerations, we prove in the next result that the BTC-property holds for all binary set of words, giving here an elementary, somewhat simpler proof of this result of [13].

Theorem 6.3.1 ([13]). *Let F be a binary set, $1 \notin F$ and X a set of words commuting with F.*

(i) If F is periodic, $F \subseteq u^+$, with u primitive word, then X is also periodic, $X \subseteq u^*$.

(ii) If F is a code, then there is $I \subseteq \mathbb{N}$ such that $X = \cup_{i \in I} F^i$.

Proof. The case (i) is trivial since $\mathcal{C}(F) = u^*$. In case (ii), since X commutes with F, then clearly, $X \subseteq \mathcal{C}(F)$, and so, by Theorem 4.3.4, $X \subseteq F^*$. Let $n \geq 0$ be such that $F^n \cap X \neq \emptyset$. To conclude the theorem, it is enough to prove that $F^n \subseteq X$.

Let $\alpha_1, \ldots, \alpha_n \in F$ be such that $\alpha_1 \ldots \alpha_n \in X$. Let also $\beta_1, \ldots, \beta_n$ be arbitrary words of F and assume that $\alpha_{i+1} \ldots \alpha_n \beta_1 \ldots \beta_i \in X$, for some $0 \leq i < n$. Then, since $XF = FX$, there are $y_{i+1} \in X$, $\gamma_{i+1} \in F$ such that

$$\alpha_{i+1} \ldots \alpha_n \beta_1 \ldots \beta_i \cdot \beta_{i+1} = \gamma_{i+1} \cdot y_{i+1}.$$

Since F is a code, it follows that $\alpha_{i+1} = \gamma_{i+1}$, and so, $\alpha_{i+2} \ldots \alpha_n \beta_1 \ldots \beta_{i+1} \in X$. Iterating the argument, we obtain for $i = n$ that $\beta_1 \ldots \beta_n \in X$, and so, $F^n \subseteq X$. Thus, $X = \cup_{i \in I} F^i$, where $I = \{i \in \mathbb{N} \mid F^i \cap X \neq \emptyset\}$, concluding the proof of the theorem. □

6.4 The commutation with ternary codes

In this section we characterize all sets commuting with a given three-element code. We prove that any set of words commuting with a three-element code X is a union of powers of X, i.e., the BTC-property holds for X. Using the combinatorial approach, employed already on the periodic and binary cases of Chapters 4 and 6.3, we prove that the centralizer of any ternary code F is F^*. The characterization of the commutation with F follows then straightforwardly.

We need the following lemma, see also [13] and [66]. Its proof is essentially the proof of Theorem 6.3.1.

Lemma 6.4.1. *Let $X \subseteq \Sigma^*$ be a code such that its centralizer is X^*, and let $Y \subseteq \Sigma^*$ be a language commuting with X. If $Y \cap X^n \neq \emptyset$, for some $n \geq 0$, then $X^n \subseteq Y$.*

We recall that for any word $x \in \Sigma^*$, $x = a_1 a_2 \ldots a_n$, with $a_i \in \Sigma$, $\forall 1 \leq i \leq n$, the *reverse* of x is the word $\overline{x} = a_n \ldots a_2 a_1$. For a language L, we denote by $\overline{L}$ the *reverse* of L, i.e., the language $\overline{L} = \{\overline{x} \mid x \in L\}$. Note that $\overline{\overline{x}} = x$, for any word x, and thus, $\overline{\overline{L}} = L$, for any language L. It turns out, and this is useful in our considerations, that the centralizer operator commutes with the bar operator defined above, as proved in the next result.

Lemma 6.4.2. *For any language L, $\mathcal{C}(\overline{L}) = \overline{\mathcal{C}(L)}$.*

Proof. $\mathcal{C}(L)$ is the centralizer of L and so, $L\mathcal{C}(L) = \mathcal{C}(L)L$. Reversing this equation, we obtain $\overline{L}\,\overline{\mathcal{C}(L)} = \overline{\mathcal{C}(L)}\,\overline{L}$. Thus, $\overline{\mathcal{C}(L)}$ commutes with $\overline{L}$, and so, $\overline{\mathcal{C}(L)} \subseteq \mathcal{C}(\overline{L})$. Similarly, $\overline{L}\mathcal{C}(\overline{L}) = \mathcal{C}(\overline{L})\overline{L}$, and so, $L\overline{\mathcal{C}(\overline{L})} = \overline{\mathcal{C}(\overline{L})}L$. Thus, $\overline{\mathcal{C}(\overline{L})} \subseteq \mathcal{C}(L)$ or, equivalently, $\mathcal{C}(\overline{L}) \subseteq \overline{\mathcal{C}(L)}$. Consequently, $\mathcal{C}(\overline{L}) = \overline{\mathcal{C}(L)}$. □

The following result is a simple consequence of Lemma 6.4.2, since $\overline{R}$ is a rational language for any rational language R.

Corollary 6.4.3. *For any language L, $\mathcal{C}(L)$ is rational if and only if $\mathcal{C}(\overline{L})$ is rational.*

It is an intriguing open problem, see Section 4.6, to prove that for any nonperiodic ternary set F, if $1 \notin F$, then $\mathcal{C}(F) = F^*$. We proved in Chapter 4 that for any such F, $\mathcal{C}(F)$ is a rational set, having a form very close to F^*. We prove in the next theorem that for codes F, we have $\mathcal{C}(F) = F^*$, thus giving a solution for Problem 4.6.6, in the case of codes. Besides its intrinsic interest, this is the most important result used in the proof of Theorem 6.4.5, stating that the BTC-property holds for all ternary codes.

Theorem 6.4.4. *The centralizer of a three-word code F is F^*.*

Proof. Let $\mathcal{C}(F)$ be the centralizer of F. We distinguish three cases, depending on the prefix decomposition of F.

1. $F = u_1 + u_2 + u_3$, with u_1 a prefix of both u_2 and u_3.
2. $F = u_1 + u_2 + u_3$, with u_1, u_2, u_3 pairwise prefix incomparable.
3. $F = u_1(1 + u_2) + u_3$, with u_1 and u_3 prefix incomparable, and $u_2 \neq 1$.

We treat each of these separately.

Case 1. $F = u_1 + u_2 + u_3$, with u_1 a prefix of both u_2 and u_3. Equivalently, F is of the form $F = \{u, uv, uw\}$, with $v, w \neq 1$ and $v \neq w$. We assume that F^* is a proper subset of $\mathcal{C}(F)$, and let x be *minimal* with respect to the length in $\mathcal{C}(F) \setminus F^*$.

Claim 1. $xu = ux$.

Proof of Claim 1. Assume the contrary and set $x_1 = x$. We define the sequence $(x_i)_{i \leq n}$, for some $n \geq 1$, such that $x_i u = u x_{i+1}$, for all $1 \leq i \leq n-1$, and either $x_n u = u x_k$, for some $1 \leq k \leq n$, or $x_n u = ty$, with $t \in \{uv, uw\}$, and $y \in \mathcal{C}(F)$. In the former case we obtain that $x_k u^{n-k+1} = u^{n-k+1} x_k$, i.e., $x_k u = u x_k$. It then follows that $x_1, \ldots, x_n$ all commute with u; in particular, we obtain that $xu = ux$, a contradiction. In the latter case, we obtain that $|y| < |x|$. Thus, since x is minimal in $\mathcal{C}(F) \setminus F^*$, it follows that $y \in F^*$. Consequently, $xu^n = u^{n-1}ty \in F^*$. Similarly we can prove that there is

$m \geq 1$ such that $u^m x \in F^*$. But F is a code, i.e., F^* is free, and then, the Schützenberger criterium of a free monoid ([47]) implies that $x \in F^*$. This is a contradiction.

As a consequence of Claim 1, we obtain that x is the only minimal element in $\mathcal{C}(F) \setminus F^*$. Thus, if $y \in \mathcal{C}(F) \setminus F^*$ and $|y| = |x|$, then necessarily $y = x$.

To make a choice, let us assume that $|v| \leq |w|$, without loss of generality. It is important to observe that neither v, nor w can commute with u since F is a code.

Claim 2: There exist words $\alpha, \beta \in F^*$ such that $x\alpha, \beta x \in F^*$.
Proof of Claim 2. We prove that there is $\alpha \in F^*$ such that $x\alpha \in F^*$; the second part of the claim can be proved using symmetric arguments.

Let us assume that $x\alpha \notin F^*$, for all $\alpha \in F^*$, and consider the word xuv. We prove first that there exists $n \geq 0$ such that

$$xvu^n = u^n wx. \tag{6.1}$$

If $xuv = uvy$, for some $y \in \mathcal{C}(F)$, we have that $|x| = |y|$, and so, either $y \in \mathcal{C}(F) \setminus F^*$, i.e., $y = x$, or $y \in F^*$. In the former case, we have that $xuv = uvx$ and then, by Claim 1, $uv = vu$, a contradiction. In the latter case, $xuv \in F^*$, contradicting our assumption.

If $xuv = uwy$, we obtain that $|y| \leq |x|$. Thus, as above, either $y \in \mathcal{C}(F) \setminus F^*$, and so, $y = x$, or $y \in F^*$. The latter case leads to $xuv \in F^*$, a contradiction. In the former case, we obtain $xv = wy$, which is what we wanted to prove first.

Finally, if $xuv = uy$, then by Claim 1, $y = xv \in \mathcal{C}(F)$. In this case, we consider the word $y_1 = y$, and then, the word $y_1 u = xvu \in \mathcal{C}(F)F$. There must be an integer $n \geq 1$, such that $y_i u = uy_{i+1}$, for all $1 \leq i \leq n-1$, and either $y_n u = uy_k$, for some $1 \leq k \leq n$, or $y_n u = uvt$, or $y_n u = uwt$, for some $t \in \mathcal{C}(F)$.

If $y_n u = uy_k$, $1 \leq k \leq n$, then $y_k u^{n-k+1} = u^{n-k+1} y_k$, i.e., $y_k u = uy_k$. This implies that $yu = uy$ and so, $uv = vu$, a contradiction.

If $y_n u = uvt$, $t \in \mathcal{C}(F)$, then we derive that $xvu^n = u^n vt$ and so, $|t| = |x|$. Thus, since x is the only minimal element in $\mathcal{C}(F) \setminus F^*$, either $t \in F^*$, or $t = x$. In the former case we obtain that $xuvu^n = u^n uvt \in F^*$, contradicting our assumption. In the latter case it follows that $uv = vu$, again a contradiction.

If $y_n u = uwt$, $t \in \mathcal{C}(F)$, then $xvu^n = u^n wx$, proving (6.1).

Consequently, there is $n \geq 0$ such that $xvu^n = u^n wx$. In particular, it follows that $|v| = |w|$.

Using similar arguments as above, with w instead of v, one can prove that there is $m \geq 0$ such that

$$xwu^m = u^m vx. \tag{6.2}$$

Indeed, although it was essential in the above considerations that $|v| \leq |w|$, note that (6.1) implies that $|v| = |w|$.

Without loss of generality, let us assume that $m \leq n$.

If $x \leq u^m$, then $u^m = xz$, for some nonempty word z. By Claim 1, $zu = uz$, and from (6.1) and (6.2) we derive that

$$vu^{n-m}z = u^{n-m}zw, \qquad wz = zv, \qquad uz = zu.$$

Applying Graph Lemma on these three relations, for the set of unknowns $\{z, u, v, w\}$, we obtain that F is periodic, a contradiction.

Consequently, $u^m \leq x$. Let ρ be the primitive root of u and x, $u = \rho^j$, $x = \rho^i$, for some positive integers i and j, $i > mj$. It follows from (6.2) that

$$\rho^{i-mj}w = v\rho^{i-mj}, \tag{6.3}$$

i.e., v and w are conjugates. We discuss separately the following two cases:

$$(i)\ \rho^{i-mj} \leq v, \qquad \text{or} \qquad (ii)\ v \leq \rho^{i-mj}.$$

(i) If $\rho^{i-mj} \leq v$, then there is a word δ such that $v = \rho^{i-mj}\delta$ and $w = \delta\rho^{i-mj}$. In this case,

$$F = \{\rho^j, \rho^{i-(m-1)j}\delta, \rho^j\delta\rho^{i-mj}\},$$

$u = \rho^j$, and $x = \rho^i \in \mathcal{C}(F)$.

Since $x \in \mathcal{C}(F)$, it follows that $s = xuv(uw)^\omega \in F^\omega$, i.e.,

$$s = \rho^{2i-(m-1)j}\delta(\rho^j\delta\rho^{i-mj})^\omega \in F^\omega.$$

Clearly, if $s = \rho^\omega$, then $\rho\delta = \delta\rho$ and $v = w$, a contradiction. Then, either

$$s = (\rho^j)^r\rho^{i-(m-1)j}\delta s', \quad \text{or} \quad s = (\rho^j)^r\rho^j\delta\rho^{i-mj}s',$$

for some $r \geq 0$ and $s' \in F^\omega$.

Since any nontrivial relation on ρ and δ leads to $\rho\delta = \delta\rho$ and then to contradiction as above, it follows that in the former case, $2i - (m-1)j = rj + i - (m-1)j$ and so, $i = rj$. This implies that $x = u^r \in F^*$, a contradiction. Similarly, in the latter case we obtain $2i - (m-1)j = rj + j$, i.e., $2i = (r+m)j$. Thus,

$$(\rho^j\delta\rho^{i-mj})^\omega = \rho^{i-mj}s',$$

with $s' \in F^\omega$. If s' starts with ρ^j or with $\rho^j\delta\rho^{i-mj}$ then we obtain a nontrivial relation on ρ and δ. Thus, s' starts with $\rho^{i-(m-1)j}\delta$ and to avoid a nontrivial relation on ρ and δ, we must have $j = i - (m-1)j$. Thus, $i = mj$, and so, $x = u^m \in F^*$, again a contradiction.

(ii) If $v \leq \rho^{i-mj}$, then $v = \rho^k\rho'$, with $k \leq i - mj - 1$ and $\rho' \leq \rho$, $\rho' \neq 1$. Thus, (6.3) can be rewritten as

$$\rho^{i-mj-k}w = \rho'\rho^{i-mj}, \tag{6.4}$$

with $\rho' \leq \rho$, $\rho' \neq 1$. Since ρ is primitive, ρ cannot be a factor of ρ^2, other than as a prefix, or as a suffix, see [11] and [47]. Thus, it follows from (6.4) that $i - mj - k = 1$. Consequently, $\rho^{i-mj-1} \leq v \leq \rho^{i-mj}$, and so, there are some nonempty words ρ_1, ρ_2, ρ_3 such that $v = \rho^{i-mj-1}\rho_1$, $w = \rho_3\rho^{i-mj-1}$,

$$\rho = \rho_1\rho_2, \qquad \text{and} \qquad \rho = \rho_2\rho_3. \tag{6.5}$$

In this case,

$$F = \{\rho^j, \rho^{i-(m-1)j-1}\rho_1, \rho^j\rho_3\rho^{i-mj-1}\}.$$

Since $x \in \mathcal{C}(F)$, it follows that $s = xuv(uw)^\omega \in F^\omega$, i.e.,

$$s = \rho^{2i-(m-1)j-1}\rho_1(\rho^j\rho_3\rho^{i-mj-1})^\omega \in F^\omega.$$

Clearly, if $s = \rho^\omega$, then $\rho_1 = 1$, and so, $v = w$, which is impossible. Thus, either

$$s = (\rho^j)^r\rho^{i-(m-1)j-1}\rho_1 s', \quad \text{or} \quad s = (\rho^j)^r\rho^j\rho_3\rho^{i-mj-1}s',$$

with $r \geq 0$ and $s' \in F^\omega$.

In the former case, if $2i - (m-1)j - 1 \neq rj + i - (m-1)j - 1$, then it clearly follows that $\rho_1\rho = \rho\rho_1$, leading to contradiction. Thus, we obtain that $i = rj$, i.e., $x = (\rho^j)^r \in F^*$, again a contradiction.

In the latter case, if $2i - (m-1)j - 1 > (r+1)j$, then

$$\rho^{2i-(m+r)j-1}\rho_1(\rho^j\rho_3\rho^{i-mj-1})^\omega = \rho_3\rho^{i-mj-1}s',$$

with $s' \in F^\omega$. The Graph Lemma applied on this equation and on (6.5) implies that F is periodic, a contradiction. Thus, $2i-(m-1)j-1 \leq (r+1)j$, and so,

$$\rho_1(\rho^j\rho_3\rho^{i-mj-1})^\omega = \rho^{(r+m)j-2i+1}\rho_3\rho^{i-mj-1}s'.$$

Since ρ is a primitive word, ρ is not a factor of ρ^2, other than as a prefix and a suffix, see [11], [47]. Thus, recalling that $\rho_1 \leq \rho$, it follows that $(r+m)j - 2i + 1 = 1$, i.e., $2i = (r+m)j$. Moreover, since $\rho_1\rho = \rho\rho_3$, we obtain

$$\rho^{j-1}\rho_3\rho^{i-mj-1}(\rho^j\rho_3\rho^{i-mj-1})^\omega = \rho^{i-mj-1}s'. \tag{6.6}$$

If s' starts with ρ^j or with $\rho^j\rho_3\rho^{i-mj-1}$ we can apply Graph Lemma on (6.6) and (6.5) to conclude that F must be periodic, which is a contradiction. Thus, $s' = \rho^{i-(m-1)j-1}\rho_1 s''$, with $s'' \in F^\omega$. Consequently,

$$\rho^{j-1}\rho_3\rho^{i-(m-1)j-1}(\rho^j\rho_3\rho^{i-(m-1)j-1})^\omega = \rho^{2i-(2m-1)j-2}\rho_1 s''.$$

Repeating the above arguments, we conclude that necessarily, $j - 1 = 1 + (2i - (2m-1)j - 2)$, i.e., $i = mj$. Thus, $x = u^m \in F^*$, a contradiction.

Since all the alternatives lead to contradictions, the conclusion is that there must be $\alpha \in F^*$ such that $x\alpha \in F^*$. Using similar symmetric arguments, we can also conclude that there is $\beta \in F^*$ such that $\beta x \in F^*$. This completes the proof of Claim 2.

Since F is a code, from Claim 2, using the Schützenberger criterium of a free monoid, we obtain that $x \in F^*$. This is impossible: x was chosen in $\mathcal{C}(F) \setminus F^*$.

Consequently, $\mathcal{C}(F) = F^*$.

Case 2. $F = u_1 + u_2 + u_3$, and u_1, u_2, u_3 are prefix incomparable. If one of the words in F is a suffix of both the other two, then the set $\overline{F}$ has a prefix decomposition as in Case 1. Moreover, $\overline{F}$ is a code. Thus, $\mathcal{C}(\overline{F}) = \overline{F}^*$ and so, $\mathcal{C}(F) = F^*$.

Assume now that no word of F is a proper suffix of both the other two words of F. Thus, by Lemma 4.4.4, F has a suffix distinguishable word. It follows then by Lemma 4.4.6 that $\mathcal{C}(F) = F^*$.

Case 3. $F = u_1(1 + u_2) + u_3$, with u_1 and u_3 prefix incomparable, and $u_2 \neq 1$. As in Case 2, if one of the words of F is a suffix of both the other two words of F, then the problem can be reduced to Case 1, concluding that $\mathcal{C}(F) = F^*$. Thus, we can assume by Lemma 4.4.4 that F has a suffix distinguishable word. It follows then by Lemma 4.4.7 that $\mathcal{C}(F) = F^*$. □

Using Theorem 6.4.4, we can now characterize all sets commuting with a three-element code.

Theorem 6.4.5. *Let F be a three-word code and X a set of words commuting with F. Then $X = \bigcup_{i \in I} F^i$, for some $I \subseteq \mathbb{N}$.*

Proof. Since $XF = FX$, and $\mathcal{C}(F)$ is the largest set commuting with F, it follows by Theorem 4.1.1 that $X \subseteq \mathcal{C}(F)$. Thus, by Theorem 6.4.4, $X \subseteq F^*$. Then, by Lemma 6.4.1, it follows that for any $n \geq 0$ such that $X \cap F^n \neq \emptyset$, we have $F^n \subseteq X$. Thus, $X = \bigcup_{i \in I} F^i$, where $I = \{n \in \mathbb{N} \mid X \cap F^n \neq \emptyset\}$. □

We emphasize that the case of ternary sets is the boundary point for the validity of the BTC-property on all sets of a given cardinal. Indeed, the BTC-property holds for all binary sets, but not for sets with at least four words. It is conjectured in [38] that in fact, all ternary sets, i.e., also the ternary non-codes, satisfy the BTC-property, see also Section 4.6 of this thesis. This, however, remains a difficult open problem.

6.5 The commutation with arbitrary ternary sets of words

Having characterized in Section 6.4 the commutation with a ternary code, we consider now the commutation with an arbitrary ternary set of words

$F \subseteq \Sigma^+$. It is known that the BTC-property holds for all binary sets of words $G \subseteq \Sigma^+$ and on the other hand, it does not holds for all sets with at least four words, as shown in Example 6.2.1. Thus, the ternary case is the boundary point for the validity of the BTC-property on all finite sets of words of a given cardinal. It has been conjectured in [38] that the BTC-property holds for all ternary sets and we discussed in Section 4.6 several possible strategies to prove it. While the problem remains open, we reduce it in this section to proving that $\mathcal{C}(F) = F^*$, for all ternary sets $F \subseteq \Sigma^+$, conceivably an important step towards a complete solution of the problem.

We first recall the known results on the commutation with a ternary set of words, i.e., the case of periodic sets, and that of codes.

Theorem 6.5.1. *Let $F \subseteq \Sigma^+$ be a three word set.*

(i) If F is periodic, $F \subseteq t^*$, for some primitive word $t \in \Sigma^*$, then $\mathcal{C}(F) = t^*$. Moreover, any language X commuting with F is of the form $X = \bigcup_{i \in I} t^i$, for some $I \subseteq \mathbb{N}$.

(ii) If F is a code, then $\mathcal{C}(F) = F^*$. Moreover, any language X commuting with F is of the form $X = \bigcup_{i \in I} F^i$, for some $I \subseteq \mathbb{N}$.

We prove in the next result that the BTC-property holds for a ternary set of words $F \subseteq \Sigma^+$ if and only if $\mathcal{C}(F) = F^*$. We thus reduce the problem of characterizing the commutation of F with an *arbitrary* set of words to the problem of establishing a combinatorial property of F itself.

Let F be a ternary set of words. We say that a word $x \in X$ *satisfies the property* $\mathcal{P}_F$, or that $\mathcal{P}_F(x)$ *holds*, if for all $n \in \mathbb{N}$, if $x \in F^n$, then necessarily, $F^n \not\subseteq X$. Note that for any F, $\mathcal{P}_F(1)$ does not hold.

Theorem 6.5.2. *Let $F \subseteq \Sigma^+$ be a nonperiodic three word noncode. Then the BTC-property holds for F if and only if $\mathcal{C}(F) = F^*$.*

Proof. If the BTC-property holds for F, then, since $F\mathcal{C}(F) = \mathcal{C}(F)F$, it follows that $F = R^i$ and $\mathcal{C}(F) = \cup_{j \in J} F^j$, for some set of words R and $i \in \mathbb{N}$, $J \subseteq \mathbb{N}$. Note however that any nonperiodic ternary set of words is primitive, and so, $i = 1$ and $R = F$. Also, since $\mathcal{C}(F)$ is the largest set commuting with F, it follows that $J = \mathbb{N}$, and thus,

$$\mathcal{C}(F) = F^*.$$

To prove the reverse implication, assume that $\mathcal{C}(F) = F^*$, and let X be a language commuting with F:

$$FX = XF.$$

We prove that in this case, $X = \cup_{i \in I} F^i$, for some $I \subseteq \mathbb{N}$. If $X = \emptyset$, then the claim is trivially true. Let us assume that $X \neq \emptyset$.

Since F is not a code, there is a nontrivial relation on the elements of F:

$$u_1 u_{i_2} \dots u_{i_m} = u_2 u_{i_{m+1}} \dots u_{i_n}, \tag{6.7}$$

with $u_1 \neq u_2$ and $u_1, u_2, u_{i_k} \in F$, for all $2 \leq k \leq n$. Let u_3 be the third element of F:

$$F = \{u_1, u_2, u_3\}.$$

By Theorem 4.1.1, $\mathcal{C}(F)$ includes any set commuting with F, and so, $X \subseteq \mathcal{C}(F) = F^*$. We prove the following two claims.

Claim 1. Let $x \in X$ be such that $\mathcal{P}_F(x)$ holds. If $y \in X$, $v \in F$ are such that

$$u_3 x = yv,$$

then $\mathcal{P}_F(y)$ holds.
Proof of Claim 1. Since $y \in X \subseteq F^*$, let $m \geq 1$ be such that $y \in F^m$, say $y = v_1 \dots v_m$, with $v_i \in F$, for all $1 \leq i \leq m$. We have the equation

$$u_3 x = v_1 \dots v_m v. \tag{6.8}$$

If $v_1 \neq u_3$, then, as $x \in \mathcal{C}(F) = F^*$, applying the Graph Lemma on (6.8) and on (6.7), for the set of unknowns $\{u_1, u_2, u_3\}$, we obtain that F is periodic, a contradiction. Thus, $v_1 = u_3$, and so, $x = v_2 \dots v_m v \in F^m$. Since $\mathcal{P}_F(x)$ holds, it follows that $F^m \not\subseteq X$. Consequently, $\mathcal{P}_F(y)$ holds, proving Claim 1.

Claim 2. If there is $x_1 \in X$ such that $\mathcal{P}_F(x_1)$ holds, then $\mathcal{P}_F(u_3^q)$ holds, for some positive integer q.
Proof of Claim 2. Since $FX = XF$, there are $v_n \in F$ and $x_{n+1} \in X$ such that

$$u_3 x_n = x_{n+1} v_n, \qquad \text{for all } n \geq 1. \tag{6.9}$$

Moreover, by Claim 1, $\mathcal{P}_F(x_n)$ holds, for all $n \geq 1$. Since $\mathcal{P}_F(1)$ does not hold, it follows that $x_n \neq 1$, for all $n \geq 1$.

Assume now that $x_n \notin u_3^+$, for all $n \geq 1$. We prove by induction on n that $x_n = u_3^{n-1} y_n$, for some $y_n \in F^+$, for all $n \geq 1$. This is trivially true for $n = 1$, since $x_1 \in X \subseteq \mathcal{C}(F) = F^*$.

Consider now $n \geq 1$ and assume that $x_n = u_3^{n-1} y_n$, with $y_n \in F^+$. Since $X \subseteq \mathcal{C}(F) = F^*$, it follows that $x_{n+1} \in F^+$, say $x_{n+1} = u_3^k w_{k+1} \dots w_l$, for some $k \geq 0$, with $w_i \in F$, for all $k+1 \leq i \leq l$, and $w_{k+1} \neq u_3$. Equation (6.9) is thus equivalent with

$$u_3^n y_n = u_3^k w_{k+1} \dots w_l v_n. \tag{6.10}$$

If $k < n$, then the equation (6.10) is rewritten as

$$u_3^{n-k} y_n = w_{k+1} \dots w_l w_n. \tag{6.11}$$

Thus, since $w_{k+1} \in \{u_1, u_2\}$, it follows by Graph Lemma on the equations (6.7) and (6.11), for the set of unknowns $\{u_1, u_2, u_3\}$, that F is periodic, a contradiction.

Thus, $k \geq n$, i.e., $x_{n+1} = u_3^n z$, with $z = u_3^{k-n} w_{k+1} \dots w_l v_n \in F^+$.

Consequently, $x_n \in u_3^{n-1} F^+$, for all $n \geq 1$, and so,

$$|x_n| \geq |u_3^{n-1}| = (n-1)|u_3|,$$

for all $n \geq 1$. On the other hand, by (6.9),

$$|x_n| = |x_{n-1}| + |u_3| - |v_{n-1}| \leq |x_{n-1}| + |u_3| - l_F.$$

Thus, $|x_n| \leq |x_1| + (n-1)(|u_3| - l_F)$. Altogether, we obtain that

$$(n-1)|u_3| \leq |x_n| \leq |x_1| + (n-1)(|u_3| - l_F),$$

for all $n \geq 1$. This further implies that $n \leq 1 + \frac{|x_1|}{l_F}$, for all $n \geq 1$, which is impossible.

Our assumption is thus false: there is $n \geq 1$ such that $x_n \in u_3^+$, i.e., $x_n = u_3^q$, for some positive integer q. Consequently, by Claim 1, $\mathcal{P}_F(u_3^q)$ holds, proving Claim 2.

Assume now that there is $x \in X$ such that $\mathcal{P}_F(x)$ holds. Then, by Claim 2, there is a positive integer q such that $\mathcal{P}_F(u_3^q)$ holds. For such a positive integer q, consider some arbitrary words $v_1, \dots, v_q \in F$. We prove that $u_3^{q-i} v_1 \dots v_i \in X$, for all $0 \leq i \leq q$.

Since $\mathcal{P}_F(u_3^q)$ holds, necessarily $u_3^q \in X$, proving the claim for $i = 0$. Let now $i \geq 0$ be such that $u_3^{q-i} v_1 \dots v_i \in X$, $0 \leq i < q$. We prove that $u_3^{q-(i+1)} v_1 \dots v_i v_{i+1} \in X$.

Since $XF = FX$, there are $w \in F$ and $y \in X$ such that

$$u_3^{q-i} v_1 \dots v_i \cdot v_{i+1} = w \cdot y. \tag{6.12}$$

If $w \neq u_3$, then by Graph Lemma on (6.7) and (6.12) we obtain that F is periodic: a contradiction. Thus, $w = u_3$ and so, $y = u_3^{q-(i+1)} v_1 \dots v_i v_{i+1} \in X$.

Consequently, for $i = q$, we obtain that $v_1 \dots v_q \in X$, for all $v_1, \dots, v_q \in F$. Thus, $F^q \subseteq X$, contradicting $\mathcal{P}_F(u_3^q)$.

The conclusion is that there is no $x \in X$ such that $\mathcal{P}_F(x)$ holds. Equivalently, for any $x \in X$, there is $m \in \mathbb{N}$ such that $x \in F^m \subseteq X$. Consequently, X is of the form

$$X = \bigcup_{i \in I} F^i,$$

with $I = \{i \in \mathbb{N} \mid \exists x \in X : x \in F^i \subseteq \mathbb{N}\}$, i.e., the BTC-property holds for F. □

6.6 The commutation with ω-codes

We consider in this section the commutation with ω-codes and reversed ω-codes. We recall that a set of words L is an *ω-code* if any *right-infinite* word has *at most one* L-factorization. We say that L is a *reversed ω-code* if any *left-infinite* word has at most one L-factorization.

We give here a solution to Conjecture 3 in the case of ω-codes and reversed ω-codes. We prove that the BTC property holds for such codes C, i.e., for any set of words X commuting with C, there is a set of words L such that

$$C = L^m \qquad \text{and} \qquad X = \bigcup_{i \in I} L^i,$$

for some $m \in \mathbb{N}$ and $I \subseteq \mathbb{N}$. To this aim, we use the so-called *multiplicity approach*, as detailed in the following.

6.6.1 The multiplicity approach

The commutation of two polynomials, or of two formal power series, in noncommuting variables, with coefficients in a field, is characterized in two results due to Bergman, 1969, [6], for polynomials, and to Cohn, 1962, [16], for formal power series. These two results, recalled in Section 6.1, characterize the commutation of two multisets of words similarly as in free monoids. In this sense, the BTC-property and Conjecture 3 are the equivalent of Bergman's and Cohn's results to sets of words. Our multiplicity approach takes advantage precisely of this corresponding. We consider a relation on sets of words, in this case, the commutation, and translate it into a corresponding relation on formal power series. We then solve the problem in terms of formal power series and transfer the result back to sets of words, obtaining a solution of the original problem. More specifically, we consider the commutation of two sets of words: $XY = YX$. Let r_X and r_Y be the *characteristic formal power series* of X and Y, respectively:

$$r_X = \sum_{u \in X} u, \qquad r_Y = \sum_{v \in Y} v.$$

If we translate the commutation $XY = YX$ to their characteristic formal power series, i.e., if we prove that $r_X r_Y = r_Y r_X$, then it follows by Cohn's theorem that there are some formal power series r, s, t such that

$$r_X = s(r), \qquad r_Y = t(r).$$

Denoting $R = \mathrm{supp}(r)$, i.e., $R = \{u \mid (r, u) \neq 0\}$, it follows that X and Y are combinations of R,

$$X = \bigcup_{i \in I} R^i, \qquad Y = \bigcup_{j \in J} R^j,$$

for some $I, J \subseteq \mathbb{N}$, thus solving the commutation $XY = YX$.

Note however that in general, the commutation of two sets of words does not necessarily imply the commutation of their characteristic formal power series.

Example 6.6.1. Consider

$$X = \{aa, ab, ba, bb, aaa\} \text{ and } Y = \{a, b, aa, ab, ba, bb, aaa\}.$$

Then $r_X = aa + ab + ba + bb + aaa$ and $r_Y = a + b + aa + ab + ba + bb + aaa$. Note that

$$XY = YX = \Sigma^3 + \Sigma^4 + \Sigma^2 a^3 + a^3 \Sigma^2 + a^6,$$

where $\Sigma = \{a, b\}$, while $r_X r_Y \neq r_Y r_X$:

$$r_X r_Y = (a+b)^3 + (a+b)^4 + (a+b)^2 a^3 + a^3 (a+b)^2 + a^6 + a^4 + \mathbf{a^3 b},$$
$$r_Y r_X = (a+b)^3 + (a+b)^4 + (a+b)^2 a^3 + a^3 (a+b)^2 + a^6 + a^4 + \mathbf{b a^3},$$

The situation is much simpler when the commutation of two sets of words, $XY = YX$, is unambiguous, in the sense that both products $XY = YX$ are unambiguous. In this case, it follows immediately that the characteristic formal power series of X and Y commute and thus, the commutation $XY = YX$ is solved using Cohn's theorem, as discussed above. Moreover, it turns out that in this case, one product is unambiguous if and only if the other one is unambiguous. This surprising result is a generalization of a property of prefix codes, see [66].

Lemma 6.6.1. *Let L_1, L_2 be two languages such that $L_1 L_2 = L_2 L_1$. Then the product $L_1 L_2$ is unambiguous if and only if the product $L_2 L_1$ is unambiguous.*

Proof. For a language T and a nonnegative integer n, we denote $T_n = \{t \in T \mid |t| = n\}$. Also, let

$$A_n = \{(l_1, l_2) \mid l_1 \in L_1, l_2 \in L_2, |l_1 l_2| = n\},$$
$$B_n = \{(l_2, l_1) \mid l_2 \in L_2, l_1 \in L_1, |l_2 l_1| = n\}.$$

Clearly, $L_1 L_2$ is unambiguous if and only if $(L_1 L_2)_n$ is unambiguous for any nonnegative integer n. Moreover, $(L_1 L_2)_n$ is unambiguous if and only if there is a bijection between $(L_1 L_2)_n$ and A_n.

Assuming that $L_1 L_2$ is unambiguous, we prove that $(L_2 L_1)_n$ is unambiguous as well, for all n.

Clearly, there is a bijection between A_n and B_n. Also, as $(L_1 L_2)_n$ is unambiguous, there is a bijection between A_n and $(L_1 L_2)_n$. Since $(L_1 L_2)_n = (L_2 L_1)_n$, it follows that there is a bijection between $(L_2 L_1)_n$ and B_n. Consequently, $L_2 L_1$ is unambiguous. □

6.6.2 The results

We apply in this section the multiplicity approach to the commutation with an ω-code. The following result provides the basis for this.

Lemma 6.6.2. *For an ω-code C, let X be a language such that $CX = XC$. Then the product CX is unambiguous.*

Proof. Assume that CX is ambiguous, i.e., there are distinct words $u, v \in C$, such that $ux = vy$, for some $x, y \in X$. Since $CX = XC$, for any $z \in X$ and any $w \in C$, we have $zw^\omega \in C^\omega$. Thus, for any $w \in C$, there are $\alpha_n, \beta_n \in C$, $n \geq 1$, such that $xw^\omega = \alpha_1\alpha_2 \ldots$ and $yw^\omega = \beta_1\beta_2 \ldots$. Then $\gamma = uxw^\omega \in C^\omega$ has two distinct factorizations over C: $\gamma = u\alpha_1\alpha_2 \ldots = v\beta_1\beta_2 \ldots$. This is impossible since C is an ω-code. □

Using Lemma 6.6.2 and the multiplicity approach, the commutation with an ω-code can be stated now in terms of formal power series and characterized using Cohn's theorem.

Theorem 6.6.3. *The BTC-property holds for all ω-codes. In other words, for any ω-code C and any language L, the following conditions are equivalent:*

(i) $LC = CL$;

(ii) there is a language R such that $C = \cup_{i\in I} R^i$ and $L = \cup_{j\in J} R^j$, for some sets I and J of nonnegative integers;

(iii) there is a language R such that $C = R^i$, for some positive integer i, and $L = \cup_{j\in J} R^j$, for some set J of nonnegative integers;

(iv) $L = \bigcup_{j\in J} \rho(C)^j$, *for some $J \subseteq \mathbb{N}$, where $\rho(C)$ denotes the primitive root of C.*

Proof. (i)⇒(ii) Using Lemmata 6.6.2 and 6.6.1, we obtain that both LC and CL are unambiguous products. Thus, the characteristic formal power series of L and C commute, and it follows by Cohn's Theorem that they can be both expressed in terms of another series r. Let R be the support of r. Then there are $I, J \subseteq \mathbb{N}$ such that $C = \cup_{i\in I} R^i$ and $L = \cup_{j\in J} R^j$.

(ii)⇒(iii) Let R be a non-empty language such that $C = \cup_{i\in I} R^i$, for some non-empty set of positive integers I. If there are two distinct integers $i, j \in I$, then $u^i, u^j \in C$, for any word $u \in R$. Thus, $(u^i)^j = (u^j)^i$ is a non-trivial relation on C, which is impossible since C is a code.

(iii)⇒(iv) This is immediate, since by Theorem 5.5.7, any code has a unique primitive root.

(iv)⇒(i) This is obvious. □

Since the commutation of languages, i.e., the language equation $L_1L_2 = L_2L_1$, is symmetric in the two unknowns, it follows that the commutation with a reversed ω-code can be characterized similarly as for ω-codes. We have the following result.

Corollary 6.6.4. *The BTC-property holds for all reversed ω-codes. Equivalently, for any reversed ω-code C and any language L, $CL = LC$ if and only if $L = \bigcup_{j\in J} \rho(C)^j$, for some $J \subseteq \mathbb{N}$, where $\rho(C)$ denotes the primitive root of C.*

Theorem 6.6.3 is the most general result on the commutation of languages, up to date. In particular, Theorem 6.6.3 generalizes (and provides shorter proofs for) all the other known results on the commutation of languages. The trade-off is that we rely in the proof of Theorem 6.6.3 on a deep algebraic result, Cohn's theorem, on the commutation of formal power series. The following results are simple corollaries of Theorem 6.6.3.

Corollary 6.6.5. *(i) The BTC-property holds for all codes with bounded-decoding delay.*
(ii) ([66]) The BTC-property holds for all prefix codes.
(iii) ([13]) The BTC-property holds for all elementary codes. Moreover, for any elementary code E, $\mathcal{C}(E) = E^*$

We also obtain a new prof, the third in this thesis, for the commutation with a binary set.

Corollary 6.6.6. *The BTC-property holds for any binary set F. Moreover, if F is a code, then $\mathcal{C}(F) = F^*$.*

Proof. A unique property of the binary case is that a binary set F is a code if and only if it is an ω-code, see [11], [47]. If F is not a code, then it is periodic. The first part of the claim thus follows by Theorems 4.3.2 and 6.6.3. Moreover, by Theorem 6.6.3, if F is a code, then there is a language R such that $F = R^i$, for some positive integer i, and $\mathcal{C}(F) = \cup_{j\in J} F^j$. Since F is a binary code, necessarily $i = 1$, and $F = R$. Moreover, since $\mathcal{C}(F)$ is the maximal set commuting with F, $J = \mathbb{N}$, and therefore $\mathcal{C}(F) = F^*$. □

The commutation with a ternary code has been characterized in Theorem 6.4.5 using some rather involved combinatorial techniques and the Graph Lemma. We give this result a simple proof here, using the multiplicity approach and Theorem 6.6.3.

Corollary 6.6.7. *The BTC-property holds for any ternary code F. Moreover, in this case $\mathcal{C}(F) = F^*$.*

Proof. Let F be a ternary code. It is a specific property of the ternary case that any ternary code is either an ω-code or a reversed ω-code, see [34], [35].

Thus, the first claim follows by Theorem 6.6.3 and Corollary 6.6.4. The second claim follows similarly as for binary codes. Indeed, if R is such that $F = R^i$, for some $i \geq 0$, then necessarily, $i = 1$. Moreover, since $\mathcal{C}(F)$ is the largest set commuting with F, it follows that $\mathcal{C}(F) = F^*$. □

6.7 Connections to Conway's Problem

We return in this section to Conway's Problem. Using the results obtained so far in this chapter, we establish some interesting connections between the centralizer and the primitive root of a code. In particular, we prove that Conway's problem has a positive answer for rational ω-codes and rational reversed ω-codes, as well as for several other families of codes.

6.7.1 Rational ω-codes

We first establish some connections between the centralizer and the primitive root of a code, and between Conway's problem and the BTC-property.

Theorem 6.7.1. *For any rational code C, the primitive root of C, $\rho(C)$, is a rational code.*

Proof. By Theorem 5.3.2, any rational subset of Σ^+ has a rational primitive root. Moreover, by Theorem 5.5.7, a code has a unique primitive root. Thus, for any rational code C, its unique primitive root $\rho(C)$ is rational. By Corollary 5.5.3, it follows that $\rho(C)$ is also a code. □

The following result establishes the connection between Conway's Problem and the BTC-property. In particular, this result shows that a property concerning the commutation of a code with an arbitrary set of words is equivalent with a property concerning the code itself.

Theorem 6.7.2. *Let X be a code and let $\rho(X)$ be its primitive root. The BTC-property holds for X if and only if $\mathcal{C}(X) = \rho(X)^*$.*

Proof. Let us assume that $\mathcal{C}(X) = \rho(X)^*$ and let Y be a language Y commuting with X. Then, clearly, $Y \subseteq \rho(X)^*$. We prove that if $Y \cap \rho(X)^k \neq \emptyset$, for some $k \geq 0$, then $\rho(X)^k \subseteq Y$, which then implies that the BTC-property holds for X.

Let $m > 0$ be such that $X = \rho(X)^m$ and let $k \geq 0$ be such that $Y \cap \rho(X)^k \neq \emptyset$. Thus, there are $x_i \in X_0$, $1 \leq i \leq k$ such that $x_1 \dots x_k \in Y$. Consequently, for any $y_j \in \rho(X)$, $1 \leq j \leq m$, we have $x_1 \dots x_k y_1 \dots y_m \in YX = XY$. Thus, $x_1 \dots x_k y_1 \dots y_m = \rho_1 \dots \rho_m y$, with $\rho_i \in \rho(X)$ and $y \in Y \subseteq \rho(X)^*$. This is a relation on $\rho(X)$. Since $\rho(X)$ is a code, it follows that the relation must be a trivial one. Thus, $y = y_{m-k+1} \dots y_m \in Y$, if $m \geq k$, and $y = x_{m+1} \dots x_k y_1 \dots y_m \in Y$, if $m < k$. In the first case, the claim follows. In the second one, we iterate the argument.

For the reverse implication, if the BTC-property holds for X, then, since $\mathcal{C}(X)$ commutes with X and any root of X is a power of $\rho(X)$, it follows that $\mathcal{C}(X) = \cup_{i \in I} \rho(X)^i$. Due to the maximality of the centralizer, we must have $I = \mathbb{N}$ and thus, $\mathcal{C}(X) = \rho(X)^*$. □

It is a simple consequence of Theorems 6.7.1 and 6.7.2, that in the case of rational codes, the BTC-property implies a positive answer to Conway's problem.

Theorem 6.7.3. *If the BTC-property holds for a rational code X, then $\mathcal{C}(X)$ is rational. Moreover, in this case, $\mathcal{C}(X) = \rho(X)^*$.*

Using these results, it follows now that Conway's Problem has a positive answer for rational (reversed) ω-codes, since these families of codes have the BTC-property, as shown in Theorem 6.6.3 and Corollary 6.6.4.

Theorem 6.7.4. *Conway's Problem has an affirmative answer in the case of rational ω-codes and rational reversed ω-codes. Moreover, for any such code X, $\mathcal{C}(X) = \rho(X)^*$.*

We also obtain the following consequence, similar to Corollaries 6.6.5, 6.6.6, and 6.6.7.

Corollary 6.7.5. *Conway's Problem has an affirmative answer in the case of rational prefix codes, binary, ternary, and elementary codes. Moreover, if F is a binary, ternary, or elementary code, then $\mathcal{C}(F) = F^*$.*

6.7.2 Some special cases

In this subsection, we solve Conway's problem for some other types of codes as well.

Observe that for any language not containing the empty word, the set of minimal length words (and also the set of all words of any given length) is a finite (biprefix) code. For a language L, let L_{min} be the set of its minimal length words.

We can also answer affirmatively to Conway's Problem for another class of codes, those having a primitive set of minimal length words.

Theorem 6.7.6. *Let X be a code such that X_{min} is a primitive code. Then X is primitive, $\mathcal{C}(X) = X^*$, and the BTC-property holds for X.*

Proof. If $X = R^n$, $n \geq 1$, then $X_{min} = R^n_{min}$, and so, $n = 1$. Thus, X is primitive.

If all languages commuting with X are subsets of X^*, then $\mathcal{C}(X) = X^*$ and the claim follows by Theorem 6.7.2. Let us assume that there is a language L commuting with X, such that $L \not\subseteq X^*$. Then $L \setminus X^*$ also commutes with X, see [66], and thus, we can assume without loss of

generality that $L \cap X^* = \emptyset$. Then clearly, $L_{min}X_{min} = X_{min}L_{min}$ is a commutation of codes and so, by Theorem 5.5.8, L_{min} and X_{min} have the same primitive root. Since X_{min} is primitive, we obtain that $L_{min} \subseteq X_{min}^* \subseteq X^*$, and so, $L \cap X^* \neq \emptyset$, a contradiction. □

Corollary 6.7.7. *Let X be a code such that either one of the following conditions is satisfied:*

(i) X_{min} contains primitive words only, or

(ii) $|X_{min}| \neq n^k$, for any $n \geq 1$, $k \geq 2$.

Then X is primitive, $\mathcal{C}(X) = X^*$, and the BTC-property holds for X.

Proof. In both cases, X_{min} is primitive and the claim follows by Theorem 6.7.6. □

6.8 Discussion

We have discussed in this chapter the commutation of two sets of words. Although the subsets of Σ^* do not form a free monoid, there are classes of languages for which the commutation with an arbitrary language can be characterized as in free monoids. We called this the BTC-property, i.e., Bergman-type of characterization.

Using the combinatorial approach, we proved that all periodic and binary sets of words have the BTC-property. This, however, is not true for sets with at least four words, as shown in Example 6.2.1. On the boundary point, that of ternary sets, we proved that the BTC-property holds for all ternary codes.

A different strategy, the so called multiplicity approach, turned out to be very fruitful. We translated the commutation of two sets of words to the commutation of their characteristic formal power series. Then, using a result of Cohn characterizing the commutation of two formal power series, we solved this new problem and translated the result back to words, obtaining a solution of the original problem. We thus proved that the BTC-property holds for all ω-codes and reversed ω-codes. The case of codes that are not ω-codes or reversed ω-codes remains open.

Problem 6.8.1. Prove that the BTC-property holds for all codes. Equivalently, prove that for any code C, a set of words X commutes with C if and only if

$$X = \cup_{i \in I} \rho(C)^i,$$

for some $I \subseteq \mathbb{N}$, where $\rho(C)$ is the primitive root of C.

As a matter of fact, it is conceivable that the multiplicity approach can give a solution to this problem, in its full generality. One has to prove that

in a commutation $XY = YX$, if X is a code, then both products XY and YX are unambiguous, or in any case, they have the same ambiguities. This is true if Y is also a code, see [66], which was the essential tool in Chapter 5 in establishing the uniqueness of the primitive root of a code. The result would then follow using Cohn's theorem, similarly as for ω-codes.

Another possible approach to Problem 6.8.1 is to use Theorem 6.7.2. We proved in that result that the BTC property is equivalent for codes X with proving that $\mathcal{C}(X) = \rho(X)^*$. This is a particularly interesting result since it proves that a property concerning the commutation of a code with an *arbitrary* set of words is equivalent with a property concerning only the *code itself* and two notions associated to it: its centralizer and its primitive root. These notions have been discussed in details in Chapters 4 and 5, respectively.

A similar result, proving that the BTC-property for a ternary set of words F is equivalent with proving that $\mathcal{C}(F) = F^*$, has been established in Theorem 6.5.2, using some different arguments. However, it remains open whether or not the BTC-property holds for all ternary set.

Problem 6.8.2. Prove that the BTC-property holds for all ternary sets of words $F \subseteq \Sigma^+$. Equivalently, prove that for any set of words X commuting with F, there is $I \subseteq \mathbb{N}$, such that

$$X = \bigcup_{i \in I} F^i.$$

Most likely, a solution to this problem should first establish that $\mathcal{C}(F) = F^*$, for all ternary sets of words and then conclude the result using Theorem 6.5.2. We refer to the discussion in Section 4.6 for several possible strategies to prove that $\mathcal{C}(F) = F^*$, for an arbitrary three word set $F \subseteq \Sigma^+$.

We conclude this chapter with a very interesting and seemingly very challenging problem concerning the commutation of two sets of words.

Problem 6.8.3. Characterize all sets of words satisfying the BTC-property.

Chapter 7

Conclusions

We have investigated in this thesis some problems related to commutation on sets of words and on formal power series. In the first part of the thesis we initiated the study of semilinearity for formal power series. We introduced two families of formal power series, both natural generalizations of the semilinear languages, and we investigated their behaviour under some elementary operations. As it turned out, the two families of series do not have the same behaviour under rational operations, morphisms, and Hadamard product. The operation of difference proved to be challenging on semilinear series, just as it is on rational series: the $\mathbb{N}$-linear series are not closed under difference and we believe that the same holds also for semilinear series, even when one of the series has bounded coefficients; we proved however, that in this case, the difference is always a rational series of star-height one.

Turning to commutation on languages in the second part of the thesis, we proved that any code has a unique primitive root that is also a code, and that two codes commute if and only if they have the same primitive root, similarly as in free monoids, thus solving two conjectures of Ratoandromanana, [66], 1989. We also proved that a language commutes with an ω-code X if and only if it is a union of powers of $\rho(X)$, where $\rho(X)$ is the primitive root of X, giving a partial solution to another conjecture of [66]; the case of codes that are not ω-codes remains open. We answered Conway's problem - asking whether or not the centralizer of any rational set is rational - in the case of periodic, binary, and ternary sets of words, as well as for rational ω-codes, the most general results on this problem.

At the end of each of the chapters of this thesis, we included some discussion and several open problems, showing that many interesting issues are still to be solved. E.g., the theory of semilinearity for formal power series is clearly only at the beginning and many contributions are still needed. For instance, one clearly needs some tools to prove that a series is not semilinear, and this can be achieved through some characterizations of the semilinear series and sequences. We only investigated in this thesis the difference oper-

ation for semilinear power series and it appears that this operation is equally interesting, though seemingly more difficult, for bounded power series. We recall here only one of the open problems stated in Chapter 3, regarding a possibly infinite hierarchy of rational series in commuting variables.

Problem. How many of the following inclusions are strict:

$$\mathbb{N}_0^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle \subseteq \mathbb{N}_1^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle \subseteq \mathbb{N}_2^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle \subseteq \ldots \subseteq \mathbb{N}^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle,$$

where $\mathbb{N}_i^{Rat}\langle\langle\Sigma^{\oplus}\rangle\rangle$ is the set of the $\mathbb{N}$-rational series over $\Sigma^{\oplus}$, having star-height i ?

For Conway's problem and the other related questions on commutation of languages we have developed a variety of different strategies, involving various combinatorial and algebraic techniques: the fixed-point approach, the combinatorial approach, the equational method, the branching point approach, and the multiplicity approach. Despite all these efforts, Conway's problem remains unanswered in its full generality, and even much weaker questions, such as whether or not the centralizer of any finite set is rational, or at least recursively enumerable, are still to receive a satisfactory answer. We believe that some of our techniques can be refined to give further insight on these problems. The most promising ones in our opinion are the branching point approach - for Conway's problem, and the multiplicity approach - for the commutation with codes and for Conway's problem in the case of rational codes. We conclude this discussion by recalling the BTC-problem and Conway's question.

Problem. Prove that the BTC-property holds for all codes. Formally, prove that for any code C, the set of words X commutes with C if and only if $X = \cup_{i\in I}\rho(C)^i$, for some $I \subseteq \mathbb{N}$, where $\rho(C)$ is the primitive root of C.

Problem (Conway's Problem). Is it true that for any rational language, its centralizer is rational ?

Amazingly, Conway's question remains unanswered up to date, even in much weaker instances, such as, e.g., when it is asked whether or not the centralizer of any finite set of words is recursively enumerable.

Bibliography

[1] L. Aceto, Z. Ésik, A. Ingólfsdóttir, A fully equational proof of Parikh's theorem, BRICS technical report, http://www.brics.dk/, 2001.

[2] J.M. Autebert, J. Berstel, L. Boasson, Context-Free Languages and Pushdown Automata. In G. Rozenberg, A. Salomaa (eds.), *Handbook of Formal Languages*, vol. 1: 329-438, Springer-Verlag, 1997.

[3] J.M. Autebert, L. Boasson, M. Latteux, Motifs et bases de langages, *RAIRO Inform. Theor.*, 23(4): 379-393, 1989.

[4] F. Bassino, Nonnegative companion matrices and star-height of $\mathbb{N}$-rational series, Theoret. Comput. Sci. 180, 61–80, 1997.

[5] G. Bergman, Centralizers in free associative algebras, *Transactions of the American Mathematical Society* 137: 327–344, 1969.

[6] J. Berstel, *Transductions and Context-free Languages*, B.G. Teubner, Stuttgart, 1979.

[7] J. Berstel, D. Perrin, *Theory of Codes*, Academic Press, New York, 1985.

[8] J. Berstel, C. Reutenauer, *Rational series and their languages*, Springer-Verlag, 1988.

[9] N. Bourbaki, *Éléments de mathématique. Algèbre.* Chapitres 1 à 3, Hermann, Paris, 1970.

[10] N. Bourbaki, *Éléments de mathématique. Algèbre commutative.* Chapitres 1 à 7, Hermann, Paris, 1961, 1964, 1965.

[11] C. Choffrut, J. Karhumäki, Combinatorics of Words. In G. Rozenberg, A. Salomaa (eds.), *Handbook of Formal Languages*, vol. 1: 329-438, Springer-Verlag, 1997.

[12] C. Choffrut, J. Karhumäki, On Fatou properties of rational languages, in C.Martin-Vide, V.Mitrana (Eds.), *Where mathematics, Computer Science, Linguistics and Biology Meet*, Kluwer, Dordrecht, 2000.

[13] C. Choffrut, J. Karhumäki, N.Ollinger, The commutation of finite sets: a challenging problem, *Theoret. Comput. Sci.*, 273 (1-2): 69–79, 2002.

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

[15] P.M.Cohn, *Algebra*, Vol. 1-3, John Wiley & Sons, Chichester, 1982, 1989, 1991.

[16] P.M. Cohn, Factorization in noncommuting power series rings, *Proc. Cambridge Philos. Soc.* 58: 452–464, 1962.

[17] P.M. Cohn, Centralisateurs dans les corps libres, in J. Berstel (Ed.), *Séries formelles*: 45–54, Paris, 1978.

[18] J.H. Conway, *Regular Algebra and Finite Machines*, Chapman Hall, 1971.

[19] M. Droste and P. Gastin, On recognizable and rational formal power series in partially commuting variables. *Automata, languages and programming (Bologna, 1997)*, 682–692, LNCS 1256, Springer, Berlin, 1997.

[20] L.C. Eggan, Transition graphs and star height of regular events, *Michigan Math. J.* 10, 385–397, 1963.

[21] S. Eilenberg, *Automata, Languages and Machines*, Academic Press, 1974.

[22] S. Eilenberg, M.P. Schützenberger, Rational sets in commutative monoids, J. Algebra 13, 173-191, 1969.

[23] *Encyclopedia of Mathematics*, Kluwer Academic Publishers, 1988.

[24] S. Ginsburg, *The Mathematical Theory of Context-free Languages* , McGraw-Hill Book Company, 1966.

[25] M. A. Harrison, *Introduction to formal language theory*, Addison-Wesley Publishing Co., Reading, Mass., 1978.

[26] K. Hashiguchi, Relative star-height, star-height and finite automata with distance functions, in J.E. Pin (Ed.), *Formal Properties of Finite Automata and Applications*, LNCS 386, Springer Berlin: 74–88, 1989.

[27] T. Harju, O. Ibarra, J. Karhumäki, A.Salomaa, Decision questions concerning semilinearity, morphisms and commutation of languages, in *Proc. of ICALP 2001*, Springer LNCS, 573–591, 2001; complete version in J. Comput. Syst. Sci., to appear.

[28] T. Harju, L. Ilie, G. Rozenberg, A. Salomaa, Notes on language equations, manuscript, 2001.

[29] T. Harju, I. Petre, On commutation and primitive roots of codes, submitted. A preliminary version of this paper has been presented at WORDS 2001, Palermo, Italy.

[30] M.W. Hopkins, D. Kozen, Parikh's theorme in commutative Kleene algebra, in *Proc. IEEE Conf. Logic in Computer Science (LICS'99)*, IEEE Press, 394–401, 1999.

[31] L. Ilie, I. Petre, G. Rozenberg, Uniformly scattered factors, in C. Calude, Gh. Paun (Eds.), *Finite Versus Infinite. Mathematical Contributions to an Eternal Dilemma*, Springer-Verlag, London: 187–198, 2000.

[32] N. Jacobson, *Structure of rings*, American Mathematical Society, 1956.

[33] N. Jacobson, *Basic Algebra*, Vol. 1-2, W. H. Freeman and Company, New York, 1985, 1989.

[34] J. Karhumäki, On three-element codes, *Theoret. Comput. Sci.* (40), 3–11, 1985.

[35] J. Karhumäki, A property of three-element codes, *Theoret. Comput. Sci.* (41), 215–222, 1985.

[36] J. Karhumäki, Challenges of commutation: an advertisement, in *Proc. of FCT 2001*, LNCS 2138, 15–23, Springer, 2001.

[37] J. Karhumäki, I. Petre, On the centralizer of a finite set, In *Proc. ICALP 2000*, LNCS 1853 536–546, Springer, 2000.

[38] J. Karhumäki, I. Petre, Conway's Problem for three word sets, *Theoret. Comput. Sci.*, to appear.

[39] J. Karhumäki, I. Petre, Conway's problem and the commutation of languages, *Bulletin of EATCS* 74, 171–177, 2001.

[40] J. Karhumäki, I. Petre, The branching point approach to Conway's problem, LNCS 2300 69–76, Springer, 2002.

[41] S. Kleene, Representation of events in nerve nets and finite automata, in *Automata Studies*, Annals of Mathematical Studies 34, 1956.

[42] W. Kuich, The Kleene and the Parikh theorem in complete semirings, *Automata, Languages and Programming* (Karlsruhe 1987), 211–215, Lecture Notes in Computer Science, Springer-Verlag, 1987.

[43] W. Kuich, Semirings and formal power series: their relevance to formal languages and automata. In G. Rozenberg, A.Salomaa (eds.), *Handbook of Formal Languages*, Springer-Verlag, Berlin, Heidelberg: 609–678, 1997.

[44] W. Kuich, On the entropy of context-free languages, Inf. Control 16, 173–200, 1970.

[45] W. Kuich, A. Salomaa, *Semirings, Automata, Languages*, Springer-Verlag, 1986.

[46] E. Leiss, *Language Equations*, Springer, 1998.

[47] M. Lothaire, *Combinatorics on Words*, Addison-Wesley, Reading, MA., 1983.

[48] M. Lothaire, *Algebraic Combinatorics on Words*, Cambridge University Press, to appear.

[49] B. Mahr, Iteration and summability in semirings, in *Algebraic and combinatorial methods in operations research*, 229–256, North-Holland, Amsterdam, 1984.

[50] B. Mandelbrot, On recurrent noise limiting coding, *Proc. Symp. on Inf. Networks*, Polytechn. Inst. of Brooklyn, 205–221, 1954.

[51] A. Mateescu, A. Salomaa, Formal Languages: an Introduction and a Synopsis. In G. Rozenberg, A. Salomaa (eds.), *Handbook of Formal Languages*, vol. 1: 329-438, Springer-Verlag, 1997.

[52] A. Mateescu, A. Salomaa, Aspects of Classical Language Theory. In G. Rozenberg, A. Salomaa (eds.), *Handbook of Formal Languages*, vol. 1: 329-438, Springer-Verlag, 1997.

[53] A. Mateescu, A. Salomaa, S. Yu, On the decomposition of finite languages, TUCS Technical Report 222, http://www.tucs.fi/, 1998.

[54] M. Minski, *Computation: finite and infinite machines*, Prentice-Hall, 1967.

[55] E. Ochmański, Regular behaviour of concurrent systems, Bulletin of the EATCS, 27:56–67, 1985.

[56] R. J. Parikh, Language generating devices, M.I.T. Res. Lab. Electron. Quart. Prog. Rep. 60, 1961.

[57] D. Perrin, Codes conjugués, Information and Control 20: 222–231, 1972.

[58] D. Perrin, Finite automata. In In J. van Leeuwen (ed.), *Handbook of Theoretical Computer Science*, Vol. B, North-Holland, 1–57, 1990.

[59] I. Petre, The Parikh's theorem does not hold for multiplicities, Journal of Automata, Languages, and Combinatorics **4**(1): 17–30, 1999.

[60] I. Petre, On semilinearity in formal power series, In G.Rozenberg, W.Thomas (eds.), *Proceedings of DLT 1999*, *Developments in Language Theory: foundations, applications, and perspectives*, World Scientific: 220–231, 2000.

[61] I. Petre, On the difference problem for semilinear power series, In V. Mitrana, C. Martin-Vide (eds.), *Grammars and Automata for String Processing: from Mathematics and Computer Science to Biology, and Back*, to appear.

[62] I. Petre, Recent results on the semilinear formal power series, Bulletin de la Societe Mathematique de Belgique **8**(2001): 323–333, 2001. A previous version of this paper was presented at *Journées Montoises d'Informatique Théorique*, Marne-la-Vallée, 2000.

[63] D.L. Pilling, Commutative regular equations and Parikh's theorem, J. London Math. Soc. 6: 663-666, 1973.

[64] E.L. Post, Absolutely unsolvable problems and relatively undecidable propositions - account of an anticipation, unpublished manuscript, 1941.

[65] E.L. Post, Formal reductions of the general combinatorial decision problem, Am. Journal of Math. 65: 197–268, 1943.

[66] B. Ratoandromanana, Codes et motifs, *RAIRO Inform. Theor.*, 23(4): 425-444, 1989.

[67] C. Reutenauer, Inversion height in free fields, *Selecta Mathematica. New Series* 2(1), 1–18, 1996.

[68] G. Rozenberg, A. Salomaa, *Cornerstones of Undecidability*, Prentice Hall, 1994.

[69] J. Sakarovitch, *Éléments de théorie des automates*, manuscript.

[70] A. Salomaa, *Formal Languages*, Academic Press, 1973.

[71] A. Salomaa, M. Soittola, *Automata-theoretic Aspects of Formal Power Series*, Springer-Verlag, 1978.

[72] M.P. Schützenberger, On the definition of a family of automata, *Information and Control*, 4: 245–270, 1961.

[73] H.J. Shyr, *Free monoids and languages*, Hon Min Book Company, 1991, 2001.

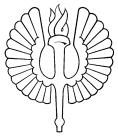
[74] A. Turing, On computable numbers, with an application to the Entscheidungsproblem, *Proceedings of the London Mathematical Society* (2), 42: 230–265, 1937.

[75] S. Yu, Regular Languages. In G. Rozenberg, A. Salomaa (eds.), *Handbook of Formal Languages*, vol. 1: 329-438, Springer-Verlag, 1997.

Index

Turku Centre for Computer Science
Lemminkäisenkatu 14
FIN-20520 Turku
Finland

http://www.tucs.abo.fi



University of Turku
- **Department of Mathematical Sciences**

Åbo Akademi University
- **Department of Computer Science**
- **Institute for Advanced Management Systems Research**

Turku School of Economics and Business Administration
- **Institute of Information Systems Science**