ON PRIMITIVE SOLUTIONS OF THE POST CORRESPONDENCE PROBLEM

by

Marjo Lipponen

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 April 27th, 1996, at 12 noon

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

1996

Supervised by

Professor Arto Salomaa
Department of Mathematics
University of Turku
Turku, Finland

Reviewed by

Professor Sheng Yu
Department of Computer Science
University of Western Ontario
London, Ontario, Canada

Professor Gheorghe Păun
Institute of Mathematics
Romanian Academy
Bucharest, Romania

ISBN 951-650-717-4
ISSN 1239-1883

Acknowledgements

I owe my deepest gratitude to my supervisor, Professor Arto Salomaa. His friendly guidance and expert advice have been invaluable throughout this work. Yet, he has always encouraged me to do things in my own way.

I would also like to thank Professor Sheng Yu and Professor Gheorghe Păun for reading the manuscript and giving valuable comments and encouraging remarks. This work has also benefited from the inspiring discussions with Professor Alexandru Mateescu.

Special thanks are due to my friend Virpi Kauko, for drawing the "primitive" picture on the cover. The instance (g, h) in the cave painting is a slightly changed version of Emil Post's original example (1946).

Excellent working conditions for this work were provided by Turku Centre for Computer Science, TUCS, and the Department of Mathematics of the University of Turku. I take this opportunity to thank my teachers and friends in both places for creating a very pleasant working environment. For financial support I would like to thank the Academy of Finland and Suomen Kulttuurirahasto (Elina, Sofia and Yrjö Arffman foundation).

Finally, I would like to acknowledge my debt to Keijo, not only for his support, but also for making some of my algorithms into programs. Real patience was needed when one of the computer runs took three weeks.

Turku, February 21st, 1996

Marjo Lipponen

Contents

Chapter 1

Introduction

Decision problems play an essential role in formal language theory. To prove a problem decidable it suffices to find an appropriate algorithm, while undecidability means that there is no algorithm which outputs an answer for each instance. Thus, the notion of an algorithm has to be formalized for proofs of undecidability. The most common way to show a problem undecidable is to 'reduce' to it another problem already known to be undecidable. In language theory, as well as in many other areas of computer science, the most convenient among such problems is the Post Correspondence Problem, due to E. Post [27], because it involves strings of letters. Indeed, the Post Correspondence Problem is one of the "cornerstones of undecidability", [28]. The study of primitive solutions of PCP gives one way to analyze this important problem.

The suitability of the Post Correspondence Problem for reduction arguments lies also in the fact that in some sense it captures the essence of computations and grammatical derivations. In [25] a variant of the PCP, the machine-oriented Post Correspondence Problem is even used to represent deterministic and nondeterministic time and space complexities leading to a hierarchy of complexity classes which is nontrivial if and only if $P \neq NP$. Thus simple solutions of PCP mean simplifications of computations and the results contribute on an abstract level to the understanding of computations.

Prime solutions have turned out to be theoretically important in various other contexts, too. For example, in the purely morphic characterization of recursively enumerable languages given in [2], attention is restricted to solutions whose proper prefixes are not allowed to be solutions. Because of this restriction, it becomes unnecessary to have the intersection with a regular language in the characterization. Such solutions are in the sequel called *F-prime solutions*: no proper final subword can be removed such that the remainder is still a solution. Similarly, a solution for an instance of the PCP is termed an *S-prime solution* if no subword can be removed, and a *P-prime solution* if no scattered subword can be removed. (In both of these cases the remainder is now the catenation of the remaining parts.) These three types of prime solutions were first introduced by A. and K. Salomaa and S. Yu [32]. In [24] A. Mateescu and A. Salomaa extended the investigation to concern analogous problems for the target alphabet, which we do not consider in this thesis.

A more general point of view is taken by considering words that can appear as a prime solution, for some instance of PCP. The study of *P-, S-* and *F-words*, initiated by A. Mateescu and A. Salomaa [23], can be viewed also as a contribution to the general combinatorics of words. The property of being an S-word, for instance, is an abstract property that is so far rather poorly understood: of two similar-looking words, one is an S-word and the other is not.

An interesting subclass of prime words are the *strongly prime PCP-words* which are prime solutions for every instance of PCP for which they are solutions.

There are also other methods to define how some solutions are more primitive than other solutions. Partial orders give a natural way to make comparisons between solutions. We can also combine languages (and thus language theory) to the basic definitions of P-, S- and F-prime solutions by requiring that the part to be removed (resp. the remaining part) belongs to the given language.

Prime solutions can be considered also from the point of view of prime languages. There are actually two possible ways, introduced by A. Mateescu and A. Salomaa [23], to define them. In the equality method, which we find more interesting because of its strict relation to the instances of PCP, L is termed a *prime language* if for some instance of PCP it equals to the set of its prime solutions, while in the inclusion method the smaller subsets are also taken into account.

We now take a closer look at the structure of the thesis.

In Sections 1.1–1.3 we define the frequently used concepts of formal language theory as well as the Post Correspondence Problem.

Chapter 2 analyzes the basic primality types, P-, S- and F-words. We prove that these three sets are all recursive and that for F-words we have even a practical algorithm. For P- and S-words we have only partial algorithms, we know that some of them have the given property but not all of them. The hierarchy of these three types is also closely studied. In our main results we prove that the inclusions are strict in every alphabet except the binary case for P- and S-words which remains open. One section is devoted to a closely related problem, periodicity forcing words. Such a property is the most convenient way to show that a word is not an S-word or a P-word.

Chapter 3 deals with the special primality types, such as strongly prime PCP-words, partial orders and L-prime solutions as defined above. We study the hierarchies of these new types as well as their relation to the basic types of Chapter 2. In connection with L-prime solutions we also investigate how different set-theoretic operations affect them; for instance, how can $L_1 \cup L_2-$ or $L_1 \cap L_2-$prime solutions be expressed in terms of L_1- and L_2-prime solutions.

Chapter 4 is dedicated to the theory of prime languages defined in both senses. We analyze them especially in the binary alphabet though some results can be extended to larger alphabets, as well. The relation to the Chomsky hierarchy is completely settled in both equality and inclusion sense. The finiteness of P-languages makes it natural to look for the maximal numbers of P-prime solutions in terms of the size of the alphabet. We present an algorithm which settles the maximal numbers in periodic instances and present evidence for these numbers being maximal also in the general case.

1.1 Basics of language theory

A language, whether a natural language such as Finnish or a programming language such as C, can be considered to be a set of sentences, that is, finite strings of elements of some basic vocabulary. Thus the following definitions for the formal language theory are very natural. For more details of formal language theory, see [30, 29, 11].

An *alphabet* Σ is a finite nonempty set, the elements of which are assumed to be *indivisible symbols* and called *letters*.

A *word* over an alphabet Σ is a finite string consisting of zero or more letters of Σ, whereby the same letter may occur several times. A typical nonempty word is written as

$$w = a_1 \dots a_k, \quad a_i \in \Sigma,$$

and the *empty word*, a string with zero letters, as λ. The *mirror image* of w is obtained by writing w backwards

$$\mathrm{mi}\,(w) = a_k \dots a_1.$$

The set of all words over an alphabet Σ is denoted by Σ^* and the set of all nonempty words, $\Sigma^* - \{\lambda\}$, by Σ^+.

The *catenation* uv of two words $u, v \in \Sigma^*$ is also a word over Σ. Under this associative operation the set Σ^* is a monoid where the empty word λ is an identity: $w\lambda = \lambda w = w$ for any word w. We use w^i to indicate the word obtained by catenating i copies of the word w, with the convention $w^0 = \lambda$.

A *subword* of a word w is a word x such that $w = uxv$ for some $u, v \in \Sigma^*$. If $u = \lambda$ (respectively $v = \lambda$) then x is termed a *prefix* (resp. a *suffix*) of w. A *scattered subword* of w is a word $x = x_1 x_2 \dots x_n$ such that $w = u_1 x_1 u_2 x_2 \dots u_n x_n u_{n+1}$ for some $u_i \in \Sigma^*$. In all these notions x is said to be *proper* if $x \neq \lambda$ and $x \neq w$.

The *length* of the word $w = a_1 a_2 \dots a_k$, $a_i \in \Sigma$, denoted by $|w|$, is the number of occurrences of letters of Σ in w, $|w| = k$. Obviously, $|\lambda| = 0$ and $|uv| = |u| + |v|$.

We may also count how many times a letter occurs in w. Denoting the number of occurrences of a by $|w|_a$, we have clearly

$$|w| = \sum_{a \in \Sigma} |w|_a.$$

By $\mathrm{alph}\,(w)$ we denote the subset of the alphabet Σ consisting of the letters actually occurring in w. Hence

$$\mathrm{alph}\,(w) = \{a \in \Sigma \mid |w|_a \geq 1\}.$$

In most of our study we assume that $\Sigma = \mathrm{alph}\,(w)$.

Combining these two notions, a *Parikh mapping* or a *Parikh vector* describes the structure of the word in terms of its letters when their actual locations are ignored. For $\Sigma = \{a_1, \dots, a_n\}$ we define a mapping from Σ^* into $\mathbb{N}^n$ by

$$\psi(w) = (|w|_{a_1}, \dots, |w|_{a_n}).$$

If this vector is divided by the greatest common divisor of its components, we talk about a *basic Parikh vector*, which, in a binary alphabet, is closely related to the notion of the *ratio* of w, $r(w) = \frac{|w|_{a_1}}{|w|_{a_2}}$. We have, for instance,

$r(a^2) = \infty$, $r(b) = 0$ and $r(a^2bab) = \frac{3}{2}$, whereas the corresponding basic Parikh vectors are $(1,0)$, $(0,1)$ and $(3,2)$.

A word $w \in \Sigma^+$ is said to be *primitive* if it is not a power of another word; that is, if $w = z^i$ for some $z \in \Sigma^+$ then $i = 1$ (and $w = z$). It is well known (see for instance [20]) that for each $w \in \Sigma^+$ there exists a unique primitive word $\rho(w) \in \Sigma^+$, the *primitive root* of w, such that $w \in \rho(w)^i$ for some $i \in \mathbb{N}$.

The following combinatorial results of words are frequently needed in the sequel. For the proofs, see for instance [21, 20].

Lemma 1.1 *If $uv = vz$ holds for some words u, v, z, where u is nonempty, then*

$$u = xy,\ v = (xy)^k x,\ z = yx,$$

for some words x, y and integer $k \geq 0$. If $uv = vu$ holds for some words u and v then u and v are powers of the same word.

Lemma 1.2 *For nonempty words u and v the following three conditions are equivalent:*

1. *$uv = vu$,*
2. *$u = w^i$ and $v = w^j$ for some word w and integers $i, j \geq 1$,*
3. *$u^m = v^n$ for some integers $m, n \geq 1$.*

Lemma 1.3 *For u, v, w in Σ^+ and $m, n, p \geq 2$, if $u^m v^n = w^p$ then u, v and w are powers of the same word.*

Lemma 1.4 *For u, v, w in Σ^* and $i \geq 1$, the equation $uwv = w^i$ implies that u, v and w are powers of the same word.*

For alphabets Σ and Δ, a mapping $\delta : \Sigma^* \rightarrow \Delta^*$ is termed a *morphism* if $\delta(uv) = \delta(u)\delta(v)$ is satisfied for all words $u, v \in \Sigma^*$. Clearly, to specify a morphism δ it suffices to specify all values $\delta(a)$, where a ranges over the letters of Σ. If none of these values equals the empty word λ, the morphism is called *nonerasing*.

A *language* over Σ is any subset of Σ^*. We denote its cardinality, the number of its elements, by card(L) or $|L|$.

Example 1.1 The sets

$$\begin{aligned} L_1 &= \{0, 00, 001\}, \\ L_2 &= \{0^i \mid i \text{ is a prime}\}, \text{and} \\ L_3 &= \{\lambda\} \end{aligned}$$

are all languages over the alphabet $\Sigma = \{0, 1\}$. Notice that L_3 is not the same as the *empty language* $\emptyset$. □

Since languages are sets, we may define the set-theoretic operations of union, intersection, difference and complement in the usual way:

$$\begin{aligned} L_1 \cup L_2 &= \{w \mid w \in L_1 \text{ or } w \in L_2\}, \\ L_1 \cap L_2 &= \{w \mid w \in L_1 \text{ and } w \in L_2\}, \\ L_1 - L_2 &= \{w \mid w \in L_1 \text{ and } w \notin L_2\}, \\ L^c &= \Sigma^* - L, \end{aligned}$$

where L is over the alphabet Σ. The catenation L_1L_2 of two languages L_1 and L_2 is defined by

$$L_1L_2 = \{uv \mid u \in L_1 \text{ and } v \in L_2\}.$$

In the same way as for words, this operation is associative, L^i is the catenation of i copies of L and L^0 is the language consisting of the empty word λ. Furthermore, for any language L,

$$L\emptyset = \emptyset L = \emptyset \quad \text{and} \quad L\lambda = \lambda L = L$$

where the singleton $\{\lambda\}$ is identified with its element λ, a usual practice also in the sequel.

The *catenation closure* or *Kleene star* is the union of all powers of L,

$$L^* = \bigcup_{i=0}^{\infty} L^i,$$

and the *λ-free catenation closure* or *Kleene plus* is the union of all positive powers of L,

$$L^+ = \bigcup_{i=1}^{\infty} L^i.$$

A *decision problem* consists of an infinity of instances to each of which the answer is either "yes" or "no". Such a problem is *decidable* if and only if there exists an algorithm (effective procedure) which for any instance of the problem outputs the answer for that instance. If no such algorithm exists, the problem is *undecidable*. The customary way of establishing the undecidability of a certain problem P is to 'reduce' to it a problem Q, which is known to be undecidable. In language theory the most convenient among such problems is the Post Correspondence Problem.

1.2 Post Correspondence Problem

By an *instance* of the Post Correspondence Problem we mean a pair (g,h) of nonerasing morphisms $g,h:\Sigma^*\to\Delta^*$ where Σ and Δ are finite alphabets, and by its *solutions* the words in the *equality set*

$$E(g,h)=\{w\in\Sigma^+ \mid g(w)=h(w)\}.$$

Hence a solution is a nonempty word whose images under g and h coincide. If $g(a)=h(a)$ for some $a\in\Sigma$, we call such an instance (g,h) *trivial*.

In our examples the domain and range alphabets Σ and Δ of the morphisms will be clear from the definitions. We will often use numerals as letters of Σ to point out the customary definition of PCP as two lists of words. Indeed, according to the original definition, [27], the Post Correspondence Problem is an ordered quadruple PCP=(Σ,n,α,β) where Σ is an alphabet, $n\geq 1$, and

$$\alpha=(\alpha_1,\ldots,\alpha_n),\quad \beta=(\beta_1,\ldots,\beta_n)$$

are ordered n-tuples of nonempty words over Σ. A solution of this instance is any nonempty finite sequence of indices $i_1,\ldots,i_k$ such that

$$\alpha_{i_1}\ldots\alpha_{i_k}=\beta_{i_1}\ldots\beta_{i_k}.$$

It is well known that the Post Correspondence Problem is *undecidable*: no algorithm decides for an arbitrary instance PCP whether or not it has a solution. Indeed, the Post Correspondence Problem is one of the "cornerstones of undecidability", [28], because of its importance in reduction arguments.

If we assume that the domain alphabet Σ is of a bounded size n, we consider a simplification of PCP, namely PCP(n). In this case we have a borderline between decidability and undecidability. The problem PCP(n), $n \leq 2$, is decidable, see [6], whereas PCP(n), $n \geq 9$, is an undecidable problem, see [27, 26]. For the remaining n, $3 \leq n \leq 8$, the status of PCP(n) is still open, see [13].

Example 1.2 Let g and h be defined by the table

	1	2
g	a^2b	a
h	a	ba^2

Any solution must begin with 1. Since $h(1)$ has one a less than $g(1)$ we must choose 1 again. Comparing $h(11)$ with $g(11)$, we see that for the next letter, only 2 is available. In the same way, the following letter must be 2 again. Thus $E(g,h) = (1122)^+$. $\square$

Since the catenation of two solutions is always a solution, the equality set is either empty or has infinitely many elements. This is why we try to consider solutions that are somehow simpler than others. For instance, if w_1 and w_2 are solutions, they are clearly both simpler than w_1w_2.

But first we consider an interesting subclass of the Post Correspondence Problem, known to be decidable, where all the values $h(w)$ and $g(w)$ are powers of the same word.

1.3 Periodic instances

Following the terminology of [8, 9], we call a morphism $h : \Sigma^* \to \Delta^*$ *elementary* if it cannot be decomposed as $h = gf$,

$$\begin{tikzcd} \Sigma^* \arrow[rr, "h"] \arrow[dr, "f"'] & & \Delta^* \\ & \Gamma^* \arrow[ur, "g"'] & \end{tikzcd}$$

so that $|\Gamma| < |\Sigma|$.

A notion somewhat opposite is the *periodic* morphism where all values $h(w)$ are powers of a single word u, referred to as the *period* of h. For each letter a there is an integer i with the property $h(a) = u^i$. The period is unique if we consider the word with the minimal length.

It is clear that for $\Sigma = \{1,2\}$ every morphism on Σ^* is either elementary or periodic. The following lemma characterizes elementary morphisms further. We note that without loss of generality we can restrict ourselves to a binary target alphabet Δ since any alphabet can be encoded over a binary alphabet such that $E(g,h) = E(g',h')$ where g', h' are the compositions of g and h with the encoding. On the other hand, we clearly cannot similarly encode the source alphabet Σ. Therefore results obtained for a binary alphabet need not be valid in the general case.

Lemma 1.5 *A morphism $h : \{1,2\}^* \to \{1,2\}^*$ is elementary if and only if it is injective if and only if $h(12) \neq h(21)$.*

An instance (g,h) is said to be *periodic* if both g and h are periodic. The following lemma established in [32] shows that in this case the periods are the same.

Lemma 1.6 *Assume that g and h are periodic and $E(g,h)$ is not empty. Then g and h are periodic with the same period.*

Example 1.3 Let the instance (g,h) be defined by the table

	1	2
g	a^2	a
h	a	a^3

Every word that has 1's twice as much as 2's is a solution and words with any other ratio are not. Thus

$$E(g,h) = \{w \in \Sigma^+ \mid \psi(w) = k(2,1),\ k \geq 1\}.$$

□

In a binary alphabet the equality sets of periodic instances are characterized by the following lemma, due to [4].

Lemma 1.7 *Let g and h be distinct periodic morphisms over a binary alphabet with minimal periods p and q, respectively. Then $E(g,h)$ is of the form*

$$\{w \in \Sigma^+ \mid \psi(w) = k(i_1, i_2),\ k \geq 1\}, \tag{1.1}$$

where the components i_1 and i_2 are relatively prime if $p = q$, and $E(g,h)$ is empty if $p \neq q$. Every set of form (1) is an equality set for some periodic morphisms.

This shows that the equality set of a periodic instance consists of all words with the same basic Parikh vector and also that for any basic Parikh vector there exists a periodic instance.

Some words have an interesting property in regard to periodicity. Following the terminology of [3], we say that a word w is *periodicity forcing* if every instance (g, h) for which w is in $E(g, h)$ is periodic or trivial. (We consider only instances defined on the minimal alphabet of w, that is, $\text{alph}\,(w) = \Sigma$.) In the same way a set of words can be periodicity forcing. The second property was not originally included in the definition since in a binary alphabet (g, h) being trivial implies that $g = h$. But in larger alphabets, as shown later on in Section 2.4, this property is necessary.

We end this section by showing an interesting property of equality sets in a binary alphabet. The following lemma due to [5] shows that every solution must have the same ratio or the same basic Parikh vector. Hence any set containing two words with different ratios is in $E(g, h)$ only if $g = h$. In [5] these sets are called *rich*. By our definition, where the instances are allowed to be trivial, they can also be viewed periodicity forcing.

Lemma 1.8 *Let g, h be distinct morphisms over a binary alphabet. If $u, v \in E(g, h)$, then $r(u) = r(v)$.*

Chapter 2

Prime Words: Basic Types

By basic types of prime words we mean P-prime, S-prime and F-prime solutions of the Post Correspondence Problem introduced in [32]. More specifically, a solution is termed F-/S-/P-prime if no suffix/subword/scattered subword can be removed such that the remaining part is also a solution.

The notion of prime words introduced in [23] is an extension of the prime solutions: every word that is F-/S-/P-prime for some instance of the Post Correspondence Problem is an F-/S-/P-word. In this first paper many problems were left open, now we solve some of them.

In Section 2.2 we prove that the membership problems of the sets of P-, S- and F-words are all decidable by using Makanin's result, [22], for solving a finite system of equations in a free monoid. Since the algorithm in the proof is far from practical we examine also other characterization methods. The algorithm for F-words is based on the lemma the proof of which was found by Dr. Iiro Honkala, while for S- and P-words we have only partial algorithms. All these characterization results were first published in [18].

In Section 2.3 we study the hierarchy of the three types of prime words. The relation between S- and F-words in all alphabets as well as that of P- and S-words in a three-letter alphabet is due to [17], whereas the extension of the latter case to larger alphabets is due to [18]. In a binary alphabet it still remains open whether S- and P-words are equal or not.

The last section of this chapter is devoted to periodicity forcing words, the study of which is strongly related to the one of the prime solutions. Most of the results here are due to [18].

2.1 Primitive solutions

For a word w over Σ^*, we consider three sets of words obtained from w by removing a final subword, a subword or a scattered subword, respectively. More specifically, define

$$\begin{aligned} \mathit{fin}\,(w) &= \{v \mid w = vx, \text{ for some } x \in \Sigma^*\}, \\ \mathit{sub}\,(w) &= \{v_1v_2 \mid w = v_1xv_2, \text{ for some } v_1, v_2, x \in \Sigma^*\}, \\ \mathit{scatsub}\,(w) &= \{v_1 \dots v_k \mid w = x_1v_1 \dots x_kv_kx_{k+1}, \text{ for some } x_i, v_i \in \Sigma^*\}. \end{aligned}$$

We want to emphasize that $\mathit{sub}\,(w)$ is obtained by catenating the *remaining parts* when a subword of w is removed. The set $\mathit{scatsub}\,(w)$ can be viewed in an analogous way but also as the set of scattered subwords of w.

We can now determine three subsets of the equality set,

$$\begin{aligned} F(g,h) &= \{w \in E(g,h) \mid \mathit{fin}\,(w) \cap E(g,h) = \{w\}\}, \\ S(g,h) &= \{w \in E(g,h) \mid \mathit{sub}\,(w) \cap E(g,h) = \{w\}\}, \\ P(g,h) &= \{w \in E(g,h) \mid \mathit{scatsub}\,(w) \cap E(g,h) = \{w\}\}. \end{aligned}$$

Words in these three sets are called *F-prime*, *S-prime* and *P-prime solutions* for the instance PCP=(g,h). Thus, a solution w is termed F-prime if no nonempty final subword can be removed, S-prime if no nonempty subword can be removed, and P-prime if no scattered subword can be removed in such a way that the remainder is still a solution.

It follows immediately that every P-prime is an S-prime, and every S-prime is an F-prime, i.e.,

$$P(g,h) \subseteq S(g,h) \subseteq F(g,h) \subseteq E(g,h).$$

One or both of the first inclusions may be strict whereas the third inclusion is always strict, provided $E(g,h)$ is nonempty. If $E(g,h)$ is nonempty then so must be the three other sets. In addition, each of the four sets is recursive.

The triple (p,s,f), where p, s and f are the cardinalities of the sets $P(g,h)$, $S(g,h)$ and $F(g,h)$, respectively, is defined to be the *primality type* of the instance PCP=(g,h). Thus p, s and f are nonnegative integers or ∞. These triples were fully characterized in [32].

Theorem 2.1 *A triple (p,s,f) is a primality type if and only if either $p = s = f = 0$, or else each of the following conditions (i)–(iii) holds: (i) $1 \le p \le s \le f$, (ii) p is finite, (iii) if $s < f$ then $f = \infty$. An example for each possible type can be effectively constructed.*

Example 2.1 Let g and h be the same as in Example 1.2; hence, $E(g,h) = (1122)^+$. The only F-prime is 1122 and so, by Theorem 2.1, the primality type of (g,h) is $(1,1,1)$. □

The notion of prime solutions is extended to prime words, as follows. We say that a word w over Σ is a *P-word* if, for some instance (g,h), w is in $P(g,h)$. In the same way *S-words* and *F-words* correspond to S-prime and F-prime solutions. It follows directly from the definition that the set of P-words is included in the set of S-words which, in turn, is included in the set of F-words. We return to this hierarchy in Section 2.3.

The following lemma due to [23] gives a way of constructing words that are not F-words (and thus not S-words and P-words either).

Lemma 2.2 *If w_1 is a nontrivial prefix of w satisfying $\psi(w_1) = r\psi(w)$ for some (rational) number r, then w is not an F-word.*

The lemma gives, among others, the following examples: $ab^2a^2b^4$, $ab^6a^3b^2$. It also guarantees that P-, S- and F-words are primitive even in the original sense (defined on page 11).

2.1.1 Primitive solutions in periodic instances

Lemma 1.7 characterized the equality sets of periodic instances in a binary alphabet: if (g,h) is periodic then $E(g,h) = \{w \mid \psi(w) = k(i_1,i_2),\ k \geq 1\}$ for some relatively prime numbers i_1 and i_2. Clearly, the corresponding set of P-prime solutions is

$$P(g,h) = \{w \in \Sigma^+ \mid \psi(w) = (i_1,i_2)\}.$$

The next lemma, originally presented in [32], helps to characterize the S-prime solutions.

Lemma 2.3 *Let $\Sigma = \{1,2\}$ be an alphabet and $M, N \geq 0$ be integers. If the Parikh vector $\psi(w)$ of $w \in \Sigma^*$ is (CM, CN), for some integer $C \geq 1$, then there exists a subword w' of w such that $\psi(w') = (M,N)$.*

Notice that this result does not hold for alphabets with more than two letters.

Example 2.2 Let $\Sigma = \{1, 2, 3\}$ and $w = 11112222223333333333$. Then w is of length $2(2+3+5)$ and its Parikh vector $\psi(w)$ is $(2 \cdot 2, 2 \cdot 3, 2 \cdot 5)$ thus satisfying the conditions of Lemma 2.3. There exists, however, no subword w' such that $\psi(w') = (2, 3, 5)$. □

Lemma 2.4 *If the instance (g, h) is periodic in a binary alphabet, then*

$$P(g, h) = S(g, h).$$

Proof: Lemma 1.7 shows that $w \in E(g, h)$ if and only if $\psi(w) = s(k_1, k_2)$ for some $\gcd(k_1, k_2) = 1$ and $s \geq 1$. On the other hand, by Lemma 2.3 each solution w has a subword w' such that $\psi(w') = (k_1, k_2)$. But this means that if $s > 1$ then w cannot be an S-prime solution. Consequently, $S(g, h) = \{w \in E(g, h) \mid \psi(w) = (k_1, k_2)\} = P(g, h)$. □

In larger alphabets this is not the case any more.

Example 2.3 Define

	1	2	3
g	a^2	a^3	a
h	a	a	a^3

Then $P(g, h) = \{113, 131, 311, 23, 32\}$ but $12^n 13^{n+1}$, $n \geq 0$, are all S-prime solutions. □

We end this section with a result of 'almost' periodicity forcing words. It was originally suggested in [23] and later corrected in [17] by adding the condition concerning v. In a binary alphabet it shows that the words of this form are periodicity forcing, while to obtain this in bigger alphabets v must be further restricted, as seen in Section 2.4.

Lemma 2.5 *Consider $w = avava$ where a is a letter and v is a word over Σ, $|\Sigma| \geq 2$, and, in addition, v begins or ends with a. If $w \in E(g, h)$ for some nontrivial instance (g, h) then $g(a)$, $g(v)$, $h(a)$ and $h(v)$ are powers of the same word.*

Proof: Since the instance (g, h) is nontrivial, $g(a) \neq h(a)$ for all $a \in \Sigma$. Without loss of generality, we denote

$$g(av) = \alpha, \; h(av) = \beta = \alpha x, \quad |x| \geq 1,$$
$$g(a) = \alpha_1, \; h(a) = \beta_1, \quad \alpha = \alpha_1\alpha_2, \; \beta = \beta_1\beta_2.$$

It follows that $|x| < |\beta_2|$ and, moreover, $\alpha_1 = \beta_1 y$ where $|y| \geq 1$.

Writing the equation $g(w) = h(w)$ in the new notation

$$\beta_1 y\alpha_2\beta_1 y\alpha_2\beta_1 y = \beta_1\beta_2\beta_1\beta_2\beta_1 = \beta_1 y\alpha_2 x\beta_1 y\alpha_2 x\beta_1,$$

we obtain

$$\beta_1 y\alpha_2\beta_1 y = x\beta_1 y\alpha_2 x\beta_1.$$

Thus $|y| = 2|x|$ and we may write $y = y_1y_2$ where $|y_1| = |y_2| = |x|$. Our preceding equation reads now

$$\beta_1 y_1 y_2\alpha_2\beta_1 y_1 y_2 = x\beta_1 y_1 y_2\alpha_2 x\beta_1.$$

Considering subwords of the same length, we obtain

$$\beta_1 y_1 = x\beta_1, \quad y_1 = y_2, \quad \alpha_2\beta_1 y_1 y_1 = y_1\alpha_2 x\beta_1,$$

where we have already used the second equation in the third one. By Lemma 1.1 there exist words u and q such that $x = uq$, $y_1 = qu$ and $\beta_1 = (uq)^k u$. Consequently, from the last equation above $\alpha_2 uq = qu\alpha_2$, and so $qu = rs$, $uq = sr$, $\alpha_2 = (rs)^l r$. Writing, without loss of generality, $q = rp$, we have $s = pu$ and $urp = pur$, hence by Lemma 1.1

$$p = z^i, \quad ur = z^j, \tag{2.1}$$

for some z. Furthermore, $uq = urp = z^{i+j}$.

Combining our knowledge, we have

$$\begin{aligned}
\alpha_1 &= (uq)^{k+2}u = (z^{i+j})^{k+2}u, \\
\alpha_2 &= r(uq)^l = r(z^{i+j})^l, \\
\beta_1 &= (uq)^k u = (z^{i+j})^k u, \\
\beta_2 &= r(uq)^{l+3} = r(z^{i+j})^{l+3}.
\end{aligned} \tag{2.2}$$

To complete the proof we shall use the fact that v begins with a or ends with it. This is exhibited in two parts.

Let $v = ca$. By the last two equations in (2),

$$r(z^{i+j})^{l+3} = \beta_2 = h(v) = h(c)h(a) = h(c)\beta_1 = h(c)(z^{i+j})^k u. \qquad (2.3)$$

We first suppose that $|u| > |z|$. Thus $u = zd$ and $u = ez$ for some d, e by the equations (1) and (3), and there exist n, o such that $e = no$, $d = on$, $z = (no)^t n$ for some t. Equation (1) now gives us

$$\begin{aligned} (no)^t nonr &= [(no)^t n]^j \\ onr &= [(no)^t n]^{j-1}. \end{aligned}$$

Reading the words from the left implies that $on = no$; hence, there exists an f such that $o = f^{i_1}$ and $n = f^{i_2}$, as intended.

If, however, $t = 0$ then

$$onr = n^{j-1}$$

and, by Lemma 1.4, there is a word f such that $o = f^{i_1}$, $r = f^{i_2}$ and $n = f^{i_3}$, as intended. So the claim follows if $|u| > |z|$. (Obviously, it follows also in the case $|u| = |z|$.)

Next we suppose that $|u| < |z|$. In the same way as before, $z = ud$ and $z = eu$ for some d, e by the equations (1) and (3), and, further, $e = no$, $d = on$, $u = (no)^t n$. Equation (3) now gives us

$$r(z^{i+j})^{l+2}[(no)^t non]^{i+j} = h(c)(z^{i+j})^{k-1}[(no)^t non]^{i+j}(no)^t n,$$

which, by reading the words from the right, implies that $no = on$.

If, however, $k = l = 0$ then the equation (1) yields that $r = on[(no)^{t+1}n]^{j-1}$ and

$$on[(no)^{t+1}n]^{j-1} = r = \alpha_2 = g(c)\alpha_1 = g(c)z^{2i+2j}u = g(c)[(no)^{t+1}n]^{2i+2j}(no)^t n.$$

Considering the suffixes we obtain $no = on$, as intended.

The second case $v = ac$ goes in the same way. We just interchange the roles of u and r. □

The original lemma in [23] tried to establish the same as Lemma 2.5 in a binary alphabet but without an additional restriction for the word v. Its proof obtained equations (2) which is not sufficient for the claim, as seen in the following example.

Example 2.4 Let g and h be defined by the table

	a	b
g	$(xy)^4x$	yxy
h	$(xy)^2x$	$y(xy)^4$

Then the word $w = ababa$ is a solution even though g and h are not periodic. The choice of (g, h) is easily constructed from equations (2) by choosing $u = x$, $r = y$, $z = xy$, $i = 0$, $j = l = 1$ and $k = 2$. □

Nevertheless, the following example shows that Lemma 2.5 does not characterize v completely, i.e., v may start and end with b and w is still periodicity forcing. In fact, it seems that the generalizations of the previous example, that is, the words of the form $(ab)^i a$, $i \geq 2$, are the only counter-examples to the original statement, Lemma 7 in [23].

Example 2.5 Let $w = avava$ where $v = baabab$. We have

$$\begin{aligned} g(v) &= g(b)g(a)g(a)g(b)g(a)g(b), \\ h(v) &= h(b)h(a)h(a)h(b)h(a)h(b), \end{aligned}$$

which together with equations (2) imply

$$\begin{aligned} r(z^{i+j})^l &= g(b)(z^{i+j})^{k+2}u(z^{i+j})^{k+2}ug(b)(z^{i+j})^{k+2}ug(b), \\ r(z^{i+j})^{l+3} &= h(b)(z^{i+j})^k u(z^{i+j})^k uh(b)(z^{i+j})^k uh(b). \end{aligned}$$

Combining these two equations we obtain

$$\begin{aligned} &g(b)(z^{i+j})^{k+2}u(z^{i+j})^{k+2}ug(b)(z^{i+j})^{k+2}ug(b)(z^{i+j})^3 \\ &\quad = h(b)(z^{i+j})^k u(z^{i+j})^k uh(b)(z^{i+j})^k uh(b). \end{aligned} \tag{2.4}$$

Comparing the lengths of both sides we infer that $|g(b)(z^{i+j})^3| = |h(b)|$. Thus reading the suffixes in equation (4) gives $h(b) = g(b)(z^{i+j})^3$ and

$$u(z^{i+j})^{k+2}ug(b)(z^{i+j})^2 = z^{i+j}u(z^{i+j})^k uh(b),$$

from which $uz^{i+j} = z^{i+j}u$. In addition,

$$z^{i+j}g(b)(z^{i+j})^2 = h(b) = g(b)(z^{i+j})^3.$$

Hence $z^{i+j}g(b) = g(b)z^{i+j}$ and g and h are periodic. □

When constructing Lemma 7 in [23], the goal was to find an S-word that is not a P-word. The closer details as well as the reason why this is not possible will be discussed in this context in Section 2.3.

2.2 Characterization

We start this section by showing that given $w \in \Sigma^+$ it is decidable whether w is a P-prime, an S-prime or an F-prime solution for some instance (g, h). These problems are reduced to the problem of the solvability of a finite system of equations shown decidable by Makanin in [22]. Essential is the possibility to translate an inequality into several systems of equations as in [3] and [4].

Theorem 2.6 *Each of the properties of being a P-word, an S-word or an F-word is decidable.*

Proof: Let $w \in \Sigma^+$. To check whether w is a P-word we examine its scattered subwords, the set $scatsub\,(w)$. This set is easy to find and it is always finite. Using the result of Makanin [22] we can test for each of the following systems whether it has a solution, i.e., whether there are g and h such that

$$\begin{aligned} &g(w) = h(w) \\ &g(v) \neq h(v) \quad \text{for all } v \in scatsub\,(w) - \{w, \lambda\}. \end{aligned}$$

If so, then w is a P-prime solution for this instance (g, h), otherwise w is not a P-word.

To check whether w is an S-word or an F-word we use similarly the sets $sub\,(w)$ and $fin\,(w)$, respectively. □

Example 2.6 Consider the word $abab$. Now

$$scatsub\,(abab) = \{a, aa, aab, ab, aba, abb, b, ba, bab, bb\}$$

but for Theorem 2.6 it suffices to study only the words ab and ba because of Lemma 1.8. The equations

$$\begin{aligned} g(abab) &= h(abab) \\ g(ab) &\neq h(ab) \\ g(ba) &\neq h(ba) \end{aligned} \tag{2.5}$$

are converted into Makanin's form as follows. First we assume, in view of the discussion in Section 1.3, that $\Delta = \{1, 2\}$. The words $g(a)$, $g(b)$, $h(a)$,

$h(b)$ are denoted by x_1, x_2, y_1, y_2, respectively. Other unknowns are the words s, t, u, v, y and z. For each $\alpha, \beta, \gamma, \delta \in \{1, 2\}$, where $\alpha \neq \beta$ and $\gamma \neq \delta$, we construct the following three systems of equations over Δ^*:

$$x_1x_2x_1x_2 = y_1y_2y_1y_2,$$
$$x_1x_2 = s\alpha t, \quad y_1y_2 = s\beta u,$$
$$x_2x_1 = v\gamma y, \quad y_2y_1 = v\delta z.$$

But now the first two equations imply that $s\alpha ts\alpha t = s\beta us\beta u$ from which $\alpha = \beta$, a contradiction. Thus the system of equations (5) has no solution and $abab$ is not a P-word. □

Theorem 2.6 tells actually very little of the nature of the three types of prime words. Also Makanin's general algorithm is very complicated. This is why we are interested in other possibilities to characterize these words.

We start with the set of F-words and a new notion. A proper prefix w' of the word w satisfying $\psi(w') = r\psi(w)$, for some rational number r, is called an *F-prefix*. Thus w' is an F-prefix if it has the same basic Parikh vector as w. We also say, following the terminology of [3], that a word is *ratioprimitive* if it has no F-prefixes.

We prove that ratioprimitive words are just another form of F-words. The following lemma shows how to construct an instance such that no word which is Parikh shorter than a given word and has different basic Parikh vector is a solution. Here a word u is *Parikh shorter* than v if $\psi(u) \leq \psi(v)$ componentwise.

Lemma 2.7 *For every vector $K(k_1, \ldots, k_n)$, where $K, k_i \in \mathbb{N} - \{0\}$ and $\gcd(k_1, \ldots, k_n) = 1$, there exist numbers $l, m_2, \ldots, m_n \in \mathbb{N} - \{0\}$ such that $k_1 l = k_2m_2 + \ldots + k_nm_n$ and if*

$$s_1 l = s_2m_2 + \ldots + s_nm_n$$

for some $s_i \in \mathbb{N}$, $s_1 \neq 0$, $s_i \leq Kk_i$, then

$$s_i = rk_i \quad \text{for some } r \leq K.$$

Proof: Let $M = k_1(K \max(k_1k_i) + 1)$ and choose $m_2 = M, \ldots, m_n = M^{n-1}$. (Throughout this proof $\max(k_1k_i)$ is an abbreviation of the maximal number of the set $\{k_1k_i \mid 1 \leq i \leq n\}$.) The number l can now be defined from the equation $k_1l = k_2M + \ldots + k_{n-1}M^{n-2} + k_nM^{n-1}$.

Assume that

$$s_1l = s_2M + \ldots + s_{n-1}M^{n-2} + s_nM^{n-1}, \quad s_i \leq Kk_i. \tag{2.6}$$

Multiplying the previous equations with s_1 and k_1, we obtain

$$\begin{cases} s_1k_1l = s_1k_2M + \ldots + s_1k_{n-1}M^{n-2} + s_1k_nM^{n-1} \\ k_1s_1l = k_1s_2M + \ldots + k_1s_{n-1}M^{n-2} + k_1s_nM^{n-1} \end{cases}$$

and, further,

$$0 = \underbrace{(k_1s_2 - s_1k_2)M + \ldots + (k_1s_{n-1} - s_1k_{n-1})M^{n-2}}_{\alpha} + \underbrace{(k_1s_n - s_1k_n)}_{\beta}M^{n-1}.$$

(The abbreviations α and β will be used in the sequel.)

We prove that this is possible only if all the coefficients are zero. Since by (6)

$$0 \leq k_1s_i \leq k_1Kk_i \leq K\max(k_1k_i),$$
$$0 \leq s_1k_i \leq k_1Kk_i \leq K\max(k_1k_i),$$

also

$$|k_1s_i - s_1k_i| \leq K\max(k_1k_i).$$

Thus

$$\begin{aligned} |\alpha| &\leq K\max(k_1k_i)(M + \ldots + M^{n-2}) \\ &\leq K\max(k_1k_i)(\frac{M^{n-1}-1}{M-1} - 1) < \frac{K\max(k_1k_i)}{M-1}M^{n-1}. \end{aligned}$$

Comparing this coefficient with β, we first notice that

$$\frac{K\max(k_1k_i)}{M-1} = \frac{K\max(k_1k_i)}{k_1(K\max(k_1k_i)+1)-1} \leq \frac{K\max(k_1k_i)}{k_1K\max(k_1k_i)} = \frac{1}{k_1} \leq 1.$$

So $|\alpha| < M^{n-1}$ which is true only if $|\beta| < 1$. Thus we must have $\alpha = 0 = \beta$. Proceeding in the same way with the equation

$$0 = \underbrace{(k_1s_2 - s_1k_2)M + \ldots + (k_1s_{n-2} - s_1k_{n-2})M^{n-3}} + \underbrace{(k_1s_{n-1} - s_1k_{n-1})}M^{n-2},$$

we finally obtain

$$k_1 s_i - s_1 k_i = 0 \quad \text{or} \quad k_1 s_i = s_1 k_i \quad \text{for all } i.$$

If k_1 does not divide s_1 then $k_1 = pq$ for some p, q where $q \neq 1$ and $q \mid k_i$ for all i; that is, $q = \gcd(k_1, \ldots, k_n)$, a contradiction. Therefore $s_1 = rk_1$ for some r and, further,

$$s_i = rk_i \quad \text{for all } i.$$

□

We are now ready to state our main result.

Theorem 2.8 *Let $w \in \Sigma^+$. Then w is an F-word if and only if w is ratioprimitive.*

Proof: By Lemma 2.2 we know that F-words are all ratioprimitive. Conversely, let w be ratioprimitive and $\psi(w) = K(k_1, \ldots, k_n)$. Define the instance (g, h) by the table

	1	2	...	n
g	a^{l+1}	a	...	a
h	a	a^{m_2+1}	...	a^{m_n+1}

According to Lemma 2.7 the nonnegative integers $l, m_2, \ldots, m_n$ can now be defined in such a way that only the prefixes w_1 of w for which $\psi(w_1) = r(k_1, \ldots, k_n)$, $r < K$, are solutions. But because w is ratioprimitive, no such prefixes exist; hence, $w \in F(g, h)$.

□

Example 2.7 Let $w = 112233$, hence, $\psi(w) = 2(1,1,1)$. We apply Lemma 2.7 to construct an instance (g, h) for which w is an F-prime solution. Now

$$M = 1 \cdot (2 \cdot 1 + 1) = 3 = m_2, \quad m_3 = 9, \quad l = 3 + 9 = 12.$$

We define

	1	2	3
g	a^{13}	a	a
h	a	a^4	a^{10}

It is easy to calculate that none of the prefixes $1, 11, 112, 1122, 11223$ are solutions; so w is an F-prime solution.

In fact, Lemma 2.7 shows that any word whose Parikh vector is $2(1,1,1)$ and whose prefix is not a permutation of 123, such as $121233, 122133$, is an F-prime solution for this instance (g,h). □

Theorem 2.8 gives an effective algorithm for F-words: it is easy to check for a given word whether it has F-prefixes or not, even in polynomial time.

For S-words we have found only a partial algorithm. We say that a word w is *subratioprimitive* if none of its proper subwords x satisfies the property $\psi(x) = r\psi(w)$, for any rational number r.

Theorem 2.9 *Every subratioprimitive word is an S-word.*

Proof: Let w be subratioprimitive and $\psi(w) = K(k_1, \ldots, k_n)$. Then w is an S-prime solution for the instance (g,h) defined as in the proof of Theorem 2.8. □

The converse of the previous theorem does not hold.

Example 2.8 Let the instance (g,h) be defined by the table

	1	2	3
g	a^6b	a	a^2
h	a^2	ba^6	a

Then, among others, the words

$$311\underbrace{132}223, \ 311\underbrace{312}212332, \ 13\underbrace{312}1123223$$

are S-prime solutions but not subratioprimitive. □

Also for P-words the algorithm is partial. The following two theorems due to [23] show what kind of words can appear as P-words.

Theorem 2.10 *The word*

$$w = a_1^{i_1} a_2^{i_2} \ldots a_n^{i_n}$$

is a P-word if $\Sigma = \{a_1, \ldots, a_n\}$ *and* $n \geq 2$.

Theorem 2.11 *Every word w over the alphabet $\{a_1, a_2\}$ such that the two components i_1 and i_2 in $\psi(w)$ are relatively prime, is a P-word. More generally, let w be over $\Sigma = \{a_1, \ldots, a_n\}$ and assume that Σ can be divided into two disjoint nonempty parts Σ_1 and Σ_2 such that the two numbers, obtained by summing up the components of $\psi(w)$ corresponding to the letters of Σ_1, as well as those corresponding to the letters of Σ_2, are relatively prime. Then w is a P-word.*

Theorem 2.11 can actually be improved. We notice that if w satisfies the conditions of this theorem then $\psi(w) = (k_1, \ldots, k_n)$ and $\gcd(k_1, \ldots, k_n) = 1$. For the other direction this is not true as seen from the vector $(3, 5, 22)$. Still we have the following result.

Theorem 2.12 *Every word $w \in \Sigma^*$ for which $\psi(w) = (k_1, \ldots, k_n)$ and $\gcd(k_1, \ldots, k_n) = 1$, $n \geq 2$, is a P-word.*

Proof: Let the instance (g, h) be defined as in the proof of Theorem 2.8. Then Lemma 2.7 guarantees that no scattered subword of w is a solution, so w is a P-prime solution. $\square$

Our next example shows that these results are not exhaustive.

Example 2.9 Let us consider the words w for which $\psi(w) = 2(1, 1, 1)$. Clearly, they do not fulfill the assumptions of Theorem 2.12; still, some of them are P-words. If we disregard the words which are not ratioprimitive or obtained by mirror images or permutations we have only the following possibilities

112233 121233 122313 122133 122331 121323

each of which is a P-prime solution for the following nonperiodic instances presented in the same order as the words.

1.	1	2	3
g	a^4b	a	a^2
h	a^2	ba^4	a

2.	1	2	3
g	a^2	a^4b	a^3
h	a	a^2	ba^6

3.	1	2	3
g	a	a^8b	a
h	a^2	a^2	a^3ba^3

4.	1	2	3
g	a^2	a^4b	a
h	a	a^2	ba^4

5.	1	2	3
g	a^4	a^4b	a
h	a^2	a^3	ba^4

6.	1	2	3
g	a^4b	a^4	a
h	a	a^2	ba^6

The fifth instance has also other P-prime solutions that do not satisfy the conditions of Theorem 2.12, such as 223311, 112223331 and 121232313. □

In a binary alphabet, however, we conjecture that Theorems 2.10 and 2.11 characterize the sets of P-words and S-words completely.

Conjecture 2.13 *In a binary alphabet w is a P-word (an S-word) if and only if*

$$\gcd(|w|_1, |w|_2) = 1 \quad \text{or} \quad w \in 1^+2^+ \cup 2^+1^+.$$

By Theorems 2.10 and 2.11, the "if"-part holds true. For the "only if"-part we will present some evidence. All the proofs are based on the following two lemmas due to [3] that characterize all the instances agreeing on a given word over $\{a, b\}$.

Lemma 2.14 *The word aab is in $E(g,h)$, with $|g(a)| > |h(a)|$, if and only if there exist an integer $t \geq 0$ and words α, β and γ such that $\alpha\beta \neq \lambda$ and*

	a	b
g	$(\alpha\beta)^t\alpha\alpha\beta$	γ
h	$(\alpha\beta)^t\alpha$	$\beta\alpha\alpha\beta\gamma$

Lemma 2.15 *The word $aabb$ is in $E(g,h)$, with $|g(a)| > |h(a)|$, if and only if there exist nonnegative integers t_1, t_2 and t_3 and words α and β such that $\alpha\beta \neq \lambda$ and*

	1	2
g	$\alpha(\beta\alpha)^{t_1}\alpha(\beta\alpha)^{t_2}\beta$	$\alpha(\beta\alpha)^{t_3}$
h	$(\alpha\beta)^{t_1}\alpha$	$\beta(\alpha\beta)^{t_2}\alpha(\alpha\beta)^{t_3}\alpha$

Lemma 2.16 *The word w is not an S-word if $\psi(w) = 3(1,1)$ and $w \notin 1^+2^+ \cup 2^+1^+$.*

Proof: The words whose Parikh vector is $3(1,1)$ are

111222	121122	122121	211221	221112
112122	121212	122211	212112	221121
112212	121221	211122	212121	221211
112221	122112	211212	212211	222111

After eliminating the words which are not even F-words we still have

$$111222, \quad 112122, \quad 221211, \quad 222111$$

the last two of which are mirror images of the first two ones. Hence it is enough to prove that the word 112122 is not an S-word.

By denoting $a = 1$ and $b = 2122$ Lemma 2.14 becomes applicable,

$$\begin{aligned}
g(1) &= g(a) = (\alpha\beta)^t\alpha\alpha\beta, \\
g(2122) &= g(b) = \gamma, \\
h(1) &= h(a) = (\alpha\beta)^t\alpha, \\
h(2122) &= h(b) = \beta\alpha\alpha\beta\gamma.
\end{aligned}$$

The first two equations imply that $\gamma = g(2)(\alpha\beta)^t\alpha\alpha\beta g(2)g(2)$ which together with the last two equations yields

$$h(2)(\alpha\beta)^t\alpha h(2)h(2) = \beta\alpha\alpha\beta g(2)(\alpha\beta)^t\alpha\alpha\beta g(2)g(2).$$

Considering the lengths of each side we obtain $|h(2)| = |g(2)\alpha\beta|$.

So $h(2) = \beta\alpha g(2)$ and

$$g(2)(\alpha\beta)^t\alpha\beta\alpha g(2)\beta\alpha = \alpha\beta g(2)(\alpha\beta)^t\alpha\alpha\beta g(2).$$

Reading the suffixes, we have $g(2)\beta\alpha = \alpha\beta g(2)$ which applied to the prefixes of the preceding equation implies that $\alpha\beta = \beta\alpha$ and

$$\rho(\alpha) = \rho(\beta) = \rho(g(2)) = \rho(h(2)).$$

Hence the instance (g, h) is periodic, and a subword 12 is also a solution. □

Lemma 2.17 *The word w is not an S-word if $\psi(w) = 4(1,1)$ and $w \notin 1^+2^+ \cup 2^+1^+$.*

Proof: In the same way as in the previous lemma it suffices to prove that the words

$$11121222, \; 11122122, \; 11211222, \; 11212122$$

are periodicity forcing.

For the word 11211222 we apply Lemma 2.15 by denoting $a = 112$ and $b = 2$. Thus

$$\begin{aligned}
g(112) &= g(a) = \alpha(\beta\alpha)^{t_1}\alpha(\beta\alpha)^{t_2}\beta,\\
g(112) &= g(1)g(1)g(b) = g(1)g(1)\alpha(\beta\alpha)^{t_3},\\
h(112) &= h(a) = (\alpha\beta)^{t_1}\alpha,\\
h(112) &= h(1)h(1)h(b) = h(1)h(1)\beta(\alpha\beta)^{t_2}\alpha(\alpha\beta)^{t_3}\alpha,
\end{aligned}$$

from which

$$\rho(\alpha) = \rho(\beta) = \rho(g(1)) = \rho(h(1)).$$

For the remaining words we apply similarly Lemma 2.14. □

Also Lemma 2.5 presents many words giving evidence to the conjecture.

We end this section by presenting a lemma from [23], a useful tool for constructing words that are not P-words or S-words. The proof is also presented because the original result was only for P-words. Clearly it holds also for F-words but this is omitted since the characterization is already settled.

Lemma 2.18 *If w is not a P-word (resp. an S-word) and ϕ is a nonerasing morphism, then $\phi(w)$ is not a P-word (resp. an S-word).*

Proof: Assume the contrary: for some instance (g, h), $\phi(w)$ is in $P(g, h)$ (resp. $S(g, h)$). Consequently, w is in $E(g\phi, h\phi)$. Since w is not a P-word (resp. an S-word), one of its scattered subwords (resp. subwords) y is in $E(g\phi, h\phi)$. This implies that $\phi(y)$, which is not empty since ϕ is nonerasing, is in $E(g, h)$. Thus, $\phi(w)$ is not in $P(g, h)$ (resp. $S(g, h)$), a contradiction. □

This presents further evidence for our conjecture since

$$\gcd(|w|_1, |w|_2) > 1 \;\Rightarrow\; \gcd(|\phi(w)|_1, |\phi(w)|_2) > 1$$

for any nonerasing morphism ϕ.

2.3 Hierarchy

In this section we study the hierarchy

$$P \subseteq S \subseteq F$$

where P, S and F denote the sets of P-words, S-words and F-words, respectively. These two inclusions turn out to be strict in every alphabet except the binary case which remains open for P-words and S-words.

We start with the relation between F-words and S-words. In [23] the binary case was already settled with the words a^2bvab^2 where v is a word over $\{a, b\}$ containing equally many a's and b's in such a way that no prefix of v contains more b's than a's. Nevertheless, we use a different type of word since its argument can be extended to larger alphabets.

Theorem 2.19 *The word* 111212 *is an F-word but not an S-word.*

Proof: By Theorem 2.8, 111212 is an F-word.

In order to prove that it is not an S-word we use Lemma 2.15. Arguing indirectly, we assume that 111212 is an S-word and, hence, in the equality set of some instance (g, h). By denoting $a = 1$ and $b = 12$, Lemma 2.15 becomes applicable. We have

$$\begin{aligned} g(1) = g(a) &= \alpha(\beta\alpha)^{t_1}\alpha(\beta\alpha)^{t_2}\beta, \\ g(12) = g(b) &= \alpha(\beta\alpha)^{t_3}, \\ h(1) = h(a) &= \alpha(\beta\alpha)^{t_1}, \\ h(12) = h(b) &= \beta(\alpha\beta)^{t_2}\alpha(\alpha\beta)^{t_3}\alpha. \end{aligned}$$

The first two equations imply

$$\alpha(\beta\alpha)^{t_3} = \alpha(\beta\alpha)^{t_1}\alpha(\beta\alpha)^{t_2}\beta g(2)$$

and the last two

$$\beta(\alpha\beta)^{t_2}\alpha(\alpha\beta)^{t_3}\alpha = \alpha(\beta\alpha)^{t_1}h(2).$$

Therefore $\alpha\beta = \beta\alpha$, independently of the exponents t_1 and t_2. But this means that $112 \in sub\,(111212)$ is also a solution, a contradiction. $\square$

The above argument can be applied similarly to alphabets with more than two letters to show that the word

$$(123\ldots(k-1))^2(123\ldots k)^2 \in \{1, 2, 3, \ldots, k\}^*$$

is an F-word but not an S-word.

We have the following result.

Theorem 2.20 *Over every alphabet Σ, $|\Sigma| \geq 2$, there exists an F-word that is not an S-word.*

Consider now the relation between S-words and P-words. We claim that in a binary alphabet they are equal and, what is more, the set of S-prime solutions is exactly the set of P-prime solutions in every instance (g, h). In most of the cases, depending on the type of the instance (g, h), this has been verified as seen in the sequel.

The next two lemmas, due to [14] and [7], characterize the equality sets when g, h or both are injective morphisms. For the second lemma we need a new notion. Let (β, γ, δ) be a triple of words such that γ is primitive and is neither a suffix of β nor a prefix of δ. Moreover, $\gamma \neq \delta\beta$. We call such a triple *reduced* and define the language $L(\beta, \gamma, \delta)$ by setting

$$L(\beta, \gamma, \delta) = \beta\gamma^*\delta.$$

Lemma 2.21 *Let g and h be morphisms over a binary alphabet, g elementary and h periodic. Then there exists effectively a word w so that $E(g, h) = \{w\}^+$.*

Lemma 2.22 *Let (g, h) be a pair of injective morphisms over a binary alphabet. The equality set $E(g, h)$ is either of the form*

1. *$\{u, v\}^+$ for some (possibly empty) words u and v, or of the form*
2. *$(L(u, w, v))^+$ for some reduced triple (u, w, v).*

Observe that in [7] the equality set $E(g, h)$ includes also the empty word λ contrary to our definition. Lemma 2.22 has been stated according to our definition. Also the condition $\gamma \neq \delta\beta$ was not originally included in the definition of reduced triples in [7]. That this condition is really needed is seen from the following example.

Example 2.10 Let the pair (g, h) be defined by

	1	2
g	a	a^2ba^2b
h	ba^2	a^2

Then $E(g, h) = \{211\}^+ = (2(112)^*11)^+ = (21(121)^*1)^+$ is a reduced language in the old sense. $\square$

With the new condition we have no examples of reduced languages. In fact, we believe – as was also the case in the original article [7] – that no such equality sets exist.

Theorem 2.23 *Let g and h be morphisms over a binary alphabet such that at least one of them is periodic. Then $P(g,h) = S(g,h)$.*

Proof: If g is injective and h periodic (or vice versa) then $E(g,h) = \{w\}^+$ by Lemma 2.21 and thus

$$P(g,h) = \{w\} = S(g,h).$$

On the other hand, if g and h are both periodic the claim follows from Lemma 2.4. □

Assume now that g and h are both injective. Then Lemma 2.22 implies that the equality set is either of the form $\{u,v\}^+$, $\{w\}^+$ or $(\beta\gamma^*\delta)^+$. In the last two cases the sets of S-prime and P-prime solutions are equal ($\{w\}$ or $\{\beta\delta\}$). And if the equality set is generated by two words u, v then we know, by Theorem 2.1, that they are both S-prime solutions; still, it is conceivable that $u \in scatsub(v)$. This is exactly the possibility that was omitted in [16] where the proof was inaccurate.

Nevertheless, the only known example of this situation is the following one originally presented in [3].

Example 2.11 Define, for a fixed $i \geq 1$, morphisms g and h by

	1	2
g	a^iba^i	$(ba^{2i})^{i-1}b$
h	a	$(ba^{2i})^{i-1}ba^i(ba^{2i})^{i-1}b$

It is straightforward to see that

$$E(g,h) = \{1^i2, 21^i\}^+.$$

Thus the primitive solutions are

$$F(g,h) = S(g,h) = P(g,h) = \{1^i2, 21^i\}.$$

□

So we strongly believe that the following conjecture holds.

Conjecture 2.24 *Let g and h be morphisms over a binary alphabet. Then*

$$P(g,h) = S(g,h).$$

There have also been attempts to find a counter-example, an S-word which is not a P-word. As mentioned earlier, Lemma 2.5 was originally constructed for this purpose. Considering words of the form $w = avava$, we first choose some ratio between a's and b's, in which the number of a's is odd and the number of b's even, such as 7:4. Then we multiply these numbers by some odd n, and try to combine letters in v in such a way that no subword of w has the original ratio, for a scattered subword always has. In our example, we may choose $n = 3$ and consider the word w which has 21 a's and 12 b's; hence, we try to arrange 9 a's and 6 b's in v so that no subword of w has the ratio 7:4. But, as Lemma 2.3 shows, this is not possible for any choice of ratio and n.

The binary alphabet turns out to be an exception: in larger alphabets P-words are strictly included in the set of S-words.

Theorem 2.25 *The word $w = a^2b^3c^3a^4b^3c^3a^3$ is an S-word but not a P-word.*

Proof: Since w is subratioprimitive, it is by Theorem 2.9 an S-word.

To prove that it is not a P-word we use Lemma 2.5. Assume now that w is a solution for some instance (g,h). Lemma 2.5 becomes applicable by denoting $v = ab^3c^3a^2$. Thus there exists a word $u \in \Delta^+$ such that

$$g(a) = u^j, \quad g(v) = u^i, \quad h(a) = u^{j'}, \quad h(v) = u^{i'}$$

for some integers i, j, i', j'. Hence

$$\begin{aligned} u^i &= g(v) = g(a)g(b)^3g(c)^3g(a)^2 = u^jg(b)^3g(c)^3u^{2j} \\ u^k &= g(b)^3g(c)^3, \quad k = i - 3j. \end{aligned}$$

We denote briefly $u^k = o^3p^3$.

In the same way,

$$u^{k'} = h(b)^3h(c)^3 =: r^3s^3.$$

This leads to four subcases: $k = k' = 1$, $k' > 1 = k$, $k > 1 = k'$ or $k, k' > 1$.

In the first case, $k = k' = 1$, the instance (g,h) is defined by

	a	b	c
g	u^j	o	p
h	u^t	r	s

The equations $o^3p^3 = u = r^3s^3$ and $(o^3p^3)^{9j+2} = g(w) = h(w) = (r^3s^3)^{9t+2}$ now imply that $j = t$ and $g(a) = h(a)$. Thus the instance (g, h) is trivial and $w \notin P(g, h)$.

Assume now that $k = 1$ and $k' > 1$. By Lemma 1.3 the words u, $h(b)$ and $h(c)$ are powers of the same word. If $u = q^l = o^3p^3$ then either q, o and p are powers of the same word or $q = o^3p^3$. In the first case the instance (g, h) is periodic, whereas the latter case yields to the instance (g, h) defined, for some $j, i_1, i_2, i_3 \in \mathbb{N}$, by

	a	b	c
g	$(o^3p^3)^j$	o	p
h	$(o^3p^3)^{i_1}$	$(o^3p^3)^{i_2}$	$(o^3p^3)^{i_3}$

Since $g(w) = (o^3p^3)^{9j+2}$ and $h(w) = (o^3p^3)^{9i_1+6i_2+6i_3}$ we must have

$$9j + 2 = 9i_1 + 6i_2 + 6i_3.$$

But this is not possible for any integers since $2 \not\equiv 0 \pmod 3$.

The third case goes in the same way and in the last case the instance (g, h) is periodic. Thus the corresponding instance (g, h) is always periodic or trivial and therefore w cannot be a P-word since $\psi(w) = 3(3, 2, 2)$. $\square$

In bigger alphabets we consider the morphic images of the word w. For instance, for the morphism defined by

$$\varphi(a) = a, \quad \varphi(b) = b, \quad \varphi(c) = c_1 \dots c_n,$$

the word $\varphi(w) = a^2b^3(c_1 \dots c_n)^3a^4b^3(c_1 \dots c_n)^3a^3$ cannot be a P-word because of Lemma 2.18 but it is an S-word by Theorem 2.9.

We have thus obtained:

Theorem 2.26 *In every alphabet of cardinality at least three there exists an S-word that is not a P-word.*

2.4 Periodicity forcing words and sets

We first introduce some results in alphabets with more than two letters and then concentrate on the binary case in the next section.

The first example of periodicity forcing words in a three-letter alphabet was the word $w = a^2b^3c^3a^4b^3c^3a^3$ in Theorem 2.25, originally presented in [18]. Applying the same argument for an arbitrary prime p (instead of 3) it is easy to prove the following.

Theorem 2.27 *The words $w = avava$, where $v = a^{i_1}b^{j_1}c^{j_2}a^{i_2}$ and*

$$\begin{aligned}
&3 + 2(i_1 + i_2) \equiv 0 \pmod p \\
&i_1 + i_2 \geq 1 \\
&j_1 \equiv 0 \pmod p, \quad j_2 \equiv 0 \pmod p
\end{aligned}$$

for some prime $p \neq 2$, are periodicity forcing.

Notice that all the conditions in Theorem 2.27 are necessary.

Example 2.12 Let $i_1 = 2, i_2 = 1, j_1 = 2, j_2 = 3$ and $p = 3$ (satisfying the first two conditions). Then the word $w = avava$ is a solution for the instance

	a	b	c
g	$(o^2p^3)^3$	o	p
h	o^2p^3	$(o^2p^3)^2$	$(o^2p^3)^2$

With $i_1 = 2, i_2 = 1$ and $j_1 = j_2 = 2$ the word $w = avava$ is in the equality set of the instance

	a	b	c
g	$(o^2p^2)^3$	o	p
h	o^2p^2	$(o^2p^2)^2$	$(o^2p^2)^3$

Defining finally $i_1 = 1, i_2 = 0$ and $j_1 = j_2 = 3$ (satisfying all but the first condition), the word $w = avava$ is a solution for the instance

	a	b	c
g	$(o^3p^3)^3$	o	p
h	o^3p^3	o^3p^3	o^3p^3

□

Consider now words in larger alphabets, $\Sigma = \{a, b_2, \ldots, b_n\}$, $n \geq 2$.

Theorem 2.28 *The words $w = avava$, where*

$$\begin{aligned}
&v = a^{i_1} b_2^{j_2} a^{i_2} b_3^{j_3} \ldots a^{i_{n-1}} b_n^{j_n} a^{i_n},\\
&i_1 + i_n \neq 0,\\
&j_k \neq 0 \quad \textit{for } k = 2, \ldots, n\\
&i_l \neq 0 \quad \textit{for } l = 2, \ldots, n-1
\end{aligned}$$

are periodicity forcing.

Proof: Assume that $w \in E(g, h)$ for some instance (g, h). Then the words $g(a), g(v), h(a)$ and $h(v)$ are, by Lemma 2.5, powers of the same word. Since

$$\begin{aligned}
g(v) &= g(a)^{i_1} g(b_2)^{j_2} g(a)^{i_2} \ldots g(b_n)^{j_n} g(a)^{i_n}\\
h(v) &= h(a)^{i_1} h(b_2)^{j_2} h(a)^{i_2} \ldots h(b_n)^{j_n} h(a)^{i_n}
\end{aligned}$$

it is easy to prove by induction on n and applying Lemma 1.4 that the words $g(a), g(b_2), \ldots, g(b_n), h(a), h(b_2), \ldots, h(b_n)$ are powers of the same word. □

The letters b_k can actually occur in any order in v as long as a separates two different b's. To show why this is important and why Theorem 2.27 is not valid in alphabets with at least four letters, we consider the following example.

Example 2.13 Let $v = ab^3c^3ad^3a$. By Lemma 2.5, the words $g(a)$, $g(b)^3g(c)^3$, $g(d)$, $h(a)$, $h(b)^3h(c)^3$ and $h(d)$ are powers of the same word. Still $w = avava$ is in the equality set of the following nonperiodic and nontrivial instance

	a	b	c	d
g	$((o^3p)^3o^3)^4$	o^3p	o	$((o^3p)^3o^3)^3$
h	$((o^3p)^3o^3)^2$	o	po^3	$((o^3p)^3o^3)^6$

□

Theorems 2.27 and 2.28 rely both on Lemma 2.5. Example 2.12 shows that we have to set restrictions on the word v to make w periodicity forcing. Still these two results are in no way exhaustive: other types of restrictions are possible. We can also change the basic form $avava$.

For instance, Lemma 2.16 where 112122 was proved periodicity forcing can also be viewed as

$$aavavv \in E(g,h) \;\Rightarrow\; g(a), g(v), h(a), h(v) \in u^+ \text{ for some } u \in \Delta^+.$$

Applying now the same argument as in Theorem 2.28 we obtain the following result.

Theorem 2.29 *The words $w = aavavv$, where*

$$\begin{aligned} &v = a^{i_1} b_2^{j_2} \ldots a^{i_{n-1}} b_n^{j_n} a^{i_n},\\ &j_k \neq 0 \quad \textit{for } k = 2, \ldots, n\\ &i_l \neq 0 \quad \textit{for } l = 2, \ldots, n-1 \end{aligned}$$

are periodicity forcing.

In this case, too, v cannot be chosen freely.

Example 2.14 The word $w = aavavv$ where $v = bc$ is a solution for the following nonperiodic instance (g,h)

	a	b	c
g	$(op)^2$	o	p
h	op	op	op

□

Examples 2.12 and 2.14 show that the property of being periodicity forcing is not usually preserved by morphisms though there are some obvious exceptions.

Lemma 2.30 *If w is periodicity forcing and φ is a permutation or φ is an injection and for all $a \in \Sigma$ there exists $b \in \Delta$ such that $\varphi(a) = b^{i_a}$ for some $i_a \geq 1$ then $\varphi(w)$ is also periodicity forcing.*

The study of periodicity forcing words in larger alphabets sheds new light on attempts to characterize P-words and S-words. The following two results are easy to prove.

Lemma 2.31 *If w is periodicity forcing and $\psi(w) = K(k_1, \ldots, k_n)$ where $K, n \geq 2$ then w is not a P-word.*

Lemma 2.32 *If w is periodicity forcing but not subratioprimitive then w is not an S-word.*

In a binary alphabet these two results amount to the same thing because of Lemma 2.3. We notice that the words in Theorem 2.27 satisfy the conditions of Lemma 2.31 but are subratioprimitive; hence, they are S-words but not P-words. In Theorems 2.28 and 2.29 some words, such as $a^3b^3ac^3a^3b^3ac^3a$ and $a^2bacabacbac$, are not even S-words by Lemma 2.32 but they are all F-words.

Nevertheless, Example 2.9 shows that this kind of study is not exhaustive, we have to consider also the words that are not periodicity forcing in order to characterize P-words and S-words.

2.4.1 Binary case

In this section we take an overview of [3] where the study of periodicity forcing words was actually initiated. We start with words, go then to some sets and finally we introduce some open problems which have a strong relation to the conjectures mentioned earlier. Lemmas 2.33, 2.40–2.42, 2.45–2.46 and Theorems 2.34, 2.38–2.39, 2.47 are due to [3].

In Sections 2.1.1 and 2.2 we have already considered several periodicity forcing words in a binary alphabet. The shortest one had six letters.

Lemma 2.33 *There are no periodicity forcing words of length shorter than five.*

Theorem 2.34 *A word $w \in \{a, b\}^*$ of length five is periodicity forcing if and only if $w = abaab$ or $w = babaa$ or w is obtained from these by interchanging letters or by taking mirror images or by making both the operations.*

Also words that are not ratioprimitive have interesting properties.

Lemma 2.35 *A word w is periodicity forcing if and only if w^2 is periodicity forcing.*

Lemma 2.36 *The word $w = uv$ is periodicity forcing if u is an F-prefix and u or v is periodicity forcing.*

From Lemmas 2.21 and 2.22 we already know what equality sets are like in a binary alphabet. If they are finitely generated they have at most two generators. If we assume that a set X is not a subset of a reduced language (see Lemma 2.22) then we have the following result.

Theorem 2.37 *The set $X = \{u, v, x\}$ where u and v are ratioprimitive and $x \notin \{u, v\}^+$ is periodicity forcing.*

We now consider some periodicity forcing sets with two elements.

Theorem 2.38 *Each of the following sets is periodicity forcing*

$$\begin{array}{rl} \{ab, ax\}, & \textit{where } x \neq b \textit{ and } ax \textit{ is ratioprimitive,} \\ \{aab, ax\}, & \textit{where } x \neq ab \textit{ and } ax \textit{ is ratioprimitive,} \\ \{aaab, ax\}, & \textit{where } x \neq aab \textit{ and } ax \textit{ is ratioprimitive.} \end{array}$$

Theorem 2.39 *Each of the following sets is periodicity forcing*

$$\begin{array}{rl} \{aabb, x\}, & \textit{where } \rho(x) \neq aabb, \\ \{aba, x\}, & \textit{where } \rho(x) \neq aba, \\ \{aaba, x\}, & \textit{where } \rho(x) \neq aaba. \end{array}$$

Each of these results is obtained by finding first all pairs of morphisms agreeing on a given word such as in Lemmas 2.14 and 2.15.

Lemma 2.40 *The word ab is in $E(g, h)$, with $|h(a)| > |g(a)|$, if and only if there exist words α, β and γ such that*

	a	b
g	α	$\beta\gamma$
h	$\alpha\beta$	γ

Lemma 2.41 *The set $\{ab, ba\}$ belongs to $E(g, h)$, with $|h(a)| > |g(a)|$, if and only if there exist nonnegative integers t_1, t_2 and t_3 and words α and β such that $t_2 > 0$, $\alpha\beta \neq \lambda$ and*

	a	b
g	$\alpha(\beta\alpha)^{t_1}$	$\beta(\alpha\beta)^{t_2+t_3}$
h	$\alpha(\beta\alpha)^{t_1+t_2}$	$\beta(\alpha\beta)^{t_3}$

Moreover, if $\{ab, ba\} \subset E(g,h)$, $g \neq h$, *and* h *or* g *is elementary, then* $E(g,h) = \{ab, ba\}^+$.

Lemma 2.42 *The word* aba *is in* $E(g,h)$, *with* $|h(a)| > |g(a)|$, *if and only if there exist an integer* $t \geq 0$ *and words* α, β *and* γ *such that* $\alpha\beta \neq \lambda$ *and*

	a	b
g	$\alpha(\beta\alpha)^t$	$\beta\alpha\gamma\alpha\beta$
h	$\alpha(\beta\alpha)^t\beta\alpha$	γ

Lemma 2.43 *The word* $aaab$ *is in* $E(g,h)$, *with* $|h(a)| > |g(a)|$, *if and only if there exist an integer* $t \geq 0$ *and words* α, β *and* γ *such that* $\alpha\beta \neq \lambda$ *and*

	a	b
g	$(\alpha\beta)^t\alpha$	$\beta\alpha(\alpha\beta)^t\alpha(\alpha\beta)^{t+1}\alpha(\alpha\beta)^t\alpha\alpha\beta\gamma$
h	$(\alpha\beta)^t\alpha(\alpha\beta)^t\alpha\alpha\beta$	γ

Lemma 2.44 *The word* $aaba$ *is in* $E(g,h)$, *with* $|h(a)| > |g(a)|$, *if and only if there exist an integer* $t \geq 0$ *and words* α, β *and* γ *such that* $\alpha\beta \neq \lambda$ *and*

	a	b
g	$(\alpha\beta)^t\alpha$	$\beta\alpha(\alpha\beta)^t\alpha(\alpha\beta)^{t+1}\alpha\gamma(\alpha\beta)^t\alpha\alpha\beta$
h	$(\alpha\beta)^t\alpha(\alpha\beta)^{t+1}\alpha$	γ

This seems to hold also in general. Words of the form a^iba^j, $i+j \geq 1$, are not periodicity forcing but they have the same properties as in Theorems 2.38 and 2.39.

Our next aim is to characterize periodicity forcing sets included in $a^*b^* \cup b^*a^*$.

Lemma 2.45 *Let* i, j, m *and* n *be natural numbers satisfying* $i \cdot n \neq 1$ *and* $j \cdot m \neq 1$. *Then the set* $\{a^ib^j, b^ma^n\}$ *is periodicity forcing.*

Notice that with $i \cdot n = 1$ and $j \neq m$ (or $j \cdot m = 1$ and $i \neq n$) the set is also periodicity forcing by Lemma 1.8.

Lemma 2.46 *Any two-element set in a^+b^+ is periodicity forcing.*

Theorem 2.47 *The subset X of $a^+b^+ \cup b^+a^+$ is periodicity forcing if and only if $|X| \geq 2$ and X is not of the form $\{a^ib, ba^i\}$, $i \geq 1$ (or symmetrically of the form $\{ab^i, b^ia\}$, $i \geq 1$).*

In [3] also the following generalization was mentioned: any two-element set in $a^+b^*a^*$ is periodicity forcing. We conjecture further that the only kind of sets not periodicity forcing are the trivial case and Example 2.11.

Conjecture 2.48 *The subset $X = \{u, v\}$ of $\{a, b\}^+$ is periodicity forcing if u and v are ratioprimitive and $X \neq \{a, b\}$ or $X \neq \{a^ib, ba^i\}$ (or symmetrically $\{ab^i, b^ia\}$).*

Conjecture 2.49 *Any word u over $\{a, b\}$ is periodicity forcing if it is ratioprimitive, $u \notin a^+b^+ \cup b^+a^+$ and $\gcd(|u|_1, |u|_2) > 1$.*

If these conjectures are shown to be true, the first result implies the equality of P- and S-words and the second one their characterization in a binary alphabet.

Chapter 3

Prime Words: Special Types

So far we have studied only three types of prime solutions based on removing a suffix, a subword or a scattered subword such that the remaining part is not a solution. There are also other ways to define the primality.

In Section 3.1 we study the usual types of P-, S- and F-words with the exception that they should now be prime solutions in every instance (g, h) where they are solutions. This restriction narrows the amount of prime solutions; on the other hand, some of the problems left open in the first chapter can now be solved in this restricted setup. For instance, P- words and S-words defined in this restricted way turn out to be equal. We also compare these restricted types to the usual P-, S- and F-words, which, as well as the other results, were originally published in [19].

Section 3.2 is devoted to prime solutions defined by partial orders. Hence a solution is primitive in this sense if there is no solution which is "smaller" with respect to the given order. In this way we can define the usual F-prime and P-prime solutions but not S-prime solutions. We consider several types of partial orders introduced in [33] and examine what kind of hierarchical structure the corresponding prime solutions have. The study of the finite primality type is due to [17].

In Section 3.3 we consider ways to add an arbitrary language L to the definitions of P-, S- and F-prime solutions. Especially interesting is then to study what the effect of the different set-theoretic operations of languages is on the corresponding L-prime solutions. Also prime solutions related to some special languages, such as flatwords, are considered. Since L-prime solutions can be defined in several ways, our study is in no way exhaustive.

3.1 Strongly prime PCP-words

Throughout the previous chapter a word is termed primitive if it is a prime solution for *some* instance (g, h). Now we study words that are primitive for *all* instances (g, h) for which they are solutions. We shall restrict ourselves to instances (g, h) where $g \neq h$; otherwise, none of the words w, $|w| > 1$, is a prime solution in this sense.

Let us denote by $\mathcal{H}$ the set of all instances (g, h), $g \neq h$. Then the usual prime words can be defined as follows: A word $w \in \Sigma^+$ is an *F-word* if $w \in \cup_{(g,h)\in\mathcal{H}} F(g, h)$. *S-words* and *P-words* are obtained similarly.

We denote by F, S and P the sets of all F-words, S-words and P-words, respectively, and by F_n, S_n and P_n the corresponding sets when the alphabet Σ is of a bounded size, card$(\Sigma) = n$. Denote also by E_n the set of the solutions of PCP for such an alphabet Σ.

The idea in defining strongly prime PCP-words is to replace - in some sense - the union in the definition of F-, S- and P-words by the intersection.

More specifically, we say that $w \in \Sigma^*$ is *strongly X-prime*, $X \in \{F, S, P\}$, if

1. there is an instance $(g, h) \in \mathcal{H}$ such that $w \in E(g, h)$, and
2. for all $(g, h) \in \mathcal{H}$ for which $w \in E(g, h)$ we have $w \in X(g, h)$.

We denote by X'_n the set of all strongly X-prime words over alphabets with n symbols, $X \in \{F, S, P\}$.

From definitions, we have

$$F'_n \subseteq F_n,\ S'_n \subseteq S_n,\ P'_n \subseteq P_n,$$
$$P'_n \subseteq S'_n \subseteq F'_n.$$

We first show that for alphabets with at least three letters the definition is too restrictive.

Theorem 3.1 *$F'_n = \emptyset$ for $n \geq 3$.*

Proof: Let w be a word over an n-letter alphabet, $n \geq 3$. Assume that $w = iw'$ and consider two letters j, k different from i. Denote

$$r = |w|_j,\ s = |w|_k.$$

For the morphisms g, h defined by

	i	j	k	l_1	l_2	$\ldots$
g	a	a^{2s}	a^{2r}	a	a	$\ldots$
h	a	a^{s}	a^{3r}	a	a	$\ldots$

we obtain

$$g(w) = h(w) = a^m, \quad m = 4rs + \sum_{p \notin \{j,k\}} |w|_p.$$

Since $g(i) = h(i)$, we have $w \in E(g,h) - F(g,h)$. Hence, by definition, $w \notin F'$. $\square$

Note that the previous proof is essentially based on the possibility to use periodic morphisms (with the same period). Is the result true also when we ignore such morphisms? **Conjecture**: Not. To make the conjecture plausible we notice that by ignoring also trivial instances, that is, imposing that $g(a) \neq h(a)$ for all $a \in \Sigma$, the set F_3', defined in this restricted way, is not empty. Indeed, $123 \in F_3'$ since both $1, 123 \in E(g,h)$ and $12, 123 \in E(g,h)$ imply that the corresponding instance is trivial.

Anyway, for the set-up we have started with, only the case $n = 2$ remains to be of interest. (The case $n = 1$ is trivial: $F_1 = F_1' = S_1 = S_1' = P_1 = P_1'$.) The following results rely heavily on the theory developed in Chapter 2.

Theorem 3.2 *The word w is in $P_2' = S_2'$ if and only if* $\gcd(|w|_1, |w|_2) = 1$.

Proof: A word w of the form $\gcd(|w|_1, |w|_2) = 1$ is indeed strongly P-prime by Lemma 1.8, since its proper scattered subwords are all of different ratio.

On the other hand, if $\gcd(|w|_1, |w|_2) = t > 1$ we can distinguish a subword v (by Lemma 2.3) for which $\gcd(|v|_1, |v|_2) = \gcd(r,s) = 1$ and $\psi(w) = t\psi(v)$. For the morphisms g, h defined by

	1	2
g	a	a^{r+1}
h	a^{s+1}	a

both v and w are solutions ($g(v) = h(v) = a^{r+s+rs}$ and $g(w) = h(w) = a^{t(r+s+rs)}$). Hence $w \notin S_2'$. $\square$

Together with Theorem 2.10 this implies that the sets P_2 and $S_2' = P_2'$ are not equal since $1^i2^i \notin P_2'$ for $i \geq 2$.

The relation between F_2 and F_2' still remains to be settled.

Theorem 3.3 $F_2 = F_2'$.

Proof: We have already pointed out that $F_2' \subseteq F_2$. Conversely, take $w \in F_2$. If $w \notin F_2'$ then there is a nontrivial instance (g, h) such that $w = vx$, for some $v \in E(g, h)$, $v \neq w$. From Lemma 1.8 it follows that $\frac{|v|_1}{|v|_2} = \frac{|w|_1}{|w|_2}$. Choosing $r = \frac{|v|_1}{|w|_1}$ we obtain $\psi(v) = r\psi(w)$, hence by Lemma 2.2, $w \notin F_2$, a contradiction. Consequently, $w \in F_2'$ and $F_2 = F_2'$. □

From Theorem 2.8 we now obtain the following characterization.

Theorem 3.4 *A word w is in $F_2 = F_2'$ if and only if w is ratioprimitive.*

Summarizing the results above, we have

$$P_2' = S_2' \subset P_2 \subseteq S_2 \subset F_2 = F_2' \subset E_2$$

all inclusions being proper except possibly $P_2 \subseteq S_2$. Furthermore, these sets have all been characterized except the sets P_2 and S_2 for which we proposed Conjecture 2.13.

In the original paper [19] we assumed that $P_2 = S_2$ as in [16]. Now we know that the proof of this result was inaccurate and only the relation $P_2 \subseteq S_2$ is used here. This affects, apart from the hierarchy, also Theorem 3.2 where the proof has been modified.

3.2 Primitive solutions with respect to partial orders

For many purposes it is useful to regard the languages $F(g, h)$ and $P(g, h)$ as obtained from the equality set $E(g, h)$ by certain operations.

We recall that a *partial order* defined on a monoid is a relation $\leq$, which is reflexive, antisymmetric and transitive. The notion "partial" means that the words u and v can also be *incomparable*: neither $u \leq v$ nor $v \leq u$. Otherwise we talk about *total order*.

Consider a partial order $\leq$ on Σ^* and a language $L \subseteq \Sigma^*$. Then $MIN_{\leq}(L)$ is the subset of L consisting of elements that are minimal with respect to $\leq$, in symbols,

$$MIN_{\leq}(L) = \{x \in L \mid \text{whenever } y \in L \text{ satisfies } y \leq x, \text{ then } y = x\}.$$

Thus the words in the *minimal set* (or *independent set*), $MIN_{\leq}(L)$, are incomparable (with respect to $\leq$).

For the partial orders

$$uFw \text{ if } u \in \mathit{fin}(w),$$
$$uPw \text{ if } u \in \mathit{scatsub}(w),$$

we have

$$F(g,h) = MIN_F(E(g,h)),$$
$$P(g,h) = MIN_P(E(g,h)).$$

But the set $S(g,h)$ cannot be expressed in this way since the relation

$$ySw \text{ if } y \in \mathit{sub}(w)$$

is not a partial order. Indeed, it lacks the property of being transitive: we have $acSabc$ and $abcSabcb$ but not $acSabcb$. However, the following theorem from [23] holds.

Theorem 3.5 *The transitive closure S^+ of S is a partial order and, moreover, $S^+ = P$.*

Several types of partial orders were defined in [33, pp. 64–65]. In some sense, all of them have already appeared in our considerations.

Let $v, w \in \Sigma^*$. We define the relations $\leq_h$, $\leq_u$, $\leq_p$, $\leq_s$, $\leq_b$, $\leq_d$, $\leq_c$ and $\leq_e$ as follows:

1. $v \leq_h w$ if $v = v_1 \ldots v_n$ and $w = x_0 v_1 x_1 \ldots v_n x_n$ for some $n \geq 0$, $v_i, x_i \in \Sigma^*$.
2. $v \leq_u w$ if $v = w$ or $|v| < |w|$.
3. $v \leq_p w$ if $w = vx$ for some $x \in \Sigma^*$.
4. $v \leq_s w$ if $w = xv$ for some $x \in \Sigma^*$.
5. $v \leq_b w$ if $w = xvy$ for some $x, y \in \Sigma^*$.
6. $v \leq_d w$ if $w = vx = yv$ for some $x, y \in \Sigma^*$.
7. $v \leq_c w$ if $w = xv = vx$ for some $x \in \Sigma^*$.

8. $v \leq_e w$ if $w = v^n$ for some $n \geq 1$.

9. Denote by $\leq_l$ the lexicographic order.

Also the following partial order, introduced in [18], is natural from the point of view of the simplicity of solutions:

$$v \leq_\psi w \text{ if } v = w \text{ or } \psi(v) < \psi(w).$$

We have actually applied this relation already in Section 2.2 where the notion Parikh shorter was used. In the definition above v is Parikh shorter than w.

We notice immediately that the relations $\leq_h$ and $\leq_p$ are the same as P and F defined above. An interesting observation is also that $\leq_l$ is actually a total order. Any two words are comparable with respect to this order.

For any two partial orders, say, $\leq_\alpha$ and $\leq_\beta$ defined on Σ^*, we define

$$(\leq_\alpha) \subseteq (\leq_\beta) \quad \text{if} \quad u \leq_\alpha v \text{ implies } u \leq_\beta v \text{ for all } u, v \in \Sigma^*.$$

By $(\leq_\alpha) \subset (\leq_\beta)$ we mean that $(\leq_\alpha) \subseteq (\leq_\beta)$ but $\leq_\alpha$ is not identical to $\leq_\beta$. The partial orders 1–9 defined above can be compared as follows: If Σ contains more than one letter, then

$$\begin{gathered}(\leq_e) \subset (\leq_c) \subset (\leq_d) \subset (\leq_p) \subset (\leq_b) \subset (\leq_h) \subset (\leq_u) \subset (\leq_l), \\ (\leq_d) \subset (\leq_s) \subset (\leq_b);\end{gathered} \tag{3.1}$$

and if Σ contains only one letter, all the partial orders are identical.

3.2.1 Hierarchy

We now consider the primitive solutions corresponding to the partial orders defined above. The minimal sets $MIN_\leq(E(g,h))$ obtained in this way are compared to each other and to the usual sets of P-prime, S-prime and F-prime solutions. Clearly, for any two partial orders $\leq_\alpha$ and $\leq_\beta$,

$$(\leq_\alpha) \subset (\leq_\beta) \quad \Rightarrow \quad MIN_{\leq_\beta}(E(g,h)) \subseteq MIN_{\leq_\alpha}(E(g,h)).$$

So the inclusions in (1) are reversed for the minimal sets, with the exception that some of them turn out to be equalities.

As mentioned earlier, $MIN_{\leq_h}(E(g,h)) = P(g,h)$ and $MIN_{\leq_p}(E(g,h)) = F(g,h)$ for every instance (g,h). Consider now the other partial orders.

$\leq_\psi$: This partial order compares the Parikh vectors of words componentwise. Thus, intuitively, we regard a solution w primitive if no word obtained from the letters of w but not using all of them, is a solution. If w is primitive in this sense it is also P-prime because for any scattered subword w_0 of w we have also $w_0 \leq_\psi w$.

$\leq_u$: The point of this partial order is to find the shortest words in the equality sets. We notice that if v is Parikh shorter than w then the length of v must also be shorter than that of w.

$\leq_s$: The primitive solutions with respect to this order are exactly the F-prime solutions. Indeed, if $w = vx \in E(g,h)$ then $v \in E(g,h)$ if and only if $x \in E(g,h)$.

$\leq_b$: This relation is, in a sense, converse to the usual S-type though in periodic instances they are equal. Consider the instances (g_1, h_1) and (g_2, h_2) defined by

	1	2	3
g_1	ab	ab	e
h_1	a	ba	be

	1	2	3
g_2	$(ab)^3$	a	ba
h_2	ab	aba	$(ba)^2$

Then

$$P(g_1,h_1) = S(g_1,h_1) = \{13\},$$
$$MIN_{\leq_b}(E(g_1,h_1)) = 12^*3,$$

but

$$S(g_2,h_2) = \{1^i 23^{2i-1} \mid i \geq 1\},$$
$$MIN_{\leq_b}(E(g_2,h_2)) = P(g_2,h_2) = \{123\}.$$

So the sets $S(g,h)$ and $MIN_{\leq_b}(E(g,h))$ cannot be compared with each other in general.

$\leq_d$: This partial order broadens the F-type. If the instance (g,h) is defined by

	1	2
g	a^2	a
h	a	a^2

then the minimal set $MIN_{\leq_d}(E(g,h))$ contains, in addition to F-prime solutions, the words

$$1221,\ 2112,\ 121122,\ 211212, \ldots$$

$\leq_c$: Here we extend the previous type further. Let the instance (g,h) be the same as above. Then the set $MIN_{\leq_c}(E(g,h))$ contains words, such as

$$122112,\ 211221,\ 12112212,\ 12221112,$$

none of which belongs to the previous set.

$\leq_e$: The point of this type is to find words that are not powers of any other words in the equality set; nevertheless, its minimal set is the same as $MIN_{\leq_c}(E(g,h))$. Indeed, by Lemma 1.1 we know that if v is primitive then $w = vx = xv$ if and only if $w = v^n$ for some $n \geq 1$.

$\leq_l$: We use a modification of this partial order to make it fit better into other sets. We say that the word w belongs to the set $MIN_{\leq_l}(E(g,h))$ if and only if $w \in E(g,h)$ and you cannot remove a subword (respectively a scattered subword) in such a way that the word v formed from the remaining parts has the properties $v \leq_l w$ and $v \in E(g,h)$.

For the instance (g_1,h_1) defined in "$\leq_b$", $MIN_{\leq_l}(E(g_1,h_1)) = 12^*3$ no matter whether we consider removing a subword or a scattered subword. On the other hand, the instance in "$\leq_d$" separates the two sets. Let S_l equal the minimal set in connection with removing a subword and P_l with removing a scattered subword. Then

$$\begin{aligned} S_l(g,h) &\supset \{1^i2^i, 1^i(21)^k2^i, 1^i12212^i, 1^i1122212^i, 1^i1221212^i\} \\ P_l(g,h) &\supset \{1^i2^i, 1^i212^i, 1^i12212^i, 1^i1122212^i\}. \end{aligned}$$

The hierarchy $P_l(g,h) \subset S_l(g,h) \subset F(g,h)$ follows from the definition and the fact that if w is not an F-prime solution, i.e. $w = vx$ for some $v \in E(g,h)$, then $w \notin S_l(g,h)$ since $v \leq_l w$.

The inclusions between the different families are depicted by the following two diagrams (an inclusion is denoted by an arrow):

$$\begin{tikzcd}
& L = E(g,h) & \\
& MIN_{\leq_e}(L) = MIN_{\leq_c}(L) \arrow[u] & \\
& MIN_{\leq_d}(L) \arrow[u] & \\
& MIN_{\leq_p}(L) = MIN_{\leq_s}(L) = F(g,h) \arrow[u] & \\
MIN_{\leq_b}(L) \arrow[ur] & & S(g,h) \arrow[ul] \\
& MIN_{\leq_h}(L) = P(g,h) \arrow[ul] \arrow[ur] & \\
& MIN_{\leq_\psi}(L) \arrow[u] & \\
& MIN_{\leq_u}(L) \arrow[u] &
\end{tikzcd}$$

For the lexicographic order we obtain

$$\begin{tikzcd}
& F(g,h) & \\
& S_l(g,h) \arrow[u] & \\
P_l(g,h) \arrow[ur] & & S(g,h) \arrow[ul] \\
& P(g,h) \arrow[ul] \arrow[ur] &
\end{tikzcd}$$

In "$\leq_b$" we showed that $MIN_{\leq_b}(L)$ and $S(g,h)$ are incomparable with each other. The fact that $S(g,h)$ is incomparable also with $P_l(g,h)$, is shown in the following example.

Example 3.1 In our earlier instance we had $S(g,h) \subset P_l(g,h)$. Consider now

	1	2	3	4
g	ab	cb	cde	def
h	abc	$bcde$	de	f

This time $S(g,h) = \{13, 1234\} = S_l(g,h)$ and $P(g,h) = \{13\} = P_l(g,h)$. $\square$

3.2.2 Sharpening in the finite case

The set of P-prime solutions is, by Theorem 2.1, known to be finite for every instance (g, h). This is actually due to the following result given explicitly in [12] as a corollary of a stronger statement.

Theorem 3.6 *Any subset of words over a finite alphabet that are not comparable with respect to $\leq_h$ is finite.*

Hence from the hierarchy of minimal sets it follows that also the sets

$$\begin{aligned} M(g,h) &= MIN_{\leq_u}(E(g,h)), \\ \Psi(g,h) &= MIN_{\leq_\psi}(E(g,h)), \end{aligned}$$

are finite and if their cardinalities are denoted by μ and ψ, respectively, then

$$\mu \leq \psi \leq p.$$

The triple (μ, ψ, p) is defined to be the *finite primality type* of the instance PCP=(g, h). Recall that the usual primality types (p, s, f) where p, s and f are the cardinalities of P-prime, S-prime and F-prime solutions were characterized in Theorem 2.1.

We prepare the characterization of the finite primality types with a result concerning the pairs (ψ, p).

Lemma 3.7 *For each pair $(1, n)$, $n \geq 1$, we can effectively construct an instance (g, h) for which $\psi = 1$ and $p = n$.*

Proof: The main idea in constructing the instance (g_n, h_n) giving rise to the pair $(1, n)$ is to enlarge the alphabet Σ with two letters and the alphabet Δ with one letter at a time.

We begin with a solution 12. Next we define the images of the letters 3 and 4 such that 3214 is also a solution. Clearly, $12 \notin \mathit{scatsub}(3214)$ whereas $\psi(12) = (1,1) < (1,1,1,1) = \psi(3214)$. Similarly, we define 5 and 6 such that 521436 is a solution, and so on. Thus every time we obtain one new P-prime solution whereas the number of Ψ-prime solutions remains the same.

The first four instances (g_i, h_i) are as follows.

	1	2
g_1	ab	a
h_1	a	ba

	1	2	3	4
g_2	ab	a	cba^2	c
h_2	a	ba	c	a^2bc

	1	2	3	4	5	6
g_3	ab	a	cba^2	c	dba^4bc^2	d
h_3	a	ba	c	a^2bc	d	$a^2bc^2ba^2d$

	1	...	6	7	8
g_4	ab	...	d	$fba^4bc^2a^2bc^2ba^2d^2$	f
h_4	a	...	$a^2bc^2ba^2d$	f	$a^2bc^2ba^2d^2ba^4bc^2f$

We clearly have

$$
\begin{aligned}
P(g_1,h_1) &= \{12\} = \Psi(g_1,h_1),\\
P(g_2,h_2) &= \{12, 3214\},\\
P(g_3,h_3) &= \{12, 3214, 521436\},\\
P(g_4,h_4) &= \{12, 3214, 521436, 72143658\}.
\end{aligned}
$$

Notice also the following equations

$$
\begin{gathered}
g_2(3) = h_2(321),\ g_2(214) = h_2(4),\ g_2(4) = h_2(3) \notin \Delta_2 = \{a, b\},\\
g_3(5) = h_3(52143),\ g_3(21436) = h_3(6),\ g_3(6) = h_3(5) \notin \Delta_3 = \{a, b, c\},\\
g_4(7) = h_4(7214365),\ g_4(2143658) = h_4(8),\ g_4(8) = h_4(7) \notin \Delta_4 = \{a, b, c, d\}.
\end{gathered}
$$

Proceeding inductively, we show that $|P(g_n, h_n)| = n$. This is clearly the case for the instance (g_1, h_1). Since $|P(g_i, h_i)| = i$ the letters $\{1, \ldots, 2i\}$ cannot produce any other P-primes in the instance (g_{i+1}, h_{i+1}) either. On the other hand, if we use the letters $2i+1$ and $2i+2$, the only possibility is that $2i+1$ begins and $2i+2$ ends the solution. Together with the equations

$$
\begin{gathered}
g(2i+1) = h(2i+1)h(2)h(1)\ldots h(2i-1)\\
g(2)g(1)\ldots g(2i-1)g(2i+2) = h(2i+2)
\end{gathered}
$$

the proof is completed. □

We are now ready to state our characterization result.

Theorem 3.8 *The triple (μ, ψ, p) is a finite primality type if and only if either $0 = \mu = \psi = p$ or $1 \le \mu \le \psi \le p$ and p is finite. An example for each possible type can be effectively constructed.*

Proof: We divide the argument into four subcases.

Case I: $\mu = \psi = p$.

Consider the instance (g, h) defined by

	1_1	$\ldots$	1_k
g	a_1	$\ldots$	a_k
h	a_1	$\ldots$	a_k

The triple is clearly (k, k, k), $k \geq 1$.

Case II: $\mu < \psi = p$

Define the morphisms g and h by

	1_k	2_i	3_i
g	a_k	$b_i c_i$	b_i
h	a_k	b_i	$c_i b_i$

Now $2_i 3_i \notin M(g, h)$ since $|2_i 3_i| > |1_k|$. Thus our triple is $(k, k+i, k+i)$, $k, i \geq 1$.

Case III: $\mu = \psi < p$

Let the instance (g, h) be defined by

	1_i	2_i	1	2	3	4	5	6	$\ldots$
g	$a_i b_i$	a_i	ab	a	cba^2	c	dba^4bc^2	d	$\ldots$
h	a_i	$b_i a_i$	a	ba	c	a^2bc	d	$a^2bc^2ba^2d$	$\ldots$

By Lemma 3.7 we obtain the triple $(i+1, i+1, i+n)$, $i \geq 1$, $n \geq 2$, whereas the instance (g, h) without the letters 1_i and 2_i gives the triples $(1, 1, n)$, $n \geq 2$.

Case IV: $\mu < \psi < p$

Consider the instance

	1_k	2_i	3_i	1	2	3	4	5	6	$\ldots$
g	a_k	$b_i c_i$	b_i	ab	a	cba^2	c	dba^4bc^2	d	$\ldots$
h	a_k	b_i	$c_i b_i$	a	ba	c	a^2bc	d	$a^2bc^2ba^2d$	$\ldots$

Again by Lemma 3.7 we obtain the triples $(k, k+i+1, k+i+n)$, $k, i \geq 1$, $n \geq 2$ and without the letters 2_i and 3_i the triples are $(k, k+1, k+n)$, $k \geq 1$, $n \geq 2$. (Observe that in the last two cases the letters 5 and 6 are not needed if $n = 2$.) $\square$

The statement of Theorem 3.8 has been slightly changed here from the original one in [17]. If we just assume that $\mu \leq \psi \leq p$ then also the triples $(0, 0, k)$ and $(0, k, k+i)$ would be potential contradicting the definition of prime solutions.

Any other combination of the cardinalities of the minimal sets in Section 3.2.1 may also give interesting properties. As an example we end this section with the mixture of the two known types, (μ, ψ, p) and (p, s, f).

Theorem 3.9 *The five-tuple (μ, ψ, p, s, f) is acceptable if and only if either $0 = \mu = \psi = p = s = f$ or each of the conditions (i)-(iii) is satisfied: (i) $1 \leq \mu \leq \psi \leq p \leq s \leq f$, (ii) p is finite, (iii) if $s < f$ then f is infinite. For each acceptable quintuple, an instance (g, h) having (p, s, f) as its primality type and (μ, ψ, p) as its finite primality type can be effectively constructed.*

3.3 Prime solutions relative to languages

In Section 2.1 a solution was termed primitive if no suffix/subword/scattered subword could be removed such that the remainder is also a solution. Now we consider ways to fit an arbitrary language L into the definition. We may require that the part to be removed is a member of L or, alternatively, the remaining part should belong to L. Even the original solution can be required to be a member of L. As in the usual definition, one checks also here that the remaining part is not a solution.

We start with the definition, due to prof. Alexandru Mateescu, where a solution is termed primitive if it cannot be divided in such a way that the part to be removed belongs to L and at the same time the remaining part is a solution. More specifically, let $L \subset \Sigma^*$. A solution $w \in E(g, h)$ is F_L-*prime* if there is no proper suffix w' of w such that $w' \in L$ and the remainder is a solution. Analogously, S_L-*prime* and P_L-*prime solutions* correspond to checking whether any subword or a scattered subword belongs to L.

In the next two sections we study how the set-theoretic operations, such as union, intersection, catenation and complement, affect the sets of L-prime solutions defined in the previous way. After this, in the following three sections, we consider prime solutions corresponding to some special languages: the equality set, the set of F-prime solutions, flat words and tense words. All of these imply results for the hierarchies of L-prime solutions. Finally, the last section is devoted to other definitions of L-prime solutions and their effects upon the results obtained earlier.

We have clearly $P_L(g,h) \subseteq S_L(g,h) \subseteq F_L(g,h)$.

The usual P-prime, S-prime and F-prime solutions are obtained by choosing $L = \Sigma^*$, while, in the other extreme, if L is empty or consists only of the empty word then

$$F_L(g,h) = S_L(g,h) = P_L(g,h) = E(g,h).$$

In what follows we usually prove the results only for F_L-prime solutions. Since any solution $w = w_1 w_2$ has the property that $w_1 \in E(g,h)$ if and only if $w_2 \in E(g,h)$, the definition for F_L-prime solutions can be written as

$$w \in F_L(g,h) \Leftrightarrow \nexists w_2 \in E(g,h) \cap L \text{ such that } w = w_1 w_2,\ w_1 \neq \lambda.$$

This observation simplifies the proofs though it is not true for S_L- and P_L-prime solutions.

Lemma 3.10 *If $L_1 \subseteq L_2$ then*

$$F_{L_2}(g,h) \subseteq F_{L_1}(g,h),\ S_{L_2}(g,h) \subseteq S_{L_1}(g,h),\ P_{L_2}(g,h) \subseteq P_{L_1}(g,h).$$

Proof: If $w \notin F_{L_1}(g,h)$ then $w = w_1 w_2$ for some $w_2 \in L_1 \cap E(g,h)$. Hence w_2 belongs to $L_2 \cap E(g,h)$, too, and $w \notin F_{L_2}(g,h)$. □

Lemma 3.10 together with the definition implies the following hierarchy for every language L (an inclusion is denoted by an arrow):

$$\begin{array}{ccccccc}
P_L(g,h) & \rightarrow & S_L(g,h) & \rightarrow & F_L(g,h) & \rightarrow & E(g,h) \\
\uparrow & & \uparrow & & \uparrow & & \\
P(g,h) & \rightarrow & S(g,h) & \rightarrow & F(g,h) & &
\end{array}$$

3.3.1 Some basic operations

We start the study of set-theoretic operations with union. The following result shows that the union of two languages is related to the intersection of the corresponding prime solutions.

Theorem 3.11 *For all languages L_1 and L_2,*

$$\begin{aligned}
F_{L_1 \cup L_2}(g,h) &= F_{L_1}(g,h) \cap F_{L_2}(g,h), \\
S_{L_1 \cup L_2}(g,h) &= S_{L_1}(g,h) \cap S_{L_2}(g,h), \\
P_{L_1 \cup L_2}(g,h) &= P_{L_1}(g,h) \cap P_{L_2}(g,h).
\end{aligned}$$

Proof: Assume that $w \notin F_{L_1 \cup L_2}(g,h)$, that is $w = w_1w_2$ for some $w_2 \in (L_1 \cup L_2) \cap E(g,h)$. This can be written as $w_2 \in L_1 \cap E(g,h) \cup L_2 \cap E(g,h)$. Hence $w \notin (F_{L_1}(g,h) \cap F_{L_2}(g,h))$.

For the other direction, assume that $w \notin (F_{L_1}(g,h) \cap F_{L_2}(g,h))$. Then $w = w_1w_2 = w_3w_4$ for some $w_2 \in L_1 \cap E(g,h)$ and $w_4 \in L_2 \cap E(g,h)$. Hence $w_2, w_4 \in (L_1 \cup L_2) \cap E(g,h)$ and $w \notin F_{L_1 \cup L_2}(g,h)$. □

Example 3.2 Let the instance (g,h) be defined by

	1	2
g	a^2b	a^2
h	a	$(ba^2)^2$

It is easy to see that $E(g,h) = \{112\}^+$ and $F(g,h) = \{112\}$. Assume now that $L_1 = \{112\}$ and $L_2 = \{112112\}$. Then $F_{L_1}(g,h) = \{112\}$, $F_{L_2}(g,h) = \{112, 112112\}$ and $F_{L_1 \cup L_2}(g,h) = \{112\} = F_{L_1}(g,h) \cap F_{L_2}(g,h)$, as intended.

□

The result is weaker for the intersection.

Theorem 3.12 *For all languages L_1 and L_2,*

$$\begin{aligned} F_{L_1}(g,h) \cup F_{L_2}(g,h) &\subseteq F_{L_1 \cap L_2}(g,h), \\ S_{L_1}(g,h) \cup S_{L_2}(g,h) &\subseteq S_{L_1 \cap L_2}(g,h), \\ P_{L_1}(g,h) \cup P_{L_2}(g,h) &\subseteq P_{L_1 \cap L_2}(g,h). \end{aligned}$$

Proof: Assume that $w \notin F_{L_1 \cap L_2}(g,h)$; that is, there exists $w_2 \in L_1 \cap L_2 \cap E(g,h)$ such that $w = w_1w_2$. But this implies that $w_2 \in L_1 \cap E(g,h) \cap L_2 \cap E(g,h)$ and $w \notin (F_{L_1}(g,h) \cup F_{L_2}(g,h))$. □

Reverse inclusions are not valid, at least not for the general case. Consider the word $w \notin (F_{L_1}(g,h) \cup F_{L_2}(g,h))$. By the definition there are words $w_2 \in L_1 \cap E(g,h)$ and $w_4 \in L_2 \cap E(g,h)$ such that $w = w_1w_2 = w_3w_4$. It is quite possible, however, that neither w_2 nor w_4 belongs to the set $L_1 \cap L_2 \cap E(g,h)$.

Example 3.3 Let the morphisms g and h as well as the languages L_1 and L_2 be the same as in Example 3.2. Then $L_1 \cap L_2 = \emptyset$ and, hence, $F_{L_1 \cap L_2}(g,h) = E(g,h)$ but $F_{L_1}(g,h) \cup F_{L_2}(g,h) = \{112, 112112\} \neq E(g,h)$. □

By adding some restrictions for L_1 and L_2, the result in Theorem 3.12 can be improved. For this we need the following lemma which is a straight consequence from the definition, as explained earlier, but is still not true for P_L- and S_L-prime solutions.

Lemma 3.13 $F_L(g,h) = F_{L\cap E(g,h)}(g,h)$.

Theorem 3.14 *If* $(L_1 \cup L_2) \cap E(g,h) = L_1 \cap L_2 \cap E(g,h)$ *then*

$$F_{L_1}(g,h) \cup F_{L_2}(g,h) = F_{L_1 \cap L_2}(g,h).$$

For S_L- and P_L-prime solutions we need a stronger restriction. The example below shows why the assumption in Theorem 3.14 is not enough.

Theorem 3.15 *If* $L_1 \cap L_2 = L_1$ *(or* L_2*) then*

$$\begin{aligned} S_{L_1\cap L_2}(g,h) &= S_{L_1}(g,h) \cup S_{L_2}(g,h), \\ P_{L_1\cap L_2}(g,h) &= P_{L_1}(g,h) \cup P_{L_2}(g,h). \end{aligned}$$

Proof: If $w \notin (S_{L_1}(g,h) \cup S_{L_2}(g,h))$ then $w = v_1x_1v_2 = v_3x_2v_4$ for some $x_1 \in L_1$, $x_2 \in L_2$ and $v_1v_2, v_3v_4 \in E(g,h)$. The assumption now implies that x_1 or x_2 belongs to $L_1 \cap L_2$ and, hence, $w \notin S_{L_1\cap L_2}(g,h)$. The same is true for P_L-prime solutions. □

Example 3.4 Consider the instance (g,h) defined by

	1	2	3
g	ab	cb	d
h	a	bc	bd

We have $E(g,h) = \{12^*3\}^+$, $F(g,h) = \{12^*3\}$ and $P(g,h) = S(g,h) = \{13\}$. If $L_1 = \{2, 13\}$ and $L_2 = \{22, 13\}$ then $(L_1 \cup L_2) \cap E(g,h) = \{13\} = L_1 \cap L_2 \cap E(g,h)$. So the condition in Theorem 3.14 is satisfied. Even so $S_{L_1}(g,h) \cup S_{L_2}(g,h) = \{13, 123\}$ and $S_{L_1\cap L_2}(g,h) = \{13\} \cup \{12^+3\}^+$. □

We now study the effect of complementation. The complement of L is denoted by L^c.

Theorem 3.16 *The conditions* $F_L(g,h) = E(g,h)$*,* $S_L(g,h) = E(g,h)$ *and* $P_L(g,h) = E(g,h)$ *imply that* $F_{L^c}(g,h) = F(g,h)$*,* $S_{L^c}(g,h) = S(g,h)$ *and* $P_{L^c}(g,h) = P(g,h)$*, respectively.*

Proof: Assume that $F_L(g,h) = E(g,h)$ and let $w \in E(g,h) - F(g,h)$. Since w is an F_L-prime solution but not an F-prime solution, we have $w = w_1 w_2$ for some $w_2 \notin L$. But this means that $w_2 \in E(g,h) \cap L^c$ and $w \notin F_{L^c}(g,h)$. Hence $F_{L^c}(g,h) \subseteq F(g,h)$. The converse is obvious. □

The conditions in the previous theorem are only sufficient. If we assume that $F_L(g,h) = F(g,h)$ it is still quite possible that $F_{L^c}(g,h)$ is a proper subset of $E(g,h)$.

Example 3.5 Let the instance (g,h) and the language L_1 be the same as in Example 3.2. Then $F_{L_1}(g,h) = F(g,h)$ but $F_{L_1^c}(g,h) = \{112, 112112\}$. Indeed, no word $(112)^i 112112$, $i \geq 1$, belongs to $F_{L_1^c}(g,h)$ because $112112 \in E(g,h) - L_1$. □

If we assume that the intersection $L \cap E(g,h)$ is empty then the conditions are also necessary.

3.3.2 Catenation

By Lemma 3.10, the condition $L_1 \subseteq L_1 L_2$ implies that $F_{L_1 L_2}(g,h) \subseteq F_{L_1}(g,h)$. This happens, for instance, if $\lambda \in L_2$. It turns out that this is the only general result for catenation. Various possibilities concerning how the set $F_{L_1 L_2}(g,h)$ is related to the sets $F_{L_1}(g,h)$ and $F_{L_2}(g,h)$ are presented in the next example.

Example 3.6 Let $L_1 = \{12\}$, $L_2 = \{121\}$ and consider the instances (g_1, h_1), (g_2, h_2) and (g_3, h_3) defined by

	1	2
g_1	ab	c
h_1	a	bc

	1	2
g_2	a	$b(ab)^2$
h_2	aba	b

	1	2
g_3	a	$b(ab)^3$
h_3	$(ab)^2a$	b

Then

$$E(g_1,h_1) = \{12\}^+ = F_{L_2}(g_1,h_1) = F_{L_1L_2}(g_1,h_1);\ F_{L_1}(g_1,h_1) = \{12\},$$
$$E(g_2,h_2) = \{121\}^+ = F_{L_1}(g_2,h_2) = F_{L_1L_2}(g_2,h_2);\ F_{L_2}(g_2,h_2) = \{121\},$$
$$E(g_3,h_3) = \{12121\}^+ = F_{L_1}(g_3,h_3) = F_{L_2}(g_3,h_3);\ F_{L_1L_2}(g_3,h_3) = \{12121\}.$$

□

Consider now the operations of Kleene star and Kleene plus. The next theorem shows that there is actually no difference between the corresponding L-prime solutions.

Theorem 3.17 *For all languages L,*

$$\begin{aligned} F_{L^*}(g,h) &= \textstyle\bigcap_{i=0}^{\infty} F_{L^i}(g,h) = \bigcap_{i=1}^{\infty} F_{L^i}(g,h) = F_{L^+}(g,h), \\ S_{L^*}(g,h) &= \textstyle\bigcap_{i=0}^{\infty} S_{L^i}(g,h) = \bigcap_{i=1}^{\infty} S_{L^i}(g,h) = S_{L^+}(g,h), \\ P_{L^*}(g,h) &= \textstyle\bigcap_{i=0}^{\infty} P_{L^i}(g,h) = \bigcap_{i=1}^{\infty} P_{L^i}(g,h) = P_{L^+}(g,h). \end{aligned}$$

Proof: By Theorem 3.11 and induction, the first and last equalities are obvious. On the other hand, $F_{L^0}(g,h) = E(g,h)$, so it does not affect the intersection. □

Example 3.7 Let the instance (g,h) be the same as (g_1,h_1) in the previous example.

We have $E(g,h) = \{12\}^+$ and $F(g,h) = \{12\}$. If $L = \{1212\}$ then $L^i = \{(12)^{2i}\}$ and

$$\begin{aligned} F_L(g,h) &= \{12, 1212\}, \\ F_{L^2}(g,h) &= \{12, 1212, 121212, 12121212\}, \\ F_{L^i}(g,h) &= \{(12)^s \mid 1 \le s \le 2i\}, \\ F_{L^*}(g,h) &= \{12, 1212\}. \end{aligned}$$

□

Contrary to the previous example $F_{L^*}(g,h)$ may also be infinite. This is clearly the case if $F(g,h)$ is infinite. The following lemma shows another condition.

Lemma 3.18 *$F_L(g,h) = E(g,h)$ if and only if $L \cap E(g,h) = \emptyset$.*

Proof: By Lemma 3.13 the "if"-part is obvious. For the "only if"-part we assume the contrary: there is a word w in the set $L \cap E(g,h)$. Then, by definition, w^n is not an F_L-prime solution and $F_L(g,h) \neq E(g,h)$, a contradiction. □

By the above argument, the conditions $S_L(g,h) = E(g,h)$ and $P_L(g,h) = E(g,h)$ imply that $L \cap E(g,h) = \emptyset$. However, the converse does not hold. If the instance (g,h) is the same as in Example 3.4 and we choose $L = \{2\}$ then $L \cap E(g,h) = \emptyset$ but $S_L(g,h) = P_L(g,h) = \{13\}$.

3.3.3 Prime solutions restricted to $E(g,h)$ and $F(g,h)$

In this section we present some results and a hierarchy when the language L is the equality set of the given instance (g,h) or the set of its F-prime solutions. For simplicity, the sets $E(g,h)$ and $F(g,h)$ are denoted by E and F when used as subscripts.

Theorem 3.19 *For all instances (g,h),*

$$F_F(g,h) = F_E(g,h) = F(g,h).$$

Proof: By the definition and Lemma 3.13, we have $F(g,h) = F_E(g,h)$ and $F_E(g,h) \subseteq F_F(g,h)$. Assume now that $w \in E(g,h) - F(g,h)$, hence $w = w_1w_2$ for some $w_1, w_2 \in E(g,h)$. If $w_2 \notin F(g,h)$ then $w_2 = w_3w_4$ for some $w_3, w_4 \in E(g,h)$. In finitely many steps we find a suffix w' of w which is an F-prime solution and the remainder, being a catenation of solutions, is a solution. But this means that $w \notin F_F(g,h)$. □

Lemma 3.20 *For all instances (g,h),*

$$P_F(g,h) \subseteq S_E(g,h).$$

Proof: Assume that a word w belongs to the set $E(g,h) - S_E(g,h)$. Then $w = uw_1v$ for some $uv, w_1 \in E(g,h)$. To prove that w is not a P_F-prime solution we have two cases to consider. If uv is an F-prime solution then the claim follows directly from the definition. On the other hand, if $uv \notin F(g,h)$ then $uv = xy$ for some $x, y \in E(g,h)$. But this means that w begins with x or ends with y, and w is not even an F-prime solution which by Theorem 3.19 implies that $w \notin P_F(g,h)$. □

We are now ready to present the hierarchy.

Theorem 3.21 *For all instances (g,h),*

$$\begin{aligned} P(g,h) \subseteq P_E(g,h) \subseteq P_F(g,h) \subseteq S_E(g,h) \subseteq S_F(g,h) \\ \subseteq F_E(g,h) = F_F(g,h) = F(g,h). \end{aligned}$$

The following five examples show that the five inclusions in Theorem 3.21 can be strict.

Example 3.8 Let the instance (g, h) be the same as in Example 3.4. Then $P(g, h) = S(g, h) = \{13\}$ but $P_E(g, h) = \{12^*3\} = S_E(g, h)$. □

Example 3.9 Let the morphisms g and h be defined by

	1	2	3	4	5	6	7	8
g	a	bec	$bbecb$	b	bbc	d	cd	f
h	ab	ec	b	$becbb$	b	bcd	c	df

Then the words $12, 34, 56, 78$ and 13254768 are F-prime solutions. The word 13254768 is not a P_E-prime solution since its scattered subword 1256 and the remainder 3478 are solutions. They are not, however, F-prime solutions, so the word 13254768 is a P_F-prime solution. □

Example 3.10 Let the morphisms g and h be defined by

	1	2	3	4
g	a	bc	bbc	d
h	ab	c	b	bcd

Then the word 1324 is not a P_F-prime solution since 12 and 34 are F-prime solutions. It is, however, an S-prime solution and, hence, an S_E-prime solution. □

Example 3.11 Let the morphisms g and h be defined by

	1	2	3	4	5	6
g	$aabcdef$	a	bc	d	ef	q
h	a	ab	c	de	f	$abcdefq$

We have $E(g, h) = \{16, 23, 45, 123456\}^+$, $F(g, h) = \{16, 23, 45, 123456\}$ and $P(g, h) = S(g, h) = \{16, 23, 45\} = P_E(g, h) = S_E(g, h)$. Now $123456 \in S_F(g, h)$ since 2345 is not an F-prime solution and removing any other subword that belongs to $F(g, h)$ does not lead to a solution. So $S_F(g, h) \neq S_E(g, h)$. However, $P_E(g, h) = P_F(g, h)$ since by choosing $u_1 = 1$, $u_2 = 6$ and $v_1 = 2345$, we have $123456 = u_1 v_1 u_2$ where $u_1 u_2 \in F(g, h)$ and $v_1 \in E(g, h)$.

□

Example 3.12 Let the morphisms g and h be defined by

	1	2	3	4
g	$aabc$	ab	c	d
h	a	a	bc	$abcd$

The word 1234 is an F-prime solution but not an S_F-prime solution since 14 and 23 are solutions. □

Examples 3.8 and 3.10 show that the set $S(g,h)$ is incomparable with the sets $P_E(g,h)$ and $P_F(g,h)$ and is therefore left out from the hierarchy. An interesting observation is also that, in contrary to the usual case, the set of P_E-prime solutions can be infinite.

3.3.4 Flat prime solutions

We start the study of flat prime solutions by defining flatwords and k-flatwords. For motivations and more details, see [10].

We say that a word w is a *k-flatword* if and only if all its subwords of length r, $r \leq k+1$, contain each letter at most once. More specifically, assume that $\Sigma = \{a_1, a_2, \ldots, a_n\}$ and let $k \geq 1$ be a fixed number. A word $w \in \Sigma^*$ is a *k-flatword* if and only if every division $w = uw'v$ where $|w'| \leq k+1$ implies that $|w'|_{a_i} \leq 1$ for $i = 1, \ldots, n$.

Example 3.13 Let $w_1 = abcbcad$ and $w_2 = cdabcad$ be the words over the alphabet $\Sigma = \{a,b,c,d\}$. Then w_1 is a 1-flatword but not a 2-flatword (the subword $w' = bcb$ has $|w'|_b = 2$) whereas the word w_2 is a 2-flatword but not a 3-flatword ($w' = abca$ has $|w'|_a = 2$). □

So k is actually the minimal distance between two letters of the same type. If $k = 1$ we just talk about flatwords instead of 1-flatwords. Hence, a word w is a flatword if and only if it does not contain square letters.

The set of all k-flatwords over an alphabet Σ is denoted by $Fl_k(\Sigma)$. Obviously, if $k \leq k'$, then $Fl_{k'}(\Sigma) \subseteq Fl_k(\Sigma)$.

It also follows from the definition that for any k, $k \geq n = \mathrm{card}\,(\Sigma)$, we have $Fl_k(\Sigma) = Fl_n(\Sigma)$. For instance, if $\Sigma = \{a,b\}$ then

$$Fl_2(\Sigma) = \{\lambda, a, b, ab, ba\} = Fl_3(\Sigma).$$

We are now ready to define flat prime solutions. A solution $w \in E(g,h)$ is *$F_{k-\text{flat}}$-prime* if and only if no k-flat suffix of w, that is, a suffix which is a k-flatword, can be removed and still obtain a solution. In the same way we obtain *$S_{k-\text{flat}}$-prime* and *$P_{k-\text{flat}}$-prime solutions*. Hence in the definition of L-prime solutions the language L is now exactly the set of k-flatwords, $Fl_k(\Sigma)$.

The sets of 1-flat prime solutions are denoted briefly by $F_{\text{flat}}(g,h)$, $S_{\text{flat}}(g,h)$ and $P_{\text{flat}}(g,h)$.

The following hierarchy of k-flat prime solutions follows directly from the definition of k-flatwords and Lemma 3.10.

Theorem 3.22 *In every alphabet* Σ, $\operatorname{card}(\Sigma) = n$, *we have:*

$$\begin{tikzcd}
P_{n-\text{flat}}(g,h) \arrow[r] & S_{n-\text{flat}}(g,h) \arrow[r] & F_{n-\text{flat}}(g,h) \arrow[r] & E(g,h) \\
\vdots \arrow[u] & \vdots \arrow[u] & \vdots \arrow[u] & \\
P_{2-\text{flat}}(g,h) \arrow[r] \arrow[u] & S_{2-\text{flat}}(g,h) \arrow[r] \arrow[u] & F_{2-\text{flat}}(g,h) \arrow[u] & \\
P_{\text{flat}}(g,h) \arrow[r] \arrow[u] & S_{\text{flat}}(g,h) \arrow[r] \arrow[u] & F_{\text{flat}}(g,h) \arrow[u] & \\
P(g,h) \arrow[r] \arrow[u] & S(g,h) \arrow[r] \arrow[u] & F(g,h) \arrow[u] &
\end{tikzcd}$$

Consider now the hierarchy of corresponding words. We prove that it is the same as in Theorem 3.22 except that all the inclusions are proper (which is by Theorems 2.26 and 2.20 already known for P, S and F) and that no other inclusions exist; that is, the sets $P_{k-\text{flat}}$ and $S_{(k-1)-\text{flat}}$ as well as the sets $S_{k-\text{flat}}$ and $F_{(k-1)-\text{flat}}$ are incomparable.

Example 3.14 Consider

	1	2
g_1	ab	c
h_1	a	bc

	1	2
g_2	a^2b	a^2
h_2	a	$(ba^2)^2$

	1	2
g_3	a^3b	a^3
h_3	a	$(ba^3)^3$

Clearly, $E(g_1,h_1) = (12)^+$. Consider the word $(12)^n$, $n \geq 2$. It cannot be an $F_{k-\text{flat}}$-solution because 12 is a k-flatword and $(12)^{n-1}$ is a solution. Thus none of the words $(12)^n$, $n \geq 2$, is $F_{k-\text{flat}}$-prime and

$$\begin{aligned}
\{12\} &= F_{k-\text{flat}}(g_1,h_1) = S_{k-\text{flat}}(g_1,h_1) = P_{k-\text{flat}}(g_1,h_1) \\
&= F(g_1,h_1) = S(g_1,h_1) = P(g_1,h_1).
\end{aligned}$$

For the second instance $E(g_2,h_2) = (112)^+$. A word $(112)^n$ is an $F_{k-\text{flat}}$-solution for all n because no k-flat suffix can be removed. On the other hand,

$$(112)^n = 1\underbrace{121}12(112)^{n-2}$$

is not S_{flat}-prime because 121 is a flatword (but not a k-flatword, $k \geq 2$) and $(112)^{n-1}$ is a solution. Thus for $k \geq 2$,

$$F_{k-\text{flat}}(g_2, h_2) = S_{k-\text{flat}}(g_2, h_2) = P_{k-\text{flat}}(g_2, h_2) = F_{\text{flat}}(g_2, h_2) = (112)^+,$$
$$S_{\text{flat}}(g_2, h_2) = P_{\text{flat}}(g_2, h_2) = F(g_2, h_2) = S(g_2, h_2) = P(g_2, h_2) = \{112\}.$$

For the last instance the words $(1112)^n$ are all P_{flat}-primes because none of the flat scattered subwords can be removed. Therefore,

$$P_{k-\text{flat}}(g_3, h_3) = S_{k-\text{flat}}(g_3, h_3) = F_{k-\text{flat}}(g_3, h_3) = (1112)^+,$$
$$P(g_3, h_3) = S(g_3, h_3) = F(g_3, h_3) = \{1112\}.$$

□

Theorem 3.23 *The word w^n, $n \geq 2$, is $F_{k-\text{flat}}$-prime if and only if w is not a k-flatword and w is $F_{k-\text{flat}}$-prime.*

Proof: Clearly, w^n is a solution if and only if w is a solution. Thus if a k-flat suffix of w can be removed, it can be done for w^n too. On the other hand, if w^n has a k-flat suffix u and $|u| > |w|$ then also w is a k-flatword. □

Theorem 3.24 *If $w = a_1 \ldots a_k a_{k+1}^2$, where $a_i \in \Sigma$ and $a_1 \ldots a_{k+1}$ is a k-flatword, then w^n is not $S_{k-\text{flat}}$-prime, for $n \geq 2$.*

Proof: The technique is similar as in Example 3.14. Consider

$$w^2 = a_1 \ldots a_k a_{k+1} \underbrace{a_{k+1} a_1 \ldots a_k a_{k+1}}_{\text{can be removed}} a_{k+1}.$$

□

Theorems 3.23 and 3.24 imply that there exists an $F_{k-\text{flat}}$-word that is not an $S_{k-\text{flat}}$-word. The same is true for $S_{k-\text{flat}}$-words and $P_{k-\text{flat}}$-words.

Theorem 3.25 *For each k there exists an $S_{k-\text{flat}}$-word that is not a $P_{k-\text{flat}}$-word.*

Proof: Let n be the smallest number divisible by 3 and greater than k. Consider the following word $w = avava$ of the same form as in Theorem 2.28. Let $i_l = 1$ for $l = 1, \ldots, n$, and $j_s = 2r + 1$ for $s = 2, \ldots, n$, where $n = 3r$. Then

$$\psi(w) = (6r + 3, 2(2r + 1), \ldots, 2(2r + 1)) = (2r + 1)(3, 2, \ldots, 2).$$

If we now remove the scattered subword $u = ab_2b_3 \ldots b_n ab_2b_3 \ldots b_n a$, which is an $(n-1)$-flatword and thus a k-flatword, then the remaining part is also a solution and, hence, w is not a $P_{k-\text{flat}}$-solution. It is, however, an $S_{k-\text{flat}}$-word because it is an S-word by Theorem 2.9. $\square$

To illustrate the previous proof we consider an example.

Example 3.15 Assume that $n = 3$. Hence $r = 1$ and $j_2 = j_3 = 2r + 1 = 3$. Consider the word $w = avava$ where $v = ab_2^3ab_3^3a$, i.e.,

$$w = a^2b_2^3ab_3^3a^3b_2^3ab_3^3a^2.$$

Since its scattered subword $u = ab_2b_3ab_2b_3a$ is a 2-flatword and the remaining part $ab_2^2ab_3^2a^2b_2^2ab_3^2a$ is, by Theorem 2.28, a solution, w cannot be $P_{2-\text{flat}}$-prime or P_{flat}-prime.

The word u is not, however, a 3-flatword, so it is not sufficient to choose $n = k$ in the proof of Theorem 3.25. $\square$

So far we have shown that the horizontal inclusions in our hierarchy are all proper. The following result now proves the vertical cases.

Theorem 3.26 *In every alphabet Σ,* card$(\Sigma) = n$*, there exists a word w that is a $P_{n-\text{flat}}$-word but not an $F_{(n-1)-\text{flat}}$-word.*

Proof: Let $\Sigma = \{a_1, \ldots, a_n\}$. The word $w = (a_1a_2 \ldots a_na_1)^2$ is not an $F_{(n-1)-\text{flat}}$-solution because $a_1a_2 \ldots a_na_1$ is an $(n-1)$-flatword; nevertheless, since no n-flat scattered subwords can be removed, w is a $P_{n-\text{flat}}$-solution for the instance

	a_1	a_2	a_3	a_4	$\ldots$	a_{n-1}	a_n
g	$b_1b_2b_1$	b_3	b_4	b_5	$\ldots$	b_n	b_{n+1}
h	b_1	b_2b_1	b_3	b_4	$\ldots$	b_{n-1}	$b_nb_{n+1}b_1b_2$

where $\Delta = \{b_1, \ldots, b_{n+1}\}$. □

Next we prove that no diagonal dependencies exist.

Theorem 3.27 *The sets $P_{k-\text{flat}}$ and $S_{(k-1)-\text{flat}}$ are incomparable.*

Proof: By Theorem 3.26 there are words in $P_{k-\text{flat}}$ that are not in $S_{(k-1)-\text{flat}}$. On the other hand, the word w in the proof of Theorem 3.25 is an S-word and thus an $S_{k-\text{flat}}$-word for all k but not a $P_{k-\text{flat}}$-word. □

Theorem 3.28 *The sets $S_{k-\text{flat}}$ and $F_{(k-1)-\text{flat}}$ are incomparable.*

Proof: By the proof of Theorem 3.26 there are words in $S_{k-\text{flat}}$ that are not in $F_{(k-1)-\text{flat}}$. On the other hand, the word $w = (a_1 \ldots a_k a_{k+1}^2)^2$ in Theorem 3.24 is not an $S_{k-\text{flat}}$-word but it is an $F_{k-\text{flat}}$-solution for the instance

	a_1	a_2	a_3	$\ldots$	a_{n-1}	a_n	a_{n+1}
g	b_1b_2	b_3	b_4	$\ldots$	b_n	b_{n+1}^2	b_{n+1}
h	b_1	b_2	b_3	$\ldots$	b_{n-1}	b_n	b_{n+1}^2

where $\Delta = \{b_1, \ldots, b_{n+1}\}$.

The relation between F-words and S_{flat}-words still remains to be settled. By Lemma 2.2, we know that no word of the form w^i, $i \geq 2$, is an F-word. Hence the word $(1112)^2$ in Example 3.14 is an S_{flat}-word but not an F-word. For the other direction, consider the word $w = 112122$ which is, by Lemma 2.16, periodicity forcing. This means that if $112122 \in E(g,h)$ then also $12 \in E(g,h)$ and w cannot be an S_{flat}-word. It is, however, an F-word because of Theorem 2.8. □

The following two results show some interesting properties concerning the hierarchy of Theorem 3.22.

Theorem 3.29 *If $|F_{k-\text{flat}}(g,h)| < \infty$ then $F_{k-\text{flat}}(g,h) = S_{k-\text{flat}}(g,h) = F(g,h) = S(g,h)$.*

Proof: If $|F_{k-\text{flat}}(g,h)| < \infty$ then $f < \infty$ and $s = f = n$ by Theorem 3.22 and Theorem 2.1. This implies that $E(g,h) = \{w_1, \ldots, w_n\}^+$. By Theorem 3.23 each word w_i must be a k-flatword; hence, $|F_{k-\text{flat}}(g,h)| = n$ and $F_{k-\text{flat}}(g,h) = S_{k-\text{flat}}(g,h)$. □

Theorem 3.30 *If $f < \infty$ then either $S_{k-\text{flat}}(g,h) = F(g,h) = S(g,h)$, or $|S_{k-\text{flat}}(g,h)| = \infty$.*

Proof: In the same way as in the previous proof, $f < \infty$ implies that $s = f = n$ and $E(g,h) = \{w_1, \ldots, w_n\}^+$. Assume now that $|S_{k-\text{flat}}(g,h)| < \infty$. If there exists an $S_{k-\text{flat}}$-solution that is not an F-prime solution, it must be of the form $w_i w_j$ or w^i for some $w_i, w_j, w \in E(g,h)$. Consider first the case $w_i w_j$. If $w_i w_j$ is an $S_{k-\text{flat}}$-solution but $(w_i w_j)^2$ is not, then $(w_i w_j)^2 = v_1 x v_2$ where $v_1 v_2$ is a solution and x is a k-flatword.

By Theorem 3.23, the words w_i, w_j cannot be k-flatwords; hence, $w_i w_j \in sub\,(v_1 v_2)$. But this means that $v_1 v_2$ cannot be an F-prime solution w nor a catenation of F-prime solutions w since in both cases either w_i is a subword of w or vice versa, which is not possible since $S(g,h) = F(g,h)$.

The same argument is true for $S_{k-\text{flat}}$-solutions of the type w^i. Therefore, if there exists an $S_{k-\text{flat}}$-solution that is not an F-prime solution, then $|S_{k-\text{flat}}(g,h)| = \infty$. □

We end this section with an algorithm that characterizes $F_{k-\text{flat}}$-words.

Theorem 3.31 *The property of being an $F_{k-\text{flat}}$-word is decidable.*

Proof: Let $w \in \Sigma^+$. The following algorithm shows whether w is an $F_{k-\text{flat}}$-word or not.

Step I: Find the shortest word u for which $w = u^n$.

Step II: If $n > 1$ and u is a k-flatword then w is not an $F_{k-\text{flat}}$-word by Theorem 3.23.

Step III: If u is an F-word (as in Theorem 2.8) then w is an $F_{k-\text{flat}}$-word by Theorem 3.23.

Step IV: Check for each division $u = u_1 u_2$, where $\psi(u_1) = r\psi(u)$ for some $r < 1$, whether u_2 is a k-flatword. If this is not true for any u_2 then w is an $F_{k-\text{flat}}$-word; otherwise, w is not an $F_{k-\text{flat}}$-word.

The morphisms g and h for which $w \in F_{k-\text{flat}}(g,h)$ in Steps III and IV can be effectively constructed by the methods of Lemma 2.7. □

3.3.5 Tense prime solutions

Tense words are in a way opposite to flatwords. For a k-flatword each subword of certain size should contain no letter more than once, while the idea in tense words is to have all letters of the alphabet in it.

The following definition is due to prof. Tero Harju. Let $k \in \mathbb{N}$. A word $w \in \Sigma^*$ is *k-tense*, if each subword of w of length k contains all letters of Σ, i.e., each division $w = w_1w_2w_3$ where $|w_2| \geq k$ implies that $|w_2|_a \geq 1$ for all $a \in \Sigma$. Words shorter than k over the same alphabet are k-tense, if they are l-tense for some $l < k$.

Let T_k^n be the set of all k-tense words over Σ, $\text{card}\,(\Sigma) = n$. The following properties are obvious.

Lemma 3.32 *Let* $\text{card}\,(\Sigma) = n$. *Then* $T_k^n \subseteq T_{k+1}^n$ *and for* $k < n$, $T_k^n = \emptyset$.

Lemma 3.33 *If* $k \geq \max(n, m)$ *then the sets* T_k^n *and* T_k^m *are incomparable.*

Example 3.16 In a binary alphabet $\Sigma = \{a, b\}$ the set of 2-tense words is

$$T_2^2 = (ab)^+ \cup (ab)^+a \cup (ba)^+ \cup (ba)^+b.$$

Indeed, any other word over Σ contains a subword aa or bb and, hence, cannot be 2-tense.
□

In what follows we usually associate each word w with its minimal alphabet Σ and minimal tenseness, that is, the smallest integer k for which $w \in T_k^n$. In this sense we may lay an upper bound for the tenseness in spite of Lemma 3.32. We also talk about words with biggest or lowest tenseness and mean the integers corresponding to the minimal tenseness.

Our first two results show how the catenation affects the tenseness.

Lemma 3.34 *Let* $u, v \in \Sigma$, $\text{card}\,(\Sigma) = n$. *If u is k_1-tense and v k_2-tense then uv is k-tense where* $\max(k_1, k_2) \leq k \leq k_1 + k_2 - 1$.

Proof: The lower bound follows from the fact that the prefix of uv is still k_1-tense and the suffix of it k_2-tense. Hence uv is at least $\max(k_1, k_2)$-tense.

Consider now the upper bound. If the suffix of u is k_1-tense and the prefix of v is k_2-tense containing the same letters in reverse order then uv is $(k_1 + k_2 - 1)$-tense. This is seen from the following example. If $u = abcc$ and $v = cccba$ then $uv = abcccccba$. So u is 4-tense, v is 5-tense and uv is 8-tense. The idea is to make the distance between two a's as long as possible but this is exactly $k_1 + k_2 - 1$.
□

In spite of the previous lemma the tenseness of uwv may be smaller than that of uv. So adding a subword does not have the same effect as in Lemma 3.34.

Example 3.17 Let $u = 2211$, $v = 1122$ and $w = 12$. Then $uv = 22111122$ is 5-tense but $uwv = 2211121122$ is 4-tense. □

The next result follows from the fact that adding the same word the second time does not affect the tenseness.

Lemma 3.35 *If w^2 is k-tense then also w^i, $i \geq 2$, are k-tense.*

We are now ready to define the primitive solutions associated to tense words. In the same way as in the previous sections, a word $w \in E(g,h)$ is $F_{k-\text{tense}}$-*prime* if no proper k-tense suffix can be removed such that the remainder is also a solution. Hence the language L is now T_k^n.

Each k-tense prime solution is associated with its minimal alphabet and minimal instance (g,h). The sets of $F_{k-\text{tense}}$-, $S_{k-\text{tense}}$- and $P_{k-\text{tense}}$-prime solutions over an n-letter alphabet are denoted by $F_k^n(g,h)$, $S_k^n(g,h)$ and $P_k^n(g,h)$, respectively.

It follows directly from Lemmas 3.10 and 3.32 that

$$F_{k+1}^n(g,h) \subseteq F_k^n(g,h),\ S_{k+1}^n(g,h) \subseteq S_k^n(g,h),\ P_{k+1}^n(g,h) \subseteq P_k^n(g,h),$$

and if $k < n$ then $P_k^n(g,h) = S_k^n(g,h) = F_k^n(g,h) = E(g,h)$.

We have thus obtained the hierarchies

$$\begin{aligned}
F(g,h) &\subseteq \ldots \subseteq F_{n+2}^n(g,h) \subseteq F_{n+1}^n(g,h) \subseteq F_n^n(g,h) \subseteq E(g,h),\\
S(g,h) &\subseteq \ldots \subseteq S_{n+2}^n(g,h) \subseteq S_{n+1}^n(g,h) \subseteq S_n^n(g,h) \subseteq E(g,h),\\
P(g,h) &\subseteq \ldots \subseteq P_{n+2}^n(g,h) \subseteq P_{n+1}^n(g,h) \subseteq P_n^n(g,h) \subseteq E(g,h),
\end{aligned}$$

which we call an F-, S- and a P-hierarchy, respectively. By definition, we have also $P_k^n(g,h) \subseteq S_k^n(g,h) \subseteq F_k^n(g,h)$ for all k, n.

We will now study these hierarchies closer. First we notice that in alphabets with at least three letters the instances may contain solutions over alphabets of different sizes which by Lemma 3.33 gives difficulties in comparing their tenseness. This is why we shall consider only nontrivial instances in a binary alphabet. We start with the F-hierarchy and consider what it is like in different instances.

Example 3.18 Let the instances (g_1, h_1) and (g_2, h_2) be defined by

	1	2
g_1	a^2b	a^2
h_1	a	$(ba^2)^2$

	1	2
g_2	a^3b	a^3
h_2	a	$(ba^3)^3$

It is easy to see that

$$\begin{aligned}
E(g_1, h_1) &= \{112\}^+ = F_2^2(g_1, h_1) \text{ since } 112 \text{ is not } 2 - \text{tense},\\
F(g_1, h_1) &= \{112\} = \ldots = F_3^2(g_1, h_1),\\
E(g_2, h_2) &= \{1112\}^+ = F_2^2(g_2, h_2) = F_3^2(g_2, h_2),\\
F(g_2, h_2) &= \{1112\} = \ldots = F_4^2(g_2, h_2).
\end{aligned}$$

□

Theorem 3.36 *If g or h is injective in a binary alphabet then the corresponding F-hierarchy is finite.*

Proof: By Lemmas 2.21 and 2.22 we have three possibilities to consider.

Case I: $E(g, h) = \{u\}^+$

Let k be the minimal number for which $u \in T_k^2$. Then uu is not an F_k^2-prime solution but it is F_{k-1}^2-prime. So

$$\begin{aligned}
F_k^2(g, h) &= F_{k+1}^2(g, h) = \ldots = \{u\} = F(g, h),\\
F_{k-1}^2(g, h) &= \ldots = F_2^2(g, h) = \{u\}^+ = E(g, h).
\end{aligned}$$

Case II: $E(g, h) = \{u, v\}^+$

Let k_1, k_2 be the minimal numbers for which $u \in T_{k_1}^2$ and $v \in T_{k_2}^2$. Without loss of generality we may assume that $k_1 \le k_2$. Then $u, v \in T_{k_2}^2$ and $u, v \notin T_{k_1-1}^2$. But this means that

$$\begin{aligned}
F_{k_1-1}^2(g, h) &= \{u, v\}^+ = E(g, h),\\
F_{k_2}^2(g, h) &= \{u, v\} = F(g, h).
\end{aligned}$$

Case III: $E(g, h) = \{uw^*v\}^+$

By definition and Lemma 3.35 there exist minimal numbers k_1, k_2, k_3 such that $uv \in T_{k_1}^2$, $uwv \in T_{k_2}^2$ and $uw^iv \in T_{k_3}^2$, $i \ge 2$. Denote the biggest number of k_i's by k_b and the smallest one by k_s. Then $uv^iw \in T^2_{k_b}$, $uw^iv \notin T^2_{k_s-1}$ for all $i \geq 0$ and

$$F^2_{k_b}(g,h) = \{uw^*v\} = F(g,h),$$
$$F^2_{k_s-1}(g,h) = \{uw^*v\}^+ = E(g,h).$$

This completes the proof of Case III and at the same time the proof of the theorem. □

In periodic instances the situation is different.

Theorem 3.37 *If g and h are both periodic in a binary alphabet then the corresponding F-hierarchy is infinite.*

Proof: We start with the instance (g,h) defined by

	1	2
g	a^2	a
h	a	a^2

Then $E(g,h) = \{w \mid \psi(w) = k(1,1),\ k \geq 1\}$, and
$F(g,h) = \{w \in E(g,h) \mid \nexists w_1 \in E(g,h)$ such that $w = w_1w_2$ and $w_2 \neq \lambda\}$. Thus none of the words w^2 are F-prime solutions. Considering the k-tense solutions, we have

$$F^2_2(g,h) \neq E(g,h) \text{ because } (12)^2 \notin F^2_2(g,h),$$
$$F^2_3(g,h) \neq F^2_2(g,h) \text{ because } (1122)^2 \in F^2_2(g,h) - F^2_3(g,h)$$

and, in general, $F^2_k(g,h) \neq F^2_{k-1}(g,h)$ by the word $(1^{k-1}2^{k-1})^2$.

So the hierarchy $F(g,h) \subset \ldots \subset F^2_3(g,h) \subset F^2_2(g,h) \subset E(g,h)$ is proper and infinite.

In general, we consider the periodic instances where the equality set consists of solutions of the form $k(M,N)$ for some relatively prime integers M,N. Without loss of generality, we may assume that $M \leq N$. Then for each k, the set $F^2_{kN+1}(g,h)$ differs form $F^2_{kN}(g,h)$ by the word $(1^{kM}2^{kN})^2$ and the hierarchy is infinite. □

We have thus characterized the F-hierarchy completely.

Consider now the S-hierarchy.

Theorem 3.38 *In a binary alphabet the S-hierarchy is finite.*

Proof: If g or h is injective then, by Theorem 3.36, the F-hierarchy is finite and, hence, also the S-hierarchy is finite.

Assume now that g and h are both periodic. Then $E(g,h) = \{w \mid \psi(w) = k(M,N),\ k \geq 1\}$ for some relatively prime integers $M, N \geq 1$, and $S(g,h) = \{w \mid \psi(w) = (M,N)\}$. By Lemma 2.3, for each $w \in E(g,h)$ there exists a subword $w' \in S(g,h)$. So it suffices to study the tenseness of the S-prime solutions.

The word with biggest tense, denoted by k_1, is clearly $1^M 2^N$, which is $(\max(M,N)+1)$-tense. The lowest tense (denoted by k_0) appears when 1's and 2's are evenly divided in w. For instance, if $M = N-1$ then the lowest tense is 2 and if $M = 1$ the lowest tense is either $\frac{N}{2}+1$ or $\frac{N+1}{2}+1$.

We have now $S^2_{k_0-1}(g,h) = E(g,h)$ and $S^2_{k_1}(g,h) = S(g,h)$. Thus the hierarchy is finite and has at most $\max(M,N)+1$ levels. □

Example 3.19 Let the instance (g,h) be defined by

	1	2
g	a^3	a
h	a	a^4

Then $E(g,h) = \{w \mid \psi(w) = k(3,2),\ k \geq 1\}$ and $S(g,h) = \{11122, 11212, 11221, \ldots, 21211, 22111\}$.

Now

$$S^2_2(g,h) \neq E(g,h) \text{ because } (12121)^2 \notin S^2_2(g,h),$$
$$S^2_3(g,h) \neq S^2_2(g,h) \text{ because } (12211)^2 \in S^2_2(g,h) - S^2_3(g,h),$$
$$S^2_4(g,h) \neq S^2_3(g,h) \text{ because } (11122)^2 \in S^2_3(g,h) - S^2_4(g,h),$$

and $S(g,h) = S^2_4(g,h)$.

Thus we have the hierarchy $S(g,h) \subset S^2_3(g,h) \subset S^2_2(g,h) \subset E(g,h)$ of four levels. □

Corollary 3.39 *In a binary alphabet the P-hierarchy is finite.*

3.3.6 Some alternative ways to define L-prime solutions

We now return to the general type of L-prime solutions with no restrictions for the language L and consider other definitions for a word to be an L-prime solution.

To start with, we view a solution primitive if no suffix/subword/scattered subword can be removed such that the remaining part is a solution and a member of L. The three types of prime solutions defined in this way for each instance (g, h) are denoted by $F_L^{(1)}(g, h)$, $S_L^{(1)}(g, h)$ and $P_L^{(1)}(g, h)$.

This definition gives new variations for $S_L^{(1)}$- and $P_L^{(1)}$-prime solutions with Sections 3.3.1–3.3.3 in mind. For $F_L^{(1)}$-prime solutions the results stay the same. Since Lemma 3.13 holds now for $S_L^{(1)}$- and $P_L^{(1)}$-prime solutions, also Theorem 3.14 concerning the intersection is valid. In the same way Lemma 3.18 and Theorem 3.19 now hold for these two types. Hence the latter result makes the hierarchy in Theorem 3.21 collapse to the usual case

$$P(g, h) \subseteq S(g, h) \subseteq F(g, h).$$

Our second type is completely different. So far we have more or less tried to avoid the property of being a member of L; now we try to maintain it. We say that a solution w is $F_L^{(2)}$-prime if $w \in L$ and no suffix can be removed such that the remainder is both a solution and a member of L.

In this definition the properties of L-prime solutions with respect to set-theoretic operations reduce to the properties of the language L, unless $L \cap E(g, h) = \emptyset$ which implies that also $F_L^{(2)}(g, h)$, $S_L^{(2)}(g, h)$ and $P_L^{(2)}(g, h)$ are empty. So the earlier results for L-prime solutions are not valid any more. The only exception is Theorem 3.21 where the hierarchy is the same as above.

Chapter 4

Prime languages

There are two ways, based on equality and inclusion, to define the prime languages of the Post Correspondence Problem: given L, we require the existence of an instance (g, h) for which L is either equal to or included in the set of its prime solutions. This leads to different properties with respect to decidability, closure under basic operations and Chomsky hierarchy.

In the first section we prove that using inclusion definition the property of being a P-/S-/F-language can be reduced to Makanin's result, [22], at least in the finite case, and is thus decidable. The infinite case is also discussed here. For the equality method we have no similar result. In this case we concentrate on the binary alphabet, where we present the prime languages we know and conjecture that there are no others.

The second section is devoted to the closure properties of the families of prime languages. Most of the actions of morphisms, catenation, union and intersection follow from the previous section. A very basic observation is that even a complete understanding of F-words throws very little light on F-languages. Some of the results presented in this chapter should be viewed as only preliminary.

An interesting observation concerning P-/S-/F-languages is that they do not form an increasing hierarchy unless defined in inclusion sense. Thus their relation to Chomsky hierarchy, studied in the third section, is not as obvious as one might think. However, we are able to settle the mutual inclusion relations in all cases.

In the final section we show how to count the maximal number of P-prime solutions in terms of the size of the instance and the cardinality of the domain alphabet.

4.1 P-, S- and F-languages

We may define the prime languages by equality or inclusion. More specifically, we say that a language L is a P-, S- or F-language if, for some instance (g,h), L either equals or is a part of the set $P(g,h)$, $S(g,h)$ or $F(g,h)$, respectively. Defined by equality we notice that prime languages are extensions of the sets of prime solutions in the same way as prime words were of prime solutions.

Our first result is, in a way, an extension of Theorem 2.6.

Theorem 4.1 *For finite languages each of the properties of being a P-, an S- and an F-language in inclusion sense is decidable.*

Proof: In order to decide whether a given language $L = \{w_1, \ldots, w_n\}$ is a P-language we apply Makanin's result, [22], as well as the technique to change the inequalities to equalities, [3, 4], for the finite system of equations,

$$\begin{aligned}
g(w_1) &= h(w_1), \\
&\vdots \\
g(w_n) &= h(w_n), \\
g(u_1) &\neq h(u_1) \quad \text{for all } u_1 \in \mathit{scatsub}\,(w_1) - \{w_1, \lambda\}, \\
&\vdots \\
g(u_n) &\neq h(u_n) \quad \text{for all } u_n \in \mathit{scatsub}\,(w_n) - \{w_n, \lambda\}.
\end{aligned} \tag{4.1}$$

If this system has a solution; that is, the equations hold for some g and h, then L is a subset of $P(g,h)$; otherwise, L cannot be a P-language.

For S- or F-languages we use similarly the sets *sub* and *fin*. □

Consider now infinite languages. By Theorem 2.1 none of them can be a P-language. Thus P-languages are decidable also in this case. To decide whether an infinite language L is an S- or F-language we cannot apply the previous argument since the system of equations is now infinite.

By Ehrenfeucht's conjecture, established in [1] (see [31] for details), every language L possesses a finite subset D, called a test set, such that, whenever g and h are two morphisms defined on Σ^* and satisfying $g(w) = h(w)$ for every w in D, then $g(w) = h(w)$ holds for every w in L. The construction of D is not effective in general but is effective, for instance, for context-free languages. With this in mind we cannot, however, be sure that all the members of L are F- or S-prime solutions even if this is the case for D.

In the same context, [31], a more general result was proved: Every system of word equations possesses a finite subsystem equivalent to the original system. Hence, if we construct the similar system as (1), the previous result implies that there exists a finite subsystem which has exactly the same solutions and Makanin's result is again applicable. However, the finite subsystem is not effective here either. Nor do we have any obvious subclasses for which the construction is effective since it is possible that the subsystem is not the same as the test set for the given language.

For the prime languages defined by equality method the argument in Theorem 4.1 is not sufficient. If the system of equations fails to have a solution then the language cannot be a prime language in this sense either; otherwise, we do not know whether the language under examination contains all or only some of the prime solutions of $P(g, h)$, $S(g, h)$ or $F(g, h)$.

Consider now other, more practical, ways to identify a non-primitive language. The following three situations are quite obvious: some words of the set can be compared by P/S/F-relations, the set contains a word that is not a P/S/F-word such as the ones in Lemmas 2.16–2.17, Theorems 2.27–2.29 and Theorem 2.8, and finally any two words over a binary alphabet have different ratios (see Lemma 1.8).

From now on we consider prime languages defined by equality if not otherwise announced.

In a binary alphabet we have some possibilities to characterize the three prime languages since we know by Lemmas 2.21–2.22 that any finitely generated equality set is generated by at most two words. This implies the following two results concerning the cardinalities of the sets $P(g, h)$, $S(g, h)$ and $F(g, h)$ denoted by p, s and f, respectively.

Lemma 4.2 *If the instance (g, h) over a binary alphabet is not periodic then*

$$p \leq s = f \leq 2$$

and if (g, h) is periodic then

$$p = s < f = \infty.$$

We first focus the attention on P- and S-languages. From the earlier results we know that they are both finite and in periodic instances they consist of all words with the same basic Parikh vector

$$P(g,h) = \{w \in \Sigma^+ \mid \psi(w) = (k_1, k_2)\} = S(g,h)$$

where $\gcd(k_1, k_2) = 1$. This means that periodic P-languages are closed under permutations, in symbols

$$P(g,h) = c(\{1^i 2^j\}(i,j) = 1)$$

where, following the terminology of [15], $c(\Sigma^*)$ is a notion of a *free commutative monoid* generated by the alphabet Σ.

Consequently, we can determine all the P-languages with cardinality more than two. Denoting $\Sigma = \{1, 2\}$ we have, up to transposition of 1 and 2, only the following languages with cardinalities from three to nine:

$$\begin{aligned}
&\{112, 121, 211\}\\
&\{1112, 1121, 1211, 2111\}\\
&\{11112, \ldots, 21111\}\\
&\{111112, \ldots, 211111\}\\
&\{1111112, \ldots, 2111111\}\\
&\{11111112, \ldots, 21111111\}\\
&\{111111112, \ldots, 211111111\}
\end{aligned}$$

The condition $p = 10$ produces two languages

$$\begin{aligned}
&\{1111111112, \ldots, 2111111111\}\\
&\{11122, \ldots, 22111\}
\end{aligned}$$

and the very next languages with more than one appearance of 2 are

$$\begin{aligned}
&\{1111122, \ldots, 2211111\} \text{ with cardinality } \frac{7!}{5!2!} = 21,\\
&\{1111222, \ldots, 2221111\} \text{ with cardinality } \frac{7!}{4!3!} = 35.
\end{aligned}$$

Thus we can easily determine whether a language L with $\operatorname{card}(L) \leq 3$ is a P-language.

Theorem 4.3 *The properties of being a P- or an S-language are both decidable for languages with cardinality at least three over a binary alphabet.*

What is more, we can widen our result to alphabets with more than two letters in case the language L under examination has a subset $L_1 \subset L$ which fulfills the assumptions of Theorem 4.3. In other words, if L_1 is not a P-language in a binary alphabet then L cannot be a P-language in the original alphabet either. However, we cannot conclude anything from the fact that L_1 is a P-language, as the following example shows.

Example 4.1 Let the instance (g, h) be defined by

	1	2	3
g	a^3	a	b
h	a	a^2	a^2b

Now $P(g, h) = \{122, 212, 221, 13\}$ so the instance (g, h) need not be periodic.

□

For F-languages the circumstances are different. We know that periodic F-languages are always infinite but this is also the case for reduced languages

$$L(\beta, \gamma, \delta) = \beta\gamma^*\delta$$

introduced in Lemma 2.22. Nevertheless, as mentioned earlier, we strongly believe that no such reduced F-languages exist. Indeed, not only Conjecture 2.48 would suggest it but the results in Section 2.4.1 guarantee that none of the following reduced languages are F-languages

$$\begin{gathered}
1v^+2,\ 1v^+12,\ 11v^+2,\ 1v^+21,\ 12v^+1,\ 1v^+122,\ 11v^+22,\ 112v^+2,\\
1v^+112,\ 11v^+12,\ 111v^+2,\ 1v^+121,\ 11v^+21,\ 112v^+1,\\
1v^+1112,\ 11v^+112,\ 111v^+12,\ 1111v^+2,\\
1v^+1121,\ 11v^+121,\ 111v^+21,\ 1112v^+1,\\
1v^+1211,\ 11v^+211,\ 112v^+11,\ 1121v^+1,
\end{gathered}$$

where v fulfills the assumptions for γ. For instance, since $\{12, 1v2\} \subset 1v^+2$, the corresponding instance (g, h) is always periodic by Theorem 2.38; hence, $F(g, h)$ cannot be of the form $1v^+2$. In the same way the mirror images and transposition of 1 and 2 of the previous sets cannot be F-languages.

If our hypothesis is true; that is, the languages of the previous type do not exist, we are able to characterize the infinite F-languages. We know that each periodic F-language consists of all ratioprimitive words w with the same basic Parikh vector or, in other terms, the same ratio $r(w) = K$ for some $K \geq 0$.

In any case we have the following result by Lemma 4.2.

Theorem 4.4 *If L is a finite language with cardinality at least three in a binary alphabet then L is not an F-language.*

Consider now finite F-languages. We recall that Example 2.11 is the only case known (in addition to the trivial language $\{1, 2\}$) where $p = s = f = 2$ and there are no examples of the case $p < s = f = 2$ which we strongly believe does not exist though this is not the situation in larger alphabets.

Example 4.2 For the instance (g, h) in Example 3.1 we have $P(g, h) = \{13\}$ and $S(g, h) = F(g, h) = \{13, 1234\}$. □

So only the case $p = s = f = 1$ remains. We want to emphasize that the instances corresponding to the following languages are in no way unique.

Example 4.3 Let the instance (g, h) be defined, for $m, n \geq 1$, by the table

	1	2
g	$(a^{mn}b)^n$	a^m
h	a^n	$(ba^{mn})^m$

We have $E(g, h) = \{1^m 2^n\}^+$ and $P(g, h) = S(g, h) = F(g, h) = \{1^m 2^n\}$. □

Example 4.4 If the pair (g, h) is defined, for $i \geq 1$, by

	1	2
g	a	$b(ab)^{i+1}$
h	$(ab)^i a$	b

then $E(g, h) = \{1(21)^i\}^+$ and $P(g, h) = S(g, h) = F(g, h) = \{1(21)^i\}$. □

Example 4.5 The instance (g, h) defined, for an odd integer $i \geq 1$, by

	1	2
g	ba^ib	b
h	a	$b(ba^ib)^{\frac{i-1}{2}}b$

determines $E(g,h) = \{21^i2\}^+$ and $P(g,h) = S(g,h) = F(g,h) = \{21^i2\}$. □

Example 4.6 Let the instance (g,h) be defined, for $i, j \geq 1$, by

	1	2
g	a	$ba^j(a^iba^j)^{i-1}\gamma(a^iba^j)^{j-1}a^ib$
h	a^iba^j	γ

Then $E(g,h) = \{1^i21^j\}^+$ and $P(g,h) = S(g,h) = F(g,h) = \{1^i21^j\}$. □

Our final examples cannot be generalized in the same way as the previous ones.

Example 4.7 For the pair (g,h) defined by

	1	2
g	a	$ababa^4bab$
h	a^3baba	ba

we have $E(g,h) = \{11221\}^+$ and $P(g,h) = S(g,h) = F(g,h) = \{11221\}$. □

Example 4.8 Consider the instance

	1	2
g	a	$a^3ba^3(ba^{10}ba^3)^3b$
h	$a^3ba^3ba^3$	ba

Then $E(g,h) = \{1^4221^3\}^+$ and $P(g,h) = S(g,h) = F(g,h) = \{1^4221^3\}$. □

We conjecture that these are the only kind of prime languages (though there are probably more singletons of the types 4.7 and 4.8) where f is finite if we exclude the trivial languages as well as the transposition of 1 and 2.

Conjecture 4.5 *In the binary alphabet $\{1,2\}$ the only P- and S-languages are*

1. *two-word sets from Example 2.11 and*

2. $c(\{1^i2^j\}(i,j) = 1)$

and the only F-languages, in addition to 1.,

2′. $\{w \in \Sigma^+ | \, r(w) = K$ *and w is ratioprimitive*$\}$ *for some* $K \geq 0$.

In larger alphabets the situation is essentially more complicated. It is also somewhat surprising that the characterization of F-words gives very little information about F-languages.

Yet, for the inclusion definition we have the following result.

Theorem 4.6 *$L \subset \{1,2\}^*$ is an F-language (in inclusion sense) if and only if its members are ratioprimitive and of the same ratio.*

Proof: By Lemmas 2.2 and 1.8 the "only if"-part is obvious: every F-language is of the form indicated. On the other hand, if the words of L have the same ratio then, by Lemma 1.7, they are solutions for some periodic instance (g,h) whereas ratioprimitiveness now guarantees that they are, indeed, F-prime solutions. Using Lemma 2.7 any such instance (g,h) can be effectively constructed. $\square$

For S- and P-languages the result is weaker.

Theorem 4.7 *$L \subset \{1,2\}^*$ is a P- and an S-language (in inclusion sense) if either its members are of the same ratio and for each $w \in L$, $\gcd(|w|_1, |w|_2) = 1$ or $L \subset \{1^i2, 21^i\}$ (or symmetrically $\{12^i, 2^i1\}$).*

We conjecture that this is true also for the other direction.

In larger alphabets the situation remains unknown. We do not have any similar method, as ratio is in the binary case, to restrict the words.

4.2 Closure properties

In this section we study how the basic operations - morphisms, catenation, intersection and union - affect the families of three prime languages denoted by $\mathcal{P}$, $\mathcal{S}$ and $\mathcal{F}$. Most of our results are direct consequences of the previous section so if Conjecture 4.5 is showed to be true, most of the following results can be improved.

We start with mirror images.

Theorem 4.8 *Mirror image preserves the property of being a P-, S- or F-language.*

Proof: Let (g,h) be the instance for which $L = P(g,h)$. We construct a new instance (g',h') by setting $g'(a) = \mathrm{mi}\,(g(a))$ and $h'(a) = \mathrm{mi}\,(h(a))$ for all $a \in \Sigma$. Evidently, if $w \in P(g,h)$ then $\mathrm{mi}\,(w) \in E(g',h')$ and if there exists $w_1 \in \mathit{scatsub}\,(\mathrm{mi}\,(w))$ then $\mathrm{mi}\,(w_1) \in \mathit{scatsub}\,(w)$. Hence $\mathrm{mi}\,(w) \in P(g',h')$. On the other hand, since $\mathrm{mi}\,(\mathrm{mi}\,(w)) = w$ the converse follows from the same argument. □

The previous result implies that also the properties of being a P-, S- or F-word are preserved by mirror images. If we, however, change the definition of primality and consider, for instance, $F_{k-\mathrm{flat}}$-words this is not the case any more. The word 112212 is not an $F_{k-\mathrm{flat}}$-word since $112212 \in E(g,h)$ implies that also $1122 \in E(g,h)$ and 12 is a k-flatword. Nevertheless, it is easy to construct an instance (g,h) for which the mirror image 212211 is $F_{k-\mathrm{flat}}$-prime.

Example 4.9 For the pair (g,h) defined by

	1	2
g	a^2b	a^2
h	a	ba^2ba^2

we have $P(g,h) = S(g,h) = F(g,h) = \{112\}$, while the mirror image 211 is a solution for the instance

	1	2
g'	ba^2	a^2
h'	a	a^2ba^2b

□

4.2.1 Morphisms

We first show that any permutation of the domain alphabet preserves the language families.

Theorem 4.9 *If $\sigma : \Sigma^* \to \Sigma^*$ is a permutation of Σ and L is a P-, S- or F-language then $\sigma(L)$ is also a P-, S- or F-language, respectively.*

Proof: Considering the instance (g,h) corresponding to the language L, we simply change the labels of letters. □

Example 4.10 Let the instance (g,h) be defined by

$$\begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline g & ab & a & a \\ h & a & ba & ba \end{array}$$

for which $P(g,h) = S(g,h) = F(g,h) = \{12, 13\}$.

For the permutation $\sigma : 1 \mapsto 2,\ 2 \mapsto 3,\ 3 \mapsto 1$ we have $\sigma(P(g,h)) = \sigma(S(g,h)) = \sigma(F(g,h)) = \{31, 32\}$ which is obtained also by the instance

$$\begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline g' & a & a & ab \\ h' & ba & ba & a \end{array}$$

□

Our next result which follows directly from Theorem 2.8 exhibits a morphism for which the image is never an F-language (and thus not an S- or a P-language either).

Theorem 4.10 *If the morphism σ does not preserve the ratioprimitiveness of the words in L then $\sigma(L)$ is not an F-, S- or P-language.*

Example 4.11 For the language $L = \{1122\}$ and the morphism $\sigma(1) = 12$, $\sigma(2) = 21$ the set $\sigma(L) = \{12122121\}$ is not an F-, S- or P-language. □

A special case of the previous theorem is the morphism $\sigma : \Sigma^* \to \Sigma_1^*$ for which $\sigma(a) = \sigma(b)$ for all $a, b \in \Sigma$.

In the following we show that the cardinality of L being more than two plays a crucial role in our considerations, whereas the case $|L| \leq 2$ has no such property.

Example 4.12 Let $L_1 = \{12\}$, $L_2 = \{12, 21\}$ and consider a morphism $\sigma : \Sigma^* \to \Sigma^*$ such that $\sigma(1) = 11$ and $\sigma(2) = 2$. Then both $\sigma(L_1)$ and $\sigma(L_2)$ are P-, S- and F-languages. □

Theorem 4.11 *If $L \subset \{1,2\}^*$ is a P- or S-language with cardinality at least three then the only morphisms $\sigma : \{1,2\}^* \to \{1,2\}^*$ for which $\sigma(L)$ is a P- or S-language, respectively, are permutations.*

Proof: The claim follows from the fact that L is, by Lemma 4.2, periodic and only permutations preserve periodic languages. $\square$

Theorem 4.11 is valid also for periodic F-languages but we do not have any results for reduced languages.

It is not possible to obtain all P-languages of the same cardinality as a morphic closure of a single language. For instance, on page 80 we had two P-languages $L_1 = \{11122, \ldots, 22111\}$ and $L_2 = \{1^9 2, \ldots, 21^9\}$ of cardinality 10. There exists, however, no morphism σ for which $L_1 = \sigma(L_2)$ or $L_2 = \sigma(L_1)$.

Theorem 4.11 can be applied also to larger alphabets, as long as L has a subset L_1 which satisfies the conditions.

Example 4.13 The language

$$L = \{122, 212, 221, 133, 313, 331, 123, 132, 213, 231, 312, 321\}$$

is a P-language for the instance

	1	2	3
g	a^3	a	a
h	a	a^2	a^2

If we consider the morphism $\sigma : \Sigma^* \to \Sigma^*$ where $\sigma(1) = 1$, $\sigma(2) = 22$ and $\sigma(3) = 3$, then the image

$$\sigma(L) = \{12222, 22122, 22221, 133, 313, 331, 1223, 1322, 2213, 2231, 3122, 3221\}$$

cannot be a P-language since the words 21222 and 22212 are missing from the set. $\square$

4.2.2 Catenation

We begin with an obvious condition:

Lemma 4.12 *If $w_1 w_2$ is an F-word in a binary alphabet then $r(w_1) \neq r(w_2)$.*

In order to study the catenation of two languages over a binary alphabet, we may assume that all the words of L_i, $i = 1, 2$, have the same ratio, denoted by $r(L_i)$, $i = 1, 2$. If this is not the case we directly obtain the following result.

Lemma 4.13 *If the words of $L \subset \{1,2\}^*$ have different ratios then no catenation L_1L or LL_1 with another language $L \subset \{1,2\}^*$ is an F-, S- or P-language.*

Proof: The claim follows immediately from Lemma 1.8 if L_1 is empty or consists only of the empty word. So assume that there is a nonempty word $u \in L_1$ and $r(w_1) \neq r(w_2)$ for some $w_1, w_2 \in L$. Then the words w_1u and w_2u (resp. uw_1 and uw_2) belong to the set LL_1 (resp. L_1L). Since $r(w_1u) \neq r(w_2u)$ (resp. $r(uw_1) \neq r(uw_2)$), Lemma 1.8 implies that LL_1 (resp. L_1L) is not an F-language. □

Theorem 4.14 *If $r(L_1) = r(L_2)$ for $L_1, L_2 \subset \{1,2\}^*$ then L_1L_2 is not an F-, S- or P-language.*

A special case of the previous theorem is to choose $L_1 = L_2$.

Combining these two results we notice that if L_1L_2 is an F-, S- or P-language then the words of L_i, $i = 1, 2$, must have the same ratio and, moreover, $r(L_1) \neq r(L_2)$.

It turns out that catenation has a similar relation to the cardinality of the languages as morphisms had.

Example 4.14 Let $L_1 = \{211\}$ and $L_2 = \{12\}$. Then $L_1L_2 = \{21112\}$ is a P-language as is the catenation L_3L_4 of the languages $L_3 = \{121\}$ and $L_4 = \{21\}$. □

Example 4.15 If $L_1 = \{12, 21\}$ and $L_2 = \{1\}$ then $L_1L_2 = \{121, 211\}$ is not a P-language. In fact, it seems that no catenation of L_1 with another P-language L_2 is a P-language. □

Theorem 4.15 *If $L_1, L_2 \subset \{1,2\}^*$ are P-, S- or F-languages and $|L_1| \geq 3$, the catenation L_1L_2 is not a P-, S- or F-language.*

4.2.3 Union and intersection

For the union of two languages we clearly have the following condition by Lemma 1.8.

Lemma 4.16 *If $L_1 \cup L_2$ is an F-language in a binary alphabet then $r(w) = r(u)$ for all $w \in L_1$ and $u \in L_2$.*

In some cases the union $L_1 \cup L_2$ is a P-language.

Example 4.16 Let $L_1 = \{1^i 2\}$ and $L_2 = \{21^i\}$ for some $i \geq 1$. Then $L_1 \cup L_2 = \{1^i 2, 21^i\}$ is a P-language, as is the union of the languages $L_1 = \{112, 211\}$ and $L_2 = \{121\}$. □

In fact, the previous example shows the only case where $|L_1| = 2$ and the consequence of the following theorem is not valid, whereas, by Lemma 4.2, the result for F-languages is satisfied also here.

Theorem 4.17 *If $L_1, L_2 \subset \{1,2\}^*$ are P- or S-languages (resp. F-languages) and $|L_1| \geq 3$ (resp. $|L_1| \geq 2$) then the union $L_1 \cup L_2$ is a P- or S-language (resp. an F-language) if and only if $L_1 \subset L_2$ or $L_2 \subset L_1$.*

The intersection has a very similar result.

Theorem 4.18 *If $L_1, L_2 \subset \{1,2\}^*$ are P- or S-languages (resp. F-languages) and $|L_1|, |L_2| \geq 3$ (resp. $|L_1| \geq 3$) then the nonempty intersection $L_1 \cap L_2$ is a P- or S-language (resp. an F-language) if and only if $L_1 \subset L_2$ or $L_2 \subset L_1$.*

It seems that also for P- and S-languages the bound $|L_1| \geq 3$ is sufficient. Consider, for instance, the case $|L_1| = 3$. Then $L_1 = \{112, 121, 211\}$ and the only possible P-languages L_2 for which the intersection $L_1 \cap L_2$ is not empty are $\{112\}$, $\{121\}$, $\{211\}$ and $\{112, 211\}$. Even the bound $|L_1| \geq 2$ seems to be enough. At least we do not have any counter-examples for which $L_1 = \{u, w\}$, $L_2 = \{w, v\}$ and $u \neq v$.

For the difference we easily obtain the following result.

Theorem 4.19 *If $L_1, L_2 \subset \{1,2\}^*$ are P- or S-languages (resp. F-languages) and $|L_1| \geq 4$ (resp. $|L_1| \geq 3$) then the difference $L_1 - L_2$ is a P- or S-language (resp. an F-language) if and only if $L_1 = L_2$ or $L_1 \cap L_2 = \emptyset$.*

The case $|L_1| \geq 5$ follows by considering the possible cardinalities of $L_1 \cap L_2$, whereas the case $|L_1| = 4$ is shown in the next example.

Example 4.17 Let $L_1 = \{1112, 1121, 1211, 2111\}$ and $L_2 = \{1112, 2111\}$. The difference $L_1 - L_2 = \{1121, 1211\}$ is not a P-language since any two words from the set $1^+2^*1^*$ are periodicity forcing. □

If we consider P-, S- and F-languages defined in inclusion sense the closure properties are different. Instead of Theorems 4.11, 4.15, 4.17, 4.18 and 4.19 we have the following results.

Theorem 4.20 *If $L_1, L_2 \subset \{1,2\}^*$ are P-, S- or F-languages and $r(L_1) \neq r(L_2)$ then the catenation L_1L_2 is a P-, S- or F-language, respectively.*

Example 4.18 Let $L_1 = \{12, 21\}$ and $L_2 = \{112, 121, 211\}$. Then they are both P-languages, $r(L_1) \neq r(L_2)$ and their catenation

$$L_1L_2 = \{12112, 12121, 12211, 21112, 21121, 21211\}$$

is also a P-language, being a subset of the set $P(g, h)$ for the instance

	1	2
g	a^3	a
h	a	a^4

□

Theorem 4.21 *If $L_1, L_2 \subset \{1,2\}^*$ are F-languages (resp. P- or S-languages and $|L_1| \geq 2$) then the union $L_1 \cup L_2$ is an F-language (resp. a P- or S-language) if and only if $r(w) = r(u)$ for all $w \in L_1, u \in L_2$.*

The reason why $|L_1| = 1 = |L_2|$ is not sufficient for P- and S-languages is shown below.

Example 4.19 Though $L_1 = \{1122\}$ and $L_2 = \{2211\}$ are both P-languages by Example 4.3 and $r(L_1) = r(L_2)$, the union $L_1 \cup L_2 = \{1122, 2211\}$ is not a P-language because of Theorem 2.39. □

The next two theorems follow from the fact that $L_1 \cap L_2$ and $L_1 - L_2$ are both subsets of L_1.

Theorem 4.22 *If $L_1, L_2 \subset \{1,2\}^*$ are P-, S- or F-languages then the intersection $L_1 \cap L_2$ is also a P-, S- or F-language, respectively.*

Theorem 4.23 *If $L_1, L_2 \subset \{1,2\}^*$ and L_1 is a P-, S- or F-language then the difference $L_1 - L_2$ is also a P-, S- or F-language, respectively.*

4.3 Chomsky hierarchy

We now study the relation of language families $\mathcal{P}$, $\mathcal{S}$ and $\mathcal{F}$ to the Chomsky hierarchy. Exhaustive characterizations are obtained both in the equality and inclusion sense. For more details of Chomsky hierarchy, see [30].

We start with regular languages.

Proposition 4.24

(*a*) *P-languages are regular.*

(*b*) *There exists a regular P-language that is neither an S- nor an F-language.*

(*c*) *There exists a regular language that is both a P- and an S-language but not an F-language.*

(*d*) *There exists a finite language that belongs to all classes $\mathcal{P}$, $\mathcal{S}$ and $\mathcal{F}$.*

(*e*) *There exists a regular language that is both an S- and an F-language but not a P-language.*

(*f*) *There exists a regular S-language that is neither a P- nor an F-language.*

(*g*) *There exists a regular F-language that is neither a P- nor an S-language.*

(*h*) *Every P-language that is not an S-language is not an F-language either.*

(*i*) *Every F-language that is not an S-language is not a P-language either.*

Proof: (a): See Theorem 2.1.
(b): Consider

	1	2	3
g	a	a^4	a
h	a^3	a	a^4

where $P(g,h) = \{11122, \ldots, 22111, 23, 32\}$. The fact that the word 12112 is periodicity forcing by Theorem 2.34 together with the words 23 and 32 implies that any instance (g,h) corresponding to this language is periodic. But then $S(g,h) \neq P(g,h)$ because, for instance, $12^{i+2}113^i1 \in S(g,h)$, $i \geq 1$.

(c): Apply Theorem 4.4.
(d): Several such languages were shown in Section 4.1.
(e): See Example 4.2.
(f): Consider the instance

	1	2	3	4	5	6
g	ab	cb	cde	def	s	s^4
h	abc	$bcde$	de	f	s^3	s

We have $S(g,h) = \{13, 1234, 55566, 55656\ldots, 66555\}$ which is neither a P- nor an F-language.

(g): We use a slightly modified version of Post's famous instance, [27],

	1	2	3
g	ab	ab	c
h	a	ba	bc

for which $F(g,h) = 12^*3$. This is not an S- or a P- language since $13 \in sub(12^*3)$.

(h): If L is a P-language but not an S-language then in each instance (g,h) where $P(g,h) = L$ we have $S(g,h) - L \neq \emptyset$. But this means that also $F(g,h) - L \neq \emptyset$ and L is not an F-language either.

(i): Let L be an F-language which is not an S-language. Then there are $w, w' \in L$ such that $w \in sub(w')$; hence, L is not a P-language either. $\square$

Next we consider context-free languages that are not regular. Notice that the word "context-free" is used in this restricted sense.

Proposition 4.25

(a) *There exists a context-free language that is both an S- and an F-language.*

(b) *There exists a context-free S-language that is not an F-language.*

(c) *There exists a context-free F-language that is not an S-language.*

Proof: (a): For the instance

	1	2	3
g	$(ab)^3$	a	ba
h	ab	aba	$(ba)^2$

we have $S(g,h) = F(g,h) = \{1^i 2 3^{2i-1} : i \geq 1\}$.

(b): Add to the previous instance two more letters 4 and 5 such that $g(4) = c$, $g(5) = c^4$, $h(4) = c^3$ and $h(5) = c$. Then $S(g,h) = \{1^i 2 3^{2i-1} : i \geq 1\} \cup \{44455, \ldots, 55444\}$ is context-free but not an F-language.

(c): Let (g,h) be the basic periodic instance

	1	2
g	a^2	a
h	a	a^2

Then $F(g,h)$ is context-free since $E(g,h)$ is context-free but, being infinite, it cannot be an S-language. □

Finally we study languages that are context-sensitive but not context-free. As before the word "context-sensitive" is used in this restricted sense.

Proposition 4.26

(a) *There exists a context-sensitive language that is both an S- and an F-language.*

(b) *There exists a context-sensitive S-language that is not an F-language.*

(c) *There exists a context-sensitive F-language that is not an S-language.*

Proof: (a): For the instance

	1	2	3
g	a^2	b	b
h	a	a	b^2

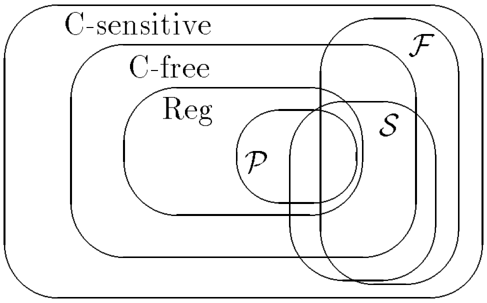
we have $S(g,h) = F(g,h) = \{1^i 2^i 3^i, 3^i 2^i 1^i : i \geq 1\}$.

(b): Add to the previous instance (g,h) two new letters 4 and 5 such that $g(4) = c^3$, $g(5) = c$, $h(4) = c$ and $h(5) = c^4$. Then $S(g,h) = \{44455, \ldots, 55444\} \cup \{1^i 2^i 3^i, 3^i 2^i 1^i : i \geq 1\}$ is context-sensitive but not an F-language.

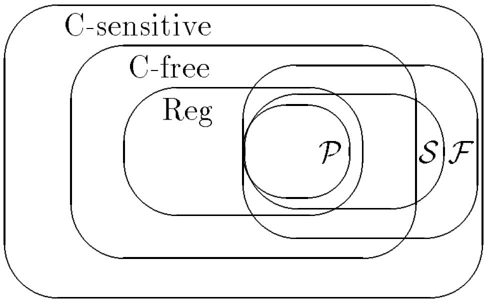
(c): Add now to the first instance (g,h) three letters 4, 5 and 6 such that $g(4) = g(5) = cd$, $g(6) = e$, $h(4) = c$, $h(5) = dc$ and $h(6) = de$. Then the language $F(g,h) = \{1^i 2^i 3^i, 3^i 2^i 1^i : i \geq 1\} \cup \{45^i 6 : j \geq 1\}$ cannot be an S-language. □

Before summarizing the results, we want to emphasize that there exist regular, context-free and context-sensitive languages that do not belong to any of the classes $\mathcal{P}$, $\mathcal{S}$ or $\mathcal{F}$. This is the case, for example, if the word 1212 is added to any of the languages we have had so far.

According to Propositions 4.24–4.26 we have the following figure where none of the areas is empty:



The situation is totally different if P-, S- and F-languages are defined by inclusion method. Now the language families form an increasing hierarchy. Hence Proposition 4.24 (b)–(c), Proposition 4.25 (b) and Proposition 4.26 (b) are not valid anymore and Proposition 4.24 (f) holds only for the part of P-words. The figure is as follows:



4.4 Maximal number of P-prime solutions

Since the set $P(g, h)$ is, by Theorem 3.6, finite for all instances (g, h), it is especially interesting to count the greatest possible number in terms of the cardinality of the domain alphabet Σ. The remainder of the thesis is devoted to this task. We start the study with periodic instances and try to infer their importance in our constructions. We also show that a period of length one gives the best result.

So far we have studied and used periodic instances in many occasions. In particular, we recall that any permutation of a solution is also a solution and the only significant matter in the definition of g and h are the differences

$$n_i - m_i, \ i \in \Sigma$$

where n_i and m_i are the powers of $g(i)$ and $h(i)$, respectively.

For instance, if

$$n_1 - m_1 = \pi > 0 \quad \text{and} \quad n_2 - m_2 = -\mu < 0$$

in a binary alphabet $\{1, 2\}$ then any word containing $\frac{\mu}{\gcd(\pi,\mu)}$ 1's and $\frac{\pi}{\gcd(\pi,\mu)}$ 2's is a P-prime solution.

Consider the following problem. If an alphabet has n symbols then the number of k-tuples which have k_1 symbols of type 1, k_2 symbols of type 2, $\ldots$, and k_n symbols of type n is given by the *multinomial coefficient*

$$\frac{k!}{k_1!k_2!\ldots k_n!}, \tag{4.2}$$

where $k_1 + k_2 + \cdots + k_n = k$.

With this formula we are able to count all the permutations of a given P-prime solution. For instance, for the previous notations, $n_1 - m_1 = \pi > 0$ and $n_2 - m_2 = -\mu < 0$, the total number of the P-prime solutions is

$$\frac{(\frac{\mu}{\gcd(\pi,\mu)} + \frac{\pi}{\gcd(\pi,\mu)})!}{(\frac{\mu}{\gcd(\pi,\mu)})!(\frac{\pi}{\gcd(\pi,\mu)})!}$$

Thus the number of P-prime solutions depends not only on the cardinality of Σ but also on the size of the instance, that is, the lengths of $g(i)$ and $h(i)$, $i \in \Sigma$.

Example 4.20 Let the pair (g, h) be defined by

	1	2
g	a^2	a
h	a	a^k

Since $\gcd(1, k-1) = 1$, every word consisting of $k-1$ 1's and one 2 is a P-prime solution. Hence we have the power of $h(2)$ P-prime solutions. □

Consider now the binary alphabet and assume that there is $l \in \mathbb{N}$ for which $|g(1)|, |g(2)|, |h(1)|, |h(2)| \leq l$. Using the previous notations, we seek for the values $\pi, \mu < l$ with the most P-prime solutions. In case $|\Delta| = 1$ the best choice turns out to be $l-2$ and $l-1$ which are the largest relatively prime numbers less than l. It suffices to prove that any choice of $\pi < l-2$ produces less P-prime solutions.

Theorem 4.27 *Let $g, h : \Sigma^* \to \Delta^*$ be periodic morphisms over a binary alphabet such that $|\Delta| = 1$ and $|g(i)|, |h(i)| \leq l$ for all $i \in \Sigma$. The greatest number of P-prime solutions the instance (g, h) can have is*

$$\frac{[(l-2)+(l-1)]!}{(l-2)!(l-1)!}$$

Proof: We first show that if $t = \gcd(b, l-1)$ and $b < l-2$ then

$$\frac{(b+l-1)!}{b!(l-1)!} \geq \frac{(\frac{b}{t}+\frac{l-1}{t})!}{(\frac{b}{t})!(\frac{l-1}{t})!} \tag{4.3}$$

By multiplying the left side with the inverse of the right side, we have

$$\frac{(b+l-1)(b+l-2)\cdots(\frac{b}{t}+\frac{l-1}{t}+1)}{b(b-1)\cdots(\frac{b}{t}+1)(l-1)(l-2)\cdots(\frac{l-1}{t}+1)} \geq 1.$$

There are now equally many factors in the divisor and dividend. If we first put the divisors in decreasing order and then separate the whole expression into fractions, we notice that each fraction is bigger than 1. Thus inequality (3) holds.

With this in mind it now suffices to prove that if $b < l-2$ then

$$\frac{(2l-3)!}{(l-2)!} \geq \frac{(b+l-1)!}{b!}$$

This is accomplished by induction on l, the case $l = 4$ being obvious. If $b < l-3$ then

$$\frac{(2l-3)!}{(l-2)!} = \frac{2l-3}{l-2} \cdot \frac{(2l-4)!}{(l-3)!} \geq \frac{2l-3}{l-2} \cdot \frac{(b+l-1)!}{b!} \geq \frac{(b+l-1)!}{b!}$$

by the induction hypothesis. On the other hand, if $b = l - 3$ then

$$\frac{(2l-3)!}{(l-2)!} = \frac{2l-3}{l-2} \cdot \frac{(2l-4)!}{(l-3)!} \geq \frac{(2l-4)!}{(l-3)!}$$

which concludes the proof. □

Assume now that the period of g and h is over more than one letter. Then the maximal number depends not only on the upper bound l but also on the length of the period w, leading to at most

$$\frac{[(s-2)+(s-1)]!}{(s-2)!(s-1)!} \quad \text{where} \quad s = \lfloor \frac{l}{|w|} \rfloor$$

P-prime solutions.

Example 4.21 Let the instances (g_1, h_1) and (g_2, h_2) be defined by the tables

	1	2
g_1	$(abc)^2$	abc
h_1	abc	$(abc)^3$

	1	2
g_2	a^8	a
h_2	a	a^9

In both cases the upper bound for the lengths is 9; still, the difference between the numbers of P-prime solutions is huge

$$\begin{aligned} p_1 &= 3, \\ p_2 &= \frac{(8+7)!}{8!7!} = 6435. \end{aligned}$$

□

Consequently, with the lengths of $g(i)$ and $h(i)$ bounded, it is best to choose $|w| = 1$ in order to obtain the maximal p. We also notice that since nonperiodic instances have at most two P-prime solutions in a binary alphabet (by Lemma 4.2), Theorem 4.27 provides actually the maximal number of P-prime (and S-prime) solutions.

Theorem 4.28 *In a binary alphabet the instance (g, h) with $|g(i)|, |h(i)| \leq l$, $i \in \Sigma$, has at most*

$$\frac{(2l-3)!}{(l-2)!(l-1)!}$$

P- and S-prime solutions.

This priority of periodic instances seems to hold also in larger alphabets. After all, the nonperiodic instances lack the permutation property of solutions. Some comparison between these two types can be seen in the following example as well as in Section 4.4.4.

Example 4.22 Let the pairs (g_1, h_1) and (g_2, h_2) be defined by

	1	2	3
g_1	a^5b	a	a^2
h_1	a^2	a^2ba^3	a

	1	2	3
g_2	a^6	a^6	a
h_2	a	a	a^5

Both instances are of the same size, $l = 6$. Analyzing possible solutions, we see that 132, 311223 and 333111222 are the only P-prime solutions for the first instance (g_1, h_1), whereas the instance (g_2, h_2) has the total of 4032 P-prime solutions, as seen later on. □

From now on, in determining an upper bound for p, we examine only the instances that are periodic over one letter.

4.4.1 Differences and basic matrices

In alphabets with at least three letters we have some new aspects to consider. The following two examples show how the choice of positive and negative differences affects the number of P-prime solutions.

Example 4.23 Let the instances (g_1, h_1) and (g_2, h_2) be defined by

	1	2	3
g_1	a^2	a^3	a
h_1	a	a	a^4

	1	2	3
g_2	a^3	a^3	a
h_2	a	a	a^4

We see that 1113, 22233, 123 and all their permutations are the P-prime solutions of (g_1, h_1), whereas 11133, 22233, 11233, 12233 are P-prime solutions for (g_2, h_2). Using (2), the formula of the multinomial coefficients, we obtain

$$\begin{aligned} p_1 &= \frac{4!}{3!1!} + \frac{5!}{3!2!} + 3! = 20, \\ p_2 &= \frac{5!}{3!2!} + \frac{5!}{3!2!} + \frac{5!}{2!2!1!} + \frac{5!}{2!2!1!} = 80. \end{aligned}$$

□

Example 4.24 Let the pairs (g_3, h_3) and (g_4, h_4) be defined by

	1	2	3	4
g_3	a^3	a^3	a	a
h_3	a	a	a^3	a^4

	1	2	3	4
g_4	a^3	a^3	a	a
h_4	a	a	a^4	a^4

Now 13, 23, 11144, 22244, 11244 and 12244 are P-prime solutions of (g_3, h_3), whereas in (g_4, h_4) they are 11133, 11134, 11144, 11233, 11234, 11244, 12233, 12234, 12244, 22233, 22234 and 22244. Thus

$$\begin{aligned} p_3 &= 2! + 2! + \frac{5!}{3!2!} + \frac{5!}{3!2!} + \frac{5!}{2!2!1!} + \frac{5!}{2!2!1!} = 84, \\ p_4 &= \frac{5!}{3!2!} \cdot 4 + \frac{5!}{3!} \cdot 2 + \frac{5!}{2!2!} \cdot 4 + \frac{5!}{2!} \cdot 2 = 320. \end{aligned}$$

□

Theorem 4.29 *Let $|\Sigma| = n$ and $|g(i)|, |h(i)| \leq l$ for all $i \in \Sigma$. If $|g(i_j)| - |h(i_j)| > 0$, for $j = 1, \ldots, k$, and $|g(i_j)| - |h(i_j)| < 0$, for $j = k+1, \ldots, n$, then by choosing the positive differences $l-2$ and the negative ones $-(l-1)$, or vice versa, we obtain more P-prime solutions than by any other choice. Moreover, in this case all the P-prime solutions are of equal length.*

Proof: Since $|g(i)|, |h(i)| \leq l$, $i \in \Sigma$, the positive differences have an upper bound $l-1$ whereas the negative differences have a lower bound $-(l-1)$, respectively. According to Theorem 4.27, for each pair of letters i_r and i_q, where $r \in \{1, \ldots, k\}$ and $q \in \{k+1, \ldots, n\}$, that is, $|g(i_r)| - |h(i_r)| =: b_r > 0$ and $|g(i_q)| - |h(i_q)| =: -b_q < 0$, the case with most P-prime solutions is achieved by choosing $b_r = l-2$ and $b_q = l-1$ (or vice versa). On the other hand, if this is the case for the letters $i_{r_1}, \ldots, i_{r_t}$ and $i_{q_1}, \ldots, i_{q_s}$, then we have to choose exactly $l-1$ letters of $i_{r_1}, \ldots, i_{r_t}$ and $l-2$ letters of $i_{q_1}, \ldots, i_{q_s}$ to obtain a P-prime solution. Thus every P-prime solution is of equal length. Since the differences b_{r_j} and b_{q_j} are restricted by $l-1$, this length is the maximum and cannot be exceeded by any other choice of b_{r_j} and b_{q_j}. Consequently, this choice produces the most P-prime solutions. □

Thus the best choice for the positive differences is $l-2$ and for the negative ones $-(l-1)$. Another important aspect in larger alphabets is to decide how the domain alphabet should be divided into these two differences, that is, how to determine the k in the previous theorem.

If we arrange the P-prime solutions so that each group contains words that are permutations of each other and then take the first ones in alphabetical order, we obtain a *basic matrix* from which all the P-prime solutions may be achieved. This is exhibited in the following example.

Example 4.25 Let the pairs (g_1, h_1) and (g_2, h_2) be defined by the tables

	1	2	3	4
g_1	a^3	a^3	a	a
h_1	a	a	a^4	a^4

	1	2	3	4	5
g_2	a^3	a^3	a^3	a	a
h_2	a	a	a	a^4	a^4

Example 4.24 has already shown that the permutations of the following words are the P-prime solutions of (g_1, h_1)

11133	11233	12233	22233
11134	11234	12234	22234
11144	11244	12244	22244

whereas the P-prime solutions of the second instance (g_2, h_2) are

11144	11244	11344	12244	12344	13344	22244	22344	23344	33344
11145	11245	11345	12245	12345	13345	22245	22345	23345	33345
11155	11255	11355	12255	12355	13355	22255	22355	23355	33355

□

With these new notions we are now ready to construct the algorithm in terms of the cardinality of the domain alphabet and the lengths of $g(i)$ and $h(i)$, $i \in \Sigma$.

4.4.2 Algorithm for maximal p

Let $|g(i)|, |h(i)| \leq l$ and $|\Sigma| = n$. In view of the previous section we may assume that for the first k letters of Σ, $|g(a_j)| - |h(a_j)| = l - 2 > 0$ and for the last $n - k$ letters, $|g(a_j)| - |h(a_j)| = -(l-1) < 0$.

We first count the k which produces the most P-prime solutions or, at this point, the largest basic matrix, and after this we count the maximal p. To start with consider all the combinations of $a_1, \ldots, a_k$, and then the combinations of $a_{k+1}, \ldots, a_n$.

In Example 4.24 we have already used the fact that the P-prime solutions having equally many different letters produce equally many permutations or

P-prime solutions outside the basic matrix. For instance, both 11144 and 22255 produce $\frac{5!}{3!2!} = 10$ P-prime solutions. This comes up in counting the total number of P-prime solutions. If we have vectors $(i_1, \ldots, i_k)$ in which i_j indicates the number of a_j's to be taken, it turns out that the P-prime solutions mentioned above have the same vectors up to the order of the indices i_j. Consequently, when considering the combinations of $a_1, \ldots, a_k$ it suffices to determine only the vectors $(i_1, \ldots, i_k)$ where

$$i_1 + \ldots + i_k = l - 1 \quad \text{and} \quad i_1 \geq i_2 \geq \ldots \geq i_k,$$

and keep in mind how many other vectors they represent. Clearly, in this way we are able to find all the vectors corresponding to the P-prime solutions of the basic matrix. Such *basic vectors* are constructed as follows:

$$\begin{array}{l}
i_1 = l \longrightarrow 1 \\
\quad i_2 = \min(i_1, l - i_1) \longrightarrow 1 \\
\quad\quad i_3 = \min(i_2, l - i_1 - i_2) \longrightarrow 1 \\
\quad\quad\quad \ldots \\
\quad\quad\quad\quad i_s = \min(i_{s-1}, l - \sum_{j=1}^{s-1} i_j) \longrightarrow 1 \\
\quad\quad\quad\quad\quad \ldots \\
\quad\quad\quad\quad\quad\quad i_k = \min(i_{k-1}, l - \sum_{j=1}^{k-1} i_j) \longrightarrow 1
\end{array}$$

We form several loops one within the other, since the values of the indices $i_1, \ldots, i_{s-1}$ determine the value of i_s completely. The notation $i_1 = l \longrightarrow 1$ means that i_1 gets all the values $l, l-1, l-2, \ldots, 2, 1$ one at a time. Clearly, when one of the indices i_s equals 0, then also the indices $i_{s+1}, \ldots, i_k$ equal 0.

The first four basic vectors are

$$(l, 0, \ldots, 0), \ (l-1, 1, 0, \ldots, 0), \ (l-2, 2, 0, \ldots, 0), \ (l-2, 1, 1, 0, \ldots, 0)$$

whereas the last one is always

$$(\underbrace{1, \ldots, 1}_{l-1}, 0, \ldots, 0).$$

Even though k is not yet known we act as if $k \geq l-1$. If this is not the case, we have to drop away some of the basic vectors as discussed later on.

Having obtained all the basic vectors $(i_1, \ldots, i_k)$, we are now ready to count how many other vectors or the P-prime solutions of the basic matrix correspond to them. We first count the number of times each number $1, \ldots, l$

appears in each basic vector. If k_j indicates the number of j's, the total number of combinations for each basic vector is

$$\frac{k!}{k_1!k_2!\dots k_l!(k-\sum_{j=1}^{l}k_j)!}$$

where $k-\sum_{j=1}^{l}k_j$ indicates the number of zeros in the vector. After cancelling a fraction, we have a polynomial of k; for instance, the combinations of our previous vectors are

$$\begin{array}{rcl}
(l,0,\dots,0) & : & \dfrac{k!}{1!(k-1)!}=k \\
(l-1,1,0,\dots,0) & : & \dfrac{k!}{1!1!(k-2)!}=k(k-1) \\
(l-2,2,0,\dots,0) & : & \dfrac{k!}{1!1!(k-2)!}=k(k-1) \\
(l-2,1,1,0,\dots,0) & : & \dfrac{k!}{1!2!(k-3)!}=\dfrac{1}{2}k(k-1)(k-2) \\
(1,\dots,1,0,\dots,0) & : & \dfrac{k!}{(l-1)!(k-l+1)!}
\end{array}$$

In the same way we carry out the previous operations to the indices $(i_{k+1},\dots,i_n)$. This time, however, the basic vectors are of the form

$$i_{k+1}+\dots+i_n=l-2 \quad \text{and} \quad i_{k+1}\geq i_{k+2}\geq\dots\geq i_n,$$

with the total of

$$\frac{(n-k)!}{k_1!k_2!\dots k_l!(n-k-\sum_{j=1}^{l}k_j)!}$$

combinations.

We have thus two lists of polynomials over k of the combinations of the basic vectors $(i_1,\dots,i_k)$ and $(i_{k+1},\dots,i_n)$. In order to find the best possible k, we first sum up the polynomials in each list and then multiply these two sums. Denote the product by P. Before the multiplication the first sum indicates actually the columns, the second one the rows and their product P the size of the basic matrix, respectively. To maximize this size, we evaluate P at $k=\frac{n+1}{2},\frac{n+1}{2}+1,\dots,n-1$ if n is odd, and at $k=\frac{n}{2},\frac{n}{2}+1,\dots,n-1$ if n is even. Usually only one of these values produces the largest basic matrix. If, however, two successive values end up with the same size, we choose the bigger one since in this case the corresponding basic matrix has more words over several letters and thus produces more P-prime solutions outside the basic matrix. So the structures of the two matrices are different even though their sizes are the same.

Example 4.26 Let $l = 4$ and $n = 6$. Choosing $k = 4$ we obtain the basic matrix of size 20×3,

11155	11156	11166
11255	11256	11266
11355	11356	11366
11455	11456	11466
12255	12256	12266
12355	12356	12366
12455	12456	12466
13355	13356	13366
13455	13456	13466
14455	14456	14466
22255	22256	22266
22355	22356	22366
22455	22456	22466
23355	23356	23366
23455	23456	23466
24455	24456	24466
33355	33356	33366
33455	33456	33466
34455	34456	34466
44455	44456	44466

and with $k = 3$ the basic matrix of size 10×6,

11144	11145	11146	11155	11156	11166
11244	11245	11246	11255	11256	11266
11344	11345	11346	11355	11356	11366
12244	12245	12246	12255	12256	12266
12344	12345	12346	12355	12356	12366
13344	13345	13346	13355	13356	13366
22244	22245	22246	22255	22256	22266
22344	22345	22346	22355	22356	22366
23344	23345	23346	23355	23356	23366
33344	33345	33346	33355	33356	33366

Though the sizes of the matrices are the same, the first one produces 2560 and the second one 2430 P-prime solutions. (Notice that, contrary to our usual practise, we have indicated in both matrices the vectors $(i_1, \ldots, i_k)$ as the rows and $(i_{k+1}, \ldots, i_n)$ as the columns, the former number being too large to fit the width of the page.) $\square$

Having found the best k, we are now able to count the maximal p. First we reconstruct the two lists of combinations by evaluating the polynomials at k. If any of these combinations equals to zero then we drop it away as well as the corresponding basic vector. In this case k has been smaller than expected, $k < l-1$ (or $n-k < l-2$), yielding, for instance, that the sum of the first k indices in $(1,\ldots,1,0,\ldots,0)$ does not equal to $l-1$.

With the two lists of basic vectors as well as their combinations, we next construct the basic matrix. For every pair of basic vectors $(i_1,\ldots,i_k)$ and $(i_{k+1},\ldots,i_n)$ we obtain the P-prime solution of the basic matrix that has i_1 a_1's, i_2 a_2's, $\ldots$, i_k a_k's, i_{k+1} a_{k+1}'s, $\ldots$, i_n a_n's, in this order. Thus it corresponds to

$$\frac{(2l-3)!}{i_1!\cdots i_k! i_{k+1}!\cdots i_n!} \tag{4.4}$$

P-prime solutions (by our definitions $2l-3$ is the same as $i_1+\ldots+i_n$). Since several other P-prime solutions in the basic matrix produce (4), we must multiply this number by the product of the combinations corresponding to the basic vectors $(i_1,\ldots,i_k)$ and $(i_{k+1},\ldots,i_n)$. The maximal number of P-prime solutions or p is finally obtained by adding up all these products.

4.4.3 Example of basic matrices

As an example we apply the previous algorithm for the case $|g(i)|, |h(i)| \leq 4$, $i \in \Sigma$. The longest P-prime solution is here of length 5 and consists of 3 letters with positive and 2 letters with negative differences. We start with the combinations of $a_1,\ldots,a_k$.

This means that if we have, for instance, three letters 1, 2, 3 ($k=3$) we seek for the permutations of (i_1,i_2,i_3) where i_1 indicates the number of 1's to be taken, i_2 the number of 2's, and i_3 the number of 3's such that

$$i_1+i_2+i_3=3.$$

In this case the possibilities are

(3,0,0), (0,3,0), (0,0,3),
(2,1,0), (2,0,1), (1,2,0), (1,0,2), (0,2,1), (0,1,2),
(1,1,1),

(where the first column contains the basic vectors) and the corresponding words are

111, 222, 333,
112, 113, 122, 133, 223, 233,
123

as we have already seen in Example 4.25.

In general, with k letters we check the permutations of $(i_1, \ldots, i_k)$, in which each i_j indicates the number of letters a_j to be taken, and

$$i_1 + i_2 + \ldots + i_k = 3.$$

By the multinomial coefficients we can calculate the total number of words having either three letters of the same type or two letters of one type and one letter of another type or three entirely different letters, respectively. In the first case we obtain $\frac{k!}{1!(k-1)!} = k$ words, in the second case $\frac{k!}{1!(k-2)!1!} = k(k-1)$ words, and finally in the third case $\frac{k!}{3!(k-3)!} = \frac{k(k-1)(k-2)}{6}$ words. For $k = 3$ the numbers are 3, 6, 1, as seen above.

Similarly, we count the combinations of $a_{k+1}, \ldots, a_n$. In this case we either choose two letters of the same type or two letters of different types. The first case produces $\frac{(n-k)!}{1!(n-k-1)!} = n - k$ words and the second case $\frac{(n-k)!}{2!(n-k-2)!} = \frac{(n-k)(n-k-1)}{2}$ words.

By combining our results we can calculate the size of the basic matrix for every k and n. We just regard the permutations of the letters $a_1, \ldots, a_k$ as the columns and $a_{k+1}, \ldots, a_n$ as the rows of the basic matrix, respectively. Thus we have the total of $k + k(k-1) + \frac{1}{6}k(k-1)(k-2)$ columns and $n - k + \frac{1}{2}(n-k)(n-k-1)$ rows. The largest matrix produces the most permutations and thus the maximal p. It suffices to find

$$\max_{k \leq n}[(k + k(k-1) + \frac{k(k-1)(k-2)}{6})(n - k + \frac{(n-k)(n-k-1)}{2})]$$

which equals

$$\max_{k \leq n}[\frac{k^5 + k^4(-2n+2) + k^3(n^2 - 5n - 1) + k^2(3n^2 - n - 2) + k(2n^2 + 2n)}{12}].$$

Since the first k letters have always more combinations, we need to consider only the values $k \geq \frac{n}{2}$.

Let us take, for example, $n = |\Sigma| = 5$ for which the largest matrix is obtained by choosing $k = 3$. The P-prime solutions in the basic matrix consist of i_1 a_1's, i_2 a_2's, i_3 a_3's, i_4 a_4's and i_5 a_5's for each two vectors (i_1, i_2, i_3) and

(i_4, i_5) obtained above. The number of P-prime solutions corresponding to each two basic vectors is obtained by multiplying the numbers of their combinations. Above we calculated that 3, 6, 1 are the combinations of vectors (i_1, i_2, i_3) and 2, 1 of vectors (i_4, i_5). Taking all together we have

$$3 \cdot 2 \cdot \frac{5!}{3!2!} + 3 \cdot 1 \cdot \frac{5!}{3!} + 6 \cdot 2 \cdot \frac{5!}{2!2!} + 6 \cdot 1 \cdot \frac{5!}{2!} + 1 \cdot 2 \cdot \frac{5!}{2!} + 1 \cdot 1 \cdot 5! = 1080$$

P-prime solutions.

4.4.4 Computing the maximal p

Our study has shown that if the cardinality of the range alphabet equals 1, the maximal number of P-prime solutions is effectively computable and an instance with the maximal number of P-prime solutions can be effectively constructed. Moreover, the time complexity of our algorithm is polynomial.

The following table shows the maximal k in some cases when the length l and the cardinality of Σ vary.

	$l = 3$	$l = 4$	$l = 5$	$l = 6$
$\|\Sigma\| = 2$	1	1	1	1
3	2	2	2	2
4	3	2	2	2
5	3	3	3	3
6	4	4	3	3
7	5	4	4	4
8	5	5	5	4
9	6	5	5	5
10	7	6	6	6
11	7	7	6	6
12	8	7	7	7
13	9	8	7	7
14	9	8	8	8
15	10	9	9	8
16	11	10	9	9

Since $k = \frac{n}{2}, \frac{n}{2}+1, \frac{n}{2}+2, \ldots$ when n is even, and $k = \frac{n+1}{2}, \frac{n+1}{2}+1, \frac{n+1}{2}+2, \ldots$ when n is odd we have underlined those cardinalities after which k changes (with respect to n). The following table shows the corresponding numbers of P-prime solutions, i.e., the maximal p in each case.

	$l=3$	$l=4$	$l=5$	$l=6$
$\vert\Sigma\vert=2$	3	10	35	126
3	12	80	560	4 032
4	27	320	4 480	64 512
5	54	1 080	22 680	489 888
6	96	2 560	76 545	2 480 058
7	150	5 760	241 920	10 450 944
8	225	11 250	590 625	33 030 144
9	324	20 000	1 400 000	100 800 000
10	441	34 560	2 903 040	250 822 656
11	588	54 880	5 670 000	612 360 000
12	768	85 750	10 504 375	1 323 551 250
13	972	128 000	18 151 560	2 744 515 872
14	1 215	184 320	30 965 760	5 350 883 328
15	1 500	262 440	49 601 160	9 913 171 968
16	1 815	360 000	78 764 805	17 863 857 774

We conclude this section with an interesting example. In order to show the priority of the periodic instances we have collected some values of a nonperiodic instance over a three-letter alphabet and compare them to the maximal values of the previous algorithm. We consider an instance (g,h) defined by

	1	2	3
g	$a^{l-1}b$	a	a^2
h	a^2	ba^{l-1}	a

For instance, if $l=5$ then the P-prime solutions are 112233, 311322, 313121232, 331112223 and 333311112222. In the following table the first column contains the number of P-prime solutions of this instance and the second column the corresponding maximal numbers with the same length l.

	$\vert P(g,h)\vert$	$\max p$
$l=3$	2	12
4	3	80
5	5	560
6	8	4 032
7	13	29 568
8	21	219 648
9	41	1 647 360
10	66	12 446 720
11	160	94 595 072
12	350	722 362 368
13	594	5 538 111 488
14	1 485	42 600 857 600
15	2 700	328 635 187 200

4.5 On test sets

An interesting aspect is to study how completely a set of prime solutions determines the equality set. F-prime solutions clearly have this property since any $w \in E(g,h)$ is obtained by catenating the solutions of $F(g,h)$.

For P-prime solutions the situation is not so clear. For instance, $P(g,h) = \{12, 21\}$ corresponds to both

$$E(g,h) = \{12, 21\}^+ \text{ and } E(g,h) = \{w \mid |w|_1 = |w|_2 \geq 1\}.$$

In a binary alphabet we know, however, that $|P(g,h)| \geq 3$ implies that (g,h) is periodic and hence $E(g,h)$ is unique. On the other hand, if (g,h) is not periodic then $p \leq s = f \leq 2$. Hence $F(g,h)$ is finite and again $E(g,h)$ is unique.

Thus in a binary alphabet the equality set has always a finite test set, $F(g,h)$ or $P(g,h)$, in the same sense as in Ehrenfeucht's conjecture, [31].

Bibliography

[1] M.H. Albert and J. Lawrence: A proof of Ehrenfeucht's conjecture, *Theoret. Comput. Sci.* **41** (1985) 121–123.

[2] K. Culik II: A purely homomorphic characterization of recursively enumerable sets, *J. Assoc. Comput. Mach.* **26** (1979) 345–350.

[3] K. Culik II, J. Karhumäki: On the equality sets for homomorphisms on free monoids with two generators, *RAIRO Inform. Théor.* **14** (1980) 349–369.

[4] K. Culik II, J. Karhumäki: Systems of equations over a free monoid and Ehrenfeucht's Conjecture, *Discrete Math.* **43** (1983) 139–153.

[5] K. Culik II, A. Salomaa: Test sets and checking words for homomorphic equivalence, *J. Comput. System Sci.* **20** (1980) 379–395.

[6] A. Ehrenfeucht, J. Karhumäki, G. Rozenberg: The (generalized) Post Correspondence Problem with lists of two words is decidable, *Theoret. Comput. Sci.* **21** (1982) 119–144.

[7] A. Ehrenfeucht, J. Karhumäki, G. Rozenberg: On binary equality sets and a solution to the test set conjecture in the binary case, *J. Algebra* **85** (1983) 76–85.

[8] A. Ehrenfeucht, G Rozenberg: Simplifications of homomorphisms, *Inform. and Control* **38** (1978) 298–305.

[9] A. Ehrenfeucht, G. Rozenberg: Elementary Homomorphisms and a solution to the DOL sequence equivalence problem, *Theoret. Comput. Sci.* **7** (1978) 169–183.

[10] T. Harju, M. Lipponen, A. Mateescu: Flatwords and Post Correspondence Problem, *Theoret. Comput. Sci.*, to appear.

[11] M.A. Harrison: *Introduction to Formal Language Theory*, Addison-Wesley (1978).

[12] G. Higman: Ordering by divisibility in abstract algebras, *Proc. London Math. Soc.* **2** (1952) 326–336.

[13] J. Karhumäki: On recent trends in formal language theory, *Proc. 14th ICALP, Springer LNCS* **267** (1987) 138–162.

[14] J. Karhumäki, I. Simon: A note on elementary homomorphisms and the regularity of equality sets, *EATCS Bull.* **9** (1979) 16–24.

[15] W. Kuich, A. Salomaa: *Semirings, Automata, Languages*, Springer-Verlag, Berlin, Heidelberg (1986).

[16] M. Lipponen: Primitive words and languages associated to PCP, *EATCS Bull.* **53** (1994) 217–226.

[17] M. Lipponen: Post Correspondence Problem: words possible as primitive solutions, *Proc. 22nd ICALP, Springer LNCS* **944** (1995) 63–74.

[18] M. Lipponen: On F-prime solutions of the Post Correspondence Problem, *2nd Internat. Conf. on Developments in Language Theory*, Magdeburg, 1995, to appear.

[19] M. Lipponen, Gh. Păun: Strongly prime PCP words, *Discrete Appl. Math.* **63** (1995) 193–197.

[20] M. Lothaire (ed.): *Combinatorics on Words*, Addison-Wesley (1983).

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

[22] G.S. Makanin: The problem of solvability of equations in a free semigroup (in Russian), *Mat. Sb.* **103** No. 145 (1977) 148–236.

[23] A. Mateescu, A. Salomaa: PCP-prime words and primality types, *RAIRO Inform. Théor.* **27** (1993) 57–70.

[24] A. Mateescu, A. Salomaa: On simplest possible solutions for Post Correspondence Problems, *Acta Inform.* **30** (1993) 441–457.

[25] A. Mateescu, A. Salomaa, K. Salomaa, Sheng Yu: P, NP and Post Correspondence Problem, *Inform. and Control* **121** (1995) 135–142.

[26] J.J. Pansiot: A note on Post's Correspondence Problem, *Inform. Proc. Lett.* **12** (1981) 233.

[27] E. Post: A variant of a recursively unsolvable problem, *Bull. Amer. Math. Soc.* **53** (1946) 264–268.

[28] G. Rozenberg, A. Salomaa: *Cornerstones of Undecidability*, Prentice Hall (1994).

[29] G. Rozenberg, A. Salomaa (ed.): *Handbook of Formal Languages, I–III*, Springer-Verlag, forthcoming.

[30] A. Salomaa: *Formal Languages*, Academic Press, New York (1973).

[31] A. Salomaa: The Ehrenfeucht conjecture: a proof for language theorists, *EATCS Bull.* **27** (1985) 71–82.

[32] A. Salomaa, K. Salomaa, Sheng Yu: Primality types of instances of the Post Correspondence Problem, *EATCS Bull.* **44** (1991) 226–241.

[33] H.J. Shyr: *Free Monoids and Languages*, Hon Min Book Company, Taichung (1991).