Preface

The theory of formal languages is widely accepted as the backbone of theoretical computer science. It mainly originated from mathematics (combinatorics, algebra, mathematical logic) and generative linguistics. Later, new specializations emerged from areas of either computer science (concurrent and distributed systems, computer graphics, artificial life), biology (plant development, molecular genetics), linguistics (parsing, text searching), or mathematics (cryptography). All human problem solving capabilities can be considered, in a certain sense, as a manipulation of symbols and structures composed by symbols, which is actually the stem of formal language theory. Language – in its two basic forms, natural and artificial – is a particular case of a symbol system.

This wide range of motivations and inspirations explains the diverse applicability of formal language theory Ű and all these together explain the very large number of monographs and collective volumes dealing with formal language theory.

In 2004 Springer-Verlag published the volume Formal Languages and Applications, edited by C. Martín-Vide, V. Mitrana and G. Păun in the series Studies in Fuzziness and Soft Computing 148, which was aimed at serving as an overall course-aid and self-study material especially for PhD students in formal language theory and applications. Actually, the volume emerged in such a context: it contains the core information from many of the lectures delivered to the students of the International PhD School in Formal Languages and Applications organized since 2002 by the Research Group on Mathematical Linguistics from Rovira i Virgili University, Tarragona, Spain.

During the editing process of the aforementioned volume, two situations appeared:

Some important aspects, mostly extensions and applications of classical formal language theory to different scientific areas, could not be covered, by different reasons. New courses were promoted in the next editions of the PhD School mentioned above.

To intend to fill up this gap, the volume Recent Advances in Formal Languages and Applications, edited by Z. Ésik, C. Martín-Vide and V. Mitrana, was published in 2006 by Springer-Verlag in the series Studies in Computational Intelligence 25.

The present volume is a continuation of this comprehensive publication effort. We believe that, besides accomplishing its main goal of complementing the previous volumes in representing a gate to formal language theory and its applications, it will be also useful as a general source of information in computation theory, both at the undergraduate and research level.

For the sake of uniformity, the introductory chapter of the first volume that presents the mathematical prerequisites as well as most common concepts and notations used throughout all chapters appears in the present volume as well. However, it may happen that terms other than those in the introductory chapter have different meanings in different chapters or different terms have the same meaning. In each chapter, the subject is treated relatively independent of the other chapters, even if several chapters are related. This way, the reader gets in touch with diverse points of view on an aspect common to two or more chapters. We are convinced of the usefulness of such an opportunity to a young researcher.

Acknowledgements

Our deep gratitude is due to all the contributors, for their professional and friendly cooperation, as well as to Springer-Verlag, for the efficient and pleasant collaboration.

Tarragona,
October 2007

Gemma Bel-Enguix
M. Dolores Jiménez-López
Carlos Martín-Vide

2

Open Problems on Partial Words*

Francine Blanchet-Sadri

Department of Computer Science, University of North Carolina
P.O. Box 26170, Greensboro, NC 27402–6170, USA
blanchet@uncg.edu

2.1 Introduction

Combinatorics on *words*, or sequences or strings of symbols over a finite alphabet, is a rather new field although the first papers were published at the beginning of the 20th century [120, 121]. The interest in the study of combinatorics on words has been increasing since it finds applications in various research areas of mathematics, computer science, and biology where the data can be easily represented as words over some alphabet. Such areas may be concerned with algorithms on strings [38, 48, 50, 51, 52, 69, 72, 84, 102, 118], semigroups, automata and languages [2, 45, 55, 75, 82, 92, 93], molecular genetics [78], or codes [5, 73, 79].

Motivated by molecular biology of nucleic acids, Berstel and Boasson introduced in 1999 the notion of *partial words* which are sequences that may contain a number of "do not know" symbols or "holes" [4]. *DNA molecules* are the carriers of the genetic information in almost all organisms. Let us look into the structure of such a molecule. A *single stranded DNA molecule* or a *DNA strand* may be viewed as a sequence over the alphabet consisting of the four *nucleotides*: a (adenine), c (cytosine), g (guanine), and t (thymine). Each strand has two different ends: the $3'$ end, and the $5'$ end. The familiar double helix of DNA, which was discovered by Watson and Crick, arises by the bonding of a strand in the $5' - 3'$ direction with another strand in the $3' - 5'$ direction with the restriction that adenine bonds with thymine, and cytosine bonds with guanine. Such a bonding gives rise to a *double stranded DNA molecule* as in the figure

$$5'\text{ - }c\ c\ a\ c\ c\ t\ c\ g\ a\ c\ c\ c\ t\ c\text{ - }3'$$
$$3'\text{ - }g\ g\ t\ g\ g\ a\ g\ c\ t\ g\ g\ g\ a\ g\text{ - }5'$$

Because of Watson-Crick's complementarity (a bonds to only t, and c bonds only to g), we can view double stranded DNA sequences as single stranded

*This material is based upon work supported by the National Science Foundation under Grant Nos. CCF–0207673 and DMS–0452020.

F. Blanchet-Sadri: *Open Problems on Partial Words*, Studies in Computational Intelligence (SCI) **113**, 11–58 (2008)
www.springerlink.com

strings by keeping the strand in the $5' - 3'$ direction. The molecule in the example above can be viewed as

$$5' \text{ - } ccacctcgaccctc \text{ - } 3'$$

or simply as $ccacctcgaccctc$, a word over the alphabet $\{a, c, g, t\}$. However bonding is not always perfect in nature as in the figure

$$\begin{array}{c} 5' \text{ - } c\ c\ a\ c\ c\ t\ c\ g\ a\ c\ c\ c\ t\ c \text{ - } 3' \\ 3' \text{ - } g\ g\ t\ t\ g\ a\ g\ c\ c\ g\ g\ g\ a\ g \text{ - } 5' \end{array}$$

where there is an occurrence of c paired with t, and an occurrence of a paired with c. In such a case, we can view the molecule as $5' - cca\diamond ctcg\diamond ccctc - 3'$ or as $cca\diamond ctcg\diamond ccctc$, where the $\diamond$'s stand for "do not know" symbols also called "holes". Thus, the latter example gives rise to a partial word with two holes over the alphabet $\{a, c, g, t\}$. Processes in molecular biology can be seen as operations on DNA sequences [72, 112]. If a set of DNA molecules fulfilling a certain property has changed a little bit after some time or under some influence, it is important to know whether the desired property still holds [91].

Several interesting combinatorial properties of partial words have been investigated and connections have been made with problems in graph theory and number theory, in particular, with problems concerning primitive sets of integers [23, 24], lattices [23, 24], partitions of integers and their generalizations [14], chromatic polynomials [23], Sudoku games [107], vertex connectivity in graphs [12, 29], etc. Partial words are useful in a new generation of pattern matching algorithms that search for local similarities between sequences. In this area, they are called "spaced seeds" and a lot of work has been dedicated to their influence on the algorithms' performance [40, 66, 83, 97, 103, 104]. Partial words have the potential for impacts in bio-inspired computing where they have been considered, in particular, for finding good encodings for DNA computations [90].

We provide here a few bibliographic remarks. Lothaire's first book *Combinatorics on Words* appeared in 1983 [92], while recent developments culminated in a second book *Algebraic Combinatorics on Words* which appeared in 2002 [93] and in a third book which appeared in 2005 [94]. Several books have appeared quite recently that emphasize connections of combinatorics on words to several research areas. We mention the book of Allouche and Shallit where the emphasis is on automata theory [2], the book of Crochemore and Rytter where the emphasis is on string algorithms [52], the book of Gusfield where the emphasis is on algorithms related to biology [72], the book of de Luca and Varrichio where the emphasis is on algebra [55], and finally the book of Blanchet-Sadri where the emphasis is on partial words [10].

Research in combinatorics on partial words is underway where there are several open problems that lay unexplored. After reviewing basic concepts on words and partial words in Section 2.2, we will discuss some of these open

problems which we have divided into sections: 2.3–2.5 study extensions to partial words of three basic classical results on periodicity of words: The theorem of Fine and Wilf which considers the simultaneous occurrence of different periods in one word [67], the critical factorization theorem which relates local and global periodicity of words [43], and a theorem of Guibas and Odlyzko which gives the structure of the set of periods of a word [71]. Section 2.6 deals with the two word properties of primitiveness and borderedness and is concerned, in particular, with the counting of primitive and unbordered partial words. Section 2.7 solves some equations on partial words. Here the notion of "equality" is replaced by that of "*compatibility*". Section 2.8 studies the concept of unavoidable set of partial words, while Section 2.9 develops square- and overlap-freeness of partial words. Finally, Section 2.10 discusses some other open problems related to codes of partial words, punctured languages, and tiling periodicity.

2.2 Preliminaries

This section is devoted to reviewing basic concepts on words and partial words.

2.2.1 Words

Let A be a nonempty finite set of symbols called an *alphabet*. Symbols in A are called *letters* and any finite sequence of letters is called a *word* over A. The *empty word*, that is, the word containing no letter, is denoted by ε. For any word u over A, $|u|$ denotes the number of letters occurring in u and is called the *length* of u. In particular, $|\varepsilon| = 0$. The set of all words over A is denoted by A^*. If we define the operation of two words u and v of A^* by juxtaposition (or concatenation), then A^* is a monoid with identity ε. We call $A^+ = A^* \setminus \{\varepsilon\}$ the *free semigroup generated by* A and A^* *the free monoid generated by* A. The set A^* can also be viewed as $\bigcup_{n\geq 0} A^n$ where $A^0 = \{\varepsilon\}$ and A^n is the set of all words of length n over A.

A word of length n over A can be defined by a total function $u : \{0, \ldots, n-1\} \to A$ and is usually represented as $u = a_0a_1 \ldots a_{n-1}$ with $a_i \in A$. A *period* of u is a positive integer p such that $a_i = a_{i+p}$ for $0 \leq i < n - p$. For a word u, the powers of u are defined inductively by $u^0 = \varepsilon$ and, for any $i \geq 1$, $u^i = uu^{i-1}$. The set of symbols occurring in a word u is denoted by $\alpha(u)$. The *reversal* of u, denoted by $\mathrm{rev}(u)$, is defined as follows: If $u = \varepsilon$, then $\mathrm{rev}(\varepsilon) = \varepsilon$, and if $u = a_0a_1 \ldots a_{n-1}$, then $\mathrm{rev}(u) = a_{n-1} \ldots a_1a_0$. A word u is a *factor* of the word v if there exist words x, y such that $v = xuy$. The factor u is called *proper* if $u \neq \varepsilon$ and $u \neq v$. The word u is a *prefix* (respectively, *suffix*) of v if $x = \varepsilon$ (respectively, $y = \varepsilon$).

A nonempty word u is *primitive* if there exists no word v such that $u = v^i$ with $i \geq 2$. Note the fact that the empty word is not primitive. If u is a nonempty word, then there exist a unique primitive word v and a unique positive integer i such that $u = v^i$.

2.2.2 Partial Words

A *partial word* u of length n over A is a partial function $u : \{0, \ldots, n-1\} \to A$. For $0 \leq i < n$, if $u(i)$ is defined, then i belongs to the *domain* of u, denoted by $i \in D(u)$, otherwise i belongs to the *set of holes* of u, denoted by $i \in H(u)$. A word over A is a partial word over A with an empty set of holes (we sometimes refer to words as *full* words). The length of u or n is denoted by $|u|$.

If u is a partial word of length n over A, then the *companion* of u, denoted by $u_\diamond$, is the total function $u_\diamond : \{0, \ldots, n-1\} \to A \cup \{\diamond\}$ defined by

$$u_\diamond(i) = \begin{cases} u(i) \text{ if } i \in D(u) \\ \diamond \quad \text{otherwise} \end{cases}$$

The bijectivity of the map $u \mapsto u_\diamond$ allows us to define for partial words concepts such as concatenation, powers, etc in a trivial way. The word $u_\diamond = abb\diamond bbcb$ is the companion of the partial word u of length $|u| = 8$ where $D(u) = \{0, 1, 2, 4, 5, 6, 7\}$ and $H(u) = \{3\}$. For convenience, we will refer to a partial word over A as a word over the enlarged alphabet $A_\diamond = A \cup \{\diamond\}$, where the additional symbol $\diamond$ plays the special role of a "do not know" symbol or "hole". This allows us to say for example "the partial word $aba\diamond aa\diamond$" instead of "the partial word with companion $aba\diamond aa\diamond$". The set of all partial words over A with an arbitrary number of holes is denoted by $A_\diamond^*$ which is a monoid under the operation of concatenation where ε serves as the identity.

A *(strong) period* of a partial word u over A is a positive integer p such that $u(i) = u(j)$ whenever $i, j \in D(u)$ and $i \equiv j \bmod p$. In such a case, we call u *(strongly) p-periodic*. Similarly, a *weak period* of u is a positive integer p such that $u(i) = u(i+p)$ whenever $i, i+p \in D(u)$. In such a case, we call u *weakly p-periodic*. The partial word $abb\diamond bbcbb$ is weakly 3-periodic but is not strongly 3-periodic. The latter shows a difference between partial words and full words since every weakly p-periodic full word is strongly p-periodic. Another difference worth noting is the fact that even if the length of a partial word u is a multiple of a weak period of u, then u is not necessarily a power of a shorter partial word. The minimum period of u is denoted by $p(u)$, and the minimum weak period by $p'(u)$. The set of all periods (respectively, weak periods) of u is denoted by $\mathcal{P}(u)$ (respectively, $\mathcal{P}'(u)$).

For a partial word u, positive integer p and integer $0 \leq i < p$, define

$$u_{i,p} = u(i)u(i+p)u(i+2p)\ldots u(i+jp)$$

where j is the largest nonnegative integer such that $i + jp < |u|$. Then u is (strongly) p-periodic if and only if $u_{i,p}$ is (strongly) 1-periodic for all $0 \leq i < p$, and u is weakly p-periodic if and only if $u_{i,p}$ is weakly 1-periodic for all $0 \leq i < p$. Strongly 1-periodic partial words as well as the full factors of weakly 1-periodic partial words are over a singleton alphabet.

If u and v are two partial words of equal length, then u is said to be contained in v, denoted by $u \subset v$, if all elements in $D(u)$ are in $D(v)$ and

$u(i) = v(i)$ for all $i \in D(u)$. The order $u \subset v$ on partial words is obtained when we let $\diamond < a$ and $a \leq a$ for all $a \in A$. For example, $a\diamond b\diamond \not\subset a\diamond\diamond b$ and $a\diamond b\diamond \not\subset a\diamond ab$, while $a\diamond b\diamond \subset a\diamond bb$.

A partial word u is *primitive* if there exists no word v such that $u \subset v^i$ with $i \geq 2$. Note that if v is primitive and $v \subset u$, then u is primitive as well. It was shown in [9] that if u is a nonempty partial word, then there exist a primitive word v and a positive integer i such that $u \subset v^i$. However uniqueness does not hold as seen with the partial word $u = \diamond a$ where $u \subset a^2$ and $u \subset ba$ for distinct letters a, b.

Partial words u and v are *compatible*, denoted by $u \uparrow v$, if there exists a partial word w such that $u \subset w$ and $v \subset w$. In other words, $u(i) = v(i)$ for every $i \in D(u) \cap D(v)$. Note that for full words, the notion of compatibility is simply that of equality. For example, $a\diamond b\diamond a\diamond \not\uparrow a\diamond\diamond cbb$ but $a\diamond bbc\diamond \uparrow \diamond bb\diamond c\diamond$.

In the rest of this section, we discuss commutativity and conjugacy in the context of partial words.

Let us start with commutativity. The case of full words is well known and is stated in the following theorem.

Theorem 1. *Let x and y be nonempty words. Then $xy = yx$ if and only if there exists a word z such that $x = z^m$ and $y = z^n$ for some integers m, n.*

For nonempty partial words x and y, if there exist a word z and integers m, n such that $x \subset z^m$ and $y \subset z^n$, then $xy \subset z^{m+n}$, $yx \subset z^{m+n}$, and $xy \uparrow yx$. The converse is not true in general: if $x = \diamond bb$ and $y = abb\diamond$, then

$$xy = \diamond bbabb\diamond \uparrow abb\diamond\diamond bb = yx$$

but no desired z exists.

Let us first examine the case of one hole.

Theorem 2. *[4] Let x, y be nonempty partial words such that xy has at most one hole. If $xy \uparrow yx$, then there exists a word z such that $x \subset z^m$ and $y \subset z^n$ for some integers m, n.*

Now, for the case of an arbitrary number of holes, let k, l be positive integers satisfying $k \leq l$. For $0 \leq i < k + l$, define

$$\mathrm{seq}_{k,l}(i) = (i_0, i_1, i_2, \ldots, i_n, i_{n+1})$$

where $i_0 = i = i_{n+1}$; for $1 \leq j \leq n$, $i_j \neq i$; and for $1 \leq j \leq n + 1$,

$$i_j = \begin{cases} i_{j-1} + k \text{ if } i_{j-1} < l \\ i_{j-1} - l \text{ otherwise} \end{cases}$$

For example, $\mathrm{seq}_{6,8}(0) = (0, 6, 12, 4, 10, 2, 8, 0)$.

Definition 1. *[11] Let k, l be positive integers satisfying $k \leq l$ and let z be a partial word of length $k + l$. We say that z is (k, l)-special if there exists $0 \leq i < k$ such that $seq_{k,l}(i) = (i_0, i_1, i_2, \ldots, i_n, i_{n+1})$ contains (at least) two positions that are holes of z while $z(i_0)z(i_1)z(i_2)\ldots z(i_{n+1})$ is not 1-periodic.*

Example 1. Let $z = cbca\diamond\diamond cbc\diamond caca$, and let $k = 6$ and $l = 8$ so $|z| = k+l$. We wish to determine if z is $(6, 8)$-special. We already calculated $\text{seq}_{6,8}(0)$ and

$$\begin{array}{cccccccc} z(0) & z(6) & z(12) & z(4) & z(10) & z(2) & z(8) & z(0) \\ c & c & c & \diamond & c & c & c & c \end{array}$$

This sequence does not satisfy the definition, and so we must continue with calculating $\text{seq}_{6,8}(1) = (1, 7, 13, 5, 11, 3, 9, 1)$. The corresponding letter sequence is

$$\begin{array}{cccccccc} z(1) & z(7) & z(13) & z(5) & z(11) & z(3) & z(9) & z(1) \\ b & b & a & \diamond & a & a & \diamond & b \end{array}$$

Here we have two positions in the sequence which are holes, and the sequence is not 1-periodic. Hence, z is $(6, 8)$-special.

Under the extra condition that xy is not $(|x|, |y|)$-special, an extension of Theorem 2 holds when xy has an arbitrary number of holes.

Theorem 3. *[11] Let x, y be nonempty partial words such that $|x| \leq |y|$. If $xy \uparrow yx$ and xy is not $(|x|, |y|)$-special, then there exists a word z such that $x \subset z^m$ and $y \subset z^n$ for some integers m, n.*

Now, let us discuss conjugacy. Again, the case of full words is well known.

Theorem 4. *Let x, y, z ($x \neq \varepsilon$ and $y \neq \varepsilon$) be words such that $xz = zy$. Then $x = uv$, $y = vu$, and $z = (uv)^n u$ for some words u, v and integer $n \geq 0$.*

For example, if $x = abcda$, $y = daabc$, and $z = abc$, then $xz = zy$ because $(abcda)(abc) = (abc)(daabc)$. Here $u = abc$, $v = da$, and $n = 0$.

The case of partial words is more subtle.

Theorem 5. *[28] Let x, y, z be partial words with x, y nonempty. If $xz \uparrow zy$ and $xz \vee zy$ is $|x|$-periodic, then there exist words u, v such that $x \subset uv$, $y \subset vu$, and $z \subset (uv)^n u$ for some integer $n \geq 0$.*

To illustrate Theorem 5, let $x = \diamond ba$, $y = \diamond b\diamond$, and $z = b\diamond ab\diamond\diamond\diamond$. Then we have

$$\begin{aligned} xz &= \diamond\ b\ a\ b\ \diamond\ a\ b\ \diamond\ \diamond\ \diamond\ \diamond \\ zy &= b\ \diamond\ a\ b\ \diamond\ \diamond\ \diamond\ \diamond\ \diamond\ b\ \diamond \\ xz \vee zy &= b\ b\ a\ b\ \diamond\ a\ b\ \diamond\ \diamond\ b\ \diamond \end{aligned}$$

It is clear that $xz \uparrow zy$ and $xz \vee zy$ is $|x|$-periodic. Putting $u = bb$ and $v = a$, we can verify that the conclusion does indeed hold.

Corollary 1. *[28] Let x, y be nonempty partial words, and let z be a full word. If $xz \uparrow zy$, then there exist words u, v such that $x \subset uv$, $y \subset vu$, and $z \subset (uv)^n u$ for some integer $n \geq 0$.*

Note that the above Corollary does not necessarily hold if z is not full even if x, y are full. The partial words $x = a, y = b$, and $z = \diamond bb$ provide a counterexample.

Two conjugacy theorems follow without any restriction on z.

Theorem 6. *[13] Let x, y and z be partial words such that $|x| = |y| > 0$. Then $xz \uparrow zy$ if and only if xzy is weakly $|x|$-periodic.*

Theorem 7. *[13]*

Let x, y and z be partial words such that $|x| = |y| > 0$. Then the following hold:

1. *If $xz \uparrow zy$, then xz and zy are weakly $|x|$-periodic.*
2. *If xz and zy are weakly $|x|$-periodic and $\lfloor \frac{|z|}{|x|} \rfloor > 0$, then $xz \uparrow zy$.*

The assumption $\lfloor \frac{|z|}{|x|} \rfloor > 0$ is necessary. To see this, consider $x = aa$, $y = ba$ and $z = a$. Here, xz and zy are weakly $|x|$-periodic, but $xz \not\uparrow zy$.

2.3 Periods in Partial Words

Notions and techniques related to periodic structures in words find important applications in virtually every area of theoretical and applied computer science, notably in text processing [51, 52], data compression [49, 119, 123], coding [5], computational biology [39, 72, 100, 112], string searching and pattern matching algorithms [38, 50, 51, 52, 69, 72, 84, 102]. Repeated patterns and related phenomena in words have played over the years a central role in the development of combinatorics on words, and have been highly valuable tools for the design and analysis of algorithms [45, 92, 93, 94]. In many practical applications, such as DNA sequence analysis, repetitions admit a certain variation between copies of the repeated pattern because of errors due to mutation, experiments, etc. Approximate repeated patterns, or repetitions where errors are allowed, are playing a central role in different variants of string searching and pattern matching problems [85, 86, 87, 88, 111]. Partial words have acquired great importance in this context.

The notion of *period* is central in combinatorics on words and there are many fundamental results on periods of words. Among them is the well known periodicity result of Fine and Wilf [67] which intuitively determines how far two periodic events have to match in order to guarantee a common period. More precisely, any word u having two periods p, q and length at least $p + q - \gcd(p, q)$ has also the greatest common divisor of p and q, $\gcd(p, q)$, as a period. The bound $p + q - \gcd(p, q)$ is optimal since counterexamples can be

provided for shorter lengths, that is, there exists an *optimal* word of length $p + q - \gcd(p, q) - 1$ having p and q as periods but not having $\gcd(p, q)$ as period [45]. Extensions of Fine and Wilf's result to more than two periods are given in [42, 47, 80, 122]. For instance, in [47], Constantinescu and Ilie give an extension for an arbitrary number of periods and prove that their bounds are optimal.

Fine and Wilf's result has been generalized to partial words in two ways:

- First, any partial word u with h holes and having two weak periods p, q and length at least the so-denoted $l(h, p, q)$ has also strong period $\gcd(p, q)$ provided u satisfies the condition of not being (h, p, q)-*special* (this concept will be defined below). This extension was done for one hole by Berstel and Boasson where the class of $(1, p, q)$-special partial words is empty [4]; for two or three holes by Blanchet-Sadri and Hegstrom [25]; and for an arbitrary number of holes by Blanchet-Sadri [8]. Elegant closed formulas for the bounds $l(h, p, q)$ were given and shown to be optimal.
- Second, any partial word u with h holes and having two strong periods p, q and length at least the so-denoted $L(h, p, q)$ has also strong period $\gcd(p, q)$. The study of the bounds $L(h, p, q)$ was initiated by Shur and Gamzova [114]. In particular, they gave a closed formula for $L(h, p, q)$ in the case where $h = 2$ (the cases where $h = 0$ or $h = 1$ are implied by the above mentioned results). In [12], Blanchet-Sadri, Bal and Sisodia gave closed formulas for the optimal bounds $L(h, p, q)$ in the case where $p = 2$ and also in the case where q is "large". In addition, they gave upper bounds when q is "small" and $h = 3, 4, 5$ or 6. Their proofs are based on connectivity in a graph $G_{(p,q)}(u)$ associated with a given p- and q-periodic partial word u. More recently, in [29], Blanchet-Sadri, Mandel and Sisodia pursue by studying two types of vertex connectivity on $G_{(p,q)}(u)$: the so-called modified degree connectivity and r-set connectivity where $r = q \bmod p$. As a result, they give an efficient algorithm for computing $L(h, p, q)$, and manage to give closed formulas in several cases including the $h = 3$ and $h = 4$ cases.

In this section, we discuss in details the two ways Fine and Wilf's periodicity result has been extended to partial words: Section 2.3.1 discusses the weak periodicity generalizations and Section 2.3.2 the strong periodicity generalizations. For easy reference, we recall Fine and Wilf's result.

Theorem 8. *[67]*

Let p and q be positive integers. If a full word u is p-periodic and q-periodic and $|u| \geq p + q - \gcd(p, q)$, then u is $gcd(p, q)$-periodic.

2.3.1 Weak Periodicity

In this section, we review the generalizations related to weak periodicity.

We first recall Berstel and Boasson's result for partial words with exactly one hole where the bound $p + q$ is optimal.

Theorem 9. *[4]*

Let p and q be positive integers satisfying $p < q$. Let u be a partial word with one hole. If u is weakly p-periodic and weakly q-periodic and $|u| \geq l(1, p, q) = p + q$, then u is strongly $\gcd(p, q)$*-periodic.*

When we discuss partial words with $h \geq 2$ holes, we need the extra assumption of u not being (h, p, q)-*special* for a similar result to hold true. Indeed, if p and q are positive integers satisfying $p < q$ and $\gcd(p, q) = 1$, then the infinite sequence $(ab^{p-1}\diamond b^{q-p-1}\diamond b^n)_{n>0}$ consists of $(2, p, q)$-special partial words with two holes that are weakly p-periodic and weakly q-periodic but not $\gcd(p, q)$-periodic.

In order to define the concept of (h, p, q)-speciality, note that a partial word u that is weakly p-periodic and weakly q-periodic can be represented as a 2-dimensional structure. Consider for example the partial word

$$w = ababa\diamond\diamond\diamond bab\diamond bb\diamond bbbbbbbbb$$

where $p = 2$ and $q = 5$. The array looks like:

$$\begin{array}{ccccc} u(0) & u(5) & u(10) & u(15) & u(20) \\ u(2) & u(7) & u(12) & u(17) & u(22) \\ u(4) & u(9) & u(14) & u(19) & \\ u(1) & u(6) & u(11) & u(16) & u(21) \\ u(3) & u(8) & u(13) & u(18) & u(23) \end{array}$$

and its corresponding array of symbols looks like:

$$\begin{array}{ccccc} a & \diamond & b & b & b \\ a & \diamond & b & b & b \\ a & a & \diamond & b & \\ b & \diamond & \diamond & b & b \\ b & b & b & b & b \end{array}$$

In general, if $\gcd(p, q) = d$, we get d arrays. Each of these arrays is associated with a subgraph $G = (V, E)$ of $G_{(p,q)}(u)$ as follows: V is the subset of $D(u)$ comprising the defined positions of u within the array, and $E = E_p \cup E_q$ where $E_p = \{\{i, i-p\} \mid i, i-p \in V\}$ and $E_q = \{\{i, i-q\} \mid i, i-q \in V\}$. For $0 \leq j < \gcd(p, q)$, the subgraph of $G_{(p,q)}(u)$ corresponding to

$$D(u) \cap \{i \mid i \geq 0 \text{ and } i \equiv j \bmod \gcd(p, q)\}$$

will be denoted by $G^j_{(p,q)}(u)$. Whenever $\gcd(p, q) = 1$, $G^0_{(p,q)}(u)$ is just $G_{(p,q)}(u)$. Referring to the partial word w above, the graph $G_{(2,5)}(w)$ is disconnected (w is $(5, 2, 5)$-special). Here, the $\diamond$'s isolate the a's from the b's.

We now define the concept of speciality.

Definition 2. *[8]*
Let p and q be positive integers satisfying $p < q$, and let h be a nonnegative integer. Let

$$l(h,p,q) = \begin{cases} (\frac{h}{2}+1)(p+q) - \gcd(p,q) & \text{if } h \text{ is even} \\ (\lfloor \frac{h}{2} \rfloor + 1)(p+q) & \text{otherwise} \end{cases}$$

Let u be a partial word with h holes of length at least $l(h,p,q)$. Then u is (h,p,q)-special if $G^j_{(p,q)}(u)$ is disconnected for some $0 \leq j < \gcd(p,q)$.

It turns out that the bound $l(h,p,q)$ is optimal for a number of holes h.

Theorem 10. *[8]*
Let p and q be positive integers satisfying $p < q$, and let u be a non (h,p,q)-special partial word with h holes. If u is weakly p-periodic and weakly q-periodic and $|u| \geq l(h,p,q)$, then u is strongly $gcd(p,q)$-periodic.

In [33], progress was made towards allowing an arbitrary number of holes and an arbitrary number of weak periods. There, the authors proved that any partial word u with h holes and having weak periods $p_1, \ldots, p_n$ and length at least the so-denoted $l(h,p_1,\ldots,p_n)$ has also strong period $\gcd(p_1,\ldots,p_n)$ provided u satisfies some criteria. In addition to speciality, they discovered that the concepts of *intractable* period sets and *interference* between periods play a role.

Open problem 1 *Give an algorithm which given a number of holes h and weak periods $p_1,\ldots,p_n$, computes the optimal bound $l(h,p_1,\ldots,p_n)$ and an optimal partial word for that bound (a partial word u with h holes of length $l(h,p_1,\ldots,p_n)-1$ is optimal for the bound $l(h,p_1,\ldots,p_n)$ if $p_1,\ldots,p_n$ are weak periods of u but $\gcd(p_1,\ldots,p_n)$ is not a strong period of u).*

Open problem 2 *Give closed formulas for the bounds $l(h,p_1,\ldots,p_n)$.*

The optimality proof will probably be based on results of graphs associated with bounds and tuples of weak periods.

2.3.2 Strong Periodicity

In this section, we review the generalizations related to strong periodicity. There exists an integer L such that if a partial word u with h holes has strong periods p,q satisfying $p < q$ and $|u| \geq L$, then u has strong period $\gcd(p,q)$ ($L(h,p,q)$ is the smallest such integer L) [115]. Recall that $L(0,p,q) = p + q - \gcd(p,q)$.

The following result is a direct consequence of Berstel and Boasson's result.

Theorem 11. *[4] The equality $L(1,p,q) = p+q$ holds.*

For $h = 2, 3$ or 4, we have the following results.

Theorem 12. *[114, 115] The equality* $L(2,p,q) = 2p + q - \gcd(p,q)$ *holds.*

Theorem 13. *[29] The following equality holds:*

$$L(3,p,q) = \begin{cases} 2q+p & \text{if } q-p < \frac{p}{2} \\ 4p & \text{if } \frac{p}{2} < q-p < p \\ 2p+q & \text{if } p < q-p \end{cases}$$

Theorem 14. *[29] The following equality holds:*

$$L(4,p,q) = \begin{cases} q+3p-\gcd(p,q) & \text{if } q-p < \frac{p}{2} \\ q+3p & \text{if } \frac{p}{2} < q-p < p \\ q+3p-\gcd(p,q) & \text{if } p < q-p \end{cases}$$

Other results follow.

Theorem 15. *[12, 113, 114, 115] The equality* $L(h,2,q) = (2\lfloor \frac{h}{q} \rfloor + 1)q + h \bmod q + 1$ *holds.*

Setting $h = nq+r$ where $0 \leq r < q$, $L(h,2,q) = (2n+1)q+r+1$. Consider the word $u = \diamond^r w(\diamond^q w)^n$ where w is the unique full word of length q having periods 2 and q but not having period 1. Note that u is an optimal word for the bound $L(h,2,q)$. Indeed, $|u| = (2n+1)q + r$, u has h holes, and since w is not 1-periodic, we also have that u is not strongly 1-periodic. It is easy to show that u is strongly 2- and q-periodic.

In [114], the authors proved that if $q > p \geq 3$ and $\gcd(p,q) = 1$ and h is large enough, then

$$\tfrac{pq}{p+q-2}(h+1) \leq L(h,p,q) < \tfrac{pqh}{p+q-2} + 4(q-1)$$

Open problem 3 *Give closed formulas for the bounds* $L(h,p,q)$ *where* $h > 4$.

Any partial word with h holes and having n strong periods $p_1, \ldots, p_n$ and length at least the so-denoted $L(h,p_1,\ldots,p_n)$ has also $\gcd(p_1,\ldots,p_n)$ as a strong period.

Open problem 4 *Give an algorithm which given a number of holes* h *and strong periods* $p_1,\ldots,p_n$*, computes the optimal bound* $L(h,p_1,\ldots,p_n)$ *and an optimal partial word for that bound (a partial word* u *with* h *holes of length* $L(h,p_1,\ldots,p_n)-1$ *is optimal for the bound* $L(h,p_1,\ldots,p_n)$ *if* $p_1,\ldots,p_n$ *are strong periods of* u *but* $\gcd(p_1,\ldots,p_n)$ *is not a strong period of* u*).*

Open problem 5 *Give closed formulas for the bounds* $L(h,p_1,\ldots,p_n)$.

2.4 Critical Factorizations of Partial words

Results concerning periodicity include the well known and fundamental critical factorization theorem, of which several versions exist [43, 45, 60, 61, 59, 92, 93]. It intuitively states that the minimal period (or global period) of a word of length at least two is always locally detectable in at least one position of the word resulting in a corresponding *critical factorization*. More specifically, given a word w and nonempty words u, v satisfying $w = uv$, the *minimal local period* associated to the factorization (u, v) is the length of the shortest square at position $|u|-1$. It is easy to see that no minimal local period is longer than the global period of the word. The critical factorization theorem shows that critical factorizations are unavoidable. Indeed, for any word, there is always a factorization whose minimal local period is equal to the global period of the word.

More precisely, we consider a word $a_0a_1 \dots a_{n-1}$ and, for any integer i ($0 \leq i < n-1$), we look at the shortest repetition (a *square*) *centered* in this position, that is, we look at the shortest (virtual) suffix of $a_0a_1 \dots a_i$ which is also a (virtual) prefix of $a_{i+1}a_{i+2} \dots a_{n-1}$. The minimal local period at position i is defined as the length of this shortest square. The critical factorization theorem states, roughly speaking, that the global period of $a_0a_1 \dots a_{n-1}$ is simply the maximum among all minimal local periods. As an example, consider the word $w = babbaab$ with global period 6. The minimal local periods of w are 2, 3, 1, 6, 1 and 3 which means that the factorization $(babb, aab)$ is critical.

Crochemore and Perrin showed that a critical factorization can be found very efficiently from the computation of the maximal suffixes of the word with respect to two total orderings on words: the lexicographic ordering related to a fixed total ordering on the alphabet $\preceq_l$, and the lexicographic ordering obtained by reversing the order of letters in the alphabet $\preceq_r$ [50]. If v denotes the maximal suffix of w with respect to $\preceq_l$ and v' the maximal suffix of w with respect to $\preceq_r$, then let u, u' be such that $w = uv = u'v'$. The factorization (u, v) turns out to be critical when $|v| \leq |v'|$, and the factorization (u', v') is critical when $|v| > |v'|$. There exist linear time (in the length of w) algorithms for such computations [50, 51, 101] (the latter two use the suffix tree construction). Returning to the example above, order the letters of the alphabet by $a \prec b$. Then the maximal suffix with respect to $\preceq_l$ is $v = bbaab$ and with respect to $\preceq_r$ is $v' = aab$. Since $|v| > |v'|$, the factorization $(u', v') = (babb, aab)$ of w is critical.

In [22], Blanchet-Sadri and Duncan extended the critical factorization theorem to partial words with one hole. In this case, the concept of *local period*, which characterizes a local periodic structure at each position of the word, is defined as follows.

Definition 3. *[22] Let w be a nonempty partial word. A positive integer p is called* a local period *of w at position i if there exist partial words u, v, x, y such that $u, v \neq \varepsilon$, $w = uv$, $|u| = i + 1$, $|x| = p$, $x \uparrow y$, and such that one of the following conditions holds for some partial words r, s:*

1. $u = rx$ *and* $v = ys$ *(internal square),*
2. $x = ru$ *and* $v = ys$ *(left-external square if* $r \neq \varepsilon$*),*
3. $u = rx$ *and* $y = vs$ *(right-external square if* $s \neq \varepsilon$*),*
4. $x = ru$ *and* $y = vs$ *(left- and right-external square if* $r, s \neq \varepsilon$*).*

In this context, a factorization is called *critical* if its minimal local period is equal to the minimal weak period of the partial word. As an example, consider the partial word with one hole $w = ba\diamond baab$ with minimal weak period 3. The minimal local periods of w are 2 (left-external square), 1 (internal square), 1 (internal square), 3 (internal square), 1 (internal square) and 3 (right-external square), and both $(ba\diamond b, aab)$ and $(ba\diamond baa, b)$ are critical.

It turns out that for partial words, critical factorizations may be avoidable. Indeed, the partial word *babdaab* has no critical factorization. The class of the so-called *special* partial words with one hole has been described that possibly avoid critical factorizations. Refining the method based on the maximal suffixes with respect to the lexicographic/ reverse lexicographic orderings leads to a version of the critical factorization theorem for the *nonspecial* partial words with one hole whose proof provides an efficient algorithm which, given a partial word with one hole, outputs a critical factorization when one exists or outputs "no such factorization exists".

In [35], Blanchet-Sadri and Wetzler further investigated the relationship between local and global periodicity of partial words: (1) They extended the critical factorization theorem to partial words with an arbitrary number of holes; (2) They characterized precisely the class of partial words that do not admit critical factorizations; and (3) They developed an efficient algorithm which computes a critical factorization when one exists.

Some open problems related to the critical factorization theorem follow.

Open problem 6 *Discover some good criterion for the existence of a critical factorization of an unbordered partial word defined as follows: A nonempty partial word* u *is unbordered if no nonempty partial words* x, v, w *exist such that* $u \subset xv$ *and* $u \subset wx$*.*

Open problem 7 *In the framework of partial words, study the periodicity theorem on words, which has strong analogies with the critical factorization theorem, that was derived in [102].*

In [62], the authors present an $O(n)$ time algorithm for computing *all* local periods of a given word of length n, assuming a constant-size alphabet. This subsumes (but is substantially more powerful than) the computation of the global period of the word and the computation of a critical factorization. Their method consists of two parts: (1) They show how to compute all *internal* minimal squares; and (2) They show how to compute *left-* and *right-external* minimal squares, in particular for those positions for which no internal square has been found.

Open problem 8 *Find the time complexity for the computation of* all *the local periods of a given partial word.*

Now, consider the language

$$CF = \{w \mid w \text{ is a partial word over } \{a, b\} \text{ that has a critical factorization}\}$$

What is the position of CF in the Chomsky hierarchy? It has been proved that CF is a context sensitive language that is not regular. The question whether or not CF is context-free remains open.

Theorem 16. *[36] The language CF is not regular.*

Theorem 17. *[21] The language CF is context sensitive.*

Open problem 9 *Is the language CF context-free?*

2.5 Correlations of Partial Words

In [71], Guibas and Odlyzko consider the period sets of words of length n over a finite alphabet, and specific representations of them, *(auto)correlations*, which are binary vectors of length n indicating the periods. Among the possible 2^n bit vectors, only a small subset are valid correlations. There, they provide characterizations of correlations, asymptotic bounds on their number, and a recurrence for the *population size* of a correlation, that is, the number of words sharing a given correlation. In [108], Rivals and Rahmann show that there is redundancy in period sets and introduce the notion of an *irreducible* period set. They prove that Γ_n, the set of all correlations of length n, is a lattice under set inclusion and does not satisfy the Jordan-Dedekind condition. They propose the first efficient enumeration algorithm for Γ_n and improve upon the previously known asymptotic lower bounds on the cardinality of Γ_n. Finally, they provide a new recurrence to compute the number of words sharing a given period set, and exhibit an algorithm to sample uniformly period sets through irreducible period sets.

In [24], the combinatorics of possible sets of periods and weak periods of partial words were studied in a similar way. There, the notions of binary and ternary correlations were introduced, which are binary and ternary vectors indicating the periods and weak periods of partial words. Extending the result of Guibas and Odlyzko, Blanchet-Sadri, Gafni and Wilson characterized precisely which binary and ternary vectors represent the period and weak period sets of partial words and proved that all valid correlations may be taken over the binary alphabet (the one-hole case was proved earlier in [16]). They showed that the sets of all such vectors of a given length form distributive lattices under suitably defined partial orderings extending results of Rivals and Rahmann. They also showed that there is a well defined minimal set of generators for any binary correlation of length n, called an *irreducible* period

set, and demonstrated that these generating sets are the primitive subsets of $\{1, 2, \ldots, n-1\}$. These primitive sets of integers have been extensively studied by many researchers including Erdös [65]. Finally, they investigated the number of partial word correlations of length n. More recently, recurrences for computing the size of populations of partial word correlations were obtained as well as random sampling of period and weak period sets [23].

We first define the *greatest lower bound* of two given partial words u and v of equal length as the partial word $u \wedge v$, where $(u \wedge v) \subset u$ and $(u \wedge v) \subset v$, and if $w \subset u$ and $w \subset v$, then $w \subset (u \wedge v)$. The following example illustrates this new concept which plays a role in this section:

$$\begin{array}{rl} u & = a\ b \diamond c\ a\ a\ b \diamond \diamond a\ a \\ v & = a\ c\ b\ c\ a\ a\ b \diamond b\ b\ a \\ \hline u \wedge v & = a \diamond \diamond c\ a\ a\ b \diamond \diamond \diamond a \end{array}$$

The contents of Section 2.5 is as follows: In Section 2.5.1, we give characterizations of correlations. In Section 2.5.2, we provide structural properties of correlations. And in Section 2.5.3, we consider the problem of counting correlations.

2.5.1 Characterizations of Correlations

Full word correlations are vectors representing sets of periods as stated in the following definition.

Definition 4. *Let u be a (full) word. Let v be the binary vector of length $|u|$ for which $v_0 = 1$ and*

$$v_i = \begin{cases} 1 & \text{if } \ i \in \mathcal{P}(u) \\ 0 & \text{otherwise} \end{cases}$$

We call v the correlation of u.

For instance, the word *abbababbab* has periods 5 and 8 (and 10) and thus has correlation 1000010010.

Binary vectors may satisfy some propagation rules.

Definition 5. *1. A binary vector v of length n is said to satisfy the forward propagation rule if for all $0 \leq p < q < n$ such that $v_p = v_q = 1$ we have that $v_{p+i(q-p)} = 1$ for all $2 \leq i < \frac{n-p}{q-p}$.*

2. A binary vector v of length n is said to satisfy the backward propagation rule if for all $0 \leq p < q < \min(n, 2p)$ such that $v_p = v_q = 1$ and $v_{2p-q} = 0$ we have that $v_{p-i(q-p)} = 0$ for all $2 \leq i \leq \min(\lfloor \frac{p}{q-p} \rfloor, \lfloor \frac{n-p}{q-p} \rfloor)$.

Note that a binary vector v of length 12 satisfying $v_7 = v_9 = 1$ and the forward propagation rule also satisfies $v_{7+2(9-7)} = v_{11} = 1$. Note also that setting $p = 0$ in the forward propagation rule implies that $v_{iq} = 1$ for all i whenever $v_q = 1$.

Fundamental results on periodicity of words include the following unexpected result of Guibas and Odlyzko which gives a characterization of full word correlations.

Theorem 18. *[71] For correlation v of length n the following are equivalent:*

1. *There exists a word over the binary alphabet with correlation v.*
2. *There exists a word over some alphabet with correlation v.*
3. *The correlation v satisfies the forward and backward propagation rules.*

Corollary 2. *[71] For any word u over an alphabet A, there exists a binary word v of length $|u|$ such that $\mathcal{P}(v) = \mathcal{P}(u)$.*

Now, partial word correlations are defined according to the following definition.

Definition 6. *[24]*

1. *The binary correlation of a partial word u satisfying $\mathcal{P}(u) = \mathcal{P}'(u)$ is the binary vector of length $|u|$ such that $v_0 = 1$ and*

$$v_i = \begin{cases} 1 & \text{if } i \in \mathcal{P}(u) \\ 0 & \text{otherwise} \end{cases}$$

2. *The ternary correlation of a partial word u is the ternary vector of length $|u|$ such that $v_0 = 1$ and*

$$v_i = \begin{cases} 1 & \text{if } i \in \mathcal{P}(u) \\ 2 & \text{if } i \in \mathcal{P}'(u) \setminus \mathcal{P}(u) \\ 0 & \text{otherwise} \end{cases}$$

Considering the partial word $abaca\diamond\diamond acaba$ which has periods 9 and 11 (and 12) and strictly weak period 5, its ternary correlation vector is 100002000101.

A characterization of binary correlations follows.

Theorem 19. *[24] Let n be a nonnegative integer. Then for any finite collection $u_1, u_2, \ldots, u_k$ of full words of length n over an alphabet A, there exists a partial word w of length n over the binary alphabet with $\mathcal{P}(w) = \mathcal{P}'(w) = \mathcal{P}(u_1) \cup \mathcal{P}(u_2) \cup \cdots \cup \mathcal{P}(u_k)$.*

Corollary 3. *[24] The set of valid binary correlations over an alphabet A with $\|A\| \geq 2$ is the same as the set of valid binary correlations over the binary alphabet. Phrased differently, if u is a partial word over an alphabet A, then there exists a binary partial word v of length $|u|$ such that $\mathcal{P}(v) = \mathcal{P}(u)$.*

Follows is a characterization of valid ternary correlations.

Theorem 20. *[24] A ternary vector v of length n is the ternary correlation of a partial word of length n over an alphabet A if and only if $v_0 = 1$ and*

1. *If $v_p = 1$, then for all $0 \leq i < \frac{n}{p}$ we have that $v_{ip} = 1$.*
2. *If $v_p = 2$, then there exists some $2 \leq i < \frac{n}{p}$ such that $v_{ip} = 0$.*

The proof is based on the following construction: For $n \geq 3$ and $0 < p < n$, let $n = kp + r$ where $0 \leq r < p$. Then define

$$\omega_p = \begin{cases} (ab^{p-1})^k & \text{if } r = 0 \\ (ab^{p-1})^k ab^{r-1} & \text{if } r > 0 \end{cases}$$

$$\psi_p = ab^{p-1} \diamond b^{n-p-1}$$

Then given a valid ternary correlation v of length n, the partial word

$$\left(\bigwedge_{p>0|v_p=1} \omega_p\right) \wedge \left(\bigwedge_{p|v_p=2} \psi_p\right)$$

has ternary correlation v.

For example, given $v = 100002000101$, then $abbbb\diamond bbb\diamond b\diamond$ has correlation v as computed in the following figure:

$$\begin{array}{rl} \omega_9 & = a\ b\ b\ b\ b\ b\ b\ b\ b\ a\ b\ b \\ \omega_{11} & = a\ b\ b\ b\ b\ b\ b\ b\ b\ b\ b\ a \\ \psi_5 & = a\ b\ b\ b\ b \diamond b\ b\ b\ b\ b\ b \\ \hline & a\ b\ b\ b\ b \diamond b\ b\ b \diamond b \diamond \end{array}$$

The following corollary implies that every partial word has a "binary equivalent".

Corollary 4. *[24] The set of valid ternary correlations over an alphabet A with $\|A\| \geq 2$ is the same as the set of valid ternary correlations over the binary alphabet. Phrased differently, if u is a partial word over an alphabet A, then there exists a binary partial word v such that*

$$1.\ |v| = |u| \quad 2.\ \mathcal{P}(v) = \mathcal{P}(u) \quad 3.\ \mathcal{P}'(v) = \mathcal{P}'(u)$$

In [74], Halava, Harju and Ilie gave a simple constructive proof of Theorem 18 which computes v in linear time. This result was later proved for partial words with one hole by extending Halava et al.'s approach [16]. More specifically, given a partial word u with one hole over an alphabet A, a partial word v over the binary alphabet exists such that Conditions 1–3 hold as well as the following condition

$$4.\ H(v) \subset H(u)$$

However, Conditions 1–4 cannot be satisfied simultaneously in the two-hole case. For the partial word $abaca\diamond\diamond acaba$ can be checked by brute force to have no such binary equivalent (although it has the binary equivalent $abbbb\diamond bbb\diamond b\diamond$ as discussed above).

Open problem 10 *Characterize the partial words that have an equivalent over the binary alphabet $\{a, b\}$ satisfying Conditions 1–4.*

Open problem 11 *Design an efficient algorithm for computing a binary equivalent satisfying Conditions 1–4 when such equivalent exists.*

Open problem 12 *Can we always find an equivalent over the ternary alphabet $\{a, b, c\}$ that satisfies Conditions 1–4?*

2.5.2 Structural Properties of Correlations

A result of Rivals and Rahmann [108] states that Γ_n, the set of full word correlations of length n, is a lattice under set inclusion which does not satisfy the Jordan-Dedekind condition, a criterion which stipulates that all maximal chains between two elements of a poset are of equal length. Violating the Jordan-Dedekind condition implies that Γ_n is not distributive.

We now discuss corresponding results for partial words.

Theorem 21. *[24] The set Δ_n of partial word binary correlations of length n is a distributive lattice under $\subset$ where for $u, v \in \Delta_n$, $u \subset v$ if $\mathcal{P}(u) \subset \mathcal{P}(v)$, and thus satisfies the Jordan-Dedekind condition. Here*

1. *The meet of u and v, $u \cap v$, is the unique vector in Δ_n such that $\mathcal{P}(u \cap v) = \mathcal{P}(u) \cap \mathcal{P}(v)$.*
2. *The join of u and v, $u \cup v$, is the unique vector in Δ_n such that $\mathcal{P}(u \cup v) = \mathcal{P}(u) \cup \mathcal{P}(v)$.*
3. *The null element is 10^{n-1}.*
4. *The universal element is 1^n.*

The union of u and v, $u \cup v$, is the vector in Δ'_n defined as $(u \cup v)_i = 0$ if $u_i = v_i = 0$, 1 if either $u_i = 1$ or $v_i = 1$, and 2 otherwise. However, Δ'_n is not closed under union. Considering the example

$$\begin{aligned} u &= 1\,0\,2\,0\,0\,0\,1\,0\,1 \\ v &= 1\,0\,0\,0\,1\,0\,0\,0\,1 \\ (u \cup v) &= 1\,0\,2\,0\,1\,0\,1\,0\,1 \end{aligned}$$

there is no $i \geq 2$ such that $(u \cup v)_{i2} = 0$, and therefore $(u \cup v)$ is not a valid ternary correlation. However 101010101 is valid.

Theorem 22. *[24] The set Δ'_n of partial word ternary correlations of length n is a distributive lattice under $\subset$ where for $u, v \in \Delta'_n$, $u \subset v$ if $u_i = 1$ implies $v_i = 1$ and $u_i = 2$ implies $v_i = 1$ or $v_i = 2$. Here*

1. *The meet of u and v, $u \wedge v$, is the vector $(u \cap v)$ in Δ'_n defined by $\mathcal{P}(u \wedge v) = \mathcal{P}(u) \cap \mathcal{P}(v)$ and $\mathcal{P}'(u \wedge v) = \mathcal{P}'(u) \cap \mathcal{P}'(v)$.*
2. *The join of u and v, $u \vee v$, is the vector in Δ'_n defined by*

$$\mathcal{P}'(u \vee v) = \mathcal{P}'(u) \cup \mathcal{P}'(v)$$
$$\mathcal{P}(u \vee v) = \mathcal{P}(u) \cup \mathcal{P}(v) \cup B(u \cup v)$$

where $B(u \cup v)$ *is the set of all* $0 < p < n$ *such that* $(u \cup v)_p = 2$ *and there exists no* $i \geq 2$ *satisfying* $(u \cup v)_{ip} = 0$.

In the case of full words, some periods are implied by other periods because of the forward propagation rule. If a twelve-letter full word has periods 7 and 9 then it must also have period 11 since $11 = 7 + 2(9 - 7)$, so $\{7, 9, 11\}$ corresponds to the irreducible period set $\{7, 9\}$. Another result of Rivals and Rahmann shows that the set Λ_n of these irreducible period sets is not a lattice but does satisfy the Jordan-Dedekind condition as a poset [108].

However, forward propagation does not hold in the case of partial words as can be seen with the partial word $abbbbbb\diamond b\diamond bb$ which has periods 7 and 9 but does *not* have period 11. The set $\{7, 9, 11\}$ is irreducible in the sense of partial words, but not in the sense of full words.

This leads us to the definition of generating sets.

Definition 7. *[24] A set* $P \subset \{1, 2, \ldots, n-1\}$ *generates the correlation* $v \in \Delta_n$ *provided that for each* $0 < i < n$ *we have that* $v_i = 1$ *if and only if there exists* $p \in P$ *and* $0 < k < \frac{n}{p}$ *such that* $i = kp$.

For instance, if $v = 1001001101$, then $\{3, 6, 7, 9\}$, $\{3, 6, 7\}$, $\{3, 7, 9\}$, and $\{3, 7\}$ generate v. However, the set $\{3, 7\}$ is the minimal generating set of v.

For every $v \in \Delta_n$ there is a minimal generating set $R(v)$ for v which we call the irreducible period set of v. Namely, this is the set of $p \in \mathcal{P}(v)$ such that for all $q \in \mathcal{P}(v)$ with $q \neq p$ we have that q does not divide p. Denoting by Φ_n the set of irreducible period sets of partial words of length n, we see that there is an obvious bijective correspondence between Φ_n and Δ_n given by

$$\begin{aligned} R : \Delta_n &\to \Phi_n \\ v &\mapsto R(v) \end{aligned}$$

$$\begin{aligned} E : \Phi_n &\to \Delta_n \\ P &\mapsto \bigcup_{p \in P} \langle p \rangle_n \end{aligned}$$

For $n \geq 3$, we see immediately that the poset $(\Phi_n, \subset)$ is not a join semilattice since the sets $\{1\}$ and $\{2\}$ will never have a join because $\{1\}$ is always maximal. On the other hand, the following holds.

Proposition 1. *[24] The set* Φ_n *of irreducible period sets of partial words of length* n *is a meet semilattice under set inclusion which satisfies the Jordan-Dedekind condition. Here the null element is* $\emptyset$*, and the meet of two elements is simply their intersection.*

Open problem 13 *Is there an efficient enumeration algorithm for* Δ_n*?*

2.5.3 Counting Correlations

In this section, we look at the number of valid correlations of a given length. In the case of binary correlations, we give bounds and link the problem to one in number theory. In the case of ternary correlations, we give an exact count.

A primitive set of integers is a subset $S \subset \{1, 2, \ldots\}$ such that for any two distinct elements $s, s' \in S$ we have that neither s divides s' nor s' divides s. The irreducible period sets of correlations $v \in \Delta_n$ are precisely the finite primitive subsets of $\{1, 2, \ldots, n-1\}$.

A result of Erdös can be stated as follows.

Theorem 23. *[65] Let S be a finite primitive set of size k with elements less than n. Then $k \leq \lfloor \frac{n}{2} \rfloor$. Moreover, this bound is sharp.*

This bound shows that the number of binary correlations of length n is at most the number of subsets of $\{1, 2, \ldots, n-1\}$ of size at most $\lfloor \frac{n}{2} \rfloor$. Moreover, the sharpness of the bound gives us that

$$\|\Delta_n\| \geq 2^{\lfloor \frac{n}{2} \rfloor}$$

Thus

$$\frac{\ln 2}{2} \leq \frac{\ln \|\Delta_n\|}{n} \leq \ln 2$$

Open problem 14 *Refine this bound on the cardinality of Δ_n, the set of all partial word binary correlations of length n.*

Guibas and Odlyzko [71] showed that as $n \to \infty$

$$\tfrac{1}{2 \ln 2} + o(1) \leq \tfrac{\ln \|\Gamma_n\|}{(\ln n)^2} \leq \tfrac{1}{2 \ln(\frac{3}{2})} + o(1)$$

and Rivals and Rahmann [108] improved the lower bound to

$$\tfrac{\ln \|\Gamma_n\|}{(\ln n)^2} \geq \tfrac{1}{2 \ln 2} \left(1 - \tfrac{\ln \ln n}{\ln n}\right)^2 + \tfrac{0.4139}{\ln n} - \tfrac{1.47123 \ln \ln n}{(\ln n)^2} + O\left(\tfrac{1}{(\ln n)^2}\right)$$

where Γ_n is the set of all full word correlations of length n. Thus the bounds we give, which show explicitly that $\ln \|\Delta_n\| = \Theta(n)$, demonstrate that the number of partial word binary correlations is much greater than the number of full word correlations.

Lemma 1. *[24]*

1. *Let u be a partial word of length n. Then $p \in \mathcal{P}(u)$ if and only if $ip \in \mathcal{P}'(u)$ for all $0 < i \leq \lfloor \frac{n}{p} \rfloor$.*
2. *If $S \subset \{1, 2, \ldots, n-1\}$, then there exists a unique correlation $v \in \Delta'_n$ such that $\mathcal{P}'(v) \setminus \{n\} = S$.*

Consequently, the cardinality of Δ'_n, the set of valid ternary correlations of length n, is the same as the cardinality of the power set of $\{1, 2, \ldots, n-1\}$, and thus

$$\|\Delta'_n\| = 2^{n-1}$$

We end this section with the following open problem.

Open problem 15 *Exhibit an algorithm to sample uniformly (weak) period sets through irreducible (weak) period sets.*

2.6 Primitive and Unbordered Partial Words

The two fundamental concepts of *primitiveness* and *borderedness* play an important role in several research areas including coding theory [5, 6, 117], combinatorics on words [45, 92, 93, 94, 96], computational biology [39, 100], data communication [41], data compression [49, 119, 123], formal language theory [57, 58], and text algorithms [38, 50, 51, 52, 69, 72, 84, 102, 118]. A primitive word is one that cannot be written as a power of another word, while an unbordered word is a primitive word such that none of its proper prefixes is one of its suffixes. For example, $\underline{ab}a\underline{ab}$ is bordered with border ab while $abaabb$ is unbordered. The number of primitive and unbordered words of a fixed length over an alphabet of a fixed size is well known, the number of primitive words being related to the Möbius function [92].

In this section, we discuss, in particular, the problems of counting primitive and unbordered partial words.

2.6.1 Primitiveness

A word u is primitive if there exists no word v such that $u = v^i$ with $i \geq 2$. A natural algorithmic problem is how can we decide efficiently whether a given word is primitive. The problem has a brute force quadratic solution: divide the input word into two parts and check whether the right part is a power of the left part. But how can we obtain a faster solution to the problem? Fast algorithms for testing primitivity of words can be based on the combinatorial result that a word u is primitive if and only if u is not an inside factor of its square uu, that is, $uu = xuy$ implies $x = \varepsilon$ or $y = \varepsilon$ [45]. Indeed, any linear time string matching algorithm can be used to test whether the string u is a proper factor of uu. If the answer is no, then the primitiveness of u has been verified [51]. So testing whether or not a word is primitive can be done in linear time in the length of the word.

Primitive partial words were defined in [9]: A partial word u is primitive if there exists no word v such that $u \subset v^i$ with $i \geq 2$. It turns out that a partial word u with one hole is primitive if and only if $uu \uparrow xuy$ for some partial words x, y implies $x = \varepsilon$ or $y = \varepsilon$ [9]. A linear time algorithm for testing primitivity of partial words with one hole can be based on this combinatorial result. As an application, the existence of a binary equivalent for any partial word with one hole satisfying Conditions 1–4 discussed in Section 2.5 was obtained [16]. In [11], a linear time algorithm was described to test primitivity on

partial words with more than one hole. Here the concept of speciality related to commutativity on partial words, which was discussed in Section 2.2, is foundational in the design of the algorithm. More precisely, it was shown that if u is a primitive partial word with more than one hole satisfying $uu \uparrow xuy$ for some nonempty partial words x and y such that $|x| < |y|$, then u is $(|x|, |y|)$-special. The partial words $u = ab\diamond bbb\diamond b$, $x = a\diamond$, and $y = c\diamond bbcb$ illustrate the fact that the condition of speciality plays a role when dealing with partial words with more than one hole.

In [19], the very challenging problem of counting the number $P_{h,k}(n)$ (respectively, $P'_{h,k}(n)$) of primitive (respectively, nonprimitive) partial words with h holes of length n over a k-size alphabet was considered. There, formulas for $h = 1$ and $h = 2$ in terms of the well known formula for $h = 0$ were given. Denote by $T_{h,k}(n)$ the sum of $P_{h,k}(n)$ and $P'_{h,k}(n)$.

We first recall the counting for primitive full words. Since there are exactly k^n words of length n over a k-size alphabet and every nonempty word w has a unique primitive root v for which $w = v^{n/d}$ for some divisor d of n, the following relation holds:

$$k^n = \sum_{d|n} P_{0,k}(d)$$

Using the Möbius inversion formula, we obtain the following well-known expression for $P_{0,k}(n)$ [92, 105]:

$$P_{0,k}(n) = \sum_{d|n} \mu(d) k^{n/d}$$

where the Möbius function, denoted by μ, is defined as

$$\mu(n) = \begin{cases} 1 & \text{if } n = 1 \\ (-1)^i & \text{if } n \text{ is a product of } i \text{ distinct primes} \\ 0 & \text{if } n \text{ is divisible by the square of a prime} \end{cases}$$

The cases where $h = 1$ and $h = 2$ are stated in the next two theorems.

Theorem 24. *[19] The equality $P'_{1,k}(n) = nP'_{0,k}(n)$ holds.*

Theorem 25. *[19]*

1. *For an odd positive integer n:*

$$P'_{2,k}(n) = \binom{n}{2} P'_{0,k}(n)$$

2. *For an even positive integer n:*

$$P'_{2,k}(n) = \binom{n}{2} P'_{0,k}(n) - (k-1) T_{1,k}(\frac{n}{2})$$

Open problem 16 *Count the number $P'_{h,k}(n)$ of nonprimitive partial words with h holes of length n over a k-size alphabet for $h > 2$.*

Another problem to investigate is the following.

Open problem 17 *Study the language of primitive partial words as is done for full primitive words in [105].*

We end this section with the following remark. In [18], the authors obtained consequences of the generalizations of Fine and Wilf's periodicity result to partial words. In particular, they generalized the following combinatorial property: "For any word u over $\{a, b\}$, ua or ub is primitive." This property proves in some sense that there exist very many primitive words.

2.6.2 Borderedness

Unbordered partial words were also defined in [9]: A nonempty partial word u is unbordered if no nonempty partial words x_1, x_2, v, w exist such that $u = x_1v = wx_2$ and $x_1 \uparrow x_2$. If such nonempty words exist and x is such that $x_1 \subset x$ and $x_2 \subset x$, then we call u *bordered* and x a *border* of u. A border x of u is called *minimal* if $|x| > |y|$ implies that y is not a border of u. Note that there are two types of borders: x is an *overlapping* border if $|x| > |v|$, and a *nonoverlapping* border otherwise. The partial word $u = a\diamond\diamond ab$ is bordered with the nonoverlapping border ab and overlapping border aab, the first one being minimal, while the partial word $ab\diamond c$ is unbordered.

We call a bordered partial word u *simply bordered* if a minimal border x exists satisfying $|u| \geq 2|x|$.

Proposition 2. *[21] Let u be a nonempty bordered partial word. Let x be a minimal border of u, and set $u = x_1v = wx_2$ where $x_1 \subset x$ and $x_2 \subset x$. Then the following hold:*

1. *The partial word x is unbordered.*
2. *If x_1 is unbordered, then $u = x_1u'x_2 \subset xu'x$ for some u'.*

Note that Proposition 2 implies that if u is a full bordered word, then $x_1 = x$ is unbordered. In this case, $u = xu'x$ where x is the minimal border of u. Hence a bordered full word is always simply bordered.

Corollary 5. *[21] Every bordered full word of length n has a unique minimal border x. Moreover, x is unbordered and $|x| \leq \lfloor \frac{n}{2} \rfloor$.*

In [20], the problem of enumerating all unbordered partial words with h holes of length n over a k-letter alphabet was considered, a problem that yields some open questions for further investigation. We will denote by $U_{h,k}(n)$ the number of such words.

Let us start with the problem of enumerating all unbordered full words of length n over a k-letter alphabet which gives a conceptually simple and elegant recursive formula: $U_{0,k}(0) = 1$, $U_{0,k}(1) = k$, and for $n > 0$,

$$U_{0,k}(2n) = kU_{0,k}(2n-1) - U_{0,k}(n)$$
$$U_{0,k}(2n+1) = kU_{0,k}(2n)$$

These equalities can be seen from the fact that if a word has odd length $2n+1$ then it is unbordered if and only if it is unbordered after removing the middle letter. If a word has even length $2n$ then it is unbordered if and only if it is obtained from an unbordered word of length $2n-1$ by adding a letter next to the middle position unless doing so creates a word that is a perfect square.

Using these formulas and Proposition 2, we can easily obtain a formula for counting bordered full words. Let $B_k(j,n)$ be the number of full words of length n over a k-letter alphabet that have a minimal border of length j:

$$B_k(j,n) = U_{0,k}(j)k^{n-2j}$$

If we let $B_k(n)$ be the number of full words of length n over a k-letter alphabet with a border of any length, then we have that

$$B_k(n) = \sum_{j=1}^{\lfloor \frac{n}{2} \rfloor} B_k(j,n)$$

When we allow words to have holes, counting bordered partial words is made extremely more difficult by the failure of Corollary 5 since there is now the possibility of a minimal border that is overlapping as in $a\diamond bb$. We will first concern ourselves with the simply bordered partial words. Note that because borderedness in partial words is defined via containment, it no longer makes sense to talk about *the* minimal border of a partial word, there could be many possible borders of a certain length.

To see inside the structure of the partial words we are trying to count we first define a function. Let $f_{h,k}(i,j,n)$ be the number of partial words of length n with $h > 0$ holes over a k-letter alphabet that have a hole in position i and a minimal border of length j. When $i = 0$:

$$f_{h,k}(0,j,n) = \begin{cases} \binom{n-1}{h-1}k^{n-h} & \text{if } j = 1 \\ 0 & \text{if } j > 1 \end{cases}$$

It is clear that $f_{h,k}(i,j,n)$ has some symmetry, namely that, $f_{h,k}(i,j,n) = f_{h,k}(n-1-i,j,n)$. Throughout this section we will rely on this to consider only $i \leq \lfloor \frac{n}{2} \rfloor$.

We have some general formulas for the evaluation of $f_{h,k}(i,j,n)$.

Proposition 3. *[20]*
If $0 < i < j-1$ and $j < \frac{n}{2}$, then

$$f_{h,k}(i,j,n) = \sum_{h'=1}^{\min(h,2j)} f_{h',k}(i,j,2j)\binom{n-2j}{h-h'}k^{n-2j-h+h'}$$

It is possible to see from the formula in Proposition 3 that we need only really concern ourselves with the case when $j = \lfloor \frac{n}{2} \rfloor$.

There is a similar simplification that can be made if $j - 1 < i$.

Proposition 4. *[20]*
If $j-1<i$, then

$$f_{h,k}(i,j,n)=2\sum_{i'=0}^{j-1}\sum_{h'=0}^{\min(h-1,2j)}f_{h',k}(i',j,2j)\binom{n-2j-1}{h-1-h'}k^{n-2j-h+h'}$$

If we restrict our attention to the case when $h=1$, then we can present many explicit formulas for the values $f_{1,k}(i,j,n)$. The exceptional case when $i=0$ is easily dispensed with:

$$f_{1,k}(0,j,n)=\begin{cases}k^{n-1} & \text{if } j=1\\ 0 & \text{if } j>1\end{cases}$$

Note that in the case where $0<i<j-1$ and $j<\frac{n}{2}$, the formula in Proposition 3 reduces to the very simple equality

$$f_{1,k}(i,j,n)=f_{1,k}(i,j,2j)k^{n-2j}$$

Similarly, in the case where $j-1<i$, the formula in Proposition 4 reduces to

$$f_{1,k}(i,j,n)=U_{0,k}(j)k^{n-2j-1}$$

By the above discussion we can restrict our attention to the cases when $i>0$, $n=2m$ and $j=m$. These are partial words with a border that takes up exactly half the length of the word. We wish to find a complete formula for $f_{1,k}(i,m,2m)$ where $i=m-1-i'$.

We proceed by induction on i'. When $i'=0$, we have the following.

Lemma 2. *[20] For all $m\geq 2$, $f_{1,k}(m-1,m,2m)=U_{0,k}(m)$.*

Continuing with the first interesting case $i'=1$, we have the following lemma.

Lemma 3. *[20]*
For all $m\geq 3$, $f_{1,k}(m-2,m,2m)=U_{0,k}(m)-k(k-1)$.

This kind of analysis quickly becomes much more complicated though. The evaluation breaks up into cases depending on how the periodicity of the words interacts with the length of the border in modular arithmetic.

Lemma 4. *[20] For all $m\geq 4$, the following holds:*

$$f_{1,k}(m-3,m,2m)=\begin{cases}U_{0,k}(m)-k^2(k-1)-k(k-1) & \text{if } m\equiv 1 \bmod 2\\ U_{0,k}(m)-k(k-1)^2-k(k-1) & \text{if } m\equiv 0 \bmod 2\end{cases}$$

Lemma 5. *[20] For all $m\geq 5$, the following holds:*

$$f_{1,k}(m-4,m,2m)=U_{0,k}(m)-k(k-1)-g_1(m)-g_2(m)$$

where

$$g_1(m) = \begin{cases} k(k-1)^2 & \text{if } m \equiv 0 \bmod 2 \\ 0 & \text{if } m \equiv 1 \bmod 2 \end{cases}$$

and

$$g_2(m) = \begin{cases} k^2(k-1)^2 & \text{if } m \equiv 0 \bmod 3 \\ U_{0,k}(4) & \text{if } m \equiv 1 \bmod 3 \\ k^2(k-1)^2 & \text{if } m \equiv 2 \bmod 3 \end{cases}$$

To give an idea of how the values for $f_{1,k}(i, m, 2m)$ behave unpredictably, here is a table of values that has been put together through a brute force count:

i	0	1	2	3	4	5	6	7
$f_{1,2}(i,2,4)$	0	2						
$f_{1,2}(i,3,6)$	0	2	4					
$f_{1,2}(i,4,8)$	0	2	4	6				
$f_{1,2}(i,5,10)$	0	6	6	10	12			
$f_{1,2}(i,6,12)$	0	10	12	16	18	20		
$f_{1,2}(i,7,14)$	0	22	26	32	34	38	40	
$f_{1,2}(i,8,16)$	0	42	52	60	66	70	72	74

Open problem 18 *Compute the values $f_{1,k}(m-i, m, 2m)$ for $m > i$.*

Let $S_{h,k}(n)$ be the number of simply bordered partial words of length n with h holes over a k-letter alphabet. Clearly if $h > n$, then $S_{h,k}(n) = 0$. Note that when $h = 0$, $S_{h,k}(n) = B_k(n)$.

Theorem 26. *[20]*
If $0 < h \leq n$, then a formula for $S_{h,k}(n)$ is given by:

$$S_{h,k}(2m+1) = S_{h-1,k}(2m) + kS_{h,k}(2m)$$

$$S_{h,k}(2m) = \frac{2\sum_{i=0}^{m-1}\sum_{j=1}^{m} f_{h,k}(i,j,2m)}{h}$$

We can check that

$$S_{1,k}(n) = \sum_{i=0}^{n-1}\sum_{j=1}^{\lfloor \frac{n}{2} \rfloor} f_{1,k}(i,j,n)$$

Let $N_{h,k}(n)$ be the number of partial words with h holes, of length n, over a k-letter alphabet that are not simply bordered. Obviously we can find the value of this function by subtracting the value of $S_{h,k}(n)$ from the total number of partial words with those parameters, but it would be of interest to find a direct formula for $N_{h,k}(n)$. If $h = 0$, then

$$N_{0,k}(n) = U_{0,k}(n)$$

since a bordered full word that is not simply bordered is an unbordered full word. It is easy to see that $N_{1,k}(0) = 0$, $N_{1,k}(1) = 1$, $N_{1,k}(2) = 0$, and for $h > 1$ that $N_{h,k}(1) = 0$ and $N_{h,k}(2) = 0$. Now, for $h > 0$, the following formula holds for odd $n = 2m + 1$:

$$N_{h,k}(2m+1) = kN_{h,k}(2m) + N_{h-1,k}(2m)$$

Open problem 19 *What is $N_{h,k}(2m)$?*

If we simplify the problem down to the $h = 1$ case, then we can again use the values $f_{1,k}(i, j, n)$ to give a formula for $N_{1,k}(n)$:

$$N_{1,k}(2m) = kN_{1,k}(2m-1) + 2U_{0,k}(2m-1) - \sum_{i=1}^{m} f_{1,k}(i, m, 2m)$$

but it rests on the evaluation of the $f_{1,k}(i, j, 2j)$'s as well.

Other interesting questions include the following.

Open problem 20 *Count the number $O_{h,k}(n)$ of overlapping bordered partial words of length n with h holes over a k-letter alphabet for $h > 0$.*

Open problem 21 *Count the number $U_{h,k}(n)$ of unbordered partial words of length n with h holes over a k-letter alphabet for $h > 0$.*

Another open question is suggested by the fact that every partial word of length 5 that has more than two holes is simply bordered. The partial word $a\diamond\diamond bb$ shows that this bound on the number of holes for length 5 is tight. For length 6, every partial word with more than 2 holes is simply bordered as well.

Open problem 22 *What is the maximum number of holes $M(n)$ a partial word of length n can have and still fail to be simply bordered? Some values for small n follow.*

n	$M(n)$
5	2
6	2
7	3
8	4
9	5
10	5
11	6
12	7
13	8
14	8
15	9

We end this section by discussing another open problem related to borderedness in the context of partial words.

In 1979, Ehrenfeucht and Silberger initiated a line of research to explore the relationship between the minimal period of a word w of length n, $p(w)$, and the maximum length of its unbordered factors, $\mu(w)$ [64]. Clearly, $\mu(w) \leq p(w)$. They conjectured that if $n \geq 2\mu(w)$, then $\mu(w) = p(w)$. In [3], a counterexample was given and it was conjectured that $3\mu(w)$ is the precise bound. In 1982, it was established that if $n \geq 4\mu(w) - 6$, then $\mu(w) = p(w)$ [61]. In 2003, the bound was improved to $3\mu(w) - 2$ in [76] where it is believed that the precise bound can be achieved with methods similar to those presented in that paper.

Open problem 23 *Investigate the relationship between the minimal weak period of a partial word and the maximum length of its unbordered factors.*

2.7 Equations on Partial Words

As was seen in Section 2.2, some of the most basic properties of words, like the commutativity and the conjugacy properties, can be expressed as solutions of the word equations $xy = yx$ and $xz = zy$ respectively. It is also well known that the equation $x^m = y^n z^p$ has only periodic solutions in a free semigroup, that is, if $x^m = y^n z^p$ holds with integers $m, n, p \geq 2$, then there exists a word w such that x, y, z are powers of w. This result, which received a lot of attention, was first proved by Lyndon and Schützenberger for free groups [96]. Their proof implied the case for free semigroups since every free semigroup can be embedded in a free group. Direct proofs for free semigroups appear in [46, 77, 92].

In this section, we study *equations on partial words*. When we speak about them, we replace the notion of equality with compatibility. But compatibility is not transitive! We already solved the commutativity equation $xy \uparrow yx$ as well as the conjugacy equation $xz \uparrow zy$ in Section 2.2. As an application of the commutativity equation, we mention the linear time algorithm for testing primitivity on partial words that was discussed in Section 2.6 [11], and as an application of the conjugacy equation, we mention the efficient algorithm for computing a critical factorization when one exists that was discussed in Section 2.4 [22, 35]. Here, we solve three equations: $x^m \uparrow y^n$, $x^2 \uparrow y^m z$, and $x^m \uparrow y^n z^p$.

First, let us consider the equation $x^m \uparrow y^n$, also called the "good pairs" equation. If x and y are full words, then $x^m = y^n$ for some positive integers m, n if and only if there exists a word z such that $x = z^k$ and $y = z^l$ for some integers k, l. When dealing with partial words x and y, if there exists a partial word z such that $x \subset z^k$ and $y \subset z^l$ for some integers k, l, then $x^m \uparrow y^n$ for some positive integers m, n.

For the converse, we need a couple of lemmas.

Lemma 6. *[13]*
Let x, y be partial words and let m, n be positive integers such that $x^m \uparrow y^n$ with $\gcd(m,n) = 1$. Call $|x|/n = |y|/m = p$. If there exists an integer i such that $0 \le i < p$ and $x_{i,p}$ is not 1-periodic, then $D(y_{i,p})$ is empty.

Lemma 7. *[13]*
Let x be a partial word, let m, p be positive integers, and let i be an integer such that $0 \le i < p$. Then the relation

$$x_{i,p}^m = x_{i,p} x_{(i-|x|) \bmod p,p} \cdots x_{(i-(m-1)|x|) \bmod p,p}$$

holds.

The "good pairs" theorem is stated as follows.

Theorem 27. *[13]*
Let x, y be partial words and let m, n be positive integers such that $x^m \uparrow y^n$ with $\gcd(m,n) = 1$. Assume that (x, y) is a good pair, that is,

1. *For all $i \in H(x)$ the word ${y^n}_{i,|x|}$ is 1-periodic,*
2. *For all $i \in H(y)$ the word ${x^m}_{i,|y|}$ is 1-periodic.*

Then there exists a partial word z such that $x \subset z^k$ and $y \subset z^l$ for some integers k, l.

The assumption of (x, y) being a good pair is necessary in the "good pairs" theorem. Indeed, $x^2 = (a\diamond b)^2 \uparrow (acbadb)^1 = y^1$ but $y(1)y(4) = cd$ is not 1-periodic, and there exists no partial word z as desired.

Corollary 6. *[13]*
Let x and y be primitive partial words such that (x, y) is a good pair. If $x^m \uparrow y^n$ for some positive integers m and n, then $x \uparrow y$.

Note that if both x and y are full words, then (x, y) is a good pair. The corollary hence implies that if x, y are primitive full words satisfying $x^m = y^n$ for some positive integers m and n, then $x = y$.

Second, we consider the "good triples" equation $x^2 \uparrow y^m z$. Here, it is assumed that m is a positive integer and z is a prefix of y.

Nontrivial solutions exist! A solution is trivial if x, y, z are contained in powers of a common word. The equation $x^2 \uparrow y^m z$ has nontrivial solutions. For instance, $(a\diamond\diamond a)^2 \uparrow (aab)^2 aa$ where $x = a\diamond\diamond a$, $y = aab$, and $z = aa$.

The "good triples" theorem follows.

Theorem 28. *[13]*
Let x, y, z be partial words such that z is a proper prefix of y. Then $x^2 \uparrow y^m z$ for some positive integer m if and only if there exist partial words

$$u, v, u_0, v_0, \ldots, u_{m-1}, v_{m-1}, z_x$$

such that $u \neq \varepsilon$, $v \neq \varepsilon$, $y = uv$,

$$x = (u_0v_0)\dots(u_{n-1}v_{n-1})u_n \tag{2.1}$$
$$= v_n(u_{n+1}v_{n+1})\dots(u_{m-1}v_{m-1})z_x \tag{2.2}$$

where $0 \leq n < m, u \uparrow u_i$ *and* $v \uparrow v_i$ *for all* $0 \leq i < m$, $z \uparrow z_x$, *and where one of the following holds:*

1. $m = 2n$, $|u| < |v|$, *and there exist partial words* u', u'_n *such that* $z_x = u'u_n$, $z = uu'_n$, $u \uparrow u'$ *and* $u_n \uparrow u'_n$.
2. $m = 2n + 1$, $|u| > |v|$, *and there exist partial words* v'_{2n} *and* z'_x *such that* $u_n = v_{2n}z_x$, $u = v'_{2n}z'_x$, $v_{2n} \uparrow v'_{2n}$ *and* $z_x \uparrow z'_x$.

A triple of partial words (x, y, z) *which satisfy these properties we will refer to as a good triple.*

Two corollaries can be deduced.

Corollary 7. *[13]*
Let x, y, z *be partial words such that* z *is a prefix of* y. *Assume that* x, y *are primitive and that* $x^2 \uparrow y^m z$ *for some integer* $m \geq 2$. *If* x *has at most one hole and* y *is full, then* $x \uparrow y$.

Corollary 8. *[13]*
Let x, y, z *be words such that* z *is a prefix of* y. *If* x, y *are primitive and* $x^2 = y^m z$ *for some integer* $m \geq 2$, *then* $x = y$.

Note that the corollaries do not hold when $m = 1$. Indeed, the words $x = aba$, $y = abaab$ and $z = a$ provide a counterexample. Also, the first corollary does not hold when x is full and y has one hole as is seen by setting $x = abaabb$, $y = ab\diamond$ and $z = \varepsilon$.

Third, let us consider the equation $x^m y^n \uparrow z^p$. The case of full words is well known.

Theorem 29. *[96]*
Let x, y, z *be full words and let* m, n, p *be integers such that* $m \geq 2, n \geq 2$ *and* $p \geq 2$. *Then the equation* $x^m y^n = z^p$ *has only trivial solutions, that is,* x, y, *and* z *are each a power of a common element.*

When we deal with partial words, the equation $x^m y^n \uparrow z^p$ certainly has a solution when x, y, and z are contained in powers of a common word (we call such solutions the trivial solutions). However, there may be nontrivial solutions as is seen with the compatibility relation

$$(a\diamond b)^2(b\diamond a)^2 \uparrow (abba)^3$$

We will classify solutions as Type 1 (or trivial) solutions when there exists a partial word w such that x, y, z are contained in powers of w, and as Type 2 solutions when the partial words x, y, z satisfy $x \uparrow z$ and $y \uparrow z$. Note that if z is full, then Type 2 solutions are trivial solutions.

The case $p \geq 4$ is stated in the following theorem.

Theorem 30. *[13] Let x, y, z be primitive partial words such that (x, z) and (y, z) are good pairs. Let m, n, p be integers such that $m \geq 2, n \geq 2$ and $p \geq 4$. Then the equation $x^m y^n \uparrow z^p$ has only solutions of Type 1 or Type 2 unless one of the following holds:*

1. *$x^2 \uparrow z^k z_p$ for some integer $k \geq 2$ and nonempty prefix z_p of z,*
2. *$z^2 \uparrow x^l x_p$ for some integer $l \geq 2$ and nonempty prefix x_p of x.*

Open problem 24 *Solve the equation $x^m y^n \uparrow z^p$ on partial words for integers $m \geq 2, n \geq 2$ and $p \in \{2, 3\}$.*

2.8 Unavoidable Sets of Partial Words

A set of (full) words X over a finite alphabet A is *unavoidable* if no two-sided infinite word over A avoids X, that is, X is unavoidable if every two-sided infinite word over A has a factor in X. For instance, the set $X = \{a, bbb\}$ is unavoidable (if a two-sided infinite word w does not have a as a factor, then w consists only of b's). This concept was explicitly introduced in 1983 in connection with an attempt to characterize the rational languages among the context-free ones [63]. It is clear from the definition that from each unavoidable set we can extract a finite unavoidable subset, so the study can be reduced to finite unavoidable sets. There is a vast literature on unavoidable sets of words and we refer the reader to [44, 93, 109, 110] for more information.

Unavoidable sets of partial words were introduced recently in [15], where the problem of classifying such sets of small cardinality was initiated, in particular, those with two elements. The authors showed that this problem reduces to the one of classifying unavoidable sets of the form

$$\{a\diamond^{m_1} a \ldots a\diamond^{m_k} a, b\diamond^{n_1} b \ldots b\diamond^{n_l} b\}$$

where $m_1, \ldots, m_k, n_1, \ldots, n_l$ are nonnegative integers and a, b are distinct letters. They gave an elegant characterization of the special case of this problem when $k = 1$ and $l = 1$. They proposed a conjecture characterizing the case where $k = 1$ and $l = 2$ and proved one direction of the conjecture. They then gave partial results towards the other direction and in particular proved that the conjecture is easy to verify in a large number of cases. Finally, they proved that verifying this conjecture is sufficient for solving the problem for larger values of k and l. In [27], the authors built on the previous work by examining, in particular, unavoidable sets of size three.

In [15], the question was raised as to whether there is an efficient algorithm to determine if a finite set of partial words is unavoidable. In [26], it was shown that this problem is NP-hard by using techniques similar to those used in a recent paper on the complexity of computing the capacity of codes that avoid forbidden difference patterns [37]. This is in contrast with the well known feasibility results for unavoidability of a set of full words [93].

The contents of Section 2.8 is as follows: In Section 2.8.1, we review basics on unavoidable sets of partial words. In Section 2.8.2, we discuss classifying such sets of size two. And in Section2.8.3, we discuss testing unavoidability of sets of partial words.

2.8.1 Unavoidable Sets

We first define some basic terminology. A two-sided infinite word over A is a total function $w : \mathbb{Z} \to A$. A finite word u is a factor of w if there exists some $i \in \mathbb{Z}$ such that $u = w(i)w(i+1)\dots w(i+|u|-1)$. A period of w is a positive integer p such that $w(i) = w(i+p)$ for all $i \in \mathbb{Z}$. If w has period p for some p, then we call w periodic. If v is a finite word, then $v^{\mathbb{Z}}$ denotes the two-sided infinite word w with period $|v|$ satisfying $w(0)\dots w(|v|-1) = v$. If X is a set of partial words, then $\hat{X}$ denotes the set of all full words compatible with a member of X. For instance, if $X = \{a\diamond\diamond a, b\diamond b\}$, then $\hat{X} = \{aaaa, aaba, abaa, abba, bab, bbb\}$.

The concept of an unavoidable set of full words is defined as follows.

Definition 8. *Let $X \subset A^*$.*

1. *A two-sided infinite word w avoids X if no factor of w is a member of X.*
2. *The set X is unavoidable if no two-sided infinite word over A avoids X, that is, X is unavoidable if every two-sided infinite word over A has a factor in X.*

If $A = \{a, b\}$, then the following sets are unavoidable: $X_1 = \{\varepsilon\}$ (ε is a factor of every two-sided infinite word); $X_2 = \{a, bbb\}$; $X_3 = \{aa, ab, ba, bb\}$ (this is the set of all words of length 2); and for any $n \in \mathbb{N}$, A^n is unavoidable.

If $X \subset A^*$ is finite, then the following three statements are equivalent: (1) X is unavoidable; (2) There are only finitely many words in A^* with no member of X as a factor; and (3) No periodic two-sided infinite word avoids X.

An unavoidable set of partial words is defined as follows.

Definition 9. *Let $X \subset A_\diamond^*$.*

1. *A two-sided infinite word w avoids X if no factor of w is a member of $\hat{X}$.*
2. *The set X is unavoidable if no two-sided infinite word over A avoids X, that is, X is unavoidable if every two-sided infinite word over A has a factor in $\hat{X}$.*

If $A = \{a, b\}$, then the following sets are unavoidable: $X_1 = \{a\diamond, \diamond b\}$; $X_2 = \{\diamond^n\}$ for any nonnegative integer n as well as any set containing X_2 as a subset (let us call such sets the *trivial* unavoidable sets); and $X_3 = \{a, bbb\}$ since of course Definition 9 is equivalent to Definition 8 if every member of X is full. We will explore some less trivial examples soon.

By the definition of $\hat{X}$, a two-sided infinite word w has a factor in $\hat{X}$ if and only if that factor is compatible with a member of X. Thus the two-sided infinite words which avoid $X \subset A_\diamond^*$ are exactly those which avoid $\hat{X} \subset A^*$, and $X \subset A_\diamond^*$ is unavoidable if and only if $\hat{X} \subset A^*$ is unavoidable. Thus with regards to unavoidability, a set of partial words serves as a representation of a set of full words. The set $\{a\diamond\diamond a, b\diamond b\}$ represents

$$\{aaaa, aaba, abaa, abba, bab, bbb\}$$

We will shortly prove that this set is unavoidable.

The smaller X is, the more information is gained by identifying X as unavoidable. Thus it is natural to begin investigating the unavoidable sets of partial words of small cardinality. Of course, every two-sided infinite word avoids the empty set and thus, there are no unavoidable sets of size 0. Unless the alphabet is unary, the only unavoidable sets of size 1 are trivial. If the alphabet is unary, then every nonempty set is unavoidable and in that case there is only one two-sided infinite word. Thus the unary alphabet is not interesting so we will not consider it further. Classifying the unavoidable sets of size 2 is the focus of the next section.

2.8.2 Classifying Unavoidable Sets of Size Two

If X is a two-element unavoidable set, then every two-sided infinite unary word has a factor compatible with a member of X. In particular, X cannot have fewer elements than the alphabet. Thus if X has size 2, then the alphabet is unary or binary. We hence assume that the alphabet is binary say with distinct letters a and b since we said above that the unary alphabet is not interesting. So one element of X is compatible with a factor of $a^{\mathbb{Z}}$ and the other element is compatible with a factor of $b^{\mathbb{Z}}$, since this is the only way to guarantee that both $a^{\mathbb{Z}}$ and $b^{\mathbb{Z}}$ will not avoid X. Thus we can restrict our attention to sets of the form

$$X_{m_1,\ldots,m_k|n_1,\ldots,n_l} = \{a\diamond^{m_1}a\ldots a\diamond^{m_k}a, b\diamond^{n_1}b\ldots b\diamond^{n_l}b\} \tag{2.3}$$

for some nonnegative integers $m_1,\ldots,m_k$ and $n_1,\ldots,n_l$. For which integers $m_1,\ldots,m_k,n_1,\ldots,n_l$ is $X_{m_1,\ldots,m_k|n_1,\ldots,n_l}$ unavoidable?

A simplification is stated in the next lemma.

Lemma 8. *[15] If p is a nonnegative integer, then set*

$$X = X_{m_1,\ldots,m_k|n_1,\ldots,n_l}$$

and

$$Y = X_{p(m_1+1)-1,\ldots,p(m_k+1)-1|p(n_1+1)-1,\ldots,p(n_l+1)-1}$$

Then X is unavoidable if and only if Y is unavoidable.

The easiest place to start is with small values of k and l. Of course, the set $\{a, b\diamond^{n_1} b \ldots b\diamond^{n_l} b\}$ is always unavoidable for if w is a two-sided infinite word which does not have a as a factor, then $w = b^{\mathbb{Z}}$. This handles the case where $k = 0$ (and symmetrically $l = 0$).

An elegant characterization for the case where $k = l = 1$ is stated in the following theorem.

Theorem 31. *[15] The set $X_{m|n} = \{a\diamond^m a, b\diamond^n b\}$ is avoidable if and only if $m + 1$ and $n + 1$ have the same greatest power of 2 dividing them.*

The next natural step is to look at $k = 1$ and $l = 2$, that is, sets of the form

$$X_{m|n_1,n_2} = \{a\diamond^m a, b\diamond^{n_1} b\diamond^{n_2} b\}$$

On the one hand, we have identified a large number of avoidable sets of the form $\{a\diamond^m a, b\diamond^n b\}$. For $X_{m|n_1,n_2}$ to be avoidable it is sufficient that $\{a\diamond^m a, b\diamond^{n_1} b\}$, $\{a\diamond^m a, b\diamond^{n_2} b\}$ or $\{a\diamond^m a, b\diamond^{n_1+n_2+1} b\}$ be avoidable. On the other hand, the structure of words avoiding $\{a\diamond^m a, b\diamond^{n_1} b\diamond^{n_2} b\}$ is not nearly as nice as those avoiding $\{a\diamond^m a, b\diamond^n b\}$. Thus a simple characterization seems unlikely, unless perhaps there are no unavoidable sets of this form at all. But there are! The set

$$\{a\diamond^7 a, b\diamond b\diamond^3 b\}$$

is unavoidable. Seeing that it is provides a nice example of the techniques that can be used. Referring to the figure below,

...	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	...
...							b		b					...
...													a	...
...					b									...
...											a			...
...									a					...
...	b													...
...							a							...

suppose instead that there exists a two-sided infinite word w which avoids it. We know from Theorem 31 that $\{a\diamond^7 a, b\diamond b\}$ is unavoidable, thus w must have a factor compatible with $b\diamond b$. Say without loss of generality that $w(0) = w(2) = b$. This implies that $w(6) = a$, which in turn implies that $w(-2) = b$. Then we have that $w(-2) = w(0) = b$, forcing $w(4) = a$. This propagation continues: $w(-4) = w(-2) = b$ and so $w(2) = a$, which makes $w(-6) = b$ giving $w(0) = a$, a contradiction.

The perpetuating patterns phenomenon of the previous example is a special case of a more general result.

Theorem 32. *[15] If $m = n_2 - n_1 - 1$ or $m = 2n_1 + n_2 + 2$, and the highest power of 2 dividing $n_1 + 1$ is less than the highest power of 2 dividing $m + 1$, then $X_{m|n_1,n_2}$ is unavoidable.*

Here are other unavoidability results for $k = 1$ and $l = 2$.

Proposition 5. *[15] The set $X_{m|n_1,n_2}$ is unavoidable if Conditions 1 or Conditions 2 or Conditions 3 hold:*

1. *$\{a\diamond^m a, b\diamond^{n_1} b\}$ is unavoidable, $m = 2n_1 + n_2 + 2$ or $m = n_2 - n_1 - 1$, and $n_1 + 1$ divides $n_2 + 1$.*
2. *$n_1 < n_2$, $2m = n_1 + n_2$ and $m - n_1$ divides $m + 1$.*
3. *$m = 6$, $n_1 = 1$ and $n_2 = 3$.*

Extensive experimentation suggests that these results (and their symmetric equivalents) give a complete characterization of when $X_{m|n_1,n_2}$ is unavoidable.

Conjecture 1 *[15] The set $X_{m|n_1,n_2}$ is unavoidable if and only if one of the following conditions (or symmetric equivalents) holds:*

1. *$\{a\diamond^m a, b\diamond^{n_1} b\}$ is unavoidable, $m = 2n_1 + n_2 + 2$ or $m = n_2 - n_1 - 1$, and $n_1 + 1$ divides $n_2 + 1$.*
2. *$m = n_2 - n_1 - 1$ or $m = 2n_1 + n_2 + 2$, and the highest power of 2 dividing $n_1 + 1$ is less than the highest power of 2 dividing $m + 1$.*
3. *$n_1 < n_2$, $2m = n_1 + n_2$ and $m - n_1$ divides $m + 1$.*
4. *$m = 6$, $n_1 = 1$ and $n_2 = 3$.*

Open problem 25 *Is Conjecture 1 true or false?*

If true, then Conjecture 1 implies that the unavoidable sets of size two have been completely classified as stated in the following proposition.

Proposition 6. *[15] If Conjecture 1 is true, then $X_{m_1,\ldots,m_k|n_1,\ldots,n_l}$ is avoidable for $k = 1$ and $l \geq 3$, and for $k > 1$ and $l \geq 2$.*

In order to prove the conjecture, only one direction remains. We must show that if none of the aforementioned conditions hold, then $X_{m|n_1,n_2}$ is avoidable. There are some partial results towards this goal. In particular there is an easy way of verifying the conjecture for even values of m.

Proposition 7. *[15] Assume m is even and $2m \leq \min(n_1, n_2)$. Then $X_{m|n_1,n_2}$ is avoidable.*

Thus for any fixed even m we only need to verify the conjecture for finitely many values of n_1 and n_2, which is generally easy. For

1. $m = 0$: $X_{0|n_1,n_2}$ is always avoidable, and indeed this is the case.
2. $m = 2$: $X_{2|n_1,n_2}$ is avoidable except for $n_1 = 1, n_2 = 3$ or $n_1 = 3, n_2 = 1$. It is easy to find avoiding two-sided infinite words for other values of n_1 and n_2 less than 5 when $m = 2$. This is all that is necessary to confirm the conjecture for $m = 2$.

In this way the conjecture has been verified for all even m up to very large values via computer.

The odd values of m seem to be much more difficult. The following proposition shows that the conjecture is true for $m = 1$.

Proposition 8. *[15] Conjecture 1 is true for $m = 1$, that is, $X_{1|n_1,n_2}$ is unavoidable if and only if n_1 and n_2 are even numbers with $|n_1 - n_2| = 2$.*

Other results on the avoidability of $X_{m|n_1,n_2}$ include the following.

Proposition 9. *[15]*

1. *Let $s \in \mathbb{N}$ with $s < m - 2$. Then for $n > 2(m+1)^2 + m - 1$, $X_{m|m+s,n} = \{a\diamond^m a, b\diamond^{m+s} b\diamond^n b\}$ is avoidable. Intuitively this means that if m and n_1 are relatively close in value, then the set of integers n_2 which make $X_{m|n_1,n_2}$ unavoidable is finite.*
2. *If $\max(n_1, n_2) < m < n_1 + n_2 + 2$, then $X_{m|n_1,n_2}$ is avoidable.*
3. *The set $X = \{a\diamond^m a, bbb\}$ is avoidable.*

Classifying the unavoidable sets of partial words of size greater than or equal to two remains an open question.

Open problem 26 *Classify the unavoidable sets of partial words of size $l \geq 2$ over a k-letter alphabet where $k \leq l$.*

2.8.3 Testing Unavoidability

Efficient algorithms to determine if a finite set of full words is unavoidable are well known [45, 93]. For example, we can check whether there is a loop in the finite automaton of Aho and Corasick [1] recognizing $A^* \setminus A^*XA^*$. Another approach is the following. We say that a set of words Y is obtained from a finite set of words X by an elementary derivation if

1. *Type 1 elementary derivation:* There exist words $x, y \in X$ such that x is a proper prefix of y, and $Y = X \setminus \{y\}$ (this will be denoted by $X \xrightarrow{1} Y$).
2. *Type 2 elementary derivation:* There exists a word $x = ya \in X$ with $a \in A$ such that, for each letter $b \in A$ there is a suffix z of y such that $zb \in X$, and $Y = (X \setminus \{x\}) \cup \{y\}$ (this will be denoted by $X \xrightarrow{2} Y$).

A *derivation* is a sequence of elementary derivations. We say that Y is derived from X if Y is obtained from X by a derivation. If Y is derived from X, then X is unavoidable if and only if Y is unavoidable.

Example 2. The following sequence of elementary derivations shows that $X = \{aaaa, aaba, abaa, abba, bab, bbb\}$ derives $\{\varepsilon\}$:

$$
\begin{aligned}
X &\xrightarrow{2} \{aaaa, aaba, aba, abba, bab, bbb\} \\
&\xrightarrow{2} \{aaaa, aaba, aba, abb, bab, bbb\} \\
&\xrightarrow{2} \{aaaa, aaba, ab, bab, bbb\} \\
&\xrightarrow{2} \{aaa, aaba, ab, bab, bbb\} \\
&\xrightarrow{2} \{aa, aaba, ab, bab, bbb\} \\
&\xrightarrow{1} \{aa, ab, bab, bbb\} \\
&\xrightarrow{2} \{a, ab, bab, bbb\} \\
&\xrightarrow{1} \{a, bab, bbb\} \\
&\xrightarrow{2} \{a, ba, bbb\} \\
&\xrightarrow{2} \{a, ba, bb\} \\
&\xrightarrow{2} \{a, b, bb\} \\
&\xrightarrow{2} \{a, b\} \\
&\xrightarrow{2} \{\varepsilon, b\} \\
&\xrightarrow{1} \{\varepsilon\}
\end{aligned}
$$

The notion of a derivation gives an algorithm to check whether a set is unavoidable: A finite set X is unavoidable if and only if there is a derivation from X to the set $\{\varepsilon\}$. The above derivation shows that $\{aaaa, aaba, abaa, abba, bab, bbb\}$ is unavoidable.

These algorithms to determine if a finite set of full words is unavoidable, like the one just described, can be used to decide if a finite set of partial words X is unavoidable by determining the unavoidability of $\hat{X}$. However this incurs a dramatic loss in efficiency, as each partial word u in X can contribute as many as $\|A\|^{\|H(u)\|}$ elements to $\hat{X}$. The above derivation shows that $\{a\diamond\diamond a, b\diamond b\}$ is unavoidable as is confirmed by Theorem 31 since $m+1 = 2+1 = 3 = 2^0 3$ and $n+1 = 1+1 = 2 = 2^1$.

In [15], the question was raised as to whether there is an efficient algorithm to determine if a finite set of partial words is unavoidable. In [26], it was proved that testing the unavoidability of a finite set of partial words is much harder to handle than the similar problem for full words. Indeed, the following theorem holds (note that the case $k = 1$ is trivial).

Theorem 33. *[26] The problem of deciding whether a finite set of partial words over a k-letter alphabet where $k \geq 2$ is unavoidable is NP-hard.*

The proof proceeds by reduction from the 3SAT problem that is known to be NP-complete (see [70]). In the 3SAT problem, we are given n binary variables $x_1, \ldots, x_n$ and m clauses that each contain three literals (a literal can be a variable or its negation), and we search a truth assignment for the variables such that each clause has at least one true literal.

In [26], the following related questions on avoidability of sets of partial words were raised.

Open problem 27 *Is the decision problem of the avoidability of a set of partial words in NP?*

A similar (stronger) question is the following one.

Open problem 28 *For any set of partial words X, does there always exist a two-sided infinite periodic word that avoids X, whose period is polynomial in the size of X?*

2.9 Freeness of Partial Words

In [99], Manea and Mercaş introduce freeness of partial words. There, they extend in a natural way the concepts of square- and overlap-freeness of words to partial words. In [31, 30], some more basic freeness properties of partial words are investigated generalizing the well-known freeness properties of full words.

A one-sided infinite word over the alphabet A is a function from $\mathbb{N}$ to A. The Thue-Morse word is an example of a one-sided infinite word defined by iterating a morphism. Let $\phi : \{a, b\}^* \to \{a, b\}^*$ be the morphism defined by $\phi(a) = ab$ and $\phi(b) = ba$. We define $t_0 = a$ and $t_i = \phi^i(a)$, for all $i \geq 1$. Note that $t_{i+1} = \phi(t_i)$ and that $t_{i+1} = t_i\overline{t_i}$, where $\bar{x}$ is the word obtained from x by replacing each occurrence of a with b and each occurrence of b with a. Thus, the limit (the infinite word) $t = \lim_{i\to\infty} t_i$ exists. The Thue-Morse word is defined as t, a fixed point for the morphism ϕ. Computations show that $t_1 = ab$, $t_2 = abba$, $t_3 = abbabaab$, $t_4 = abbabaabbaababba$, and

$$t_5 = abbabaabbaababbabaabbaababbaabbabaab \tag{2.4}$$

and so on.

A one-sided infinite word w is *k-free* if there is no word x such that x^k is a factor of w (a word that is 2-free is also called square-free and a word that is 3-free is called cube-free). It is called *overlap-free* if it does not contain any factor of the form *cycyc* with $c \in A$. Any overlap-free word is clearly k-free for all $k \geq 3$.

Theorem 34. *[120, 121] The Thue-Morse infinite word t is overlap-free and hence k-free for all $k \geq 3$.*

A one-sided infinite partial word w over the alphabet A is a partial function from $\mathbb{N}$ to A. We call w *k-free* if for any nonempty factor $x_1 \dots x_k$ of w, no partial word x exists such that $x_i \subset x$ for all $1 \leq i \leq k$. And it is said to be *overlap-free* if for any factor $c_1y_1c_2y_2c_3$ of w no letter $c \in A$ and partial word y over A exist such that $c_i \subset c$ for all $1 \leq i \leq 3$ and $y_j \subset y$ for all $1 \leq j \leq 2$. In [99], the authors propose an efficient algorithm to test whether or not a partial word of length n is k-free. Both the time and space complexities of the algorithm are $O(\frac{n}{k})$. In case of full words, the time complexity can be

reduced to $O(n \log n)$ using suffix arrays [98]. In [99], the authors also give an efficient algorithm to construct in $O(n)$ time a cube-free (and hence k-free for all $k \geq 3$) partial word with n holes, and modify the algorithm in the case of a four-letter alphabet to produce such a partial word of minimal length $3n-2$ (which is the minimal length among all the possible cube-free words with n holes regardless of the alphabet over which these words are constructed).

Theorem 35. *[99] For $k \geq 3$, there exist infinitely many k-free infinite partial words over a two-letter alphabet containing an arbitrary number of holes.*

Note that it is enough to show the result for $k = 3$. The idea of the proof is to show that there exist infinitely many cube-free infinite partial words containing exactly one hole over a two-letter alphabet. In order to do this, observe that if the underlined b in Equality 2.4 is replaced by $\diamond$, then the resulting partial word is still cube-free. Since there is an infinite number of occurrences of t_5 in t, any replacement of the underlined b in such occurrences leads to an infinite partial word with one hole that is cube-free. The result follows since there is an infinite number of nonoverlapping occurrences of t_5 in t.

A surprising result holds for an alphabet of size four.

Theorem 36. *[99] There exists an infinite cube-free word over a four-letter alphabet in which we can randomly replace letters by holes and obtain in this way an infinite partial word that is cube-free as long as each pair of two consecutive holes are separated by at least two letters of the alphabet. Moreover, such a word does not exist over a three-letter alphabet.*

We discuss the concept of square-freeness of partial words in Section 2.9.1 and of overlap-freeness of partial words in Section 2.9.2.

2.9.1 Square-Freeness

Let us now consider the $k = 2$ case. A well known result from Thue states that over a three-letter alphabet there exist infinitely many infinite words that are square-free [120, 121]. To generalize Thue's result, we wish to find a square-free partial word with infinitely many holes, and an infinite full word that remains square-free even after replacing an arbitrary selection of letters with holes. Unfortunately, every partial word containing at least one hole and having length at least two contains a square (either $a\diamond$ or $\diamond a$ cannot be avoided, where a denotes a letter from the alphabet). Furthermore, it is obvious that if we replace $2n$ consecutive letters in a full word with holes, then the corresponding factor of the resulting partial word will be a square.

Motivated by these observations, we call a word *non-trivial square*-free if it contains no factors of the form $w^k, k \geq 2$, except when $|w| \in \{1, 2\}$ and $k = 2$. Notice that the cube *aaa* is considered to be a non-trivial square. For the sake of readability, we shall use the terms *non-trivial square* and *square*

interchangeably. The study of non-trivial squares is not new. In [106], several iterating morphisms are given for infinite words avoiding non-trivial squares. In particular, the authors give an infinite binary word avoiding both cubes xxx and squares yy with $|y| \geq 4$ and an infinite binary word avoiding all squares except 0^2, 1^2, and $(01)^2$ using a construction that is somewhat simpler than the original one from Fraenkel and Simpson [68].

Remark 1. When we introduce holes into arbitrary positions of a word, we impose the restriction that every two holes must have at least two non-hole symbols between them.

With this restriction, the study of square-free partial words becomes much more subtle and interesting.

Theorem 37. *[31] There exists an infinite word over an eight-letter alphabet that remains square-free after replacing an arbitrary selection of its letters with holes, and none exists over a smaller alphabet.*

A suggested problem for investigation is the following. Let $g(n)$ be the length of a longest binary full word containing at most n distinct squares. How does the sequence $\{g(n)\}$ behave? A complete answer appears in [68].

Open problem 29 *Compute the maximum number of distinct squares in a partial word with h holes of length n over a k-letter alphabet.*

2.9.2 Overlap-Freeness

A well known result of Thue states that over a binary alphabet there exist infinitely many overlap-free words [120, 121]. In [99], the question was raised as to whether there exist overlap-free infinite partial words, and to construct them over a binary alphabet if such exist. The following result settles this question.

Theorem 38. *[31] There exist overlap-free infinite partial words with one hole over a two-letter alphabet, and none exists with more than one hole.*

The following result relates to a three-letter alphabet.

Theorem 39. *[31] There exist infinitely many overlap-free infinite partial words with an arbitrary number of holes over a three-letter alphabet.*

For the following result, we adhere to the restriction described in Remark 1 when replacing an arbitrary selection of letters in a word with holes.

Theorem 40. *[31] There exists an infinite overlap-free word over a six-letter alphabet that remains overlap-free after an arbitrary selection of its letters are changed to holes, and none exists over a four-letter alphabet.*

The case of a five-letter alphabet remains open.

Open problem 30 *Does there exist an infinite word over a five-letter alphabet that remains overlap-free after an arbitrary insertion of holes?*

Other problems are suggested in [31].

Open problem 31 *Extend the concept of square-free (respectively, overlap-free or cube-free) morphism to partial words.*

From [31, 99], some of the properties of this kind of morphisms already start to be obvious. A further analysis might give additional properties that such morphisms should fulfill. Following the approach of Dejean [56], another interesting problem to analyze is the following.

Open problem 32 *Identify the exact value of k (related to k-freeness) for a given alphabet size. This value would represent the repetitiveness threshold in an n-letter alphabet.*

If for full words this value is known for alphabets up to size 11 and it is conjectured that for bigger size alphabets the value is $\frac{n+1}{n}$, for partial words this value has not yet been investigated.

2.10 Other Open Problems

The theory of *codes* has been widely developed in connection with combinatorics on words [5]. In [7, 32], a new line of research was initiated by introducing *pcodes* in connection with combinatorics on partial words, and a theoretical framework for pcodes was developed by revisiting the theory of codes of words, as exposited in [5], starting from pcodes of partial words. Pcodes are defined in terms of the compatibility relation as follows.

Definition 10. *[7] Let X be a nonempty set of partial words over an alphabet A. Then X is called a* pcode *over A if for all positive integers m, n and partial words $u_1, \ldots, u_m, v_1, \ldots, v_n \in X$, the condition*

$$u_1 u_2 \ldots u_m \uparrow v_1 v_2 \ldots v_n$$

implies $m = n$ and $u_i = v_i$ for $i = 1, \ldots, m$.

An area of current interest for the study of pcodes is data communication where some information may be missing, lost, or unknown. While a code of words X does not allow two distinct decipherings of some word in X^+, a pcode of partial words Y does not allow two distinct compatible decipherings in Y^+. Various ways have been described for defining and analyzing pcodes. In particular, many pcodes can be obtained as antichains with respect to certain partial orderings. Adapting a graph technique related to dominoes [6, 73, 79], the pcode property was shown to be decidable for finite sets of partial words.

For example, the set $X = \{a\diamond, a\diamond b\}$ is a pcode over $\{a, b\}$, but the set $Y = \{u_1, u_2, u_3, u_4\}$ where $u_1 = a\diamond b, u_2 = aa\diamond bba, u_3 = \diamond b$, and $u_4 = ba$ is not a pcode over $\{a, b\}$ since $u_1u_3u_3u_4u_3 \uparrow u_2u_3u_1$ is a nontrivial compatibility relation over Y.

It is well known that the two-element set of words $\{u, v\}$ is a code if and only if $uv \neq vu$. However, this is not true in general for partial words. For instance, the set $\{u, v\}$ where $u = a\diamond b$ and $v = abbaab$ satisfies $uv \not\uparrow vu$, but $\{u, v\}$ is not a pcode since $u^2 \uparrow v$.

Open problem 33 *Find a necessary and sufficient condition for a two-element set of partial words to be a pcode.*

Other suggested problems are the following.

Open problem 34 *Investigate the concept of* tiling periodicity *introduced recently by Karhumäki, Lifshits and Rytter's [81]. There, the authors suggest a number of questions for further work on this new concept.*

Punctured languages are sets whose elements are partial words. In [91], Lischke investigated to which extent restoration of punctured languages is possible if the number of holes or the proportion of holes per word, respectively, is bounded, and studied their relationships for different boundings. The considered restoration classes coincide with similarity classes according to some kind of similarity for languages. Thus all results he can also formulate in the language of similarity. He shows some hierarchies of similarity classes for each class of the Chomsky hierarchy, and proves the existence of linear languages which are not δ-similar to any regular language for any $\delta < \frac{1}{2}$.

Open problem 35 *For $\frac{1}{2} \leq \delta$, do there exist linear languages which are not δ-similar to any regular language? If they exist, then they must be* non-slender.

References

1. A.V. Aho and M.J. Corasick. Efficient string machines, an aid to bibliographic research. *Comm. ACM*, 18:333–340, 1975.
2. J.P. Allouche and J. Shallit. *Automatic Sequences: Theory, Applications, Generalizations.* Cambridge University Press, 2003.
3. R. Assous and M. Pouzet. Une caractérisation des mots périodiques. *Discrete Math.*, 25:1–5, 1979.
4. J. Berstel and L. Boasson. Partial words and a theorem of fine and wilf. *Theoret. Comput. Sci.*, 218:135–141, 1999.
5. J. Berstel and D. Perrin. *Theory of Codes.* Academic Press, Orlando, FL, 1985.
6. F. Blanchet-Sadri. On unique, multiset, and set decipherability of three-word codes. *IEEE Trans. Inform. Theory*, 47:1745–1757, 2001.
7. F. Blanchet-Sadri. Codes, orderings, and partial words. *Theoret. Comput. Sci.*, 329:177–202, 2004.

8. F. Blanchet-Sadri. Periodicity on partial words. *Comput. Math. Appl.*, 47:71–82, 2004.
9. F. Blanchet-Sadri. Primitive partial words. *Discrete Appl. Math.*, 148:195–213, 2005.
10. F. Blanchet-Sadri. *Algorithmic Combinatorics on Partial Words*. Chapman & Hall/CRC Press, 2007.
11. F. Blanchet-Sadri and A.R. Anavekar. Testing primitivity on partial words. *Discrete Appl. Math.*, 155:279–287, 2007. (`www.uncg.edu/mat/primitive`).
12. F. Blanchet-Sadri, D. Bal, and G. Sisodia. Graph connectivity, partial words and a theorem of Fine amd Wilf. *Information and Computation*, to appear(`www.uncg.edu/mat/research/finewilf3`).
13. F. Blanchet-Sadri, D.D. Blair, and R.V. Lewis. Equations on partial words. In R. Královic and P. Urzyczyn, editors, *MFCS 2006. 31st International Symposium on Mathematical Foundations of Computer Science*, LNCS, vol. 4162, Springer, pp. 167-178.
14. F. Blanchet-Sadri, L. Bromberg, and K. Zipple. Tilings and quasiperiods of words. Preprint (`www.uncg.edu/cmp/research/tilingperiodicity`).
15. F. Blanchet-Sadri, N.C. Brownstein, and J. Palumbo. Two element unavoidable sets of partial words. In T. Harju, J. Karhumäki, and A. Lepistö, editors, *DLT 2007. 11th International Conference on Developments in Language Theory*, LNCS, Vol. 4588, Springer, pp. 96-107.
16. F. Blanchet-Sadri and A. Chriscoe. Local periods and binary partial words: an algorithm. *Theoret. Comput. Sci.*, 314:189–216, 2004. (`www.uncg.edu/mat/AlgBin`).
17. F. Blanchet-Sadri, E. Clader, and O. Simpson. Border correlations of partial words. Preprint (`www.uncg.edu/cmp/research/bordercorrelation`).
18. F. Blanchet-Sadri, K. Corcoran, and J. Nyberg. Fine and wilf's periodicity result on partial words and consequences. In *LATA 2007, 1st International Conference on Language and Automata Theory and Applications*, GRLMC Report 35/07, Tarragona.
19. F. Blanchet-Sadri and M. Cucuringu. Counting primitive partial words. Preprint.
20. F. Blanchet-Sadri, M. Cucuringu, and J. Dodge. Counting unbordered partial words. Preprint.
21. F. Blanchet-Sadri, C.D. Davis, J. Dodge, R. Mercaş, and M. Moorefield. Unbordered partial words. Preprint (`www.uncg.edu/mat/border`).
22. F. Blanchet-Sadri and S. Duncan. Partial words and the critical factorization theorem. *J. Combin. Theory Ser. A*, 109:221–245, 2005. (`www.uncg.edu/mat/cft`).
23. F. Blanchet-Sadri, J. Fowler, J. Gafni, and K. Wilson. Combinatorics on partial word correlations. Preprint (`www.uncg.edu/cmp/research/correlations2`).
24. F. Blanchet-Sadri, J. Gafni, and K. Wilson. Correlations of partial words. In W. Thomas and P. Weil, editors, *STACS 2007*, volume 4393, pages 97–108, Berlin, 2007. (`www.uncg.edu/mat/research/correlations`).
25. F. Blanchet-Sadri and R.A. Hegstrom. Partial words and a theorem of fine and wilf revisited. *Theoret. Comput. Sci.*, 270:401–419, 2002.
26. F. Blanchet-Sadri, R. Jungers, and J. Palumbo. Testing avoidability of sets of partial words is hard. Preprint.
27. F. Blanchet-Sadri, A. Kalcic, and T. Weyand. Unavoidable sets of partial words of size three. Preprint (`www.uncg.edu/cmp/research/unavoidablesets2`).

28. F. Blanchet-Sadri and D.K. Luhmann. Conjugacy on partial words. *Theoret. Comput. Sci.*, 289:297–312, 2002.
29. F. Blanchet-Sadri, T. Mandel, and G. Sisodia. Connectivity in graphs associated with partial words. Preprint (`www.uncg.edu/cmp/research/finewilf4`).
30. F. Blanchet-Sadri, R. Mercaş, and G. Scott. Counting distinct squares in partial words. Preprint (`www.uncg.edu/cmp/research/freeness`).
31. F. Blanchet-Sadri, R. Mercaş, and G. Scott. A generalization of thue freeness for partial words. Preprint (`www.uncg.edu/cmp/research/freeness`).
32. F. Blanchet-Sadri and M. Moorefield. Pcodes of partial words. Preprint(`www.uncg.edu/mat/pcode`).
33. F. Blanchet-Sadri, T. Oey, and T. Rankin. Partial words and generalized fine and wilf's theorem for an arbitrary number of weak periods. Preprint (`www.uncg.edu/mat/research/finewilf2`).
34. F. Blanchet-Sadri, B. Shirey, and G. Gramajo. Periods, partial words, and a result of guibas and odlyzko. Preprint (`www.uncg.edu/mat/bintwo`).
35. F. Blanchet-Sadri and N.D. Wetzler. Partial words and the critical factorization theorem revisited. *Theoret. Comput. Sci.*, to appear. (`www.uncg.edu/mat/research/cft2`).
36. F. Blanchet-Sadri and J. Zhang. On the critical factorization theorem. Preprint.
37. V.D. Blondel, R. Jungers, and V. Protasov. On the complexity of computing the capacity of codes that avoid forbidden difference patterns. *IEEE Trans. Inform. Theory*, 52:5122–5127, 2006.
38. R.S. Boyer and J.S Moore. A fast string searching algorithm. *Comm. ACM*, 20:762–772, 1977.
39. D. Breslauer, T. Jiang, and Z. Jiang. Rotations of periodic strings and short superstrings. *J. of Algorithms*, 24:340–353, 1997.
40. J. Buhler, U. Keich, and Y. Sun. Designing seeds for similarity search in genomic dna. *J. Comput. System Sci.*, 70:342–363, 2005.
41. P. Bylanski and D.G.W. Ingram. Digital transmission systems. *IEE*, 1980.
42. M.G. Castelli, F. Mignosi, and A. Restivo. Fine and wilf's theorem for three periods and a generalization of sturmian words. *Theoret. Comput. Sci.*, 218:83–94, 1999.
43. Y. Césari and M. Vincent. Une caractérisation des mots périodiques. *C.R. Acad. Sci. Paris*, 268:1175–1177, 1978.
44. C. Choffrut and K. Culik II. On extendibility of unavoidable sets. *Discrete Appl. Math.*, 9:125–137, 1984.
45. C. Choffrut and J. Karhumäki. Combinatorics of words. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*, volume 1, pages 329–438. Springer, Berlin, 1997.
46. D.D. Chu and H.S. Town. Another proof on a theorem of lyndon and schützenberger in a free monoid. *Soochow J. Math.*, 4:143–146, 1978.
47. S. Constantinescu and L. Ilie. Generalised fine and wilf's theorem for arbitrary number of periods. *Theoret. Comput. Sci.*, 339:49–60, 2005.
48. M. Crochemore, C. Hancart, and T. Lecroq. *Algorithms on Strings*. Cambridge University Press, 2007.
49. M. Crochemore, F. Mignosi, A. Restivo, and S. Salemi. Text compression using antidictionaries. *LNCS*, 1644:261–270, 1999.
50. M. Crochemore and D. Perrin. Two-way string matching. *J. of the ACM*, 38:651–675, 1991.

51. M. Crochemore and W. Rytter. *Text Algorithms*. Oxford University Press, 1994.
52. M. Crochemore and W. Rytter. *Jewels of Stringology*. World Scientific, NJ, 2003.
53. A. de Luca. On the combinatorics of finite words. *Theoret. Comput. Sci.*, 218:13–39, 1999.
54. A. de Luca and S. Varricchio. *Regularity and Finiteness Conditions*, volume 1, chapter 11, pages 747–810. Springer, Berlin, 1997.
55. A. de Luca and S. Varricchio. *Finiteness and Regularity in Semigroups and Formal Languages*. Springer, Berlin, 1999.
56. F. Dejean. Sur un théorème de thue. *J. Combin. Theory Ser. A*, 13:90–99, 1972.
57. P. Dömösi. Some results and problems on primitive words. In *11th International Conference on Automata and Formal Languages*, 2005.
58. P. Dömösi, S. Horváth, and M. Ito. *Primitive Words and Context-Free Languages*.
59. J.P. Duval. Périodes locales et propagation de périodes dans un mot. *Theoret. Comput. Sci.*, 204: 87-98, 1998
60. J.P. Duval. Périodes et répétitions des mots du monoïde libre. *Theoret. Comput. Sci.*, 9:17–26, 1979.
61. J.P. Duval. Relationship between the period of a finite word and the length of its unbordered segments. *Discrete Math.*, 40:31–44, 1982.
62. J.P. Duval, R. Kolpakov, G. Kucherov, T. Lecroq, and A. Lefebvre. Linear-time computation of local periods. *Theoret. Comput. Sci.*, 326:229–240, 2004.
63. A. Ehrenfeucht, D. Haussler, and G. Rozenberg. On regularity of context-free languages. *Theoret. Comput. Sci.*, 27:311–322, 1983.
64. A. Ehrenfeucht and D.M. Silberger. Periodicity and unbordered segments of words. *Discrete Math.*, 26:101–109, 1979.
65. P. Erdös. Note on sequences of integers no one of which is divisible by another. *J. London Math. Soc.*, 10:126–128, 1935.
66. M. Farach-Colton, G.M. Landau, S.C. Sahinalp, and D. Tsur. Optimal spaced seeds for approximate string matching. In L. Caires, G.F. Italiano, L. Monteiro, C. Palanidessi, and M. Yung, editors, *ICALP 2005*, LNCS, vol. 3580, Springer, pp. 1251-1262, 2005.
67. N.J. Fine and H.S. Wilf. Uniqueness theorems for periodic functions. In *Proc. Amer. Math. Soc.*, volume 16, pages 109–114, 1965.
68. A.S. Fraenkel and R.J. Simpson. How many squares must a binary sequence contain? *Electron. J. Combin.*, 2, 1995.
69. Z. Galil and J. Seiferas. Time-space optimal string matching. *J. Comput. System Sci.*, 26:280–294, 1983.
70. M.R. Garey and D.S. Johnson. *Computers and Intractability -A Guide to the Theory of NP-Completeness*. Freeman, 1979.
71. L.J. Guibas and A.M. Odlyzko. Periods in strings. *J. Combin. Theory Ser. A*, 30:19–42, 1981.
72. D. Gusfield. *Algorithms on Strings, Trees, and Sequences*. Cambridge University Press, Cambridge, 1997.
73. F. Guzmán. Decipherability of codes. *J. Pure Appl. Algebra*, 141:13–35, 1999.
74. V. Halava, T. Harju, and L. Ilie. Periods and binary words. *J. Combin. Theory Ser. A*, 89:298–303, 2000.

75. T. Harju. *Combinatorics on Words*, chapter 19, pages 381–392. Springer, Berlin, 2006.
76. T. Harju and D. Nowotka. Periodicity and unbordered segments of words. *Bull. Eur. Assoc. Theor. Comput. Sci. EATCS*, 80:162–167, 2003.
77. T. Harju and D. Nowotka. The equation $x^i = y^j z^k$ in a free semigroup. *Semigroup Forum*, 68:488–490, 2004.
78. T. Head, G. Paun, and D. Pixton. *Language Theory and Molecular Genetics*, volume 2, chapter 7, pages 295–360. Springer, Berlin, 1997.
79. T. Head and A. Weber. Deciding multiset decipherability. *IEEE Trans. Inform. Theory*, 41:291–297, 1995.
80. J. Justin. On a paper by castelli, mignosi, restivo. *Theoret. Inform. Appl.*, 34:373–377, 2000.
81. J. Karhumäki, Y. Lifshits, and W. Rytter. Tiling periodicity. In *CPM 2007, 18th Annual Symposium on Combinatorial Pattern Matching*, 2007.
82. L. Kari, G. Rozenberg, and A. Salomaa. *L Systems*, volume 1, chapter 7, pages 253–328. Springer, Berlin, 1997.
83. U. Keich, M. Li, B. Ma, and J. Tromp. On spaced seeds for similarity search. *Discrete Appl. Math.*, 138:253–263, 2004.
84. D.E. Knuth, J.H. Morris, and V.R. Pratt. Fast pattern matching in strings. *SIAM J. on Comput.*, 6:323–350, 1977.
85. R. Kolpakov and G. Kucherov. Finding approximate repetitions under hamming distance. *Lecture Notes in Computer Science*, 2161:170–181, 2001.
86. R. Kolpakov and G. Kucherov. Finding approximate repetitions under hamming distance. *Theoret. Comput. Sci.*, 33:135–156, 2003.
87. G. Landau and J. Schmidt. An algorithm for approximate tandem repeats. *Lecture Notes in Computer Science*, 684:120–133, 1993.
88. G.M. Landau, J.P. Schmidt, and D. Sokol. An algorithm for approximate tandem repeats. *J. Comput. Biology*, 8:1–18, 2001.
89. P. Leupold. Languages of partial words - how to obtain them and what properties they have. *Grammars*, 7:179–192, 2004.
90. P. Leupold. Partial words for dna coding. *LNCS*, 3384:224–234, 2005.
91. G. Lischke. Restorations of punctured languages and similarity of languages. *Math. Logic Quart.*, 52:20–28, 2006.
92. M. Lothaire. *Combinatorics on Words.* Cambridge University Press, Cambridge, 1997.
93. M. Lothaire. *Algebraic Combinatorics on Words.* Cambridge University Press, Cambridge, 2002.
94. M. Lothaire. *Applied Combinatorics on Words.* Cambridge University Press, Cambridge, 2005.
95. R.C. Lyndon and P.E. Schupp. *Combinatorial Group Theory.* Springer, Berlin, 2001.
96. R.C. Lyndon and M.P. Schützenberger. The equation $a^m = b^n c^p$ in a free group. *Michigan Math. J.*, 9; 289–298, 1962.
97. B. Ma, J. Tromp, and M. Li. Patternhunter: faster and more sensitive homology search. *Bioinformatics*, 18:440–445, 2002.
98. U. Manber and G. Myers. Suffix arrays: a new method for on-line string searches. *SIAM J. on Comput.*, 22:935–948, 1993.
99. F. Manea and R. Mercas. Freeness of partial words. Preprint.

100. D. Margaritis and S. Skiena. Reconstructing strings from substrings in rounds. In *FOCS 1995, 36th Annual Symposium on Foundations of Computer Science*, pages 613–620, 1995.
101. E.M. McCreight. A space-economical suffix tree construction algorithm. *J. of the ACM*, 23:262–272, 1976.
102. F. Mignosi, A. Restivo, and S. Salemi. A periodicity theorem on words and applications. *LNCS*, 969:337–348, 1995.
103. F. Nicolas and E. Rivals. Hardness of optimal spaced seed design. In Apostolico, A., Crochemore, M. and Park, K., editors, *CPM 2005, 16th Annual Symposium on Combinatorial Pattern Matching*, LNCS, vol. 3537, Srpinger, pages 144–155, 2005.
104. L. Noé and G. Kucherov. Improved hit criteria for dna local alignment. *BMC Bioinformatics*, 5, 2004.
105. H. Petersen. On the language of primitive words. *Theoret. Comput. Sci.*, 161:141–156, 1996.
106. N. Rampersad, J. Shallit, and M. w Wang. Avoiding large squares in infinite binary words. *Theoret. Comput. Sci.*, 339:19–34, 2005.
107. G. Richomme. Sudo-lyndon. *Bull. Eur. Assoc. Theor. Comput. Sci. EATCS*, 92:143–149, 2007.
108. E. Rivals and S. Rahmann. Combinatorics of periods in strings. *J. Combin. Theory Ser. A*, 104:95–113, 2003.
109. L. Rosaz. Unavoidable languages, cuts and innocent sets of words. *RAIRO Theoret. Inform. Appl.*, 29:339–382, 1995.
110. L. Rosaz. Inventories of unavoidable languages and the word-extension conjecture. *Theoret. Comput. Sci.*, 201:151–170, 1998.
111. J.P. Schmidt. All highest scoring paths in weighted grid graphs and their application to finding all approximate repeats in strings. *SIAM J. Comput.*, 27:972–992, 1998.
112. J. Setubal and J.Meidanis. *Introduction to Computational Molecular Biology.* PWS Publishing Company, Boston, MA, 1997.
113. A.M. Shur and Y.V. Gamzova. Periods' interaction property for partial words. In T. Harju and J. Karhümaki, editors, *Words 2003*, volume 27, pages 75–82, 2003.
114. A.M. Shur and Y.V. Gamzova. Partial words and the periods' interaction property. *Izv. RAN*, 68:199–222, 2004. (see Shur, A.M., Gamzova, Y.V.: Partial words and the interaction property of periods. Izv. Math. **68** (2004) 405–428, for the English translation).
115. A.M. Shur and Y.V. Konovalova. On the periods of partial words. *LNCS*, vol. 2136, Springer, pp. 657–665, 2001.
116. H.J. Shyr. *Free Monoids and Languages.* Hon Min Book Company, Taichung, Taiwan, 1991.
117. H.J. Shyr and G. Thierrin. Disjunctive languages and codes. *LNCS*, vol.56, Springer, pp. 171–176, 1977.
118. W.F. Smyth. *Computing Patterns in Strings.* Pearson Addison-Wesley, 2003.
119. J.A. Storer. *Data Compression: Methods and Theory.* Computer Science Press, Rockville, MD, 1988.
120. A. Thue. Uber unendliche zeichenreihen. *Norske Vid. Selsk. Skr. I, Mat. Nat. Kl. Christiana*, 7:1–22, 1906. Reprinted in Nagell, T., Selberg, A., Selberg, S., Thalberg, K. (eds.): Selected Mathematical Papers of Axel Thue. Oslo, Norway, Universitetsforlaget (1977) 139–158.

121. A. Thue. Uber die gegenseitige lage gleicher teile gewisser zeichenreihen. *Norske Vid. Selsk. Skr. I, Mat. Nat. Kl. Christiana*, 12:1–67, 1912. Reprinted in Nagell, T., Selberg, A., Selberg, S., Thalberg, K. (eds.): Selected Mathematical Papers of Axel Thue. Oslo, Norway, Universitetsforlaget (1977) 139–158.
122. R. Tijdeman and L. Zamboni. Fine and wilf words for any periods. *Indag. Math.*, 14:135–147, 2003.
123. J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Trans. Inform. Theory*, 23:337–343, 1977.