Combinatorics on Words

Tero Harju

Department of Mathematics
University of Turku
FIN-20014 Turku, Finland
email: harju@utu.fi

1 Introduction

Words (strings of symbols) are fundamental in computer processing. Indeed, each bit of data processed by a computer is a string, and nearly all computer software use algorithms on strings. There are also abundant supply of applications of these algorithms in other areas such as data compression, DNA sequence analysis, computer graphics, cryptography, and so on.

Combinatorics on words belongs to discrete mathematics and theoretical computer science, and it is, also historically, close to many branches of mathematics and computer science. Related areas in discrete mathematics include automata, formal languages, probability, semigroups, groups, dynamical systems, combinatorial topology and number theory.

The history of combinatorics on words goes back almost 100 years to Axel Thue and his work on repetitions in words. Systematic study of combinatorial properties on words was initiated in late 1950s by Marcel Schützenberger. Good overviews of the present state of art can be found in the books of Lothaire [11, 12] and in the survey chapter [5] by Choffrut and Karhumäki in the Handbook of Formal Languages.

The generic topic of the combinatorics on words is to study general properties of words (strings or sequences of discrete events), as well as sets, functions and sequences of words. The theory covers both finite and infinite words. In these lectures we concentrate mostly on (special problems of) finite words.

Combinatorics on words has the Mathematical Reviews classification **68R15**.

2 First notions

Let A be a finite alphabet. Then A^* denotes the set of all words (or strings) over A, that is, A^* consists of all concatenations $a_1a_2\dots a_n$ of letters $a_i \in A$. We include the empty word λ in A^*. For a subset $X \subseteq A^*$ of words, let X^+ and $X^* = X^+ \cup \{\lambda\}$ be the *semigroup* and the *monoid*, respectively, generated by X. Hence X^+ is the set of all concatenations of words from X. If $\lambda \in X$, then $X^* = X^+$. For $w \in A^*$, we shall write simply $w^* = \{w\}^*$. $A^{\mathbb{N}}$ will denote the set of all *infinite words* in the alphabet A:

$$w = a_1a_2\dots \quad (a_i \in A).$$

$A^{\mathbb{Z}}$ denotes the set of all *bi-infinite words* in the alphabet A:

$$w = \dots a_{-2}a_{-1}a_0a_1a_2\dots \quad (a_i \in A).$$

Recall that the length $|w|$ of a word $w \in A^*$ is the number of occurrences of letters in it. If $w = a_1a_2\dots a_n$ with $a_i \in A$, then $|w| = n$. For the empty word, we have $|\lambda| = 0$. We write w^n for the nth *power* $ww\dots w$ (n times).

Let $u, v \in A^+$ be two words. Then u is a *factor* (or a *substring*) of v, if $v = w_1uw_2$ for some $w_1, w_2 \in A^*$; u is a *prefix* of v, if $v = uw$ for some $w \in A^*$; u is a *suffix* of v, if $v = wu$ for some $w \in A^*$; u is a *subsequence* (or a *subword*) of v if $v = v_1u_1v_2u_2\dots u_nv_{n+1}$, where $u = u_1u_2\dots u_n$.

Example 2.1. Let $u, v \in A^*$ be such that $uw = wv$ for some word w. Then one can show that u and v are *conjugates*, that is, there are words $x, y \in A^*$ such that $u = xy$ and $v = yx$. In this case, also $w = (xy)^kx$ for some $k \geq 0$.

A mapping $\alpha\colon A^* \to B^*$, where A and B are alphabets, is a *morphism*, if it satisfies the condition

$$\alpha(uv) = \alpha(u)\alpha(v) \quad \text{for all } u, v \in A^*.$$

In particular, $\alpha(a_1a_2\dots a_n) = \alpha(a_1)\alpha(a_2)\dots\alpha(a_n)$ for all $a_i \in A$ and $n \geq 1$, and thus a morphism is determined by the images of the letters.

3 A short cavalcade of topics

We shall now mention some examples of popular problems in combinatorics on words. The list in the below is by no means exhaustive.

3.1 Unavoidable patterns

By a *pattern* we mean a word $p \in X^*$, where X is an alphabet. We say that the pattern p *occurs* in a word $w \in A^*$ (over an alphabet A), if there exists a morphism $\alpha\colon X^* \to A^*$ such that $\alpha(p)$ is a factor of w.

Example 3.1. Consider a sequence of events of throwing a coin. In this case, the alphabet can be chosen to be $A = \{h, t\}$, where h denotes 'head' and t denotes 'tail'. Typically we have words such as $tththht$ that represent sequences resulting in coin throwing. Let $X = \{x, y\}$, and consider the pattern $p = xyxy$. Then p occurs in $w = htthhtthhtt$, since $w = ht(thh)(t)(thh)(t)t$, and so the required morphism can be defined by $\alpha(x) = thh$ and $\alpha(y) = t$. Now $\alpha(xyxy) = thhtthht$ is a factor of w.

If the pattern p does not occur in w, then w is said to *avoid* p. We shall also say that a pattern p is *unavoidable* in an alphabet A, if p occurs in every sufficiently long word $w \in A^*$, that is, if there exists a constant c (depending only on card(A)) such that $|w| \geq c$ implies that p occurs in w.

Of particular interest are the patterns that are the powers, $p = x^k$ for $x \in X$. As a special case, a word is said to be *square-free* if it avoids the pattern x^2, and it is *cube-free* if it avoids the pattern x^3.

Example 3.2. There are only finitely many square-free words in $\{a,b\}^*$. Hence repetition in coin throwing is unavoidable. Indeed, the pattern $p = xx$ occurs in every word $w \in \{a,b\}^*$ of length at least 4.

On the other hand, Thue showed that x^2 can be avoided in larger alphabets. For a survey of Thue's work, see Berstel [2].

Theorem 3.3 (Thue (1906)). *There are arbitrarily long square-free words over any alphabet A of at least 3 letters.*

Hence, e.g., repetition is avoidable in throwing dice (where the alphabet has six letters, $A = \{1, 2, \ldots, 6\}$).

It is an open problem in general to determine which patterns $p \in X^*$ are unavoidable in an alphabet of k letters?

Bean, Ehrenfeucht and McNulty [1], and, independently, Zimin [15] gave a characterization of those words which are avoidable on sufficiently large alphabets. According to this characterization, a pattern $p \in X^*$ with $\text{card}(X) = n$ is avoidable on *some alphabet*, if the word Z_n avoids p, where the words $Z_n \in \{1, 2, \ldots, n\}^*$ are defined as follows:

$$Z_1 = 1, \qquad Z_{n+1} = Z_n(n+1)Z_n.$$

Example 3.4. Consider the pattern $p = xyzxzy$ (with $n = 3$). Then $Z_3 = 1213121$ avoids p, and therefore p is avoidable for sufficiently large alphabets. (How large is sufficiently large here?)

There are no known avoidable patterns that are not avoidable on a 4-letter alphabet. The avoidable binary patterns $p \in \{x,y\}^*$ have been completely characterized by Casseigne and Roth, see [4].

3.2 Local properties of words

Quite many problems relate global properties of words to local ones. In this problem setting one can ask which local properties determine the words (with given properties).

Example 3.5. Let S be a set of words, and define $f_w \colon S \to \mathbb{N}$ as follows:

$$f_w(s) = \text{the number of factors } s \text{ in w.}$$

For which sets S, does the function f_w determine w? That is, when does $f_u = f_v$ for all $u, v \in S$ imply $u = v$? If $S = \{ab, ba\}$, then the answer is negative, since for $u = abba$ and $v = baab$ we have

$$f_u(ab) = 1 = f_v(ab) \text{ and } f_u(ba) = 1 = f_v(ba).$$

On the other hand, for $S = A^+$, of course, f_w determines every word $w \in A^*$.

Example 3.6. It is not always easy to show that a word is not *ultimately periodic*, that is, of the form $w = uvv\ldots$, where the word v repeats itself infinitely many times. As an example, consider the infinite *Kolakoski word*:

$$w = 22112122\ldots$$

which is a self-similar word $w = w_1w_2\ldots$ where $w_{2i} \in \{1, 11\}$ and $w_{2i+1}\{2, 22\}$ such that $|w_i| =$ the i^{th}letter of w.

For $w \in A^*$ or $w \in A^{\mathbb{N}}$,

$$C_n(w) = \text{the number of factors of length } n \text{ in } w$$

Example 3.7. The infinite *Fibonacci word* f is defined as the limit of the following sequence: $f_1 = 1, f_2 = 12, \quad f_{n+2} = f_{n+1}f_n$, so that $f = 12112121\ldots$ Then it can be shown that $C_n(f) = n+1$. An infinite word is *Sturmian* if it satisfies $C_n(w) = n+1$. Sturmian words have many characterizations, see [12].

Theorem 3.8 (Morse and Hedlund (1940)). *An infinite word $w \in \{a,b\}^{\mathbb{N}}$ is not ultimately periodic if and only if $C_n(w) \geq n+1$ for all n.*

3.3 Periodicity

Let $w \in A^+$ be a nonempty word. An integer $1 \leq p \leq |w|$ is a *period* of w if $a_i = a_{i+p}$ for all i, $1 \leq i \leq |w| - p$. Let $\partial(w)$ be the *minimum period* of w:

$$\partial(w) = \min(\mathrm{Per}(w)) \quad \text{where} \quad \mathrm{Per}(w) = \{p \mid p \text{ is a period of } w\}.$$

Note that we always have $|w| \in \mathrm{Per}(w)$, and hence $\partial(w) \leq |w|$ for all words w.

Example 3.9. An integer p with $1 \leq p \leq |w|$ is a period of w iff there is a word v such that $|v| = p$ and w is a factor of v^n for some n. (1) Let $w = aabaaa$ with $|w| = 6$. Then $\mathrm{Per}(w) = \{4,5,6\}$, and $\partial(w) = 4$. Here w is a factor of $(aaba)^2$, $(aabaa)^2$, and w. (2) Let $w = abacabaaabacaba$ with $|w| = 15$. Then

$$w = abacabaa \cdot abac \cdot ab \cdot a$$

and $\mathrm{Per}(w) = \{8, 12, 14, 15\}$, and $\partial(w) = 8$.

The following theorem is a corner stone of periodicity studies. It is due to Fine and Wilf (1965). The result improves the fact that if $f \colon \mathbb{N} \to \mathbb{N}$ is a function with two periods p and q, then also $\gcd(p,q)$ is a period of f.

Theorem 3.10 (Fine and Wilf). *Let $p, q \in \mathrm{Per}(w)$ and $d = \gcd(p,q)$. If $|w| = p + q - d$ then also $d \in \mathrm{Per}(w)$.*

Among other things Guibas and Odlyzko proved in 1981 that each word in an arbitrarily large alphabet has a corresponding word in the binary alphabet having exactly the same periods. For a short proof of this result, see [7].

Theorem 3.11. *Let $w \in A^*$ be a word. There exists a binary word $w' \in \{a,b\}^*$ such that $\mathrm{Per}(w') = \mathrm{Per}(w)$.*

3.4 Critical factorizations

The critical factorization theorem relates local and global periodicity of words. It was first considered by Schützenberger in 1976, and proved by Césari and Vincent in 1978. The present stronger form of the result is due to Duval (1979, 1983). For a short proof, see Crochemore and Perrin in 1991, or the modification given in [10].

Let $w \in A^+$ be a fixed word. A nonempty word $z \neq \lambda$ is called a *central repetition at point k* if $w = uv$ with $|u| = k$ and there are words x, y such that z is a suffix of xu and a prefix of vy. Let

$$\Gamma(w,k) = \{z \mid z \text{ central repetition at } k \text{ in } w\}.$$

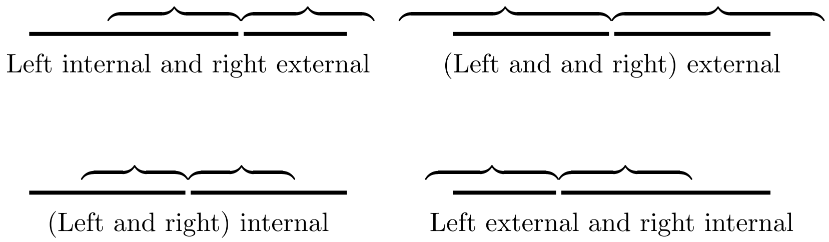
Left internal and right external
(Left and and right) external
(Left and right) internal
Left external and right internal

Figure 1: Choices for the central repetitions

In Figure 1 we have illustrated the four possibilities for a central repetition to occur at a given position.

Example 3.12. Let $w = abaabab$. Then the central repetitions of w at the position 3 are $\Gamma(w, 3) = \{a, aba, ababa, \ldots\}$. Of these a and aba are internal: $ab(a.a)bab$ and $(aba.aba)b$. The rest are (left and right) external. For $k = 1$, $ba \in \Gamma(w, 1)$ is left external and right internal $(\mathtt{b}a.ba)abab$.

For a point k in w, let

$$\partial(w, k) = \min\{\ |z| \mid z \in \Gamma(w, k)\ \}.$$

Hence $\partial(w, k)$ is the length of the shortest central repetition of w at k.

Example 3.13. Let $w = abaaaba$. The following table gives the values for $\partial(w, i)$ and the smallest central repetitions at each point i:

$\partial(w, 1) = 2$	$(z_1 = ba)$	$\partial(w, 2) = 4$	$(z_2 = aaab)$
$\partial(w, 3) = 1$	$(z_3 = a)$	$\partial(w, 4) = 1$	$(z_4 = a)$
$\partial(w, 5) = 4$	$(z_5 = baaa)$	$\partial(w, 6) = 2$	$(z_6 = ab)$

A point k in w is *critical* if

$$\partial(w, k) = \partial(w).$$

Theorem 3.14 (Critical Factorization Theorem). *Every word w, $|w| \geq 2$, has a critical point.*

The proof of Theorem 3.14 states that a critical point of a word w can be obtained as follows: Let the alphabet A be ordered, $a_1 \lhd a_2 \lhd \ldots \lhd a_n$. Then the inverse order is $a_n \lhd^{-1} a_{n-1} \lhd^{-1} \ldots \lhd^{-1} a_1$. Consider the lexicographic ordering of the words w.r.t $\lhd$:

$$u \lhd v \iff u \text{ prefix of } v \ \text{ or } u = xau', v = xbv' \text{ with } a \lhd b.$$

Let s and r be the maximal suffixes of $w = us = vr$ w.r.t. the orders $\lhd$ and $\lhd^{-1}$, respectively. If $|s| < |r|$, then $|u|$ is a critical point of w, and otherwise $|v|$ is a critical point of w.

Example 3.15. Let $w = acbabb$ with $a \lhd b \lhd c$. Then $cbabb$ is the maximal suffix w.r.t. $\lhd$. However, the point 1 is not critical. Indeed, here $\partial(w, 1) = 3$, since w has a central repetition cba at 1: $(\mathtt{cb}a.cba)bb$. But $\partial(w) = 6$, since the word w is unbordered. W.r.t. the inverse order $\lhd^{-1}$ (that is, $c \lhd^{-1} b \lhd^{-1} a$), the maximal suffix is abb, and a critical point is given by this.

Theorem 3.16. *Let $|w| \geq 2$. Each set of $\partial(w) - 1$ consecutive points in w has a critical point.*

Corollary 3.17. *Every primitive word w has a conjugate that is unbordered. That is, $w = uv$ such that vu is unbordered.*

Example 3.18. Let $w = aabaaabaa$, and let the ordering of $A = \{a, b\}$ be $a \lhd b$. The words $aaabaa$ and $baaabaa$ are the maximal suffixes w.r.t. $\lhd$ and $\lhd^{-1}$. Here $\partial(w) = 4$, and $k = 3$ is a critical point, $aab.aaabaa$, and $k = 6$ is a critical point, $aabaaa.baa$.

Let $X \subseteq A^*$ be a set of words. A factorization $w = x_1 x_2 \dots x_n$ is an X-*interpretation* of w if $x_i \in X$ for $1 < i < n$, and x_1 is a suffix of some $x \in X$, x_n is prefix of some $y \in X$. Two X-interpretations $x_1 x_2 \dots x_n$ and $y_1 y_2 \dots y_m$ of w are said to be *disjoint* if $|x_1 x_2 \dots x_i| \neq |y_1 y_2 \dots y_j|$ for all $i < n,\ j < m$.

Example 3.19. Let $X = \{cab, caca, bcabba\}$. Then $w = abbacacacabca$ has the following two disjoint X-interpretations:

$$ab \cdot baca \cdot caca \cdot bca,$$
$$abba \cdot caca \cdot cab \cdot ca.$$

Here ab ($abba$, resp.) is a suffix of cab ($bcabba$), and bca (ca, resp.) is a prefix of $bcabba$ (cab).

Theorem 3.20. *Let $X \subseteq A^+$ with* $\operatorname{card}(X) = n$*. If $w \in A^+$ has $n + 1$ disjoint X-interpretations, then the minimum period of w is at most the maximum of the minimum periods of $x \in X$.*

Example 3.21. The above result is optimal. Indeed, see [8], consider

$$w = (ba^{n+1})^k b \qquad \text{for } k \geq 1, n \geq 0, \quad X = \{a^i b a^{n-i+1} \mid i = 1, 2, \dots, n\}.$$

Then $\operatorname{card}(X) = n$, $\partial(w) = n + 2$ and $\partial(u) \leq n + 1$ for all $u \in X$. The word w has n disjoint X-factorizations:

$$\begin{aligned} w &= ba \cdot a^n ba \dots a^n ba \cdot a^n b \\ &= ba^2 \cdot a^{n-1} ba^2 \dots a^{n-1} ba^2 \cdot a^{n-1} b \\ &\ \ \vdots \\ &= ba^n \cdot aba^n \dots aba^n \cdot ab. \end{aligned}$$

3.5 Reconstruction of events

How to reconstruct a word w from its 'short' factors? This kind of problem appear, for instance, in analysing DNA molecules, because the automated sequencing machines can decipher only relatively short DNA fragments (about 500 base pairs long).

In *shotgun sequencing* of (very long) DNA molecules are first (1) randomly cleaved into short (overlapping) fragments, then (2) the fragments are sequenced into strings, that is words in the alphabet {`A`, `C`, `G`, `T`}, and, finally, (3) a superstring is constructed that represents the original DNA molecule.

In combinatorics of words the reconstruction problem is stated as follows: Given a set X of words, how to construct a shortest word w such that w has the elements of X as its factors. The problem is known to be NP-hard.

There are many related problems that are algorithmically important:

- Finding longest common factor of two words.
- Finding longest repeated factors in a word, see [3] and de Luca [6].
- Finding longest common subsequence of two words. The Unix program `diff` compares two versions of the same file by finding a longest common subsequence of the lines of the files. This problem occurs also in updating display screens, where the two words correspond to the contents of the current screen and its update.
- Finding an alignment between two words u and v. This is needed especially in analysing DNA. For instance, the two words $u =$ `CAGCGTA` and $v =$ `CAGCACTTGGATTCTCA` can be aligned as in the below.

```
---CAGCGT---A------
CAGCA-C-TTGGATTCTCA
```

String matching is a central topic in the domain of text processing. Indeed, string matching algorithms are utilized in practical implementations of all text processing software.

In this problem area we look for given factors or subsequences from input words: Given two words u and v from A^*, how to compute the number of times that u occurs as a factor (or a subsequence) in v? How to find the first occurrence of u as a factor (or a subsequence) in v?

These questions are algorithmic in nature, that is, one seeks for fast(er) algorithms to solve the problems. The problems have also *approximate* variants where one allows some mismatches in the words.

3.6 Uniqueness conditions (freeness)

The freeness problem concerns sets of words. It is stated as follows: Does a given set of words $x_1, \ldots, x_n$ satisfy a nontrivial relation? That is, is there a sequence of events $w = x_{i_1} x_{i_2} \ldots x_{i_k}$ which has an obscure history: also $w = x_{j_1} x_{j_2} \ldots x_{j_m}$. This problem is important in transmission of coded information:

$$\begin{tikzcd}
\boxed{\text{sender}} \arrow[r, "\text{code}", "\text{message}"'] & \boxed{\text{channel}} \arrow[r, "\text{decode}", "\text{message}"'] & \boxed{\text{reader}} \\
i_1 \ldots i_n & w & j_1 \ldots j_m
\end{tikzcd}$$

Let $S \subseteq A^+$ be a subset of words. Then S is an *F-semigroup* if S is closed under concatenation, that is, if $S = S^+$. A subset $X \subseteq A^*$ is a *code* and the F-semigroup X^+ is called *free* if each word $w \in X^+$ has a unique factorization in terms of X: if $x_i, y_i \in X$, then

$$x_1 x_2 \ldots x_n = y_1 y_2 \ldots y_m \implies n = m \text{ and } x_i = y_i.$$

Example 3.22. (1) The set $X = \{a, ab, ba\}$ is not a code, since aba has two different factorizations, $a \cdot ba = ab \cdot a$, in terms of X.

(2) The set $X = \{a, ab, bb\}$ is a code. Indeed, if X were not a code, then there would exist a word $w \in X^+$ with two different X-factorizations,

$$x_1 x_2 \ldots x_n = y_1 y_2 \ldots y_m \quad \text{with } x_1 \neq y_1 \text{ and } x_n \neq y_m.$$

However, every word $w \in X^+$ has a unique suffix from X.

Let $S \subseteq A^+$ be an F-semigroup. The *base* of S is defined by

$$\mathrm{base}(S) = S - S^2,$$

that is, $\mathrm{base}(S)$ is the set of all words $w \in S$ that cannot be expressed as a product of two or more words from S. The *rank* of S is the size of the base,

$$\mathrm{rank}(S) = \mathrm{card}(\mathrm{base}(S)).$$

Notice that the rank of an F-semigroup can be infinite, that is, $\mathrm{base}(S)$ can be an infinite subset of S.

Example 3.23. (1) Let $X = \{ab, ba, abab, abb, bab\}$. Then X is not a code, since $abab = ab \cdot ab$. Most certainly, $\mathrm{base}(X^+) \subseteq X$. Moreover, $abab \notin \mathrm{base}(X^+)$, since it can be factored to two shorter words of X, $ab \cdot ab$. We have $\mathrm{base}(X^+) = \{ab, ba, abb, bab\}$ and hence $\mathrm{rank}(X^+) = 4$.

(2) Let $S = \{w \mid w \text{ contains } a\}$. This is an F-semigroup, and now

$$\mathrm{base}(S) = \{w \mid w \text{ contains exactly one } a\}$$

is infinite, that is, $\mathrm{rank}(S) = \infty$.

Later we consider mostly only those cases where the ranks are finite.

Lemma 3.24. *Let $S = S^+$. Then* $\mathrm{base}(S)$ *is the minimum generating set of S:*

$$S = \mathrm{base}(S)^+ \quad \textit{and} \quad S = X^+ \implies \mathrm{base}(S) \subseteq X.$$

If a word $w \in S$ can be written as

$$w = uxv \ \text{ where } u, ux \in S \ \text{ and } \ xv, v \in S \quad (x \neq \lambda)$$

then x is called an *overflow* of S, see Figure 2.

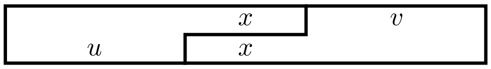

Figure 2: Overflow x

The following criterium of freeness is due to Schützenberger.

Theorem 3.25. *An F-semigroup $S \subseteq A^+$ is free if and only if every overflow of S is in S.*

3.7 The defect effect

The defect theorem is often referred to as a folklore result. In print it has appeared at least in the article by Skordev and Sendov [14] in 1961. The result is very basic to combinatorics on words and its applications. It also bears resemblance to the notion of dimension in linear algebra, see [9].

The defect theorem states that if a set of n words satisfies a nontrivial relation, then these words can be expressed as products of at most $n-1$ (possibly other) words. For the proof of the defect theorem, we need Tilson's result:

Theorem 3.26. *If S_i $(i \in I)$ are free F-semigroups of A^+, then $S = \bigcap_{i\in I} S_i$ is free or it is empty.*

By Theorem 3.26, for any $X \subseteq A^*$,

$$\widehat{X} = \bigcap\{S \mid X \subseteq S,\ S \text{ is a free subsemigroup}\}$$

is a free semigroup, called the *free hull* of X. It is clearly *the smallest free F-semigroup containing X.*

Example 3.27. Let $A = \{a, b\}$ and $X = \{bb, bbaba, abaa, baaba, baa\}$. Consider any free F-semigroup S with $X^+ \subseteq S$. Notice that X^+ is not free, since

$$bbaba \cdot abaa = bb \cdot abaa \cdot baa.$$

Therefore a and aba are overflows of X^+, and thus also of S. So $a \in S$. Next observe that $baaba \cdot abaa = baa \cdot baa \cdot baa$, and so ba is an overflow of S. Consequently, $ba \in S$.

Hence $\{a, ba, bb\} \subseteq S$, and since $\{a, ba, bb\}$ is a code, $\widehat{X} = \{a, ba, bb\}^+$. Notice that

$$\operatorname{rank}(X) = \operatorname{card}(X) = 5 > 3 = \operatorname{rank}(\widehat{X}).$$

Theorem 3.28 (Defect theorem). *Let $X \subseteq A^+$ be finite. If X is not a code, then*

$$\operatorname{rank}(\widehat{X}) \leq \operatorname{card}(X) - 1.$$

A word $w \in A^+$ is said to be *primitive*, if it is not a proper power of another word, that is,

$$w = u^k \implies k = 1 \text{ and } u = w.$$

Corollary 3.29. *Each word $w \in A^+$ is a power of a unique primitive word.*

Corollary 3.30. *Two words $u, v \in A^*$ commute, $uv = vu$ iff they are powers of a common word.*

3.8 Test sets

The result of this section was conjectured by Ehrenfeucht in the beginning of the 1970s in the following language theoretic setting:

Theorem 3.31. *Let $L \subseteq A^*$ be any set of words in a (finite) alphabet A. Then there exists a* finite *subset $F \subseteq L$ such that if two morphisms α and β agree on T (that is, $\alpha(w) = \beta(w)$ for all $w \in T$) then α and β agree on the whole L.*

Such a subset $T \subseteq L$ is called a *test set* of L. By the theorem, in order to check whether two morphisms agree on L, it suffices to check if they agree on the finite subset T.

The conjecture was reformulated for equations by Culik and Karhumäki in 1983, and this formulation opened the gates for the solution of the conjecture by Albert and Lawrence and independently by Guba in 1985. The proof techniques were originated already by Markov in the 1950s.

The result is nowadays known also as *Compactness theorem* (*for equations*) and *Noetherian property* (*of equations*).

References

[1] Bean, D. R., Ehrenfeucht, A., and McNulty, G., Avoidable patterns in strings of symbols, *Pasific J. Math.* **85** (1979), 261 – 294.

[2] Berstel, J., Axel Thues work on repetitions in words, in P. Leroux and C. Reutenauer, eds, Séries Formelles et Combinatoire Algébrique. Number 11 in Publications du LACIM, pp. 6580. Université du Québec á Montréal, 1992.

[3] Carpi, A. and de Luca, A., Words and special factors, *Theoret. Comput. Sci.* **259** (2001), 145 - 182.

[4] Casseigne, J., Unavoidable binary patterns, *Acta Inf.* **30** (1993), 385 – 395.

[5] Choffrut, C. and Karhumäki, J., Combinatorics of words, in Handbook of Formal Languages, Vol. 1, (A. Salomaa and G. Rozenberg, eds.), Springer-Verlag, 1997, pp. 329 – 438.

[6] de Luca, A., On the combinatorics of finite words, *Theoret. Comput. Sci.*, **218** (1999), 13 - 39.

[7] Halava, V., Harju, T. and Ilie, L., Periods and binary Words, *J. Combin. Theory Ser. A* **89** (2000), 298 – 303.

[8] Harju, T., On factorizations of words, *Bulletin EATCS* **24** (1984), 217.

[9] Harju, T. and Karhumäki, J., On the many faces of the defect theorem, submitted for publication. Technical Report 358, Turku Centre for Computer Science, TUCS, 1998.

[10] Harju, T. and Nowotka, D., On the density of critical factorizations, submitted, 2001.

[11] Lothaire, M., *Combinatorics on Words*, Addison-Wesley, 1983.

[12] Lothaire, M., *Algebraic Combinatorics on Words*, Cambridge University Press, 2002.

[13] Mignosi, F., Restivo, A. and Salemi, S., Periodicity and the golden ratio, *Theoret. Comput. Sci.* **204** (1998), 153 – 167.

[14] Skordev, D. and Sendov, Bl., On equations in words (in russian), *Z. Math. Logic Grundlagen Math.* **7** (1961), 289 – 297.

[15] Zimin, A., Blocking sets of terms, *Mat. Sb. (N.S.)* **119** (161) (1982); *Math. USSR Sb.* **47** (1984), 353 – 364.