Axel Thue's papers on repetitions in words: a translation

J. Berstel
L.I.T.P.
Institut Blaise Pascal
Université Pierre et Marie Curie
Paris, France

November 27, 1994

Contents

Introduction

In a series of four papers which appeared during the period 1906–1914, Axel Thue considered several combinatorial problems which arise in the study of sequences of symbols. Two of these papers [48, 50] deal with word problems for finitely presented semigroups (these papers contain the definition of what is now called a “Thue system”). He was able to solve the word problem in special cases. It was only in 1947 that the general case was shown to be unsolvable independently by E. L. Post [32] and A. A. Markov [28].

The other two papers [47, 49] deal with repetitions in finite and infinite words. Perhaps because these papers were published in a journal with restricted availability (this is guessed by G. A. Hedlund [22]), this work of Thue was widely ignored during a long time, and consequently some of his results have been rediscovered again and again. Axel Thue’s papers on sequences are now more easily accessible since they are included in the “Selected Papers” [51] which were edited in 1977.

It is the purpose of the present text to give a translation of Axel Thue’s papers on repetitions in sequences, both in more recent terminology and in relation with new results and directions of research.

It appears that there is a noticeable difference, both in style and in amount of results, between the 1906 paper (22 pages) and the 1912 paper (67 pages). The first of these papers mainly contains the construction of an infinite square-free word over three letters. Thue gives also an infinite square-free word over four letters obtained by what is now called an iterated morphism, whilst the three letter word is constructed in a slightly more complicated way (a uniform tag-system, in the terminology of Cobham [14]).

The second paper attacks the more general problem of what Thue calls *irreducible* words. He devotes special attention to the case of two and three letters. In particular, he introduces what is now called the *Thue-Morse sequence*, and shows that all twosided infinite overlap-free words are derived from this sequence. There are several aspects he did not consider: first, many combinatorial properties of the Thue-Morse sequence (such as the number of factors, the recurrence index, and so on) were only investigated by M. Morse [29] or later; next, the characterization of all onesided infinite overlap-free words — which is much more difficult than that of twosided words — was only given later by Fife [17]. However, Thue gives a complete description of circular overlap-free words.

Axel Thue's investigation of square-free words over three letters is even more detailed. He gives, in this paper, another construction of an infinite square-free word, by iterated morphism, and then initiates, in a 30 pages development, a tentative to describe all square-free words over three letters. He observes that every infinite square-free word is an infinite product of words chosen in a set of six words, and classifies those infinite square-free words that are products of four among these six words. His classification, he observes, is similar both in statement and in proof technique to what is found in diophantine equations: the solutions are parametrized by some variables which are easier to manage.

This text is organized as follows: in the first chapter, we give some preliminary definitions and notation. We introduce the notions of square-free, overlap-free words, avoidable pattern, morphisms and codes. These are useful to present Thue's results in a somewhat more concise manner. As an example, we give some combinatorial properties of the Thue-Morse sequence.

The two following chapters contain a translation of Thue's papers. We have tried to formulate Thue's results as faithfully as possible. For the proofs, some easy parts have been simplified, and more frequently some difficult steps have been developed. In these chapters, footnotes only concern technical details. A longer chapter of notes contains more general remarks and developments both about the contents of Thue's papers and about the actual state of the art.

Chapter 1

Preliminaries

In this preliminary chapter, we first introduce some definitions and notation and then present the so-called Thue-Morse sequence and some of its properties.

1.1 Notation

An *alphabet* is a finite set (of *symbols* or *letters*). A *word* over some alphabet A is a (finite) sequence of elements in A. The length of a word w is denoted by $|w|$. The *empty word* of length 0 is denoted by ε. We denote by $\mathrm{alph}(w)$ the set of letters that occur at least once in the word w. An *infinite* word is a mapping from $\mathbb{N}$ into A, and a *twosided* infinite word is a mapping from $\mathbb{Z}$ into A. A *circular* word or *necklace* is the equivalence class of a finite word under conjugacy (or circular permutation). We shall write $u \simeq w$ if u and w define the same circular word. Sometimes, we identify a circular word with one of its representatives.

A *factor* of a word w is any word u that occurs in w, i. e. such that there exist words x, y with $w = xuy$. A *square* is a nonempty word of the form uu. A word is *square-free* if none of its factors is a square. Similarly, an *overlap* is a word of the form $xuxux$, where x is nonempty. The terminology is justified by the fact that xux has two occurrences in $xuxux$, one as a *prefix* (initial factor) one as a *suffix* (final factor) and that these occurrences have a common part (the central x). As before, a word is *overlap-free* if none of its factors is an overlap. The *reversal* of a word $u = a_1 \cdots a_n$, where $a_1, \ldots, a_n$ are letters, is the word $\tilde{u} = a_n \cdots a_1$. If $u = \tilde{u}$, then u is a *palindrome*. The reversal of an infinite word to the right is an infinite word to the left.

The set of words over A is the free monoid generated by A and is denoted by A^*. The set of nonempty words over A is denoted by A^+. It is the free semigroup generated by A. A function $h : A^* \to B^*$ is a *morphism* if $h(uv) = h(u)h(v)$ for all words u, v. If $|h(w)| \geq |w|$ for all words w, then h is *nonerasing* or *length increasing*. It is equivalent to say that $h(w) \neq \varepsilon$ for $w \neq \varepsilon$. If there is a letter a such that $h(a)$ starts with the letter a, then $h^n(a)$ starts with the word $h^{n-1}(a)$ for all $n > 0$. If the set of words $\{h^n(a) \mid n \geq 0\}$ is infinite, the morphism defines a unique infinite word say $\mathbf{x}$ by the requirement that all $h^n(a)$ are prefixes of $\mathbf{x}$. The word $\mathbf{x}$ is said to be obtained by iterating h on a and is called a *morphic word*. Sometimes, $\mathbf{x}$ is also denoted by $h^\omega(a)$. Clearly, $\mathbf{x}$ is a fixed point of h. The Thue-Morse sequence of section 1.4 is an example of a morphic word. A morphism $h : A^* \to B^*$ easily extends to onesided infinite words. If $\mathbf{x} = a_0 a_1 \cdots a_n \cdots$ is an infinite word, then $h(\mathbf{x}) = h(a_0)h(a_1) \cdots h(a_n) \cdots$. The resulting word is infinite iff the set of indices n such that $h(a_n) \neq \varepsilon$ is infinite. This holds in particular if h is nonerasing. The extension to twosided infinite words is similar. The only ambiguity is in the convention adopted to fix the origin of the image. We agree that any origin is convenient. In other words, we consider, insofar as homomorphic images are concerned, the equivalence class under the *shift operator* T that is defined by $T(\mathbf{x})(n) = \mathbf{x}(n+1)$. If u is a finite word, then the infinite periodic word $u^\omega = uuu \cdots$ verifies $u^\omega = T^{|u|}(u^\omega)$.

1.2 Codes and encodings

A *code* over A is a set X of nonempty words such that each word over A admits at most one factorization as a product of words in X. In other words, for all $n, m \geq 1$, $x_1, \ldots, x_n, y_1, \ldots, y_m \in X$,

$$x_1 \cdots x_n = y_1 \cdots y_m, \quad \Rightarrow \quad n = m \text{ and } x_i = y_i \ (1 \leq i \leq n) \ .$$

It is equivalent to say that the submonoid X^* generated by X is free and that X is its base.

A set X is *prefix* if no word in X is a prefix of any other word in X; thus $x, xu \in X$ implies $u = \varepsilon$. *Suffix* sets are defined symmetrically. Prefix and suffix sets are codes. A *biprefix* code is a code that is both prefix and suffix.

An *encoding* is a morphism $h : A^* \to B^*$ that is injective. If h is an encoding, then the set $X = h(A)$ is a code. Conversely, if X is a code over an alphabet B, then an encoding of X is obtained by taking a bijection h from an alphabet A onto X. This extends to an injective morphism from A^* into B^*. It is convenient to implicitly transfer terminology between codes and encodings. Thus, we may speak about prefix encodings, or about composition of codes.

Several special properties of codes are useful, and will be introduced when they are needed.

1.3 The Thue-Morse sequence

In this section, we recall some basic properties concerning the Thue-Morse sequence. Other properties and proofs can be found in Lothaire [26] and Salomaa [38], and of course in Thue's second paper.

Let $A = \{a, b\}$ be a two letter alphabet. Consider the morphism μ from the free monoid A^* into itself defined by

$$\mu(a) = ab, \qquad \mu(b) = ba\ .$$

Setting, for $n \geq 0$,

$$u_n = \mu^n(a), \qquad v_n = \mu^n(b)$$

one gets

$$\begin{array}{ll} u_0 = a & v_0 = b \\ u_1 = ab & v_1 = ba \\ u_2 = abba & v_2 = baab \\ u_3 = abbabaab & v_3 = baababba \\ \quad\cdots & \end{array}$$

and more generally

$$u_{n+1} = u_n v_n, \quad v_{n+1} = v_n u_n$$

and

$$u_n = \overline{v}_n, \qquad v_n = \overline{u}_n$$

where $\overline{w}$ is obtained from w by exchanging a and b. Words u_n and v_n are frequently called *Morse blocks*. It is easily seen that u_{2n} and v_{2n} are palindromes, and that $u_{2n+1} = \tilde{v}_{2n+1}$, where $\tilde{w}$ is the reversal of w. The morphism μ can be extended to infinite words; it has two fixed points

$$\mathbf{t} = abbabaabbaababbabaab\cdots = \mu(\mathbf{t})$$

$$\overline{\mathbf{t}} = baababbaabbabaababba\cdots = \mu(\overline{\mathbf{t}})$$

and u_n (resp. v_n) is the prefix of length 2^n of $\mathbf{t}$ (resp. of $\overline{\mathbf{t}}$). It is equivalent to say that $\mathbf{t}$ is the *limit* of the sequence $(u_n)_{n\geq 0}$ (for the usual topology on finite and infinite words), obtained by iterating the morphism μ.

The *Thue-Morse sequence* is the word $\mathbf{t}$. There are several other characterizations of this word. Let t_n be the n-th symbol in $\mathbf{t}$, starting with $n = 0$. Then it is easily shown by induction that

$$t_n = \begin{cases} a & \text{if } d_1(n) \equiv 0 \pmod 2 \\ b & \text{if } d_1(n) \equiv 1 \pmod 2 \end{cases}$$

where $d_1(n)$ is the number of bits equal to 1 in the binary expansion $bin(n)$ of n. For instance, $bin(19) = 10011$, consequently $d_1(19) = 3$, and indeed $t_{19} = a$.

As a consequence, there is a finite automaton computing the values t_n as a function of $bin(n)$. This automaton has two states 0 and 1. It reads the string $bin(n)$ from left to right, starting in state 0. At the end, the state reached is 0 or 1 according to $t_n = b$ or $t_n = a$. In fact, the automaton computes $d_1(n)$ modulo 2. For a general discussion along these lines, see Cobham [14] and Allouche [2]. Another description is given by Christol, Kamae, Mendes France, Rauzy in [13]. There are many generalizations of the Thue-Morse sequence, motivated by its simplicity, and by its numerous properties. One quite general definition was in fact already given by Prouhet in 1851 ! (see [33, 1].)

As we shall see, the Thue-Morse sequence is overlap-free. What Thue actually showed, is that a word w over the two letter alphabet $A = \{a, b\}$ is overlap-free iff $\mu(w)$ is overlap-free.

1.4 Symbolic dynamical systems

Although the notion of (symbolic) dynamical system is not essential for understanding the papers of Thue, it gives some insight into what Thue perhaps had in mind when he tried to "parametrize" the square-free words.

A *symbolic dynamical system* or *subshift* is a set X of infinite words over some alphabet A that is closed for the shift operator, defined by $T(\mathbf{x})(n) = \mathbf{x}(n+1)$, and that is closed for the usual topology on infinite words. The *language* of X is the set $L(X)$ (or $\text{Fact}(X)$) of finite words that are factors of some element in X. It is not difficult to show that $\mathbf{x}$ is in X iff $L(\mathbf{x}) \subset L(X)$. A dynamical system X is *minimal* if it does not contain strictly any other dynamical system. This means that X is equal to the dynamical system generated by any of its elements, and also that $L(\mathbf{x}) = L(X)$ for any $\mathbf{x} \in X$. It has been shown that a dynamical system is minimal iff each of its elements is *uniformly recurrent* in the following sense. A word $\mathbf{x}$ is uniformly recurrent if there exists a function $\kappa : \mathbb{N} \to \mathbb{N}$ such that for all $u, w \in L(\mathbf{x})$, if $|w| \geq \kappa(|u|)$, then u is a factor of w. Other people say that factors appear with "bounded gaps". M. Morse [29] says simply *recurrent*. The property that the dynamical system generated by the (twosided) Thue-Morse sequence is minimal was explicitly proved by Gottschalk and Hedlund [18]. Axel Thue only mentions that every factor appears infinitely often.

Chapter 2

Thue's First Paper : About infinite sequences of symbols

Let u be a word over some alphabet A, and let w be a word over some alphabet B. We consider the question whether, given u and w, there always exists a nonerasing morphism $h : A^* \to B^*$ such that $h(u)$ is a factor of w. We shall prove that this does not hold, as a consequence of a theorem which answers the question for a large class of problems.

In the sequel, we call *irreducible*[1] a word without two adjacent equal factors.

§1

THEOREM 1.1. (Satz 1) *There exist arbitrarily long square-free words over four letters.*

In order to prove this result, we show that, given any square-free word of length k over four letters, one can always build a longer square-free word over the same alphabet.

Let p be any word over three letters — for instance a, b and c — of length at least 4, and such that p^2 contains no other square than itself. By inserting a new letter, say d, between two letters in p at four different places, we obtain four words x, y, z, t which all contain a single d, and which reduce to p when this letter is erased.

As an example, starting with

$$p = abacbc$$

[1] we shall write *square-free*.

we can set for instance

$$\begin{array}{ll} x = adbacbc & y = abdacbc \\ z = abadcbc & t = abacdbc \end{array}$$

and define a morphism

$$h : \{a, b, c, d\}^* \to \{a, b, c, d\}^*$$

by

$$h(a) = x, \quad h(b) = y, \quad h(c) = z, \quad h(d) = t \; .$$

We shall prove that h is a square-free morphism, i.e. that $h(u)$ is a square-free word whenever u is square-free. In order to do this, we need two lemmas.

LEMMA 1.2. *A word that contains an overlap also contains a square.*

Proof[2]. Let w be a word that has two overlapping occurrences of some nonempty word u. Then

$$w = xuy = x'uy'$$

for some words x, x', y, y'. We may assume that x is shorter than x', and since the occurrences overlap, one has $|x| < |x'| < |xu| < |x'u|$. Thus, setting $xs = x'$ and $x'q = xu$, one gets $xu = x'q = xsq$, whence $u = sq$, and

$$w = x'uy' = xssqy'$$

showing that w contains a square, namely ss. ∎

LEMMA 1.3. *Let p be a word such that p^2 contains no other square than itself. For all $n \geq 2$, if p^n contains a square u^2, then $|u| \equiv 0 \bmod |p|$.*[3]

Proof. Let u^2 be a factor of p^n. We first show that there exist prefixes x and x' of p, and words y, y' and an integer k such that

$$p^k = xuy = x'uy'$$

Indeed, assume $|x| < |x'|$. If xu is shorter than x', this means that the first occurrence of u is a factor of p. But then u^2 is a factor of p^2. Thus, the two occurrences of u in p^k overlap.

Thus, setting $xs = x'$, the (proof of the) preceding lemma shows that s^2 is a factor of p^2. Thus $s = p$. ∎

Observe that the preceding lemma also holds for any two distinct occurrences of u in a power of p, provided that $2|u| \geq |p|$.

---

[2] For the relationship between overlaps and squares, see the introductory chapter.
[3] and consequently, u is a conjugate of a power of p.

We now come back to the theorem. Let u be a square-free word. Set $w = h(u)$, where h is the morphism defined above, and assume, arguing by contradiction, that w contains a square, say v^2. Then

$$w = h(u) = \alpha v^2 \beta$$

for some words α, β. Let v' be obtained from v by erasing all occurrences of the letter d.

First, v contains at least one occurrence of the letter d. Indeed, otherwise $v = v'$, and v'^2 is a factor of w, and consequently v'^2 is a proper factor of p^2, contrary to the assumption on p. Next, by the preceding lemma, v' is the conjugate of some power of p, i. e.

$$v' = p_2 p^\ell p_1, \quad \ell \geq 0,\ p = p_1 p_2$$

thus v contains exactly $1 + \ell$ occurrences of the letter d. We set

$$v = s r_1 \cdots r_\ell \bar{s} = s' r'_1 \cdots r'_\ell \bar{s}'$$

where $r_1, \ldots, r_\ell, r'_1, \ldots, r'_\ell, \bar{s}s'$ are all in the set $X = \{x, y, z, t\}$. If $s \neq s'$, then it is easily seen that p^2 contains a proper square. Thus $s = s'$, $r_i = r'_i$ for $1 \leq i \leq \ell$, and $\bar{s} = \bar{s}'$. Since $\bar{s}s'$ contains one d, either s (and s') or $\bar{s}$ (and $\bar{s}'$) contains the letter d. But a suffix or a prefix of a word in X containing the letter d determines the word in X. This means that u contains a square. ■

We observe that the argument also holds for p of length 4. Thus, we may as well consider

$$p = abcb$$

and

$$x = adbcb \quad y = abdcb$$
$$z = abcdb \quad t = abcbd\ .$$

The previous theorem can be generalized to the following statement:

FACT. *Let X be a code of four nonempty words over a 4-letter alphabet satisfying*

(1) *if $x \in X$ and $uxv \in X^*$, then $u, v \in X^*$*[4]*;*
(2) *if $x, y, z \in X$, and $x \neq y \neq z$, then xyz is square-free;*
(3) *if $\alpha\beta, \alpha\gamma, \delta\beta \in X$, then $\alpha = \delta$ or $\beta = \gamma$.*[5]

and define a morphism h by assigning the four words in X to the four letters in the alphabet. Then $h(u)$ is square-free if u is square-free. (See **Notes 4.1.**)

The proof is by contradiction: let u be a word, and assume $h(u)$ contains a square ss. By (2), ss is not a factor of a product of three words in X. Consequently,

[4] This is the definition of a *comma-free* code; see the next chapter.
[5] As we shall see, this condition is superfluous.

ss contains a product xy, with $x, y \in X$. Thus one of the occurrences of s (and by (1) also the other one), contains an occurrence of a word of X. This implies, again by (1), that

$$ss = \beta x_1 \cdots x_n \alpha \beta x_1 \cdots x_n \alpha$$

with $\alpha\beta \in X$. It follows that u contains a factor $avbvc$, with

$$h(a) = p\beta,\ h(b) = \alpha\beta,\ h(c) = \alpha p',\ h(v) = x_1 \cdots x_n$$

for some p, p', whence

$$h(abc) = p\beta\alpha\beta\alpha p'$$

and by (2), $a = b$ or $b = c$. But then u contains a square.[6] ∎

THEOREM 1.4. (Satz 2) *There exists an infinite square-free word over four letters. More precisely, there exists a sequence* $(w_n)_{n \geq 0}$ *of square-free words such that* w_n *is a prefix of* w_{n+1}.

Indeed, it suffices to choose the morphism h such that $h(a)$, say, starts with the letter a. Then, there is an infinite word $\mathbf{x}$ that is a fixpoint of h, i.e. such that $\mathbf{x} = h(\mathbf{x})$. As an example, if we use the second set of words, we obtain the following infinite square-free word:

$$(adbcb)(abcbd)(abdcb)(abcdb)(abdcb)(adbcb)(abdcb)\cdots$$

In a very similar way, one may construct twosided infinite square-free words, or circular square-free words of arbitrary length.

§2

THEOREM 2.1. (Satz 3) *There exist arbitrarily long square-free words over three letters.*

We will prove the following more general result:

THEOREM 2.2. (Satz 4) *Over a three-letter alphabet* $\{a, b, c\}$, *there exist arbitrarily long square-free words without factors* aca *or* bcb.

[6] This should be compared with Satz 17 of the next paper.

These words can be obtained from the periodic word

$$ababababababababababababababab\cdots$$

by inserting the letter c at well chosen places between a's and b's.

Proof. The construction is in several steps[7]. Let u be a square-free word over a, b, c without factors aca or bcb.

(1) In the first step, we replace each occurrence of c preceded by a by the word $\beta\alpha$, and each occurrence of c preceded by b by $\alpha\beta$. In other words, a factor ac is replaced by $a\beta\alpha$ and bc is replaced by $b\alpha\beta$. Denote the resulting word by u'. For instance, if $u = acb$, then $u' = a\beta\alpha b$. Observe that we get u back from u' by erasing all α's and replacing each β by c.

We prove that u' is square-free and has no factor of the form $s\alpha s$ or $s\beta s$. Indeed, if u' contains a square ss, then, erasing all α's and replacing each β by c, one obtains a square contained in u. Thus, u' is square-free. Next, assume that u' contains a factor $s\beta s$. The central β is preceded or followed by an α. Thus, e.g. $s = \alpha t$, and $s\beta\sigma = \alpha t\beta\alpha t\beta$. Thus, erasing α's and replacing β's by c gives a factor of the form $xcxc$ of u. This proves the claim.

(2) In the second step, a letter γ is inserted after any letter of the word u'. Denote the resulting word by u''. For example, if $u' = a\beta\alpha b$ then $u'' = a\gamma\beta\gamma\alpha\gamma b\gamma$. Clearly, the word u'' has no factor of the form ss (since otherwise u' would contain a square).

(3) In the last step, we replace each a in u'' by $\alpha\beta\alpha$, and each b by $\beta\alpha\beta$. Denote the resulting word by w. Thus, for the word u'' of the example, we get $w = \alpha\beta\alpha\gamma\beta\gamma\alpha\gamma\beta\alpha\beta\gamma$.

We claim that the word w is square-free and has no factors of the form $\alpha\gamma\alpha$ and $\beta\gamma\beta$. To prove the second fact, observe that in u', letters a or α alternate with letters b or β. Thus, the factors of length 3 with a central γ in u'' are $a\gamma b$, $a\gamma\beta$, $\alpha\gamma b$, $\alpha\gamma\beta$ and their reversals. Consequently, the corresponding factors in w are $\alpha\gamma\beta$ and $\beta\gamma\alpha$.

Assume next that w contains a square ss. Since, between two consecutive γ's, the only factors are α, β, $\alpha\beta\alpha$ and $\beta\alpha\beta$, the word ss and consequently s contains at least one γ. If s contains only one γ and this letter is not, say, the last letter of s, then it is followed by α (or by β and the argument is the same). This means that ss contains the factor $\gamma\alpha\gamma\alpha$ or $\gamma\alpha\beta\alpha\gamma\alpha$, and thus w contains a factor $\alpha\gamma\alpha$, contradiction. Thus s contains at least two occurrences of the letter γ. Consequently, setting $X = \{\alpha, \beta, \alpha\beta\alpha, \beta\alpha\beta\}$, one gets

$$s = p\gamma x_1\gamma\cdots\gamma x_m\gamma q$$

[7] The next Satz contains a more compact construction.

for some integer $m \geq 1$, where $x_1, \ldots, x_m \in X$, $qp \in X$, and $p'p$, $qq' \in X$ for some p', q'.

If $q = \varepsilon$, then $p \in X$, and replacing in $p\gamma x_1 \gamma \cdots \gamma x_m \gamma$ each $\alpha\beta\alpha$ by a and each $\beta\alpha\beta$ by b, one gets a square contained in u''. The same conclusion holds if $p = \varepsilon$. Thus $p \neq \varepsilon$, $q \neq \varepsilon$ and $qp = \alpha\beta\alpha$ or $qp = \beta\alpha\beta$. It suffices to consider the first alternative. Then $(q,p) = (\alpha, \beta\alpha)$ or $(q,p) = (\alpha\beta, \alpha)$. These are symmetric. Consider the first case. The word w cannot start with $p\gamma = \beta\alpha\gamma$. Thus, there is at least a letter α preceding this factor, and consequently $qp\gamma x_1 \gamma \cdots \gamma x_m \gamma$ is a factor of w. But then u'' contains a square. This proves the claim.

The construction shows that, starting with a square-free word u over three letters a, b and c without factors aca and bcb, we get a longer square-free word w over the three letter α, β and γ without the factors $\alpha\gamma\alpha$ and $\beta\gamma\beta$. This concludes the proof. ■

THEOREM 2.3. (Satz 5) *There exists an infinite square-free word over three letters. More precisely, there exists a sequence $(w_n)_{n \geq 0}$ of square-free words over three letters such that w_n is a prefix of w_{n+1}.*

Proof. Let u be a square-free word over the letters a, b and c with no factor aca or bcb and starting with a or b. We obtain a new word by applying to u the function σ defined by

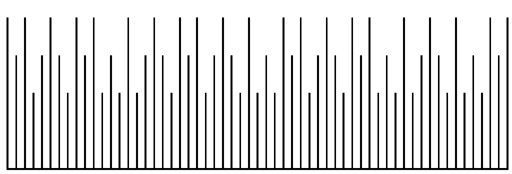
$$\sigma : \begin{array}{ll} a \mapsto abac & \\ b \mapsto babc & \\ c \mapsto bcac & \text{if } c \text{ is preceded by } a \\ c \mapsto acbc & \text{if } c \text{ is preceded by } b \end{array}$$

It is easily seen that the word $\sigma(u)$ is the same as the word w deduced from u in the preceding proof, when α, β, γ are replaced by a, b, c respectively. Thus $\sigma(u)$ is square-free and has no factor aca or bcb. Consequently, starting with $w_0 = a$, one gets a sequence $w_n = \sigma^n(w_0)$ of square-free words with the required property. ■

As an example, one gets the infinite square-free word

$$abac|babc|abac|bcac|babc|abac|babc|acbc|\ldots$$

If the letters a, b and c are replaced by vertical sticks of unequal length in this infinite word, one gets an infinite palisade without two equal consecutive parts:



We now give another construction of infinite square-free words over three letters a, b and c. For this, consider three fixed words

$$p = acab, \quad r = acb, \quad q = abcb$$

and the two sets of words

$$\begin{array}{ll} A_1 = p\alpha r\beta q & A = prq \\ B_1 = p\alpha' r\beta q & B = pcrq \\ C_1 = p\alpha r\beta' q & C = prcq \\ D_1 = p\alpha' r\beta' q & D = pcrcq \end{array}$$

Here $\alpha, \beta, \alpha', \beta'$ are new letters. The second column of words is obtained from the first by applying the morphism θ defined by:

$$\begin{aligned} \theta(\alpha) &= \theta(\beta) = \varepsilon; \\ \theta(\alpha') &= \theta(\beta') = c\ . \end{aligned}$$

Set $X_1 = \{A_1, B_1, C_1, D_1\}$ and $X = \{A, B, C, D\}$. It is easy to check that the product of two distinct words in X is square-free. Observe also that exchanging a and b converts A and D into their reversals.

The construction is in three steps, and starts with an infinite square-free word $\mathbf{s}$ over the letters α, α' and β' without factors $\alpha'\alpha\alpha'$ and $\beta'\alpha\beta'$. We have already seen that such a word exists. As an example, consider

$$\mathbf{s} = \alpha'\beta'\alpha'\alpha\beta'\alpha'\beta'\alpha\alpha'\cdots$$

(1) In the first step, we insert a letter β between any two consecutive occurrences of α and α' in the word s. Denote by $\mathbf{u}$ the resulting word. In our example,

$$\mathbf{u} = \alpha'\beta'\alpha'\beta\alpha\beta'\alpha'\beta'\alpha\beta\alpha'\cdots$$

If ρ is the projection that erases β, then $\rho(\mathbf{u}) = \mathbf{s}$. Clearly, $\mathbf{u}$ is square-free. Also, it has no factor of the form $w\beta w$, because $\mathbf{s}$ is square-free. We also show that $\mathbf{u}$ has no factor of the form $w\alpha w$. For this, observe that every α in $\mathbf{s}$ is preceded or followed by a letter α'. Indeed, otherwise, there would be a $\beta'\alpha\beta'$. Thus, every α in $\mathbf{u}$ is also preceded or followed by a β. This implies that, if we define a morphism τ by

$$\tau : \begin{array}{l} \alpha \mapsto \varepsilon \\ \beta \mapsto \alpha \\ \alpha' \mapsto \alpha' \\ \beta' \mapsto \beta' \end{array}$$

then $\tau(\mathbf{u}) = \mathbf{s}$. Thus, if $\mathbf{u}$ contains a factor $w\alpha w$, then $\mathbf{s}$ contains a square.

(2) In the second step, we replace every factor $\alpha\beta$, $\alpha'\beta$, $\alpha\beta'$, $\alpha'\beta'$ in $\mathbf{u}$ respectively by A_1, B_1, C_1, D_1, and denote the resulting word by $\mathbf{w}_1$. In our example, we get

$$\mathbf{w}_1 = D_1B_1C_1D_1A_1\cdots$$

Formally, if π denotes the projection of $\{a,b,c,\alpha,\alpha',\beta,\beta'\}^*$ onto the monoid $\{\alpha,\alpha',\beta,\beta'\}^*$, then $\pi(\mathbf{w}_1) = \mathbf{u}$. The word $\mathbf{w}_1$ is square-free, and contains no factor of the form $w\alpha w$ or $w\beta w$, since otherwise $\mathbf{u}$ would contain such a factor.

(3) Finally, let $\mathbf{w}$ be the word $\mathbf{w} = \theta(\mathbf{w}_1)$, where θ was defined above. We show that $\mathbf{w}$ is square-free. Assume the contrary. Then $\mathbf{w}$ contains a square, say uu. We have already seen that uu is not a factor of a product of two words in X. Consequently, uu contains as a factor at least one word in X. This implies that u itself contains one of the words p or q as a factor, and also, setting $t = qp$, that u contains r or t as a factor. Two consecutive occurrences of r and t in $\mathbf{w}$ are either adjacent or separated by the letter c. Thus, u can be factorized into

$$u = ws_1d_1s_2\cdots d_{m-1}s_mv$$

for some $m \geq 1$, where $s_1,\ldots s_m$ are in $\{r,t\}$, $d_1,\ldots,d_{m-1} \in \{\varepsilon,c\}$, and $vw \in \{\varepsilon,c\}$ or $vw = dsd'$, with $d,d' \in \{\varepsilon,c\}$ and $s \in \{r,t\}$. There are two adjacent factors U_1 and U_2 in $\mathbf{w}$ such that $\theta(U_1) = \theta(U_2) = u$. We may assume that U_1 does not start with α or β and U_2 does not end with α or β. This implies that

$$U_1 = w_1s_1\delta_1s_2\delta_2\cdots\delta_{m-1}s_mv_1$$
$$U_2 = w_2s_1\delta_1s_2\delta_2\cdots\delta_{m-1}s_mv_2$$

where δ_i is entirely determined by $s_id_is_{i+1}$. Now, v_1w_2 is neither α nor β, since otherwise $\mathbf{w}_1$ would have a factor of the form $v\alpha v$ or $v\beta v$. Also, v_1w_2 is neither α' nor β', since otherwise $U_1 = U_2$. Thus $\theta(v_1w_2) = dsd'$, with $s = r$ or $s = t$. However, this determines d and d', and implies that $U_1 = U_2$. The proof is complete. ∎

THEOREM 2.4. (Satz 6) *There exists an infinite cube-free word over two letters.*

As we shall see, we obtain such a cube-free word over a and b by replacing, in any infinite square-free word over the letters x, y and z, every x by a, every y by ab, and every z by abb[8]. In other terms, the cube-free infinite word is the image of a square-free infinite word under the morphism $f : \{x,y,z\}^* \to \{a,b\}^*$ defined by

$$f: \begin{array}{l} x \mapsto a \\ y \mapsto ab \\ z \mapsto abb \end{array}$$

Let $X = \{a, ab, abb\}$. This set is a suffix code[9].

[8] See also the 1912 paper.

[9] As we shall see, this observation basically suffices to prove the following elementary lemmas.

LEMMA 2.5. (Hülfssatz 1) *If u and v are words over the letters x and y such that $f(u) = f(v)$, then $u = v$.* ■

LEMMA 2.6. (Hülfssatz 2) *The morphism f is injective.*

Proof. This holds because X is a code. ■

Let $\mathbf{x}$ be an infinite square-free word over the letters x, y and z, and set $\mathbf{y} = f(\mathbf{x})$.

LEMMA 2.7. (Hülfssatz 3) *If $\mathbf{y}$ contains a factor uuu, then u does not start with the letter a.*

Proof. If u starts with the letter a, then there is a (unique) factor v of $\mathbf{x}$ such that $f(v) = u$. But then $\mathbf{y}$ contains the square vv. ■

LEMMA 2.8. (Hülfssatz 4) *If $\mathbf{y}$ contains a factor uuu, then u does not start with the word bb.*

Proof. If u does not begin with the word bb, then any occurrence of u is preceded by the letter a, and also u ends with an a. Thus, setting $u = u'a$, the word $\mathbf{y}$ has a factor $au'au'au'$, contrary to the preceding lemma. ■

LEMMA 2.9. (Hülfssatz 5) *If $\mathbf{y}$ contains a factor uuu, then u does not end with the letter b.*

Proof. In view of the preceding lemmas, u must start with ba. Thus, assuming the contrary and setting $u = bau'b$, one obtains in $\mathbf{y}$ the factor $bbau'bbau'bbau'b$. But then $\mathbf{y}$ contains the factor $au'bbau'bb$, showing that $\mathbf{x}$ contains a square. ■

We now can prove the theorem. Assume that $\mathbf{y}$ contains a cube uuu. Then u starts with ba and ends with a. If $u = ba$, then $\mathbf{x}$ contains the square yy. If $u = bau'a$ for some word u', then $uuu = bau'abau'abau'a$ and $\mathbf{y}$ contains the factor $abau'abau'$, showing that $\mathbf{x}$ contains a square. ■

It is easily verified that a word $f(\mathbf{x})$, where $\mathbf{x}$ is square-free, may have overlaps, but if $xuxux$ is an overlap, then x is a letter.[10]

[10] Compare with square-free words of type (I) in the 1912 paper.

Chapter 3

Thue's Second Paper : On the relative position of equal parts in certain sequences of symbols

For the development of logical sciences it will be important, without consideration for possible applications, to find large domains for speculation about difficult problems. In this paper, we present some investigations in the theory of sequences of symbols, a theory that has some connections with number theory.

3.1 Introductory Remarks

1.— A *word* over an alphabet

$$A = \{a_1, a_2, \ldots, a_n\}$$

of n letters (symbols) may have several meanings. For instance, a book can be viewed as a sequence of typographic symbols. The letters of the alphabet A can also be interpreted as mathematical entities or as substitutions for example. Let p be a positive integer. Then it is straightforward that any word $w \in A^*$ of length $m \geq n^p + p$ has two identical factors of length p. Observe that if w is viewed as a book, these unavoidable repetitions may not be meaningless. Without considering the meaning of words, it is of interest to investigate whether finite or infinite words can be constructed that have prescribed properties concerning the apparition of symbols. We expect that the results of such investigations have applications to usual mathematical problems. As an example, the existence of nonperiodic decimal developments proves that irrational numbers exist. The following is a general problem of this kind concerning the existence of identical factors in a word.

Let A and B be finite disjoint alphabets. A morphism $h : (A \cup B)^* \to A^*$ is called an *extension* if $h(a) = a$ for all $a \in A$. The problem is: given n words

$w_1, \ldots, w_n$ over $A \cup B$, does there exist an infinite word $\mathbf{x}$ over A such that, for any extension h, the word $\mathbf{x}$ has no factor in the set $\{h(w_1), \ldots, h(w_n)\}$? (See also **Notes 4.3**)

In the sequel, we will consider onesided infinite words, twosided infinite words, circular words, and ordinary finite words. Finite and onesided infinite words are called *open* words, twosided infinite and circular words are said to be *closed.*

2.— We are concerned with the construction of words with the property that any two occurences of the same factor are as far as possible one from each other. In any word w of length at least $n+2$ over an alphabet of size n, two equal factors cannot always be separated by a word of length greater than $n-2$. More precisely, if $|w| \geq n+2$, then w admits a factor of the form uvu, with $u \neq \varepsilon$ and

$$|v| \leq n-2.$$

Indeed, assume on the contrary that there is a word $w = a_1 \cdots a_n a_{n+1} a_{n+2}$ without a factor of this kind. Then the letters $a_1, \ldots, a_n$ are all distinct, and moreover $a_1 = a_{n+1}$ and $a_2 = a_{n+2}$. But then $w = a_1 a_2 v a_1 a_2$ with $|v| = n-2$.

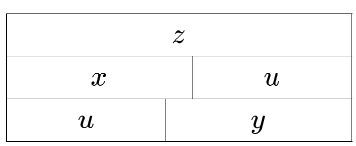
We shall see later how to construct, for $n > 1$, arbitrarily long closed words, and infinite words, such that any two equal factors are always separated by at least $n-3$ symbols.

A word over an n-letter alphabet is called *irreducible* if two occurences of a factor are always separated by at least $n-2$ letters. The word is called *reducible* otherwise[1]. Formally, w is irreducible if for any factor

$$z = xu = uy \qquad (x, y, u \neq \varepsilon)$$

one has

$$|z| - 2|u| = |x| - |u| \geq n-2.$$



One reason for this terminology is the following. Say that two words are equivalent if one word is obtained from the other by deleting or replacing factors of a given form by some fixed shorter words. Then, if factors of this prescribed class are unavoidable in sufficiently long words, this implies that there exist only finitely many classes for this equivalence relation.

[1] Examples : For $n = 3$, a word w is irreducible iff it is square-free ; for $n = 2$, it is irreducible iff it is overlap-free. Observe that the definition given in the previous paper applies in this context only for a three letter alphabet.

As an example, consider words that are composed of numbers which are alternatively positive and negative. Assume now that such a word u has two factors x and y which are the same up to the signs of the numbers, and which are separated by a factor z of odd length if x and y have even length, and with z of even length if x and y have odd length. Then u has the same algebraic value[2] after removing both x and y. Moreover, the resulting sequence is still formed of numbers with alternating signs.

Another example is the following. Consider a sequence u of parallel glass prisms arranged in such a way that a perdendicular light ray passes through all prisms. Let v be a similar sequence of prisms with the additional property that outcoming rays are always parallel to ingoing ones. If u contains v as a factor, this means that the deletion of v does not modify the angle of the lightrays.

Let us list some simple facts. We consider alphabets with n letters, assuming $n \geq 2$. Let u be a word of length $r = d + n - 3$, where $d \geq 2$ if $n = 2$ and $d \geq n - 2$ otherwise. In other terms, $r + 1 = d$ if $n = 2$, and $r \geq d$ if $n \geq 3$.

FACT. *Any factor of length $k \geq r + d$ in the infinite word u^ω is reducible.*

Indeed, such a factor has the form $w = u'v$, where u' is some conjugate of u, and v is a prefix of u'^ω. If $|v| > |u|$, then w is an overlap. Otherwise, $u' = vy$ for some word y, and $w = vyv$, with

$$|y| = |u| - |v| \leq r - d = n - 3 \qquad \blacksquare$$

FACT. *If all factors of length d of u^2 are irreducible, then any reducible factor of u^3 or of u^ω has length at least $r + d$.*

Proof. Let z be a reducible factor of u^3 of minimal length. If $|z| \leq d$, then $|z| \leq 1 + r$ and z is a factor of u^2, contrary to the assumption. Thus $|z| > d$. Assume, arguing by contradiction, that

$$d < |z| < r + d \ .$$

Since z is reducible, there are nonempty words x, y, t such that

$$z = xt = ty$$

and moreover

$$|z| - 2|t| = |x| - |t| \leq n - 3 \ .$$

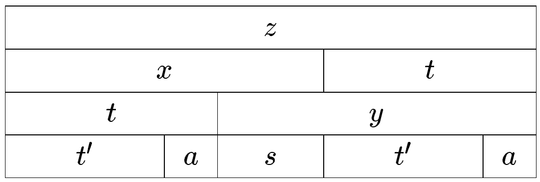
We show first[3] that

$$|x| - |t| = n - 3 \ .$$

[2] Thue means of course the sum of the numbers composing u.

[3] This is not done in the original paper.

Indeed, consider first the case where $n \geq 3$. If, contrary to the claim, $|x| < |t|$, then $xs = t$ for some nonempty word s. Thus $z = xxs$, showing that xx is a reducible factor which is shorter than z. Thus $|x| \geq |t|$, which proves the equality for $n = 3$. Assume now $n > 3$. Since $|x| \geq |t|$, we have $x = ts$ for some word s, whence $y = st$ and $z = tst$.



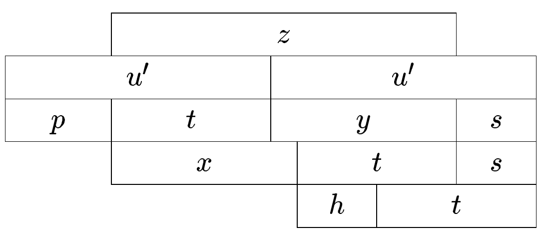
If $|s| \leq n - 4$, let a be the last letter of t, and let $t = t'a$, $s' = as$. Then $z' = t's't'$ is a shorter reducible factor than z, except for the case where $t' = \varepsilon$. Thus $|t| = 1$. This implies that $|z| = 2 + |s| \leq n - 2 \leq d$, again a contradiction. We thus have proved that $|s| = n - 3$.

Consider now the (easier) case $n = 2$. Then $|x| < |t|$, and consequently $xs = t = sy$ for some nonempty word s. Let a be the first letter of t (and of x and of s), and let $x = ax'$. Then $z = xxs$ starts with $ax'ax'a$ which is a reducible prefix. Thus this word is equal to z, showing that $|s| = 1$. This completes the proof.

We now come back to our initial claim. Since $|x| = |t| + n - 3$, and

$$|z| = 2|t| + n - 3 < r + d = 2d + n - 3$$

one has $2|t| < 2d$, whence $|t| < d$, $|x| < r$. Let p be the word of length $r - |t|$ such that $pt = u'$ is a conjugate of u, and let s be the word of length $r - |y| > 0$ such that $u' = ys$.



Then

$$u'u' = ptpt = pzs = ptys = pxts = pxht$$

where h is some word of same the length as s. Consequently $ts = ht$. This word clearly is reducible, and has length $r - n + 3 = d$, a contradiction. ■

A closed word of length r over an n-letter alphabet is called *irreducible* if $r > 2n - 6$ and if every open factor of length $r - n + 3$ is irreducible. Otherwise, the word is reducible.

For $n > 2$ and arbitrary r, a closed word of length r is a *closed irreducible* word if any two disjoint occurrences of the same factor are separated by at least $n - 2$

symbols. The word is *reducible* if it contains two disjoint occurences of the same factor separated by fewer than $n-2$ symbols[4].

3.— To each set S of words which all start with the same letter, say a, one associates a *tree* that represents the set in a simple way : the root is a vertex labelled with the common initial letter a of the words in S. Next, let $T = a^{-1}S = \{w \mid aw \in S\}$ and set

$$T = \bigcup_{b \in A} T_b, \quad T_b = T \cap bA^* \ .$$

For each $b \in A$ with $T_b \neq \emptyset$, the root of the tree of S is connected to the root of the tree associated with T_b. As an example, Fig. 1 shows an initial part of the tree of words over the n-letter alphabet $\{a, b, \ldots, h, k\}$ ($n \geq 8$) starting with $abcdef \ldots h$ and which are irreducible.

Figure No. 1

4.— Let A be an alphabet with n letters. We observe the following immediate facts. In an open irreducible word w over A, any $n-1$ consecutive letters are distinct.

Next, if w and wa, with a a letter, are irreducible words and $|w| \geq n-2$, then a is distinct from the $n-2$ rightmost letters in w.

Let w be an irreducible word. The word wa, with a a letter, is called a *right extension* of w if wa is irreducible. If $|w| \geq n-2$, then w has at most two right extensions.

Let w be an irreducible word of length $\geq n$, and assume that it has two right extensions wa and wb. Then setting $w = w'cdu$ with $|u| = n-2$, one has $\{c, d\} = \{a, b\}$.

Consider a tree containing all irreducible words starting with a given letter a, and let b be an arbitrary letter. If the path starting at the root and ending in

[4] There seems to be a third case, namely where the two occurrences are overlapping. But this also implies that the word is reducible.

a vertex labelled with b is composed of at least $n-2$ symbols, then the vertex labelled b has at most two sons.

FACT. *An irreducible word w has at most one right extension if and only if it has a suffix of the form upu, with $u \neq \varepsilon$ and $|p| = n-2$.*

Proof. If w has at most one right extension, then wa is reducible for all letters a with at most one exception. Take such a letter a which is distinct from the $n-2$ last letters of w. Then $wa = w'upuw''$ for some words w', w'', u, p, with $u \neq \varepsilon$ and $|p| \leq n-3$. Since w is irreducible, the word w'' is empty and thus the last letter of u is an a. Set $u = va$. Then $w = w'vapv$. If v is not empty, then since w is irreducible, one gets $|ap| \geq n-2$, which, combined with the first inequality, gives $|ap| = n-2$ and the announced suffix. Assume finally that $v = \varepsilon$. Then $w = w'ap$, and since a was chosen in an appropriate way, $|p| \geq n-2$.

Conversely, assume that $w = qupu$ for some word q. Then w has at most two right extensions wa and wb, and a, b are different from the last $n-2$ letters of pu. They are also different from the first letter of p. This shows the result if $|u| \geq n$, and also if u is a single letter. Thus the claim follows from the next fact. ■

FACT. *Any word of the form upu with $2 < |u| < n$ and $|p| = n-2$ is reducible.*

Assuming the contrary, let c be the first letter of p and let $A - \text{alph}(p) = \{a, b\}$. Then by considering the word pu, the first letter of u must be either a or b, and the second letter is either b or a ; it cannot be c since otherwise up would be reducible. Thus

$$upu = abu'pabu'$$

for u' defined by $u = abu'$, and $1 \leq |u'| \leq n-3$. The last letter u' is none of the letters in $\text{alph}(p)$, and is neither a nor b, a contradiction.

FACT. *If a word qvq is a proper suffix of an irreducible word pup and $|u| = |v| = n-2$, then qvq is a suffix of p.*

Indeed, q is a suffix of p, and consequently qvq is a suffix of qup. But the left occurences of q in these two words must be separated by at least $n-2 = |u|$ symbols. The claim follows.

Thus $p = tqvq$ for some word t. This implies that u and v start with different symbols. Indeed, if $u = au'$, $v = av'$, then

$$pup = tqav'qau'p$$

has the reducible factor $qav'qa$. Observe also that any word with suffix pup cannot be extended to the right into an irreducible word. By a previous remark, we know that the $n-2$ last symbols of q are all different ; they are also distinct from the first letter of u and from the first letter of v. Thus these letters altogether form the alphabet. Assume now that pup can be extended by a letter c. Then c can be neither one of the $n-2$ last letters of q, nor the first letter of u or of v. Thus c cannot exist.

As a consequence, an irreducible word cannot have three distinct suffixes pup, qvq, rwr, with $|u| = |v| = |w|$. Indeed, otherwise, and assuming $|r| < |q| < |p|$, the first occurence of p in pup has as suffixes both qvq, rwv, and is extensible to the right.

FACT. *If* $\mathbf{x}$ *is an infinite irreducible word, then for each integer* m*, there exists an irreducible word* w *of length* m *that admits at least two right extensions.*

Indeed, otherwise there is an integer m such that any extensible word has only one right extension. This would hold also for words longer than m, since each such word has a suffix of length m. However, this means that the infinite word $\mathbf{x}$, which then is completely characterized by its first m letters, is ultimately periodic, which is contrary to the assumption that it is irreducible.

5.— Again, we consider a fixed alphabet A with n letters; we first assume $n \geq 3$.

FACT. *Let* u *and* p *be words, with* $|p| = n-3$*, and such that* up *and* pu *are irreducible. If the (reducible) word* upu *has a (reducible) proper factor of the form* wqw*, with* $|q| = n-3$*, then* $|wqw| \leq |u| + |p|$ *(i.e.* $|w| \leq |u|/2$*).*

Proof. Since wqw is a proper factor of upu, there exist words x, z, with $|x|+|z| > 0$ such that

$$upu = xwqwz \ .$$

We may assume that xw is a prefix of u (otherwise wz is a suffix of u). Let t be such that $u = xwt$.

In order to prove the claim, assume now, arguing by contradiction, that $|wqw| > |up|$. Then $|xz| < |u|$. Therefore, there is a nonempty word y such that $u = xyz$. From this, it follows that

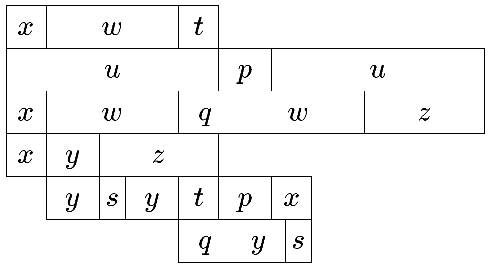
$$wqw = yzpxy \ .$$

Since $|q| = |p|$, one gets $|y| \leq |w|$, and equality cannot hold because otherwise $|xz| = 0$. Thus y is a proper prefix and a proper suffix of w. Since w is a factor of the irreducible word u, there is a factorization

$$w = ysy$$

with $|s| \geq n-2$. Let t be such that $wt = yz$. Recall that $u = xwt$. Thus,

$$qys = tpx \ .$$



If $|tp| \le |qy|$, then $|x| \ge |s|$, and consequently $x = x's$ for some x'. But then

$$pu = pxyz = px'swt = px'sysyt$$

is reducible. Consequently $|tp| > |qy|$, and $tp = qyy'$ for some y'. But then

$$up = xyzp = xwtp = xysyqyy'$$

has the reducible factor yqy, again a contradiction. ■

FACT. *Let w be an irreducible circular word, let p be a factor of length $n-3$ of w, and let u be the rest of w, i.e. such that $w = up$. Then upu has no proper irreducible factor. If furthermore $|u| \ge 2$, let $u = ha$, with a in A, and let $k = ap$. Then hkh is irreducible.*

Indeed, if there is an irreducible factor in upu, then we may assume, by a previous remark, that it has the form vqv for some v, and some q with $|q| = n-3$. But then $|vqv| \le |up| = |w|$, and vqv is a factor of w. This proves the first part. Next

$$hkha = hapha = upu$$

and therefore hkh is an irreducible factor of upu.

FACT. *Let w be a circular irreducible word, and suppose that w has a factor of the form vqv with $|q| = n-2$. Suppose further that $w \simeq vqvr$ with $|r| \ge n-2$. Then, there is a word $u \simeq w$ which has no right extension.*

Indeed, let $r = pas$ with $|p| = n-2$ and $a \in A$. Then $svqvqasvqv$ has no right extension.

6.— Two words $\mathbf{x}, \mathbf{y} \in A^{\mathbb{Z}}$ are called *congruent*[5] if there exists $k \in \mathbb{Z}$ such that $\mathbf{x}(i) = \mathbf{y}(i+k)$ for all $i \in \mathbb{Z}$. A word $\mathbf{x} \in A^{\mathbb{Z}}$ is *simply recurrent* if every factor of $\mathbf{x}$ has infinitely many occurences in $\mathbf{x}$. A word has only *h-bounded overlaps* if for every factor of the form $xuxux$ with $u \ne \varepsilon$, one has $|x| \le h$. We say that the word has bounded overlaps if it has h-bounded overlaps for some h. Finally, we say that $\mathbf{x}$ *avoids* a finite set $X \subset (A \cup B)^*$, where $A \cap B = \emptyset$ if there is no extension $h : (a \cup B)^* \to A^*$ such that all words $h(x)$, $(x \in X)$ are factors of $\mathbf{x}$.

[5]Morse, Hedlund call them *similar*.

THEOREM 1.1. (Satz 1) *Let* $\mathbf{x} \in A^{\mathbb{Z}}$ *be an infinite word that satisfies the following conditions:*

(i) $\mathbf{x}$ *is simply recurrent;*
(ii) $\mathbf{x}$ *has bounded overlap;*
(iii) $\mathbf{x}$ *avoids a fixed set* $X \subset (A \cup B)^*$.

Then there exist infinitely many twosided infinite words with the same three properties.

Proof. We construct a sequence $(u_k)_{k \geq 0}$ of factors of $\mathbf{x}$ as follows :

(i) u_0 is an arbitrary nonempty factor of $\mathbf{x}$.

(ii) assume u_k is constructed. Then to a given occurence of u_k, there is another occurence of u_k, to the right or to the left. Suppose it is to the right. Thus there is a word v_k such that

$$u_k v_k u_k$$

is a factor of $\mathbf{x}$. Consider one occurence of this word, and consider any factor α_k that extends the occurence to the left, and a factor w_{k+1} that extends the occurence to the right and that, furthermore, has the property that w_{k+1} is not a prefix of $u_k w_{k+1}$[6]. We thus have obtained a factor

$$u_{k+1} = \alpha_k u_k v_k u_k w_{k+1} = w_{-(k+1)} u_k w_{k+1}$$

with $w_{-(k+1)} = \alpha_k u_k v_k$. A symmetric definition holds in the symmetric case. It is not very difficult to check that the infinite word

$$\mathbf{y} = \ldots w_{-3} w_{-2} w_{-1} u_0 w_1 w_2 w_3 \ldots$$

is not congruent to $\mathbf{x}$ but has the same factors as $\mathbf{x}$ and therefore has the properties claimed. ■

The construction can be used for deriving similar results on infinite words. We consider the following problem. Let $h : A^* \to A^*$ be a fixed nonerasing morphism. Does there exist a twosided infinite word $\mathbf{x}_0 \in A^{\mathbb{Z}}$ such that there is a sequence

$$\mathbf{x}_1, \ldots, \mathbf{x}_m, \ldots$$

of twosided infinite words over A with

$$h(\mathbf{x}_{m+1}) = \mathbf{x}_m \ .$$

If this holds, and if furthermore $\mathrm{alph}(h(a)) = A$ for $a \in A$ then every factor of $\mathbf{x}_0$ appears infinitely often in $\mathbf{x}_0$.

Let u_0 be a nonempty word, and assume that

$$h(u_0) = v_0 u_0 w_0$$

[6] This is possible because $\mathbf{x}$ has bounded overlaps.

for nonempty words v_0, w_0. Then setting, for $m \geq 0$,

$$u_{m+1} = h(u_m),\ v_{m+1} = h(v_m),\ w_{m+1} = h(w_m),$$

we get

$$u_{m+1} = v_m u_m w_m = v_m v_{m-1} \cdots v_0 u_0 w_0 \cdots w_m\ .$$

Therefore, the twosided infinite word $\mathbf{x}$ defined by

$$\mathbf{x} = \cdots v_2 v_1 v_0 u_0 w_0 w_1 w_2 \cdots$$

is a fixed point for h, i.e. $h(\mathbf{x}) = \mathbf{x}$.

After these introductory remarks, we will consider in more detail irreducible words for special values of n, the number of letters. As we shall see, closed or twosided infinite irreducible words have some analogy with Diophantine equations.

3.2 Sequences over two symbols

7.— We now consider a fixed alphabet $A = \{a, b\}$. A finite or infinite word w over A is irreducible if it has no overlap; in the sequel, we[7] call it *overlap-free.* A circular word w is overlap-free iff the open word ww is overlap-free. For any finite or infinite word w, we denote by $\bar{w}$ the word obtained by exchanging the a's and b's in w.

EXAMPLE. The circular words aa and $abab$ have overlaps. The circular word aab is overlap-free.

It is not difficult to verify that a circular word of length r is overlap-free iff all factors of length $1 + r$ of the open word ww are overlap-free.[8]

Figure 2 shows all overlap-free words of length at most 12 starting with the letter a. The final letters of words which cannot be extended are marked with a circle.

Through a sequence of statements we will in particular prove the existence of infinite overlap-free words. We begin with some lemmas.

LEMMA 2.1. (Satz 2) *Let $X = \{ab, ba\}$. For any $x \in X^*$, one has $axa \notin X^*$ and $bxb \notin X^*$.*

[7] the translator

[8] For, assume that $ww = ycxcxcz$ with $c \in A$, x of minimal length, and $|cxcxc| > 1 + |w|$. Then either $ycxc$ is a prefix of w or, symmetrically, $cxcz$ is a suffix of w. In the first case, the word $cxcz$ has another occurrence of cxc, and the length condition implies that these occurrences overlap.

Proof. By induction on $|x|$, the case $|x| = 0$ being trivial. Let $x \in X^*$, $x \neq \varepsilon$, and assume that $u = axa$ is in X^* (the case $bxb \in X^*$ is similar). Then the first and the last letters of x must be b. Thus $x = byb$ for some word, and consequently

$$u = abyba.$$

Since $u \in X^*$, one has $y \in X^*$, and by induction $u = byb$ is not in X^*, contrary to the assumption. ∎

Figure No. 2

We consider the two morphisms

$$\mu : \begin{array}{l} a \mapsto ab \\ b \mapsto ba \end{array} \qquad \bar{\mu} : \begin{array}{l} a \mapsto ba \\ b \mapsto ab \end{array}$$

LEMMA 2.2. (Satz 3) *If w is an overlap-free word, then $\mu(w)$ and $\bar{\mu}(w)$ are overlap-free.*

Proof. Assume that $\mu(w)$ has an overlap. Then

$$\mu(w) = xcvcvcy$$

for some words x, v, y and a letter c. Since $|\mu(w)|$ is even and $|cvcvc|$ is odd, it follows that $|xy|$ is odd and therefore one of $|x|$ or $|y|$ is even and the other is odd. By symmetry, we may assume that $|x|$ is odd and $|y|$ is even.

Set $X = \{ab, ba\}$. Then $y \in X^*$, and furthermore $|v|$ is odd. Indeed, since $vcvc \in X^*$, the contrary would imply that both v and cvc are in X^*, in contradiction to the previous lemma. It follows that vc is in X^*, and xc is in X^*. Thus $w = rsst$ with $\mu(r) = xc$, $\mu(s) = vc$, $\mu(t) = y$. But r and s have the same final letter, showing that w has an overlap. ∎

A similar proof gives the following lemma.

LEMMA 2.3. (Satz 4) *If w is an overlap-free circular word, then $\mu(w)$ and $\bar{\mu}(w)$ are overlap-free.* ∎

By induction, $\mu^p(w)$ is overlap-free for any overlap-free word w and for any positive integer p. Set for $n \geq 0$

$$u_n = \mu^n(a),\ v_n = \mu^n(b).$$

THEOREM 2.4. (Satz 5) *There exists an overlap-free infinite word over two letters.*

Proof. Let

$$\mathbf{t} = av_0v_1v_2...v_n...$$

By induction on n, $u_{n+1} = av_0v_1...v_n$ for $n \geq 0$. Thus

$$\mu(\mathbf{t}) = \mathbf{t}$$

and $\mathbf{t}$ is overlap-free. ∎

COROLLARY 2.5. (Satz 6) *Let $\mathbf{x}$ and $\mathbf{y}$ be infinite words with $\mathbf{x} = \mu(\mathbf{y})$. Then $\mathbf{x}$ is overlap-free iff $\mathbf{y}$ is overlap-free.*

Proof. It is easily seen that if $\mathbf{x}$ is overlap-free, then $\mathbf{y}$ is overlap-free. The converse follows from Satz 3. ∎

Observe that u_{2n} and v_{2n} are palindromes and that $\tilde{u}_{2n+1} = v_{2n+1}$ for $n >= 0$. Indeed, by induction,

$$u_{2n+2} = \mu^2(u_{2n}) = u_{2n}v_{2n}v_{2n}u_{2n} = \tilde{u}_{2n}\tilde{v}_{2n}\tilde{v}_{2n}\tilde{u}_{2n} = \tilde{u}_{2n+2}.$$

The other verifications are similar.

THEOREM 2.6. (Satz 7) *Let $w_n = \tilde{v}_n$ for $n \geq 0$. The twosided infinite word*

$$\mathbf{u} = \cdots w_n \cdots w_2w_1w_0aav_0v_1 \cdots v_n \cdots$$

is overlap-free.

Proof. Of course, $\mathbf{u} = \tilde{\mathbf{t}}\,\mathbf{t}$. From the relations above, it follows that

$$w_n \cdots w_1 w_0 a a v_0 \cdots v_n = \begin{cases} v_{n+2} & n \text{ even} \\ u_{n+1} u_{n+1} & n \text{ odd.} \end{cases}$$

This holds indeed for $n = 0, 1$; next, if n is even, then $w_n = v_n$ and

$$w_n \cdots w_1 w_0 a a v_0 \cdots v_n = v_n u_n u_n v_n = v_{n+2} \ .$$

If n is odd, then $w_n = u_n$ and

$$w_n \cdots u_n = u_n v_{n+1} v_n = u_n v_n v_n u_n = u_{n+1}^2 \ .$$

The result follows. ∎

Observe that

$$\mu(\tilde{\mathbf{t}})\mathbf{t} = \tilde{\tilde{\mathbf{t}}}\mathbf{t}$$

is also an overlap-free twosided infinite word.

8. — Let w be an overlap-free word over A. If $|w| \geq 5$, then w has at least one factor in the set $Y = \{aa, bb\}$. Consequently , if $|w| \geq 9$, then w has at least two occurrences of factors in Y.

If w is an overlap-free word circular with at least 4 letters, then w has at least two occurrences of factors in Y.

PROPOSITION 2.7. (Satz 8) *Let w be a word over A of the form*

$$w = cddxeef$$

where c, d, e, f are letters and x is a word. If w is overlap-free, then w and dxe are in X^*, where $X = \{ab, ba\}$.[9]

Proof. By induction on the length of x. Without loss of generality, we may assume that $c = a$, whence $d = b$. If $x = \varepsilon$, then $c \neq d \neq e \neq f$ and $w = abbaab$ which is in X^*.

Assume that $x \neq \varepsilon$. Then $x = ay$ for some $y \neq \varepsilon$, and

$$w = abbayeef.$$

If y starts with the letter a the result holds by induction. Thus assume that $y = bz$ for some z. If $z = \varepsilon$, then $w = abbabeef$, whence $e = a$ and $f = b$ and w is in X^*. If $z \neq \varepsilon$, and z starts with b, the result again follows by induction. Finally, we assume that $z = at$, and thus

$$w = abbabateef \ .$$

Observe that $t \neq \varepsilon$ since otherwise w contains an overlap. Thus t starts with a b. The result follows by induction. ∎

[9] This means that (a, a) and (b, b) are synchronizing pairs.

PROPOSITION 2.8. (Satz 9) *If* $\mathbf{w}$ *is a twosided infinite overlap-free word, then* $\mathbf{w} = \mu(\mathbf{u})$ *for some infinite overlap-free word* $\mathbf{u}$.

Proof. Let $\mathbf{w}$ be a twosided infinite overlap-free word. As observed above, any long enough factor has two distinct occurrences of a factor aa or bb. The result follows then from the previous proposition. ∎

PROPOSITION 2.9. (Satz 9) *For any overlap-free circular word* w *of length at least 4, there exists a unique circular word* u *such that* $w = \mu(u)$. ∎

PROPOSITION 2.10. (Satz 10) *If* $\mathbf{w}$ *is a twosided infinite overlap-free word, then for any integer* $k \geq 1$, *there is a unique infinite overlap-free word* $\mathbf{u}$ *such that* $\mathbf{w} = \mu^k(\mathbf{u})$. ∎

In taking k sufficiently large in the previous proposition, one gets:

COROLLARY 2.11. (Satz 11) *Let* $\mathbf{w}$ *be a twosided infinite overlap-free word. Every factor of* $\mathbf{w}$ *appears infinitely often in* $\mathbf{w}$. ∎

Observe that, according to Satz 1, this shows that there exist infinitely many congruence classes of overlap-free words. (See **Notes 4.2**)

Another consequence is the following:

THEOREM 2.12. (Satz 12) *Let* $\mathbf{x}$, $\mathbf{y}$ *and* $\mathbf{z}$ *be (onesided) infinite words over* A, *and consider the twosided infinite words*

$$\mathbf{u} = \tilde{\mathbf{x}}\mathbf{y}, \ \mathbf{v} = \tilde{\mathbf{x}}\mathbf{z} \ .$$

Assume that $\mathbf{y}$ *and* $\mathbf{z}$ *start with different letters. If* $\mathbf{u}$ *and* $\mathbf{v}$ *are both overlap-free, then* $\mathbf{y} = \bar{\mathbf{z}}$ *and furthermore either* $\mathbf{x} = \mu(\mathbf{x})$ *or* $\mathbf{x} = \bar{\mu}(\mathbf{x})$ *and* $\mathbf{z} = \mu(\mathbf{z})$ *or* $\mathbf{z} = \bar{\mu}(\mathbf{z})$ *and thus* $\mathbf{x}$, $\mathbf{y}$ *and* $\mathbf{z}$ *are equal to* $\mathbf{t}$ *or* $\bar{\mathbf{t}}$.

Proof. It suffices to observe that for any $k \geq 0$, the infinite words $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ have $\mu^k(a)$ or $\mu^k(b)$ as prefixes. ∎

PROPOSITION 2.13. (Satz 13) *Every circular overlap-free word with length at least 2 is of the form* $\mu^n(aab)$, $\mu^n(bba)$, $\mu^n(ab)$ *for some integer* $n \geq 0$.

Proof. If w has length at least 4, then $w = \mu(u)$ for some overlap-free circular word u, and of course $|w| = 2|u|$. Thus, it suffices to consider the overlap-free circular words of length 2 or 3. ∎

COROLLARY 2.14. (Satz 14) *Any circular overlap-free word has length* 2^n *or* $3\cdot 2^n$ *for some* $n \geq 0$. ■

Let w be an overlap-free word of length at least 10. Then there exist letters x, y, z, u and words p, s, t such that

$$w = pyxxtzzus$$

with $x \neq y$, $z \neq u$, and

$$p \in \{\varepsilon, x, y, yx, xx, yyx\}, \quad s \in \{\varepsilon, z, u, zu, zz, zuu\}$$

and

$$xtz \in \{ab, ba\}^*.$$

This observation is useful in the proof of the following theorem:

THEOREM 2.15. (Satz 15) *Let* $n \geq 1$ *and let* w *be an overlap-free word of length* n. *If there exist words* u, v *of length at least* $8n$ *such that* uwv *is overlap-free, then any overlap-free word of length at least* $26n$ *contains* w *as a factor.*

Proof. Let uwv be overlap-free, with $|w| = n$ and $|u|, |v| \geq 8n$. Let $k = 1 + \lfloor \log_2 n \rfloor$. We construct a decreasing sequence of words

$$s_h = u_h w_h v_h \qquad (0 \leq h \leq k)$$

with $u_0 = u$, $w_0 = w$, $v_0 = v$, such that $\mu(s_{h+1})$ is a factor of s_h and w_h is a factor of $\mu(w_{h+1})$:

$\mu(u_{h+1})$	$\mu(w_{h+1})$	$\mu(v_{h+1})$
u_h	w_h	v_h

Assume that $|w_h| > 10$. Then, according to the preceding observation, there is a factor s' of w_h of length at least $|w_h| - 6$ in $\{ab, ba\}^*$. Define $s_{h+1} = u_{h+1}w_{h+1}v_{h+1}$ in such a way that $s' = \mu(s_{h+1})$ and furthermore w_h is a factor of $\mu(w_{h+1})$. Clearly

$$2|u_{h+1}| \geq |u_h| - 3, \quad 2|v_{h+1}| \geq |v_h| - 3$$

$$|w_h| \leq 2|w_{h+1}| \leq |w_h| + 2$$

whence by induction

$$|u_{h+1}| > \frac{|u|}{2^{h+1}} - 3, \quad \frac{|w|}{2^{h+1}} \leq |w_{h+1}| < \frac{|w|}{2^{h+1}} + 2 \ .$$

Since $|u| \geq 8n \geq 8 \cdot 2^{k-1}$, it follows that

$$|u_{k-1}| > \frac{|u|}{2^{k-1}} - 3 > 5 \ .$$

Thus s_{k-1} has length greater than 10, and consequently the word s_k exists. It follows that w is a factor of $\mu^k(a)$ or of $\mu^k(b)$.

Consider now a word f of length $26n$. By the observation above, if f is overlap-free, then there are words p, s, and a word g such that

$$f = p\mu(g)s$$

and $|p|, |q| \leq 2$. Thus there is a sequence of words $f_0, f_1, \ldots$ such that f_h is a factor of $\mu(f_{h+1})$ and

$$2|f_{h+1}| \geq |f_h| - 4 \ .$$

This implies that

$$|f_{h+1}| > \frac{|f|}{2^{h+1}} - 4 \ .$$

Since $|f| \geq 26n \geq 13 \cdot 2^k$, one has

$$|f_k| > \frac{|f|}{2^k} - 4 \geq 9 \ .$$

Thus f_k contains also s_k. This proves the result. ∎

Observe that since the word w of the preceding theorem is a factor of some $\mu^k(a)$ or $\mu^k(b)$, this means that w is extensible to a twosided infinite word.

A morphism h is called *overlap-free* if $h(w)$ is overlap-free for all overlap-free words w. The next result gives a characterisation of overlap-free morphisms. (See **Notes 4.2**)

THEOREM 2.16. (Satz 16) *For any overlap-free morphism h over two letters, there is an integer $k \geq 0$ such that $h(a) = \mu^k(a)$, $h(b) = \mu^k(b)$ or $h(a) = \mu^k(b)$, $h(b) = \mu^k(a)$.*

Proof. Set $h(a) = u$, $h(b) = v$. The result holds if $|u| = |v| = 1$.

We prove first that if $|u| > 1$, then $|u|$ is even. Indeed, assume first that $|u| = 3$. If $u = aab$ or $u = baa$, that $vuuv$ has a factor $b(aab)(aab)$ or $(baa)(baa)b$. Similarly, if $u = aba$, then $vvuuvv$ or vuv have an overlap, according to the first and the last letter of v. Thus $|u| > 3$. If $|u| > 4$, then u has the form

$$u = paas \text{ or } u = pbbs$$

for some nonempty words p, s. Assume the former. The word

$$vuuv = vpaaspaas$$

fulfills the requirements of Satz 8. Thus the central factor $aspa$ has even length, showing that u has even length. This proves the claim.

Next we show that $|u| = 1$ implies $|v| = 1$. Indeed, if say $u = a$, and $|v| > 1$, then v has even length. Moreover, since $vuuv$ is overlap-free, $v = bwb$ for some word w, and $w \neq \varepsilon$ because vv must be overlap-free. But then, in

$$vvuuv = bwbbwbaabwb$$

the central factor $bwba$ has even length, again by Satz 8. Thus $|v|$ is odd, a contradiction. This shows the second claim.

We now prove the result by induction on $|u| + |v|$, assuming $|u| > 1$, $|v| > 1$. We already know that u and v have even length. Without loss of generality, we may assume that u starts with the letter a.

If $u = awa$, then w is not empty. Thus $w = bzb$ for some word z, since otherwise uu contains an overlap. Moreover, w contains a factor aa or bb. Indeed, otherwise $w = (ba)^n b$ for some n, which is impossible because $|w|$ is even. Thus w has the form $w = xddy$ for some letter d and some words x, y, and

$$uu = axddyaaxddy$$

showing that dya and axd are in X^*, with $X = \{ab, ba\}$. Thus, u also is in X^*.

If $u = awb$, then $v = bza$ for some z. The word

$$vuvu = awbbzaawbbza$$

is overlap-free, and as above, this shows that u is in X^*. Similarly, v is in X^*.

It follows that $u = \mu(u')$, $v = \mu(v')$ for some words u', v', and that the morphism h' defined by $h'(a) = u'$, $h'(b) = v'$ also is overlap-free. Since $h = \mu \circ h'$, the result follows. ■

10.— We now give some results about the tree of overlap-free words over two letters a and b. We set $X = \{ab, ba\}$.

LEMMA 2.17. *Let ux, uy be two overlap-free words, with $|x|$, $|y| \geq 2$, and assume that x and y start with different letters. If u is of the form $u = abbu'$ for some word u', then $u \in X^*$. If furthermore,*

$$x = x'eef$$
$$y = y'ggh$$

where e, f, g, h are letters and x', y' are words, then $x, y \in X^*$.

Proof. Since ux and uy are overlap-free and x and y start with different letters, the word u' is not empty. Set $u' = vc$ where c is a letter. Then either ux or uy is of the form

$$u = abbvccdw$$

for some letter d and some word w. By Satz 8, the word bvc is in X^*. Thus $u \in X^*$.

Since $ux = abbu'x'eef$, the same proposition shows that $bu'x'e$ is in X^*. It follows that bu' is in X^* and finally $x \in X^*$.

LEMMA 2.18. *Let u be a prefix of $\mathbf{t}$ of length $m = |u| \geq 3$, and set $\mathbf{t} = u\mathbf{x}$. If uy is a finite overlap-free word with $|y| \geq m-2$, and if $\mathbf{x}$ and y start with different letters, then m is a power of 2.*

Proof. If $m = abb$ then y starts with b and uy contains a cube. Thus, $m \geq 4$. By the lemma above, u is in X^*. Furthermore, and still by the lemma, there is a prefix z of y which differs from y by at most 2 letters and which is in X^*. Consider now the words

$$u' = \mu^{-1}(u), \quad \mathbf{x}' = \mu^{-1}(\mathbf{x}), \quad z' = \mu^{-1}(z).$$

Again $u'\mathbf{x}' = \mathbf{t}$, $u'z'$ is overlap-free, and $\mathbf{x}'$ and z' start with different letters. Since $|z'| \geq (m-2)/2$, the lemma follows by induction. ■

LEMMA 2.19. *Let u be a word of length at least 4 such that $auuc$ is overlap-free, with c a letter. Then $u \in X^*$.*

Proof. We first observe that u cannot end with an a, and that the first letter of u is not c. We shall see that in fact u starts with baa or abb or with $babaa$ or with $ababb$.

We first show that bb is not a prefix of u. Indeed, otherwise $u = bbu'b$ for some word u' and uu contains a cube. Clearly, aa is not a prefix of u.

Next, we show that $babb$ is not a prefix of u. Indeed, otherwise $u = babbu'$ for some u', and since uu is overlap-free, u' is not empty. More precisely, u' starts with a and ends with ab, thus $u' = avab$ or $u' = ab$. In the first case, $uu = babbavabbabbavab$ contains the factor $avabbabbava$ which contains an overlap. In the second case, $uu = babbabbabbab$ contains an overlap.

Next, if $u = abaau'$, then u' is not empty and $u' = vb$ for some v. Thus $uuc = abaavbabaavbb$. Clearly, v is not empty, and ends neither with a nor b.

It follows from this that if the first letter of u is b, then u starts either with baa or with $babaa$. Similarly if u starts with a, it starts with abb or $ababb$. In the first case,

$$uu = baavbaav = ba(avba)av$$

showing (even if $v = \varepsilon$) that $avba \in X^*$. In the second case,

$$uu = babaavbabaav = baba(avbaba)av$$

showing that $avbaba \in X^*$. ∎

Finally, let

$$\mathbf{x} = a_0 a_1 \cdots a_n \cdots$$

be an infinite overlap-free word. Then not every suffix $a_n a_{n+1} \cdots$ starts with a square. In other words, there exists an integer p such that, setting $\mathbf{y} = a_p a_{p+1} \cdots$, both $a\mathbf{y}$ and $b\mathbf{y}$ are overlap-free. Indeed, $\mathbf{x}$ has infinitely many occurrences of the word $ababbaab$, and contains no square that starts with $babbaab$.

3.3 Sequences over three symbols

11.— A word w over a three-letter alphabet is irreducible if it is square-free. Clearly, if w contains an overlap, it also contains a square. A circular word w of length r is square-free iff it contains no square of length less than r.

Figure No. 3

Figure 3 shows all square-free words of length at most 12. Again, a small circle around a letter means that the corresponding branch in the tree cannot be extended. A morphism h is called square-free if $h(w)$ is a square-free word for every square-free word w.

It is convenient[10] to call a morphism h over some alphabet A a *factor-free* morphism if, whenever $h(a)$ is a factor of $h(b)$ for some letters a and b, then $a = b$. This implies of course that h is injective, and in fact that $h(A)$ is a biprefix code. The set $X = h(A)$ itself will be called factor-free. Next, a set $X = h(A)$, and by extension the morphism h, is *comma-free* if, whenever $x \in X$ and $uxv \in X^*$ for some words u, v, then $u, v \in X^*$. Clearly, a comma-free morphism is factor-free (the converse is false, consider $\{a, bab\}$).[11]

THEOREM 3.1. (Satz 17) *Let A be a three-letter alphabet, and let $h : A^* \to A^*$ be a nonerasing factor-free morphism. If $h(w)$ is square-free for all square-free words of length 3, then h is a square-free morphism.*

For the proof, we first give a lemma of independent interest:

LEMMA 3.2. *Let A be a three-letter alphabet, and let $h : A^* \to A^*$ be a nonerasing factor-free morphism. If $h(w)$ is square-free for all square-free words of length 2, then h is comma-free.*

Proof. Set $X = h(A)$. Assume that X is not comma-free. Then there is a shortest word $uxv \in X^*$ with $x \in X$ and u or v not in X^*. Since X is a biprefix code, the minimality condition implies that u is the proper prefix of some word in X and similarly for v. Moreover, h being factor-free, the word x has no factor in X. Thus $uxv = yz$ for two elements y, z in X, and there are three letters a_1, a_2, a such that $h(a_1a_2) = uh(a)v$. Since the occurrences of $x = h(a)$, and $z = h(a_2)$ overlap, the word xz contains a square and therefore $a_1 = a$. Similarly, $a_2 = a$. But then x is a nontrivial factor of x^2 and thus, x itself contains a square, a contradiction. ∎

Proof of the theorem. Set $X = h(A)$. Assume now that the conclusion of the theorem is false. Then there is a shortest square-free word $w = a_1a_2 \cdots a_n$, where $a_1, \ldots, a_n$ are letters, such that $h(w)$ contains a square, say

$$h(w) = yuuz = x_1x_2 \cdots x_n$$

where $x_i = h(a_i)$ for $1 \le i \le n$. By the hypotheses, $n \ge 4$, and by the minimality of w, y is a proper prefix of x_1 and z is a proper suffix of x_n. Thus, there are words $s' \ne \varepsilon$ and $p' \ne \varepsilon$ with

$$x_1 = ys', \qquad x_n = p'z.$$

Next, u is not a prefix of s', since otherwise $x_2 \cdots x_{n-1}$ is a factor of u, thus also of x_1, contrary to the assumption that h is factor-free. Thus, there exists an index j with $1 < j < n$ and a factorization $x_j = ps$ such that

$$yu = x_1 \cdots x_{j-1}p, \qquad uz = sx_{j+1} \cdots x_n$$

[10] for the translator. Sometimes, such a code is called *infix*.

[11] Observe that the two statements that follow are true for arbitrary finite alphabets.

or, also,

$$u = s'x_2 \cdots x_{j-1}p = sx_{j+1} \cdots x_{n-1}p' \ .$$

Since $n \geq 4$, one has $j \geq 3$ or $n - j \leq 2$, i.e. at least one of the two occurrences of u contains one of the x_k's. By symmetry, we may assume $j \geq 3$. Thus

$$puz = ps'x_2 \cdots x_{j-1}pz = x_j x_{j+1} \cdots x_n \ .$$

Since X is comma-free and $x_2 \cdots x_{j-1} \neq \varepsilon$, this implies that ps' is in X^*, and since no element in X is a prefix of p nor a suffix of s', in fact ps' is in X. Thus $ps' = x_j$ and $s = s'$. It follows that $x_2 \cdots x_{j-1}p = x_{j+1} \cdots x_{n-1}p'$, which in turn implies $x_2 \cdots x_{j-1} = x_{j+1} \cdots x_{n-1}$ and $p = p'$. Altogether, we have obtained that

$$x_1 = ys, \quad x_j = ps, \quad , x_n = pz,$$

$$a_2 \cdots a_{j-1} = a_{j+1} \cdots a_{n-1} \ .$$

Now

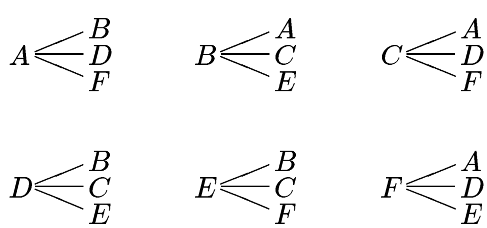
$$h(a_1 a_j a_n) = yspspz$$

contains a square, and thus $a_1 = a_j$ or $a_j = a_n$. But then w contains a square, a contradiction. ■

Every square-free word over three letters a, b, c that starts with the letter a and ends with b or c can be factorized into a product of words A, B, C, D, E, F, where

$$A = ab, \qquad C = abc, \qquad E = abcb$$
$$B = ac, \qquad D = acb, \qquad F = acbc \ .$$

The words AC, AE, BD, BF, CE, DF, CBa, DAa, EAa, EDa, FBa,FCa all contain squares. The same holds for the words ADB, BCA, CFD, DEC. On the contrary, the 18 words in the following diagram



all are square-free.

We observe also that in a twosided infinite square-free word, the words ABA and BAB do not appear as factors. Any occurrence of AFA, FAF, BEB, EBE, CDC, DCD always occurs as a factor of $BAFAB$, $CFAFD$, $ABEBA$, $DEBEC$, $BCDCA$, $ADCDB$ respectively.

These considerations lead to the morphisms h and g defined by

$$\begin{aligned} h(a) &= CA = abcab \\ h(b) &= BE = acabcb \\ h(c) &= FD = acbcacb \end{aligned}$$

and

$$\begin{aligned} g(a) &= AD = abacb \\ g(b) &= EB = abcbac \\ g(c) &= CF = abcacbc\ . \end{aligned}$$

It is immediately seen [12] that these morphisms have the properties required by the theorem. This proves that there exist arbitrarily long square-free words, and infinite square-free words over three letters.

COROLLARY 3.3. (Satz 18) *Let*

$$\mathbf{x} = (abcab)(acabcb)(acbcacb)(abcab)(acabcb)(abcab)\cdots$$

be the infinite word over a, b, c such that $h(\mathbf{x}) = \mathbf{x}$, where h is the morphism given above. Then $\mathbf{x}$ is square-free.

As we shall see, square-free words frequently are almost completely defined by the requirement that they do no contain factors in a certain set. We make some observations. First, every square-free word w over a, b, c of length at least 4 contains all three letters. If $|w| > 13$, then w contains each of the six possible two letter words ab, ac, ba, bc, ca, cb as factors. If $|w| > 30$, then each of the words abc, acb, bca, bac, cab and cba obtained by permuting the three letters is a factor of w.

12.— We now investigate in more detail those square-free words which contain 4 of the 6 words A, B, C, D, E, F given above in their decomposition. Each square-free word should lack one pair of factors among the following 15 pairs:

1) aba, aca	6) $aca, abca$	10) $abca, acba$
2) $aba, abca$	7) $aca, acba$	11) $abca, abcba$
3) $aba, acba$	8) $aca, abcba$	12) $abca, acbca$
4) $aba, abcba$	9) $aca, acbca$	13) $acba, abcba$
5) $aba, acbca$		14) $acba, acbca$
		15) $abcba, acbca$

The pairs of words (6), (7), (8), (9), (13) and (14) transform into the pairs (3), (2), (5), (4), (12) and (11) respectively by exchanging b and c. Thus, we do not need to consider the first group. Next, any square-free word w of length $|w| > 32$ necessarily contains one of the factors $abca$ or $abcba$ of group (11). Indeed, the

[12] A. Thue says.

prefix of length 31 of w contains abc which is followed by a or by ba. Similarly, one of the factors $abca$ or $acbca$ of group (12) must appear in w.

Also, any square-free word of length at least 60 [13] contains one of the words aba or $abca$ of group (2) as a factor, and the same holds for the words aba and $acba$ of group (3).

Finally, a square-free word of length more than 47 contains aba or $abcba$ as a factor. Indeed, the prefix of length 31 contains an occurrence of abc, and the next factor of length 16 must contain aba or $abcba$.

Thus, our investigation is reduced to the 4 cases (1), (5), (10) and (15). Now, we reduce case (10) to case (5). If a square-free word w does not have $abca$ or $acba$ as a factor, then w has no factor of the form $\alpha bab\beta$ or $\gamma cac\delta$, where $\alpha, \beta, \gamma, \delta$ are words of length at least 3[14]. Conversely, if w contains no factor of the form bab and cac, then it contains no factor of the form $\alpha abca\beta$ nor $\gamma acba\delta$, where α, β, γ, δ are letters. This reduces case (10) to case (5).

Thus, we restrict our investigation to square-free words over three letters a, b, c, where the pair of factors

$$aca \quad \text{and} \quad bcb \tag{I}$$

or

$$aba \quad \text{and} \quad aca \tag{II}$$

or

$$aba \quad \text{and} \quad bab \tag{III}$$

is missing.

3.4 First Case : *aca* and *bcb* are missing

13.— We shall call a word over a, b, c that is both square-free and has no factor of the form aca and bcb a *word of type* (I). Every infinite word $\mathbf{x}$ of type (I) is obtained from the periodic word

$$\cdots abababa ba \cdots$$

by interleaving it with the letter c[15]. Any factor of length at least 11 contains the word $caba$ or $cbab$. Let p denote a or b and q denote the other letter. Then we get the following ramification starting with $cpqp$:

---

[13] I found 41.

[14] Indeed, consider for instance $\alpha bab\beta$. Then α ends with c, thus with bc, thus with abc and symmetrically, β starts with cba. But then $abcbabcba$ contains a square.

[15] See also Thue's first paper.

Figure No. 4

This shows that every twosided infinite word of type (I) can be factorized into a product of words

$$\begin{aligned} x &= caba \\ y &= cbab \\ z &= cacb \\ u &= cbca\ . \end{aligned}$$

The same holds for circular words of type (I) of length at most 12. Next, the words xz, ux, yu, zy, zu, uz contain squares. Therefore, one obtains the following ramification (starting with zx):

Figure No. 5

Every twosided infinite word (or circular word of length at least 32) of type (I) is a product of the three words

$$\begin{aligned} A &= z &&= cacb \\ B &= xuy &&= cabacbcacbab \\ C &= xy &&= cabacbab\ . \end{aligned}$$

Define a morphism h from $\{a, b, c\}^*$ into itself by:

$$h\ :\ \begin{aligned} a &\mapsto A \\ b &\mapsto B \\ c &\mapsto C \end{aligned}$$

PROPOSITION 4.1. *If* $\mathbf{x}$ *is a twosided infinite word of type* (I), *then* $\mathbf{y} = h^{-1}(\mathbf{x})$ *is also of type* (I). *The same holds for circular words of length at least* 32.

Proof. It suffices to check that neither ACA nor BCB are factors of $\mathbf{x}$. Indeed, $BCB = xuyxuxuy$ contains a square, and if ACA is a factor of $\mathbf{x}$, then also $BACAB$, and therefore $yACAx = yzxyzx$. ∎

Observe that the morphism h is not factor-free because A is a factor of B. However, any occurrence of A, B or C in a word $h(w)$ coincides with an occurrence of $h(a)$, $h(b)$ or $h(c)$. Furthermore, the six words AB, AC, BC, BA, CA, CB are easily checked to be square-free.

THEOREM 4.2. (Satz 19) *If* $\mathbf{x}$ *is a twosided infinite word of type* (I), *then so is* $h(\mathbf{x})$.

Proof. A simple verification shows that $h(w)$ is square-free for all square-free words of length 3 except aca and bcb.

Assume that $h(\mathbf{x})$ contains a square tt. Then tt is not a factor of a word $h(v)$, where v is a factor of length 3 of $\mathbf{x}$. Thus there are words $p, q, s \in \{A, B, C\}$ and $r \in \{A, B, C\}^*$ such that

$$p = \gamma\beta,\ \ s = \alpha\beta,\ \ q = \alpha\delta,\ \ t = \beta r \alpha,\ \ prsrq = \gamma tt\delta$$

and $prsrq$ is a factor of $h(\mathbf{x})$:

<table>
<tr><td colspan="2">p</td><td>r</td><td colspan="3">s</td><td>r</td><td colspan="2">q</td></tr>
<tr><td>γ</td><td colspan="3">t</td><td colspan="4">t</td><td>δ</td></tr>
<tr><td></td><td>β</td><td>r</td><td colspan="2">α</td><td>β</td><td>r</td><td>α</td><td></td></tr>
</table>

Since $\mathbf{x}$ is square-free, one has $p \neq s \neq q$. Next, $psq = \gamma\beta\alpha\beta\alpha\delta$ has a square. Consequently, $psq = ACA$ or $psq = BCB$. If $p = A$, then either $r = B$ or r starts and ends with B, and $\mathbf{x}$ contains BCB, which is a contradiction. Similarly, $p \neq B$. This proves the proposition. ■

This result gives a method for constructing words of type (I). However, there is a relation between words of type (I) and overlap-free words which gives a more direct construction. For this, we consider a morphism

$$\tau : \{a, b, c\}^* \to \{\alpha, \beta\}^*$$

where α and β are two letters, defined by:

$$\tau\ :\ \begin{array}{l} a \mapsto \alpha \\ b \mapsto \alpha\beta\beta \\ c \mapsto \alpha\beta \end{array}$$

THEOREM 4.3. (Satz 20 & 21) *Let* $\mathbf{x}$ *be a twosided infinite overlap-free word over the two letters* α, β. *Then there exists a unique infinite word* $\mathbf{y}$ *over the three letters* a, b, c *such that* $\tau(\mathbf{y}) = \mathbf{x}$, *and moreover* $\mathbf{y}$ *is of type* (I). *Conversely, if* $\mathbf{y}$ *is of type* (I), *then* $\tau(\mathbf{y})$ *is overlap-free.*

Proof. Le $\mathbf{x}$ be an infinite overlap-free word over α and β. Clearly, there exists a unique word $\mathbf{y}$ such that $\tau(\mathbf{y}) = \mathbf{x}$. Assume that $\mathbf{y}$ contains a square uu. Then $\tau(u)$ starts with the letter α, and $\tau(uu)\alpha$ is an overlapping factor of $\mathbf{x}$. Thus, $\mathbf{y}$ is square-free.

Next, $\tau(bcb) = \alpha\beta\beta\alpha\beta\alpha\beta\beta$ contains an overlap, so bcb is not a factor of $\mathbf{y}$. If aca is a factor of $\mathbf{y}$, then so is $bacab$. But $\tau(bacab) = \alpha\beta\beta\alpha\alpha\beta\alpha\alpha\beta\beta$ contains an overlap, a contradiction. This proves that $\mathbf{y}$ is of type (I).

Assume conversely that $\mathbf{y}$ is of type (I), and set $\mathbf{x} = \tau(\mathbf{y})$. If $\mathbf{x}$ contains some overlap s, then s cannot be of the form $\alpha v \alpha v \alpha$, because $\mathbf{y}$ is square-free; thus $s = \beta v \beta v \beta$ for some nonempty word v. If v starts with a β, then it ends with α, and $v = \alpha w \beta$ for some w. But then αs is a factor of $\mathbf{x}$, and since

$$\alpha s = \alpha\beta\beta w\alpha\beta\beta w\alpha\beta$$

the word $\mathbf{y}$ contains a square.

We show now that, similarly, v does not start with the letter α. Indeed, if $v = \alpha$, then $s = \beta\alpha\beta\alpha\beta$ and since $\mathbf{y}$ is square-free, bcb is a factor of $\mathbf{y}$. Thus $v = \alpha w \gamma$ with $\gamma = \alpha$ or $\gamma = \beta$. If $\gamma = \beta$, then

$$s = \beta\alpha w\beta\beta\alpha w\beta\beta\ .$$

and $\mathbf{y}$ contains the square $\tau^{-1}(\alpha w\beta\beta)^2$. Thus $\gamma = \alpha$ and $v = \alpha w \alpha$, whence

$$s = \beta\alpha w\alpha\beta\alpha w\alpha\beta$$

Neither αs nor $s\alpha$ is a factor of $\mathbf{x}$ since otherwise $\mathbf{y}$ contains a square. Thus $\beta s \beta$ and even $\alpha\beta s\beta$ is a factor of $\mathbf{x}$. Since

$$\alpha\beta s\beta = \alpha\beta\beta\alpha w\alpha\beta\alpha w\alpha\beta\beta$$

the word $\mathbf{y}$ has a factor $bzczb$, with $\tau(z) = \alpha w$. Since $z \neq \varepsilon$, it starts and ends with the letter a. But then aca is a factor of $\mathbf{y}$, again a contradiction. ∎

3.5 Second Case : *aba* and *aca* are missing

14.— Since circular words can be treated in a way similar to twosided infinite words, it suffices to consider only words of the second kind. We shall call a word over a, b, c that is both square-free and has no factor of the form aba and aca a *word of type* (II).

Figure No. 6

In the present situation, we obtain the ramification:

Figure No. 7

Thus, a twosided infinite word $\mathbf{x}$ of type (II) is the product of words

$$\begin{aligned} x &= abc \\ y &= acb \\ z &= abcb \\ u &= acbc \ . \end{aligned}$$

Next, $\mathbf{x}$ has no factor of the form

$$xyx, yxy, xux, yzx,$$
$$wxwz, wywu, uwxw, zwyw$$

where w is in $\{x, y, z, u\}^*$. Indeed, the words

$$\begin{aligned} xyxu &= xy\,xy\,c \\ yxyz &= yx\,yx\,b \\ xuy &= ab\,cy\,cy \\ yzx &= ac\,bx\,bx \\ wxwz &= wx\,wx\,b \\ wywu &= wy\,wy\,b \\ uwxwa &= ac\,bcwa\,bcwa \\ zwywa &= ab\,cbwa\,cbwa \end{aligned}$$

all contain a square. Furthermore, $xwuwy$ and $ywzwx$ are not factors of $\mathbf{x}$ since

$$\begin{aligned} xwuwy &= a\,bcwac\,bcwac\,b \\ ywzwx &= a\,cbwab\,cbwab\,c \end{aligned}$$

have squares.

Set $X = \{x, y, z, u\}$. Of course, X is a (suffix) code. Since every word in X starts with the letter a and the letter a appears nowhere else in words in X, any twosided infinite word $\mathbf{x}$ of type (II) admits a unique factorization into words in X.

LEMMA 5.1. *Let p and q be two nonempty words in X^*, with $p \neq uz$, $q \neq zu$. Then neither pxp nor qyq are factors of an infinite word $\mathbf{x}$ of type* (II).

Proof. We prove the first claim, the second is shown in the same manner by exchanging b and c.

Assume on the contrary that pxp is a factor of a twosided infinite word $\mathbf{x}$ of type (II). Since neither $pxpx$ nor $pxpz$ are factors of $\mathbf{x}$, the factor pxp can only be followed by y or u. Similarly, it can only be preceded by y or z.

We first show that $p = rxuz$ for some nonempty $r \in X^*$. The last factor in X of p is neither x nor u (because ux is not a factor of $\mathbf{x}$), and it is not y since every occurrence of y is followed by x, which would imply that $pxpx$ is a factor of $\mathbf{x}$. Thus

$$p = p'z$$

for some $p' \in X^*$, and p' is not empty because xz is not a factor of $\mathbf{x}$. Next

$$p = p'z = p''uz$$

because neither xz nor yzx are factors of $\mathbf{x}$. By assumption, $p'' \neq \varepsilon$. The last factor of p'' in X is not y, because yu is not a factor of $\mathbf{x}$. Next, the code word following pxp is u, because p ends with z and zy is not a factor of $\mathbf{x}$. This implies that the last code word of p'' is not z. Thus $p'' = rx$ for some word r, and $r \neq \varepsilon$ since otherwise pxp contains the square xx.

The first word in X of p is u: indeed, it is neither x (since otherwise pxp contains the square xx) nor z (since otherwise pxp contains the factor $uzxz$), and it cannot be y since otherwise pxp must be preceded by z, and zy must be a factor of $\mathbf{x}$, which is impossible. Putting this all together, we have $p = usxuz$ for some word $s \in X^*$ which is nonempty, and consequently pxp admits the factor

$$xuzxus\ .$$

But s neither starts with u nor with x (because ux is not a factor of $\mathbf{x}$) nor with y (because xuy is not a factor). Thus pxp contains the square $(xuz)^2$, a contradiction. ■

We now change slightly the notation: we consider the set $T = \{x, y, z, u\}$ as a new alphabet and we introduce a morphism f from T^* into $\{a, b, c\}^*$ defined by

$$f\ :\ \begin{array}{l} x \mapsto abc \\ y \mapsto acb \\ z \mapsto abcb \\ u \mapsto acbc \end{array}$$

Define a set of words over T by

$$\mathcal{F} = \{wxwz, wywu, zwyz, uwxw \mid w \in T^*\} \cup \{xyx, yxy, xuy, yzx\}$$

and denote by $\mathcal{T}$ the set of twosided infinite words over T that are square-free and that have no factor in $\mathcal{F}$.

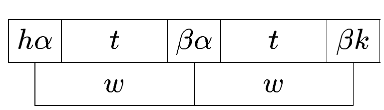
The discussion at the beginning of this section can be rephrased as: Every twosided infinite word $\mathbf{x}$ of type (II) is of the form $\mathbf{x} = f(\mathbf{y})$ for some $\mathbf{y} \in \mathcal{T}$. We now prove the converse:

THEOREM 5.2. (Satz 22) *If* $\mathbf{y}$ *is a word in* $\mathcal{T}$, *then* $f(\mathbf{y})$ *is of type* (II).

Proof. Set $X = \{f(x), f(y), f(z), f(u)\}$. Clearly, neither aba nor aca is a factor of $f(\mathbf{y})$. In order to show that $\mathbf{x} = f(\mathbf{y})$ is square-free, assume the contrary, and let ww be the shortest square in $\mathbf{x}$. Clearly, w contains at least one a. If w contains only one a, then ww is a factor of some word in X^3. However, it is easily checked that f preserves square-freeness of the factors of $\mathbf{y}$ of length 3. Thus $|w|_a \geq 2$, and consequently there are words h, α, β, t, k such that

$$h\alpha t\beta\alpha t\beta k$$

is a factor of $\mathbf{x}$, and further $h\alpha, \beta\alpha, \beta k \in X$, and $t \in X^*$, $t \neq \varepsilon$ and $w = \alpha t\beta$.



If $\beta = \varepsilon$, then $h = \varepsilon$ because X is a suffix code, and $\mathbf{y}$ contains a square. Thus $\beta \neq \varepsilon$. Let $s = f^{-1}(t)$. We prove now the contradiction by showing that $\beta\alpha$ cannot be the image of some letter in $\{x, y, z, u\}$. Assume first $\beta\alpha = f(x) = abc$. Then there are three cases, namely $(\beta, \alpha) = (abc, \varepsilon)$, $(\beta, \alpha) = (ab, c)$ and $(\beta, \alpha) = (a, bc)$. In all cases, $\mathbf{y}$ contains one of the words $xsxs$, $usxs$, $sxsz$, $sxsx$, but none of them appears as a factor in $\mathbf{y}$.

Consider now the case $\beta\alpha = f(z) = abcb$. Then, arguing as before, $\mathbf{y}$ contains as a factor one of the words $szsz$, $yszsx$, $zszs$, which is impossible.

The two cases $\beta\alpha = f(y)$ and $\beta\alpha = f(u)$ are handled by exchanging b and c. The proof is complete. ∎

We now go one step further. Consider a twosided infinite word $\mathbf{y}$ over the letters x, y, z, u that is in the set $\mathcal{T}$. The letters following some occurrence of z in $\mathbf{y}$ give rise to the following ramification

Figure No. 8

This shows that a word $\mathbf{y} \in \mathcal{T}$ can be factorized into a product of words

$$\begin{aligned} zuyxu &= A \\ zu &= B \\ zuy &= C \\ zxu &= D \\ zxy &= E\ . \end{aligned}$$

Again, the set $Z = \{A, B, C, D, E\}$ is a code, and the factorization of $\mathbf{y}$ is unique. The word $\mathbf{y}$ has no factor in the set

$$\mathcal{G} = \{AB, AD, BA, BC, CA, CD, CE, DB, DE, EC, ED, \\ BEB, EBE, DAC, DCBD, CBDC\}.$$

Also, $\mathbf{y}$ has no square of the form tt, with t in Z^*. Indeed,

$$
\begin{aligned}
ABz &= zuyxuzuz\\
ADz &= zuyxuzxuz\\
BA &= BByxu\\
BC &= BBy\\
CA &= CCxu\\
CD &= zuyzxu\\
CE &= zuyzxy\\
DBz &= zxuzuz\\
uDE &= uzxuzxy\\
ECzu &= zxyzuyzu\\
ED &= zxyzxu\\
yBEBzx &= yzuzxyzuzx\\
uEBEz &= uzxyzuzxyz\\
DAC &= zxuzuyxuzuy\\
BDCBDA &= BDzuyBDzuyxu\\
ACBDCBE &= zuyxuCBzxuCBzxy\ .
\end{aligned}
$$

We also observe that the word $\mathbf{y}$ has no factor of the form tAt with $t \in Z^*$, $t \neq \varepsilon$, $t \neq E$, and no factor of the form tBt with $t \in Z^*$, $t \neq \varepsilon$. Furthermore, any factor $\mathbf{y}$ of the form tCt or tDt, with $t \in Z^*$, $t \neq \varepsilon$, can be preceded and followed only by a E, and any factor of the form tEt can be preceded and followed only by a C.

Next, the word $\mathbf{y}$ has at least one occurrence of B, which implies the ramification

Figure No. 9

This shows that the word $\mathbf{y}$ is a product of the words

$$
\begin{aligned}
BDAEAC &= A'\\
BDC &= B'\\
BDAE &= C'\\
BEAC &= D'\\
BEAE &= E'\ .
\end{aligned}
$$

In other words, this leads to consider a new alphabet $Y = \{A, B, C, D, E\}$ and a morphism

$$\omega : Y^* \rightarrow Y^*$$

defined by

$$\omega : \begin{array}{l} A \mapsto BDAEAC \\ B \mapsto BDC \\ C \mapsto BDAE \\ D \mapsto BEAC \\ E \mapsto BEAE \end{array}$$

and a second morphism $h : Y^* \to T^*$ defined by

$$h : \begin{array}{l} A \mapsto zuyxu \\ B \mapsto zu \\ C \mapsto zuy \\ D \mapsto zxu \\ E \mapsto zxy \end{array}$$

(Remember also the morphism $f : T^* \to \{a, b, c\}^*$ defined at page 47 by

$$f : \begin{array}{l} x \mapsto abc \\ y \mapsto acb \\ z \mapsto abcb \\ u \mapsto acbc \end{array}$$

and which is intended to give words of type (II)!)

Define a set $\mathcal{Y}$ of twosided infinite words over the alphabet Y by the conditions that they are square-free, and that they have no factor in the set

$$\mathcal{G} = \{AB, AD, BA, BC, CA, CD, CE, DB, DE, EC, ED, \\ BEB, EBE, DAC, DCBD, CBDC\} .$$

We can restate the observation made above by saying that any infinite word $\mathbf{x}$ in $\mathcal{T}$ is of the form $\mathbf{x} = h(\mathbf{y})$ for some word in $\mathcal{Y}$. The following statement is concerned with $\mathcal{Y}$:

PROPOSITION 5.3. (Satz 23) *For any word* $\mathbf{y}$ *in* $\mathcal{Y}$*, there is a word* $\mathbf{z}$ *in* $\mathcal{Y}$ *such that* $\mathbf{y} = \omega(\mathbf{z})$.

Proof. We have seen already that there is an infinite word $\mathbf{z}$ over the alphabet Y such that $\mathbf{y} = \omega(\mathbf{z})$. Clearly, $\mathbf{z}$ is square-free. Next,

$$
\begin{aligned}
A'B' &= BDAEACBDC \\
A'D'B &= BDAEACBEACB \\
B'A' &= BDCBDAEAC \\
B'C' &= BDCBDAE \\
C'A' &= BDAEBDAEAC \\
C'D' &= BDAEBEAC \\
C'E' &= BDAEBEAE \\
D'B' &= BEACBDC \\
ED'E &= EBEACBEAE \\
CD'E' &= CBEACBEAE \\
E'C'BD &= BEAEBDAEBD \\
E'C'BE &= BEAEBDAEBE \\
E'D' &= BEAEBEAC \\
EB'E'B'BEA &= EB'BEAEB'BEA \\
CE'B'E'BD &= CE'BDCE'BD \\
D'A'C' &= BEACBDAEACBDAE \\
B'D'C'B'D'A' &= B'D'BDAEB'D'BDAEAC \\
A'C'B'D'C'B'E' &= BDAEACC'B'BEACC'B'BEAE \; .
\end{aligned}
$$

This proves the claim[16]. ■

The converse of the preceding proposition is more involved:

THEOREM 5.4. (Satz 24) *If* $\mathbf{z}$ *is a word in* $\mathcal{Y}$, *then* $\mathbf{y} = \omega(\mathbf{z})$ *is in* $\mathcal{Y}$.

Proof. The proof is by contradiction. It is easily seen that $\mathbf{y}$ has no factor in the set $\mathcal{G}$. It remains to prove that $\mathbf{y}$ is square-free. Assume the contrary, and let ww be a square in $\mathbf{y}$. Then w contains at least one occurrence of the letter B. In fact, w contains at least two occurrences of the letter B, since otherwise ww contains only two B's, which means that ww is a factor of a word $\omega(u)$ where u is a factor of length 3 of $\mathbf{z}$. Now, since $\mathbf{z}$ is in $\mathcal{Y}$, the factors of length 3 are ACB, AEA, AEB, BDA, BDC, BEA, CBD, CBE, DAE, DCB, EAC, EAE, EBD. It is easily checked that none of the images, by ω, of these words contain a square.

Thus w is of the form $w = \alpha t \beta$, where $t = \omega(s)$ for some nonempty factor s of $\mathbf{z}$, and where β and α are such that $\beta\alpha = \omega(N)$ for some letter N in Y, and furthermore there exist letters M, P in Y and words γ, δ such that $\gamma\alpha = \omega(M)$, $\beta\delta = \omega(P)$. In other words, setting

$$u = MsNsP$$

[16] A. Thue says. In fact, one must check that the words in the right column cannot appear as factors in $\mathbf{z}$. For instance, the first of these words ends with $CBDC$ which is in the forbidden set $\mathcal{G}$.

one has

$$\omega(u) = \gamma w w \delta, \quad w = \alpha\omega(s)\beta.$$

Since u is square-free, one has $M \neq N \neq P$. A last notation: we set $U = \omega(Y)$. The set U is a suffix code.

We first rule out the cases where $\beta = \varepsilon$ or $\alpha = \varepsilon$. If $\beta = \varepsilon$, then $\omega(N)$ is a suffix of $\omega(M)$. Since the code U is a suffix code, this implies $M = N$, a contradiction. Thus $\beta \neq \varepsilon$. If $\alpha = \varepsilon$, then $N = C$ and $P = A$ because only $\omega(C)$ is a prefix of $\omega(A)$. The only letter which can precede both C and A is D, and the only letter which can follow C is B. Thus s starts with B and ends with D, and the second letter of s (which is either D or E) is E since otherwise u contains the factor $DCBD$. Since s starts with BE, the initial letter M of u (which is either C or E) cannot be the letter E. Thus $M = C$, and $M = N$, a contradiction.

We now examine the possibilities for the letter N, and show that they all lead to a contradiction.

(i) $N = A$. Then $\beta\alpha = BDAEAC$. Since α is a suffix of another word in U, and α is a prefix of another word in U, the only factorizations are

$$(\beta, \alpha) = (BDA, EAC), \qquad (\beta, \alpha) = (BDAE, AC)$$

which both lead to $M = D$, $P = C$. But this implies that s starts with C and ends with D. Thus, u contains the factor DAC which is in $\mathcal{G}$, contradiction.

(ii) $N = B$. Here $\beta\alpha = BDC$, and in fact $\beta = BD$, $\alpha = C$ since DC is not a suffix of another word in U. Thus $M = A$ or $M = D$ (and $P = A$ or $P = C$).

If $M = A$, then $v = AsBs$ is a factor of $\mathbf{z}$. The first letter of s is E, and since EBE is not a factor, the last letter of s is C. Since C is only followed by B, this implies that $P = B$, which is impossible.

If $M = D$, then $v = DsBs$ is a factor of $\mathbf{z}$. However, there is no letter that can follow both D and B in a factor of $\mathbf{z}$, thus this case is impossible.

(iii) $N = C$. Here $\beta\alpha = BDAE$. It follows that $M = E$ and $P = A$ or $P = B$. The second case is ruled out by the fact that there is no letter preceding both B and C. Thus $u = EsCsA$. The first letter of s is B, and the last letter of s is D (the only letter that can precede both C and A). This shows that s has length at least 2. The second letter of s is not E, because EBE is not a factor, thus it is D. But this shows that $DCBD$ is a factor of u, and this is impossible since $DCBD \in \mathcal{G}$.

(iv) $N = D$. Here $\beta\alpha = BEAC$. The possible factorizations are $(\beta, \alpha) = (B, EAC)$, or $= (BE, AC)$, in which case $M = A$, and $(\beta, \alpha) = (BEA, C)$, in which case $M = A$ or $M = B$ and $P = E$.

Assume first that $M = A$, whence $u = AsDsP$. The first letter of s is C, and the last letter of s is B. Thus s has length at least 2. The second to last letter of

s is either C or E. It cannot be C since otherwise u contains the factor $CBDC$. Thus s ends with EB, and this implies that $P = D$, because EBE is not a factor. But then u contains a square, contradiction.

Assume now $u = BsDsE$. This is impossible because there is no letter that can follow both a B and a D in $\mathbf{z}$.

(v) $N = E$. Since $\beta\alpha = BEAC$, the possible factorizations are $(\beta, \alpha) = (BE, AE)$ or $= (BEA, E)$ and both lead to $M = C$ and $P = D$. Thus $u = CsEsD$. Since D is preceded only by B, the last letter of s is B. Since C is only followed by B, the first letter of s is B. Thus u contains the factor BEB, a contradiction.

The proof is complete. ■

For the characterization of words of type (II), there remains to prove that if $\mathbf{y}$ is an infinite word in $\mathcal{Y}$, then $h(\mathbf{y})$ is in $\mathcal{T}$. For this, we need a lemma.

LEMMA 5.5. (Satz 25) *A word* $\mathbf{y}$ *in* $\mathcal{Y}$ *has no factor of the form* $wAwC$, $DwAw$, $wEwD$, $CwEw$, wDw, wCw, wBw, *with* w *a nonempty word.*[17]

Proof. We argue by induction on the length of w, and show that if a word $\mathbf{y}$ in $\mathcal{Y}$ has a factor $wAwC$, then there is a word $\mathbf{y}'$ in $\mathcal{Y}$ that has a factor $DvAv$ with v shorter than w. The other proofs are similar.

Assume there is a word $\mathbf{y}$ in $\mathcal{Y}$ that has a factor $wAwC$ with $w \neq \varepsilon$. Then w ends with a D, and since AD is not a factor, $w = w_1D$ with $w_1 \neq \varepsilon$. Since DA can only be followed by the letter E, the word w_1 starts with E; thus $w_1 = Ew_2$, and $w_2 \neq \varepsilon$ because ED is not a factor. Now the letter preceding D in $wAwC = Ew_2DAEw_2DC$ is B, whence $w_2 = w_3B$. If $w_3 = \varepsilon$, then $wAwC = EBDAEBDC$, and there is no letter that can precede this word in $\mathbf{y}$. If $w_3 \neq \varepsilon$, we observe that the letter preceding the leftmost E cannot be A since this gives a square, and therefore is a B. Moreover, this initial BE can only be followed by A. Thus $w_3 = Aw_4$ for some w_4, and we get a factor

$$BwAwC = BEAw_4BDAEAw_4BDC \ .$$

Now, recall that $U = \omega(Y) = \{A', B', C', D', E'\}$. The decomposition shows that w_4 starts with the letter C, and since $CBDC$ is not a factor, $w_4 \neq C$, so that

$$BwAwC = D'w'A'w'B'$$

for some w' in U^*, and $w' \neq \varepsilon$. Thus $w' = \omega(v)$ for some v, and $DvAv$ is a factor of some word in $\mathcal{Y}$.

[17] The factor wBw is added here by the translator. It is implicit in the proof of the next Satz.

The argument is similar in the other cases, and we[18] only give the basic steps. Assume that $\mathbf{y}$ contains a factor $DwAw$ for some nonempty word w. Then it also contains the following factors:

$$DCw_1ACw_1$$
$$DCBw_2ACBw_2$$
$$DCBw_3EACBw_3EB$$

This shows that $\mathbf{y}$ contains a factor of the form $B'w'A'w'C'$, for some $w' \in U^*$. Thus some word in $\mathcal{Y}$ contains a factor of the form $BvAvC$, and since BA is not a factor, $v \neq \varepsilon$.

Assume now that $\mathbf{y}$ contains a factor

$$CwEw$$

for some nonempty word w. Then it also contains the following factors.

$$CBw_1EBw_1$$
$$CBw_2AEBw_2A$$
$$CBw_2AEBw_2ACB\ .$$

This shows that $\mathbf{y}$ contains a factor of the form $w'E'w'D'$, for some $w' \in U^*$. Thus some word in $\mathcal{Y}$ contains a factor of the form $vEvD$, and since ED is not a factor, $v \neq \varepsilon$.

Symmetrically, assume that $\mathbf{y}$ contains a factor

$$wEwD$$

for some nonempty word w. Then it also contains the following factors:

$$w_1BEw_1BD$$
$$Aw_2BEAw_2BD$$
$$BDAw_2BEAw_2BD\ .$$

This shows that $\mathbf{y}$ contains a factor of the form $C'w'E'w'$, for some $w' \in U^*$. Thus some word in $\mathcal{Y}$ contains a factor of the form $CvEv$, and since CE is not a factor, $v \neq \varepsilon$.

Assume now that $\mathbf{y}$ contains a factor

$$wDw$$

for some nonempty word w. Then it also contains the following factors.

$$w_1BDw_1BE$$
$$w_2CBDw_2CBE$$
$$Aw_3CBDAw_3CBE$$
$$EAEw_4CBDAEw_4CBE\ .$$

[18] and Axel Thue

This shows that **y** contains a factor of the form $E'w'C'w'$, for some nonempty $w' \in U^*$. Thus some word in $\mathcal{Y}$ contains a factor of the form $EvCv$, for some $v \neq \varepsilon$.

Assume next that **y** contains a factor

$$wCw$$

for some nonempty word w. Then it also contains the following factors.

$$\begin{gathered} EBw_1CBw_1 \\ EBDw_2CBDw_2 \\ EBDw_3ACBDw_3AE\ . \end{gathered}$$

This shows that **y** contains a factor of the form $w'D'w'E'$, for some nonempty $w' \in U^*$. Thus some word in $\mathcal{Y}$ contains a factor of the form $vDvE$, for some $v \neq \varepsilon$.

Assume finally that **y** contains a factor

$$wBw$$

for some nonempty word w. Then it also contains the following factors.

$$\begin{gathered} w_1EBw_1EA \\ Dw_2EBDw_2EA\ . \end{gathered}$$

and since a letter D can only be preceded by a B, the word **y** contains a square, contradiction. The proof is complete. ∎

THEOREM 5.6. (Satz 26) *For all* $\mathbf{y} \in \mathcal{Y}$, *the word* $h(\mathbf{y})$ *is in* $\mathcal{T}$.

Proof. Let $\mathbf{y} \in \mathcal{Y}$. It is easily seen that the word $\mathbf{t} = h(\mathbf{y})$ has no factors of the form

$$xz, yu, zy, ux, xyx, yxy, xuy, yzx$$

(the last because CE is not a factor of **y**). It remains to show **t** has no factors of the form

$$wxwz, wywu, zwyw, uwxw$$

for $w \neq \varepsilon$, and that it is square-free.

Recall that the set $Z = \{zuyxu, zu, zuy, zxu, zxy\} = h(Y)$ is a suffix code, and since every word in Z starts with the letter z, it has *deciphering delay* 1.

Assume first that **t** contains a factor

$$wxwz$$

for some nonempty word w. Then it contains

$$w_1yxw_1yz$$

because the only letter in T that can precede both x and z is y. Inspection of Z shows that the factor yx appears only in $zuyxu = h(A)$. Thus w_1 starts with u, and **t** contains the factor

$$uw_2yxuw_2yz\ .$$

Moreover, w_2 is nonempty because xu is only followed by z. Thus **t** contains

$$uw_3h(A)w_3h(C)z$$

where $w_3 = h(W)$ for some word $W \in Y^*$. If $W \neq \varepsilon$, this contradicts the preceding lemma, and if $W = \varepsilon$, the word **t** contains $uh(AC)$, which implies that **y** contains AAC, BAC or DAC. All these cases are impossible.

Assume now that **t** contains a factor

$$wywu$$

for some nonempty word w. Then it contains

$$w_1xyw_1xu$$

with $w_1 \neq \varepsilon$, and also

$$zw_2xyzw_2xu$$

and $w \neq \varepsilon$ since otherwise **t** contains a factor $h(ED)$. By inspecting Z, one sees that a factor xy is preceded by a z. Thus **t** contains a factor

$$zw_3zxyzw_3zxu\ .$$

Thus $zw_3 = h(W)$ for some nonempty word $W \in Y^*$, and $WEWD$ is a factor of **y**, contradiction.

Assume next that **t** contains a factor

$$q = zwyw$$

for some nonempty word w. Then it contains

$$zxw_1yxw_1$$

with $w_1 \neq \varepsilon$ because xyx is not a factor. But w_1 starts with u, and zxw_1 ends with zu. Thus $w_1 = uw_2zu$ for some w_2, and the factor q is

$$zxuw_2zuyxuw_2zu = h(DWAWN)$$

for some word $W \in Y^*$ and some letter $N \in \{A, B, C\}$. In view of the lemma, $W = \varepsilon$. But **y** is square-free and has neither AB nor DAC as a factor. Contradiction.

Assume next that **t** contains a factor

$$q = uwxw$$

for some nonempty word w. Then it contains

$$uyw_1xyw_1$$

and w_1 ends with a letter z. Thus

$$q = uyw_2zxyw_2z$$

showing that **t** contains a factor $CWEW$ for some word $W \in Y^*$, which is impossible.

We now prove that **t** is square-free, arguing by contradiction. Assume that ww is a square factor of **t**. Clearly, w contains at least one occurrence of the letter z. In fact, it contains two occurrences of z, since otherwise ww would be a factor of a word of the form $h(s)$, where $s \in Y^*$ has length 3. Now, the factors of length 3 of **y** are

$$ACB, AEA, AEB, BDA, BDC, BEA, CBD, CBE,$$
$$DAE, DCB, EAC, EAE, EBD$$

and their images are all easily checked to be square-free.

It follows that, as in the proof of Satz 24, there is a factorization

$$w = \alpha t \beta$$

and words γ, δ where

$$t = h(s),\ s \in Y^*,\ s \neq \varepsilon,$$
$$\beta\alpha = h(N),\ \gamma\alpha = h(M),\ \beta\delta = h(P),\ M, N, P \in Y$$

and

$$\gamma ww\delta = h(MsNsP)\ .$$

Of course $M \neq N \neq P$. If $\beta = \varepsilon$ then as above $M = N$ because Z is a suffix code. Next, we observe that, by the lemma, the letter N is neither B, C, nor D. If $\alpha = \varepsilon$, then $h(N)$ is a prefix of $h(P)$, and this would imply that N is B or C which was just ruled out. Let us consider the remaining cases.

(i) $N = A$. Then $\beta\alpha = zuyxu$, and the only possibility is in fact $(\beta, \gamma) = (zuy, xu)$. This implies that $M = D$, in contradiction with the fact that **y** has no factor of the form $DsAs$.

(ii) $N = E$. Either $(\beta, \alpha) = (z, uy)$ and $M = E$ or $(\beta, \alpha) = (zu, y)$ and $P = D$. The first case yields a square, and the second contradicts the lemma. ∎

3.6 Third Case : *aba* and *bab* are missing

15.— We shall call a word over a, b, c that is both square-free and has no factor of the form aba and bab a *word of type* (III).

In this case, we obtain the ramification:

Figure No. 10

As in the second case, we consider an alphabet $T = \{x, y, z, u\}$, a set of words $\mathcal{F}$ over T defined by

$$\mathcal{F} = \{wxwz, wywu, zwyz, uwxw \mid w \in T^*\} \cup \{xyx, yxy, xuy, yzx\}$$

and we denote by $\mathcal{T}$ the set of twosided infinite words over T that are square-free and that have no factor in $\mathcal{F}$.

Here, we introduce a morphism g from T^* into $\{a, b, c\}^*$ defined by

$$g \ : \ \begin{array}{l} x \mapsto ca \\ y \mapsto cb \\ z \mapsto cab \\ u \mapsto cba \end{array}$$

In view of the ramification given above, every word $\mathbf{x}$ of type (III) admits a unique inverse image by g: i. e. there is a unique infinite word $\mathbf{t}$ over T such that $g(\mathbf{t}) = \mathbf{x}$. We observe the following

FACT. *If* $\mathbf{x} = g(\mathbf{t})$ *is of type* (III), *then* $\mathbf{t}$ *is in* $\mathcal{T}$.

Proof. It suffices to show that $\mathbf{t}$ is square-free (this is clear) and that it has no factor in the set $\mathcal{F}$. And indeed, since $g(z) = g(x)b$ and $g(u) = g(y)a$, the words $g(wxwz)$ and $g(wywu)$ contain squares. Next,

$$\begin{aligned} g(zwyw)c &= cabg(w)cbg(w)c \\ g(uwxw)c &= cbag(w)cag(w)c \\ g(xyx)cb &= cacbcacb \\ g(yxy)ca &= cbcacbca \\ g(yzx) &= cbcabca \\ g(xuy) &= cacbacb \end{aligned}$$

This proves the claim. ■

Recall that, for infinite words of type (II), we considered above (page 47) the morphism f from T^* into $\{a,b,c\}^*$ defined by

$$f \;:\; \begin{array}{l} x \mapsto abc \\ y \mapsto acb \\ z \mapsto abcb \\ u \mapsto acbc \end{array}$$

In view of Satz 22, we obtain directly:

THEOREM 6.1. (Satz 27) *If* $\mathbf{x}$ *is a word of type* (III), *then* $f(g^{-1}(\mathbf{x}))$ *is a word of type* (II). ∎

The converse also holds. For the proof, we give an alternative construction. Introduce a new morphism $\bar{f}$ from T^* into $\{a,b,c\}^*$ defined by

$$\bar{f} \;:\; \begin{array}{l} x \mapsto cba \\ y \mapsto cab \\ z \mapsto cbab \\ u \mapsto caba \end{array}$$

obtained from f by exchanging the letters a and c. Then for $\mathbf{y}$ of type (II) (i.e. without factors cbc and cac), the word

$$\mathbf{x} = g(\bar{f}^{-1}(\mathbf{y}))$$

is obtained from $\mathbf{y}$ by deleting each letter that follows immediately an occurrence of c in $\mathbf{y}$.

THEOREM 6.2. (Satz 28) *If* $\mathbf{y}$ *is a word of type* (II), *(i. e. is square-free and without factors cbc and cac), then* $g(\bar{f}^{-1}(\mathbf{y}))$ *is a word of type* (III).

Proof. Set $\mathbf{x} = g(\bar{f}^{-1}(\mathbf{y}))$. It is straighforward that $\mathbf{x}$ has no factor of the form aba and bab. Assume that $\mathbf{x}$ contains a square ww. Clearly, w contains at least one occurrence of the letter c. Setting $X = \{ca, cb, cab, cba\}$, we may decompose

$$w = \gamma v c \beta$$

with $v \in X^*$ and $\gamma, \beta \in \{a,b\}^*$. Then

$$ww = \gamma v c \beta \gamma v c \beta$$

and $c\beta\gamma \in X$. If $\beta \neq \varepsilon$, we may assume that β starts with the letter b. Then there is in $\mathbf{y}$ a factor

$$\gamma u c a \beta \gamma u c a \beta$$

with u mapping on v. But this factor contains a square, contradiction. Thus $\beta = \varepsilon$. Again, we may assume that γ starts with b, so $\gamma = b$ or $\gamma = ba$. Then

$$ww = bvcbvc \quad \text{or} \quad ww = bavcbavc\ .$$

Thus $\mathbf{y}$ contains a factor of the form

$$bucabuc \quad \text{or} \quad abaucabauc$$

with u mapping on v. In the second case, we obtain a square. In the first case, the initial letter is preceded, in $\mathbf{y}$, by the letter a, so again there is a square. This completes the proof. ∎

16.— Finally, we observe:

THEOREM 6.3. (Satz 29) *Let*

$$\mathbf{x} = x_0 x_1 x_2 \cdots$$

be an infinite square-free word over three letters. Then there exists a factorization

$$\mathbf{x} = u\mathbf{y}$$

such that $\mathbf{y}$ *has no prefix of the form* waw, *where* a *is a letter and* w *is a nonempty word.*

For the proof, it will be convenient to set $\mathbf{x}_i = x_i x_{i+1} \cdots$ for $i \geq 0$. We argue by contradiction, and assume that any $\mathbf{x}_i$ admits a prefix of the form waw, with a a letter and w a nonempty word.

LEMMA 6.4. *Let* u *be a nonempty factor of* $\mathbf{x}_1$, *and let* a *be a letter such that* au *is not a factorof* $\mathbf{x}$. *If*

$$uz = wdwy$$

is a factor of $\mathbf{x}$ *for some words* z, $w \neq \varepsilon$, y *and some letter* d, *then* $d = a$ *and* $|w| < |u|$.

Proof. Let c be the first letter of u. Since u is a factor of $\mathbf{x}_1$, there is a letter b such that bu is a factor of $\mathbf{x}$, and $b \neq c$, $b \neq a$. By assumption

$$buz = bwdwy.$$

Since $\mathbf{x}$ is square-free, $d \neq b$, and since u (hence w) starts with the letter c, one has $d \neq c$. Thus $d = a$. Next, if $|w| \geq |u|$, then dw starts with au and au is a factor of $\mathbf{x}$, a contradiction. ∎

The proof of the theorem is by repeated application of the lemma. We first prove that a specific word cannot be a factor, and then, removing the initial letters, reduce this word to a short word that must appear in $\mathbf{x}$.

(i) *The word* $u = abcacbabca$ *is not a factor of* $\mathbf{x}_2$.
Indeed, observe first that $u = vcbv$ with $v = abca$. Thus cbu is not a factor of $\mathbf{x}$. This implies that bu is not a factor of $\mathbf{x}_1$ because bu can be preceded neither by a nor by b. Since u is a factor and bu is not, the assumptions of the theorem and the lemma show that there are words z, $w \neq \varepsilon$, y such that

$$uz = wbwy \ .$$

Since u has 3 occurrences of the letter b and $|w| < |u|$, one has $w = a$, $w = abcac$, or $w = abcacba$. The first and the last case are immediately ruled out. In the second case, $wbw = uc = vcbvc$, and since this factor is always followed by a b, this also is impossible.

(ii) *Set* $u_1 = cacbabca$ *(i.e.* $u = abu_1$*). Then* u_1b *is not a factor of* $\mathbf{x}_4$.
We show that bu_1b is not a factor of $\mathbf{x}_3$. The result follows because any occurrence of u_1 is preceded by a b. Assume bu_1b is a factor. Since abu_1b is not a factor, we may apply the lemma. A factorization

$$bu_1bz = wawy$$

with $|w| < |bu_1b|$ implies $w = bcacbabc$ (the two other cases are clearly impossible). But then waw contains the square $abcabc$.

(iii) *Set* $u_2 = acbabca$ *(i.e.* $u_1 = cu_2$*). Then* u_2b *is not a factor of* $\mathbf{x}_5$.
Indeed, since cu_2b is not a factor, the equation

$$u_2bz = wcwy$$

implies $w = a$ or $w = acbab$, and both are impossible.

(iv) *Set* $u_3 = cbabca$ *(i.e.* $u_2 = au_3$*.) Then* u_3b *is not a factor of* $\mathbf{x}_6$.
Indeed, since au_3b is not a factor, we obtain the equation $u_3bz = wawy$ with $|w| < |u_3b|$ which clearly is impossible.

(v) *Set* $u_4 = babca$ *(i.e.* $u_3 = cu_4$*.) Then* u_4b *is not a factor of* $\mathbf{x}_7$.
Indeed, otherwise we get the equation $u_4bz = wcwy$, whence $w = bab$, a contradiction.

(vi) *Set* $u_5 = abca$ *(i.e.* $u_4 = bu_5$*.) Then* u_5b *is not a factor of* $\mathbf{x}_8$.
Indeed, otherwise we get the equation $u_5bz = wbwy$, whence $w = a$, a contradiction.

(vii) *Set* $u_6 = bca$ *(i.e.* $u_5 = au_6$*.) Then* u_6b *is not a factor of* $\mathbf{x}_9$.
Indeed, otherwise we get the equation $u_6bz = wawy$, whence $w = bc$, a contradiction.

(viii) *Set* $u_7 = ca$ *(i.e.* $u_6 = bu_7$.*) Then* u_7b *is not a factor of* $\mathbf{x}_{10}$.
Indeed, otherwise we get the equation $u_7bz = wbw$ which has no solution.

Thus, we have shown that cab is not a factor of $\mathbf{x}_{10}$. But we have seen earlier that every square-free word of length at least 31 over three letters contains any factor of length 3 composed of the three letters. This leads to the desired contradiction and proves the theorem.

3.7 Irreducible words over four letters

17.— According to our general definition, a word w over a four letter alphabet is called *irreducible* if any two distinct occurrences of a factor in w are separated by at least two letters. For simplicity, we consider here only twosided infinite words.

Let $A = \{a, b, c, d\}$ be a four-letter alphabet, and let $B = \{x, y, z, u, v, w\}$ be a six-letter alphabet. Consider a morphism $f : A^* \to B^*$ defined by

$$f : \begin{array}{l} x \mapsto abcad \\ y \mapsto acbad \\ z \mapsto bacbd \\ u \mapsto bcabd \\ v \mapsto cabcd \\ w \mapsto cbacd \end{array}$$

The set $X = \{f(x), f(y), f(z), f(u), f(v), f(w)\}$ is a comma-free code, because the letter d occurs only at the end of each codeword. Moreover, the code has another interesting property[19]. A word α is called a *characteristic* prefix of $x \in X$ if α is a prefix of x and if no other codeword in X has α as a prefix. A symmetric definition holds for characteristic suffixes. The code X has the property that, for any $x \in X$ and any factorization $x = \alpha h \beta$, with $h \in A$, either α or β is characteristic for x.

Set

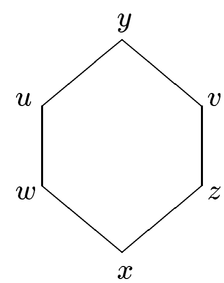
$$H = \{xz, xw, yu, yv, zx, zv, uy, uw, vy, vz, wx, wu\}$$

and

$$\mathcal{H} = f(H) = \{f(h) \mid h \in H\} \ .$$

Write the letters of the alphabet on the vertices of a polygon as follows:

[19] See also earlier.



Then the set H is composed of pairs of adjacent letters. It is easily verified that all words in $\mathcal{H}$ are irreducible.

THEOREM 7.1. (Satz 30) *Let* $\mathbf{x}$ *be a twosided infinite word over* B *such that all its factors of length* 2 *are in* $\mathcal{H}$. *If* $\mathbf{x}$ *is square-free, then* $f(\mathbf{x})$ *is irreducible.*

Proof. Assume that the word $\mathbf{y} = f(\mathbf{x})$ is reducible. Then $\mathbf{y}$ contains a factor tkt, where $|k| \leq 1$. Assume first that t has a factor that is in the code X. Then there are words α, β and $s \in X^*$, $s \neq \varepsilon$ such that $t = \beta s \alpha$, and moreover $\alpha k \beta \in X$, i.e. setting $q = \alpha k \beta$,

$$tkt = \beta s \alpha k \beta s \alpha = \beta s q s \alpha.$$

Since α or β is characteristic for q, either the prefix β of tkt is the suffix of an occurrence of q, or the suffix α of tkt is the prefix of an occurrence of q. Thus, either $qsqs$ or $sqsq$ is a factor if $\mathbf{y}$ and $\mathbf{x}$ contain a square.

Since tkt is not a factor of a word of $\mathcal{H}$, it remains to consider the case where tkt is a factor of some word $q_1q_2q_3$ in X^3. As before, one has $\alpha k \beta = q_2$ for some words α and β, and $t = \beta\alpha$. Thus $q_1 = \gamma\beta$ and $q_3 = \alpha\delta$, and since α or β is characteristic for q_2, it follows that $q_1 = q_2$ or $q_2 = q_3$. This is impossible and proves the theorem. ∎

We now show how to construct twosided infinite words of the kind described in the theorem, i.e. that are square-free and have all their factors of length two in the set H. We shall see that even five letters are sufficient. It is immediately seen that at least five letters are required.

Assume that the letter w does not appear in a twosided infinite word $\mathbf{x}$ that is both square-free and has all its factors of length two in the set H. Then, in following the cycle in the picture, one sees that any two consecutive occurrences of the letter y are separated by u, v, vzv or $vzxzv$. Thus $\mathbf{x}$ is a product of the words yu, yv, $yvzv$ and $yvzxzv$. In fact, $\mathbf{x}$ cannot contain the factor yvy, since otherwise it would also contain $vyuyvyuy$ which is a square. Thus, $\mathbf{x}$ is a product of the three words yu, $yvzv$ and $yvzxzv$. Define a morphism $\sigma : \{a,b,c\}^* \to \{x,y,z,u,v\}^*$ by

$$\sigma : \begin{array}{l} a \mapsto yu \\ b \mapsto yvzv \\ c \mapsto yvzxzv \end{array}$$

The word $\mathbf{x}$ has no factor of the form $\sigma(aba)$ or $\sigma(cbc)$, since

$$\sigma(cabac) = yvzxz(vyuyvz)(vyuyvz)xzv$$

and

$$\sigma(cbc) = yvzx(zvyv)(zvyv)zxzv$$

both contain squares.

THEOREM 7.2. (Satz 31) *Let* $\mathbf{z}$ *be a twosided infinite word over* a, b, c *that is square-free and has no factor* aba *and* cbc[20]. *Then* $\sigma(\mathbf{z})$ *is a square-free word with all its factors of length* 2 *in the set* H.

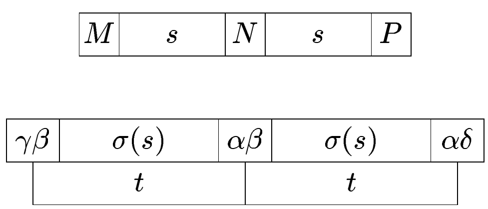
This of course implies that $f(\sigma(\mathbf{z}))$ is irreducible for every infinite word $\mathbf{z}$ of type (I).

Proof. Set $\mathbf{y} = \sigma(\mathbf{z})$. By construction, the factors of length two of $\mathbf{y}$ are all in H. It remains to show that $\mathbf{y}$ is square-free. It is easily checked that the image, by σ, of any factor of length 3 of $\mathbf{z}$ is square-free. Thus, if $\mathbf{y}$ contains a square tt, then the shortest factor p of $\mathbf{z}$ such that tt is a factor of $\sigma(p)$ has length at least 4. Thus, there are letters M, N, P in $\{a, b, c\}$, a word $s \in \{a, b, c\}^*$ and words α, β, γ, δ in $\{x, y, z, u, v\}^*$ such that

$$\sigma(M) = \gamma\beta,\ \sigma(N) = \alpha\beta,\ \sigma(P) = \alpha\delta,\ t = \beta\sigma(s)\alpha$$

and

$$\gamma tt\delta = \sigma(MsNsP)$$



If $N = a$ or $N = c$, then either α or β is characteristic for N. Indeed, if $N = a$, then either α or β contains the letter u which appears nowhere else in the words $\{\sigma(a), \sigma(b), \sigma(c)\}$. In the second case, the same holds with the letter x. In these cases, $N = M$ or $N = P$ and $\mathbf{z}$ contains a square. Thus $N = b$, whence $\alpha\beta = yvzv$. If α or β is empty, the word $\mathbf{x}$ contains a square. If $\alpha = y$, then $M = b$, again impossible. In the two remaining cases, a square is avoided only if $M = P = c$. Thus $\mathbf{x}$ contains the factor $csbsc$. This implies that s is not empty, and that it starts and ends with the letter a. But this in turn shows that aba is a factor of $\mathbf{x}$. This proves the claim. ∎

Similar arguments show how to construct arbitrarily long circular words which are irreducible.

[20]It is of type (I).

3.8 Irreducible words over more than four letters

We show here how to contruct, for any integer $n > 4$, arbitrarily long words over an alphabet with n letters such that any two occurrences of a factor are separated by at least $n - 2$ symbols.

We consider first the case where n is even, and set $n = 2h$. We consider an alphabet $\{a_1, a_2, \dots a_n\}$. Our purpose is to build a morphism that maps a square-free word over three letters into an irreducible word over $\{a_1, a_2, \dots a_n\}$. For this, we construct three sequences of words of special form. First, consider a sequence $u = u_0, u_1, \dots, u_h$ of words of length $n + 1$ defined by

$$u = u_0 = a_1 a_2 \cdots a_{n-1} a_1 a_n$$

obtained by inserting the letter a_1 in $a_1 \cdots a_n$ between a_{n-1} and a_n. Next

$$u_k = \sigma(u_{k-1}) \qquad 1 \leq k < h$$

where σ is the permutation defined by[21]

$$\sigma(a_i) = \begin{cases} a_i & \text{if } i \text{ is even} \\ a_{i+2 \bmod n} & \text{if } i \text{ is odd} \end{cases}$$

Thus

$$\begin{aligned} u_0 &= a_1 a_2 a_3 a_4 a_5 \cdots a_{n-1} a_1 a_n \\ u_1 &= a_3 a_2 a_5 a_4 a_7 \cdots a_1 a_3 a_n \\ u_2 &= a_5 a_2 a_7 a_4 a_9 \cdots a_3 a_5 a_n \\ &\cdots \\ u_{h-1} &= a_{n-1} a_2 a_1 a_4 a_3 \cdots a_{n-3} a_{n-1} a_n \\ u_h &= u \end{aligned}$$

We first prove that $u_0 u_1$ is irreducible. This implies that every word $u_k u_{k+1}$ $(0 \leq k < h)$ is irreducible over $\{a_1, a_2, \dots a_n\}$. Any factor of length at least 2 has only one occurrence in $u_0 u_1$. Indeed, this is clear for the factors $a_{n-1} a_1$ and $a_1 a_3$. All other factors of length two contain a letter with even index which is preceded or followed by a different letter of index in its two occurrences. Next, two occurrences of the same letter are separated by at least $n - 2$ letters.

Set

$$p = u_0 u_1 \cdots u_{h-1} \ .$$

This is the first word we are looking for. The words p and pu are irreducible. Indeed, the same argument as before shows that only letters have more than one occurrence in p, and occurrences of the same factor of length greater than 1 in $pu = u u_1 \cdots u_{h-1} u$ are separated by at least $(h - 1)(h + 1)$ letters.

[21] We write improperly $j \bmod n$ for $1 + (j - 1 \bmod n)$.

A second sequence $v_0, v_1, \ldots, v_h, v_{h+1}$ of words of length $n+1$ is defined by $v_0 = u$ and

$$v_k = \tau(v_{k-1}) \qquad 1 \le k \le h$$

where τ is the permutation given by

$$\begin{aligned} \tau(a_1) &= a_2 \\ \tau(a_2) &= a_3 \\ \tau(a_i) &= \begin{cases} a_i & \text{if } i \text{ is even, } i > 2 \\ a_{i+2 \bmod n} & \text{if } i \text{ is odd, } i > 1 \end{cases} \end{aligned}$$

Thus

$$\begin{aligned} v_0 &= a_1 a_2 a_3 a_4 a_5 \cdots a_{n-1} a_1 a_n \\ v_1 &= a_2 a_3 a_5 a_4 a_7 \cdots a_1 a_2 a_n \\ v_2 &= a_3 a_5 a_7 a_4 a_9 \cdots a_2 a_3 a_n \\ v_3 &= a_5 a_7 a_9 a_4 a_{11} \cdots a_3 a_5 a_n \\ &\cdots \\ v_{h-1} &= a_{n-3} a_{n-1} a_1 a_4 a_2 \cdots a_{n-5} a_{n-3} a_n \\ v_h &= a_{n-1} a_1 a_2 a_4 a_3 \cdots a_{n-3} a_{n-1} a_n \\ v_{h+1} &= u \end{aligned}$$

Observe that v_h and u_{h-1} are obtained from each other by exchanging a_1 and a_2. Next, $v_0 v_1$ is irreducible. Indeed, two occurrences of the same letter are separated by at least $n-2$ letters, and the only two factors of length 2 which appear twice in $v_0 v_1$, namely $a_2 a_3$ and $a_1 a_2$ are separated by words of length $n-2$ and $2n-3$. Thus $v_0 v_1$ and consequently all $v_k v_{k+1}$ for $0 \le k \le h$ are irreducible.

Our second word is

$$q = v_0 v_1 \cdots v_h \ .$$

This word is also irreducible. Assume indeed that q contains two distinct occurrences of the same factor. If this factor has length greater than 3, then it contains one of the letters $a_4, a_6, \ldots, a_n$. But two occurrences of these letters are never followed or preceded by the same letter. Thus, the factor has length at most 3, and contains none of $a_4, a_{,} \ldots, a_n$. Two occurrences of this type are easily checked to be separated by a word of length at least $n-2$.

Finally, we consider the word

$$r = w_0 w_1 w_2 \cdots w_{h-1}$$

where each w_k is obtained from u_k by exchanging a_1 and a_2. Since p is irreducible, so is r. Moreover, one has $v_h = w_{h-1}$ and $w_{h-1} u = v_h v_{h+1}$ is irreducible. It is convenient to write $v = v_h$.

Define a morphism $h : \{a, b, c\}^* \to \{a_1, \ldots, a_n\}^*$ by

$$f \ : \ \begin{aligned} a &\mapsto p = u u_1 \cdots u_{h-1} \\ b &\mapsto q = u v_1 \cdots v_{h-1} v \\ c &\mapsto r = w_0 \cdots w_{h-2} v \end{aligned}$$

Then the following result holds :

THEOREM 8.1. (Satz 32) *For every twosided infinite square-free word* $\mathbf{x}$ *over* $\{a, b, c\}$, *the word* $f(\mathbf{x})$ *is irreducible.*

Proof. We observe first that the words $u_{h-1}u_0$, $u_{h-1}w_0$, v_hu_0, v_hw_0 are irreducible. Thus, in the word $\mathbf{y} = f(\mathbf{x})$, a reducible factor is not contained in the product of two of the u_i's, v_i's, w_i's. Denote by S the set

$$S = \{u_0, \ldots u_{h-1}, v_1, \ldots, v_h, w_0, \ldots, w_{h-1}\} .$$

This set is a uniform code. The fact that every codeword ends with the letter x_n shows that S is a comma-free code. Moreover, every codeword is *characterized* by its prefix of length 3.

Similarly, the set $X = \{p, q, r\}$ is a comma-free code. Moreover, in any factorization $\alpha\beta$ of a word in $x \in X$ either α or β is characteristic for x. We finally observe that if ss', with $s, s' \in S$ is a factor of some word xx', with $x, x' \in X$, then $s \neq s'$ and even s and s' have different suffixes of length 2. These suffixes are of the form aa_n and $a'a_n$ for two letters a, a' in $\{a_1, \ldots a_n\}$. If ss' is a factor of p, q or r, then $a \neq a'$. Otherwise $a = a_2$ or $a = a_3$ and $a' = a_{n-1}$. This proves the claim.

Assume now that $\mathbf{y}$ is reducible. Thus $\mathbf{y}$ contains a factor tgt with $|g| \leq n - 3$. First observe that we can assume equality, i.e. that $|g| = n - 3$. Indeed, if $|g| < n - 3$, then t is not a letter, and thus, setting $t = t'a$ and $g' = ag$, with a a letter, one gets the reducible factor $t'g't'$ with a longer central word g'. The claim follows by induction on $|g|$.

We already mentioned that tgt cannot be contained in a factor of $\mathbf{y}$ which is a product of two words in S.

If tgt is contained in a factor of $\mathbf{y}$ which is a product $s_1s_2s_3$ of three words in S, then t contains an occurrence of the letter x_n, and consequently

$$s_1s_2s_3 = \gamma\beta\alpha g\beta\alpha\delta$$

with $t = \beta\alpha$, $s_1 = \gamma\beta$, $s_2 = \alpha g\beta$, $s_3 = \alpha\delta$. Note that $|\alpha\beta| = 4$. Since $s_1 \neq s_2$, one has $|\alpha| \leq 2$, whence $|\beta| \geq 2$. But we have seen that two consecutive words in S cannot have the same suffix of length 2.

If tgt is contained in a factor of $\mathbf{y}$ which is a product $s_1s_2s_3s_4$ of 4 words in S, then there are words α, β, γ, δ, α', β' such that $t = \beta\alpha$, $g = \beta'\alpha'$, and $s_1 = \gamma\beta$, $s_2 = \alpha\beta'$, $s_3 = \alpha'\beta$, $s_4 = \alpha\delta$.

<table>
<tr><td colspan="2">s_1</td><td colspan="2">s_2</td><td colspan="2">s_3</td><td colspan="2">s_4</td></tr>
<tr><td>γ</td><td>β</td><td>α</td><td>β'</td><td>α'</td><td>β</td><td>α</td><td>δ</td></tr>
<tr><td></td><td colspan="2">t</td><td colspan="2">g</td><td colspan="2">t</td><td></td></tr>
</table>

Since $n+1 = |\alpha| + |\beta'| \leq |\alpha| + n - 3$, one has $|\alpha| \geq 3$, which implies $s_2 = s_4$. But this is impossible in a word in X^*.

Thus tgt is contained in a factor of length greater that 4, and this means that $t = \beta s_1 \cdots s_m \alpha$ for words $s_1, \ldots, s_m \in S$. Let γ and δ be such that $\gamma\beta$, $\alpha\delta$ are in S. As before, there are two cases, namely either g is contained in some s, or g is overlapping over two words in S. In the first case

$$\gamma tgt\delta = \gamma\beta s_1 \cdots s_m \alpha g \beta s_1 \cdots s_m \alpha\delta$$

with $\alpha g \beta \in S$, and in the second case,

$$\gamma tgt\delta = \gamma\beta s_1 \cdots s_m \alpha\beta'\alpha'\beta s_1 \cdots s_m \alpha\delta$$

with $\alpha\beta'$, $\alpha'\beta \in S$, and $g = \beta'\alpha'$.

Consider the first case. The word $\gamma\beta s_1 \cdots s_m \alpha g \beta s_1 \cdots s_m \alpha\delta$ is a product of words in $X = \{p, q, r\}$. The word x in X in this product containing $\alpha g \beta$ does not contain two equal words in S. Thus

$$x = s_j \cdots s_m \alpha g \beta s_1 \cdots s_i$$

with $i < j$. If $i > 1$, then $s_1 \cdots s_i$ is characteristic for x, and $\gamma\beta = \alpha g \beta$. If $j < m$, then $s_j \cdots s_m$ determines x and $\alpha g \beta = \alpha\delta$. In both cases, we get a square. If $i \leq 1$ and $j \geq m$, then $x = s_m \alpha g \beta s_1$. This implies also that $m > 1$. Next, $s_m \alpha\delta$ is a suffix of a word y in X and $\gamma\beta s_1$ is a prefix of a word z in X. If $y = x$ or $z = x$, we get a square. Thus the only remaining possibility is, because x and y share the same prefix $s_m = u$, and x and z share the same suffix $s_1 = v$, that $z = r$, $x = q$, and $y = p$. However q is formed of at least 4 words in S and x contains only 3. Contradiction.

The second case is very similar. Consider the word

$$\gamma\beta s_1 \cdots s_m \alpha\beta'\alpha'\beta s_1 \cdots s_m \alpha\delta.$$

Then $n + 1 = |\alpha\beta'| \leq |\alpha| + |g| = |\alpha| + n - 3$, whence $|\alpha| \geq 4$. Thus $\alpha\beta' = \alpha\delta$. Setting $s_{m+1} = \alpha\delta$, this yields

$$\gamma\beta s_1 \cdots s_m s_{m+1} \alpha'\beta s_1 \cdots s_m s_{m+1}$$

The rest of the proof is as before. ∎

THEOREM 8.2. (Satz 33) *If every letter a_n is erased in a word $f(\mathbf{x})$ of the kind described in the previous theorem, the resulting word is irreducible over $\{a_1, \ldots, a_{n-1}\}$.*

Proof. Denote by π the projection of $\{a_1, \ldots, a_n\}^*$ onto $\{a_1, \ldots, a_{n-1}\}^*$, and let $\mathbf{y} = f(\mathbf{x})$, $\mathbf{y}' = \pi(\mathbf{y})$. Let tgt be a reducible factor of $\mathbf{y}'$. Observe first that $|tgt| \geq n-1$. Indeed, t contains a letter different from a_n, and two occurrences of this letter are separated by at least $n-3$ letters. Next, by arguing as in the preceding proof, we may assume that $|g| = n-4$. This in fact implies that $|tgt| \geq n$.

The word t contains at least one ocurrence of the letter a_{n-2}. Indeed, two consecutive occurrences of a_{n-2} in $\mathbf{y}'$ are always separated by exactly $n-1$ letters, and if the claim is wrong, then $|tgt| \leq n-1$.

Let w, ℓ and w' be words such that $w\ell w'$ is a factor of $\mathbf{y}$ and $\pi(w\ell w') = tgt$, and $\pi(w) = \pi(w') = t$, $\pi(\ell) = g$. There may be several choices for these words, and we choose w and w' of maximal length (i.e. including bordering a_n's). Since the letter a_{n-2} always occurs at the same place in words in S, namely at the fourth position from the right, the equality $\pi(w) = \pi(w')$ implies that $w = w'$. Moreover, ℓ contains at most one occurrence of the letter a_n. This proves the result. ∎

These theorems show that, as claimed above, there exist infinite irreducible words over n letters for all $n > 4$.

Chapter 4

Notes

This chapter contains several notes and comments about theorems in Thue's papers. They mainly concern further results and later developments.

4.1 Square-free morphisms

All morphisms considered are supposed nonerasing. A morphism $h : A^* \to B^*$ is *square-free* if it preserves square-free words, that is if $h(w)$ is square-free for all square-free words $w \in A^*$. As we shall see, the square-freeness of a morphism is decidable in general. Several conditions on a morphism ensure that it is square-free, and are easy to check. First, observe that one can always assume that h is injective on the alphabet, since if $h(a) = h(b)$ for $a \neq b$, then $h(ab)$ is a square.

We introduce some definitions on sets of words or codes. These have a natural extension to morphisms: a morphism $h : A^* \to B^*$ is said to have a property P if the set $h(A)$ has this property.

Let X be a set of words. A word p is a *recognizing prefix* for X (Thue says *characteristic*) if p is the prefix of one and only one word in X. Recognizing *suffixes* are defined symmetrically. As an example, a set X is a prefix code iff every $x \in X$ is a recognizing prefix for X.

A set X is a *recognizing code* (Goralčik, Vaniček) or a *ps*-code (Keränen) if, for all $x \in X$ and for every factorization $x = ps$, either p or s is recognizing. More formally, this condition can be expressed as:

$$ps, ps', p's \in X \Rightarrow p = p' \text{ or } s = s' \ .$$

As a consequence, the following fact is easily shown.

FACT. *A recognizing code is biprefix.*

A *pip* (or *recognizing factor*) for X is a word p that is a factor of exactly one word x in X and that, moreover, has only one occurrence in x. A *Melničuk code* is a set X such that every word x in X has at least one pip.

FACT. *A Melničuk code is infix.*

(A set X is *infix* if no word in X is a proper factor of another word in X.)

A word p is a *synchronizing* prefix (suffix) for X if $upv \in X^+$ implies $u \in X^*$ ($v \in X^*$). A code is *synchronizing* if, for all $x \in X$ and for every factorization $x = ps$, either p or s is synchronizing. A code X is *bissective* if it is both recognizing and synchronizing.

FACT. *A bissective code is comma-free.*

We can now state several results about morphisms that imply square-freeness. The first two are basically those of Thue. (Satz 17. Indeed, the restriction on the size of the alphabets is not relevant, see also Bean, Ehrenfeucht, McNulty.)

Let $h : A^* \to B^*$ be a morphism.

PROPOSITION 1.1. *If h is infix and preserves square-free words of length 2, then h is comma-free.*

PROPOSITION 1.2. *If h is comma-free and preserves square-free words of length 3, then h is square-free.*

An immediate corollary is:

COROLLARY 1.3. *If h is a uniform morphism (i.e. $|h(a)| = |h(b)|$ for $a, b \in A$), and if h preserves square-free words of length 3, then h is square-free.*

PROPOSITION 1.4. *If h is a bissective morphism that preserves square-free words of length 2, then h is square-free.*

This result is due to Goralčik and Vaniček. As a (negative) example, consider the morphism $g : \{a, b, c\}^* \to \{a, b, c, d\}^*$ defined by

$$g : \begin{array}{l} a \mapsto ab \\ b \mapsto cb \\ c \mapsto cd \end{array}$$

given by Brandenburg. This morphism is uniform, thus infix. It preserves square-free words of length 2. It is also easily checked to be comma-free and to be synchronizing. However, g is not recognizing since in $h(b) = cb$, neither c nor b is recognizing, and g is not square-free since $g(abc)$ contains a square.

There is a general criterion on morphisms that ensures square-freeness due to Crochemore. Define an integer $K(h)$ as follows. Set

$$M(h) = \max\{|h(a)| \mid a \in A\}, \qquad m(h) = \min\{|h(a)| \mid a \in A\}.$$

Then

$$K(h) = \max\left(3, 1 + \left\lceil \frac{M(h) - 3}{m(h)} \right\rceil\right)$$

Then one has:

THEOREM 1.5. *If h preserves square-free words of length $K(h)$, then h is square-free.*

The next two observations make it possible to build square-free morphisms over arbitrary alphabets. Examples are given in Bean, Ehrenfeucht, McNulty, Crochemore and Brandenburg (who introduced the parallel composition).

FACT. *The composition of two square-free morphisms is again square-free.*

Sometimes, the *parallel composition* $h_1 \times h_2$ of morphisms may be useful. It is defined as follows. Let $h_1 : A_1^* \to B_1^*$ and $h_2 : A_2^* \to B_2^*$ be two morphisms, where $A_1 \cap A_2 = \emptyset$. The parallel composition $h_1 \times h_2 : (A_1 \cup A_2)^* \to (B_1 \cup B_2)^*$ is defined by

$$h_1 \times h_2(a) = \begin{cases} h_1(a) & \text{if } a \in A_1 \\ h_2(a) & \text{if } a \in A_2 \end{cases}$$

FACT. *If $B_1 \cap B_2 = \emptyset$ and if h_1 and h_2 are square-free, then $h_1 \times h_2$ is square-free.*

The most difficult task is to find a square-free morphism from a four-letter alphabet into a three-letter alphabet. The example given by Bean, Ehrenfeucht, McNulty maps the letters into words of length greater than 200. Brandeburg gives (implicitly) an example of a uniform morphism of length 44. The following morphism is due to Crochemore and has length 20:

$$f : \begin{array}{l} a \mapsto abcbacabcacbabcabacb \\ b \mapsto abcbacbcabacbabcacbc \\ c \mapsto abcbacbcacbabcabacbc \\ d \mapsto abcbacbcacbacabcacbc \end{array}$$

The word *abcbac* is a synchronizing prefix for h, and the suffixes of length 14 of the four words are synchronizing suffixes. Thus h is synchronizing. Next, the prefixes of length 10 are recognizing, since they are distinct, and so are the suffixes of length 8. Thus h is bissective, and it "suffices" to check that the 12 words of length 40 obtained as images of square-free words of length 2 are square-free.

4.2 Overlap-free words

What Thue actually shows, is that a word w over the two letter alphabet $A = \{a, b\}$ is overlap-free iff $\mu(w)$ is overlap-free. Thue observes that the same result holds for circular words. More precisely, he gives a complete characterization of circular overlap-free words (Satz 13).

As a consequence of Satz 13, Thue characterizes overlap-free squares, a result that was discovered later also by [46]. T. Harju [21] gives a result which is similar, but different.

The property that the dynamical system generated by the (twosided) Thue-Morse sequence is minimal was explicitly proved by Gottschalk and Hedlund [18]. As a consequence, every factor appears with bounded gaps (is *recurrent*, in the terminology of M. Morse [29]). Axel Thue (Satz 11) only mentions that every factor appears infinitely often.

Recall (Satz 16) that Thue characterizes all overlap-free morphisms by showing basically that there is only one. This result has been completed by P. Séébold [41], who shows that the Thue-Morse word is the only *morphic* overlap-free word. Thus, the infinite words $\mathbf{t}$ and $\bar{\mathbf{t}}$ are the only infinite overlap-free words generated by iterated morphisms. There is now a simple proof of these results by Berstel and Séébold [6]. They prove that for a morphism h to be overlap-free, it suffices that $h(abbabaab)$ is overlap-free.

The structure of onesided infinite overlap-free words is more complicated. An explicit description of the tree of infinite overlap-free word by means of a finite automaton was given by E. D. Fife and deserves a mention.

Fife defines three operators on words, say α, β, γ, and he shows that every overlap-free infinite words is the "value" of some infinite word $\mathbf{f}$ in the three operators, provided the word $\mathbf{f}$ is in some rational set he gives explicitely. To be more precise, let $X_n = \{u_n, v_n\}$ be the set of Morse blocs of index n and let $X = \bigcup_{n\geq 0} X_n$. Any word $w \in A^*X_1$ admits a *canonical decomposition* $(z, y, \bar{y})$ where y is the longest word in X such that $w = zy\bar{y}$. It is equivalent to say that $(z, y, \bar{y})$ is the canonical decomposition of w if $\bar{y}y$ is not a suffix of z. As an example, the canonical decomposition of $aabaabbabaab$ is

$$(aaba, abba, baab)$$

and the decomposition of $abaabbaababbaabbabaab$ is

$$(abaab, baababba, abbabaab)\ .$$

The three functions $\alpha, \beta, \gamma : A^*X_1 \to A^*X_1$, acting on the right, are defined as follows for a word $w \in A^*X_1$ with canonical decomposition $(z, y, \bar{y})$:

$$w \cdot \alpha = zy\bar{y} \cdot \alpha = zy\bar{y}yy\bar{y} = wyy\bar{y}$$
$$w \cdot \beta = zy\bar{y} \cdot \beta = zy\bar{y}y\bar{y}\bar{y}y = wy\bar{y}\bar{y}y$$
$$w \cdot \gamma = zy\bar{y} \cdot \gamma = zy\bar{y}\bar{y}y = w\bar{y}y$$

Since w is a prefix of $w \cdot \alpha$, $w \cdot \beta$, and of $w \cdot \gamma$, it makes sense to define $w \cdot f$ by induction for all "words" f in B^*, with $B = \{\alpha, \beta, \gamma\}$. By continuity, $w \cdot \mathbf{f}$ is defined also for infinite words $\mathbf{f}$. Here are some examples:

$$
\begin{aligned}
ab \cdot \alpha &= abaab \\
ab \cdot \beta &= ababba \\
ab \cdot \gamma &= abba \\
ab \cdot \gamma^\omega &= \mathbf{t} \\
aab \cdot \alpha &= aabaab = a(ab \cdot \alpha) \\
ab \cdot \alpha\beta\gamma &= abaababbabaababbaabbabaab
\end{aligned}
$$

Observe that the last word contains an overlap. Note also that, for $w \in A^*X_1$ and $f \in B^*$, one has $\mu(w \cdot f) = \mu(w) \cdot f = w \cdot \gamma f$. A *description* of an infinite word $\mathbf{x}$ starting with ab or aab is an infinite word $\mathbf{f}$ over B such that $\mathbf{x} = ab \cdot \mathbf{f}$ or $\mathbf{x} = aab \cdot \mathbf{f}$, according to $\mathbf{x}$ starts with ab or aab.

Proposition 2.1. *Every infinite overlap-free word starting with the letter a admits a unique description.*

Let

$$F = B^\omega - B^* I B^\omega$$

be the (rational) set of infinite words over B having no factor in the set

$$I = \{\alpha, \beta\}(\gamma^2)^*\{\beta\alpha, \gamma\beta, \alpha\gamma\}$$

and let G be the set of words $\mathbf{f}$ such that $\beta\mathbf{f}$ is in F. Then:

Theorem 2.2. (Fife's Theorem) *Let $\mathbf{x}$ be an infinite word over $A = \{a, b\}$.*

(*i*) *If $\mathbf{x}$ starts with ab, then $\mathbf{x}$ is overlap-free iff its description is in F;*

(*ii*) *If $\mathbf{x}$ starts with aab, then $\mathbf{x}$ is overlap-free iff its description is in G.*

A direct consequence is the following:

Corollary 2.3. *An overlap-free word w is the prefix of an infinite overlap-free word iff w is a prefix of a word $ab \cdot f$ with $f \in W$ or of a word $aab \cdot f$ with $\beta f \in W$, where $W = B^* - B^* I B^*$.*

This implies in particular a result of Restivo et Salemi [34], namely that it is decidable whether an overlap-free word is extensible into an infinite overlap-free word. Another consequence of Fife's description is the following corollary which can also be proved directly:

COROLLARY 2.4. *The Thue-Morse word* $\mathbf{t}$ *is the greatest infinite overlap-free word, in lexicographical order, that starts with the letter* a.

Indeed, the choice of the letters α, β, et γ implies that if $\mathbf{f} \leq \mathbf{f}'$, then $ab{\cdot}\mathbf{f} \leq ab{\cdot}\mathbf{f}'$. The greatest word in F is γ^ω, and this shows the corollary. A. Carpi [10] has developed a description for finite overlap-free words by means of a finite automaton. Unfortunately, his automaton is rather big (more than 300 states). J. Cassaigne [12], using a similar but different encoding, gets a much smaller automaton.

Since overlap-free words have a strong structure, it seems natural to count them. The first result is due to Restivo and Salemi [34]. They prove that the number γ_n of overlap-free words over two letters grows polynomially in n (in fact slower than n^4). Kobayashi [25] has used Fife's theorem to derive the lower of the more precise bounds for γ_n :

THEOREM 2.5. *There are constants* C_1 *and* C_2 *such that*

$$C_1 n^\alpha < \gamma_n < C_2 n^\beta$$

where $\alpha = 1.155\ldots$ *and* $\beta = 1.5866\ldots$.

One might ask what is the "real" limit. In fact, a recent and surprising result by J. Cassaigne [12] shows that there is no limit. More precisely, he gets exact formulas for the number of overlap-free words, and setting

$$\alpha' = \sup\{r \mid \exists C > 0, \forall n, \gamma_n \geq C n^r\}$$

and

$$\beta' = \sup\{r \mid \exists C > 0, \forall n, \gamma_n \leq C n^r\}$$

he obtains:

THEOREM 2.6. *One has* $1.155 < \alpha' < 1.276 < 1.332 < \beta' < 1.587$.

This is to be compared with the situation for square-free words. Indeed, Brandenburg [7] proved that for the number $c(n)$ of square-free words of length n over three letters, there are constants $c_1 \geq 1.032$ and $c_2 \leq 1.38$ such that $6c_1^n < c(n) < 6c_2^n$. Brandenburg also proves that the number of cube-free words over two letters grows exponentially.

4.3 Avoidable patterns

The overlap-freeness of the Thue-Morse sequence, and the square-freeness of the other words we have presented can be expressed in the more general framework of avoidable and unavoidable patterns in strings. This concept has been introduced in the context of equations defining algebras. Certain unavoidable words have been used e.g. in [39] to characterize those finite semigroups S that are inherently nonfinitely based, in the sense that S is not a member of any locally finite semigroup variety definable by finitely many equations. It may be noticed that Axel Thue places his research on repetitions in strings in an even slightly more general context, since he considers avoiding patterns with constants. However, he has not stated results in this specific framework.

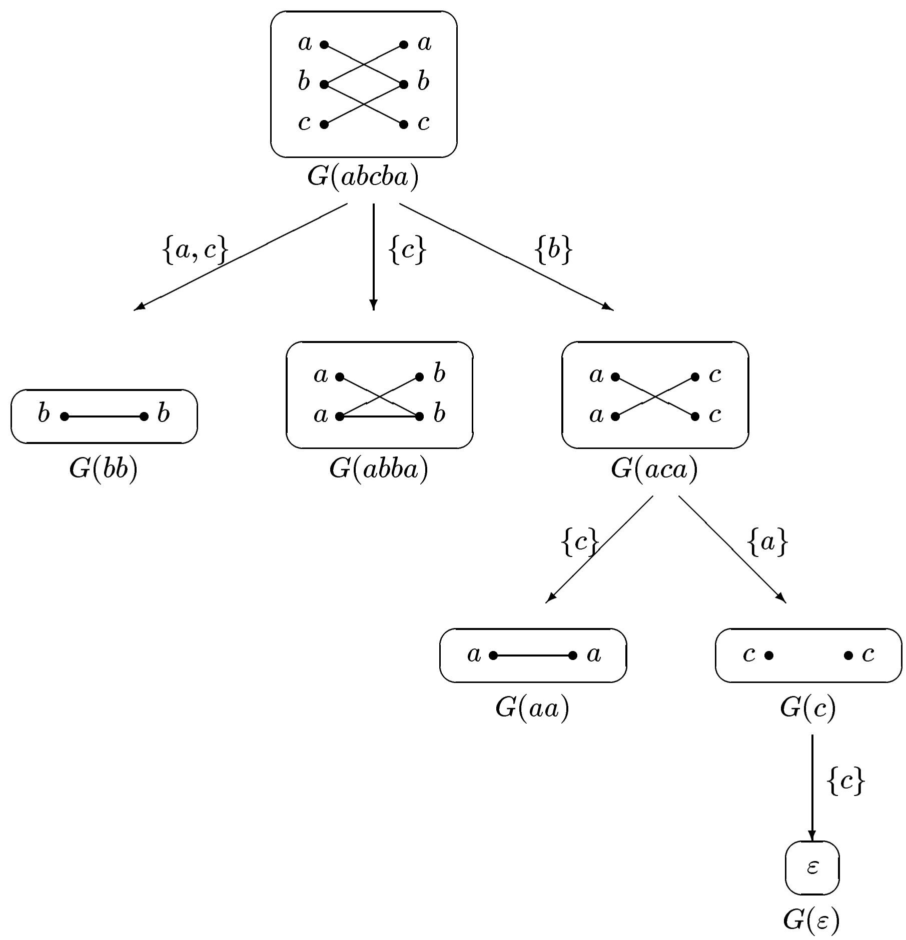
A word u is said to *appear* in a word v if there is a nonerasing morphism h such that $h(u)$ is a factor of v. Clearly, if u appears in v and if v appears in w, then u appears in w. Thus, the relation of *appearance* is a quasi-order, and it is an order if words are considered to be equal if they are the same up to a renaming of letters.

Consider an alphabet E of "pattern symbols". A word e over E is called a pattern. A pattern e is *avoidable* over k letters, or is k-avoidable, if there is an infinite word $\mathbf{x}$ over k letters such that e does not appear in $\mathbf{x}$. The Thue-Morse sequence shows that the patterns aaa and $ababa$ are (simultaneously) 2-avoidable, and square-free infinite words show that aa is 3-avoidable (but not 2-avoidable). If u appears in v and if v is unavoidable, then u is unavoidable or, equivalently, if v is avoidable, then u is avoidable. Avoidable and unavoidable patterns have been studied by several people (Zimin [52], Schmidt [40], Bean, Ehrenfeucht, McNulty [5], Roth [35], Cassaigne [11], Goralcik, Vanicek [19], Baker, McNulty, Taylor [3], Crochemore, Goralcik [15]).

A first problem is to determine whether a given pattern is avoidable. There is a nice algorithm in [5], and basically the same in [52], to decide whether a pattern is avoidable. It works as follows.

Let w be a word for which one has to decide if it is avoidable, and let $A = \mathrm{alph}(w)$. One constructs a bipartite graph $G(w)$ whose vertex set is $A_G \cup A_D$, where A_G and A_D are disjoint sets labelled with the letters in A. There is an edge from a_G to b_D iff ab is a factor of w.

EXAMPLE. For $w = abcba$, the graph $G(w)$ is given below.



A subset B of A is called *free* for w if no connected component of $G(w)$ contains both a letter of B_G and a letter of B_D. In our example, the free subsets are $\{a\}$, $\{b\}$, $\{c\}$ and $\{a, c\}$.

With these definitions, we are able to define a reduction relation as follows: $w \to w'$ iff there exists a free subset B such that $w' = \mathrm{era}_B(w)$, where era_B is the morphism that erases all letters in B and is the identity on the other letters. The following result is due to [52], and Baker, McNulty, Taylor [3]. It is contained in a slightly different form in Bean, Ehrenfeucht, McNulty [5].

THEOREM 3.1. *A word w is unavoidable iff $w \to^* \varepsilon$.*

The complexity of this algorithm is at least exponential. P. Roth (personal communication) recently has proved that the general problem is NP-complete.

There are several easy consequences of this characterization. Call a letter a in w an *isolated* letter if $|w|_a = 1$, i.e. if it occurs only once in w.

COROLLARY 3.2. *If w contains no isolated letter, then w is avoidable.*

Indeed, if $w \to w'$ and if w' contains an isolated letter, then w contains an isolated letter.

COROLLARY 3.3. *Every word w of length $|w| \geq 2^n$ over an n-letter alphabet is avoidable.*

Indeed, it is not very difficult to show that such a word contains a factor without an isolated letter. This bound is the best possible, because there exist unavoidable words of length $2^n - 1$ over an n-letter alphabet. This can be formulated as follows. Let $Z = \{z_1, z_2, \ldots, z_n, \ldots\}$ be a countable infinite alphabet, and define the *Zimin* words Z_n by

$$Z_1 = z_1, \qquad Z_n = Z_{n_1} z_n z_{n-1}, \quad n > 1 \ .$$

Thus $Z_4 = z_1z_2z_1z_3z_1z_2z_1z_4z_1z_2z_1z_3z_1z_2z_1$. Then

PROPOSITION 3.4. *For every $n \geq 1$, the Zimin word Z_n is unavoidable. Moreover, if w is an unavoidable pattern over an n-letter alphabet, then w appears in Z_n.*

The first part of the proposition has been proved by Coudrain, Schützenberger (see also Lothaire). Define a *biideal sequence* to be a sequence $(w_n)_{n\geq 1}$ of words such that w_1 is nonempty and, for all $n > 1$, $w_{n+1} = w_n v_n w_n$ for some nonempty word v_n. Then Coudrain and Schützenberger state that for any fixed n, every long enough word contains an element w_n of some biideal sequence.

For an avoidable pattern e, denote by $\mu(e)$ the smallest integer k such that e is k-avoidable. We have seen that $\mu(aa) = 3$. The first word that is 4-avoidable but not 3-avoidable has been given by [3]. It has the form $ab\alpha bc\beta ca\gamma ba\delta ac$. It is not known if, for every n, there exists a pattern that is $n+1$ avoidable but not n-avoidable. Upper bounds for μ, as a function of α are also given in [3]. Recently, Roth [35], Cassaigne [11], Goralcik, Vanicek [19] have solved the problem of determining all the 2-avoidable binary patterns. There is an unpublished result by Melničuk that states that $\mu(e) \leq \mathrm{alph}(e) + 4$.

Bibliography

[1] A. ADLER, S. LI, Magic cubes and Prouhet sequences, *American Math. Monthly* **84** (1977), 618–627.

[2] J.-P. ALLOUCHE, Automates finis en théorie des nombres, *Exposition. Math.* **5** (1987), 239–266.

[3] K. A. BAKER, G. F. MCNULTY, W. TAYLOR, Growth problems for avoidable words, *Theoret. Comput. Sci.* **69** (1989), 319–345.

[4] L. BAUM, M. SWEET, Continued fractions of algebraic power series in characteristic 2, *Ann. Math.* **103** (1976), 593–610.

[5] D. R. BEAN, A. EHRENFEUCHT, G. F. MCNULTY, Avoidable patterns in strings of symbols, *Pacific J. of Math.* **85** (1979), 261–294.

[6] J. BERSTEL, P. SÉÉBOLD, A characterization of overlap-free morphisms, *Discr. Appl. Math.* **46** (1993), 275–281.

[7] F. J. BRANDENBURG, Uniformely growing k-th powerfree homomorphisms, *Theoret. Comput. Sci.* **23** (1983), 69–82.

[8] S. BRLEK, Enumeration of factors in the Thue-Morse word, *Discr. Appl. Math.* **24**, (1989), 83–96.

[9] A. CARPI, On the size of a squarefree morphism on a three letter alphabet, *Inform. Proc. Letters* **16** (1983), 231–236.

[10] A. CARPI, Overlap-free words and finite automata, manuscrit, 1990.

[11] J. CASSAIGNE, Unavoidable binary patterns, *Acta Informatica*, to appear.

[12] J. CASSAIGNE, Counting overlap-free binary words, in: Stacs'93, *Lect. Notes Comput. Sci.* **665**, Springer Verlag, 1993, 216–225.

[13] G. CHRISTOL, T. KAMAE, M. MENDÈS FRANCE, G. RAUZY, Suites algébriques, automates et substitutions, *Bull. Soc. Math. France* **108** (1980), 401–419.

[14] A. Cobham, Uniform tag sequences, *Math. Systems Theory* **6** (1972), 164–192.

[15] M. Crochemore, P. Goralcik, Mutually avoiding ternary words of small exponents, *Intern. J. Algebra Comput.***1** (1991), 407–410.

[16] F. Dejean, Sur un théorème de Thue, *J. Combin. Th. A* **13** (1972), 90–99.

[17] E. D. Fife, Binary sequences which contain no BBb, *Trans. Amer. Math. Soc.* **261** (1980), 115–136.

[18] W.H. Gottschalk, G.A. Hedlund, A characterization of the Morse minimal set, *Proc. Amer. Math. Soc.* **15** (1964), 70–74.

[19] P. Goralcik, T. Vanicek, Binary patterns in binary words, *Intern. J. Algebra Comput.***1** (1991), 387–391.

[20] M. Hall, Generators and relations in groups – the Burnside problem, in T. L. Saaty (ed) *Lectures on Modern Mathematics* **2**, Wiley, 1964, 42–92.

[21] T. Harju, On cyclically overlap-free words in binary alphabets, *The Book of L*, Springer-Verlag, 1986, 123–130.

[22] G.A. Hedlund, Remarks on the work of Axel Thue, *Nordisk Mat. Tidskr.* **15** (1967), 148–150.

[23] V. Keränen, On k-repetition free words generated by length uniform morphisms over a binary alphabet, *Lect. Notes Comp. Sci.* **194**, 1985, 338–347.

[24] V. Keränen, On the k-freness of morphisms on free monoids, *Ann. Acad. Sci. Fennicae* **61**, 1986.

[25] Y. Kobayashi, Enumeration of irreductible binary words, *Discrete Appl. Math.***20** (1988), 221–232.

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

[27] J. H. Loxton, A. J. van der Poorten, Arithmetic properties of the solutions of a class of functional equations, *J. reine angew. Math.* **330** (1982), 159–172.

[28] A. A. Markov, Imossibility of certain algorithms in the theory of associative systems *Dokl. Akad. Nauk. SSSR* **55** (1941), 587–590.

[29] M. Morse, Recurrent geodesics on a surface of negative curvature, *Transactions Amer. Math. Soc.* **22** (1921), 84–100.

[30] J. Moulin-Ollagnier, Preuve de la conjecture de Dejean pour des alphabets à 5, 6, 7, 8, 9 lettres, Prépublication du département de mathématique et informatique, Université Paris-Nord, Nr. 89–4, 1989.

[31] J.-J. PANSIOT, A propos d'une conjecture de F. Dejean sur les répétitions dans les mots, *Discrete Appl. Math.***7** (1984), 297–311.

[32] E. L. POST, Recursive unsolvability of a problem of Thue, *J. Symbolic Logic* **11** (1947), 1–11.

[33] M. E. PROUHET, Mémoire sur quelques relations entre les puissances des nombres, *C. R. Acad. Sci. Paris.* **33** (1851), 31.

[34] A. RESTIVO, S. SALEMI, Overlap-free words on two symbols, in: *Automata on infinite words*, Nivat, Perrin (eds), Lect. Notes Comp. Sci.,**192**, Springer-Verlag, 1985, 198–206.

[35] P. ROTH, Every binary pattern of length six is avoidable on the two-letter alphabet, *Acta Informatica* (1992).

[36] G. ROZENBERG, A. SALOMAA, *The Mathematical Theory of L-Systems*, Academic Press, 1980.

[37] W. RUDIN, Some theorems on Fourier coefficients, *Proc. Amer. Math. Soc.* **10** (1959), 855–859.

[38] A. SALOMAA, *Jewels of Formal Language Theory*, Computer Science Press, 1981.

[39] M. SAPIR, Inherently nonfinitely based finite semigroups, *Mat. Sb.* **133** (1987), 154–166.

[40] U. SCHMIDT, Avoidable patterns on two letters, *Theoret. Comput. Sci.* **63** (1989), 1–17.

[41] P. SÉÉBOLD, Sequences generated by infinitely iterated morphisms, *Discrete Appl. Math.***11**, (1985), 255–264.

[42] H. S. SHAPIRO, Extremal problems for polynomials and power series, Thesis, M.I.T., 1951.

[43] R. SHELTON, Aperiodic words on three symbols I, *J. Reine Angew. Math.* **321** (1981), 195–209.

[44] R. SHELTON, Aperiodic words on three symbols II, *J. Reine Angew. Math.* **327** (1981), 1–11.

[45] R. SHELTON, R. SONI, Aperiodic words on three symbols III, *J. Reine Angew. Math.* **330** (1982), 44–52.

[46] R. SHELTON, R. SONI, Chains and fixing blocks in irreducible sequences, *Discrete Math.***54** (1985), 93–99.

[47] A. THUE, Über unendliche Zeichenreihen, *Kra. Vidensk. Selsk. Skrifter. I. Mat.-Nat. Kl.*, Christiana 1906, Nr. 7.

[48] A. THUE, Die Lösung eines Spezialfalles eines generellen logischen Problems, *Kra. Vidensk. Selsk. Skrifter. I. Mat.-Nat. Kl.*, Christiana 1910, Nr. 8.

[49] A. THUE, Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen, *Kra. Vidensk. Selsk. Skrifter. I. Mat.-Nat. Kl.*, Christiana 1912, Nr. 10.

[50] A. THUE, Probleme über Veränderungen von Zeichenreihen nach gegebenen Regeln, *Kra. Vidensk. Selsk. Skrifter. I. Mat.-Nat. Kl.*, Christiana 1914, Nr. 7.

[51] A. THUE, *Selected Mathematical Papers*, edited by T. Nagell, A. Selberg, S. Selberg, K. Thalberg, Universitetsforlaget, Oslo 1977.

[52] A. I. ZIMIN, Blocking sets of terms, *Math. USSR Sb.* **47**, (1984), 353–364.

Index