Theoretical Computer Science 95 (1992) 187–205
Elsevier

Proof of Dejean's conjecture for alphabets with 5, 6, 7, 8, 9, 10 and 11 letters

Jean Moulin Ollagnier

Département de Mathématiques et Informatique, Université Paris-Nord, Avenue J.B. Clément, F 93430 Villetaneuse, France

Communicated by D. Perrin
Received September 1989
Revised March 1990

Abstract

Moulin Ollagnier, J., Proof of Dejean's conjecture for alphabets with 5, 6, 7, 8, 9, 10 and 11 letters, Theoretical Computer Science 95 (1992) 187–205.

Axel Thue proved that overlapping factors could be avoided in arbitrarily long words on a two-letter alphabet while, on the same alphabet, square factors always occur in words longer than 3. Françoise Dejean stated an analogous result for three-letter alphabets: every long enough word has a factor, which is a fractional power with an exponent at least 7/4 and there exist arbitrary long words in which no factor is a fractional power with an exponent strictly greater than 7/4. The number 7/4 is called the repetition threshold of the three-letter alphabets.

Thereafter, she proposed the following conjecture: the repetition threshold of the k-letter alphabets is equal to $k/(k-1)$ except in the particular cases $k=3$, where this threshold is 7/4, and $k=4$, where it is 7/5.

For $k=4$, this conjecture was proved by J.J. Pansiot (1984).

In this paper, we give a computer-aided proof of Dejean's conjecture for several other values: 5, 6, 7, 8, 9, 10 and 11.

Résumé

Moulin Ollagnier, J., Proof of Dejean's conjecture for alphabets with 5, 6, 7, 8, 9, 10 and 11 letters, Theoretical Computer Science 95 (1992) 187–205.

Axel Thue a montré que l'on peut éviter les chevauchements dans des mots arbitrairement grands sur un alphabet à deux lettres alors que l'on ne peut éviter les carrés. On peut ainsi dire que l'exposant 2 est le seuil de répétition des alphabets à deux lettres. Françoise Dejean a démontré que le seuil de répétition des alphabets à trois lettres est 7/4; elle a alors proposé la conjecture suivante: le seuil s_k de répétition des alphabets à k lettres est égal à $k/(k-1)$ sauf pour les cas particuliers 3 et 4 ($s_3=7/4$, $s_4=7/5$).

Jean Jacques Pansiot a établi le résultat pour $k=4$.

Nous donnons ici une démonstration assistée par ordinateur de cette conjecture pour $k=5, 6, 7, 8, 9, 10$ et 11.

0. Introduction

Consider the alphabet $A=\{0, 1\}$. There are only finitely many square-free words on this alphabet, i.e. words no factor of which has the form uu, where u is not the empty word ε.

In order to prove this result, it suffices to enumerate all possible words in alphabetical order and to see that they constitute a finite set

$$\{\varepsilon, 0, 01, 010, 1, 10, 101\}.$$

Consider now the infinite word of Thue–Morse, which is generated by the morphism that gives to 0 the image 01 and to 1 the image 10; the finite factors of this word avoid repetitions with a strictly greater exponent. If the word uu ($u\neq\varepsilon$) is a factor of the infinite word of Morse–Thue and if x is the first letter of u, then uux is not a factor of the infinite word (results by Axel Thue [5, 6]; see, for instance, [2]).

Then, with a two-letter alphabet, square words (exponent-2 repetitions) are unavoidable as factors of long enough words, while there exist arbitrarily long words in which no overlapping factor occurs (a repetition with an exponent strictly greater than 2); the exponent 2 is called the repetition threshold of the two-letter alphabets.

We now give some definitions in order to state Dejean's conjecture.

Definition 0.1. Given a k-letter finite alphabet A, a *repetition* is a pair (p, e) of words in A^*, where p is not the empty word ε, and such that e is a left factor of pe: $w=pe=ep'$.

Definition 0.2. The *exponent* of the repetition (p, e) is the ratio $|pe|/|p|$ of the lengths of $w=pe$ and p; it could be said that w is a power word which consists of "$|w|/|p|$" consecutive copies of the word p.

The word p is said to be the *period* of the repetition and e its *excess*. A repetition (p, e) is said to *occur in* a word u if the word $w=pe$ is a factor of u.

Call s the following sequence of rational numbers, defined for $k>1$:

$$s_2=2;\quad s_3=7/4;\quad s_4=7/5;\quad k>4 \Rightarrow s_k=k/(k-1).$$

Françoise Dejean proposed the following conjecture: number s_k is the repetition threshold of the k-letter alphabet; this first means that, in every long enough word on such an alphabet, some repetition occurs, whose exponent is at least equal to s_k; that also means that there exist arbitrarily long words such that the repetitions occurring in them have an exponent at most s_k.

The established results in this direction are the following:

(i) Case $k=2$, previously quoted, was solved by Thue [5, 6].

(ii) The solution for $k=3$ was given by Dejean, who thereafter proposed the conjecture [1].

(iii) Pansiot [4] proved the result for $k=4$.

(iv) For every $k>4$, the first part of the conjecture can be easily settled: every word of length $k+2$ on a k-letter alphabet has a factor, which is a repetition with an exponent greater than or equal to $k/(k-1)$.

In the present paper, we propose a proof of the conjecture for several other particular values of k: 5, 6, 7, 8, 9, 10 and 11. A coding method due to Pansiot is used, which enables us to represent some words on k-letter alphabets by words on the alphabet $\{0, 1\}$; some morphisms on $\{0, 1\}^*$ then yield infinite sets of words. The proof afterwards reduces to finitely many verifications for which the help of a computer is necessary.

Let us begin by describing more precisely the previously established results.

1. Known results

1.1. Dejean's morphism

In order to show that long enough words on $A=\{a, b, c\}$ cannot avoid repetitions with an exponent 7/4, it suffices to enumerate, in lexicographical order, all words on this 3-letter alphabet without repetition with an exponent greater than or equal to 7/4; it appears that the tree of all these words is finite.

While, for $k=2$, reaching the length 4 was enough to see that squares were unavoidable, we must here examine all words until length 39: there exist some words of length 38 on the alphabet A no factor of which is a repetition with an exponent greater than or equal to 7/4, but no word of length 39 has this property.

To construct infinitely many words on A avoiding repetitions with an exponent strictly greater than 7/4, Dejean uses, as Thue did, a morphism.
Dejean's morphism m on $\{a, b, c\}^*$ is defined by

$$m(a)=abcacbcabcbacbcacba,$$

$$m(b)=bcabacabcacbacabacb,$$

$$m(c)=cabcbabcabacbabcbac.$$

Morphism m preserves the set of words without repetition with an exponent strictly greater than 7/4, which means that, if the word w has this property, so has its image by m; whence the result.

1.2. Pansiot's construction

Dejean also used an enumeration to show that the exponent 7/5 was unavoidable for repetitions in long enough words on a 4-letter alphabet.

A much deeper search is then necessary: all words of length 122 have at least one factor, which is a repetition with an exponent greater than or equal to 7/5 and some words of length 121 avoid such repetitions.

In order to establish that 7/5 is the repetition threshold of 4-letter alphabets, Pansiot uses a construction more involved that a mere morphism on $\{a, b, c\}^*$: some words on the 4-letter alphabet are coded by words on the two-letter alphabet $\{0, 1\}$ and a morphism is build on $\{0, 1\}^*$.

We shall give further a precise description of his coding method, that we used to settle the result for $k=5, 6, 7, 8, 9, 10$ and 11.

1.3. The first part for $k>4$

Let us state the easy first part of the conjecture as a proposition.

Proposition 1.1. *Let k be an integer greater than 4. Every word on a k-letter alphabet whose length is $k+2$ has a factor, which is a repetition with an exponent at least $k/(k-1)$.*

Proof. Indeed, a word w of length $k+2$ on such an alphabet $A=\{1,\ldots,k\}$ has three factors of length k.

If one of these three factors is not built with k different symbols, a repetition with an exponent greater than or equal to $k/(k-1)$ then occurs. Otherwise, within a permutation of the symbols, w is the word $12\ldots k12$; in this last case, repetition $(12\ldots k, 12)$ occurs in w and its exponent is $(k+2)/k$, which is greater than or equal to $k/(k-1)$. □

1.4. Pansiot's coding method

We are looking for arbitrarily long words in A^* no factor of which is a repetition with an exponent greater than s_k. Such words have an interesting property which enabled Pansiot [4] to code them with words on the 2-letter alphabet; we describe this property in the following lemma.

Lemma 1.2. *Let w be a word on a k-letter alphabet, no factor of which is a repetition with an exponent strickly greater than s_k. Then, all its factors of length $k-1$ consist of $k-1$ different symbols.*

Proof. Otherwise, it would appear a factor starting and ending with the same letter, with a length less than or equal to $k-1$; this factor would then correspond to a repetition with an exponent greater than or equal to $(k-1)/(k-2)$. This last number is always strictly greater than s_k (even for $k=3$ and 4; for $k=2$, the remark is trivial). □

Now, if any $k-1$ successive letters in a word w are different, there are only two ways to choose the letter at a given place in w when the $k-1$ preceding ones are known:

- either this new letter is the first of the $k-1$ preceding ones (the furthest),

– or it is the unique element of the alphabet which does not appear among the $k-1$ preceding letters.

The transition is coded by 0 in the first case and by 1 in the second case. Thus, "good" words with a length l greater than or equal to $k-1$ on the k-letter alphabet are described by their prefix of length $k-1$ and by a word on $\{0, 1\}$, whose length is $l-k+1$, which is the code of the transitions.

For instance, word $m(a)$ of Dejean's morphism is completely determined by its prefix ab and by the word 11010111011101011 which codes successive transitions:

$$m(a)=\boldsymbol{ab}_1 c_1 a_0 c_1 b_0 c_1 a_1 b_1 c_0 b_1 a_1 c_1 b_0 c_1 a_0 c_1 b_1 a.$$

Let us remark that a modification of the prefix without modification of the coding word corresponds to a permutation of the elements of the alphabet; that does not modify the exponent of the repetitions occurring in the coded word.

And the problem that we deal with is in fact a problem on the free monoid $\{0, 1\}^*$; this problem of course depends on the cardinal number k of the first alphabet.

2. Infinite languages on $\{0, 1\}$

2.1. The morphism σ_k from $\{0, 1\}^*$ in the group $\mathscr{S}_k$

The following construction yields a one-to-one mapping between the set of all words on the alphabet $A=\{1,\ldots,k\}$ written with $k-1$ different letters and the group $\mathscr{S}_k$ of all permutations of the set $\{1,\ldots,k\}$.

Given the word $w=a_1 a_2 \ldots a_{k-1}$, we consider the following permutation

$$\begin{pmatrix} 1 & 2 & \ldots & k-1 & k \\ a_1 & a_2 & \ldots & a_{k-1} & ? \end{pmatrix},$$

where "?" stands for the only letter of the alphabet A that does not appear in w.

With this one-to-one mapping, Pansiot's coding method corresponds to an action on the symmetric group because a transition (coded by 0 or by 1) jumps from a word in A^* consisting of $k-1$ different letters to another such word. It is rather easy to verify that this action is the following one:

– the transition, coded 0, corresponds to the right multiplication in the symmetric group $\mathscr{S}_k$ by the rotation σ_0 on the $k-1$ first elements of $\{1,\ldots,k\}$:

$$\sigma_0=\begin{pmatrix} 1 & 2 & \ldots & k-2 & k-1 & k \\ 2 & 3 & \ldots & k-1 & 1 & k \end{pmatrix},$$

– the transition, coded 1, corresponds to the right multiplication by the rotation σ_1 on the k elements of $\{1,\ldots,k\}$:

$$\sigma_1=\begin{pmatrix} 1 & 2 & \ldots & k-2 & k-1 & k \\ 2 & 3 & \ldots & k-1 & k & 1 \end{pmatrix}.$$

We thus define a monoid morphism σ_k from the free monoid $\{0, 1\}^*$ to the symmetric group $\mathcal{S}_k$; images σ_0 and σ_1 generate $\mathcal{S}_k$ so that the morphism is onto.

2.2. *Sufficient conditions*

Let us now resume Dejean's problem and express it with Pansiot's coding method, that we have just described.

Given an integer $k > 1$ and the presumed corresponding repetition threshold s_k, we look for a language $\mathcal{L}$, subset of $\{0, 1\}^*$, with the three following properties:

- if w belongs to $\mathcal{L}$ and if u is a factor of w, u also belongs to $\mathcal{L}$ (Condition FS: stability under factor taking);
- if w belongs to $\mathcal{L}$, it can be extended on the right in $\mathcal{L}$, which means that one of the two words $w0$ or $w1$ belongs to $\mathcal{L}$. If w belongs to $\mathcal{L}$, it can be extended on the left in $\mathcal{L}$, which means that one of the two words $0w$ or $1w$ belongs to $\mathcal{L}$. (Condition E: words can be extended in the language);
- if w belongs to $\mathcal{L}$, there is no repetition (p, e) on the k-letter alphabet with an exponent strictly greater than s_k such that the word pe has w for its code (Condition I: interdiction of repetitions with a too large exponent).

Proposition 2.1. *To prove the second part of Dejean's conjecture for a given integer k consists in finding a nonempty language $\mathcal{L}$ for which the three previous properties (FS), (E), (I) hold.*

Proof. Such a language $\mathcal{L}$ is then infinite and no word on the k-letter alphabet coded by an arbitrary initial factor (with pairwise different letters of course) and a word in $\mathcal{L}$ for the transitions contains a repetition with an exponent strictly greater than s_k. □

Condition (I) is of course the most difficult to get and we shall give sufficient conditions therefore. We must before distinguish repetitions with respect to the length of their excess.

2.3. *Short and kernel repetitions*

A repetition (p, e) on the alphabet $\{1, \ldots, k\}$ is called a *short repetition* if the length $|e|$ of its excess e is less than $k-1$. If the length of e is greater than or equal to $k-1$, denote by e' the left factor of length $k-1$ of e: $e = e'e''$.

Word e' is also the left factor of length $k-1$ of pe; $w = pe$ is then coded by its initial factor e', and by a transition word W of $\{0, 1\}^*$. Coding word W is the concatenation of P and E, where e' and P together code word pe', while e' and E together code the excess e of the repetition. The transition coded by P jumps from e' to e', that is to say that its image in the symmetric group is the identity of $\mathcal{S}_k$; word P belongs to the kernel of the previously defined morphism σ_k from $\{0, 1\}^*$ onto $\mathcal{S}_k$.

A repetition (P, E) on the alphabet $\{0, 1\}$, whose period belongs to the kernel of σ_k is then called a *kernel repetition.*

We shall no longer deal with the k-letter alphabet and we can use lowercase letters for words on $\{0, 1\}$.

The exponent of a kernel repetition (p, e) is the exponent of a repetition on $\{1, \ldots, k\}$ coded by an arbitrary word consisting in $k-1$ pairwise different letters together with the elements of the pair (p, e) as transitions.

The value $\rho((p, e))$ of this exponent is

$$\rho((p, e)) = (|p| + |e| + k - 1)/|p|.$$

Let us give some examples (with $k=3$). Repetition (abc, a) is short. Repetition (abc, ab) is not short; it is coded by the initial word ab and by the kernel repetition $(111, \varepsilon)$; the exponent of the kernel repetition $(111, \varepsilon)$ is $(3+0+2)/3$.

Condition (I), to be satisfied by languages in order to yield a proof of the conjecture, divides into two conditions (IS) and (IK).

- if (p, e) is a short repetition with an exponent greater than s_k, the word of $\{0, 1\}^*$ coding the transitions of $w = pe$ does not belong to $\mathscr{L}$ (Condition IS);
- if (p, e) is a kernel repetition with an exponent greater than s_k, word pe does not belong to $\mathscr{L}$ (Condition IK).

To fulfil condition (IS), it suffices to exclude a finite number of forbidden words from the language; a word is forbidden if it codes the transitions of $w = pe$, where (p, e) is a short repetition with an exponent greater than s_k. Given integer k, a mere enumeration gives the finite list of all these words.

Some sufficient conditions will imply condition (IK) for some morphism-generated languages; let us now describe the notion of a *free word*, which is related to languages only satisfying conditions (FS) and (E) (factor stability and extensibility).

2.4. *Free words*

Let the language $\mathscr{L}$ on $\{0, 1\}$ satisfy conditions (FS) and (E). A word w of $\mathscr{L}$ is said to be *right-determined* if only one of the two words $w0$ and $w1$ belongs to $\mathscr{L}$: there is a unique way to give a right extension to w in $\mathscr{L}$ with one more letter. On the contrary, if both $w0$ and $w1$ belong to $\mathscr{L}$, w is said to be *free on its right.* Corresponding definitions hold on the left. If a word is free on both sides, then it is simply said to be *free.*

In order to use free words in the reduction of the forthcoming proofs, we shall demand that language $\mathscr{L}$ satisfies condition (FF) meaning that necessary extensions of a word w of $\mathscr{L}$ (on the left and on the right) eventually end:

- every word of $\mathscr{L}$ is a factor of a free word (Condition FF).

Especially with condition (FF), to get the least upper bound of the exponents of all kernel repetitions in $\mathscr{L}$, it suffices to take into account the kernel repetitions, whose excess is a free word, as it will be shown later.

2.5. Free kernel repetitions

Let the language $\mathscr{L}$ satisfy conditions (FS), (E), (IS) and (FF). Let (p, e) be a kernel repetition such that $w = pe$ belongs to $\mathscr{L}$; if its excess e is not free in $\mathscr{L}$, it is possible to enlarge repetition (p, e) into another one with a strictly greater exponent.

Indeed, if the excess e is right-determined, this word is always followed by the same letter, say x; word $wx = pex = ep'x = exp''$ belongs to $\mathscr{L}$ and corresponds to the kernel repetition (p, ex), whose exponent is greater than the one of (p, e). If the excess e is left-determined, it is always preceded by the same letter x and $xw = xpe = xep' = xqxe = p''xe$ belongs to $\mathscr{L}$.

As $p''x = xp$, word p'', which is a conjugate of p, also belongs to the kernel of morphism σ; the exponent of the kernel repetition (p'', xe) is the greater than the exponent of (p, e). If condition (FF) holds for $\mathscr{L}$, we only need a control on the exponent of the kernel repetitions, whose excess is a free word in order to establish property (IK), as follows from the next lemma.

Lemma 2.2. *Let the language $\mathscr{L}$ satisfy condition (FF); and let (p, e) be some kernel repetition such that $w = pe$ belongs to $\mathscr{L}$. If e is not a free word, there then occurs in $\mathscr{L}$ another kernel repetition (p', e'), with a greater exponent, where e' is free and e is a factor of e'.*

Proof. A repeated use of the previous remark yields the result because the sequence of extensions of the excess of a kernel repetition eventually ends as property (FF) holds. □

In order to verify condition (IK), it then suffices to be sure that there does not occur a kernel repetition in $\mathscr{L}$, whose exponent is greater than s_k, and whose excess is a free word. Denote by (IKF) this last condition.

Nonempty languages that satisfy conditions (FS), (E), (IS), (FF) and (IKF) would then yield a proof of the conjecture for a given integer k; we shall look for them in the class of morphism-generated languages.

3. Morphism-generated languages

3.1. Definitions, first properties

Let ϕ be a morphism from $\{0, 1\}^*$ into itself; $\mathscr{L}_\phi$ stands for the language generated by ϕ, which is inductively defined as follows:

- words 0 and 1 belong to it,
- the image by ϕ of an element of $\mathscr{L}_\phi$ still belongs to $\mathscr{L}_\phi$,
- and every factor of an element of $\mathscr{L}_\phi$ is an element of $\mathscr{L}_\phi$.

According to its very construction, $\mathscr{L}_\phi$ satisfies condition (FS).

We shall further restrict ourselves to *standard* morphisms; a morphism ϕ is called standard if neither $\phi(0)$ nor $\phi(1)$ is the empty word, if one of these two words at least has a length greater than 1, which ensures that the language $\mathscr{L}_\phi$ is infinite, and if arbitrary powers of 0, 1 or 01 do not exist in the language.

The interest of this standardness assumption is explained by the following simple propositions.

Proposition 3.1. *If morphism ϕ is standard, then either* 0 *or* 1 *is a free word of $\mathscr{L}_\phi$.*

Proof. Let x be one of the two letters 0, 1. If x is not a free word, then it is either left-determined or right-determined.

Suppose that x is right-determined, i.e. that y always follows x in a word of $\mathscr{L}_\phi$. Then, y cannot be x as arbitrary powers of x are forbidden in $\mathscr{L}_\phi$; and x is followed by the other letter $\bar{x}$. Moreover, as an x cannot appear before another x, x is also left-determined.

Now, if neither 0 nor 1 is free, both words are determined on both sides and every word in $\mathscr{L}_\phi$ is a factor of $(01)^\infty$, which is excluded by the standardness hypothesis. □

Proposition 3.2. *If none of the two words $\phi(0)$ and $\phi(1)$ is empty, then, for any integer n, there exists an integer m such that every word of $\mathscr{L}_\phi$, whose length is greater than or equal to m, has a factor, which is the image by morphism ϕ of a word of $\mathscr{L}_\phi$ of length n.*

Proof. Denote by $\|\phi\|$ the maximum of the lengths $|\phi(0)|$ and $|\phi(1)|$. Integer m can be chosen as $(n+1)\|\phi\|-1$. □

Proposition 3.3. *If morphism ϕ is standard, then, for every element w of $\mathscr{L}_\phi$, there exists an integer n such that all words in $\mathscr{L}_\phi$, whose length is greater than or equal to n, have w as a factor.*

Proof. According to the standardness hypothesis, 0 and 1 are factors of all words of $\mathscr{L}_\phi$ whose length is greater than some n_0. The previous proposition then gives the proof by induction for all words $\phi^k(i)$, $i=0$, 1; and the elements of $\mathscr{L}_\phi$ are factors of them. □

Proposition 3.4. *Extensibility condition (E) holds for languages generated by standard morphisms.*

Proof. For morphism-generated languages, condition (E) is satisfied as soon as 0 and 1 can be left and right extended. Standardness implies that the language is infinite and, if one of the two letters could not be extended on some side, then arbitrary powers of the other would exist in $\mathscr{L}_\phi$. □

Standardness is then an useful tool for the proof; this assumption can be settled by effective verifications. On the other hand, looking down the finite list of all forbidden words effectively decides if condition (IS) holds.

Some more results are to be described in order to reduce the proof of the conjecture to a finite number of effective verifications on a well choosen language $\mathscr{L}_\phi$. As we said above, a nonempty language is convenient if it satisfies sufficient conditions (FS), (E), (IS), (FF) and (IKF). It remains to be shown that conditions (FF) and (IKF) can be effectively decided for some morphism-generated languages.

Let us now describe the conjugacy of morphisms, with which the verification of these conditions can become effecive.

3.2. *Conjugacy of morphisms*

After Lentin's defect theorem, if the images $\phi(0)$ and $\phi(1)$ do not constitute a code, these two words are powers of a same third word w; $\mathscr{L}_\phi$ is then the set of all factors of powers of w and condition (IK) cannot be satisfied.

We are then interested by morphisms ϕ such that $\phi(0)$ and $\phi(1)$ constitute a code; denote by (CC) this code condition. Let morphism ϕ satisfy (CC). If $\phi(0)$ and $\phi(1)$ start with the same letter x, denote by $x^{-1}\phi x$ the morphism defined by $(x^{-1}\phi x)(i)=x^{-1}\phi(i)x$ for $i=0, 1$.

As it will be shown later, the language, that this *conjugate* morphism generates, is the same as $\mathscr{L}_\phi$. After a finite number of steps, this left conjugacy is no longer possible; the two words would otherwise have equal powers and could not constitute a code.

This allows us to define a mapping L from the set of all morphisms satisfying condition (CC) to the free monoid $\{0, 1\}^*$ by induction:

- if the first letters of $\phi(0)$ and $\phi(1)$ are different, $L(\phi)$ is the empty word;
- on the contrary, if x is the first letter of these two words, $L(\phi)$ is the concatenation of x and $L(x^{-1}\phi x)$.

Right conjugacy and mapping R are defined in a similar way.

It can be inductively proved that $L(\phi)$ is the greatest left-ambiguous word for ϕ: a word is said to be *left-ambiguous* if it is a common left factor of two words $\phi(0u)$ and $\phi(1v)$, and every left-ambiguous word is a left factor of $L(\phi)$. This property is indeed true if the first letters of $\phi(0)$ and $\phi(1)$ are different, and, when assumed for $x^{-1}\phi x$, it also holds for ϕ.

It can also be inductively established that $L(\phi)$ is the greatest common left factor of all long enough words $\phi(w)$, where w belongs to $\mathscr{L}_\phi$.

Mirror properties hold for mapping R.

We give now some useful properties, related to the conjugacy of morphisms.

Remark 3.5. The conjugate of a standard morphism is standard. The conjugate of a morphism for which the code condition (CC) holds, also satisfies (CC).

Definition 3.6. Given a morphism ϕ from $\{0, 1\}^*$ into itself satisfying condition (CC), denote by ψ_ϕ or simply by ψ the so-defined mapping from $\{0, 1\}^*$ into itself:

$$\psi(w) = R(\phi)\phi(w)L(\phi).$$

Mapping ψ only depends on the conjugacy class of morphism ϕ.

Proposition 3.7. *If conditions (E) and (CC) hold for ϕ, in particular if ϕ is standard, the image $\psi(w)$ of a word w of $\mathscr{L}_\phi$ also belongs to the language.*

Proof. Take any long enough extension $w_1 w w_2$ of w on both sides; $R(\phi)$ is then a right factor of $\phi(w_1)$ and $L(\phi)$ a left factor of $\phi(w_2)$. And $\psi(w)$ belongs to $\mathscr{L}_\phi$ as a factor of $\phi(w_1 w w_2)$. □

Proposition 3.8. *If conditions (E) and (CC) hold for ϕ, language $\mathscr{L}_\phi$ can also be inductively defined with the aid of mapping ψ: $\mathscr{L}_\phi$ is the smallest language containing 0 and 1, stable under ψ and under factor taking.*

Proof. According to the previous proposition, the so-defined language is contained in $\mathscr{L}_\phi$; as it is stable under factor taking, it contains $\mathscr{L}_\phi$. □

Corollary 3.9. *If conditions (E) and (CC) hold for ϕ, language $\mathscr{L}_\phi$ depends only on the conjugacy class of ϕ.*

Proof. Language $\mathscr{L}_\phi$ can be defined from ψ, which only depends on the conjugacy class of ϕ. □

Because $\mathscr{L}_\phi$ depends only on the conjugacy class of ϕ, one of the conjugate morphisms can be distinguished to define what are the *markable* words of the language. This useful conjugate is the one, whose image by mapping R is the empty word; such a morphism is said to be *normalized*.

Proposition 3.10. *If conditions (E) and (CC) hold for a normalized morphism ϕ, the image by ψ of a free word in $\mathscr{L}_\phi$ is another free word of the language $\mathscr{L}_\phi$.*

Proof. As w is free, $w0$ and $w1$ belong to $\mathscr{L}_\phi$; they can be enlarged in $\mathscr{L}_\phi$ to $w0w_0$ and $w1w_1$. And $\psi(w)$ is a common left factor of $\phi(w0w_0)$ and $\phi(w1w_1)$. The letter following $G(\phi)$ in $\phi(0w_0)$ and $\phi(1w_1)$ cannot be the same and $\psi(w)$ is free. □

Corollary 3.11. *If conditions (E) and (CC) hold for ϕ, condition (FF) is fulfilled as soon as 0 and 1 are factors of free words.*

Proof. Every word of the language is a factor of some $\psi^k(i)$; and if i ($i = 0, 1$) is a factor of a free word w, $\psi^k(i)$ is a factor of the free word $\psi^k(w)$. □

The following proposition shows that if 0 or 1 is not a factor of a free word, then repetitions with an arbitrarily large exponent occur in $\mathscr{L}_\phi$.

Proposition 3.12. *Let ϕ be a standard morphism and suppose that a letter x (0 or 1) is not a factor of a free word, which means that every word of $\mathscr{L}_\phi$ which is not a power of $\bar{x}$ is either left or right determined. Then, there exists a nonempty word w in the language whose powers all belong to $\mathscr{L}_\phi$.*

Proof. An increasing sequence of necessary extensions in $\mathscr{L}_\phi$, $w_0 = x$, $w_1, \ldots$, where $w_{i+1} = w_i x_{i+1}$ or $w_{i+1} = x_{i+1} w_i$ can be built as x appears in every w_i.

On one side at least, suppose on the right, there are infinitely many extensions and we can consider that all extensions take place on the right:

$$w_0 = x, \ \ldots, \ w_{i+1} = w_i x_{i+1}, \ldots$$

As ϕ is standard, some x_i is equal to x (arbitrary powers of $\bar{x}$ are forbidden) and the sequence of all x_i is periodical; and all powers of the period word belong to $\mathscr{L}_\phi$. □

3.3. Interpretations

Consider a standard normalized morphism ϕ, i.e. a standard morphism satisfying (CC) for which $R(\phi) = \varepsilon$. We call *beginning*, a nonempty right factor of $\phi(0)$ or $\phi(1)$; we call *end*, a nonempty left factor of one of these two words. An *interpretation* of a word w of $\mathscr{L}_\phi$ is a triple (b, u, e), where b is a beginning, e an end and u a nonempty word in $\mathscr{L}_\phi$ such that:

- beginning b is a right factor of $\phi(f)$, where f is the first letter of u: $e'b = \phi(f)$;
- end e is a left factor of $\phi(l)$, where l is the last letter of u: $eb' = \phi(l)$;
- image $\phi(u)$ is equal to $e'wb'$.

For instance, if $\phi(0) = 11$ and $\phi(1) = 10$, $(1, 01011, 1)$ is an interpretation of the word 1 10 11 10 1: we describe how the given word of $\mathscr{L}_\phi$ is a factor of the image by ϕ of another word of the language.

A word of $\mathscr{L}_\phi$ is said to be *markable* if all its interpretations have the same beginning.

The useful normalized morphisms (with respect to the proof that we look for) will be asked to satisfy the following condition:

- every long enough word of $\mathscr{L}_\phi$ is markable (Condition M).

Property (M) is decidable for the morphisms that we deal with, as follows from the next theorem. A partial proof of a similar result can be found in [3].

Theorem 3.13. *If there exist arbitrarily long words of $\mathscr{L}_\phi$ that are not markable, i.e. if condition (M) does not hold for ϕ, then there exists a nonempty word w in $\mathscr{L}_\phi$ whose powers all belong to $\mathscr{L}_\phi$.*

Proof. If condition (M) is not fulfilled, there exists a right infinite word W, whose finite factors all belong to $\mathscr{L}_\phi$, with two different decompositions:

$$W = uu_1 \dots u_n \dots = vv_1 \dots v_n \dots,$$

where the u_i and v_j are images by ϕ of 0 or of 1, where u and v are right factors of such words, i.e. what we called beginnings ($u \neq v$).

Denote by U_i the concatenation product $uu_1 \dots u_i$ (U_0 then stands for u) and by V_j the product $vv_1 \dots v_j$ ($V_0 = v$).

For every pair (i, j) of positive integers, either U_i is a left factor of V_j, or V_j is a left factor of U_i; but, as morphism ϕ is normalized, the last letters of $\phi(0)$ and $\phi(1)$ are different and it is possible to decipher from the right so that no U_i is equal to a V_j.

The infinite word W with a double decomposition, whose existence we supposed, is completely determined by the starting pair (u, v) of beginnings, as follows from the next construction in which we go further and further in the double decomposition.

At the first step, we know the starting pair $(U_i(0), V_j(0)) = (U_0, V_0)$. One of these two words is longer than the other (it is equal to the concatenation of the other and of some end $e(0)$) and is certainly followed by the greatest left-ambiguous word $L(\phi)$; the shorter one must then be followed by $eL(\phi)$, which cannot be left-ambiguous, so that we have no choice for the image of a letter to put after this shorter word.

We are then faced with the same situation and inductively build a sequence of pairs $(U_{i(k)}, V_{j(k)})$, where the longer of the two words is equal to concatenation of the shorter with some end $e(k)$. Moreover, $e(k+1)$ only depends on $e(k)$ and our process is periodical, which yields the eventually periodical infinite word W. Whence the result. □

Corollary 3.14. *Property (M) is effectively decidable for standard morphisms: either condition (M) holds in $\mathscr{L}_\phi$ or the number of words with a given length in the language, which is an increasing function of the length, eventually becomes constant.*

Remark 3.15. It is not difficult to state a converse to Theorem 3.13: if there exists a nonempty word w whose powers all belong to $\mathscr{L}_\phi$, then property (M) does not hold for the language.

Indeed, the length of some of the shortest of these words is at least 2 and then $|\phi(w)| > |w|$. An infinite descent method then shows that w cannot be markable nor any power of it.

Proof. If condition (CC) does not hold for ϕ, then $\mathscr{L}_\phi$ consists of all factors of powers of some nonempty word w; if this condition holds while condition (M) is not fulfilled, there also exists an nonempty word w, whose powers all belong to $\mathscr{L}_\phi$, by the previous theorem.

As ϕ is standard, all elements of $\mathscr{L}_\phi$ with a given length occur as factors in every long enough word of the language, and in particular in some well chosen power of w;

the number of elements of $\mathscr{L}_\phi$ with a given length is then bounded by the length of w, which completes the proof. □

Property (M) is then decidable for a standard normalized morphism. If property (M) holds for a standard normalized morphism, it is an effective task to build the finite set of all free nonmarkable words of the corresponding language.

All free words are then known, as follows from the next proposition.

Proposition 3.16. *Let ϕ be a standard normalized morphism for which condition (M) holds. Then every markable free word w of $\mathscr{L}_\phi$ has a decomposition $w=\psi(\bar{w})=\phi(\bar{w})L(\phi)$, where $\bar{w}$ is a shorter free word of $\mathscr{L}_\phi$.*

Proof. If w is markable, all its interpretations have the same beginning. This beginning is either $\phi(0)$ or $\phi(1)$; otherwise, w would be left determined (if we know the last letter of $\phi(i)$, we know i).

This word w can then be written $w=\phi(\bar{w})w'$, where w' is free on its right and left ambiguous: such a decomposition exists; if w' were right determined, so would be w; if w' is not left ambiguous, one can find a shorter one. Word w' is then equal to the greatest left-ambiguous word $L(\phi)$.

Moreover $\bar{w}$ is free: a determination of $\bar{w}$ on some side would imply a determination of w on the same side.

It remains to be shown that $|\bar{w}|<|w|$.

The only possibility for $\bar{w}$ to be as long as w is that $L(\phi)=\varepsilon$ and that $\bar{w}=x^k$ for a letter x such that $\phi(x)=x$ or $\phi(x)=\bar{x}$.

If $\phi(x)=x$ and $L(\phi)=\varepsilon$, then $\phi(\bar{x})=\bar{x}\lambda\bar{x}$ and $w=\bar{w}$ is a factor of λ, hence a factor of $\phi(\bar{x})$.

If $\phi(x)=\bar{x}$ and $L(\phi)=\varepsilon$, then $\phi(\bar{x})=x\lambda x$ and $w=\bar{x}^k$ is a factor $\phi(\bar{x})$.

In both cases, w cannot be markable, as it is a factor of $\phi(x^k)$ and of $\phi(\bar{x})$. □

Remark 3.17. If ϕ is a standard normalized morphism for which condition (M) holds, then condition (FF) is easily fulfilled. According to previous propositions, it remains to be shown that 0 and 1 are factors of free words. Following Proposition 3.12, if either 0 or 1 is not a factor of a free word, then arbitrary powers of a nonempty word belong to $\mathscr{L}_\phi$, which is excluded by property (M).

The infinite language $\mathscr{L}_\phi$ then yields a proof of the conjecture if the standard normalized morphism ϕ satisfies sufficient conditions (IS), (FF01), (M) and (IKF).

Conditions (IS), (FF01) and (M) are decidable; in order to make condition (IKF) effectively computable for a given morphism, a new reduction is to be introduced (the first reduction was the passage from condition (IK) to condition (IKF)).

This reduction will allow us to consider kernel repetitions with a free excess, where this excess is moreover nonmarkable, to settle condition (IKF). An algebraic condition on ϕ is related to this reduction; let us discuss it.

3.4. Algebraic conditions

Let us now introduce an algebraic condition (A) on a morphism ϕ, relating ϕ to the previously defined morphism σ_k from $\{0, 1\}^*$ onto the symmetric group $\mathscr{S}_k$:

- there exists an inner automorphism α of the symmetric group such that the following equalities simultaneously hold (A).

$$\sigma_k(\phi(0))=\alpha(\sigma_k(0)); \qquad \sigma_k(\phi(1))=\alpha(\sigma_k(1)).$$

The important condition for α is to be an automorphism; but, except for the symmetric group $\mathscr{S}_6$, all automorphisms are inner ones. This restriction will not therefore be an important one.

Let us remark that condition (A) only depends on the conjugacy class for morphisms satisfying condition (CC). Condition (A) will be used as follows: if $\phi(w)$ belongs to the kernel of σ_k, so does w.

For $k>2$, condition (A) implies that the set $\{\phi(0), \phi(1)\}$ is a code (CC).

3.5. Reduction of kernel repetitions

Recall that we would like to control the exponent of kernel repetitions that occur in a language $\mathscr{L}_\phi$ generated by a morphism ϕ.

As it has been said previously, if condition (FF) holds, a control of the exponent of kernel repetitions, whose excess is free, will be sufficient (that does not depend on the generation of the language by a morphism).

In the case of a morphism-generated language $\mathscr{L}_\phi$, where the standard normalized morphism ϕ satisfies condition (A), the previously announced second reduction is given by the following propositions.

Proposition 3.18. *Let the standard normalized morphism ϕ satisfy condition (A). The mapping μ, given by*

$$\mu((p', e'))=(\phi(p'), \psi(e'))$$

transforms a kernel repetition with a free excess that appears in $\mathscr{L}_\phi$ in another one.

Proof. As ϕ satisfies condition (A), $\phi(p')$ belongs to the kernel of σ_k. Moreover, word $p'e'$ belongs to the language $\mathscr{L}_\phi$ and its image by ψ also belongs to $\mathscr{L}_\phi$. But $\psi(p'e')=\phi(p')\psi(e')$, which proves that the kernel repetition $\mu((p', e'))$ occurs in $\mathscr{L}_\phi$. □

Proposition 3.19. *Let the standard normalized morphism ϕ satisfy conditions (A). If the free excess e of a kernel repetition (p, e) is markable, then (p, e) is the image of another such repetition by μ.*

Proof. If e is a free markable word, $e=\phi(e')\,L(\phi)$, where e' is another free word; if e is moreover the excess of a kernel repetition (p, e), $pe=ep'$ is markable and can be written as

$$pe=\phi(p'e')\,L(\phi)=\phi(p')\phi(e')\,L(\phi)=\phi(e')\phi(p'')\,L(\phi).$$

According to condition (A), as $p=\phi(p')$, p' also belongs to the kernel of σ_k. The kernel repetition (p, e) is the image by the mapping μ of another kernel repetition (p', e'); e is free if and only if e' is free. □

Corollary 3.20. *Under the previous hypotheses, every kernel repetition (p, e) whose excess is a free word, is the image under some power of μ of another such repetition (p', e'), whose excess is not markable.*

Proof. A repeated use of the previous proposition yields a sequence (p_n, e_n) of kernel repetitions, whose excess is free and such that $(p, e)=\mu^n((p_n, e_n))$. This sequence goes on as long as the free excess is markable; this must eventually end because the length decreases, and the last free excess e' is not markable according to Proposition 3.12. □

If condition (M) holds, the verification of condition (IKF) consists in the control of the exponent of all repetitions $\mu^n((p, e))$, where e is a free nonmarkable word (there is a finite number of such words), where p belongs to the kernel of σ, and where pe belongs to the language.

Sufficient conditions are now to be given in order to make the calculation of the least upper bound of the exponents of all repetitions $\mu^n((p, e))$ from e and p feasible.

Denote by e_i (by p_i) the number of letters i ($i=0, 1$) in e (in p). Call α and β the numbers of 0 and of 1 in $L(\phi)$ (recall that $R(\phi)$ is empty for a normalized morphism). The notation $|w|_i$ will stand for the number of i ($i=0, 1$) in the word w.

Let then $M(\phi)$ be the 2×2 matrix

$$M(\phi)=\begin{pmatrix} a & b \\ c & d \end{pmatrix}=\begin{pmatrix} |\phi(0)|_0 & |\phi(1)|_0 \\ |\phi(0)|_1 & |\phi(1)|_1 \end{pmatrix}$$

and let $N(\phi)$ be the 3×3 matrix

$$N(\phi)=\begin{pmatrix} a & b & \alpha \\ c & d & \beta \\ 0 & 0 & 1 \end{pmatrix}.$$

The exponent of repetition $\mu^n((p, e))$ is given by the following ratio.

$$\rho(\mu^n((p, e)))=\frac{(1\ 1\ k-1)N^n\begin{pmatrix}e_0\\e_1\\1\end{pmatrix}}{(1\ 1)M^n\begin{pmatrix}p_0\\p_1\end{pmatrix}}=\frac{N_n}{D_n}.$$

Sequences (N_n) and (D_n) are increasing; the points corresponding to these elements of N^2 lie on a "convex curve", on a straight line or on a "concave curve" if the following quantity $q(\phi)$ is positive, equal to 0 or negative:

$$q(\phi)=(k-1)(1+\det(M(\phi))-a-d)+\alpha(1+c-d)+\beta(1+b-a).$$

In the convex case ($q(\phi)\geqslant 0$), the maximum of the values of $\rho(\mu^n((p, e)))$ is either $\rho((p, e))$, or $\rho_\infty((p, e))$, which represents the limit, as n tends to infinity, of $\rho(\mu^n((p, e)))$. This limit can be computed from the column vectors corresponding to p and e.

In the concave case ($q(\phi)<0$), if moreover the determinant of $M(\phi)$ is negative, this maximum of all $p(\mu^n((p, e)))$ is either $\rho((p, e))$, or $\rho(\mu((p, e)))$ (Case 1).

In the special case of an uniform morphism ($\phi(0)$ and $\phi(1)$ have the same length), the maximum is reached by $\rho((p, e))$. Indeed, for every kernel repetition, $\rho((p, e))$ is greater than or equal to $\rho(\mu((p, e)))$ (Case 2).

Thus, in these two particular cases, the verification of condition (IKF), for a standard normalized morphism satisfying (IS) and the algebraic condition (A), consists in showing, for every nonmarkable free word e, that no kernel repetition (p, e), such that its exponent or the exponent of its image by μ is greater than the conjectured threshold, occurs in the language.

On the one hand, there are finitely many free nonmarkable words and on the other hand, bad repetitions would have a short enough period, so that the verification of condition (IKF) is effectively possible.

4. Conclusions

4.1. Practical work

To fulfil the proving scheme that we have described, i.e. to find a morphism on the free monoid $\{0, 1\}^*$ that satisfies conditions (E), (FF01), (IS), (A), (M), (IKF) and then leads to a proof of Dejean's conjecture for a given value k of the cardinal number of the alphabet, the following tasks are done.

(1) Determination of all forbidden words to avoid short repetitions with an exponent strictly greater than s_k.

(2) Enumeration of all standard normalized morphisms, defined by short enough words, that are good candidates; these morphisms satisfy conditions (E), (IS), and (A).

(3) In the nonempty list of these morphisms, those that satisfy one of the nice conditions allowing an easier proof of the last steps are of special interest: cases 1 (concave case with negative determinant) and 2 (uniform morphisms).

(4) The proof is then achieved if such an example satisfies conditions (M) and (IKF).

Condition (M) is decidable thanks to Theorem 3.13 and it implies (FF); to establish the essential condition (IKF), it suffices to show, for every free nonmarkable word e, that there does not exist a kernel repetition (p, e) such that its exponent (or the exponent of its image by μ) is greater than s_k.

4.2. Results

Here are some morphisms giving a proof of Dejean's conjecture according to our proving scheme for cardinal numbers k, $2 \leqslant k \leqslant 11$.

$k=2$

Morphism $0 \to 11$, $1 \to 10$, solves the problem. The corresponding language is the one of all words coding the factors of Thue–Morse word.

$k=3$

Morphism ϕ, defined by $\phi(0)=1101$ and by $\phi(1)=11010$, leads us to the result.

The corresponding language on $\{1, 2, 3\}$ is different from Dejean's one [1]: 111011 does not belong to $\mathscr{L}_\phi$, while this word is the code of the transitions of *bcabcbac*, which is a word that belongs to Dejean's language.

An uniform morphism can also be found, which leads to the proof

$$0 \to 1010110,\ 1 \to 1011011.$$

$k=4$

The simplest example is Pansiot's one [4]:

$$0 \to 101101,\ 1 \to 10.$$

We did not find an uniform morphism, even much more complicated, that yields the proof.

Among others, the following uniform morphisms solve the conjecture for $k=5$, 6, 7, 8, 9, 10 and 11, and they are the shortest convenient uniform morphisms.

$k=5$

$$\begin{cases} 0 \to 0101011011010101101 10 \\ 1 \to 1010101011011011011 01 \end{cases}$$

$k=6$

$$\begin{cases} 0 \to 0101011011010110101 10 \\ 1 \to 1010110101101101011 01 \end{cases}$$

$k=7$

$$\begin{cases} 0\to 01101101101101011011011010 \\ 1\to 10101101101101011011011011 \end{cases}$$

$k=8$

$$\begin{cases} 0\to 10110101011010110101101010 \\ 1\to 10110101010110110110101011 \end{cases}$$

$k=9$

$$\begin{cases} 0\to 101011010110110101011010110110101010 \\ 1\to 101010110110110101011011010110101101 \end{cases}$$

$k=10$

$$\begin{cases} 0\to 1010101011011011011010101011011011010110 \\ 1\to 1010101011011011010110101010101011010101 \end{cases}$$

$k=11$

$$\begin{cases} 0\to 10101010101011011010110101101010110110 \\ 1\to 10101010101010110110110110110110101011011 \end{cases}$$

Acknowledgment

Maxime Crochemore drew my attention to this interesting problem; it is a pleasure for me to thank him for his friendly encouragements and his pertinent advices.

References

[1] F. Dejean, Sur un théorème de Thue, *J. Combin. Theory Ser. A* **13** (1972) 90–99.
[2] M. Lothaire, *Combinatorics on Words*, Encyclopedia of Mathematics and its Applications, Vol. 17 (Addison-Wesley, Reading, MA, 1983).
[3] J.C. Martin, Minimal Flows Arising from Substitutions of Nonconstant Length, *Math. Systems Theory* **7** (1) (1973) 73–82.
[4] J.J. Pansiot, A propos d'une conjecture de F. Dejean sur les répétitions dans les mots, *Discrete Appl. Math.* **7** (1984) 297–311.
[5] A. Thue, Über unendliche Zeichenreihen, *Norske Vid. Selsk. Skr. I Mat-Nat Kl. Christiana* **7** (1906) 1–22.
[6] A. Thue, Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen, *Norske Vid. Selsk. Skr. I Mat-Nat Kl. Christiana* **1** (1912) 1–67.