95. Square-free Words: the Thue–Morse Theorem

There is an infinite word on the alphabet 1, 2, 3 *in which no block occurs twice in succession.*

Proof. For $i = 1, 2, 3$, set $w_i = i(i+1)(i+2)(i+1)i$, with the letters taken modulo 3. Thus $w_1 = 12321$, $w_2 = 23132$ and $w_3 = 31213$. Furthermore, let $u_0 = 1$ and $u_{k+1} = u_k[w_1, w_2, w_3]$ for $k \geq 0$. Thus $u_1 = 12321$,

$$u_2 = 12321\ 23132\ 31213\ 23132\ 12321$$

and u_3 is

$$12321\ 23132\ 31213\ 23132\ 12321\ 23132\ 31213\ 12321\ 31213\ 23132\ \ldots.$$

Note that u_k has 5^k letters, and is the initial block of u_{k+1}. In particular, let u be the infinite word which is the 'limit' of these words: the nth letter of u is the nth letter of u_k whenever $5^k \geq n$. Thus, our aim is to prove that every u_k is square-free. (In fact, looking for an infinite word u is a bit of a red herring: an easy compactness argument tells us that if there are arbitrarily long (but finite) square-free words on a finite alphabet, then there is a two-way infinite word $\ldots x_{-2}x_{-1}x_0x_1x_2\ldots$ on the same alphabet without two identical consecutive blocks.)

Before we turn to this, let us give another description of the u_k and u. We know that each $u_k = a_1a_2\ldots$ starts with $a_1 = 1$, so, in order to describe it, it suffices to say whether a_2 is 'bigger' or 'smaller' than a_1 (i.e., whether $a_2 = a_1 + 1$ or $a_2 = a_1 - 1$ modulo 3), and then whether a_3 is bigger or smaller than a_2, and so on. Equivalently, u_1 is coded with $++--$ and u_2 with

$$++--\ +\ ++--\ +\ ++--\ -\ ++--\ -\ ++--.$$

The pattern is clear: the sign-sequence of u_{k+1} is obtained from that of u_k by adding $++--$ at the beginning, inserting the same sequence $++--$ between any two signs, and adding it at the end. (The sign-sequence of u_k has length $5^k - 1$; to obtain u_{k+1}, we take 4 signs at the beginning, 4 at the end, and $5^k - 2$ groups of 4 between the $5^k - 1$ signs, so that the new sign-sequence has $2 \cdot 4 + (5^k - 2) \cdot 4 + (5^k - 1) = 5^{k+1} - 1$ terms, as it should.) An equivalent pictorial representation of u_2 is given in Figure 77. Note that for each u_k the entire walk is a sequence of 'runs' of lengths 2 or 3, starting with an increasing run of 2: the total length of a decreasing run followed by an increasing run is always 5, so that the 'peaks' are precisely at distance 5 from each other, and the word u_k has exactly 5^{k-1} peaks. Furthermore, if $u_{k-1} = a_1a_2\cdots a_\ell$ then $(a_1 - 1)(a_2 - 1)\cdots(a_\ell - 1)$ is the sequence of peaks of u_k.

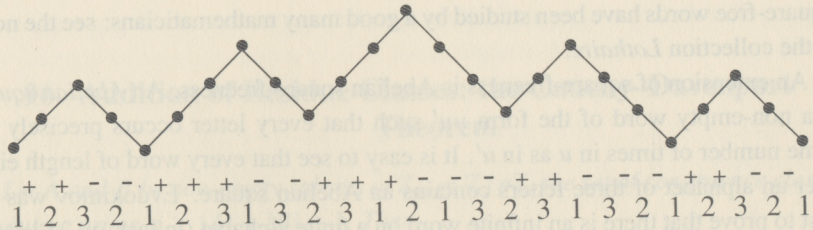

Fig. 77. A representation of u_2 as a walk going up and down by one step, according to the sign-sequence.

After these descriptions, at long last we embark on proving by induction that every u_k is square-free. This is trivial for $k \leq 1$, so assume that $k \geq 1$ and the assertion holds for u_k.

Suppose for a contradiction that u_{k+1} contains a square ww. We shall show that in this case u_k also contains a square: this contradiction will complete our induction step and the entire proof.

Write ℓ for the length of w, and $\sigma = \sigma_1\sigma_2\ldots\sigma_{2\ell-1}$ for the sign-sequence giving ww. Then the sum of the first ℓ signs (considered as $+1$ and -1, of course) is 0 modulo 3, and the subsequence of the first $\ell - 1$ signs is the same as the subsequence of the last $\ell - 1$ signs. In fact, σ gives ww if and only if these two conditions are satisfied. (For example, if $w = 2313$ then $ww = 23132313$ and $\sigma = ++-{-}++-$.)

Since the sum of the first ℓ signs is 0 modulo 3, w cannot be too short. First, trivially, $\ell \geq 3$. Even more, $\ell = 3$ cannot hold either since then σ has to start with $+++$ or $---$ and has to be followed by a run of length 2 of the opposite sign. Furthermore, if $\ell = 4$ then we must have $\sigma = ++--++-$ or $--++--+-$ neither of which can arise. Next, $\ell = 5$ cannot hold, since the sum of 5 terms, each of which is $+1$ or -1, cannot be 0 modulo 3. Hence $\ell \geq 6$. But then w contains at least one peak: having identified this peak, we know that every fifth letter is a peak, and the positions of the peaks are the same in the two copies of w. In particular, if the peaks in the first copy of w are $p_1p_2\ldots p_\ell$, then the peaks in the second copy form the same sequence $p_1p_2\ldots p_\ell$. But then, writing p for the word $(p_1 + 1)(p_2 + 2)\cdots(p_\ell + 1)$, we find that u_k contains the square pp, completing our proof. □

Notes. As we remarked at the end of the solution of Problem 95, Thue was the first to study avoidable substructures of infinite words: in particular, he was the first to prove the result in this question, using a different infinite word. The infinite word defined by Thue was rediscovered by Morse (see also Morse and Hedlund).