Journal of Universal Computer Science, vol. 8, no. 2 (2002), 292-296
submitted: 15/9/01, accepted: 31/1/02, appeared: 28/2/02 © *J.UCS*

Bridging Two Hierarchies of Infinite Words[1]

Solomon Marcus
(Romanian Academy, Mathematics Calea Victoriei 125, Bucharest, Romania
Email: solomon.marcus@imar.ro.)

Abstract: Infinite words on a finite non-empty alphabet have been investigated in various respects. We will consider here two important strategies in approaching such words; one of them proceeds from particular to general, while the other proceeds from general to particular. As we shall see, the respective hierarchies don't interfer. There is between them an empty space waiting for investigation.

Key Words: Ultimately periodic infinite, uniformly recurrent, disjunctive, random infinite words
Category: F.1, F.1.1.

1 From particular to general

Already in 1938, Morse and Hedlund [19] focused their attention on ultimately periodic infinite words on a finite alphabet, showing that they are characterized by the property of boundedness of the number of factors of the same length. A similar property for languages was considered only 55 years later [9]: a language L is said to be slender if the number of words in L of the same length is bounded. However, as one can show, in order to get the right analogy, for languages, of the property considered by Morse and Hedlund for infinite words, we should replace slenderness by strong slenderness of L, defined as the slenderness of the language of factors of L. Strongly slender languages have some features near to periodicity.

The function associating to each n the number of factors of length n is said to be the complexity function; it may refer equally to infinite words and to languages. The Morse-Hedlund theorem suggests to consider periodicity as the lowest possible complexity of an infinite word, while the complexity of infinite words which are not ultimately periodic is given by its degree of non-periodicity. It was shown that the lowest possible complexity which is not bounded corresponds to Sturmian infinite words, defined by $f(n) = n + 1$, where f is the complexity function. An example of Sturmian word is the Fibonacci word, defined, on the alphabet $\{a, b\}$, as the fixed point of the morphism replacing a by ab and b by a. One can define also Sturmian languages; an example, in this respect, is the well-known language whose words are obtained by concatenation of n occurrences of a with n occurrences of $b\,(n = 1, 2, 3, \ldots)$.

Ultimately periodic words are a particular case of Toeplitz words. Consider an infinite process starting with a pattern w as a word on the alphabet obtained by union between the initial alphabet X and an additional element denoted by the interogative sign ?, which cannot be the first term in w (a term is an

[1] C. S. Calude, K. Salomaa, S. Yu (eds.). *Advances and Trends in Automata and Formal Languages. A Collection of Papers in Honour of the 60th Birthday of Helmut Jürgensen.*

occurrence of an element of the underlying alphabet in w). In a first step, we consider the infinite word of period w. The word of n-th order is obtained from the word of $(n-1)$-th order, by replacing the first occurrence of ? with the n-th letter of the word obtained at the $(n-1)$-th step (this letter is always different from ?). The limit of the word of order n, when n is increasing to infinite is said to be the Toeplitz word generated by the pattern w and it is denoted by $T(w)$. If p is the length of w and q is the number of holes in w, (i.e., of occurrences of ?), then $T(w)$ is said to be a (p, q)-Toeplitz word. Periodic words are those Toeplitz words where w includes no hole. The complexity of non-periodic (p, q)-Toeplitz words is always polynomial and depends exclusively of p and q (as it was shown by Cassaigne and Karhumäki [7].

Both Sturmian words and Toeplitz words are particular cases of almost periodic words (De Luca and Varricchio [10], Cassaigne and Karhumäki [7]). The infinite word w is said to be almost periodic (Jacobs and Keane [13] if for any finite factor x of w there exists a decomposition of w into infinitely many factors of the same length, such that every term of the decomposition includes a copy of x. One can prove that each of the following four variants is equivalent to *almost periodicity*:

1. The infinite word w is said to be *uniformly reccurent* in the sense of [12] if for each positive integer k there exists a constant $n(k)$ such that, if x of length k occurs as a factor of w, then it occurs in any factor of w of length $n(k)$.
2. The infinite word w is said to be *uniformly recurrent* in the sense of [11] (see also [1]) if any factor x of w occurs infinitely many times and the gaps between consecutive occurrences of x are bounded in length.
3. The infinite word w is said to be *uniformly recurrent* in the sense of [10] if every finite factor x of w occurs syntetically in w, i.e., there exists an integer k such that in any factor of w of length k there is at least one occurrence of x.
4. The infinite word w is said to be *minimal*, if there exists no other infinite word v (on the same alphabet) such that every factor of v is a factor of w.

The Fibonacci word is not a Toeplitz word. The Thue-Morse word, defined as the fixed point of the morphism replacing a by ab and b by ba is almost periodic, but it is neither a Toeplitz word nor a Sturmian word; its complexity is not polynomial.

Another extension of (ultimately) periodic words is given by quasi-periodic infinite words. We define them following the model of quasi-periodic finite words. The infinite word w is *quasi-periodic* if there exists a finite factor x of w such that any term of w is included in a copy of x; the factor x is a quasi-period of w. Here is an example of a finite quasi-periodic word which is not periodic: $abaababaababa\ldots$ (the quasi-period is aba). An example of an infinite quasi-periodic word which is not periodic is: $abaababaabaababaabaababaabaababaaba\ldots$ where the generic term of order n is obtained by concatenating n iterations of aba and ba; the quasi-period is aba. Quasi-periodicity for finite words occurs in connection with molecular sequence analysis [17] and with DNA sequences [18], while in [8] some connection with musical similarity and with melodic recognition is shown.

In contrast with periodicity, where the interval corresponding to the period is concatenated with infinitely many copies of it, the quasi-period is sometimes concatenated, sometimes superposed.

2 From general to particular

We start with arbitrary infinite words on the finite alphabet X, i.e., infinite words for which no restriction is formulated. The trap of such a concept is that as soon as we try to give an individual example of an arbitrary infinite word, it is no longer arbitrary. But total arbitrariness is misleading. For instance, any infinite word includes some factors occurring infinitely many times. An example is the infinite word on $\{a, b\}$ obtained by concatenation of n occurences of a followed by n occurrences of $b\,(n = 1, 2, 3, \ldots)$. Any factor obtained by concatenation of p consecutive occurrences of a followed by q consecutive occurrences of b (with p different from q) occurs only finitely many times; but obviously, some factors, such as those including no a or no b, occur infinitely many times in the considered infinite word.

A first restriction on arbitrary infinite words is obtained by considering the sub-class of so-called *recurrent infinite words,* defined by the property that any factor occurs infinitely many times. Obviously, any almost periodic infinite word is a recurrent word, but the converse is not true, as it is shown by the infinite word obtained as the fixed point of the morphism where a is substituted by aba, while b is substituted by bb; since there are arbitrarily long factors containing no a, the recurrence is not uniform.

One can imagine, between recurrent and uniformly recurrent infinite words, some intermediate classes. A particular class of infinite recurrent words is that of *disjunctive words*; this means that any finite word on the alphabet X is a factor of the considered infinite word. It is easy to see that if w is disjunctive, then any finite word on X occurs infinitely many times as a factor of w; it follows that any infinite disjunctive word is a recurrent word. The converse is not true, as the above example of a recurrent word which is not almost periodic shows (no word enough long, including only occurrences of a, is a factor of the respective infinite word).

Disjunctive words have a history of about 60 years; see, in this respect, Jürgensen [14] and also [5, 6, 15, 16]. The concept of a disjunctive language and its connections with disjunctive infinite words is analyzed in [14, 15, 16].

Randomness is another important possible property of an infinite word, for which Calude [3] and Calude and Hromkovič [4] give a very detailed account; see also Calude and Jürgensen [5]. Roughly speaking, a finite word is random if it has maximal program-size complexity, when compared with the program-size complexity of all words of the same length ([4]:40). A random finite word cannot be algorithmically compressed, but incompressibility is not an effective property: no individual word, except finitely many, can be proven random, despite the fact that, in some sense, most finite words are random. However, one can describe a non-effective construction of random finite words ([4]: 41). An infinite word is said to be *random* if all its prefixes are random. As it was shown by Jürgensen, Shyr and Thierrin [15] (see also Jürgensen ([14]:267), randomness implies disjunctivity, but, as it was shown by Jürgensen-Thierrin [16] (see also Jürgensen [14]:267), the converse is not true; there are disjunctive infinite words which are not random.

Let $F(w, n)$ the number of factors of w of length n which appear infinitely many times as factors of w; $F(w, n)$ is considered by Nakashima, Tamura and Yasutomi [20] as the complexity function of w. It is easy to see that the function $F(w, n)$ takes, for each n, its maximum value when w is recurrent and only in this case. One can define the function $G(w, n)$ as the number of words on X of length n which appear infinitely many times as factors of w; this function takes, for each n, its maximum value when w is disjunctive and only in this case. So, one can weaken the condition of disjunctivity by permitting to F to take, at least for some values of n, values smaller than the maximum possible value, i.e., smaller than n^p, where p is the cardinal of the alphabet. The maximality of $F(w, n)$ is also related to Martin-Löf-Chaitin-Kolmogorov randomness.

3 Bridging the hierarchies

Suppose the cardinal of the alphabet is strictly larger than one. We exclude from our considerations trivial words, i.e., words which don't include all elements of the alphabet Do the above hierarchies interfer? The answer is negative, as it follows from the next two statements:

1. *No almost periodic infinite word is disjunctive (Staiger [21]).*
2. *No quasi-periodic infinite word is disjunctive.*

Indeed, let w be an infinite quasi-periodic word. It follows the existence of a factor y of w such that any term of w is included in a copy of y. This fact implies that w includes all elements of the alphabet occurring in w. Due to our preliminary assumptions, it follows that w includes at least two different elements of the alphabet, let them be a and b; so y will, in its turn, include obligatory a and b. On the other hand, no word obtained by concatenation of a with itself n times, where n is strictly larger than the length of y, can be a factor of w, so w is not disjunctive.

References

1. J. P. Allouche, Sur la complexite des suites infinies. *Bull. Belg. Math. Society* 1 (1994), 133-143.
2. V. Berthe, *Étude mathematique et dynamique des suites algorithmiques*. These d'habilitation, Univ. de Marseille, 1999.
3. C. Calude, *Information and Randomness. Algorithmic Perspective*. Berlin: Springer, 1994.
4. C. Calude, J. Hromkovič, Complexity: A language-theoretic point of view. Chapter I in volume 2 of *Handbook of Formal Languages* (eds. G. Rozenberg, A. Salomaa). Berlin: Springer,1997, 1-60.
5. C. Calude, H. Jürgensen, Randomness as an invariant for number representations. In *Results and Trends in Theoretical Computer Science* (eds. H. Maurer, J. Karhumäki, G. Păun, G. Rozenberg), Lecture Notes in Computer Science 812, Berlin: Springer, 1994, 44-66.
6. C. Calude, L. Priese, L. Staiger, *Disjunctive sequences. An overview.* CDMTCS-063, October 1997, Univ. of Auckland, New Zealand, 39 pages.
7. J. Cassaigne, J. Karhumäki, On the complexity of Toeplitz words. *European J. of Combinatorics* 18 (1997), 497-510.
8. T. Crawford, C.S. Iliopoulos, R. Raman, String matching techniques for musical similarity and melodic recognition. *Computing in Musicology*, 1998.
9. J. Dassow, G. Păun, A. Salomaa, On thinness and slenderness of languages. *EATCS Bulletin* 49 (1993), 152-158.

10. A. De Luca, S. Varricchio, *Combinatorics on Words and Regularity Conditions*, Berlin: Springer, 1998.
11. A. Ehrenfeucht, K.P. Lee, G. Rozenberg. Subwords of various classes of deterministic developmental languages. *Theoretical Computer Science* 1 (1975), 59-75.
12. H. Furstenberg, *Recurrence in Ergodic Theory and Combinatorial Number Theory.* Princeton: University Press, 1981.
13. K. Jacobs, M. Keane, 0-1 sequences of Toeplitz type. *Z. Wahr. Verw. Geb.* 13 (1969), 123-131.
14. H. Jürgensen, Disjunctivity. In *Words, Semigroups, Transductions* (eds. M. Ito, G. Păun, S. Yu), Festschrift in Honor of Gabriel Thierrin, New Jersey et al: World Scientific, 2001, 255-274.
15. H. Jürgensen, G. Shyr, G. Thierrin, Disjunctive omega-languages. *Elektron. Informationsverarb. Kybernet.* 19, 6 (1983) 267-278.
16. H. Jürgensen, G. Thierrin, Some structural properties of omega languages. *Nat. School with international participation "Applications of Math in Technology",* Sofia, 1988, 56-63.
17. S. Karlin, M. Morris, G. Ghandour, M. Leung. Efficient algorithms for molecular sequence analysis. *Proc. National Acad. Sci. USA* 85 (1988), 841-845.
18. A. Milosavljevic, J. Jurka, Discovering simple DNA sequences by the algorithmic significance method. *Comput. Appl. Biosci.* 9(1993), 407-411.
19. M. Morse, G. A. Hedlund, Symbolic dynamics, *American J. of Math.* 60 (1938), 815-866.
20. I. Nakashima, J.I. Tamura, S.I. Yasutomi, Modified complexity and Sturmian words (mentioned in [2] as a forthcoming publication).
21. L. Staiger, Personal communication, July 2001.