Numeration systems and string attractors

France Gheeraert
Joint work with Giuseppe Romana and Manon Stipulanti

In 1995, Fabre introduced a family of word morphisms using Parry numbers and the corresponding Rényi expansions of 1 [5]. More precisely, these morphisms are defined as follows. Let $\beta > 1$ be a Parry number. If β is simple and $d_\beta(1) = d_0 \cdots d_k$, we define $\sigma_\beta(i)$ as the word $0^{d_i}(i+1)$ for all $0 \leq i \leq k-1$, and $\sigma_\beta(k)$ as 0^{d_k}. We then extend σ_β to the set of finite sequences of elements in $\{0, \ldots, k\}$ (called *words* on $\{0, \ldots, k\}$) by saying that the image of a sequence is the concatenation of the images of its elements (called *letters*). If β is not simple and $d_\beta(1) = d_0 \cdots d_k (d_{k+1} \cdots d_m)^\omega$, we instead define σ_β for the words on $\{0, \ldots, m\}$ with

$$\sigma_\beta(i) = \begin{cases} 0^{d_i}(i+1) & \text{if } 0 \leq i \leq m-1 \\ 0^{d_n}(k+1) & \text{if } i = m. \end{cases}$$

Each morphism σ_β admits an infinite fixed point $\mathbf{u}_\beta$ starting with 0 obtained as the limit of $\sigma_\beta^n(0)$ when n tends to infinity. In his paper, Fabre proved that the classical Bertrand numeration system associated with β can be used to easily reconstruct the infinite word $\mathbf{u}_\beta$. This is a particular case of a more general result. Given a morphism μ admitting an infinite fixed point, we can build an automaton called the *prefix-suffix automaton* [4] and define a numeration system $\mathcal{S}_\mu$ such that, when reading $\mathrm{rep}_{\mathcal{S}_\mu}(n)$ in the automaton, we recover the n-th letter of the fixed point [12].

In the case of the morphisms σ_β, the corresponding numeration system $\mathcal{S}_{\sigma_\beta}$ is the well-known Bertrand numeration system associated with β, which is the greedy numeration system corresponding to the sequence of integers $(U_n)_n$ constructed using a linear recurrence based on $d_\beta(1)$ [3]. Observe that U_n is also the length of $\sigma_\beta^n(0)$. For a general morphism μ, the numeration system $\mathcal{S}_\mu$ is not necessarily greedy, even when defining μ similarly to σ_β using a general sequence $d_0 \cdots d_k$ instead, nor is its valuation always associated to a sequence of integers. This makes the morphisms σ_β all the more interesting in regards to numeration systems.

Later on, the fixed points $\mathbf{u}_\beta$ of the morphisms σ_β have been studied from a combinatorics-on-words point of view. In particular, several authors have studied their factor complexity [6], the palindromic factors [1] and the return words [2] in the words $\mathbf{u}_\beta$. Indeed, they offer a non-classical way of generalizing one of the most famous words in the field called the *Fibonacci word* which corresponds to $\mathbf{u}_\varphi$ where φ is the golden ratio (so $d_\varphi(1) = 11$).

When starting the present work, our goal was to study string attractors in infinite fixed points of morphisms. String attractors is a recent concept, originally introduced in the data compression field by Kempa and Prezza [8]. It also has applications in combinatorial pattern matching [10] and quickly gained traction in the combinatorics on words community. A string attractor can be conceptualized as follows: it is a set of positions within a finite word that enables to capture all distinct consecutive subsequences (called *factors*) appearing in the word. The obvious goal being to have the smallest string attractor possible. However, this question is

known to be NP-hard [8] and even the slightest modification in the word can radically impact the string attractors. Moreover, we are often not interested in string attractors for one specific words but for an infinite family of words.

It is therefore important to add some restriction on the considered words, in the hope that combinatorial properties will limit the possible behaviours. One way of doing so is to consider words that are prefixes of a common infinite word (meaning that they correspond to the subsequence from index 0 up to n for some n). Complete descriptions of the string attractors exist when considering prefixes of some famous infinite words: the Thue–Morse word [9], the period-doubling word [13], and Sturmian words [11]. We wanted to obtain similar descriptions but for a larger family of infinite words.

Preliminary results told us that, for the Fibonacci word, the Fibonacci numbers played a key role in such a description. Indeed, any prefix has a string attractor made of two consecutive Fibonacci numbers. As they form the basis of the Zeckendorf numeration system (the associated Bertrand numeration system), it was then natural to turn to the words $\mathbf{u}_\beta$ and wonder if a similar result was true. We obtained the following theorem.

Theorem 1 ([7])**.** *Let β be a simple Parry number, let k be the length of $d_\beta(1)$, and let $(U_n)_n$ denote the sequence for the Bertrand numeration system associated with β. Every prefix of $\mathbf{u}_\beta$ has a string attractor made of at most $k+1$ consecutive elements of $(U_n)_n$.*

Moreover, we have an explicit partition of $\mathbb{N}$ into intervals such that, depending on the interval containing the length of the prefix, we can tell which U_n's will be in the string attractor.

In fact, we can slightly extend this result by considering periodisations of $d_\beta(1)$ instead. Let β be a simple Parry number and $d_\beta(1) = d_0 \cdots d_{k-1}$. For all $n \geq 0$, we define $d_\beta^{(n)}(1) = (d_0 \cdots d_{k-2}(d_{k-1}-1))^n d_0 \cdots d_{k-1}$. It is an alternative representation of 1 in base β. We then define $\sigma_{\beta,n}$ and $\mathbf{u}_{\beta,n}$ in the same way as σ_β and $\mathbf{u}_\beta$, using $d_\beta^{(n)}(1)$ instead of $d_\beta(1)$. The corresponding numeration system $\mathcal{S}_{\sigma_{\beta,n}}$ is also the Bertrand numeration system associated with β. We then have the following result.

Theorem 2 ([7])**.** *Let β be a simple Parry number, let k be the length of $d_\beta(1)$, and let $(U_n)_n$ denote the sequence for the Bertrand numeration system associated with β. For all $n \geq 0$, every prefix of $\mathbf{u}_{\beta,n}$ has a string attractor made of at most $(n+1)k+1$ consecutive elements of $(U_n)_n$.*

However, this does not work if the morphism μ is defined based on a general sequence $d_0 \cdots d_{k-1}$ and if we consider the corresponding numeration system. This leads to the following open question.

Question 3. *Let μ be a morphism having an infinite fixed point $\mathbf{u}$. Does there exist a numeration system $\mathcal{S}$ such that*

1. *there exists an automaton in which reading $\mathrm{rep}_\mathcal{S}(n)$ restitutes the n-th letter of $\mathbf{u}$,*
2. *every prefix of μ has a string attractor easily described using $\mathcal{S}$?*

References

[1] Petr Ambrož et al. Palindromic complexity of infinite words associated with simple Parry numbers. *Ann. Inst. Fourier*, 56(7):2131–2160, 2006.

[2] L. Balková, E. Pelantová, and W. Steiner. Sequences with constant number of return words. *Monatsh. Math.*, 155(3-4):251–263, 2008.

[3] A. Bertrand-Mathis. Comment écrire les nombres entiers dans une base qui n'est pas entière. *Acta Math. Hungar.*, 54(3-4):237–241, 1989.

[4] V. Canterini and A. Siegel. Prefix-suffix automaton associated with a primitive substitution. *J. Théor. Nombres Bordx.*, 13(2):353–369, 2001.

[5] S. Fabre. Substitutions et β-systèmes de numération. *Theoret. Comput. Sci.*, 137(2):219–236, 1995.

[6] C. Frougny, Z. Masáková, and E. Pelantová. Complexity of infinite words associated with beta-expansions. *Theor. Inform. Appl.*, 38(2):163–185, no. 3, 269–271, 2004.

[7] F. Gheeraert, G. Romana, and M. Stipulanti. String attractors of fixed points of k-bonacci-like morphisms, 2023. Preprint available at arXiv:2302.13647.

[8] D. Kempa and N. Prezza. At the roots of dictionary compression: string attractors. In *STOC'18—Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 827–840. ACM, New York, 2018.

[9] K. Kutsukake et al. On repetitiveness measures of Thue-Morse words. In *String processing and information retrieval*, volume 12303 of *Lecture Notes in Comput. Sci.*, pages 213–220. Springer, Cham, 2020.

[10] Christiansen A. R. et al. Optimal-time dictionary-compressed indexes. *ACM Trans. Algorithms*, 17(1):8:1–39, 2021. Id/No 8.

[11] A. Restivo, G. Romana, and M. Sciortino. String attractors and infinite words. In *LATIN 2022: Theoretical informatics*, volume 13568 of *Lecture Notes in Comput. Sci.*, pages 426–442. Springer, Cham, 2022.

[12] M. Rigo. *Formal languages, automata and numeration systems. 2. Applications to recognizability and decidability.* Networks and Telecommunications Series. ISTE, London; John Wiley & Sons, Inc., Hoboken, NJ, 2014.

[13] L. Schaeffer and J. Shallit. String attractor for automatic sequences, 2022. Preprint available at arXiv:2012.06840.