Numeration systems: a bridge between formal languages and number theory

Michel Rigo

http://www.discmath.ulg.ac.be/
http://orbi.ulg.ac.be/

Arithmétique en Plat Pays : Journée printanière
30th May 2022

Université de Liège

- Recognizable sets of integers
- Possible generalizations
- Büchi–Bruyère theorem:
  applications in combinatorics and arithmetic

A computation model

We have many algorithms dealing with primality testing

Agrawal–Kayal–Saxena 2002

PRIMES is in $\mathcal{P}$

$$\textit{base } 2: \; 1, 10, 11, 100, 101, 110, 111, 1000, 1001, 1010, 1011, \ldots$$

$$\textit{base } 3: \; 1, 2, 10, 11, 12, 20, 21, 22, 100, 101, 101, \ldots$$

$\rightsquigarrow$ Concept of a *decidable* or *recursive* language:
there is a Turing Machine which, given a finite word as input, accepts it if it belongs to the language and rejects it otherwise.

The Chomsky hierarchy (from Wikipedia)

V·T·E **Automata theory: formal languages and formal grammars** [hide]

Chomsky hierarchy	Grammars	Languages	Abstract machines
Type-0	Unrestricted	Recursively enumerable	Turing machine
—	(no common name)	**Decidable**	Decider
Type-1	Context-sensitive	Context-sensitive	Linear-bounded
—	Positive range concatenation	Positive range concatenation*	PTIME Turing Machine
—	Indexed	Indexed*	Nested stack
—	—	—	Thread automaton
—	Linear context-free rewriting systems	Linear context-free rewriting language	restricted Tree stack automaton
—	Tree-adjoining	Tree-adjoining	Embedded pushdown
Type-2	Context-free	Context-free	Nondeterministic pushdown
—	Deterministic context-free	Deterministic context-free	Deterministic pushdown
—	Visibly pushdown	Visibly pushdown	Visibly pushdown
Type-3	Regular	Regular	Finite
—	—	Star-free	Counter-free (with aperiodic finite monoid)
—	Non-recursive	Finite	Acyclic finite

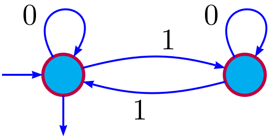
Each category of languages, except those marked by a *, is a proper subset of the category directly above it.
Any language in each category is generated by a grammar and by an automaton in the category in the same line.

Regular languages are accepted/recognized by finite automata (the simplest model of computation).

The family of regular languages is closed under Boolean operations, Kleene star, concatenation, (inverse) homomorphism, projections, mirror image, . . .

Much simpler than a Turing machine, linearly reading symbols once

Here is an example of finite automaton over $\{0, 1\}$



accepting words with an even number of 1's.

From integers to words

Choose a base $k \geq 2$, any integer $n > 0$ can be uniquely written as

$$n = \sum_{i=0}^{\ell-1} d_i \, k^i$$

with the digits $d_i \in \{0, \dots, k-1\}$ and $d_{\ell-1} \neq 0$

$$\mathrm{rep}_k(n) = d_{\ell-1} \cdots d_0.$$

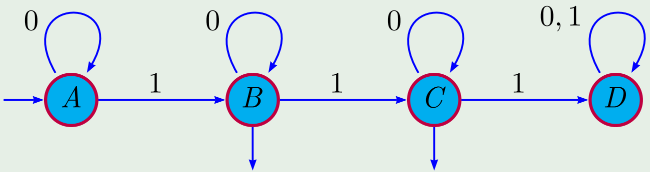
Definition — the bridge between the two worlds

A set $X \subset \mathbb{N}$ is *k-recognizable*, if the set of base-k expansions of the elements in X is accepted by some finite automaton, *i.e.*, $\mathrm{rep}_k(X)$ is a regular language.

Some examples

A 2-recognizable set

$X = \{n \in \mathbb{N} \mid \exists i, j \geq 0 : n = 2^i + 2^j\} \cup \{1\}$



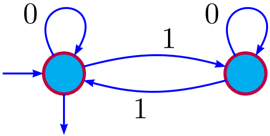
$$X = \{1, 2, 3, 4, 5, 6, 8, 9, 10, 12, 16, 17, 18, 20, 24, \ldots\}$$

$$\mathrm{rep}_2(X) = \{1, 10, 11, 100, 101, 110, 1000, 1001, 1010, 1100, \ldots\}$$

Some examples

- The set of even integers is 2-recognizable.
- More generally, the set of numbers congruent to r modulo m is 2-recognizable; and thus, any finite union of arithmetic progressions.
- The Prouhet–Thue–Morse set is 2-recognizable,

$$X = \{n \in \mathbb{N} \mid s_2(n) \equiv 0 \bmod 2\}$$



$$X = \{0, 3, 5, 6, 9, 10, 12, 15, 17, 18, \ldots\}$$

$$\mathrm{rep}_2(X) = \{\varepsilon, 11, 101, 110, 1001, 1010, 1100, 1111, 10001, \ldots\}$$

- The set of powers of 2 is 2-recognizable.

Digression about Prouhet's problem (1851):

QUESTION

Let $k \geq 1$. Can you partition the set $S_k = \{0, \dots, 2^k - 1\}$ such that

$$\sum_{i \in I} i^m = \sum_{i \in S_k \setminus I} i^m$$

for all $m \in \{0, \dots, k-1\}$?

More examples

Let $X = \{x_0 < x_1 < x_2 < \cdots\} \subseteq \mathbb{N}$. Define

$$\mathbf{R}_X := \limsup_{i\to\infty} \frac{x_{i+1}}{x_i} \text{ and } \mathbf{D}_X := \limsup_{i\to\infty} (x_{i+1} - x_i).$$

Gap theorem (Cobham'72)

Let $k \geq 2$. If $X \subseteq \mathbb{N}$ is a k-recognizable infinite subset of $\mathbb{N}$, then either $\mathbf{R}_X > 1$ or, $\mathbf{D}_X < +\infty$.

A. Cobham, Uniform tag, Theory Comput. Syst. 6, (1972), 164–192.

Corollary

Let $k, t \geq 2$. The set $\{n^t \mid n \geq 0\}$ is NOT k-recognizable.

S. Eilenberg, Automata, Languages, and Machines, 1974.

More examples

Minsky–Papert 1966

The set $\mathcal{P}$ of prime numbers is not k-recognizable.

A proof using the gap theorem :

- Since $n! + 2, \ldots, n! + n$ are composite numbers, $\mathbf{D}_{\mathcal{P}} = +\infty$
- Since $p_n \in (n \ln n, n \ln n + n \ln \ln n)$, $\mathbf{R}_{\mathcal{P}} = 1$

E. Bach, J. Shallit, Algorithmic number theory, MIT Press

Schützenberger 1968

No infinite subset of $\mathcal{P}$ can be recognized by a finite automaton.

- conversion: X is k-recognizable IFF X is k^n-recognizable
- Any ultimately periodic set is k-recognizable, for all $k \geq 2$
- Cobham 1969: Let $k, \ell \geq 2$ be two multiplicatively independent integers, i.e., if $\log k / \log \ell$ is irrational. If $X \subseteq \mathbb{N}$ is k-rec. AND ℓ-rec., then X is ultimately periodic.

T. Krebs, A more reasonable proof of Cobham's theorem `arxiv.1801.06704`

So there are sets that are recognizable

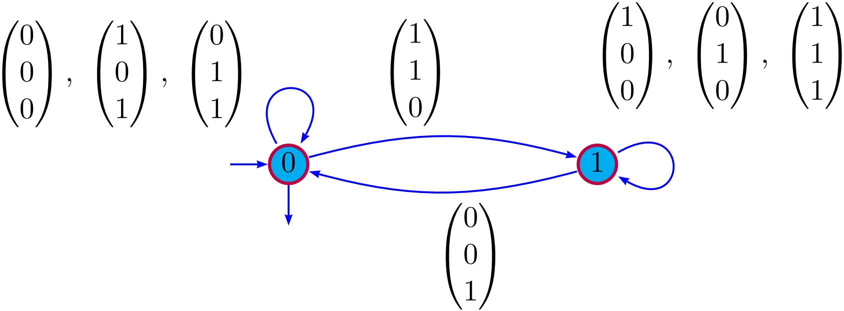
- in every base (ultimately periodic sets);
- for an equiv. class multiplicatively dependent bases;
- in no base at all.

V. Bruyère, G. Hansel, C. Michaux, R. Villemaire, Logic and p-recognizable sets of integers (1994)

In several dimensions

A set $X \subset \mathbb{N}^d$ is *k-recognizable*, if the set of base-k expansions of the d-tuples in X (*padded accordingly*) is accepted by some finite automaton (reading a d-tuple of digits at a time).

$\rightsquigarrow$ DFA reading l.s.d.f. for $+$, $X = \{(x, y, z) \in \mathbb{N}^3 \mid x + y = z\}$



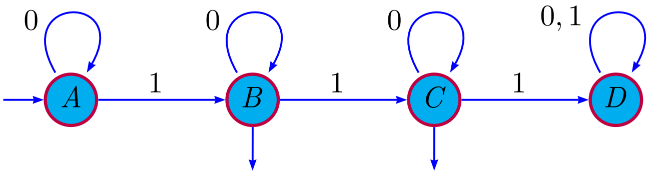
Such a DFA exists for all bases (no arbit. long carry propagation).

These operations cannot be "recognized" by finite automata

- multiplication $\{(x, y, z) \in \mathbb{N}^3 \mid x \cdot y = z\}$
- "general" base conversions, e.g., $\{(\mathrm{rep}_2(n), \mathrm{rep}_3(n)) \mid n \in \mathbb{N}\}$

Link with combinatorics on words

<u>Cobham 1972:</u> A set is k-recognizable if and only if its characteristic sequence $\mathbf{1}_X \in \{0,1\}^{\mathbb{N}}$ is k-automatic.



$$f : \begin{cases} A \mapsto AB \\ B \mapsto BC \\ C \mapsto CD \\ D \mapsto DD \end{cases} \qquad g : \begin{cases} A \mapsto 0 \\ B \mapsto 1 \\ C \mapsto 1 \\ D \mapsto 0 \end{cases}$$

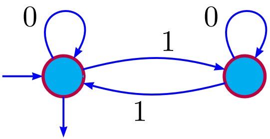
$$f^\omega(A) = ABBCBCCDBCCDCDDDBCCDCDDDCDDDDDDD\cdots$$

$$g(f^\omega(A)) = 01111110111010001110100010000000\cdots$$

$$X = \{1,2,3,4,5,6,8,9,10,12,16,17,18,20,24,\ldots\}$$

Link with combinatorics on words

$$0 \mapsto 01,\ 1 \mapsto 10$$



We get the so-called Thue–Morse word:

$$011010011001011010010110011010\cdots\ .$$

Link with combinatorics on words

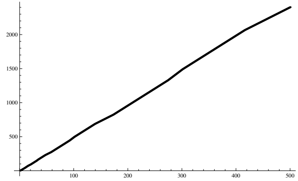
k-automatic sequences have *low complexity*.

Proposition (Cobham)

If $(x_n)_{n\geq 0}$ is k-automatic, then its factor complexity

$$p_x(n) := \#\{x_i \cdots x_{i+n-1} \mid i \geq 0\} \in \mathcal{O}(n)$$

J. Cassaigne, F. Nicolas, Factor complexity, Ch. 4 CANT'2010 (Cambridge).



$$p_{TM}(2^k + r) = \begin{cases} 3 \cdot 2^k + 4(r-1) & \text{, if } 1 \leq r \leq 2^{k-1}; \\ 4 \cdot 2^k + 2(r-1) & \text{, if } 2^{k-1} < r \leq 2^k. \end{cases}$$

Link with combinatorics on words

An application in number theory

Theorem (Adamczewski–Bugeaud 2007)

The complexity function of the k-ary expansion of every irrational algebraic number satisfies

$$\liminf_{n\to\infty} \frac{p(n)}{n} = +\infty.$$

Corollary

Aperiodic k-automatic numbers are transcendental.

The Thue–Morse number

$$\frac{0}{2} + \frac{1}{4} + \frac{1}{8} + \frac{0}{16} + \frac{1}{32} + \frac{0}{64} + \frac{0}{128} + \frac{1}{256} + \cdots \simeq 0.824907$$

is transcendental.

Link with combinatorics on words

A general question is "pattern avoidance" and "repetitions in words"

- A square : lekkerker
- An overlap : ananas
- A cube : lekkerkerkerk

Easiest theorem, a teaser

Over a binary alphabet, squares cannot be avoided.

M.R., Formal languages, Automata, Numeration Systems, ISTE 2014

Link with combinatorics on words

A general question is "pattern avoidance" and "repetitions in words"

- A square : lekkerker
- An overlap : ananas
- A cube : lekkerkerkerk

Easiest theorem, a teaser

Over a binary alphabet, squares cannot be avoided.

M.R., Formal languages, Automata, Numeration Systems, ISTE 2014

a

Link with combinatorics on words

A general question is "pattern avoidance" and "repetitions in words"

- A square : lekkerker
- An overlap : ananas
- A cube : lekkerkerkerk

Easiest theorem, a teaser

Over a binary alphabet, squares cannot be avoided.

M.R., Formal languages, Automata, Numeration Systems, ISTE 2014

ab

LINK WITH COMBINATORICS ON WORDS

A general question is "pattern avoidance" and "repetitions in words"

- A square : lekkerker
- An overlap : ananas
- A cube : lekkerkerkerk

EASIEST THEOREM, A TEASER

Over a binary alphabet, squares cannot be avoided.

M.R., Formal languages, Automata, Numeration Systems, ISTE 2014

aba

Link with combinatorics on words

Theorem (A. Thue 1906)

The Thue–Morse word is overlap free

$0 \mapsto 01,\ 1 \mapsto 10$

$$0110100110010110100101 10\cdots$$

In particular, this word is aperiodic.

Squares can be avoided on a 3-letter alphabet

$$\underbrace{011}_{a}\underbrace{01}_{b}\underbrace{0}_{c}\underbrace{011}_{a}\underbrace{0}_{c}\underbrace{01}_{b}\underbrace{011}_{a}\underbrace{01}_{b}\underbrace{0}_{c}\underbrace{01}_{b}\underbrace{011}_{a}0\cdots$$

Link with combinatorics on words

Theorem (A. Thue 1906)

The Thue–Morse word is overlap free

$0 \mapsto 01,\ 1 \mapsto 10$

$$0110100110010110100101 10 \cdots$$

In particular, this word is aperiodic.

Squares can be avoided on a 3-letter alphabet

$$\underbrace{011}_{a}\underbrace{01}_{b}\underbrace{0}_{c}\underbrace{011}_{a}\underbrace{0}_{c}\underbrace{01}_{b}\underbrace{011}_{a}\underbrace{01}_{b}\underbrace{0}_{c}\underbrace{01}_{b}\underbrace{011}_{a}0\cdots$$

Instead of integer base systems

Generalizations to other numeration systems / morphic sequences

- Zeckendorf expansions

$$49 = 1.34 + 0.21 + 1.13 + 0.8 + 0.5 + 0.3 + 1.2 + 0.1$$

$$\mathrm{rep}_F(49) = 10100010.$$

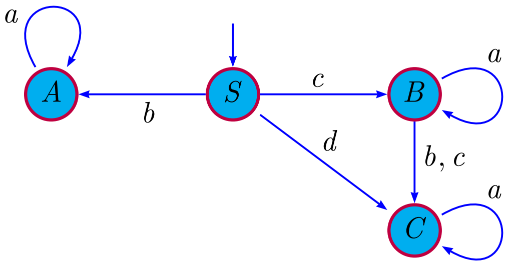
- Pisot numeration systems
- abstract numeration systems (enumerate a language)

Theorem (A. Maes, M.R. 2002)

An infinite word is morphic IFF it is S-automatic for some abstract numeration system S.

$$f : \begin{cases} S \mapsto SABC \\ A \mapsto A \\ B \mapsto BCC \\ C \mapsto C \end{cases} \qquad g : \begin{cases} S \mapsto 1 \\ A \mapsto 1 \\ B \mapsto 0 \\ C \mapsto 0 \end{cases}$$

$$g(f^{\omega}(S)) = 11001000010000000100000000\cdots$$



$\varepsilon, b, c, d, ba, ca, cb, cc, da, baa, \ldots$

Let's come back to Cobham's theorem (1969)

Recap

Let $k, \ell \geq 2$ be two multiplicatively independent integers.
If a set $X \subseteq \mathbb{N}$ is k-rec. and ℓ-rec., then X is ultimately periodic

Generalizations/extensions

- Pisot systems
- morphic sequences (F. Durand)
- multidimensional setting

Theorem (Cobham–Semenov)

Let $k, \ell \geq 2$ be two multiplicatively independent integers.
If a set $X \subseteq \mathbb{N}^d$ is k-rec. and ℓ-rec.,
then X is definable in the first order structure $\langle \mathbb{N}, + \rangle$.

$$=, 0, +, (\forall x),\ (\exists x), \neg, \rightarrow, \wedge, \vee, \leftrightarrow$$

you can add constants, multiplication by a constant, congruences

Remark ($d = 1$)

The subsets of $\mathbb{N}$ that are definable in $\langle \mathbb{N}, + \rangle$ are exactly the finite union of arithmetic progressions.

$\rightarrow$ The Thue-Morse word/set cannot be defined in $\langle \mathbb{N}, + \rangle$.

We can thus define subsets of $\mathbb{N}^d$

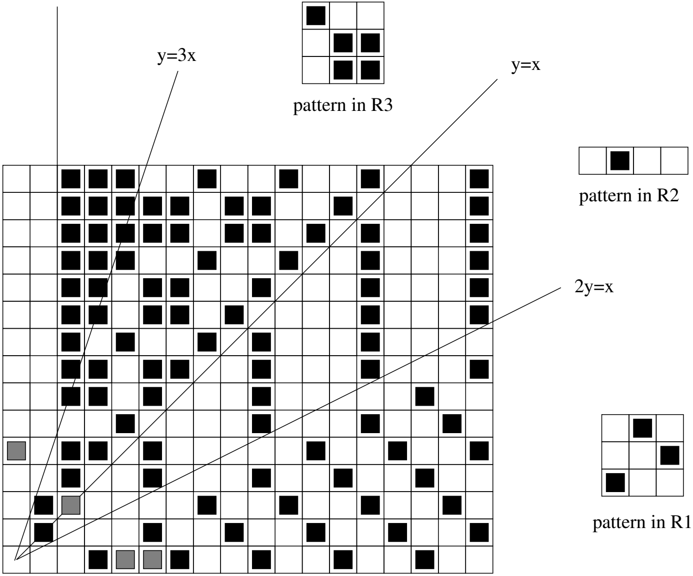
Presburger definable sets

A formula $\varphi(x_1, \ldots, x_d)$ with d free variables,

$$\{(n_1, \ldots, n_d) \in \mathbb{N}^d \mid \langle \mathbb{N}, + \rangle \models \varphi(n_1, \ldots, n_d)\}$$

$\varphi(x_1, x_2) \equiv \rho_1(x_1, x_2) \vee \rho_2(x_1, x_2) \vee \rho_3(x_1, x_2) \vee \rho_4(x_1, x_2) \vee \phi(x_1, x_2)$
where

$$\begin{aligned}
\rho_1(x_1, x_2) &\equiv (2x_2 < x_1) \wedge (x_1 + x_2 \equiv_3 0)\,, \\
\rho_2(x_1, x_2) &\equiv (2x_2 \geq x_1) \wedge (x_2 < x_1) \wedge (x_1 \equiv_4 1)\,, \\
\rho_3(x_1, x_2) &\equiv \underbrace{(x_2 > x_1) \wedge (x_2 < 3x_1)}_{\text{a region}} \wedge \underbrace{((2x_1 + x_2 \equiv_3 1) \vee (x_1 + x_2 \equiv_3 0))}_{\text{a pattern}}\,, \\
\rho_4(x_1, x_2) &\equiv (x_2 \geq 3x_1) \wedge (x_1 \geq 2)\,, \\
\phi(x_1, x_2) &\equiv (x_1 = 0 \wedge x_2 = 4) \vee (x_1 = 2 \wedge x_2 = 2) \vee (x_1 = 4 \wedge x_2 = 0) \\
&\quad \underbrace{\vee (x_1 = 5 \wedge x_2 = 0)}_{\text{a few isolated points}}\,.
\end{aligned}$$


y=3x
pattern in R3
y=x
pattern in R2
2y=x
pattern in R1

Presburger arithmetic ; decidable theory (quantifiers elimination)

A less trivial example (Frobenius' problem)

Chicken McNuggets can be purchased only in 6, 9, or 20 pieces.
The largest number of nuggets that cannot be purchased is 43.

$$(\forall n)(n > 43 \to (\exists x, y, z \geq 0)(n = 6x + 9y + 20z))$$
$$\wedge\neg((\exists x, y, z \geq 0)(43 = 6x + 9y + 20z))\,.$$

There is an algorithm with output `TRUE/FALSE`.

Let's come back to chicken Mc nuggets and Thue–Morse

- The first one is given by a sentence in $\langle \mathbb{N}, + \rangle$;
- For the second one, we need an extra function V_2.

$V_k(n)$ is the largest power of k dividing n (close to p-adic valuation).

A set $X \subseteq \mathbb{N}^d$ is k-recognizable/k-automatic if and only if it definable in $\langle \mathbb{N}, +, V_k \rangle$.

Theorem

For all $k \geq 2$, the first order theory of $\langle \mathbb{N}, +, V_k \rangle$ is decidable.

Büchi's proof: from formula to finite automata
emptiness and universality are decidable.

Let's come back to chicken Mc nuggets and Thue–Morse

- The first one is given by a sentence in $\langle \mathbb{N}, + \rangle$;
- For the second one, we need an extra function V_2.

$V_k(n)$ is the largest power of k dividing n (close to p-adic valuation).

A set $X \subseteq \mathbb{N}^d$ is k-recognizable/k-automatic if and only if it definable in $\langle \mathbb{N}, +, V_k \rangle$.

Theorem

For all $k \geq 2$, the first order theory of $\langle \mathbb{N}, +, V_k \rangle$ is decidable.

Büchi's proof: from formula to finite automata
emptiness and universality are decidable.

Shallit *et al.*

Theorem

Let $k \geq 2$. If one can express a property of a k-automatic sequence x using quantifiers, logical operations, integer variables, $+$, -, indexing into x (i.e., access to $x(n)$), and comparison of integers or elements of x, then this property is decidable.

Is x ultimately periodic?

$$(\exists N)(\exists p > 0)(\forall i \geq N) x(i) = x(i+p).$$

Does it have an overlap?

$$(\exists i \geq 0)(\exists \ell \geq 1)(\forall j \in \{0, \ldots, \ell\})(x(i+j) = x(i+\ell+j)).$$

D. Goč, D. Henshall, J. Shallit, Automatic theorem-proving in combinatorics on words (2013)

Hamoon Mousavi, Automatic Theorem Proving in Walnut.

Walnut : `https://cs.uwaterloo.ca/~shallit/walnut.html`

```
(base) mrigo@math-mrtpx1:~/Walnut/Walnut_old/Walnut/bin$ java Main.prover
eval test "~(Ei El Aj (((l>0) & (j<l)) => ( (T[i+j]=T[(i+j)+l]) & (T[i]=T[((i+l)+l)]))))":
l>0 has 2 states: 8ms
 j<l has 2 states: 0ms
  (l>0&j<l) has 2 states: 0ms
   T[(i+j)]=T[((i+j)+l)] has 12 states: 30ms
    T[i]=T[((i+l)+l)] has 6 states: 5ms
     (T[(i+j)]=T[((i+j)+l)]&T[i]=T[((i+l)+l)]) has 72 states: 5ms
      ((l>0&j<l)=>(T[(i+j)]=T[((i+j)+l)]&T[i]=T[((i+l)+l)])) has 97 states: 11ms
       (A j ((l>0&j<l)=>(T[(i+j)]=T[((i+j)+l)]&T[i]=T[((i+l)+l)]))) has 1 states: 40ms
        (E l (A j ((l>0&j<l)=>(T[(i+j)]=T[((i+j)+l)]&T[i]=T[((i+l)+l)])))) has 1 states: 0ms
         (E i (E l (A j ((l>0&j<l)=>(T[(i+j)]=T[((i+j)+l)]&T[i]=T[((i+l)+l)]))))) has 1 states: 0ms
          ~(E i (E l (A j ((l>0&j<l)=>(T[(i+j)]=T[((i+j)+l)]&T[i]=T[((i+l)+l)]))))) has 1 states: 0ms
total computation time: 108ms
```

- You can do much more, letting some free variables, enumeration
- used in more than 50 papers

Another recent example (Shallit `arXiv.2112.13627`)
the number of representations of n as the sum of two elements of a given set:

$$R_1^{(A)}(n) = \#\{(x,y) \in \mathbb{N}^2 : x, y \in A \wedge x + y = n\}$$

$$R_2^{(A)}(n) = \#\{(x,y) \in \mathbb{N}^2 : x, y \in A \wedge x + y = n \wedge x < y\}$$

$$R_3^{(A)}(n) = \#\{(x,y) \in \mathbb{N}^2 : x, y \in A \wedge x + y = n \wedge x \leq y\}$$

Sárközy asked whether there exist two sets of positive integers A and B, with infinite symmetric difference, for which $R_i^{(A)}(n) = R_i^{(B)}(n)$ for all sufficiently large n, $i = 1, 2, 3$.

G. Dombi (2002), Y.-G. Chen and B. Wang (2003).

- L. Schaeffer: the set of occurrences of abelian squares

$$(0110)(1100)$$

in the paperfolding word

$$(0 \star 1)^\omega = 00100110001101100010011100110110 \cdots$$

is not 2-recognizable, even though the paperfolding word is 2-automatic. Hence such a property cannot be captured by a formula in $\langle \mathbb{N}, +, V_2 \rangle$.
- F. Klaedtke: Nested quantifiers could, in the worst-case, lead to a tower of exponentials.

Some other related work on Sturmian words

- P. Hieronymi, D. Ma, R. Oei, L. Schaeffer, C. Schulz, J. Shallit, *Decidability for Sturmian words* (2021)
- P. Hieronymi, Alonza Terry Jr, *Ostrowski numeration systems, addition, and finite automata*, N. D. J. Form. Log. (2014)
- Reed Oei, Eric Ma, Christian Schulz, and Philipp Hieronymi, Pecan

Pecan: An Automated Theorem Prover

Pecan is an automated theorem prover. You can view the manual (which is currently rather incomplete) here. Below, you can enter a program and click "Run" to try out Pecan; there are several example programs you can try out. There are some limitations: some features of Pecan are not available, your programs must be under 10^5 characters, and your programs will time out after 5 minutes. If that's too limiting, you should look into installing Pecan on your own computer. Instructions are available at the repository: https://github.com/ReedOei/Pecan. If you use Pecan in your research, please cite this paper

Enter a Pecan program below:

Basic Arithmetic | Chicken McNuggets | Plotting | Thue-Morse Word | Sturmian Words

```
#import("SturmianWords/ostrowski_defs.pn")
#load("SturmianWords/automata/antisquare.aut", "hoa", antisquare(a,i,n))
#load("SturmianWords/automata/eventually_periodic.aut", "hoa", eventually_periodic(a, p))
#load("SturmianWords/automata/palindrome.aut", "hoa", palindrome(a, i, n))

// See the GitHub repo for more: https://github.com/ReedOei/SturmianWords

Let a be bco_standard.
Let i,n,m,p be ostrowski(a).

Theorem ("Sturmian words are not eventually periodic", {
    forall a, p. !@no_simplify[eventually_periodic(a, p)]
}).

Theorem ("Sturmian words contain palindromes of every length.", {
    forall a, n. if n > 0 then exists i. palindrome(a, i, n)
}).

Theorem ("There are finitely many antisquares", {
    forall a. exists m. forall i,n. if antisquare(a,i,n) then n <= m
}).
```