Abstract. I survey some of the connections between formal languages and number theory. Topics discussed include applications of representation in base \( k \), representation by sums of Fibonacci numbers, automatic sequences, transcendence in finite characteristic, automatic real numbers, fixed points of homomorphisms, automaticity, and \( k \)-regular sequences.

Key words. finite automata, automatic sequences, transcendence, automaticity.

AMS(MOS) subject classifications. Primary 11B85, Secondary 11A63 11A55 11J81.

1. Introduction. In this paper, I survey some interesting connections between number theory and the theory of formal languages. This is a very large and rapidly growing area, and I focus on a few areas that interest me, rather than attempting to be comprehensive. (An earlier survey of this area, written in French, is [1].) I also give a number of open questions.

Number theory deals with the properties of integers, and formal language theory deals with the properties of strings. At the intersection lies

(a) the study of the properties of integers based on their representation in some manner — for example, representation in base \( k \); and

(b) the study of the properties of strings of digits based on the integers they represent.

An example of a theorem of type (a) — perhaps the first significant one — is the famous theorem of Kummer [60, pp. 115–116], which states that the exponent of the highest power of a prime \( p \) which divides the binomial coefficient \( \binom{n}{m} \) is equal to the number of “carries” when \( m \) is added to \( n - m \) in base \( p \).

For type (b) I do not know a theorem as fundamental as Kummer’s. But here is a little problem that some may find amusing. Call a set of strings sparse if, as \( n \to \infty \), it contains a vanishingly small fraction of all possible strings of length \( n \). Then can one find a sparse set \( S \) of strings of 0’s and 1’s such that every string of 0’s and 1’s can be written as the concatenation of two strings from \( S \)? One solution is to let \( S \) be the set of all strings of 0’s and 1’s such that the number of 1’s is a sum of two squares. Then by a famous theorem in number theory — Lagrange’s theorem — every number \( n \) is the sum of four squares, so every string of 0’s and 1’s is a concatenation of two strings chosen from \( S \). The sparseness of \( S \) follows from an estimate in sieve theory [38]. Further examples of theorems of type (b) can be found in Section 8.1.
It may be objected that studying the formal language aspects of number theory is somewhat artificial, in the sense that it depends on choosing one particular representation — such as representation in base 2 — and that there is no reason to choose base 2 over any other base. For example, recall the famous objection of Hardy to certain kinds of digital problems:

These are odd facts, very suitable for puzzle columns and likely to amuse amateurs, but there is nothing in them which appeals much to a mathematician. The proofs are neither difficult nor interesting — merely a little tiresome. The theorems are not serious; and it is plain that one reason (though perhaps not the most important) is the extreme speciality of both the enunciations and the proofs, which are not capable of significant generalization. [46, p. 105]

I offer four answers to Hardy’s objection. First, we attempt to make our theorems as general as possible. For example, we can try to prove theorems for all bases $k$ rather than just a single base. Second, sometimes some bases do occur naturally in problems, and base 2 is one of them; see Section 4. Third, the area has proved to have many applications; perhaps the most dramatic examples are the recent simple proofs of transcendence in finite characteristic by Allouche and others; see Section 5. Finally, the area is “natural”, and I submit as evidence the fact that many good mathematicians throughout history have worked in it, including Kummer, Lucas, and Carlitz.

2. Notation. I begin with some notation for formal languages, for which a good reference is the book of Hopcroft and Ullman [49].

Let $\Sigma$ be a finite list of symbols, or alphabet, and let $\Sigma^*$ denote the free monoid over $\Sigma$, that is, the set of all finite strings of symbols chosen from $\Sigma$, with concatenation as the monoid operation. Thus, if $\Sigma = \{0, 1\}$, then

$$\Sigma^* = \{\epsilon, 0, 1, 00, 01, 10, 11, 000, \ldots\},$$

where $\epsilon$ is the notation for the empty string. A formal language, or just language, is defined to be any subset of $\Sigma^*$.

Let $L, L_1, L_2$ be languages. We define the concatenation of languages as follows:

$$L_1L_2 = \{x_1x_2 : x_1 \in L_1, x_2 \in L_2\}.$$

\[^1\]The two problems he cited as examples were (a) show that 8712 and 9801 are the only four-digit numbers which are nontrivial integral multiples of their reversals and (b) show that 153, 370, 371, and 407 are the only integers $> 1$ which are equal to the sum of the cubes of their decimal digits. Today, digital problems continue to attract attention and criticism; see, for example, [35].