Unique Decipherability in the Monoid of Languages: an Application of Rational Relations [⋆]

Christian Choffrut[1] and Juhani Karhumäki[2]

[1] L.I.A.F.A., Université Paris 7, 2 Pl. Jussieu – 75 251 Paris Cedex – France
[2] Dept. of Math. and TUCS, University of Turku, 20014, Turku – Finland

Abstract. We attack the problem of deciding whether a finite collection of finite languages is a code, that is, possesses the unique decipherability property in the monoid of finite languages. We investigate a few subcases where the theory of rational relations can be employed to solve the problem. The case of unary languages is one of them and as a consequence, we show how to decide for two given finite subsets of nonnegative integers, whether they are the n-th root of a common set, for some $n \geq 1$. We also show that it is decidable whether a finite collection of finite languages is a Parikh code, in the sense that whenever two products of these sets are commutatively equivalent, so are the sequences defining these products. Finally, we consider a nonunary special case where all finite sets consist of words containing exactly one occurrence of the specific letter.

Key words: unique decipherability, finite automata, regular languages

1 Introduction

The question whether or not a given morphism $h : \Sigma^* \to \Delta^*$ is injective, that is, whether or not the encoded message can be uniquely decoded, is fundamental in the theory of message transmission. More precisely, the problem asks whether or not the given set of code words possesses the *unique decipherability property.* The issue for finite sets X was already affirmatively answered in 1950 with the so-called Sardinas and Patterson algorithm, see [19]. Later, it was extended, via syntactic monoids, to all rational sets, see [3] and its complexity was analyzed in [7].

A particularly illustrative way of solving this problem is to construct a finite two-tape automaton for all double representations (factorizations) of message sequences and to reduce the testing to the emptiness problem for rational relations. In [5] the same approach was used to show how to decide whether or not, given two finite sets X and Y, the two monoids X^* and Y^* they generate

[⋆] This research was done during the second author's visit at the LIAFA, CNRS UMR 7089, Université Paris Diderot, Novembre–December 2008. The second author was supported by the Academy of Finland under the grant 126211.

are isomorphic. This, however, works for finite sets only and the general case of rational sets is still open.

The unique decipherability problem can be formulated for any associative algebra. It has been studied, e.g., in the theory of trees, [15], or in the case of multivalued encodings, [18]. In the case of the monoid of finite languages, amazingly, little seems to be known, a splendid exception being the fact that the set of prefix languages under the operation of concatenation product is free, that is, any collection of finite prefix sets is a code, see [17]. A partial explanation to the lack of such results was revealed recently when it was shown in a number of papers, how powerful or difficult language equations are. To mention a few examples, it is shown in [12] that even the question whether or not, for given finite sets A, B, C, D, E, F, the equation $AB^iC = DE^iF$ holds for all $i \geq 0$, is recursively undecidable, see also [14]; the maximal set commuting with a given finite set A need not be recursive, see [13]; or the fact that two given finite sets A and B are conjugate, i.e., that there exists a set Z such that $AZ = ZB$ holds, is known to be decidable only in the case of bifix sets, see [4].

Nonetheless, there are two related research topics which have been studied in the literature. Research on decomposition of rational languages was initiated already in Conway's book, see [8]. Later, this research was pursued in, e.g., [11], where also prime decompositions are defined. In another direction, unambiguous products of languages were studied in [1] and [16]. As we shall see, it is this, or more precisely its negation, the ambiguity, which makes our problems difficult.

As already hinted, our goal is to tackle the unique decipherability decision problem for a finite collection of finite languages. More precisely, we want to apply the theory of rational relations to solve a few special cases, even if the general problem seems to be very much beyond the reach of our tools.

The structure of our presentation is as follows. In section 2, we fix the terminology, recall the basic tools we are using and prove a simple case of our problem to be decidable, namely we show how to decide, given two finite sets of integers, whether or not some of their nonnegative powers (actually subset sums since we are working in the additive structure) coincide. The problem can be formulated as a natural decision question in additive number theory: decide whether or not two finite sets of numbers are the (additive) roots of some common set.

In section 3 we extend the above proof to the unique decipherability property for unary languages, and later in section 4 a further extension is introduced where so-called Parikh codes are considered. We say that a collection of finite languages is a Parikh code, in the sense that whenever two products of these sets are commutatively equivalent, so are the two sequences defining these products. Being a Parikh code is necessary but not sufficient for a set of finite languages to be a code. Finally, in section 5, another approach of using rational languages is introduced. It allows us to consider the case when all words of all sets of the collection contain one and only one occurrence of a fixed letter. In this case, we can decide, given two such sets, whether or not some of their powers coincide; we also outline methods of deciding whether a given collection of special finite languages is uniquely decipherable, that is a code.

2 Preliminaries and an example

In this section we fix the terminology, recall basic results and give a simple example. For a general reference to the field, we suggest [2] and [3].

We denote by Σ a finite alphabet, and by Σ^* the free monoid it generates. Elements and subsets of Σ^* are called *words* and *languages*, respectively. Other monoids considered here are submonoids of the additive monoid of nonnegative integers $\mathbb{N}$, and Cartesian products of these and of Σ^*. Our main concern is on finite languages and finite subsets of $\mathbb{N}$ and $\mathbb{N}^k$, and basic tools to deal with those rely on properties of rational, i.e., semilinear sets. We recall that a *rational* subset of a monoid is a subset obtained from finite subsets by applying finitely many times the operations of set union, product and Kleene iteration, also known as the star-operation. The result of applying the Kleene operation to the subset X is denoted by X^*. We shall also use the notation $X^+ = XX^* = X^*X$. A *linear* set, in turn, is a set of the form

$$\{a + \lambda_1 b_1 + \cdots \lambda_p b_p \mid \lambda_i \in \mathbb{N}, \text{ for } i = 1, \ldots, p\}$$

where $p \geq 0$ and $a, b_i \in \mathbb{N}^k$ for $1 \leq i \leq p$. A *semilinear* set, is a finite union of linear sets.

We recall that a subset of Σ^* (resp. the Cartesian product $\Sigma^* \times \Delta^*$) is rational if and only if it is recognized by some finite one-tape (resp. two-tape) automaton.

It is quite straightforward to check that the family of rational sets of $\mathbb{N}^k$ is identical to the family of semilinear sets. A fundamental, nontrivial property due to Ginsburg and Spanier is that this family is closed under complement. More precisely, we have, see [10], also [9]

Theorem 1. *The family of semilinear sets is an effective Boolean algebra.*

This means that not only the family is closed under the Boolean operations, but that from a specification of two semilinear sets we can compute a specification for the complement and the intersection. This theorem plays a crucial role in our considerations, as well as some other closure properties of semilinear sets such as the closure under morphic images and projections.

Now we state our basic problems. Let M be an associative algebra, that is an algebra with a single associative operation. A subset $X \subseteq M$ is *uniquely decipherable* in M if , whenever

$$x_1 \cdots x_p = y_1 \cdots y_q, \quad \text{with } x_i, y_j \in X$$

holds, then necessarily we have

$$p = q \ \text{ and } x_i = y_i, \quad \text{for } i = 1, \ldots, p\ .$$

This leads to the following decision issue: the UNIQUE DECIPHERABILITY PROBLEM for M, (UD-problem for short) asks to decide whether or not a given finite subset of M can be uniquely deciphered.

Two related simpler problems are:

The Power equality problem for M (PE for short) asks whether or not, for two given finite subsets X and Y of M, some of their powers coincide, that is whether or not $X^n = Y^m$ holds for some $n, m \geq 1$.

The Common root problem for M (CR for short) asks whether or not for two given finite subsets X and Y of M, they are distinct n-th powers of a set for a certain n, that is, whether or not $X^n = Y^n$ holds for some $n \geq 1$.

We conclude this section by illustrating our techniques with a simple example, which, we believe, is a natural problem in additive theory of numbers. By convention, we keep using the multiplicative notation though we work with the additive structure of the integers. In particular if X and Y are two subsets of integers, then XY stands for all the sums of the form $x + y$ with $x \in X$ and $y \in Y$ and the notation X^n stands for the expression $\overbrace{X + \cdots + X}^{n \text{ times}}$. E.g., with $X = \{0, 1, 2\}$ we have $X^2 = X + X = \{0, 1, 2, 3, 4\}$.

Theorem 2. *Given two finite subsets $X, Y \subseteq \mathbb{N}$, it is recursively decidable whether or not the equality $X^n = Y^m$ holds for some integers $n, m \geq 1$.*

Proof. We define a subset of $\mathbb{N}^3$

$$Z = \mathbb{N} \times (X \times 1)^+ \setminus (1 \times Y)^+ \times \mathbb{N} \,. \tag{1}$$

If $\pi_{1,3}$ is the projection of $\mathbb{N}^3$ onto $\mathbb{N}^2$ defined by $\pi_{1,3}(x, y, z) = (x, z)$, then we claim that the following holds

$$(n, m) \in \pi_{1,3}(Z) \text{ if and only if } X^n \not\subseteq Y^m \,.$$

Indeed, this follows from the construction: if $(n, m) \in \pi_{1,3}(Z)$ then there exist elements $x_1, \ldots, x_n \notin Y^m$, that is $X^n \not\subseteq Y^m$, and conversely. Consequently,

$$(n, m) \notin \pi_{1,3}(Z) \text{ if and only if } X^n \subseteq Y^m \,. \tag{2}$$

It follows that the set of pairs (n, m) satisfying the condition (2), which defines a rational relation on $\mathbb{N}^2$, characterizes the pairs of integers for which $X^n \subseteq Y^m$ holds. Similarly, the relation characterizing the set of pairs for which $X^n \supseteq Y^m$ holds, is rational. As the intersection is again rational and effective, the equality $X^n = Y^m$ holds for some integers n and m if and only if this rational relation is nonempty. □

Theorem 2, rather its proof, has the following immediate consequences

Corollary 1. *The common root problem of finite subsets for $\mathbb{N}$ is decidable.*

Corollary 2. *It is recursively decidable whether or not two finite subsets X and Y of $\mathbb{N}$ are ultimately equivalent, i.e., whether or not there exists an integer N such that*

$$X^n = Y^n$$

holds for $n \geq N$.

Of course, in either of the above corollaries, in order to get the answer “yes”, the maximal and minimal numbers of the two subsets must necessarily coincide. This condition holds for the second largest and smallest elements as well. For the others, the ambiguity comes into play, and makes the problem difficult to analyse.

A simple example from [6] showing that a square root of a set may not be unique is as follows: take $X = \{0, 2, 3, 7, 10, 12, 14, 15\}$ and $Y = \{0, 2, 3, 7, 12, 13, 14, 15\}$. Then $X^2 = Y^2 = [0, 30] \setminus \{1, 8, 11, 23\}$.

3 The unary case

In this section we extend our considerations of the previous section to cover the UD-problem for unary languages, that is we prove

Theorem 3. *The unique decipherability problem is decidable for unary languages*

Proof. Let $\Xi = \{X_1, \ldots, X_k\}$ be a collection of finite unary languages. We have to decide whether or not there exist two sequences $i_1, \ldots, i_p$ and $j_1, \ldots, j_q$ such that

$$X_{i_1} \ldots X_{i_p} = X_{j_1} \ldots X_{j_q} \text{ with } X_{i_\alpha}, X_{j_\beta} \in \Xi \text{ and } i_1, \ldots, i_p \neq j_1, \ldots, j_q$$

We fix some notations. For $i = 1, \ldots, k$ let $\pi_i : \mathbb{N}^k \to \mathbb{N}$ be the projection onto the i-th component. Furthermore, let $e_i \in \mathbb{N}^k$ be the vector having 1 in position i and 0 everywhere else. We modify the expression (1) in the proof of Theorem 2 by setting

$$Z = \mathbb{N}^k \times \big(\bigcup_{j=1}^{k} (X_j \times e_j)\big)^+ \setminus \big(\bigcup_{j=1}^{k} (e_j \times X_j)\big)^+ \times \mathbb{N}^k$$

Then we have $(z_1, x, z_2) \in Z$ with $z_1, z_2 \in \mathbb{N}^k$ and $x \in \mathbb{N}$ if and only if $x \in X(z_1) \setminus X(z_2)$, where $X(z_\alpha) = \prod_{i=1}^{k} X_i^{\pi_i(z_\alpha)}$, for $\alpha = 1, 2$. The last part of the proof mimics that of Theorem 2, the only additional feature being that at the end we have to intersect with the following rational subset of $\mathbb{N}^{2k}$

$$\{(x, y) \in \mathbb{N}^{2k} \mid x, y \in \mathbb{N}^k, x \neq y\}$$

□

4 Unique Parikh decipherability

Our method allows us to go still a step further. In the previous section we were able to solve our problem for all unary languages. Here, we can solve the general problem at the price of substituting the condition of unique Parikh decipherability to that of unique decipherability.

We say that a collection $\varXi = \{X_1, \ldots, X_k\}$ of finite languages possesses the unique Parikh decipherabilty property if the condition

$$X_{i_1} \ldots X_{i_p} \sim_c X_{j_1} \ldots X_{j_q}$$

implies that

$$i_1 \ldots i_p \sim_c j_1 \ldots j_q$$

holds, where $\sim_c$ is used to denote the commutative equivalence of languages or words, respectively. We can formulate

Theorem 4. *The unique Parikh decipherability is recursively decidable for finite collections of finite languages in $\varSigma^*$.*

Sketch of the proof This result is actually a generalization of Theorem 2. Indeed, it can be shown that this latter theorem holds for subsets of $\mathbb{N}^m$ for arbitrary $m \geq 1$, not only for subsets of $\mathbb{N}$. Now, set $\varSigma = \{a_1, \ldots, a_m\}$ and consider the morphism $\phi : \varSigma^* \to \mathbb{N}^m$ which maps each word $w \in \varSigma^*$ to the m-tuple $(|w|_{a_1}, \ldots, |w|_{a_m})$ where $|w|_{a_i}$ denotes the number of occurrences of the letter a_i in w. Then our problem reduces to the unique decipherability problem for the finite collection $\phi(X_1), \ldots, \phi(X_k) \subseteq \mathbb{N}^m$.

□

It is worthwhile emphasizing that all our results reported so far are based on strong closure properties of rational relations in the commutative case, in particular the closure under complement, and as a consequence under intersection. This leads to the following comments. First, the complexity of our algorithms are quite high, particularly due the the operation of complementation. Second, there is no hope to extend our approach at least in a naive way, since rational relations over free monoids with more than one generator are not closed under intersection. On the other hand, we do not see how to construct complicated examples of collections of finite sets which would satisfy the unique Parikh decipherability, but would not satisfy the unique decipherability property.

Two last observations. It can be readily shown as hinted in the proof of the previous theorem, that Theorem 2 carries over from $\mathbb{N}$ to $\mathbb{N}^m$ for $m \geq 1$ and actually also to $\mathbb{Z}^m$. Also, since the commutative image of a rational subset of a free monoid is a semilinear set of $\mathbb{N}^m$, Theorem 4 also holds for a finite collections of rational, not only finite, languages of a free monoid.

5 A special nonunary case

In this section we consider the UD-problem, for languages over a general alphabet, but in quite a restricted setting, namely we assume that the finite languages are subsets of

$$(\varSigma \setminus \{b\})^* b (\varSigma \setminus \{b\})^* \text{ for some fixed letter } b \in \varSigma. \tag{3}$$

We show that the problem is decidable in this case. The solution is based on classical closure properties of rational languages under the Boolean operations and the substitutions.

Theorem 5. *The unique decipherability problem is decidable for any collection of finite sets of the form (3).*

Proof The proof is based on closure properties of rational languages.

We start by solving a different problem. Given $I = \{1, \dots, N\}$ and two finite collections of finite languages of type (3)

$$\begin{array}{ll} X_i & = \{x_{i,s} \mid s = 1, \dots, s(i)\} \\ \text{and} & \\ Y_i & = \{y_{i,r} \mid r = 1, \dots, r(i)\} \end{array}$$

for $i \in I$, determine whether or not

$$X_w = Y_w \tag{4}$$

holds for some $w = i_1 \cdots i_k \in I^+$. The notation X_w stands for the product

$$X_w = X_{i_1} \cdots X_{i_k}.$$

The idea is to determine the set of $w \in I^+$ such that

$$X_w \subseteq Y_w$$

and to test whether or not its intersection with the set of w's such that $Y_w \subseteq X_w$ holds, is not empty. We show that these two languages are recognized by finite automata, therefore that the test is effective.

It suffices to prove the claim for the set of words for which the inclusion $X_w \subseteq Y_w$ holds. We define an automaton $\mathcal{A}$ as follows. Its states are words over $\Sigma \setminus \{b\}$ and inverses (considering the free monoid as embedded in the free group), the empty word 1 being both the initial and final state. Its alphabet is the set

$$J = \{(i, s) \mid i = 1, \dots, N \text{ and } s = 1, \dots, s(i)\}$$

Intuitively, the word $(i_1, s_1) \cdots (i_k, s_k)$ takes the initial state to state $\alpha \in J^*$ if

$$x_{i_1,s_1} \cdots x_{i_1,s_k} = y_{i_1,r_1} \cdots y_{i_1,r_k} \alpha$$

holds for some $r_1, \cdots, r_k$ and to state β^{-1}, with $\beta \in J^*$ if

$$x_{i_1,s_1} \cdots x_{i_1,s_k} = y_{i_1,r_1} \cdots y_{i_1,r_k} \beta$$

holds for some $r_1, \cdots, r_k$. Formally, the transitions are defined as follows: For two states α and β and $x_{i,s} \in X_i$, if

$$\alpha x_{i,s} = y_{i,r} \beta, \text{ for some } y_{i,r} \in Y_i,$$

there is a transition

$$\alpha \xrightarrow{(i,s)} \beta \tag{5}$$

Actually, there are four different possibilities depending on whether α and/or β are words, as above, or their formal inverses. These modifications are obvious. Clearly, $\mathcal{A}$ is a well defined (due to the form of sets X_i and Y_i), finite, but nondeterministic automaton. Intuitively, it checks those products of X-words which can be decomposed into Y-words as well. Note that, again according to the form of our words, such decompositions are of the same length.

Define the letter-to-letter substitution $\pi : J^* \mapsto I^*$ by posing $\pi(i,s) = i$. Let L be the language recognized by the automaton and set $P = \pi^{-1}\pi(L)$. Then the set of w's such that $X_w \not\subseteq Y_w$ holds is equal to $\pi(P \backslash L)$. Since all these operations are effective and involve rational languages, the problem (4) is decidable.

We now turn to the proof of Theorem 5. It is obtained by modifying the above construction. The noncode property, that is the existence of $w \neq w'$ such that $X_w = X_{w'}$ holds, is equivalent to the fact that the set

$$\{(w,w') \in (I \times I)^* \mid w \neq w', X_w \subseteq X_{w'} \text{ and } X_{w'} \subseteq X_w\}$$

is nonempty. Set $T = \{(w,w') \in (I \times I)^* \mid w \neq w', X_w \subseteq X_{w'}\}$ and $T' = \{(w',w) \mid (w,w') \in T\}$. Then it suffices to verify that $T \cap T' \neq \emptyset$. It remains to prove that T is a rational subset of the free monoid $(I \times I)^*$ since in that case T' is clearly also rational as well as the intersection $T \cap T'$.

Equivalently, we prove that the set

$$\begin{aligned} &\{(w,w') \in (I \times I)^* \mid w \neq w', X_w \not\subseteq X_{w'}\} \\ = &\{(w,w') \in (I \times I)^* \mid w \neq w'\} \cap \{(w,w') \in (I \times I)^* \mid X_w \not\subseteq X_{w'}\} \end{aligned} \tag{6}$$

is rational. The first term of the intersection is rational. Concerning the second term, consider the automaton $\mathcal{B}$ whose alphabet is $J \times I$, set of states, initial and final states are identical to those of $\mathcal{A}$ and whose transitions are of the form

$$\alpha \xrightarrow{((i,s),i')} \beta$$

if

$$\alpha x_{i,s} = x_{i',r}\beta,$$

holds for some $i' \in \{1, \cdots, N\}$ and $r \in \{1, \cdots, s(i')\}$ and of the three other forms depending on whether or not the states are inverses of words in the free monoid. Then the second term of expression 6 is the projection onto $(I \times I)^*$ of the language recognized by $\mathcal{B}$ which completes the proof. □

The first part of the previous proof can be reformulated as

Corollary 3. *For two finite languages $X, Y \subseteq (\Sigma\backslash\{b\})^* b (\Sigma\backslash\{b\})^*$, with $b \in \Sigma$, it is recursively decidable whether or not there exists an n such that $X^n = Y^n$ holds true.*

The above deserves a few comments. It relies very much on the special form of the X-sets, that is, on the fact that there is just one "marker" symbol in all words of X sets. On the other hand, the marker need not be a symbol. That is to say that X should satisfy the conditions that each word in X contains exactly one occurence of a word u and each occurence of X^2 contains exactly two occurences of u.

References

1. M. Anselmo and A. Restivo. Factorizing languages. In *IFIP Congress (1)*, pages 445–450, 1994.
2. J. Berstel. *Transductions and context-free languages.* B. G. Teubner, 1979.
3. J. Berstel and D. Perrin. *The theory of codes*, volume 117. Academic Press, 1985.
4. J. Cassaigne, J. Karhumäki, and P. Salmela. The conjugacy of biprefix sets. to appear.
5. C. Choffrut, T. Harju, and J. Karhumäki. A note on decidability questions on presentations of word semigroups. *Theor. Comput. Sci.*, 183(1):83–92, 1997.
6. C. Choffrut and J. Karhumäki. On Fatou properties of rational languages. In C. Martin-Vide and V. Mitrana, editors, *Where Mathematics, Computer Science, Linguistics and Biology Meet*, pages 227–235. Kluwer, Dordrecht, 2000.
7. M. Chrobak and W. Rytter. Unique deciperability for partially commutative alphabet (extended abstract). In J. Gruska, B. Rovan, and J. Wiedermann, editors, *MFCS*, volume 233 of *Lecture Notes in Computer Science.* Springer, 1986.
8. J.H. Conway. *Regular algebras and finite machines.* Chapman and Hall, 1974.
9. S. Eilenberg and M.-P. Schützenberger. Rational Sets in Commutative Monoids. *Journal of Algebra*, 13:173–191, 1969.
10. S. Ginsburg and E. H. Spanier. Semigroups, Presburger formulas, and languages. *Pacific Journal of Mathematics*, 16:285–296, 1966.
11. Y.-S. Han, A. Salomaa, K. Salomaa, Derick Wood, and Sheng Yu. On the existence of prime decompositions. *Theory Comput. Syst.*, 376:60–69, 2007.
12. J. Karhumäki and L. P. Lisovik. The equivalence problem of finite substitutions on ab*c, with applications. *Int. J. Found. Comput. Sci.*, 14(4):699–, 2003.
13. M. Kunc. The power of commuting with finite sets of words. *Theory Comput. Syst.*, 40(4):521–551, 2007.
14. M. Kunc. The simplest language where equivalence of finite substitutions is undecidable. In *FCT*, pages 365–375, 2007.
15. S. Mantaci and A. Restivo. Codes and equations on trees. *Theor. Comput. Sci.*, 255(1-2):483–509, 2001.
16. P. Massazza and A. Bertoni. On the square root of languages. In *Formal power series and algebraic combinatorics (Moscow, 2000)*, pages 125–134, 2000.
17. D. Perrin. Codes conjugués. *Information and Control*, 20(3):222–231, 1972.
18. A. Restivo. A note on multiset decipherable codes. *IEEE Transactions on Information Theory*, 35(3):662–663, 1989.
19. A. A. Sardinas and C. W. Patterson. A necessary and sufficient condition for the unique decomposition of coded messages. In *IRE Intern. Conv. Rec. 8 (1953)*, pages 104–108. Chapman and Hall, 1953.