Discrete Mathematics and Theoretical Computer Science DMTCS vol. **9:2**, 2007, 227–240

Coding Partitions †

Fabio Burderi and Antonio Restivo

Dipartimento di Matematica e Applicazioni, Università degli studi di Palermo, Via Archirafi 34, 90123 Palermo, Italy. e-mail: {burderi, restivo}@math.unipa.it

received 14 Jan 2005, revised 17 Jun 2005, accepted 9 Sep 2005.

Motivated by the study of decipherability conditions for codes weaker than Unique Decipherability (UD), we introduce the notion of *coding partition*. Such a notion generalizes that of UD code and, for codes that are not UD, allows to recover the "unique decipherability" at the level of the classes of the partition. By tacking into account the natural order between the partitions, we define the *characteristic* partition of a code X as the finest coding partition of X. This leads to introduce the *canonical decomposition* of a code in at most one "unambiguous" component and other (if any) "totally ambiguous" components. In the case the code is finite, we give an algorithm for computing its canonical partition. This, in particular, allows to decide whether a given partition of a finite code X is a coding partition. This last problem is then approached in the case the code is a rational set. We prove its decidability under the hypothesis that the partition contains a finite number of classes and each class is a rational set. Moreover we conjecture that the canonical partition satisfies such a hypothesis. Finally we consider also some relationships between coding partitions and varieties of codes.

Keywords: Codes, Unique Decipherability, Rational Languages, Variety of Codes.

1 Introduction

The theory of uniquely decipherable (UD) codes, born in the context of information theory, plays a relevant role also in language theory and combinatorics on words (see [1]). In spite of their simple definition, the structure of UD codes is still for a large extent unknown.

In recent years some papers take into account codes that are not UD. The study of the corresponding ambiguities are, in certain cases, motivated by investigations on natural languages (see [3]). From another point of view, the classification of ambiguities is related to conditions of decipherability, weaker than UD, introduced to handle some special problems in information transmission (see [7, 5, 10, 8]). More generally, the study of ambiguities can help in understanding the structure of UD codes.

In this paper we introduce the notion of *coding partition* of a code X (here we call *code* an arbitrary set of words). Given a partition $P = \{X_1, X_2, \ldots\}$ of a code X, P is, roughly speaking,

†Partially supported by Italian MURST Project of National Relevance "Linguaggi Formali e Automi: Metodi, Modelli e Applicazioni"

a coding partition of X if any word $w \in X^*$ has a *unique* factorization $w = z_1 z_2 \cdots z_t$, where each "block" z_i is the concatenation of words from one class of P, and consecutive "blocks" are concatenations of words from different classes of P. The notion of coding partition generalizes that of UD code: indeed UD codes correspond to the extremal case in which each class contains exactly one element. In general, for codes that are not UD, the notion of coding partition allows to recover "unique decipherability" at the level of classes of the partition. In other words, such notion gives a tool to *localize* the ambiguities for a code that is not UD: indeed the ambiguities are bordered inside the individual classes of the partition and a sort of mutual unambiguity holds between the different classes.

By taking into account the natural ordering between the partitions of a set X, where finer is higher, we have that the coding partitions form a complete lattice. As a consequence, given a code X, we can define the finest coding partition P of X. It is called the *characteristic* partition of X and it is denoted by $P(X)$.

The structure of $P(X)$ gives useful information about coding properties of X. In particular, an extremal case (each class of $P(X)$ contains only one element) corresponds to UD codes. The opposite extremal case ($P(X)$ contains only one class) gives rise to the definition of *totally ambiguous* code. Such considerations leads to define a *canonical decomposition* of a code in at most one unambiguous component and in a set (possibly empty) of totally ambiguous components.

In Sec.3, we consider the case of a finite code X and we present an algorithm that gives the canonical decomposition of X. In particular, the algorithm allows also to decide: 1) given a partition P of X, whether P is a coding partition of X, and 2) given a code X, whether it is totally ambiguous.

In Sec.4, we take into account the rational case. We consider a partition $P = \{X_1, X_2, \ldots, X_n\}$ of a rational code X in a finite number of classes and such that all classes X_i are rational sets. We define a rational relation related to such a partition, and we prove that the partition is coding if and only if the rational relation is a function. This allows to decide whether the partition is coding.

In the last section, we consider the relationships between coding partitions and varieties of codes (see [5]). We prove that, given a coding partition of a code X, if each class of the partition belongs to a given variety of codes, then X belongs to the same variety.

2 Partitions of a code

Let A be a finite alphabet. Let A^* denotes the free monoid generated by A, i.e. the set of words over the alphabet A, and let $A^+ = A^* \backslash \{\varepsilon\}$.

A code X over A is a subset of A^+. The words of X are called *code words*, the elements of X^* *messages*, where X^* denotes the submonoid of A^* generated by X, i.e. the set of words obtained concatenating elements of X.

A code X is said to be *uniquely decipherable* (UD) if every message has a unique factorization into code words, i.e. the equality

$$x_1 x_2 \cdots x_n = y_1 y_2 \cdots y_m,$$

$x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m \in X$, implies $n = m$ and $x_1 = y_1, \ldots, x_n = y_n$.

The theory of UD codes has been widely developed, and it is closely related also to problems in automata theory, combinatorics on words, formal languages and semigroup theory. A complete treatment of such theory can be found in [1].

Remark 1 In literature, in general, the word *code* denotes a UD code (see [1]). In this paper we take also into account conditions of decipherability weaker than UD. This motivate the choice to call *code* an arbitrary subset of A^+.

Let X be a code and let

$$P = \{X_1, X_2, \ldots, X_i, \ldots\},$$

be a partition of X i.e. : $\bigcup_{i \geq 1} X_i = X$ and $X_i \cap X_j = \emptyset$, for $i \neq j$.

We say that a partition P is *concatenatively independent* if, for $i \neq j$,

$$X_i^+ \cap X_j^+ = \emptyset.$$

Let $P = \{X_1, X_2, \ldots\}$ be a concatenatively independent partition of a code X. A P-*factorization* of an element $w \in X^+$ is a factorization $w = z_1 z_2 \cdots z_t$, where

- $\forall i \;\; z_i \in X_k^+$, for some $k \geq 1$
- if $t > 1$, $z_i \in X_k^+ \Rightarrow z_{i+1} \notin X_k^+$, for all $1 \leq i \leq t-1$.

The partition P is called a *coding partition* if it is concatenatively independent and moreover any element $w \in X^+$ has a *unique* P-*factorization*, i.e. if

$$w = z_1 z_2 \cdots z_s = u_1 u_2 \cdots u_t,$$

where $z_1 z_2 \cdots z_s$, $u_1 u_2 \cdots u_t$ are P-*factorizations* of w, then $s = t$ and $z_i = u_i$ for $i = 1, \ldots, s$.

Remark 2 If $P = \{X_1, X_2, \ldots\}$ is a coding partition of X, in general neither X nor the sets X_i are UD codes. On the other hand P may not be a coding partition though all X_i are UD. This is shown by the following examples.

Example 1 $P = \{X_1, X_2,\}$, $X_1 = \{00, 000\}$, $X_2 = \{11, 111\}$. X_1 and X_2 are not UD but all the words of X^+ have a unique P-factorization.

Example 2 $P = \{X_1, X_2,\}$, $X_1 = \{0\}$, $X_2 = \{01, 10\}$. The sets X_1, X_2 are UD but the word $w = 010 \in X^+$ has two P-factorizations : $w = (0)(10) = (01)(0)$.

Remark 3 If X is a UD code then every partition of X is a coding partition. Therefore if e.g. X is a infinite UD code, it is possible to have coding partitions with infinitely many classes.

Let X be a code and let $x_1x_2\cdots x_s = y_1y_2\cdots y_t$ be two factorizations into code words of a message $w \in X^+$. In the sequel, when no confusion arises, sometimes we will denote by w both the "word" w and the *relation* $x_1x_2\cdots x_s = y_1y_2\cdots y_t$. We say that the relation $x_1x_2\cdots x_s = y_1y_2\cdots y_t$ is *prime* if for all $i < s$ and for all $j < t$ one has $x_1x_2\cdots x_i \neq y_1y_2\cdots y_j$.

Remark 4 A relation $w = x_1x_2\cdots x_s = y_1y_2\cdots y_t$ can be uniquely factorized into prime relations $w = v_1v_2\cdots v_p$, where

$$v_1 = x_1\cdots x_{i_2-1} = y_1\cdots y_{j_2-1},$$

$$\vdots$$

$$v_p = x_{i_p}\cdots x_s = y_{j_p}\cdots y_t.$$

Remark 5 Let $P = \{X_1, X_2, \ldots\}$ be a partition of a code X and let $w \in X^+$ be a message. If there exists a unique factorization of w into code words then there exists a unique P-factorization of w: the consecutive words belonging to the same set of the partition will form a block. This means that if we have two distinct P-factorizations of a message w, we have at least two distinct factorizations of w into code words.

Theorem 1 *Let $P = \{X_1, X_2, \ldots\}$ be a partition of a code X. The partition P is a coding partition of X iff for every prime relation $x_1x_2\cdots x_s = y_1y_2\cdots y_t$ there exists an integer h such that for all $i \leq s$ and for all $j \leq t$, x_i, $y_j \in X_h$.*

Proof: $\Rightarrow$ Let $w = x_1x_2\cdots x_s = y_1y_2\cdots y_t$ be a prime relation and let $z_1z_2\cdots z_p$ be the unique P-factorization of w. We prove that $p = 1$. If $p > 1$, by the uniqueness of the P-factorization, there exist $l < s$, $m < t$, such that $z_1 = x_1x_2\cdots x_l = y_1y_2\cdots y_m$; but this contradicts the fact that the relation is prime.
$\Leftarrow$ We first prove that the sets X_i are concatenatively independent. Suppose, by contradiction, that there exists a word $w \in X_h^+ \cap X_k^+$, with $h \neq k$ and let $w = x_1x_2\cdots x_s = y_1y_2\cdots y_t$, with $x_i \in X_h,\ 1 \leq i \leq s$, and $y_j \in X_k,\ 1 \leq j \leq t$. From the Remark 4 we can factorize w into prime relations $w = v_1v_2\cdots v_p$ and let $v_1 = x_1\cdots x_l = y_1\cdots y_m$, $l \leq s, m \leq t$. This contradicts the hypothesis, since $x_i \in X_h$, $y_j \in X_k$, $h \neq k$.
We will show now that every message has an unique P-factorization. Suppose, by contradiction, that $w \in X^+$ is a message with two distinct P-factorizations: $w = z_1z_2\cdots z_s = u_1u_2\cdots u_t$. Since the sets X_i are concatenatively independent, we can assume that $z_1 \neq u_1$. We can then suppose that $|u_1| < |z_1|$ and so $t \geq 2$. From the Remark 5 there exist at least two different factorizations of w into code words. Given two different factorizations of w, we have a relation that, by Remark 4, can be uniquely factorized as a product of prime relations: $w = v_1v_2\cdots v_p$. Since $t \geq 2$, there exist $i, j \geq 1$ such that $u_1 \in X_i^+, u_2 \in X_j^+$, with $i \neq j$. Now since v_h, $1 \leq h \leq p$, is a prime relation and u_1 and u_2 belong to different sets of the partition, we have from the hypothesis that $p \geq 2$ and no factor v_h can cross u_1, i.e. v_h cannot be factorized as $v_h = xy$, with x suffix of u_1 and y prefix of u_2. In particular $|v_1| \leq |u_1| < |z_1|$ and then v_1 is a prefix of both u_1 and z_1. By hypothesis we obtain that $z_1 \in X_i^+$. Let $u_1 = v_1v_2\cdots v_k, 1 \leq k < p$. Now if $s > 1$ or $t > 2$, v_{k+1} can cross neither u_2 nor z_1 so, in any case, v_{k+1} is a prefix of u_2 and a factor of z_1. But $u_2 \in X_j^+$, $z_1 \in X_i^+$ and v_{k+1} is prime. So we have a contradiction and this concludes the proof. $\square$

The partition $P = \{X_1, X_2, \ldots\}$ of X is called *trivial* if $|P| = 1$. It is called the *discrete* partition if $|X_i| = 1$ for $i \geq 1$.

Remark 6 The trivial partition is a coding partition. Moreover X is a UD code if and only if the discrete partition of X is a coding partition. In this sense the notion of coding partition generalizes to the partitions of a set the notion of UD code.

Given a partition $P = \{X_1, X_2, \ldots\}$ of a code X, a subset $Y \subseteq X$ is a *cross-section* of P, if $|Y \cap X_i| = 1$, for $i \geq 1$.

Theorem 2 *If P is a coding partition of X, then any cross-section Y of P is a UD code.*

Proof: Let $Y \subseteq X$ be a cross section of $P = \{X_1, X_2, \ldots\}$ and let $Y = \{y_1, y_2, \ldots\}$. Assume that Y is not UD, then there exists a word $w \in Y^+$ with two distinct factorizations $w = y_1 y_2 \cdots y_s = y'_1 y'_2 \cdots y'_t$, $y_i, y'_j \in Y$ for $1 \leq i \leq s,\ 1 \leq j \leq t$ and we can assume that $y_1 \neq y'_1$. Of course $w \in X^+$ but these factorizations may be not P-factorizations because it's possible that $y_i = y_{i+1}$ for some $1 \leq i \leq s-1$ or $y'_j = y'_{j+1}$ for some $1 \leq j \leq t-1$. We obtain a P-factorizations if we turn the repetitions in a block, but we obtain two distinct P-factorizations: indeed, since $y_1 \neq y'_1$ and Y is a cross section, the first blocks are different. Since P is, by hypothesis, a coding partition, we have a contradiction and so Y must be a UD code. □

Remark 7 The converse of previous proposition does not hold in general, i.e. there exist non-coding partitions P of a code X, such that all the cross-sections of P are UD codes. This is shown by the partition in the Examples 2. As a consequence, in order to decide whether a partition P of a finite code X is coding, it does not suffice to test whether all cross-sections of P are UD codes. An algorithm to decide whether a partition is coding will be given in the next section.

Recall that there is a natural order between the partitions of a set X: if P_1 and P_2 are two partitions of X, $P_1 \leq P_2$ if the elements of P_1 are unions of elements of P_2.

Theorem 3 *The set of the coding partitions of a code X is a complete lattice.*

Proof: Let $\mathcal{L}(X)$ the set of all coding partitions of a set X and let $\mathcal{F} = \{P_i \,|\, P_i \in \mathcal{L}(X)\,,\, i \in I\}$ be a family of coding partitions of X. Since the partitions of a set form a complete lattice, it is sufficient to prove that both the meet $M = \bigwedge_{i \geq 1} P_i$ and the join $J = \bigvee_{i \geq 1} P_i$ belong to $\mathcal{L}(X)$. Let $x_1 x_2 \cdots x_s = x_{s+1} x_{s+2} \cdots x_t$ be a prime relation between code words and let $S = \{x_1, x_2, \ldots, x_t\}$. By Theorem 1 there exists $X_{h_i} \in P_i$, such that $S \subseteq X_{h_i}$, $\forall i \in I$. Since $M \leq P_i$ $\forall i \in I$, there exists $X_k \in M$ such that $S \subseteq X_k$; from another hand since J is the least upper bound there exists $X_l \in J$ such that $S \subseteq X_l$ so, by the same theorem, we obtain the thesis.

□

As a consequence of previous theorem, given a code X, we can define the finest coding partition P of X. It is called the *characteristic* partition of X and it is denoted by $P(X)$.

Then, as remarked before, X is a UD code if and only if $P(X)$ is the discrete partition. The opposite extremal case gives rise to the next definition.

A code X is called *ambiguous* if it is not UD. It is called *totally ambiguous* if $|X| > 1$ and $P(X)$ is the trivial partition.

The code $X = \{01, 10, 1\}$ is totally ambiguous, as the reader can easily verify using Theorem 1.

Remark 8 In [11], Weber and Head introduced the notion of *numerically decipherable* code: a code X is numerically decipherable (ND) if any two factorizations in code words of each message over X involve the same number of words. In the same paper the authors introduced the notion of *homophonic partition* of ND codes. The notion of *coding partition* is here introduced for any code, i.e. for an arbitrary set X of words, and differs from that of homophonic partition even in the case when X is a ND code. Indeed in [11] the authors show that $\{\{0, 10, 101\}, \{111\}\}$ is the *finest homophonic partition* of the ND code $X = \{x_1, x_2, x_3, x_4\} = \{0, 10, 101, 111\}$. Since $x_2x_4x_2 = x_3x_4x_1$ is a prime relation, we have, according to Theorem 1, that $P(X)$ is the trivial partition.

A less trivial example of totally ambiguous code is the code

$$X = \{000, 0010, 001, 10, 1\},$$

that we will study in an example of the next section.

Theorem 4 *Let $P(X) = \{X_1, X_2, \ldots\}$ be the characteristic partition of a code X. If $|X_i| > 1$ for some $i \geq 1$, then X_i is a totally ambiguous code.*

Proof: Suppose, without loss of generality, that $|X_1| > 1$. Suppose, by contradiction, that X_1 is not totally ambiguous, and let $P(X_1) = \{Y_1, Y_2, \ldots\}$ be the characteristic partition of X_1, with $|P(X_1)| \geq 2$. If we consider now $P' = \{Y_1, Y_2, \ldots, X_2, \ldots\}$, this is a coding partition of X with $P(X) < P'$. Since this contradicts the definition of $P(X)$, X_1 must be totally ambiguous. □

The property that any proper subset Y of an ambiguous code X is a UD code, is related to the property of X to be totally ambiguous, as we can see from the next proposition.

Theorem 5 *Let X be a code such that all proper subsets Y of X are UD codes. Then either X is a UD code or it is totally ambiguous.*

Proof: Let X be a non UD code. Let $P(X) = \{X_1, X_2, \ldots\}$ be the characteristic partition of X and suppose, by contradiction, that X is not totally ambiguous. Then $|P(X)| \geq 2$ and since X is not UD, there exists a X_i, $i \geq 1$ such that $|X_i| > 1$. We have that $X_i \subsetneq X$ and that, by Theorem 4, it is totally ambiguous. This contradicts the hypothesis. □

The converse implication does not hold in general, as shown by the last example just given above: it is totally ambiguous, but its proper subset $\{000, 001, 10, 1\}$ is not a UD code.

Let X be a code and let $P(X)$ be the characteristic partition of X. Let X_0 be the union of all classes of $P(X)$ having only one element, i.e. of all classes $Z \in P(X)$ such that $|Z| = 1$. The code X_0 is a UD code and is called the *unambiguous component* of X. From $P(X)$ one then derives another partition of X

$$P_C(X) = \{X_0, X_1, \ldots\},$$

where $|X_i| > 1$, for $i \geq 1$. The sets X_i, with $i \geq 1$, are, by Theorem 4, totally ambiguous. They are called the *totally ambiguous components* of X. The partition $P_C(X)$ is called the *canonical partition* of X: it defines a *canonical decomposition* of a code X in at most one unambiguous component and a (possibly empty) set of totally ambiguous components. Roughly speaking, if a code X is not UD, then its canonical decomposition, on one hand separates the unambiguous component of the code (if any), and, on the other, localizes the ambiguities inside the totally ambiguous components of the code. If, on the contrary, X is UD, then its canonical decomposition contains only the unambiguous component X_0.

3 Computing the canonical partition of a finite code

In this section we present an algorithm that computes the canonical partition of a finite code. The algorithm is very close to the Sardinas-Patterson algorithm testing whether a code is UD and to its variations like *domino graph* and *simplified domino graph* (see [6, 5]). Informally, like in Sardinas-Patterson algorithm, at each step we construct a set of suffixes of code words by comparing the code words with the suffixes constructed in the previous step. Here, in addition, for each new suffix u generated by the procedure, we record in a set S (associated to the suffix u) the set of indices of code words involved in the generation of such a suffix. We then construct a sequence of sets, whose elements are pairs of the form (u, S), where u is a suffix of code words and S is a set of indices corresponding to the code words.

Let $X = \{x_1, x_2, \ldots, x_k\}$ be a finite code. We construct a sequence $(U_i)_{i\geq 1}$, where each U_i is a set of pairs of the form (u, S), with $u \in A^*$ and $S \subseteq \{1, 2, \ldots, k\}$. The sequence $(U_i)_{i\geq 1}$ is defined inductively as follows:

$$U_1 = \{(u, \{i, j\})|\ x_i u = x_j,\ 1 \leq i, j \leq k,\ i \neq j,\ u \in A^+\},$$

and, for $n \geq 1$,

$$U_{n+1} = \{(z, S \cup \{i\})|\ x_i z = u \ \ or \ \ uz = x_i,\ \ (u, S) \in U_n, u \neq \varepsilon,\ \ 1 \leq i \leq k\},$$

Consider now the family of subsets of the set $\{1, 2, \ldots, k\}$

$$\mathcal{T}_X = \{S|\ (\varepsilon, S) \in U_i,\ i > 1\},$$

and denote by S_X the set

$$S_X = \bigcup_{S \in \mathcal{T}_X} S \subseteq \{1, 2, \ldots, k\}.$$

Let τ be the transitive closure of the relation $\backsim$ defined in the set S_X as follows:

$$r \backsim t \ \Leftrightarrow\ r, t \in S, \quad S \in \mathcal{T}_X.$$

Set $R_0 = \{1, 2, \ldots, k\} \setminus S_X$ and let $R_1, R_2, \ldots, R_n$ be the equivalence classes of τ in S_X. It is obvious that $\{R_0, R_1, \ldots, R_n\}$ defines a partition of the set $\{1, 2, \ldots, k\}$. Such a partition induces, in turns, a partition on the set $X = \{x_1, x_2, \ldots, x_k\}$: $X_i = \{x_j \in X \mid j \in R_i\}, i = 0, 1, \ldots, n$. We denote this partition produced by the algorithm, by $\mathcal{R}(X)$.

Remark 9 Since $|X| = k$, the elements of U_i's are pairs composed by a suffix of words in X and a subset of $\{1, 2, \ldots, k\}$. Then the set of different elements in the sequence $(U_i)_{i \geq 1}$ is finite. As a consequence, the partition $\mathcal{R}(X)$ can be effectively constructed.

Theorem 6 *The partition $\mathcal{R}(X)$ is the canonical partition of the code X.*

Proof: Let $P_0 = \{\{x_j\} : j \in R_0\}$ and let $P(X) = P_0 \cup \{X_1, X_2, \ldots X_n\}$. By construction $P(X)$ is a partition of X. Moreover for every prime relation $v = x_1 x_2 \cdots x_s = x_{s+1} x_{s+2} \cdots x_t$ the algorithm finds the pair (ε, S), where $S = \{1, 2, \ldots t\}$ and, starting from S, the algorithm creates the set of code words X_h. By construction we have that $x_i \in X_h, 1 \leq i \leq t$ and so, because of Theorem 1, $P(X)$ is a coding partition. It is left to the reader to convince himself that $P(X)$ is the characteristic partition of X so that $\mathcal{R}(X)$ become the canonical partition of the code X. $\square$

Corollary 7 *Let $P = \{X_1, X_2, \ldots, X_n\}$ be a partition of a finite code X. There is an algorithm to decide whether P is a coding partition.*

Proof: Using the algorithm we find the canonical partition of X, and we test after if $P \leq P_C(X)$ just verifying if the classes of P are unions of classes of $P_C(X)$. $\square$

Example 3 Consider the code $X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\} = \{00, 0010, 1000, 11, 1111, 010, 011\}$.

$U_1 = \{(10, \{1, 2\}), (11, \{4, 5\})\}$,

$U_2 = \{(\varepsilon, \{4, 5\}), (00, \{1, 2, 3\}), (11, \{4, 5\})\}$,

$U_3 = \{(\varepsilon, \{1, 2, 3\}), (\varepsilon, \{4, 5\}), (10, \{1, 2, 3\}), (11, \{4, 5\})\}$,

$U_4 = U_2$.

So we have: $R_0 = \{6, 7\}, R_1 = \{1, 2, 3\}, R_2 = \{4, 5\}$. Then $X_0 = \{010, 011\}$, $X_1 = \{00, 0010, 1000\}, X_2 = \{11, 111\}$ and $P_C(X) = \{X_0, X_1, X_2\}$ is the canonical partition of X.

Example 4 Consider the code $X = \{000, 0010, 001, 10, 1\}$. *Set:*

$$x_1 = 000, \;\; x_2 = 0010, \;\; x_3 = 001, \;\; x_4 = 10, \;\; x_5 = 1.$$

We construct the sequence:

$U_1 = \{(0, \{2, 3\}), \, (0, \{4, 5\})\}$,

$U_2 = \{(00, \{1,2,3\}), (00, \{1,4,5\}), (01, \{2,3\}), (01, \{3,4,5\}),$
$(010, \{2,3\}), (010, \{2,4,5\})\},$

$U_3 = \{(0, \{1,2,3\}), (0, \{1,4,5\}), (1, \{1,2,3\}), (1, \{1,3,4,5\}), (10, \{1,2,3\}),$
$(10, \{1,2,4,5\})\},$

$U_4 = \{(\varepsilon, \{1,2,3,4\}), (\varepsilon, \{1,2,3,5\}), (\varepsilon, \{1,2,4,5\}), (\varepsilon, \{1,3,4,5\}),$
$(0, \{1,2,3,4\}), (0, \{1,2,3,5\}), (0, \{1,2,4,5\}), (0, \{1,3,4,5\}), (00, \{1,2,3\}),$
$(00, \{1,4,5\}), (01, \{1,2,3\}), (01, \{1,3,4,5\}), (010, \{1,2,3\}), (010, \{1,2,4,5\})\},$

We claim that, in this case, we can stop the computation. Indeed since U_4 contains the pairs $(\varepsilon, \{1,2,3,4\})$, $(\varepsilon, \{1,2,3,5\}), (\varepsilon, \{1,2,4,5\}), (\varepsilon, \{1,3,4,5\})\}$, it follows that $S_X = \{1,2,3,4,5\}$ and that the partition corresponding to the equivalence τ is composed by only one class. Then the canonical partition of X is the trivial partition and then X is totally ambiguous.

4 The rational case

If a code X is infinite, one can have partitions having an infinite number of classes and, moreover, each class can contain infinitely many elements. In this section we consider first partitions having a *finite* number of classes and such that each class is a *rational* set. So we have a concatenatively independent partition

$$P = \{X_1, X_2, \ldots, X_n\}$$

such that each X_i is a rational set. In order to give an algorithm to decide whether a partition is a coding partition, we need some preliminary definitions and results from the theory of *rational relations*. Let us recall that a rational relation $\rho : A^* \to B^*$ is a mapping from A^* into the set 2^{B^*} of the subsets of B^* such that the graph

$$G(\rho) = \{(u,v) \in A^* \times B^* | v \in \rho(u)\}$$

is a rational subset of the product monoid $A^* \times B^*$.

M. Nivat has given the following characterization of a rational relation (see [4]).

Theorem 8 *Let $\rho : A^* \to B^*$ be a relation. Then ρ is a rational relation iff there exist a new finite alphabet Σ, two alphabetic morphisms $\alpha : \Sigma^* \to A^*$ and $\beta : \Sigma^* \to B^*$, and a rational subset K of Σ^* such that*

$$G(\rho) = \{(\alpha(v), \beta(v)) | v \in K\}.$$

Come now back to the partition $P = \{X_1, X_2, \ldots, X_n\}$, where each X_i is a rational set over the alphabet A. Let $A_1, A_2, \ldots, A_n$ be n disjoint copies of the alphabet A, and set $\Sigma = \bigcup_{i=1}^n A_i$. Let $\alpha_i : A_i \to A$ be the bijection of A_i and A, for $i = 1, 2, \ldots, n$ and let $\alpha : \Sigma^* \to A^*$ be the extension of the α_i to $\Sigma^* : \alpha|_{A_i} = \alpha_i$ for $i = 1, 2, \ldots, n$. Set

$$K = (\bigcup_{i=1}^n K_i)^* \subseteq \Sigma^*,$$

where $K_i = \alpha_i^{-1}(X_i)$.

Consider now another alphabet $B = \{b_1, b_2, \ldots, b_n\}$ and the alphabetic morphism $\beta : \Sigma^* \to B^*$ defined as follows:

$$\beta(a) = b_i \;\; \Leftrightarrow \;\; a \in A_i.$$

Let $\rho(P) : A^* \to B^*$ be the relation defined by the following graph

$$G(\rho(P)) = \{(\alpha(v), \beta(v)) | v \in K\}.$$

By the theorem of Nivat, $\rho(P)$ is a rational relation.

Let us now recall that a relation $\rho : A^* \to B^*$ is a function if $|\rho(u)| = 1$ for all words $u \in A^*$. The main result of this section is the following theorem.

Theorem 9 *A partition P is a coding partition iff $\rho(P)$ is a function.*

Proof: Let $\rho = \rho(P)$. From the definition of ρ we see that there is a bijection between the set of the P-factorizations of a word $w \in X^+$ and the set of the words $v \in K$ such that $\alpha(v) = w$. Moreover since β is injective, P is a coding partition if and only if for each word $w \in X^+$, $|\{\beta(v) \; : \; v \in K, \, \alpha(v) = w\}| = 1$, that is if and only if ρ is a function. □

In [9] M. P. Schutzenberger proved that it is decidable whether a rational relation is a function. As a consequence, we obtain the following corollary of the previous theorem.

Corollary 10 *Given a partition $P = \{X_1, X_2, \ldots, X_n\}$ such that X_i, for $i = 1, 2, \ldots, n$, is a rational set, then it is decidable whether P is a coding partition.*

Let us now consider the (more difficult) problem to determine the *canonical* partition of a rational code X. If the canonical partition has infinitely many classes, it is not clear what means to compute such a partition. Remark that the *characteristic* partition of a rational code X may have infinitely many classes. Consider, for instance, an infinite UD code X: the characteristic partition coincides with the discrete partition, i.e. each class contains only one element and there exist infinitely many classes. However, in such a case, the canonical partition is the trivial partition, composed by only one class. In all examples of rational codes that we know, the number of classes of the canonical partition is always finite and each class is a rational set. So we formulate the following conjecture.

CONJECTURE : *If X is rational, the number of classes of $P_C(X)$ is finite and each class of $P_C(X)$ is a rational set.*

If the conjecture is true, the restrictive conditions considered in this section are not actually a restriction, but correspond to the general case.

Remark further that, if X is not rational, then $P_C(X)$ can have infinitely many classes, as shown by the following example.

Example 5 Consider the code

$$X = \{(a^n b)^2 | n \geq 0\} \cup \{(a^n b)^3 | n \geq 0\}.$$

It is easy to verify that $P_C(X)$ contains infinitely many classes and that any class X_i is of the form $X_i = \{(a^i b)^2, (a^i b)^3\}$, $i \in \mathbb{N}$.

5 Coding partitions and varieties of codes

The notion of coding partition allows to decompose a code in such a way that the ambiguities are bordered into the single components of the code and a sort of mutual unambiguity holds between its different components. In recent years, conditions of decipherability weaker than UD have been investigated, formalized in terms of the notion of variety of codes, and a sort of classification of ambiguities of codes has been introduced. It is then interesting to study the relationships between the "type" of ambiguity of the single components in the partition, and that of the whole code.

Let us first briefly introduce the basic definitions and the motivations about the notion of variety of codes. The investigation on decipherability conditions weaker than UD was initiated in [7] by Lempel, who introduced the notion of *multiset decipherable* (MSD) codes. Here the information of interest is the multiset of code words used in the encoding process so that the order in which transmitted words are received is immaterial. In a more formal way, a code X is a MSD code if the equality

$$x_1 x_2 \cdots x_n = y_1 y_2 \cdots y_m,$$

with $x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m \in X$, implies the equality of the two multisets $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_m\}$.

In [5] Guzmán considers also the notion of *set decipherable* (SD) codes. In this case the original message is recovered up to commutativity and actual count of occurrences, i.e. two factorizations of the same message yield the same *set* of code words. Denote by $\mathcal{UD}$, $\mathcal{MSD}$ and $\mathcal{SD}$ the classes of UD, MSD and SD codes, respectively. It is clear that

$$\mathcal{UD} \subseteq \mathcal{MSD} \subseteq \mathcal{SD}$$

and has been shown that the two inclusions are strict.

In the same paper [5] Guzmán introduces a very general concept of decipherability using *varieties* of monoids. Unique decipherability, multiset decipherability and set decipherability then appear as very special cases of such general concept.

Let X be a code and let M be a monoid. We say that X is *decipherable* in M if every map $f : X \to M$ extends to a (unique) homomorphism $\overline{f} : X^* \to M$.

Let $\mathcal{K}$ be a class of codes and $\mathcal{V}$ a class of monoids. We denote by $\mathcal{M}(\mathcal{K})$ the class of all momoids in which every $X \in \mathcal{K}$ is decipherable. Conversely, $\mathcal{C}(\mathcal{V})$ represents the class of all codes decipherable in every $\mathcal{M} \in \mathcal{V}$.

A *variety of codes* is a class of codes $\mathcal{K}$ such that $\mathcal{C}(\mathcal{M}(\mathcal{K})) = \mathcal{K}$.

The following holds (see [5]):

- $\mathcal{UD}$ is a variety of codes corresponding to the variety of all monoids.
- $\mathcal{MSD}$ is a variety of codes corresponding to the variety of commutative monoids.
- $\mathcal{SD}$ is a variety of codes corresponding to the variety of semilattices i.e. the variety of commutative monoids that are idempotent.

The varieties of codes have been investigated from different points of view. In particular, the papers [5] (see also [11]) studies the problem to decide whether a code belongs to a given variety. In [2] the authors study the problem to characterize those varieties of codes where the Kraft inequality is satisfied.

Let us now come back to partitions. Let $P = \{X_1, X_2, \ldots\}$ be a partition of a code X. We are interested to investigate whether there is some relationship between the varieties corresponding to the classes X_i's and the variety corresponding to the code X. The result of this section, that generalizes a result obtained in [2], is given by the following theorem.

Theorem 11 *Let $P = \{X_1, X_2, \ldots\}$ be a coding partition of a code X and let $\mathcal{K}$ be a variety of codes. If $X_i \in \mathcal{K}$ for $i \geq 1$, then $X \in \mathcal{K}$.*

Proof: Let $\mathcal{N}$ be the variety of monoids associate to the variety $\mathcal{K}$ of codes. We have to show that for all monoids $M \in \mathcal{N}$, X is decipherable in M. Let $M \in \mathcal{N}$, and let $f : X \to M$ be a map from X to M. Denote, for $i \geq 1$, by g_i the restrictions of f to X_i : $g_i := f|_{X_i}$. Since X_i are decipherable in M, g_i extends to $\overline{g_i} : X_i^* \to M$, for $i \geq 1$. Let $w \in X^+$ and suppose that its unique P-factorization is $w = z_{i_1} z_{i_2} \cdots z_{i_n}$, where $z_{i_j} \in X_{i_j}^+$, $i_j \in \{1, 2, \ldots\}$. Then, putting $\overline{f}(w) := \overline{g_{i_1}}(z_{i_1})\overline{g_{i_2}}(z_{i_2}) \cdots \overline{g_{i_n}}(z_{i_n})$, we get the unique homomorphism extending f, and so X is decipherable in M. $\Box$

Using the fact that the varieties of codes form a complete lattice (see Corollary 1.6 in [5]) we have the next corollary.

Corollary 12 *Let $P = \{X_1, X_2, \ldots\}$ be a coding partition of a code X and let $\mathcal{K}_i$ be varieties of codes such that $X_i \in \mathcal{K}_i$, $i \geq 1$. Then X belongs to the join $\bigvee_{i\geq 1} \mathcal{K}_i$.*

Therefore, in particular, if each X_i is a UD code, then also X is a UD code. In the same way, if each X_i is a MSD code, then also X is a MSD code, etc. In a coding partition, the properties of the individual classes are transferred to the whole code.

References

[1] J. Berstel, D. Perrin: The Theory of Codes, Academic Press, New York, 1985.

[2] F. Burderi, A. Restivo: Varieties of Codes and Kraft Inequality, LNCS 3404, Proceedings of STACS 2005, 545-556.

[3] C. Del Vigna, V. Berment: Ambiguités irreductibles dans les monoides des mots, Bulletin of the Belgian Mathematical Society, Vol. 10, N.5 2003, 693-706.

[4] S. Eilenberg: Automata, Languages and Machines, Vol.A, Academic Press, New York, 1974.

[5] F. Guzmán: Decipherability of codes, Journal of Pure and Applied Algebra **141** (1999) 13-35.

[6] T. Head, A. Weber: Deciding multiset decipherability, IEEE Trans. Inform. Theory **41** (1995) 291-297.

[7] A. Lempel: On multiset decipherable codes, IEEE Trans. Inform. Theory **32** (1986) 714-716.

[8] A. Savelli: On numerically decipherable codes and their homophonic partitions, Information Processing Letters 90 (2004) 103-108.

[9] M. P. Schutzenberger: Sur les Relations Rationnelles, Lecture Notes in Computer Science, 33 1975.

[10] A. Restivo: A note on multiset decipherable code, IEEE Trans. Inform. Theory **35** (1989) 662-663.

[11] A. Weber, T.J. Head: The finest homophonic partition and related code concepts IEEE Trans. Inform. Theory **42** (1996) 1569-1575.