RAIRO-Theor. Inf. Appl. **42** (2008) 539–552
DOI: 10.1051/ita:2008016

Available online at:
www.rairo-ita.org

COMPATIBILITY RELATIONS ON CODES AND FREE MONOIDS *

Tomi Kärki[1]

Abstract. A compatibility relation on letters induces a reflexive and symmetric relation on words of equal length. We consider these word relations with respect to the theory of variable length codes and free monoids. We define an (R, S)-code and an (R, S)-free monoid for arbitrary word relations R and S. Modified Sardinas-Patterson algorithm is presented for testing whether finite sets of words are (R, S)-codes. Coding capabilities of relational codes are measured algorithmically by finding minimal and maximal relations. We generalize the stability criterion of Schützenberger and Tilson's closure result for (R, S)-free monoids. The (R, S)-free hull of a set of words is introduced and we show how it can be computed. We prove a defect theorem for (R, S)-free hulls. In addition, a defect theorem of partial words is proved as a corollary.

Mathematics Subject Classification. 68R15.

Introduction

The theory of variable length codes is firmly related to information theory and to combinatorics on words [2]. The object of the theory is to study factorization of words into sequences of words taken from a given set X. In a free monoid X^* generated by a code X there does not exist two distinct factorizations in X for any word. This coding property can be strengthened by requiring that two nearly similar, *i.e.*, *compatible* words, have the same, or at least similar, factorizations. This is attained here by introducing *word relations* and *relational codes*. More

Keywords and phrases. Compatibility relation, free monoid, stability, defect theorem, partial word.

* *The best student paper of the 11th Mons Days of Theoretical Computer Science awarded by Fondation Michel Métivier.*

[1] Department of Mathematics and Turku Centre for Computer Science, University of Turku, 20014 Turku, Finland; topeka@utu.fi

Article published by EDP Sciences

precisely, we consider here words together with a *compatibility relation* induced by a relation on letters. This notion generalizes that of *partial words*; see [1]. The theory of codes on combinatorics on words is revisited by defining (R,S)-*codes* for word relations R and S. If some of the letters in a message generated by an (R,S)-code are changed to related letters, the message can still be factorized, in other words decoded, in a proper manner. Thus these codes possess some error correction capabilities. We describe an algorithm to test whether or not a finite set of words is an (R,S)-code. In addition, coding properties of finite sets of words are explored by finding maximal and minimal relations with respect to relational codes.

The base of a free word monoid is a code. Relational codes, in turn, generate *relationally free* monoids. Results of free monoids can be generalized for relationally free monoids. For example, we generalize the stability criterion of Schützenberger and Tilson's closure result. Especially, we are interested in the so called *defect effect*: if a set of n words satisfies a nontrivial relation, then these words can be expressed simultaneously as products of less than n words. Another formulation of the defect effect is to say that the cardinality of the base of the *free hull* of X, *i.e.*, the smallest free monoid containing a set of words X is strictly smaller than the cardinality of X if and only if X is not a code. Actually, there exist several defect theorems depending on the restrictions that are put to the $n-1$ words [11]. The defect theorem of words is used in many different connections [3,7,13,15]. In this paper we generalize the defect theorem for (R,S)-*free hulls* and we give an algorithm how to compute these hulls. Moreover, a defect theorem of partial words is proved as a corollary.

We end this section with some notation. An *alphabet* A is a nonempty finite set of symbols and a *word* over A is a (finite or infinite) sequence of symbols from A. The empty word is denoted by ε. The sets of all finite words and finite nonempty words over A are denoted by A^* and A^+, respectively. With the operation of catenation A^* is a free monoid and A^+ is a free semigroup generated by the letters of A. The *length* of a word w, denoted by $|w|$, is the total number of (occurrences of) letters in w. The ith symbol of the word w is denoted by $w(i)$. A word w is a *factor* of a word u (resp. a left factor or a *prefix*, a right factor or a *suffix*), if there exist words x and y such that $u = xwy$ (resp. $u = wy, u = xw$). If $w = uv$ then we denote $v = u^{-1}w$.

For subsets $L, K \subseteq A^*$, we let

$$\begin{aligned} LK &= \{uv \mid u \in L, v \in K\},\\ L^+ &= \textstyle\bigcup_{i\geq 1} L^i, \quad L^* = L^+ \cup \{\varepsilon\},\\ L^{-1}K &= \{u^{-1}w \mid u \in L, w \in K\}. \end{aligned}$$

1. Word relations

Let $R \subseteq X \times X$ be a relation on a set X. We often write $x\,R\,y$ instead of $(x,y) \in R$. Then R is a *compatibility relation* if it is both reflexive and symmetric, *i.e.*, (i) $\forall x \in X : \; x\,R\,x$, and (ii) $\forall x,y \in X : \; x\,R\,y \implies y\,R\,x$. The *identity*

relation on a set X is defined by $\iota_X = \{(x,x) \mid x \in X\}$ and the *universal relation* on X is defined by $\Omega_X = \{(x,y) \mid x,y \in X\}$. Subscripts are often omitted when they are clear from the context. Clearly, both ι_X and Ω_X are compatibility relations on X. A compatibility relation $R \subseteq A^* \times A^*$ on the set of all words will be called a *word relation* if it is induced by its restriction on the letters, *i.e.*,

$$a_1 \dots a_m \, R \, b_1 \dots b_n \quad \Longleftrightarrow \quad m = n \text{ and } a_i \, R \, b_i \text{ for all } i = 1, 2, \dots, m$$

whenever $a_1, \dots, a_m, b_1, \dots, b_n \in A$.

Let R be a relation on A. By $\langle R \rangle$ we denote the compatibility relation *generated* by R, *i.e.*, $\langle R \rangle$ is the reflexive and symmetric closure of the relation R. Sometimes we need to consider the restriction of a relation R on a subset X of A^*. We denote $R_X = R \cap (X \times X)$. Words u and v satisfying $u \, R \, v$ are said to be *compatible* or, more precisely, *R-compatible*. If two words are not compatible, they are said to be *incompatible*.

Example 1.1. In the binary alphabet $A = \{a, b\}$ the compatibility relation

$$R = \langle \{(a,b)\} \rangle = \{(a,a), (b,b), (a,b), (b,a)\}$$

makes all words with equal length compatible with each other. In the ternary alphabet $\{a, b, c\}$, where

$$S = \langle \{(a,b)\} \rangle = \{(a,a), (b,b), (a,b), (b,a), (c,c)\},$$

we have $abba \, S \, baab$ but, for instance, words abc and cac are not S-compatible.

Clearly a word relation R satisfies the following two conditions:

$$\begin{array}{lll} \text{multiplicativity:} & u \, R \, v, u' \, R \, v' & \Longrightarrow \quad uu' \, R \, vv', \\ \text{simplifiability:} & uu' \, R \, vv', |u| = |v| & \Longrightarrow \quad u \, R \, v, \ u' R v'. \end{array}$$

However, a word relation R does not need to be transitive. *From now on the relations on words considered in this presentation are supposed to be word relations induced by some compatibility relation on letters.*

Let 2^X denote the *power set* of X, that is, the family of all subsets of X including the empty set $\emptyset$ and X itself. For a word relation R on A^*, let the corresponding function $R\colon 2^{A^*} \to 2^{A^*}$ be defined by

$$R(X) = \{u \in A^* \mid \exists \, x \in X : \, x \, R \, u\}.$$

If X contains only one word $w \in A^*$, we denote $R(X)$ shortly by $R(w)$. Note that the function R is multiplicative: $R(X)R(Y) = R(XY)$ for all $X, Y \subseteq A^*$ and $R(X)^* = R(X^*)$ for all $X \subseteq A^*$.

As another example of word relations we consider partial words. The notion of partial words was introduced by Berstel and Boasson in 1999 [1]. This subject has

been widely studied under the recent years; see, *e.g.*, the references in [4]. Motivation for the research of partial words comes partly from the study of biological sequences such as DNA, RNA and proteins; see [4,12].

Example 1.2. A *partial word* of length n over an alphabet A is a partial function

$$w\colon \{1,2,\ldots,n\} \to A.$$

The domain $D(w)$ of w is the set of positions $p \in \{1,2,\ldots,n\}$ such that $w(p)$ is defined. The set $H(w) = \{1,2,\ldots,n\} \setminus D(w)$ is the set of *holes* of w. To each partial word we may associate a total word $w_\diamondsuit$ over the extended alphabet $A_\diamondsuit = A \cup \{\diamondsuit\}$. This *companion* of w is defined by

$$w_\diamondsuit(p) \quad = \quad \begin{cases} w(p) & \text{if } p \in D(w), \\ \diamondsuit & \text{if } p \in H(w). \end{cases}$$

Thus, the holes are marked with the "do not know" symbol $\diamondsuit$. Clearly, partial words are in one-to-one correspondence with words over $A_\diamondsuit$.

The compatibility relation of partial words is defined as follows. Let x and y be two partial words of equal length. The word x is *contained* in y if $D(x) \subseteq D(y)$ and $x(k) = y(k)$ for all k in $D(x)$. Two partial words x and y are said to be *compatible* if there exists a partial word z such that z contains both x and y. Then we write $x \uparrow y$. For example, we see that the following partial words are compatible by comparing them with the total word "knowledge".

$$\begin{array}{cccccccccc} k & n & \diamondsuit & w & l & \diamondsuit & d & g & e \\ \diamondsuit & n & o & w & \diamondsuit & \diamondsuit & d & g & \diamondsuit \\ k & n & o & w & l & e & d & g & e \end{array}$$

It was shown in [9] that the compatibility relation $\uparrow$ of partial words can be considered as a word relation

$$R_\uparrow = \langle \{(\diamondsuit, a) \mid a \in A\} \rangle$$

over the alphabet $A_\diamondsuit$. Namely, compatible partial words x and y must have equal letters in the positions $i \in D(x) \cap D(y)$. This makes them also $R_\uparrow$-compatible and vice versa.

2. Relational codes

Let R and S be two word relations on the monoid A^*. A subset $X \subseteq A^*$ is an (R,S)-*code* if for all $n, m \geq 1$ and $x_1, \ldots, x_m, y_1, \ldots, y_n \in X$, we have

$$x_1 \ldots x_m \, R \, y_1 \ldots y_n \implies n = m \text{ and } x_i \, S \, y_i \text{ for } i = 1,2,\ldots,m.$$

A set X is called a *relational code* if it is an (R,S)-code for some word relations R and S. If S is the identity relation ι, then an (R,S)-code is called a *strong*

R-code, or shortly just an *R-code*. A strong R-code is always a set where the elements are pairwise R-incompatible, but the converse does not hold generally. An (R, R)-code is called a *weak R-code*. An (ι, ι)-code is simply called a *code*. The definition coincides with the original definition of a variable length code.

Example 2.1. We consider the following set $X = \{ab, c\}$. This set is clearly a (prefix) code and also an (R, ι)-code for $R = \langle\{(a, c)\}\rangle$. On the other hand, for $R' = \langle\{(a, c), (b, c)\}\rangle$, we have $ab\,R'cc$. This shows that X is not an (R', R')-code.

The following results are proved in [9]. Suppose that R_1, R_2 and S are relations on A^* satisfying $R_1 \subset R_2$. If X is an (R_2, S)-code, then X is an (R_1, S)-code. Similarly, consider relations R, S_1 and S_2 satisfying $S_1 \subset S_2$. If X is an (R, S_1)-code, then X is an (R, S_2)-code. Note that (R, S)-codes are always (ι, ι)-codes, i.e., codes in the usual meaning.

Proposition 2.2 [9]**.** *Every (R, S)-code X is a code.*

Moreover, we have the following characterization of (R, S)-codes.

Proposition 2.3 [9]**.** *Let X be a subset of A^*. X is an (R, S)-code if and only if X is an (R, R)-code and $R_X \subseteq S_X$.*

We note that (R, S)-codes are more general than *pcodes* of partial words defined by F. Blanchet-Sadri in [4]. Indeed, pcodes are (R, S)-codes where $R = R_\uparrow$ and $S = \iota$. However, the concept of an (R, S)-code is more general. First, it enables us to consider weak R-codes. This case seems to be very natural and worth studying. Actually, by Proposition 2.3, all (R, S)-codes are weak R-codes, and therefore weak R-codes are the basis of our considerations. Secondly, we may consider more complex compatibility relations on letters than in the case of partial words. Namely, the compatibility relation of partial words has a very special structure. In $A_\diamondsuit$ there is a universal letter $\diamondsuit$ compatible with all other letters. Most reflexive and symmetric but not transitive relations are not of this form. For example, the "cyclic" relation $\langle\{(a, b), (b, c), (c, d), (d, a)\}\rangle$ does not have a universal letter.

In [14] Sardinas and Patterson gave their famous algorithm for deciding whether a given finite set X of words is a code or not. Blanchet-Sadri proved in [4] that the corresponding problem for partial words is decidable. Here we give a simple algorithm for the more general problem of deciding whether a given finite set X is an (R, S)-code or not. The essential part of the algorithm is to solve the problem for (R, R)-codes.

Algorithm 2.4 modified Sardinas-Patterson**.** *Let the input be a finite set $X \subseteq A^+$. Let $U_1 = R(X)^{-1}X \setminus \{\varepsilon\}$, and define $U_{n+1} = R(X)^{-1}U_n \cup R(U_n)^{-1}X$ for $n \geq 1$. Let $i \geq 2$ satisfy $U_i = U_{i-t}$ for some $t > 0$. Then X is a weak R-code if and only if*

$$\varepsilon \notin \bigcup_{j=1}^{i-1} U_j.$$

The proof of correctness of this algorithm in [9] is a modification of the proof of the Sardinas-Patterson algorithm in [2]. Thus, we have the following theorem. In [5] a similar kind of result for partial words is proved using the same guidelines.

Theorem 2.5 [9]. *The set X is a weak R-code if and only if none of the sets U_n contains the empty word.*

Note that there exist only finitely many different sets U_n, since all the lengths of the elements of U_n are less than $\max\{|x| \mid x \in X\}$. Secondly, if $U_i = U_j$ then, for any $t \geq 0$, $U_{i+t} = U_{j+t}$. Thus once a repetition in the sequence $U_1, U_2, \ldots$ is found, all U_i sets are found as well. Now it is clear by the previous proposition and Proposition 2.3 that the (R, S)-coding property of a finite subset X of A^+ can be verified by using Algorithm 2.4 and checking that $R_X \subseteq S_X$.

3. Minimal and maximal relations

Coding properties of an (R, S)-code X can be measured by defining maximal and minimal relations. These notions describe how well information can be decoded if R-compatibility of code words is allowed, or how much changes of the message, *i.e.* compatibility of words, can be allowed in order to decode the message with a precision of S. More precisely, let X be a subset of A^*. Let $S_{\min}(X, R)$ be the set of word relations S such that X is an (R, S)-code, and for all S' with $S' \subset S$, X is not an (R, S')-code. Similarly, let $S_{\max}(X, R)$ be the set of word relations S such that X is an (R, S)-code, and for all S' with $S \subset S'$, X is not an (R, S')-code. Relations belonging to $S_{\min}(X, R)$ (resp. $S_{\max}(X, R)$) are called *minimal* (resp. *maximal*) S-relations with respect to a set X and a relation R. The minimal and maximal relations in $R_{\min}(X, S)$ and $R_{\max}(X, S)$ are defined symmetrically.

Note that $S_{\max}(X, R) = \{\Omega\}$, where $\Omega = \Omega_{A^*}$ and $R_{\min}(X, S) = \{\iota\}$, where $\iota = \iota_{A^*}$, for all (R, R)-codes X. It can be proved that, for weak R-codes, $S_{\min}(X, R)$ is a unique element, but there may be several maximal relations belonging to $R_{\max}(X, S)$. For example, if $X = \{ab, bccb, ca\}$ and $S = \langle\{(a, b), (a, c)\}\rangle$, then $R_{\max}(X, S) = \{\langle\{(b, c)\}\rangle, \langle\{(a, b), (a, c)\}\rangle\}$. These two maximal R relations are by no means isomorphic. They do not even have the same *size*, *i.e.*, the number of pairs in the corresponding compatibility relation of letters.

In [9] two algorithms were given for finding minimal and maximal relations. Finding the minimal S-relation $S_{\min}(X, R)$ can be done in a polynomial time with respect to the size of the given set X. Finding the maximal R-relations in $R_{\max}(X, S)$ is a more complicated task. For a fixed alphabet, finding all the maximal relations R with respect to a given set X and a given word relation S can be done in polynomial time. This is based on a polynomial time version of the Sardinas-Patterson algorithm; see [6]. From another viewpoint, *i.e.*, if we allow arbitrary alphabets, the problem of finding maximal R relations is actually very difficult. The corresponding decision problem is namely *NP-complete*. Let us denote the size of a word relation R by $\mathrm{sz}(R)$. Define the number $M_R(X, S)$ to

be the maximal size of the relations in $R_{\max}(X,S)$, *i.e.*, $M_R(X,S) = \max\{\mathrm{sz}(R) \mid R \in R_{\max}(X,S)\}$. We formulate the following problem:

Problem: MAXIMAL RELATION
Instance: A set $X \subseteq A^+$, a relation S on A and a positive integer k
Question: Is $M_R(X,S) \geq k$?

This problem is related to an NP-complete problem called VERTEX COVER problem of graphs ([8], GT1). In [9] it was proved that VERTEX COVER can be polynomially reduced to MAXIMAL RELATION. Thus we have the following result.

Proposition 3.1 [9]**.** *The problem MAXIMAL RELATION is NP-complete.*

4. Relationally free monoids and stability

A monoid $M \subseteq A^*$ is (R,S)-*free* if it has a subset $B \subseteq M$ (called an (R,S)-*base* of M) such that

(i) $M = B^*$;
(ii) B is an (R,S)-code.

Strong R-freeness, weak R-freeness and freeness are defined similarly using the corresponding definitions of codes.

Remark 4.1. In [5] *pfreeness*, *i.e.*, the freeness of monoids of partial words was defined using *pinjective* morphisms: A monoid M is pfree if there exists a morphism $\varphi\colon B^* \to M$ of a free word monoid B^* onto M that satisfies

$$\varphi(x) \uparrow \varphi(y) \implies x = y.$$

Although pfreeness equals $(R_\uparrow, \iota)$-freeness, the above definition seems quite different from ours. In the matter of fact, there is no straightforward way to generalize the definition of pfreeness in [5] for (R,S)-codes, if $S \neq \iota$. For example, consider a morphism $\alpha\colon B^* \to X^*$, where $B = \{a,b,c,d\}$, $R = \langle\{(a,b),(c,d)\}\rangle$, $X = \{ab, abc, cdb, db\}$ and $\alpha(a) = ab$, $\alpha(b) = abc$, $\alpha(c) = cdb$, $\alpha(d) = db$. It is easy to show that α is "(R,R)-injective", i.e,

$$\alpha(x)\, R\, \alpha(y) \implies x\, R\, y.$$

The only nontrivial relation on X^* that we have to consider is $(ab)(cdb) = (abc)(db)$. Here we have $\alpha(ac)\, R\, \alpha(bd)$ and $ac\, R\, bd$. Hence, α satisfies the condition above. However, X^* is not (R,R)-free.

A subset B of a monoid M such that $M = B^*$ is called a *generating set* of M. A generating set is called *minimal* if no proper subset of B is a generating set of M. Each monoid $M \subseteq A^*$ has a unique minimal generating set consisting of the indecomposable elements of M, that is the set $(M \setminus \{\varepsilon\}) \setminus (M \setminus \{\varepsilon\})^2$. For a subset $X \subseteq A^*$, the unique minimal generating set of X^* is $(X \setminus \{\varepsilon\}) \setminus (X^+ \setminus \{\varepsilon\})^2$. The following result holds for all relationally free monoids. We note that this

proposition and other results concerning (R,S)-free monoids can be proved also using so called *unique factorization extensions*; see [10].

Proposition 4.2. *Let $X \subseteq A^*$. The set X is an (R,S)-code if and only if X^* is (R,S)-free and X is its minimal generating set.*

Proof. Suppose first that X is an (R,S)-code. By the definition of (R,S)-free monoids, it is clear that X^* is (R,S)-free. Thus it satisfies to show that X is the minimal generating set of X^*, *i.e.*, $X = (X \setminus \{\varepsilon\}) \setminus (X^+ \setminus \{\varepsilon\})^2$. Clearly, $X \supseteq (X \setminus \{\varepsilon\}) \setminus (X^+ \setminus \{\varepsilon\})^2$. Assume now that $x \in X$. By Proposition 2.2, X is a code. Therefore, $x \in X \setminus \{\varepsilon\}$ and $x \notin (X^+ \setminus \{\varepsilon\})^2$. Hence, $X \subseteq (X \setminus \{\varepsilon\}) \setminus (X^+ \setminus \{\varepsilon\})^2$.

Suppose then that X^* is (R,S)-free and X is its minimal generating set. Let Y be an (R,S)-base of X^*. Assume that $x \in X$. Since $Y^* = X^*$, we have $x = y_1 \dots y_m$ for some words $y_i \in Y$, where $i = 1, 2, \dots, m$. Moreover, $y_i = x_{i1} \dots x_{im_i}$, where $x_{ij} \in X$ for all i and j. Since $\varepsilon \notin Y \cup X$ and the elements of X are indecomposable, we conclude that $m = 1$ and $m_i = 1$. Thus, $X \subseteq Y$. Since Y satisfies the definition of an (R,S)-code, so does X. □

As a consequence of the previous proposition and Proposition 2.3 we have a characterization of (R,S)-free monoids using weak R-free monoids and a condition on the order of the relations R and S.

Proposition 4.3. *A monoid $M \subseteq A^*$ is (R,S)-free, if and only if M is (R,R)-free with (R,R)-base B and $R_B \subseteq S_B$.*

Proof. A monoid $M \subseteq A^*$ is (R,S)-free if and only if M has a base B such that B is an (R,S)-code. By Proposition 2.3, this is possible if and only if B is an (R,R)-code and $R_B \subseteq S_B$, *i.e.*, M is (R,R)-free and $R_B \subseteq S_B$. □

A monoid $M \subseteq A^*$ is called (R,S)-*stable* if for all for all $u, v, w, u', v', w' \in A^*$ satisfying conditions

(i) $u \mathrel{R} u'$, $w \mathrel{R} w'$ and $v \mathrel{R} v'$;
(ii) $uw, v, u', w'v' \in M$,

we have $u, w \in M$ and $u \mathrel{S} u'$. This situation is illustrated in Figure 1. As above, we talk about strong and weak R-stability depending on whether $S = \iota$ or $S = R$. The definition of (R,S)-stable monoids coincides with the original definition of stable monoids in the case $R = S = \iota$.

Relational stability can be used to characterize relationally free monoids like in the case of Schützenberger's criterion for normal (ι,ι)-free monoids. For pfree monoids, the stability criterion was first given in [5]. Our proof follows the guidelines of [2].

Proposition 4.4 (generalized Schützenberger's criterion)**.** *A submonoid of A^* is (R,S)-free if and only if it is (R,S)-stable.*

Proof. Let M be an (R,S)-stable submonoid of A^*. By Proposition 4.2 it is enough to show that the minimal generating set $X = (M \setminus \{\varepsilon\}) \setminus (M \setminus \{\varepsilon\})^2$ of M is an (R,S)-code. Suppose that this is not the case. Then there exist words $x_1, \dots, x_m, y_1, \dots, y_n \in X$ such that $x_1 \dots x_m \mathrel{R} y_1 \dots y_n$ and $(x_1, y_1) \notin S$.

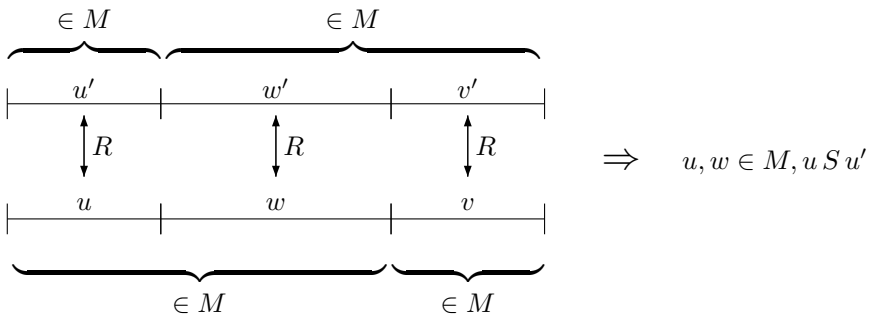

FIGURE 1. Illustration of (R, S)-stability.

We may suppose that $|x_1| \leq |y_1|$. Hence there exist words $w, z \in A^*$ such that $y_1 = wz$, $x_1 \, R \, w$ and $x_2 \dots x_m \, R \, z y_2 \dots y_n$. By (R, S)-stability, we have $w, z \in M$ and $w \, S \, x_1$. Note that $|w| = |x_1| > 0$ since $x_1 \in X \subseteq M \setminus \{\varepsilon\}$. Since y_1 is indecomposable and $w \neq \varepsilon$, we must have $z = \epsilon$. Thus it follows that $y_1 = w \, S \, x_1$. A contradiction.

Conversely, let M be (R, S)-free and let X be its (R, S)-base. Furthermore, assume that $uw, v, u', w'v' \in M$ satisfy $u \, R \, u'$ and $wv \, R \, w'v'$. We write the words $uw, v, u', w'v'$ as products of elements of the base X:

$$\begin{aligned} uw &= x_1 \dots x_k, \\ v &= v_1 \dots v_l, \\ u' &= u_1 \dots u_m, \\ w'v' &= y_1 \dots y_n. \end{aligned}$$

Since $u \, R \, u'$ and $wv \, R \, w'v'$, we have by the multiplicativity of word relations that

$$x_1 \dots x_k v_1 \dots v_l \, R \, u_1 \dots u_m y_1 \dots y_n.$$

Since X is an (R, S)-code, we conclude that $k + l = m + n$ and corresponding elements of the both sides are S-compatible and furthermore of the same length. Since $|u'| = |u| \leq |uw|$, we have

$$u' = u_1 \dots u_m \, S \, x_1 \dots x_m = u \quad \text{and} \quad w = x_{m+1} \cdots x_k \in X^*.$$

In other words, $u, w \in M$ and $u \, S \, u'$. Hence, M is (R, S)-stable. □

As a corollary of the previous proposition we get the following result concerning (R, S)-free semigroups. It is called here the generalized Tilson's result; see [16].

Proposition 4.5 (generalized Tilson's result)**.** *Any intersection of (R,S)-free monoids of A^* is (R,S)-free.*

Proof. Let M_i be (R,S)-free monoids for each $i \in \mathcal{I}$. Set $M = \cap_{i\in\mathcal{I}} M_i$. As an intersection of monoids, M is a monoid. Consider now words $u, v, w, u', v', w' \in A^*$ satisfying $u\,R\,u'$, $w\,R\,w'$, $v\,R\,v'$ and $uw, v, u', w'v' \in M$. Now $uw, v, u', w'v' \in M_i$ for each $i \in \mathcal{I}$ and by the (R,S)-stability of (R,S)-free monoids M_i (Proposition 4.4), we conclude that $u, w \in M_i$ and $u\,S\,u'$ for all $i \in \mathcal{I}$. Thus, $u, w \in M$ and M is (R,S)-stable. Hence, M is (R,S)-free by Proposition 4.4. □

5. Hulls

Let X be a set of words over A and consider now the following set of submonoids of A^*:

$$\mathcal{F}_{(R,S)}(X) = \{M \mid X^* \subseteq M \subseteq A^*, M \text{ is an } (R,S)\text{-free monoid}\}.$$

This is the set of all (R,S)-free submonoids of A^* containing X. Note that $\mathcal{F}_{(R,S)}(X)$ may be empty. For example, choose $X = \{ab, ac\}$, $R = \langle\{(b,c)\}\rangle$ and $S = \iota$. Let M be an (R,S)-free monoid containing X^* and let B be the base of M. Then $R_M \subseteq S_M$, since every R-compatible pair of words in M has a similar B-factorization and $R_B \subseteq S_B$ by Proposition 4.3. This is impossible, since considering relations $ab\,R\,ac$ and $ab \neq ac$ we see that $R_X \not\subseteq S_X$. Even more easily, we notice that there does not exists an (R,S)-free monoid containing $X = \{a, b\}$ for $R = \Omega$ and $S = \iota$.

On the other hand, it follows from Proposition 4.5 that the set $\mathcal{F}_{(R,S)}(X)$ is closed under intersection. Thus, if $\mathcal{F}_{(R,S)}(X)$ is nonempty, there exists a monoid

$$F_{(R,S)}(X) = \bigcap_{M\in\mathcal{F}_{(R,S)}(X)} M,$$

which is the smallest (R,S)-free monoid containing X. It is called the *(R,S)-free hull* of X. Unlike in the case of a normal free hull ($R = S = \iota$), the existence of $F_{(R,S)}(X)$ depends on the relations R and S and the set X itself. Proposition 4.3 implies that A^* is (R,S)-free if $R \subseteq S$. Then $A^* \in \mathcal{F}_{(R,S)}(X)$ and $F_{(R,S)}(X)$ exists. Moreover, we always have $\mathcal{F}_{(R,R)}(X) \neq \emptyset$, since A^* is (R,R)-free. The situation is characterized more precisely in the next proposition.

Proposition 5.1. *Let F_R be the weak R-free hull of X. The (R,S)-free hull of X exists, if and only if $R_{F_R} \subseteq S_{F_R}$, in which case $F_{(R,S)}(X) = F_R$.*

Proof. Suppose that $F_{(R,S)}(X)$ exists. By Proposition 4.3, it is (R,R)-free. Thus we must have $F_R \subseteq F_{(R,S)}(X)$ by the definition of F_R. Now the definition of (R,S)-freeness implies that $R_{F_{(R,S)}(X)} \subseteq S_{F_{(R,S)}(X)}$. Especially, this is valid for the subset F_R, *i.e.*, $R_{F_R} \subseteq S_{F_R}$.

Conversely, suppose that $R_{F_R} \subseteq S_{F_R}$. Let B be the base of the (R,R)-free hull F_R. Then $R_B \subseteq S_B$, since $B \subseteq F_R$ and, by Proposition 4.3, F_R is (R,S)-free. Hence, $\mathcal{F}_{(R,S)}(X)$ is not empty and $F_{(R,S)}(X)$ exists. Moreover, $F_{(R,S)}(X) \subseteq F_R$ by the definition of the (R,S)-free hull.

Furthermore, if $F_{(R,S)}(X)$ exists, we have showed that $F_R \subseteq F_{(R,S)}(X)$ and $F_{(R,S)}(X) \subseteq F_R$. Hence, $F_{(R,S)}(X) = F_R$ in this case. □

Next we consider a method to find the free hull in practice. Let X be a finite subset of A^+. In order to construct free monoids containing X, we must prevent "nontrivial" relations on X^+. We define the set

$$C_f(X) = \{(u,v) \in X \times X \mid (u,v) \notin R, uX^* \cap R(vX^*) \neq \emptyset\}.$$

By the definition, X is an (R,R)-code if and only if $C_f(X,R) = \emptyset$. Let us now define the following iterative procedure similar to the procedures introduced in [11].

Algorithm 5.2 (free hull A_f)**.** *Let the input be a finite set $X \subseteq A^+$ Set $X_0 = X$, and iterate for $j \geq 0$.*

(1) *Choose $(u,v) \in C_f(X_j,R)$ such that $u = u'u''$, where $|u'| = |v|$ and $u'' \in A^+$. If no such pair exists, then stop and return $A_f(X) = X_j$.*
(2) *Set $R'(u) = \{pref_{|u'|}(w) \mid w \in (R_{X_j})^+(u)\}$ and set $R''(u) = \{suf_{|u''|}(w) \mid w \in (R_{X_j})^+(u)\}$, where $(R_{X_j})^+$ is the transitive closure of R_{X_j}.*
(3) *Set $X_{j+1} = \big(X_j \setminus (R_{X_j})^+(u)\big) \cup R'(u) \cup R''(u)$.*

Note that in each iteration at least one of the words in X_j is factorized into two proper factors, since $\varepsilon \notin X_j$ for any $j \geq 0$. For a finite set of words there are only finitely many factors of words, and therefore the algorithm must terminate. Now we prove that the previous algorithm computes the free hull of X.

Proposition 5.3. *Let X be a finite subset of A^+. Then Algorithm 5.2 with input X returns the base B of the (R,R)-free hull of X, i.e., $B = A_f(X)$.*

Proof. As mentioned above the algorithm A_f always terminates with finite input $X \subseteq A^+$. Suppose now that the algorithm terminates after k iterations. Let us first show by induction that $X_j^* \subseteq F_{(R,R)}(X)$ for all $j = 0,1,\ldots,k$. The case $j = 0$ is clear by the definition of $F_{(R,R)}(X)$. Suppose now that $X_j^* \subseteq F_{(R,R)}(X)$ and the pair $(u,v) \in C_f(X_j,R)$ is chosen in Step (1). By the stability condition (Proposition 4.4) and the induction assumption $X_j^* \subseteq F_{(R,R)}(X)$, we conclude that the sets $R'(u)$ and $R''(u)$ must be subsets of $F_{(R,R)}(X)$. Since $(R_{X_j})^+(u) \subseteq R'(u)R''(u)$, we have $X_j^* \subseteq X_{j+1}^* \subseteq F_{(R,R)}(X)$.

Since $C_f(X_k,R) = \emptyset$, the monoid X_k^* is (R,R)-free by the definition of (R,R)-freeness. Hence $X_k^* \subseteq F_{(R,R)}(X)$ by the above and the minimality of the free hull implies that $X_k^* = F_{(R,R)}(X)$. Since X_k consists only of the indecomposable elements of X_k^*, it is the (R,R)-base B of $F_{(R,R)}(X)$. In other words, $A_f(X) = X_k = B$. □

6. Defect effect

The well know *defect theorem* of words says that if a set of n words satisfies a nontrivial relation, then these words can be expressed simultaneously as products of less than n words. This is the so called defect effect. See [3,11] for more on defect theorems of words.

We formulate now a defect effect with respect to a word relation R. Note that the original defect theorem does not hold in general and we need a new nontrivial formulation for the defect in the relational case. Let X be a finite subset of A^*. Let us consider a graph $G_R(X) = (V, E)$ defined as follows. The vertices are the words in X, and $(u, v) \in E$ if and only if $u\,R\,v$. We consider the connected components of G. Denote the transitive closure of R by R^+ as above. We note that the set of vertices in the connected component containing x is exactly $(R_X)^+(x)$. Denote the number of connected components of $G_R(X)$ by $c(X, R)$. We formulate a defect effect of words with respect to a word relation in the following way.

Theorem 6.1. *Let X be a finite subset of A^+ and let B be the base of the (R, R)-hull of X. Then $c(B, R) \leq c(X, R)$, and the equality holds if and only if X is an (R, R)-code.*

Proof. If X is an (R, R)-code, then $B = X$ by Proposition 4.2 and the equality holds trivially. Suppose now that X is not an (R, R)-code. Thus, $X^* \neq F_{(R,R)}(X)$. By Proposition 5.3, Algorithm 5.2 computes the base of the (R, R)-free hull correctly.

Let X_j be any intermediate set of the procedure such that X_j is not a code. First we prove that the number of connected components cannot increase in any iteration step from $G_R(X_j)$ to $G_R(X_{j+1})$. Assume that $(u, v) \in C_f(X_j, R)$ is chosen in Step (1) and $u = u'u''$, where $|u'| = |v|$ and $u'' \in A^+$. More precisely, suppose that $us\,R\,vt$, where $u, v \in X_j$ and $s, t \in X_j^*$. Denote $t = t_1 \dots t_n$, where $t_i \in X_j$ for all $i = 1, 2, \dots, n$. Let us denote the set of vertices in the connected component of X_j containing a vertex u by $V_j(u)$. Since u splits into two parts u' and u'', we have components $V_{j+1}(u')$ and $V_{j+1}(u'')$ in $G(X_{j+1}, R)$. Thus, the number of connected components could increase by two. On the other hand, the whole connected component $V_j(u)$ disappears. In addition, we know that $u'\,R\,v$ and therefore $V_{j+1}(u') = V_{j+1}(v) \supseteq V_j(v) \cup R'(u)$. Thus, the new vertices $R'(u)$ are connected to an old component containing v. Therefore, the number of connected components does not increase.

Now it remains to show that in some stage of the procedure the number of components strictly decreases. Suppose that X_j is the last intermediate step before the output X_{j+1} and words u and v satisfy the conditions described above. Suppose first that $t_1\,R\,u''$. This means that $V_{j+1}(u'') = V_{j+1}(t_1)$. In other words $V_j(u)$ disappears and both $V_{j+1}(u')$ and $V_{j+1}(u'')$ are joint to old components. Thus the number of connected components has decreased by one.

Suppose next that $t_1 \notin R(u'')$. If $t_1 \notin (R_{X_j})^+(u)$, then t_1 is not split and $t_1 \in X_{j+1}$. Hence, the relation $us\,R\,vt$ implies the following nontrivial relation in

$X_{j+1} \times X_{j+1}$:

$$u'' s\, R\, t_1 \dots t_n.$$

Therefore $(u'', t_1) \in C_f(X_{j+1})$ and X_{j+1} is not the final outcome of the algorithm A_f. A contradiction. Thus, we must have $t_1 \in (R_{X_j})^+(u)$. We may denote $t_1 = t_1' t_1''$, where $|t_1'| = |u'|$, $|t_1''| = |u''|$ and $t_1' \in (R_{X_{j+1}})^+(u')$. If $t_1' \notin R(u'')$, then

$$u'' s\, R\, t_1' t_1'' t_2 \dots t_n$$

is a nontrivial relation in X_{j+1}. This is again impossible, since $C_f(X_{j+1}, R)$ must be empty. Thus $t_1'\, R\, u''$. We have

$$u''\, (R_{X_{j+1}})^+\, t_1'\, (R_{X_{j+1}})^+\, u'\, (R_{X_{j+1}})^+\, v.$$

In other words, $V_{j+1}(u'') = V_{j+1}(v)$ and the number of connected components must be reduced by one. This proves the defect effect for free (R, R)-hulls. □

Finally we note that the defect effect is valid also for (R, S)-free hulls by Proposition 5.1. Namely, if B is the base of the (R, S)-free hull of X, then it is the base of the (R, R)-free hull and $c(B, R) \leq c(X, R)$. Like above the equality holds if and only if X is an (R, S)-code. Moreover, since partial words can be seen as a special case of words with word relations, the previous theorem implies a defect theorem of partial words [5]. Recall that pcodes are $(R_\uparrow, \iota)$-codes over $A_\diamond$ and pfree monoids are $(R_\uparrow, \iota)$-free. We may state:

Corollary 6.2. *Let X be a finite set of partial words, i.e., a set of words over the alphabet $A_\diamond$. Suppose that the pfree hull of X exists and let B be its base. Then $|B| \leq |X|$, and the equality holds if and only if X is a pcode.*

Proof. As mentioned above the pfree hull is the $(R_\uparrow, \iota)$-free hull of X. By Proposition 5.1, $(R_\uparrow, \iota)$-free hull is also the $(R_\uparrow, R_\uparrow)$-free hull of X. Thus, by Theorem 6.1, we have $c(B, R_\uparrow) \leq c(X, R_\uparrow)$ and the equality holds if and only if X is an $(R_\uparrow, R_\uparrow)$-code. Since $M = B^*$ is an $(R_\uparrow, \iota)$-free monoid, we have $(R_\uparrow)_M \subseteq \iota_M$. This means that all the connected components of $G_{R_\uparrow}(B)$ and $G_{R_\uparrow}(X)$ must consist of single elements. Thus $c(B, R_\uparrow) = |B|$ and $c(X, R_\uparrow) = |X|$. This implies our statement. □

Note that Theorem 6.1 is more general than the defect theorem of partial words in [5]. Firstly, the (R, R)-free hull always exists, which is not the case for pfree hulls. Secondly, we want to point out that no algorithm for finding the base of the pfree hull is given in [5] but in our paper the algorithm for finding the (R, R)-base is essential. In the Algorithm 5.2 the simultaneous splitting of the elements in a connected component is crucial to our proof of the defect effect. Such splitting is not needed in the case of partial words. Namely, if the pfree hull of partial words exists, then the connected components $V_j(u)$ appearing in our proof must be trivial, *i.e.*, $V_j(u) = u$.

References

[1] J. Berstel and L. Boasson, Partial words and a theorem of Fine and Wilf. *Theoret. Comput. Sci.* **218** (1999) 135–141.

[2] J. Berstel and D. Perrin, *Theory of Codes*. Academic press, New York (1985).

[3] J. Berstel, D. Perrin, J.F. Perrot and A. Restivo, Sur le théorème du défaut. *J. Algebra* **60** (1979) 169–180.

[4] F. Blanchet-Sadri, Codes, orderings, and partial words. *Theoret. Comput. Sci.* **329** (2004) 177–202.

[5] F. Blanchet-Sadri and M. Moorefield, *Pcodes of partial words*, manuscript. http://www.uncg.edu/mat/pcode (2005).

[6] M. Crochemore and W. Rytter, *Jewels of Stringology*. World Scientific Publishing (2002).

[7] A. Ehrenfeucht and G. Rozenberg, Elementary homomorphisms and a solution of the D0L sequence equivalence problem. *Theoret. Comput. Sci.* **7** (1978) 169–183.

[8] M.R. Garey and D.S. Johnson, *Computer and Intractability: A Guide to the Theory of NP-Completeness*. Freeman, New York (1979).

[9] V. Halava, T. Harju and T. Kärki, Relational codes of words. *Theoret. Comput. Sci.* **389** (2007) 237–249.

[10] V. Halava, T. Harju and T. Kärki, Defect theorems with compatibility relation. *Semigroup Forum* (to appear, available online).

[11] T. Harju and J. Karhumäki, Many aspects of defect theorems. *Theoret. Comput. Sci.* **324** (2004) 35–54.

[12] P. Leupold, Partial words for DNA coding. *Lect. Notes Comput. Sci.* **3384** (2005) 224–234.

[13] M. Linna, The decidability of the D0L prefix problem. *Int. J. Comput. Math.* **6** (1977) 127–142.

[14] A.A. Sardinas and G.W. Patterson, A necessary and sufficient condition for the unique decomposition of coded messages. *IRE Internat. Conv. Rec.* **8** (1953) 104–108.

[15] J.C. Spehner, Présentation et présentations simplifiables d'un monoïd simplifiable. *Semigroup Forum* **14** (1977) 295–329.

[16] B. Tilson, The intersection of free submonoids of free monoids is free. *Semigroup Forum* **4** (1972) 345–350.