# Distances between languages and reflexivity of relations

Christian Choffrut[1] and Giovanni Pighizzini[2]

[1] LIAFA, Université Paris 7
Tour 55–56, 2 Place Jussieu, 75251 Paris Cedex 05 (cc@litp.ibp.fr)
[2] Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano
via Comelico, 39 – 20135 Milano – Italy (pighizzi@dsi.unimi.it)

**Abstract.** In this paper, the notions of $k$–reflexivity and almost reflexivity of binary relations over the free monoid $A^*$ are introduced, which are natural extensions of the usual reflexivity. They can be defined relative to arbitrary distances over $A^*$. The problems of deciding whether or not a relation is almost reflexive and whether or not it is $k$–reflexive for a given integer $k$, are studied. It is shown that both problems are unsolvable in the case of deterministic rational relations. Moreover, the latter problem remains undecidable even when an oracle asserting that the relation is almost reflexive is provided. On the other hand, for the subclass of recognizable relations, both problems are shown to be solvable: as a consequence, the distance between two rational languages can be effectively computed. Further decidability results concerning the intermediate class of synchronized rational relations are proved.

## 1 Introduction

Comparing strings has been intensively studied during the past two decades, for example in connection with bibliographic search, text editing, alignments of biologic sequences *etc.* The problem usually consists in associating with two strings a non negative integer which reports how close they are from each other. This usually defines a distance between words, the most popular of which is certainly the one based on the maximal common subsequence, the so-called Levenshtein distance.

The basic question concerning the subsequence distance and its extension the edit distance [Kru83] asks whether or not we can design a sequential algorithm that computes the distance in time less than $O(n^2)$, where $n$ is the maximal length of the two words.

Here we tackle the problem of computing the distance between two subsets of words. We assume these subsets to be effectively given. Actually we restrict the effectiveness of the subsets to some types of finite memory device. This is a reasonable assumption since there is no hope of actually computing the distance outside this class as we easily show in this paper.

Let $A$ be a finite alphabet and let $A^*$ be the free monoid it generates. The passage from the distance between strings to the distance between subsets $X, Y \subseteq A^*$ of strings is done in the standard way by resorting to the Hausdorff distance. In particular, whenever $X$ and $Y$ are rational subsets of $A^*$ and whenever we endow the free monoid with the Hamming distance, the problem can be reduced to Hashiguchi's result on distance-automata (caveat: the term distance here refers to a different concept, weight would be more appropriate).

It is worthwhile observing that some fundamental problems can be stated in terms of distances. For example, asking whether or not the distance, whatever the distance, between two subsets is equal to 0 amounts to asking whether or not they are equal. This question is undecidable for context-free languages, so this gives a limit to our expectations and shows that we are quickly facing difficult problems.

Intuitively, asserting that two subsets $X$ and $Y$ have finite distance is a claim that the binary relation $X \times Y$ is "almost" reflexive with respect to the distance under consideration. Say a relation $R \subseteq A^* \times A^*$ is $k$–reflexive, for some integer $k \leq \infty$, if every element $x$ of the domain is at distance at most $k$ from some element $y$ of the range, with $(x, y) \in R$, and vice versa. It is almost reflexive if $k < \infty$. We are led to consider the more general question of asking whether or not a given relation over $A^*$ is almost reflexive. J. Johnson [Joh86] proved that reflexivity is undecidable for rational relations. In this work, we strengthen this result by showing that the problem remains undecidable even when restricted to *deterministic* rational relations. This result is a new contribution to the theory of rational relations and has some important consequences concerning the problems we are interested in in this paper. First of all, we cannot decide whether or not a deterministic rational relation is $k$–reflexive, for a given integer $k$, with respect to any distance. Moreover, in the case of the distances studied in this work, namely the Hamming, subword and edit distances, this problem remains undecidable even in the presence of an oracle asserting that the relation is almost reflexive. In addition, we prove that it is undecidable whether or not a deterministic rational relation is almost reflexive, i.e., whether or not $k < \infty$.

Subsequently, in order to identify the largest classes of relations for which the previously mentioned properties are decidable, we restrict our attention to proper subfamilies of deterministic rational relations, in particular that of synchronized rational relations [FS93] and that of recognizable relations. Recall that the class of recognizable relations is a proper subclass of that of synchronized rational relations which is itself a proper subclass of that of deterministic rational relations (assuming $A$ and $B$ are non empty).

We show that Hashiguchi's above mentioned problem is very close to the problem of deciding whether or not a synchronized rational relation is almost reflexive with respect to the Hamming distance. In fact, each problem can be immediately encoded into the other. As a consequence, we prove that the smallest $k \leq \infty$ such that a synchronized rational relation is $k$–reflexive with respect to the Hamming distance can be effectively computed.

In the case of the subword and of the edit distances, the largest class for which we show that this computation can be done effectively is that of recognizable relations. This result does not seem to be extendable to synchronized rational relations. However, we are able to prove that given an oracle asserting that a synchronized rational relation is almost reflexive, then it is decidable whether or not it is $k$–reflexive, for a given integer $k$, with respect to the subword or edit distance. These results are summarized in the following table

| | $Rec$ | $Syn$ | $DRat$ |
|---|---|---|---|
| Is $R$ $k$–reflexive? | $S$ | $S$ | $U$ |
| Is $R$ almost reflexive? | $S$ | $S/?$ | $U$ |
| Find the smallest $k$ s.t. $R$ is $k$–reflexive | $S$ | $S/?$ | $U$ |
| Given $R$ almost reflexive, find the smallest $k$ s.t. $R$ is $k$–reflexive | $S$ | $S$ | $U$ |

$S$=Solvable, $U$=Unsolvable
$S/?$=Solvable for the Hamming distance, unknown for the edit and the subword distance
$Rec$=Recognizable relations $Syn$=Synchronized rational relations
$DRat$=Deterministic rational relations

We conclude this introduction, by mentioning that in the unary case, namely in the case of relations defined over a one letter alphabet, rational relations are equivalent to Presburger formulas [GS66]. Using this fact, we have shown that in the unary case all the considered problems are solvable. For lack of space, this result is omitted in this version of the paper. For the same reason, some of the proofs have been just outlined.

## 2 Preliminary definitions and results

Given a finite alphabet $A$, let $A^*$ be the free monoid of strings (or words) over $A$, and let $\epsilon$ be the empty word. Given a string $x \in A^*$, we denote by $|x|$ its length and by $x_i$ its $i$-th symbol (or letter), $i = 1, \ldots, |x|$. We recall that a string $z \in A^*$ is a *factor* or *subword* of $x$ if there exist two strings $u, v \in A^*$ such that $x = uzv$.

A *distance* over $A^*$ is a function $d : A^* \times A^* \to \mathbb{N} \cup \{\infty\}$ which is reflexive, symmetric, and which satisfies the triangular inequality: for all $x, y, z$, $d(x, y) \leq d(x, z) + d(z, y)$ holds.

In this paper, we consider three distances over $A^*$. Given two strings $x, y \in A^*$, the *Hamming distance* reports the number of positions where the symbols of $x$ differ from those of $y$; the *subword distance* is based on the longest common subword of $x$ and $y$; the *edit distance* is defined as the minimum number of *edit operations* (substitution, insertion, deletion of a symbol) which transform $x$ into $y$. Formally

**Definition 1.** Given $x, y \in A^*$, the *Hamming*, *subword*, and *edit distance* between $x$ and $y$, denoted by $d_H(x, y)$, $d_s(x, y)$, and $d_e(x, y)$, respectively, are defined as

- $d_H(x,y) = \begin{cases} \#\{i \mid x_i \neq y_i\} & \text{if } |x| = |y| \\ \infty & \text{otherwise;} \end{cases}$
- $d_s(x, y) = |x| + |y| - 2\max\{|z| \mid z \text{ is a subword of both } x \text{ and } y\}$;
- $d_e(x, y) = \min\{k \mid x \stackrel{k}{\vdash} y\}$, where for $w, z \in A^*$, $w \vdash z$ if and only if there exist $u, v \in A^*$, $a, b \in A$, with $a \neq b$, such that either $w = uav$, $z = ubv$ (substitution), or $w = uav$, $z = uv$ (deletion), or $w = uv$, $z = ubv$ (insertion), and, as usual, $\stackrel{k}{\vdash}$ denotes the $k$-th power of $\vdash$.

Every distance over $A^*$ can be extended to a distance between a string and a language or between two languages. The distance between a string $x$ and a language $L$ is defined as the minimum of the distances between $x$ and the strings belonging to $L$; the distance between two languages can be defined in a standard way, by resorting to the Hausdorff distance. More precisely

**Definition 2.** Let $A$ be a finite alphabet and $d : A^* \times A^* \to \mathbb{N} \cup \{\infty\}$ a distance. Given a string $x \in A^*$ and a language $L \subseteq A^*$, the *distance* $d(x, L)$ between $x$ and $L$ is defined as

$$d(x, L) = \begin{cases} \min\{d(x, y) \mid y \in L\} & \text{if } L \neq \emptyset \\ \infty & \text{otherwise.} \end{cases}$$

Given two languages $L', L'' \subseteq A^*$, the *distance* $d(L', L'')$ between $L'$ and $L''$ is defined as

$$d(L', L'') = \max\{\sup\{d(x, L'') \mid x \in L'\}, \sup\{d(x, L') \mid x \in L''\}\}.$$

Recalling that the equivalence problem is undecidable for context–free languages, it is immediate that the distance between context–free languages is not computable. So, it is natural to restrict our attention to subclasses of context–free languages. On the other hand, the study of the distance between two languages can be generalized to the study of reflexivity properties of relations. To this aim, we now recall some basic definitions concerning binary relations over strings, in order to introduce the notion of $k$–reflexive relation.

Given a binary relation $R \subseteq A^* \times B^*$, let $Dom(R)$ and $Ran(R)$ be the domain and the range of $R$, respectively, namely, $Dom(R) = \{x \mid \exists y \text{ s.t. } (x, y) \in R\}$, and $Ran(R) = \{y \mid \exists x \text{ s.t. } (x, y) \in R\}$, and let $R^{-1}$ be the relation $\{(y, x) \mid (x, y) \in R\}$.

Given a language $L \subseteq A^*$, let $i_L = \{(x, x) \mid x \in L\}$ be the identity relation restricted to $L$. A relation $R \subseteq A^* \times A^*$ is *reflexive* if and only if, for any $x \in Dom(R) \cup Ran(R)$, the pair $(x, x)$ belongs to $R$, namely, $i_{Dom(R) \cup Ran(R)} \subseteq R$. We now extend the notion of reflexivity, by introducing the *$k$–reflexivity*

**Definition 3.** Given a distance $d$ over $A^*$, a relation $R \subseteq A^* \times A^*$, and an integer $k \in \mathbb{N} \cup \{\infty\}$, the relation $R$ is *$k$–reflexive* with respect to $d$, if and only if the following two statements are true

- $\forall x \in Dom(R)\ \exists y \in A^* \text{ s.t. } (x, y) \in R$ and $d(x, y) \leq k$;
- $\forall y \in Ran(R)\ \exists x \in A^* \text{ s.t. } (x, y) \in R$ and $d(x, y) \leq k$.

The relation $R$ is *almost* reflexive (with respect to $d$) if and only if it is $k$–reflexive for some $k < \infty$.

Note that, for any distance $d$, a relation is reflexive if and only if it is 0–reflexive. Furthermore, it is not difficult to observe that the distance between two languages $L'$ and $L''$ is bounded by $k$ if and only if the relation $L' \times L''$ is $k$–reflexive. Hence, $L'$ and $L''$ have finite distance if and only if the relation $L' \times L''$ is almost reflexive.

We now briefly recall the definition of the main classes of relations we are interested in. For more details the reader is referred to [Eil74, Joh86, FS93].

As usual, given a monoid $M$, we denote by $Rec(M)$ the family of *recognizable* subsets of $M$ (i.e., all subsets of $M$ that are the union of some equivalence classes of a congruence of finite index) and by $Rat(M)$ the family of *rational* subsets of $M$ (i.e., all the subsets that are obtained from singletons by a finite application of the operations of set union, concatenation and Kleene star). The family $Rat(A^* \times B^*)$ of *rational relations* over $A^* \times B^*$ can be characterized in terms of rational transducers, i.e., of two-tape automata. By considering deterministic and synchronized transducers, we define, respectively, the classes $DRat(A^* \times B^*)$ and $Syn(A^* \times B^*)$ of *deterministic* and *synchronized rational relations* over $A^* \times B^*$ [FS93].

The following proper inclusions are well known [RS59, Eil74, Joh86, FS93] (we assume $\#A \times \#B \geq 2$ where $\#A$ denotes the cardinality of $A$)

$$Rec(A^* \times B^*) \subset Syn(A^* \times B^*) \subset DRat(A^* \times B^*) \subset Rat(A^* \times B^*).$$

We conclude this section by recalling that a *distance* automaton $\mathcal{A}$ (actually a finite automaton with a *weight* function) [Has82, Has90] is an ordinary nondeterministic automaton $(A, Q, \delta, I, F)$, along with a function $\omega : Q \times A \times Q \to \mathbb{N} \cup \{\infty\}$ such that, for $q, p \in Q$, $a \in A$, $\omega(q, a, p) = \infty$ if and only if $p \notin \delta(q, a)$.

The weight function $\omega$ can be extended to strings, by setting, for $q, p \in Q$, $x, w \in A^*$, $a \in A$

- $\omega(q, \epsilon, p) = \begin{cases} 0 & \text{if } q = p \\ \infty & \text{otherwise;} \end{cases}$
- $\omega(q, wa, p) = \min\{\omega(q, w, r) + \omega(r, a, p) \mid r \in Q\}$;
- $\omega(x) = \min\{\omega(q, x, p) \mid q \in I, p \in F\}$.

The weight of $\mathcal{A}$, denoted by $\omega(\mathcal{A})$, is the maximum of the weights of the strings accepted by the underlying automaton, namely $\omega(\mathcal{A}) = \sup\{\omega(x) \mid x \in L(\mathcal{A})\}$, with the convention that if $\mathcal{A}$ accepts the empty set, then $\omega(\mathcal{A}) = \infty$. The following result was proved by Hashiguchi in [Has82, Has90].

**Theorem 4.** *Given an automaton $\mathcal{A}$ with a weight function $\omega$, the weight $\omega(\mathcal{A})$ is computable.*

## 3 Undecidability results

In this section, we begin the study by showing that $k$–reflexivity is undecidable, for deterministic rational relations. The results we state are mainly based on the following theorem, which extends to deterministic rational relations a result proved in [Joh86] for rational relations

**Theorem 5.** *It is undecidable whether or not a deterministic rational relation $R \subseteq A^* \times A^*$ is reflexive, i.e., $i_{Dom(R) \cup Ran(R)} \subseteq R$.*

*Proof.* Given a deterministic rational relation $R \subseteq A^* \times A^*$, let $X = A^* - Dom(R)$ and $Y = A^* - Ran(R)$. Then $R \cup (X \times Y)$ is deterministic too. Moreover, $R$ is reflexive if and only if the diagonal $i_{A^*}$ is contained in $R \cup (X \times Y)$. Now, we show that the problem of deciding whether or not a deterministic rational relation contains the diagonal is undecidable.

Given an instance $u_1, \ldots, u_p, v_1, \ldots, v_p \in A^*$ of Post Correspondence Problem (PCP) such that $\{u_1, \ldots, u_p\}$ and $\{v_1, \ldots, v_p\}$ are prefix codes, consider the relation $R = \{(u_i, v_i) \mid i = 1, \ldots, n\}^+$. It is not difficult to see that $R \in DRat(A^* \times A^*)$. Moreover, $DRat(A^* \times A^*)$ is closed under complementation. Hence $R^c \in DRat(A^* \times A^*)$. Note that $R^c$ contains the diagonal $i_{A^*}$ if and only if the given instance of PCP does not have a solution. Since PCP is undecidable even for prefix codes [Ruo85], we conclude that the reflexivity is undecidable. □

As an immediate consequence of Theorem 5, we can state the following undecidability result

**Corollary 6.** *Let $d$ be a fixed distance over $A^*$. It is undecidable, given a deterministic rational relation $R \in DRat(A^* \times A^*)$ and an integer $k$, whether or not $R$ is $k$–reflexive, relative to $d$.*

Now, in the case of the Hamming, subword and edit distance, we strengthen the result obtained in Corollary 6 in two directions. On the one hand, we show that it is undecidable whether or not a deterministic rational relation is almost reflexive (namely, we just require the existence of a finite bound $k$, we do not require its value). On the other hand, we show that the problem of deciding whether or not a deterministic rational relation is $k$–reflexive remains undecidable even when we know it is almost reflexive.

**Theorem 7.** *It is undecidable whether or not a deterministic rational relation is almost reflexive, relative to the Hamming, the subword, and the edit distance, respectively.*

*Proof.* (outline) Given $R \in DRat(A^* \times A^*)$, let $R_\$ = \{(\$u\$, \$v\$) \mid (u, v) \in R\}$, where $\$ \notin A$. Of course, $R_\$$ and its Kleene closure $R_\$^*$ are deterministic.

We prove that $R$ is reflexive if and only if $R_\$^*$ is almost reflexive, and then, by Theorem 5, we conclude that this property is undecidable. The "only if" part is trivial. To prove the converse implication, suppose that $R_\$^*$ is $k$–reflexive, for an integer $k < \infty$. Thus, given $u \in Dom(R)$ and $x = (\$u\$)^{k+1}$, there exists $y \in (A \cup \{\$\})^*$ such that $(x, y) \in R_\$^*$ and $d(x, y) \leq k$.

It is not difficult to see that in the case of the Hamming and the edit distances, the string $y$ can be decomposed as the product of $k+1$ factors $y_0, y_1, \ldots, y_k$ such that $d(x, y) = d(\$u\$, y_0) + d(\$u\$, y_1) + \ldots + d(\$u\$, y_k)$. This implies that $y_i = \$u\$$, for some $i$, $0 \leq i \leq k$. Thus, $(\$u\$, \$u\$) \in R_\$$ and $(u, u) \in R$.

In the case of the subword distance, for $k = 0$ it is immediate to observe that $(u, u) \in R$. For $k \geq 1$, we have $d_s(x, y) = |x| + |y| - 2l_s(x, y) \leq k$, where $l_s(x, y)$ is the length of the longest common subword of $x$ and $y$. So $2l_s(x, y) \geq |x| + |y| - k \geq (|u| + 2)(k + 1) + l_s(x, y) - k$, and then $l_s(x, y) \geq 2|u| + 3$. Thus, $\$u\$$ should be a subword of both $x$ and $y$, and, by construction, $(\$u\$, \$u\$) \in R_\$$, $(u, u) \in R$.

Dually, if $R_\$^*$ is $k$–reflexive, then $v \in Ran(R)$ implies $(v, v) \in R$. □

**Theorem 8.** *It is undecidable whether or not a $k$–reflexive deterministic rational relation $R \in DRat(A^* \times A^*)$ is $(k-1)$–reflexive, relative to the Hamming, the subword, and the edit distance, respectively.*

*Proof.* We prove the result by showing that given a 1–reflexive deterministic rational relation, it is undecidable whether it is reflexive too.

First, we observe that the problem of deciding whether a deterministic rational relation $R$ is reflexive is undecidable even if $\epsilon \notin Dom(R) \cup Ran(R)$. Let $R$ be a deterministic rational relation satisfying such a restriction, and let $\sharp$ be a symbol not belonging to $A$. Consider the relation $R^\sharp = R \cup R_0 \cup R_1 \cup R_2$, where $R_0 = \{(u, u^\sharp) \mid u \in Dom(R)\}$, $R_1 = \{(u^\sharp, u) \mid u \in Ran(R)\}$, $R_2 = \{(u^\sharp, u^\sharp) \mid u \in Dom(R) \cup Ran(R)\}$, and, for any $u \in A^*$, $i = 1, \ldots, k$, $u^\sharp$ denotes the word $\sharp u_2 \ldots u_{|u|}$.

It is not difficult to observe that $R^\sharp$ is a 1–reflexive deterministic rational relation, with respect to the considered distances. Furthermore, $R^\sharp$ is reflexive if and only if $R$ is reflexive, but, by Theorem 5, this is undecidable. □

Since the relation $R^\sharp$ used to prove Theorem 8 is almost reflexive, actually 1–reflexive, we immediately obtain

**Corollary 9.** *Given an almost reflexive deterministic rational relation $R$ and an integer $k \in \mathbb{N}$, it is undecidable whether or not $R$ is $k$–reflexive, with respect to the Hamming, the subword and the edit distance, respectively.*

## 4 $k$–reflexivity relative to the Hamming distance

In Section 3, we have shown that the problem of deciding whether a deterministic rational relation is $k$–reflexive with respect to the Hamming distance is unsolvable. Now, we prove that this problem turns out to be solvable when we restrict our attention to the class of synchronized rational relations.

**Theorem 10.** *For any synchronized rational relation $R$, it is possible to compute the smallest $k \leq \infty$, such that $R$ is $k$–reflexive relative to the Hamming distance.*

*Proof.* (outline) Given $R \in Syn(A^* \times A^*)$, let $\mathcal{A} = (A, A, Q, I, F, E)$ be a synchronized transducer accepting $R$. Thus, for any transition $(q, x, y, p) \in E$ of $\mathcal{A}$, $(x, y) \in (A \cup \{\epsilon\}) \times (A \cup \{\epsilon\})$, and $(x, y) \neq (\epsilon, \epsilon)$ holds [FS93].

We consider an automaton $\mathcal{A}'$ obtained by "forgetting" the third component of $\mathcal{A}$-transitions, namely for any $(q, a, b, p) \in E$, with $a, b \in A$, the automaton $\mathcal{A}'$ has a transition from $q$ to $p$ on $a$. If $a = b$ then the weight of this transition is 0, otherwise it is 1. This automaton recognizes $Dom(R)$. Dually, we can define an automaton $\mathcal{A}''$, which recognizes $Ran(R)$, with a suitable weight function $\omega''$.

It is not difficult to verify that $R$ is $k$–reflexive if and only if $\omega'(\mathcal{A}') \leq k$ and $\omega''(\mathcal{A}'') \leq k$. □

Actually, the problem of deciding whether or not a synchronized rational relation is $k$–reflexive with respect to the Hamming distance and the problem of computing weights of automata are closely related. Indeed, Theorem 10 shows a reduction in one direction. The converse also holds.

**Theorem 11.** *Given an automaton $\mathcal{A} = (A, Q, \delta, I, F)$, with a weight function $\omega$, there exists a synchronized rational relation $R$ such that $\omega(\mathcal{A})$ is finite if and only if $R$ is almost reflexive.*

## 5 $k$–reflexivity relative to the edit and to the subword distance

In this section we study the problem of deciding whether or not a relation is $k$–reflexive with respect to the edit and to the subword distance, respectively. In particular, we show that this problem is solvable for recognizable relations. As in the case of the Hamming distance and of synchronized relations, these results are obtained by means of reductions to the problem of Hashiguchi.

We start by showing the following lemma

**Lemma 12.** *For any recognizable relation $R \subseteq A^* \times A^*$ there exists an automaton $\mathcal{A}$ with a weight function $\omega$, such that $\mathcal{A}$ recognizes $Dom(R)$, and for any $x \in Dom(R)$, $x \neq \epsilon$, $\omega(x) = \min\{d_e(x, y) \mid (x, y) \in R\}$.*

*Proof.* (outline) We sketch the proof for $R = L' \times L''$ with $L'$ and $L''$ rational languages and $L'' \neq \emptyset$. The proof can be easily extended to the general case (all recognizable relations are finite unions of relations of this type).

Let $\mathcal{A}' = (A, Q', \delta', q'_0, F')$ and $\mathcal{A}'' = (A, Q'', \delta'', q''_0, F'')$ be two finite deterministic automata accepting the languages $L'$ and $L''$, respectively. We consider the product automaton $\mathcal{A} = (A, Q, \delta, I, F)$, with $Q = Q' \times Q''$, $I = \{(q'_0, q''_0)\}$, $F = F' \times F''$, and $\delta((q', q''), a) = \{(\delta'(q', a), \delta''(q'', y)) \mid y \in A^*\}$ for $a \in A$, $q' \in Q'$, $q'' \in Q''$. Intuitively, each $\mathcal{A}'$ transition is "padded" with $\mathcal{A}''$-paths. The language accepted by $\mathcal{A}$ coincides with the language $L'$ accepted by $\mathcal{A}'$. Moreover, each $\mathcal{A}$-accepting path corresponds to some $\mathcal{A}''$-accepting path.

We define the weight function $\omega$ of $\mathcal{A}$, in such a way that the edit distance between a string $x \in L'$ and the language $L''$ can be obtained as the minimum of the weights of accepting paths on $x$ in $\mathcal{A}$. As in the dynamic programming, we obtain the minimum from "local" minima. To this aim, we define $\omega$ for $q', p' \in Q'$, $q'', p'' \in Q''$, $a \in A$, as follows

$$\omega((q', q''), a, (p', p'')) = \begin{cases} \infty & \text{if } \delta'(q', a) \neq p' \text{ or } \{y \mid \delta''(q'', y) = p''\} = \emptyset \\ 1 & \text{if } \delta'(q', a) = p',\ q'' = p'', \text{ and } \delta''(q'', a) \neq q'' \\ n & \text{otherwise,} \end{cases}$$

where $n = \min(\{|y| \mid y \in (A - \{a\})^* \text{ and } \delta''(q'', y) = p''\} \cup \{|y| - 1 \mid y \in A^* a A^* \text{ and } \delta''(q'', y) = p''\})$.

Some comments are now in order. In the first case, the transition from $(q', q'')$ to $(p', p'')$ on $a$ in $\mathcal{A}$ is forbidden either because there is no transition from $q'$ to $p'$ on $a$ in $\mathcal{A}'$, or because $p''$ is not reachable from $q''$ in $\mathcal{A}''$.

In the second and in the third cases, the symbol $a$ is compared with each string $y$ such that, in the automaton $\mathcal{A}''$, $p''$ is reachable from $q''$ on $y$. Let us denote by $L''_{q''p''}$ the set of such strings. If $q''$ and $p''$ coincide, then $L''_{q''p''}$ contains the empty string, so the distance between $a$ and $L''_{q''p''}$ is at most 1. More precisely, if $a \notin L''_{q''p''}$, i.e., $q'' \neq \delta''(q'', a)$, then this distance is exactly 1 (second case: the symbol $a$ is deleted); otherwise, it is 0 (third case, with $y = a$).

If states $q''$ and $p''$ do not coincide, then each string $y$ belonging to $L''_{q''p''}$ contains at least one symbol. If no symbol of $y$ coincides with $a$, then $|y|$ edit operations are needed to obtain $y$ from $a$: one operation to replace $a$ by a symbol of $y$, the other $|y|-1$ operations to insert the remaining symbols of $y$. On the other hand, if a symbol of $y$ coincides with $a$, then the substitution is not needed; thus, the total number of edit operations is $|y|-1$.

It is possible to prove that, for $q', p' \in Q'$, $q'', p'' \in Q''$, $x \in A^+$, the following equality holds

$$\omega((q', q''), x, (p', p'')) = \begin{cases} \min\{d_e(x, y) \mid \delta''(q'', y) = p''\} \\ \quad \text{if } \{y \mid \delta''(q'', y) = p''\} \neq \emptyset \text{ and } \delta'(q', x) = p' \\ \infty \ \text{ otherwise.} \end{cases}$$

Thus, for $x \in L'$, $x \neq \epsilon$, the weight of $x$ in $\mathcal{A}$ can be computed as

$$\omega(x) = \min\{\omega((q'_0, q''_0), x, (p', p'')) \mid (p', p'') \in F\} = \min\{d_e(x, y) \mid y \in L''\}. \ \square$$

Now, we are able to prove the main result of this section.

**Theorem 13.** *For any recognizable relation $R$, it is possible to compute the smallest integer $k$ such that $R$ is $k$–reflexive with respect to the edit distance.*

*Proof.* (outline) Given $R \in Rec(A^* \times A^*)$, we can define two automata $\mathcal{A}$, $\mathcal{A}'$, with weight functions $\omega$, $\omega'$, respectively, as in Lemma 12. If $\epsilon \notin Dom(R) \cup Ran(R)$, then we can show that, for $k \in \mathbb{N}$, $R$ is $k$–reflexive if and only if $\omega(\mathcal{A}) \leq k$ and $\omega'(\mathcal{A}') \leq k$. This allows to compute the smallest $k$ such that $R$ is $k$–reflexive. The proof can be easily adapted to the case of recognizable relations which contain the empty word in the domain or in the range. $\square$

Now, it is natural to investigate whether or not Theorem 13 can be extended to the class of synchronized rational relations. On the one hand, in the case of the Hamming distance we have to compare symbols in corresponding positions, so the fact of having a synchronized transducer is very useful (see the proof of Theorem 10), while in the case of the edit distance we have to compare symbols in different positions: hence, dealing with synchronized relations is of no special help. On the other hand, we are able to prove that given a constant $k$, the problem of deciding whether or not a synchronized rational relation is $k$–reflexive relative to the edit distance is solvable. As a consequence, if we know that a synchronized rational relation $R$ is almost reflexive, then we are able to obtain the minimum integer $k$ such that $R$ is $k$–reflexive (recall that, by Theorem 8, this problem is unsolvable for deteterministic rational relations).

**Theorem 14.** *For any synchronized rational relation $R \subseteq A^* \times A^*$ and any integer $k \geq 0$, it is possible to decide whether or not $R$ is $k$–reflexive, with respect to the edit distance.*

*Proof.* For $j \geq 0$, let $I_j = \{(u, v) \in A^* \times A^* \mid d_e(u, v) \leq j\}$. Observe that $I_0 = i_{A^*}$, $I_1 = i_{A^*} \cdot ((A \cup \{\epsilon\}) \times (A \cup \{\epsilon\})) \cdot i_{A^*}$, and $I_j = I_{j-1} \circ I_1$, where $\circ$ denotes the composition of relations.

So, the relation $I_1$ is rational and has length difference bounded by 1, namely $||u| - |v|| \leq 1$, for $(u, v) \in I_1$. By Corollary 4.2 in [FS93], this implies that $I_1$ is a synchronized rational relation. Furthermore, the class $Syn(A^* \times A^*)$ is closed under composition and intersection. So, $R \cap I_k = \{(x, y) \in R \mid d_e(x, y) \leq k\} \in Syn(A^* \times A^*)$, for any $k \geq 0$. Hence, $R$ is $k$–reflexive if and only if the domain and the range of $R \cap I_k$ coincide with the domain and the range of $R$, respectively. Since domains and ranges of rational relations are rational languages, this property is decidable. □

The decidability results relative to the edit distance can be extended even to the subword distance. In particular, by considering a different reduction to the problem of Hashiguchi, it can be shown that the smallest integer $k$ such that a recognizable relation $R$ is $k$–reflexive, with respect to the subword distance, can be effectively computed. Furthermore, the proof of Theorem 14 can be easily adapted to the subword distance.

Finally, we point out that using the results of [Web94], it is possible to evaluate the complexity of the decision procedures presented in this paper.

## References


[Eil74] S. Eilenberg. *Automata, languages and machines*, volume A. Academic Press, 1974.

[FS93] C. Frougny and J. Sakarovitch. Synchronized rational relations of finite and infinite words. *Theoretical Computer Science*, 108:45–82, 1993.

[GS66] S. Ginsburg and E. Spanier. Semigroups, Presburger formulas, and languages. *Pacific Journal of Mathematics*, 16:285–296, 1966.

[Has82] K. Hashiguchi. Limitedness theorem on finite automata with distance functions. *Journal of Computer and System Sciences*, 24:233–244, 1982.

[Has90] K. Hashiguchi. Improved limitedness theorems on finite automata with distance functions. *Theoretical Computer Science*, 72:27–38, 1990.

[Joh86] J. Johnson. Rational equivalence relations. *Theoretical Computer Science*, 47:39–60, 1986.

[Kru83] J. Kruskal. An overview of sequence comparison. In D. Sankoff and J. Kruskal, editors, *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*, pages 1–44. Addison–Wesley, 1983.

[RS59] M.O. Rabin and D. Scott. Finite automata and their decision problems. *IBM J. Res. Develop*, 3:125–144, 1959. Also in: E.F. Moore. *Sequential machines*. Addison–Wesley, 1964.

[Ruo85] K. Ruohonen. Reversible machines and Post's correspondence problem for biprefix morphisms. *J. Inform. Process. Cybernet. EIK*, 21:579–595, 1985.

[Web94] A. Weber. Finite–valued distance automata. *Theoretical Computer Science*, 134:225–251, 1994.