Journal of Automata, Languages and Combinatorics 9 (2004) 2/3, 293–309

DESCRIPTIONAL COMPLEXITY OF ERROR/EDIT SYSTEMS[1,2]

LILA KARI
Department of Computer Science, University of Western Ontario, London, Ontario, N6A 5B7 Canada
e-mail: lila@csd.uwo.ca

and

STAVROS KONSTANTINIDIS
Department of Mathematics and Computing Science, Saint Mary's University, Halifax, Nova Scotia, B3H 3C3, Canada
e-mail: s.konstantinidis@stmarys.ca

ABSTRACT

Errors appear in a wide range of information processing and transmission applications, such as data communications, biological computing, computer typesetting, speech recognition, etc. It can be said indeed that errors are truly natural phenomena. In this work we introduce error or edit systems (e-systems, for short), which are formal languages over the alphabet of the basic edit operations. Our formalism allows one to model essentially any kind of error situations. For certain natural regular e-systems, we investigate their descriptional complexity in terms of the number of states of the automata accepting such systems. This problem is of interest in its own right as well as in the computation of maximal error-correcting capabilities of known languages.

Keywords: Error models, automata, descriptional complexity, formal languages

The capacity to blunder slightly is the real marvel of DNA. Without this attribute we still would be anaerobic bacteria and there would be no music.

(Lewis Thomas, 'The medusa and the snail')

1. Introduction

Errors appear in a wide range of information processing and transmission applications, such as data communications [11], biological computing [14, 5, 6], computer typesetting [10], speech recognition [1], etc. It can be said indeed that errors are

[1] Full version of a submission presented at the 4th Workshop on *Descriptional Complexity of Automata, Grammars and Related Structures* (London, Ontario, Canada, August 21–24, 2002).

[2] Research partially supported by Grants R2824A01 and R220259 of the Natural Sciences and Engineering Research Council of Canada.

truly natural phenomena. In the past, little has been done in representing explicitly errors as formal objects of study – in the sense of formal languages for instance. In [3], for example, there is a brief discussion on edit scripts (words over the alphabet of the basic edit operations), but no constraint is imposed on the occurrences of the various edit operations. Perhaps one of the first attempts to describe formally and systematically languages of errors appears in [4] and [8].

The present work introduces the formalism of error or edit systems (e-systems, for short), which supercedes the formalism in [8] in two ways: First, e-systems are merely formal languages over the alphabet of the basic edit operations, a fact that allows one to study e-systems using, for instance, tools from automata theory. Second, the new formalism allows one to model any kind of error situations that can arise in typical information processing applications. It should be noted that, in defining error situations, we follow the combinatorial approach as opposed to the probabilistic (or information theoretic) one. More specifically, we model only error situations with high enough probability of occurrence, and omit all the rest that one might define using the alphabet of the basic edit operations. This (combinatorial) kind of approach is followed, for instance, in the classical theory of error correcting codes [13].

Apart from the new formalism, we investigate the descriptional complexity of certain natural regular[3] e-systems. These can be used to model various constraints on the type and frequency (or density) of the permitted errors. In particular, we model e-systems with scattered as well as burst errors of any error type and density. These are regular e-systems and, therefore, the descriptional complexity of such a system can be given in terms of the number of states of the minimal automaton accepting the system. This problem is of interest in its own right as well as in the computation of maximal error-correcting capabilities of known languages [7].

The paper is structured as follows. The next section contains some basic terminology and notation from formal languages and automata. In Section 3, we introduce e-systems and error types, and define the concept of (deterministic) descriptional complexity for classes of regular e-systems. In Section 4, we define e-systems with scattered errors and provide upper- and lower-bounds on the descriptional complexity of such systems. In Section 5, we define e-systems with burst errors and provide exact expressions for the complexity of those systems. Section 6 investigates the problem of combining the errors of two e-systems in such a way that the new e-system includes the errors of both e-systems and preserves the constraints on the errors according to each of the two systems. It is shown that the complexity of the new system is no higher than the product of the complexities of the two e-systems. Finally, Section 7 contains a few concluding remarks.

2. Basic Notation

An alphabet is a finite nonempty set of symbols. In the sequel we shall use a fixed alphabet Σ. A word or string (over Σ) is a finite sequence $a_1 \dots a_n$ such that each a_i is in Σ. The length of a word w is denoted by $|w|$. The empty word, denoted λ,

[3] In the sense of formal language theory.

is the word of length zero. If X is a subset of the alphabet and w is a word then $|w|_X$ is the number of symbols in X that occur in w. For example, if $\Sigma = \{a, b, c\}$ and $X = \{a, c\}$ then $|aabbcc|_X = 4$, whereas $|aabbcc| = 6$. We write w_1w_2 for the word obtained by concatenating the words w_1 and w_2. If w is a word and n is a nonnegative integer, then w^n is the word that consists of n concatenated copies of w. In particular, $w^0 = \lambda$. A word z is a *factor* of a word w, if $w = w_1zw_2$ for some words w_1 and w_2. A language is a set of words. The language of all words is denoted by Σ^*. We shall use standard language operations, such as the concatenation and Kleene star operations, as well as the notation of regular expressions – see [15], for instance.

A deterministic finite automaton, or *DFA* for short, is a quintuple $A = (X, Q, s, F, t)$ such that X is an alphabet, Q is a finite nonempty set, the set of state symbols (or states for short), s is the start state in Q, F is a subset of Q, the set of final states, and t is a partial function of $Q \times X$ into Q, called the transition function. The automaton is said to be *complete* if t is a total function. The transition function t is extended to $Q \times X^*$ as follows: for every state q, $t(q, \lambda) = q$, and for every symbol a in X and every word w in X^*, $t(q, aw) = t(t(q, a), w)$. The language accepted by the automaton A, denoted by $L(A)$, is the set of words w such that $t(s, w)$ is a state in F. A *computation* of A is a string of the form $q_0a_1q_1 \dots a_nq_n$ such that each a_i is in X, each q_j is a state, and $t(q_{i-1}, a_i) = q_i$. It should be clear that w is in $L(A)$ if and only if there is a computation as above such that $w = a_1 \dots a_n$, $q_0 = s$ and q_n is a final state. Such a computation is called *accepting*. An *extended computation* of A is a string of the form $q_0a_1q_1 \dots a_nq_n$ such that each a_i is in $X \cup \{\lambda\}$, each q_j is a state, and $t(q_{i-1}, a_i) = q_i$ – note that $q_{i-1} = q_i$ when $a_i = \lambda$. It should be clear that w is in $L(A)$ if and only if there is an extended computation as above such that $w = a_1 \dots a_n$, $q_0 = s$ and q_n is a final state.

In this paper, the size of a DFA A, denoted by $|A|$, is the number of states of A. If A and A' are two DFAs then $A \cap A'$ is the DFA obtained by using the standard product construction such that $L(A \cap A') = L(A) \cap L(A')$. Moreover, $|A \cap A'| = |A||A'|$.

A language L is called *regular* if there is a DFA accepting L. For a regular language L we write C_L for the number of states of a minimal complete DFA accepting L. Obviously, if A is an arbitrary complete DFA accepting L then $C_L \leq |A|$.

3. Error/Edit Systems

The alphabet E of the *basic edit operations* is the set of all symbols x/y such that $x, y \in \Sigma \cup \{\lambda\}$ and at least one of x and y is in Σ. If x/y is in E and x is not equal to y then we call x/y an *error*. We write λ/λ for the empty word over the alphabet E. We note that λ is used as a formal symbol in the elements of E. For example, if x and y are in Σ then $(x/\lambda)(x/y) \neq (x/x)(\lambda/y)$. Of course, λ is the empty word over Σ; that is $w\lambda = \lambda w = w$ for all words w in Σ^*. The elements of E^* are called *e-strings*. The *weight* of an e-string h, say, is the number of errors occurring in h. The *input* and *output parts* of an e-string $h = (x_1/y_1) \dots (x_n/y_n)$ are the words (over Σ) $x_1 \dots x_n$ and $y_1 \dots y_n$, respectively. We write $\mathrm{inp}(h)$ for the input part and $\mathrm{out}(h)$ for the output part of the e-string h. The *size* of h is the length of $\mathrm{inp}(h)$. An *e-system* is a subset of E^* (or a language over E).

The term error type has been used informally in several works concerning errors. Our formalism allows us to give a precise definition of that term. We write ε for the subset $\{x/x \mid x \in \Sigma\}$ of E. An *error type* τ is a nonempty subset of E that is disjoint from ε. If D is an e-system then θ_D is the *error type of* D; that is the set of all errors that appear in the e-strings of D. Equivalently, θ_D is the smallest subset of $E \setminus \varepsilon$ such that $D \subseteq (\theta_D \cup \varepsilon)^*$.

Now we provide examples of certain error types. The SID error types are $\sigma = \{x/y \mid x, y \in \Sigma \text{ and } x \neq y\}$, $\iota = \{\lambda/x \mid x \in \Sigma\}$, and $\delta = \{x/\lambda \mid x \in \Sigma\}$; they are called substitution, insertion, and deletion types, respectively. These error types are important in various domains [8]. In the context of biomolecular computing and bioinformatics, where a DNA strand can be interpreted as a word over the four-letter alphabet $\{A, C, G, T\}$ several types of errors can occur [12]. The most common ones are insertions and deletions of one or more letters, as well as one-letter substitutions. The latter can be either *transitions* or *transversions*. Transitions occur when one purine is replaced by another purine (A/G or G/A) or one pyrimidine with another pyrimidine (T/C or C/T). Transversions result when a purine is replaced by a pyrimidine or vice versa. Because the structural changes leading to transitions are relatively small, in real-life genomic DNA they occur more frequently than transversions (which require more substantial modifications of the molecule). Thus, we define the transition error type to be (it) $= \{C/T, T/C, A/G, G/A\}$ and the transversion error type to be (ve) $= \{C/A, A/C, C/G, G/C, T/A, A/T, T/G, G/T\}$. Obviously, (it), (ve) $\subseteq \sigma$.

We note that the above definition of error type concerns only the basic, or atomic, error types. Other composite error types such as transpositions – a common kind of error in computer typesetting – can be defined by means of e-systems.

Note that every (finite) transducer T, say, in standard form can be viewed as an automaton (nondeterministic, in general) accepting an e-system $D(T)$. It should be clear, however, that the relation $R(T)$ realized by T is different from $D(T)$, as $R(T)$ consists of pairs of words. In [9], a set of pairs of words is viewed as a discrete channel. The fact that (w, z) is in the channel means that the input word w can be received as z via the channel. Thus a discrete channel in the sense of [9] models the effects of errors on words, whereas an e-system models the errors themselves. Now every e-system D defines a channel γ_D consisting of all pairs (w, z) such that $w = \text{inp}(h)$ and $z = \text{out}(h)$ for some e-string h in D.

In [9] it is shown that the problem of whether a given regular language is error-correcting (or -detecting) for a given rational channel is decidable in polynomial time. The time complexity depends of course on the efficiency of the channel description and, therefore, it is desirable to investigate the descriptional complexity of regular e-systems. This issue is also important in [7], where the efficiency of algorithms for computing maximal error-correcting capabilities of languages depends on the efficiency of representing e-systems.

Definition 1 Let P be a recursive index (or parameter) set and let $D = \{D(p) \mid p \in P\}$ be a set of regular e-systems. The *(deterministic) descriptional complexity of* D is the function $C_D : P \to \mathbf{N}$ such that $C_D(p)$ is the number of states

of a minimal complete DFA accepting $D(p)$.

In the next two sections we define certain classes of e-systems by specifying the type of errors permitted, the way of combining errors, and the frequency (or density) of the errors. In the theory of error correcting codes there are two major approaches of combining errors: scattered or burst. In the expressions for denoting e-systems we shall use the symbol **s** for scattered errors and **b** for burst errors. We shall use **x** to indicate either of **s** and **b**.

4. Complexity of E-Systems with Scattered Errors

Consider two positive integers m and n, with $m < n - 1$, and an error type τ. The e-system $[\tau\mathrm{s}](m, n)$ consists of all e-strings that contain up to m (scattered) errors of type τ in any factor of length n, or less, of the input part of the e-string. More formally, an e-string h is in $[\tau\mathrm{s}](m, n)$ if and only if for every factor g of h with $|\operatorname{inp}(g)| \leq n$, one has that $|g|_\tau \leq m$. The ratio m/n is called the *error density* of the e-system $[\tau\mathrm{s}](m, n)$. For example, let $\Sigma = \{x, y\}$ and let $h = (x/x)(x/y)(y/\lambda)(x/x)(y/y)(x/\lambda)(y/x)$ be an e-string. As the factor $(x/y)(y/\lambda)(x/x)(y/y)(x/\lambda)$ of h is of size 5 and weight 3, it follows that h is not in the e-system $[(\sigma \cup \delta)\mathrm{s}](2, 5)$. Similarly, h is not in $[(\sigma \cup \delta)\mathrm{s}](3, 6)$ either, but it is in $[(\sigma \cup \delta)\mathrm{s}](3, 5)$. Now let

$$g = (\lambda/x)(y/y)(x/y)(x/x)(y/y)(\lambda/y)(x/x)(x/x)(x/x)(\lambda/y)(\lambda/y).$$

As the factor $(\lambda/x)(y/y)(x/y)(x/x)(y/y)(\lambda/y)(x/x)$ of g is of size 5 and weight 3, we have that g is not in $[(\sigma \cup \iota)\mathrm{s}](2, 5)$, but it is in $[(\sigma \cup \iota)\mathrm{s}](3, 5)$.

In this section, we are interested in the descriptional complexity of the e-systems

$$[\tau\mathrm{s}] = \{[\tau\mathrm{s}](m, n) \mid 1 < m < n - 1\}.$$

E-systems of the form $[\tau\mathrm{s}](1, n)$ can be regarded as systems with burst errors. These are considered in the next section.

Shorthand Notation: In describing the transition function t, say, of a DFA accepting an e-system, we shall often use the notation $t(p, \theta) = q$ as a shorthand for '$t(p, e) = q$ for all e in θ', where θ is a nonempty finite set of input symbols and p and q are states of the DFA in question. Moreover, for any state q of the DFA in question, we write $L(q)$ to denote the language accepted by the DFA when q is used as the start state.

We shall use the following facts from combinatorics [2].

- For all nonnegative integers r and s, with $r \geq s$,

$$\binom{r}{s} = \binom{r}{r-s}, \quad \binom{r+1}{s+1} = \frac{r+1-s}{s+1}\binom{r+1}{s} \text{ and}$$

$$\binom{r+1}{s} = \frac{r+1}{r+1-s}\binom{r}{s} \geq \binom{r}{s}.$$

- For all positive integers r and s, with $r \geq s$,

$$\binom{r}{s} = \binom{r-1}{s} + \binom{r-1}{s-1} \text{ and } \binom{r}{s} = \sum_{i=s}^{r} \binom{i-1}{s-1}.$$

- For all nonnegative integers r, s, and t, with $s \geq t$,

$$\sum_{i=0}^{t} \binom{r}{i}\binom{s}{t-i} = \binom{r+s}{t} \text{ and } \sum_{i=0}^{t} \binom{r+i}{i} = \binom{r+1+t}{t}.$$

In the above, we assume that $\binom{r}{i} = 0$ when $i > r$. We also need the following lemma.

Lemma 1 *For all nonnegative integers r and t,*

$$\frac{2t+r+1}{4t+r+2}\binom{r+2t+2}{r+1} \leq \sum_{i=0}^{t} \binom{r+2i+1}{r} \leq \frac{r+1}{r+2}\binom{r+2t+2}{r+1}.$$

Proof. Let T be the sum $\sum_{i=0}^{2t+1} \binom{r+i}{i}$. Then $T = S_0 + S_1$, where $S_0 = \sum_{i=0}^{t} \binom{r+2i}{r}$ and $S_1 = \sum_{i=0}^{t} \binom{r+2i+1}{r}$. This implies that

$$T = S_0 + \sum_{i=0}^{t} \binom{r+2i}{2i} + \sum_{i=0}^{t} \binom{r+2i}{2i+1} = 2S_0 + \sum_{i=0}^{t} \frac{r}{2i+1}\binom{r+2i}{2i}$$

and, therefore, $T \leq 2S_0 + rS_0$ and $T \geq 2S_0 + r/(2t+1)S_0$. The claim follows now from the facts that $T = \binom{r+2t+2}{r+1}$, $S_1 = T - S_0 \leq T - 1/(r+2)T$, and $S_1 \geq T - (2t+1)/(4t+r+2)T$. □

Theorem 2 *Let τ be an error type involving no insertions, that is $\tau \cap \iota = \emptyset$. For every parameters m and n, with $1 < m < n-1$,*

$$1 + \binom{n-1}{m-1} \leq C_{[\tau s]}(m,n) \leq 1 + \sum_{i=0}^{m} \binom{n-1}{i}.$$

Proof. First we define an automaton $A = (\varepsilon \cup \tau, Q, [\bar{\varepsilon}], Q \setminus [-1], t)$ accepting exactly the e-strings of the e-system $[\tau s](m,n)$. We shall use $\bar{\varepsilon}$ and $\bar{\tau}$ as symbols representing the sets ε and τ, respectively. Symbols of the form $\bar{\eta}$ are words over the alphabet $\{\bar{\varepsilon}, \bar{\tau}\}$; that is $\bar{\eta} \in \{\bar{\varepsilon}, \bar{\tau}\}^*$. Then $\bar{\eta}$ represents the corresponding set η, a subset of $(\varepsilon \cup \tau)^*$, in a natural manner. For example, if $\bar{\eta} = \bar{\varepsilon}\bar{\tau}\bar{\varepsilon}$ then $\eta = \varepsilon\tau\varepsilon$. The states $[-1]$ and $[\bar{\varepsilon}]$ are the sink and start states, respectively. The rest of the states are of the form $[\bar{\eta}]$, where $\bar{\eta}$ is a nonempty word of length up to $n-1$ such that $\bar{\eta}$ starts with $\bar{\tau}$ and contains at most m $\bar{\tau}$'s; hence, $1 \leq |\bar{\eta}|_{\bar{\tau}} \leq m$. State $[\bar{\eta}]$ means that the last $|\bar{\eta}|$ input symbols read, say $e_1 \ldots e_{|\bar{\eta}|}$, are in η and the next input symbols to read are independent of any symbols read prior to $e_1 \ldots e_{|\bar{\eta}|}$.

Before we define the transition function t, observe that for every state $[\bar{\eta}]$ there is a word $\bar{\eta}_0$ such that $\bar{\eta} = \bar{\tau}\bar{\varepsilon}^k\bar{\eta}_0$, for some $k \geq 0$, such that $\bar{\eta}_0$ is either empty or in $\bar{\tau}\{\bar{\varepsilon}, \bar{\tau}\}^*$. Now we proceed with defining t:

- $t([\bar{\varepsilon}], \varepsilon) = [\bar{\varepsilon}]$ and $t([\bar{\varepsilon}], \tau) = [\bar{\tau}]$.
- For every state $[\bar{\eta}]$ with length of $\bar{\eta}$ less than $n-1$, $t([\bar{\eta}], \varepsilon) = [\bar{\eta}\bar{\varepsilon}]$ and

$$t([\bar{\eta}], \tau) = \begin{cases} [\bar{\eta}\bar{\tau}], & \text{if } |\bar{\eta}|_{\bar{\tau}} < m, \\ [-1], & \text{if } |\bar{\eta}|_{\bar{\tau}} = m. \end{cases}$$

- For every state $[\bar{\eta}]$ with length of $\bar{\eta}$ equal to $n-1$, $t([\bar{\eta}], \varepsilon) = [\bar{\eta}_0\bar{\varepsilon}]$ and

$$t([\bar{\eta}], \tau) = \begin{cases} [\bar{\eta}_0\bar{\tau}], & \text{if } |\bar{\eta}|_{\bar{\tau}} < m, \\ [-1], & \text{if } |\bar{\eta}|_{\bar{\tau}} = m. \end{cases}$$

It should be clear that for every state $[\bar{\eta}]$ there is an e-string h in η such that $t([\bar{\varepsilon}], h) = [\bar{\eta}]$; hence, every state in Q is reachable from the start state. Next we prove the following claims:

1. The cardinality of Q is $1 + \sum_{i=0}^{m} \binom{n-1}{i}$.
2. For any r in $\{0, \ldots, \min\{m-2, n-m-1\}\}$, the set of states

$$T_r = \{[\bar{\tau}\bar{\varepsilon}^r\bar{\psi}] \mid |\bar{\psi}|_{\bar{\tau}} \geq m-r-2,\ |\bar{\tau}\bar{\varepsilon}^r\bar{\psi}| = n-1\}$$

is of cardinality $\sum_{i=m-r-2}^{m-1} \binom{n-2-r}{i}$, such that no two (different) states of T_r are equivalent.

The required statement follows from the above claims when we note that

$$|T_0| = \binom{n-2}{m-2} + \binom{n-2}{m-1} = \binom{n-1}{m-1}.$$

For the first claim, note firstly that the cardinality of the set $Q' = Q \setminus \{[-1], [\bar{\varepsilon}], [\bar{\tau}]\}$ is equal to the number of words $\bar{\psi}$ of length l, with $1 \leq l \leq n-2$, containing up to $m-1$ symbols $\bar{\tau}$; that is, $0 \leq |\bar{\psi}|_{\bar{\tau}} \leq m-1$. Also, as $|\bar{\psi}|_{\bar{\tau}} \leq l$, it follows that

$$|Q'| = \sum_{l=1}^{n-2} \sum_{i=0}^{\min\{l, m-1\}} \binom{l}{i} = (n-2) + \sum_{l=1}^{n-2} \sum_{i=1}^{\min\{l, m-1\}} \binom{l}{i}.$$

By expanding the sums and rearranging the resulting terms, it follows that

$$|Q'| = (n-2) + \sum_{i=1}^{m-1} \sum_{l=i}^{n-2} \binom{l}{i} = (n-2) + \sum_{j=2}^{m} \sum_{l=j}^{n-1} \binom{l-1}{j-1}.$$

Hence,

$$|Q'| = (n-2) + \sum_{j=2}^{m} \binom{n-1}{j} = -2 + \sum_{j=0}^{m} \binom{n-1}{j}.$$

The first claim follows now, as $|Q| = 3 + |Q'|$.

For the second claim, note firstly that the cardinality of the set T_r is equal to the number of words $\bar{\psi}$ of length $n-r-2$, containing at least $m-r-2$ and at most $m-1$ symbols $\bar{\tau}$. Now let $[\bar{\tau}\bar{\varepsilon}^r\bar{\psi}_1]$ and $[\bar{\tau}\bar{\varepsilon}^r\bar{\psi}_2]$ be two different states in T_r

such that $|\bar{\psi}_2|_{\bar{\tau}} \leq |\bar{\psi}_1|_{\bar{\tau}}$. We need to show that these states are not equivalent. For this, define the word $\bar{\tau}_1 \dots \bar{\tau}_{r+2}$ such that each $\bar{\tau}_i$ is either $\bar{\tau}$ when the last $n-1$ symbols of $\bar{\tau}\bar{\varepsilon}^r\bar{\psi}_2\bar{\tau}_1 \dots \bar{\tau}_{i-1}$ contain fewer than m symbols $\bar{\tau}$, or $\bar{\varepsilon}$ otherwise. Then, as $|\bar{\psi}_2|_{\bar{\tau}} + (r+2) \geq m$, it follows that the word $\bar{\psi}_2\bar{\tau}_1 \dots \bar{\tau}_{r+2}$ of length n contains exactly m symbols $\bar{\tau}$. Hence, $\tau_1 \dots \tau_{r+2}$ is a subset of $L([\bar{\tau}\bar{\varepsilon}^r\bar{\psi}_2])$. If $|\bar{\psi}_2|_{\bar{\tau}} < |\bar{\psi}_1|_{\bar{\tau}}$ then the word $\bar{\psi}_1\bar{\tau}_1 \dots \bar{\tau}_{r+2}$ of length n contains more than m symbols $\bar{\tau}$, which implies that $\tau_1 \dots \tau_{r+2}$ and $L([\bar{\tau}\bar{\varepsilon}^r\bar{\psi}_1])$ are disjoint. Hence, the two states are not equivalent when $|\bar{\psi}_2|_{\bar{\tau}} < |\bar{\psi}_1|_{\bar{\tau}}$. Now suppose that $|\bar{\psi}_2|_{\bar{\tau}} = |\bar{\psi}_1|_{\bar{\tau}}$, and let k be the first position in which the words $\bar{\tau}\bar{\varepsilon}^r\bar{\psi}_1$ and $\bar{\tau}\bar{\varepsilon}^r\bar{\psi}_2$ differ. Then $k \geq r+2$ and there are symbols $\bar{\theta}_1, \bar{\theta}_2$ and words $\bar{\phi}, \bar{\phi}_1, \bar{\phi}_2$ such that $\bar{\theta}_1 \neq \bar{\theta}_2$ and $|\bar{\phi}_1| = |\bar{\phi}_2| > 0$ and $\bar{\tau}\bar{\varepsilon}^r\bar{\psi}_1 = \bar{\tau}\bar{\varepsilon}^r\bar{\phi}\bar{\theta}_1\bar{\phi}_1$ and $\bar{\tau}\bar{\varepsilon}^r\bar{\psi}_2 = \bar{\tau}\bar{\varepsilon}^r\bar{\phi}\bar{\theta}_2\bar{\phi}_2$. Without loss of generality, assume $\bar{\theta}_1 = \bar{\tau}$ and $\bar{\theta}_2 = \bar{\varepsilon}$. Each of the words $\bar{\phi}\bar{\tau}\bar{\phi}_1\bar{\tau}_1 \dots \bar{\tau}_{r+2}$ and $\bar{\phi}\bar{\varepsilon}\bar{\phi}_2\bar{\tau}_1 \dots \bar{\tau}_{r+2}$ is of length n and contains exactly m symbols $\bar{\tau}$. Hence, $\tau_1 \dots \tau_{r+2}$ is a subset of both $L([\bar{\tau}\bar{\varepsilon}^r\bar{\psi}_1])$ and $L([\bar{\tau}\bar{\varepsilon}^r\bar{\psi}_2])$. Consider also the word $\bar{\phi}_1\bar{\tau}_1 \dots \bar{\tau}_{r+2}\bar{\phi}\bar{\tau}$ of length n. This word contains exactly m symbols $\bar{\tau}$, whereas the word $\bar{\phi}_2\bar{\tau}_1 \dots \bar{\tau}_{r+2}\bar{\phi}\bar{\tau}$ contains $m+1$ symbols $\bar{\tau}$. Hence, $\tau_1 \dots \tau_{r+2}\phi\tau$ is a subset of $L([\bar{\tau}\bar{\varepsilon}^r\bar{\psi}_1])$ and disjoint from $L([\bar{\tau}\bar{\varepsilon}^r\bar{\psi}_2])$, which implies that the two states are not equivalent. □

As the behaviour of the binomial coefficient $\binom{n}{m}$ is nonmonotonic with respect to m and n, it is not meaningful to give general asymptotic estimates for the bounds obtained in the above theorem. If, however, we assume that m is a function of n, that is $m = m(n)$, then it is possible to provide such estimates and get some idea about how close the bounds are (asymptotically). We consider two cases. First take $m(n) = k$ for some constant integer $k \geq 2$. In this case, the lower bound is $\Theta(n^{k-1})$ and the upper bound is $\Theta(n^k)$. Now take the case where $m(n) = \lfloor (n-1)/2 \rfloor$. Using the fact that $\binom{2n}{n} = \Theta(2^{2n}/\sqrt{n})$, it can be shown that $\binom{n-1}{m(n)-1} = \Theta(2^n/\sqrt{n})$. On the other hand, using the facts $2^{n-1} = \sum_{i=0}^{n-1}\binom{n-1}{i} = \Theta(\sum_{i=0}^{m(n)}\binom{n-1}{i})$, it follows that the upper bound is $\Theta(2^n)$.

Theorem 3 *Let τ be an error type involving insertions, and deletions or substitutions (possibly both); that is $\tau \cap \iota \neq \emptyset$ and $\tau \cap (\sigma \cup \delta) \neq \emptyset$. For every parameters m and n, with $1 < m < n-1$,*

$$1 + \binom{n+m}{m-1} \leq C_{[\tau s]}(m,n) \leq \frac{4n+m+1}{m+1}\binom{2n+m}{m-1}.$$

Proof. The error type τ can be written as $\tau_1 \cup \iota_1$ such that $\tau_1 \subseteq \sigma \cup \delta$ and $\iota_1 \subseteq \iota$. We shall use $\bar{\varepsilon}, \bar{\iota}_1$ and $\bar{\tau}_1$ as symbols representing the sets ε, ι_1 and τ_1, respectively. Symbols of the form $\bar{\eta}$ are words over the alphabet $\{\bar{\varepsilon}, \bar{\iota}_1, \bar{\tau}_1\}$; that is $\bar{\eta} \in \{\bar{\varepsilon}, \bar{\iota}_1, \bar{\tau}_1\}^*$. Then $\bar{\eta}$ represents the corresponding set η, a subset of $(\varepsilon \cup \iota_1 \cup \tau_1)^*$, in a natural manner. For example, if $\bar{\eta} = \bar{\varepsilon}\bar{\tau}_1\bar{\varepsilon}$ then $\eta = \varepsilon\tau_1\varepsilon$. Moreover, we shall use the following shorthand notation: If $\bar{\eta}$ is a word then $|\bar{\eta}|_{\text{size}} = |\bar{\eta}|_{\bar{\varepsilon}} + |\bar{\eta}|_{\bar{\tau}_1}$ and $|\bar{\eta}|_{\text{err}} = |\bar{\eta}|_{\bar{\tau}_1} + |\bar{\eta}|_{\bar{\iota}_1}$.

First we define an automaton $A = (\varepsilon \cup \tau, Q, [\bar{\varepsilon}], Q \setminus [-1], t)$ accepting exactly the e-strings of the e-system $[\tau s](m,n)$. The states $[-1]$ and $[\bar{\varepsilon}]$ are the sink and start states, respectively. The rest of the states are of the form $[\bar{\eta}]$, where $\bar{\eta}$ is a nonempty word that starts with $\bar{\tau}_1$ or $\bar{\iota}_1$ such that $|\bar{\eta}|_{\text{size}} \leq n$ and $|\bar{\eta}|_{\text{err}} \leq m$. State $[\bar{\eta}]$ means

that the last $|\bar{\eta}|$ input symbols read, say $e_1 \dots e_{|\bar{\eta}|}$, are in η and the next input symbols to be read are independent of any symbols read prior to $e_1 \dots e_{|\bar{\eta}|}$.

Next we define the transition function t. Let $[\bar{\eta}]$ be a state with $|\bar{\eta}|_{\text{size}} = n$. Then there is a unique word $\bar{\eta}_0$ that does not start with $\bar{\varepsilon}$ such that the following conditions are satisfied: (i) If $\bar{\eta}$ starts with $\bar{\tau}_1$ then $\bar{\eta}$ is of the form $\bar{\tau}_1\bar{\varepsilon}^k\bar{\eta}_0$ for some nonnegative integer k. (ii) If $\bar{\eta}$ starts with $\bar{\iota}_1$ then there is a positive integer l such that either $\bar{\eta}$ is of the form $\bar{\iota}_1^l\bar{\varepsilon}^k\bar{\eta}_0$ for some positive integer k, or $\bar{\eta}$ is of the form $\bar{\iota}_1^l\bar{\tau}_1\bar{\varepsilon}^k\bar{\eta}_0$ for some nonnegative integer k. Now we proceed with defining t:

- $t([\bar{\varepsilon}], \varepsilon) = [\bar{\varepsilon}]$, $t([\bar{\varepsilon}], \tau_1) = [\bar{\tau}_1]$ and $t([\bar{\varepsilon}], \iota_1) = [\bar{\iota}_1]$.
- For every state $[\bar{\eta}]$, other than $[\bar{\varepsilon}]$ and $[-1]$,

$$t([\bar{\eta}], \varepsilon) = \begin{cases} [\bar{\eta}\bar{\varepsilon}], & \text{if } |\bar{\eta}|_{\text{size}} < n, \\ [\bar{\eta}_0\bar{\varepsilon}], & \text{if } |\bar{\eta}|_{\text{size}} = n; \end{cases}$$

$$t([\bar{\eta}], \iota_1) = \begin{cases} [\bar{\eta}\bar{\iota}_1], & \text{if } |\bar{\eta}|_{\text{err}} < m, \\ [-1], & \text{if } |\bar{\eta}|_{\text{err}} = m; \end{cases}$$

$$t([\bar{\eta}], \tau_1) = \begin{cases} [\bar{\eta}\bar{\tau}_1], & \text{if } |\bar{\eta}|_{\text{size}} < n \text{ and } |\bar{\eta}|_{\text{err}} < m, \\ [-1], & \text{if } |\bar{\eta}|_{\text{size}} < n \text{ and } |\bar{\eta}|_{\text{err}} = m, \\ [\bar{\eta}_0\bar{\tau}_1], & \text{if } |\bar{\eta}|_{\text{size}} = n. \end{cases}$$

It should be clear that for every state $[\bar{\eta}]$ there is an e-string h in η such that $t([\bar{\varepsilon}], h) = [\bar{\eta}]$; hence, every state in Q is reachable from the start state. Next we prove the following claims, from which the theorem follows:

1. The cardinality of Q is less than, or equal to, the upper bound that appears in the theorem.
2. Let Q' be the set of states $[\bar{\eta}]$ starting with $\bar{\iota}_1$ and satisfying the following conditions: (i) $|\bar{\eta}|_{\text{size}} = n$; (ii) if $\bar{\iota}_1\bar{\varepsilon}$ is a factor of $\bar{\eta}$, then this factor is a prefix of $\bar{\eta}$; that is, $\bar{\eta} = \bar{\eta}_1\bar{\iota}_1\bar{\varepsilon}\bar{\eta}_2$ implies that the word $\bar{\eta}_1$ is empty. Then, the cardinality of $Q' \cup \{[-1]\}$ is equal to the lower bound that appears in the theorem.
3. No two (different) states in Q' are equivalent.

For the first claim, we consider the set P, say, of states of the form $[\bar{\iota}_1\bar{\psi}]$. For any $l = 0, \dots, n$, let P_l be the set of states $[\bar{\iota}_1\bar{\psi}]$ in P with $|\bar{\psi}|_{\text{size}} = l$. Each state in P_l obtains by considering the word $\bar{\iota}_1\bar{\varepsilon}^l$ and by, firstly, choosing j out of the l positions for replacing $\bar{\varepsilon}$'s with $\bar{\tau}_1$'s and, then, choosing (with repetition) i positions out of the $l+1$ for inserting $\bar{\iota}_1$'s, where $j = 0, \dots, \min\{l, m-1\}$ and $i = 0, \dots, m-1-j$. Note that the same argument applies for states of the form $[\bar{\tau}_1\bar{\psi}]$ with $|\bar{\psi}|_{\text{size}} = l$ and $l = 0, \dots, n-1$. Hence,

$$|P_l| = \sum_{j=0}^{\min\{l,m-1\}} \binom{l}{j} \sum_{i=0}^{m-1-j} \binom{l+i}{i} = \sum_{j=0}^{\min\{l,m-1\}} \binom{l}{j}\binom{l+m-j}{m-1-j} \text{ and}$$

$$|Q| = 2 + 2\sum_{l=0}^{n-1} |P_l| + |P_n|.$$

Now for any $l \geq m-1$, one has that

$$|P_l| = \sum_{j=0}^{m-1} \binom{l}{j}\binom{l+m-j}{m-1-j} \leq \sum_{j=0}^{m-1} \binom{l}{j}\binom{l+m}{m-1-j} = \binom{2l+m}{m-1}.$$

On the other hand, for any $l = 0, \ldots, m-2$,

$$|P_l| = \sum_{j=0}^{l} \binom{l}{j}\binom{l+m-j}{m-1-j} \leq \sum_{j=0}^{m-1} \binom{l}{j}\binom{l+m}{m-1-j} = \binom{2l+m}{m-1}.$$

Hence, the sum $\sum_{l=0}^{n-1} |P_l|$ is upper-bounded by

$$\sum_{l=0}^{n-1} \binom{2l+m}{m-1} = \sum_{l=0}^{n-1} \binom{(m-1)+2l+1}{m-1} \leq \frac{m}{m+1}\binom{2n+m-1}{m}.$$

The first claim follows now when we note that

$$\frac{m}{m+1}\binom{2n+m-1}{m} = \frac{m}{m+1}\frac{2n}{m}\frac{2n+1}{2n+m}\binom{2n+m}{m-1} < \frac{2n}{m+1}\binom{2n+m}{m-1}.$$

For the second claim, we observe that each state in Q' is obtained by considering the word $\bar{\iota}_1\bar{\varepsilon}^n$ and by performing the following procedure: For some $j = 0, \ldots, m-1$ and $i = 0, \ldots, m-1-j$, choose j out of the last n positions in the word $\bar{\iota}_1\bar{\varepsilon}^n$ for replacing $\bar{\varepsilon}$'s with $\bar{\tau}_1$'s and, then, insert i symbols $\bar{\iota}_1$ such that each $\bar{\iota}_1$ is inserted at the end of the word, or to the left of one of the j $\bar{\tau}_1$'s. Note that there are $(j+1)$ positions from which one can choose (with repetition) to insert the $\bar{\iota}_1$'s. Hence,

$$|Q'| = \sum_{j=0}^{m-1} \binom{n}{j} \sum_{i=0}^{m-1-j} \binom{j+i}{i} = \sum_{j=0}^{m-1} \binom{n}{j}\binom{m}{m-1-j} = \binom{n+m}{m-1}.$$

For the third claim, we shall use the fact that if $\bar{\iota}_1\bar{\phi}\bar{\zeta}$ is a word with $|\bar{\iota}_1\bar{\phi}\bar{\zeta}|_{\text{size}} = n$ and $|\bar{\iota}_1\bar{\phi}\bar{\zeta}|_{\text{err}} = m$, then for every factor $\bar{\eta}$, say, of $\bar{\iota}_1\bar{\phi}\bar{\zeta}\bar{\tau}_1\bar{\phi}$ with $|\bar{\eta}|_{\text{size}} = n$ we have that $|\bar{\eta}|_{\text{err}} \leq m$. This can be seen if we write ϕ as $\bar{\iota}_1^{l_0}\bar{\theta}_1\bar{\iota}_1^{l_1}\ldots\bar{\theta}_k\bar{\iota}_1^{l_k}$, for some $k \geq 0$ and some $l_j \geq 0$ and $\bar{\theta}_j \in \{\bar{\varepsilon}, \bar{\tau}_1\}$. Now let $[\bar{\iota}_1\bar{\psi}_1]$ and $[\bar{\iota}_1\bar{\psi}_2]$ be two different states in Q' such that $|\bar{\iota}_1\bar{\psi}_2|_{\text{err}} \leq |\bar{\iota}_1\bar{\psi}_1|_{\text{err}}$. We need to show that these states are not equivalent. Let $d = m - |\bar{\iota}_1\bar{\psi}_2|_{\text{err}}$. Then $|\bar{\iota}_1\bar{\psi}_2\bar{\iota}_1^d|_{\text{err}} = m$ and ι_1^d is a subset of $L([\bar{\iota}_1\bar{\psi}_2])$. If $|\bar{\iota}_1\bar{\psi}_2|_{\text{err}} < |\bar{\iota}_1\bar{\psi}_1|_{\text{err}}$ then $d > 0$ and $|\bar{\iota}_1\bar{\psi}_2\bar{\iota}_1^d|_{\text{err}} > m$, which implies that ι_1^d is disjoint from $L([\bar{\iota}_1\bar{\psi}_1])$. Hence, the two states are not equivalent when $|\bar{\iota}_1\bar{\psi}_2|_{\text{err}} < |\bar{\iota}_1\bar{\psi}_1|_{\text{err}}$.

Now suppose that $|\bar{\iota}_1\bar{\psi}_2|_{\text{err}} = |\bar{\iota}_1\bar{\psi}_1|_{\text{err}}$. Then $|\bar{\iota}_1\bar{\psi}_2\bar{\iota}_1^d|_{\text{err}} = m$ and $|\bar{\iota}_1\bar{\psi}_1\bar{\iota}_1^d|_{\text{err}} = m$. Moreover, $|\bar{\iota}_1\bar{\psi}_2\bar{\iota}_1^d|_{\text{size}} = n$ and $|\bar{\iota}_1\bar{\psi}_1\bar{\iota}_1^d|_{\text{size}} = n$. We can write $\bar{\iota}_1\bar{\psi}_1$ as $\bar{\iota}_1\bar{\phi}\bar{\theta}_1\bar{\phi}_1$ and $\bar{\iota}_1\bar{\psi}_2$ as $\bar{\iota}_1\bar{\phi}\bar{\theta}_2\bar{\phi}_2$ such that $\bar{\theta}_1, \bar{\theta}_2$ are different symbols in $\{\bar{\varepsilon}, \bar{\tau}_1, \bar{\iota}_1\}$ and $\bar{\phi}, \bar{\phi}_1, \bar{\phi}_2$ are words with $|\bar{\theta}_1\bar{\phi}_1|_{\text{size}} = |\bar{\theta}_2\bar{\phi}_2|_{\text{size}}$ and $|\bar{\theta}_1\bar{\phi}_1|_{\text{err}} = |\bar{\theta}_2\bar{\phi}_2|_{\text{err}}$. Recall that every factor $\bar{\eta}$ of $\bar{\iota}_1\bar{\phi}\bar{\theta}_i\bar{\phi}_i\bar{\iota}_1^d\bar{\tau}_1\bar{\phi}$, with $|\bar{\eta}|_{\text{size}} = n$, has $|\bar{\eta}|_{\text{err}} \leq m$. Moreover, if $\bar{\theta}_i$ is in $\{\bar{\varepsilon}, \bar{\tau}_1\}$, then $|\bar{\phi}_i\bar{\iota}_1^d\bar{\tau}_1\bar{\phi}|_{\text{size}} = n$. We consider three cases for the pair $(\bar{\theta}_1, \bar{\theta}_2)$ – the other three cases are symmetric.

Firstly, $(\bar{\theta}_1, \bar{\theta}_2) = (\bar{\varepsilon}, \bar{\tau}_1)$. In this case, $|\bar{\phi}_1 \bar{\iota}_1^d \bar{\tau}_1 \bar{\phi} \bar{\iota}_1|_{\text{size}} = |\bar{\phi}_2 \bar{\iota}_1^d \bar{\tau}_1 \bar{\phi} \bar{\iota}_1|_{\text{size}} = n$ and $|\bar{\phi}_1 \bar{\iota}_1^d \bar{\tau}_1 \bar{\phi} \bar{\iota}_1|_{\text{err}} = m+1$ and $|\bar{\phi}_2 \bar{\iota}_1^d \bar{\tau}_1 \bar{\phi} \bar{\iota}_1|_{\text{err}} = m$, which implies that $\iota_1^d \tau_1 \phi \iota_1$ is a subset of $L([\bar{\iota}_1 \bar{\psi}_2])$ and disjoint from $L([\bar{\iota}_1 \bar{\psi}_1])$.

Secondly, $(\bar{\theta}_1, \bar{\theta}_2) = (\bar{\varepsilon}, \bar{\iota}_1)$. In this case, we cannot have $|\bar{\phi}_2|_{\text{size}} = 0$ and, therefore, $\bar{\phi}_2$ can be written as $\bar{\iota}_1^l \bar{\theta} \bar{\zeta}_2$ with $l \geq 0$ and $\bar{\theta} \in \{\bar{\varepsilon}, \bar{\tau}_1\}$. Also, the last factor $\bar{\eta} = \bar{\zeta}_2 \bar{\iota}_1^d \bar{\tau}_1 \bar{\phi} \bar{\iota}_1$ of the word $\bar{\iota}_1 \bar{\psi}_2 \bar{\iota}_1^d \bar{\tau}_1 \bar{\phi} \bar{\iota}_1$ has $|\bar{\eta}|_{\text{size}} = n$ and $|\bar{\eta}|_{\text{err}} \leq m$, which implies that $\iota_1^d \tau_1 \phi \iota_1$ is a subset of $L([\bar{\iota}_1 \bar{\psi}_2])$. On the other hand, $\iota_1^d \tau_1 \phi \iota_1$ is disjoint from $L([\bar{\iota}_1 \bar{\psi}_1])$.

Thirdly, $(\bar{\theta}_1, \bar{\theta}_2) = (\bar{\tau}_1, \bar{\iota}_1)$. Again $\bar{\phi}_2$ can be written as $\bar{\iota}_1^l \bar{\theta} \bar{\zeta}_2$ with $l \geq 0$ and $\bar{\theta} \in \{\bar{\varepsilon}, \bar{\tau}_1\}$. In fact, as $\bar{\iota}_1 \bar{\iota}_1^l \bar{\theta}$ is a non-prefix factor of $\bar{\iota}_1 \bar{\psi}_2$, one has that $\bar{\theta} = \bar{\tau}_1$. This implies that the last factor $\bar{\eta} = \bar{\zeta}_2 \bar{\iota}_1^d \bar{\tau}_1 \bar{\phi} \bar{\iota}_1^2$ of the word $\bar{\iota}_1 \bar{\psi}_2 \bar{\iota}_1^d \bar{\tau}_1 \bar{\phi} \bar{\iota}_1^2$ has $|\bar{\eta}|_{\text{size}} = n$ and $|\bar{\eta}|_{\text{err}} \leq m$, which implies that $\iota_1^d \tau_1 \phi \iota_1^2$ is a subset of $L([\bar{\iota}_1 \bar{\psi}_2])$. On the other hand, the last factor $\bar{\phi}_1 \bar{\iota}_1^d \bar{\tau}_1 \bar{\phi} \bar{\iota}_1^2$ of the word $\bar{\iota}_1 \bar{\psi}_1 \bar{\iota}_1^d \bar{\tau}_1 \bar{\phi} \bar{\iota}_1^2$ has $|\bar{\phi}_1 \bar{\iota}_1^d \bar{\tau}_1 \bar{\phi} \bar{\iota}_1^2|_{\text{size}} = n$ and $|\bar{\phi}_1 \bar{\iota}_1^d \bar{\tau}_1 \bar{\phi} \bar{\iota}_1^2|_{\text{err}} = m+1$, which implies that $\iota_1^d \tau_1 \phi \iota_1^2$ is disjoint from $L([\bar{\iota}_1 \bar{\psi}_1])$.

In all cases the states $[\bar{\iota}_1 \bar{\psi}_1]$ and $[\bar{\iota}_1 \bar{\psi}_2]$ are not equivalent, as required. □

Suppose that m is a function of n; that is, $m = m(n)$. Let L_n and U_n be the lower- and upper-bound, respectively, in the above theorem. If $m(n) = k$ for some constant integer $k \geq 2$, then it follows that $L_n = \Theta(n^{k-1})$ and $U_n = \Theta(n^k)$. On the other hand, if $m(n) = \lfloor n/k \rfloor$, for some constant integer $k \geq 2$, it can be shown that $U_n = O(L_n^2)$.

5. Complexity of E-Systems with Burst Errors

Let τ be an error type. A *burst* of type τ is an e-string g in $\tau \cup \tau(\varepsilon \cup \tau)^* \tau$. Let m an n be positive integers with $m < n - 1$. The e-system $[\tau \mathbf{b}](m, n)$ consists of all the e-strings h for which there is $k \geq 0$ such that $h = h_0 g_1 h_1 \dots g_k h_k$ and $h_i \in \varepsilon^*$ with $|h_i| \geq n - 1$ for $i = 1, \dots, k-1$, and each g_j is a burst of type τ, size up to m, and weight up to m. As before, the ratio m/n is the error density of the e-system $[\tau \mathbf{b}](m, n)$. Let $B_m(\tau)$ be the set of all bursts of type τ, size up to m and weight up to m. Then it follows that

$$[\tau \mathbf{b}](m, n) = (\varepsilon^* B_m(\tau) \varepsilon^{n-1})^* (\varepsilon^* \cup \varepsilon^* B_m(\tau) \cup \dots \cup \varepsilon^* B_m(\tau) \varepsilon^{n-2}).$$

For example, let Σ be the alphabet $\{x, y\}$ and let

$$h = (x/\lambda)(x/x)(x/\lambda)(x/x)^5(\lambda/x)(\lambda/x)(x/x)(x/x)(x/\lambda)(x/x)^8(\lambda/x).$$

Then h contains three bursts of type $(\iota \cup \delta)$. The first is of size 3 and weight 2, the second is of size 3 and weight 3, and the last is of size 0 and weight 1. Hence, h is in the e-system $[(\iota \cup \delta)\mathbf{b}](3, 6)$. On the other hand, h is not in $[(\iota \cup \delta)\mathbf{b}](3, 7)$, as the factor $(x/x)(x/\lambda)(x/x)^5(\lambda/x)$ of size 7 contains two bursts that are too close to each other, or one burst that is too long.

In this section, we are interested in the descriptional complexity of the e-systems

$$[\tau \mathbf{b}] = \{[\tau \mathbf{b}](m, n) : 0 < m < n - 1\}.$$

Theorem 4 *If τ is an error type involving no insertions, that is $\tau \cap \iota = \emptyset$, then*

$$C_{[\tau\mathbf{b}]}(m,n) = n + 1 + m(m-1)/2 = \Theta(m^2 + n).$$

Proof. We define the complete DFA $A = (\varepsilon \cup \tau, Q, [n-1], Q \setminus [-1], t)$ as follows. The set of states Q consists of all $[j]$, with $j = -1, 0, \ldots, n-1$, and all $[l,s]$ with $l = 1, \ldots, m-1$ and $s = 1, \ldots, l$. The start state is $[n-1]$. The state $[-1]$ is a sink state. State $[l,s]$ means that l symbols have been seen since the beginning of the current burst (including the first symbol in τ of the burst) and the size of the burst is s. In general $s < l$, as the burst might grow in case one of the next symbols to read is in τ. For $j \geq 0$, state $[j]$ means that there have been j symbols in ε after the end of the last burst. Obviously, the number of states is $n + 1 + \sum_{l=1}^{m-1} l$, as required. The transition function t is as follows – see also Figure 1.

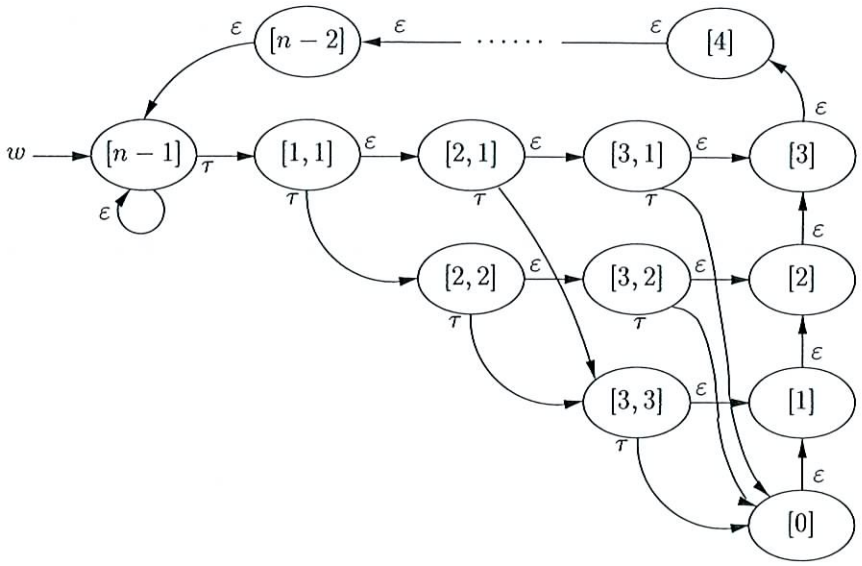

Figure 1: The minimal automaton accepting $[\tau\mathbf{b}](4,n)$ (the sink state is not shown)

- $t([n-1],\varepsilon) = [n-1]$ and $t([n-1],\tau) = [1,1]$. Note that $t([n-1],\tau) = [0]$ when $m = 1$. In this case, no states of the form $[l,s]$ exist.
- $t([j],\varepsilon) = [j+1]$ and $t([j],\tau) = [-1]$, for all $j = 0, \ldots, n-2$.
- $t([l,s],\varepsilon) = [l+1,s]$ and $t([l,s],\tau) = [l+1,l+1]$, for any s and for $l < m-1$.
- $t([l,s],\varepsilon) = [l-s+1]$ and $t([l,s],\tau) = [0]$, for any s and for $l = m-1$.

We need to show now that no two states are equivalent. Obviously, $[-1]$ is not equivalent to any other state. Firstly, note that for all states $[i]$ and $[j]$, other than $[-1]$, with $i < j$, we have that $\varepsilon^{n-1-j}\tau$ is a subset of $L([j])$ and disjoint from $L([i])$. Hence, $[i]$ and $[j]$ are not equivalent. Secondly, for all states $[j]$ with $j = 0, \ldots, n-2$ and for all states of the form $[l,s]$, we have that τ is a subset of $L([l,s])$ and disjoint

from $L([j])$. Also, τ^m is a subset of $L([n-1])$ and disjoint from $L([l,s])$. Hence, no state of the form $[j]$ is equivalent to a state of the form $[l,s]$. Thirdly, consider two states $[l,s]$ and $[l',s']$ with $l < l'$. Then, τ^{m-l} is a subset of $L([l,s])$ and disjoint from $L([l',s'])$. Finally, for any two states $[l,s]$ and $[l,s']$ with $s < s'$, we have that $\varepsilon^{n-1-l+s}\tau$ is a subset of $L([l,s])$ and disjoint from $L([l,s'])$. Hence, no two different states of the form $[l,s]$ are equivalent. □

Theorem 5 *Let τ be an error type involving insertions, that is $\tau \cap \iota \neq \emptyset$. Then $C_{[\tau\mathbf{b}]}(m,n) = \Theta(m^3+n)$. In particular,*

if τ also involves substitutions/deletions, that is $\tau \cap (\sigma \cup \delta) \neq \emptyset$, then

$$C_{[\tau\mathbf{b}]}(m,n) = n+1+(2m+1)+(m+1)(m+2)(m-2)/2;$$

if τ involves only insertions, that is $\tau \subseteq \iota$, then

$$C_{[\tau\mathbf{b}]}(m,n) = n+1+(m+1)+(m+1)(m+2)(m-2)/2.$$

Proof. Suppose $\tau = \iota_1 \cup \tau_1$ such that ι_1 is a nonempty subset of ι and τ_1 is a subset of $\sigma \cup \delta$. We define the complete DFA $A = (\varepsilon \cup \tau, Q, [n-1], Q \setminus [-1], t)$ as follows. The set of states Q consists of all $[j]$, with $j = -1, 0, \ldots, n-1$, and all $[s,l,k]$ with $k = 1, \ldots, m-1$, $l = 0, \ldots, m$ and $s = 0, \ldots, m$ such that $0 \leq s+l \leq m$, and $s = 0$ if $k = 1$ and τ_1 is empty, or $s \in \{0,1\}$ if $k = 1$ and τ_1 is not empty.

The start state is $[n-1]$. The state $[-1]$ is a sink state. State $[s,l,k]$ means that s is the size of the current burst (as known so far) and k is the weight of the burst, and l symbols in ε have been seen after the last symbol (in τ) of the burst. Note that the size of the burst might be smaller than its weight, as insertions are permitted. For example, at state $[2,1,3]$ with $m = 4$ and $\tau = \iota$ it is possible to have seen 6 symbols in $\iota\varepsilon\iota\varepsilon\iota\varepsilon \cup \iota\iota\varepsilon\varepsilon\iota\varepsilon$. In this case, the current burst is of the form $\iota\varepsilon\iota\varepsilon\iota$ or $\iota\iota\varepsilon\varepsilon\iota$. For $j \geq 0$, state $[j]$ means that there have been j symbols in ε after the end of the last burst.

We define the transition function t for the case where τ_1 is not empty. If τ_1 is empty one can simply omit the transitions involving τ_1.

- $t([n-1],\varepsilon) = [n-1]$, $t([n-1],\iota_1) = [0,0,1]$, and $t([n-1],\tau_1) = [1,0,1]$. Note that $t([n-1],\tau) = [0]$ when $m = 1$. In this case, no states of the form $[s,l,k]$ exist.
- $t([j],\varepsilon) = [j+1]$ and $t([j],\tau) = [-1]$, for all $j = 0, \ldots, n-2$.
- $t([s,l,k],\varepsilon) = [s,l+1,k]$ if $s+l < m$, and $t([s,l,k],\varepsilon) = [l+1]$ if $s+l = m$.
- $t([s,l,k],\iota_1) = [s+l,0,k+1]$ if $k < m-1$, and $t([s,l,k],\iota_1) = [0]$ if $k = m-1$.
- $t([s,l,k],\tau_1) = [s+l+1,0,k+1]$ if $s+l < m$ and $k < m-1$, and $t([s,l,k],\tau_1) = [0]$ if $s+l < m$ and $k = m-1$, and $t([s,l,k],\tau_1) = [-1]$ if $s+l = m$.

For the number of states, first note that there are $m+1$ states of the form $[0,l,1]$ when τ involves only insertions, or $(m+1)+m$ states of the form $[s,l,1]$ with $s \in \{0,1\}$ when τ_1 is not empty – recall, $s+l$ cannot exceed m. Now for each $k = 2, \ldots, m-1$, there are $\sum_{s=0}^{m}(m-s+1)$ states of the form $[s,l,k]$, and the claim about the complexity follows. We need to show now that no two states are equivalent. Obviously, $[-1]$ is

not equivalent to any other state. Firstly, note that for all states $[i]$ and $[j]$, other than $[-1]$, with $i < j$, we have that $\varepsilon^{n-1-j}\iota_1$ is a subset of $L([j])$ and disjoint from $L([i])$. Hence, $[i]$ and $[j]$ are not equivalent. Secondly, for all states $[j]$ with $j = 0, \ldots, n-2$ and for all states of the form $[s,l,k]$, we have that ι_1 is a subset of $L([s,l,k])$ and disjoint from $L([j])$. Also, ι_1^m is a subset of $L([n-1])$ and disjoint from $L([s,l,k])$. Hence, no state of the form $[j]$ is equivalent to a state of the form $[s,l,k]$. Thirdly, consider two states $q = [s,l,k]$ and $q' = [s',l',k']$. If $k < k'$ then ι_1^{m-k} is a subset of $L(q)$ and disjoint from $L(q')$. If $k = k'$ and $l < l'$ then $\varepsilon^{n-1-l'}\iota_1^m$ is a subset of $L(q')$ and disjoint from $L(q)$. Finally, if $k = k'$ and $l = l'$ and $s < s'$ then $\varepsilon^{m-(s+l)}\iota_1$ is a subset of $L(q)$ and disjoint from $L(q')$. Hence, no two different states of the form $[s,l,k]$ are equivalent. □

6. Combining E-Systems

One way of combining two error types τ_1 and τ_2 is to use the e-system $[(\tau_1 \cup \tau_2)\mathrm{x}](m,n)$. It might be desirable, however, to assume that errors of the types τ_1 and τ_2 occur independently of each other, and possibly with different densities. Consider the *two* e-systems $[(\mathrm{it})\mathrm{s}](m_1,n_1)$ and $[(\mathrm{ve})\mathrm{s}](m_2,n_2)$ describing scattered transition and transversion errors, respectively, with $m_1/n_1 > m_2/n_2$. We wish to have *one* e-system that allows errors of both types such that the error density of the type (it) is m_1/n_1 and of the type (ve) is m_2/n_2. We shall use the notation $[(\mathrm{it})\mathrm{s}](m_1,n_1) \oplus [(\mathrm{ve})\mathrm{s}](m_2,n_2)$ for such an e-system. For example, an e-string of the form $(\mathrm{ve})(\mathrm{it})\varepsilon\varepsilon(\mathrm{it})\varepsilon\varepsilon(\mathrm{it})\varepsilon\varepsilon(\mathrm{ve})$ would be in $[(\mathrm{it})\mathrm{s}](3,7) \oplus [(\mathrm{ve})\mathrm{s}](1,10)$, but it would be neither in $[(\mathrm{it})\mathrm{s}](3,7)$ nor in $[(\mathrm{ve})\mathrm{s}](1,10)$, as these systems are subsets of $((\mathrm{it}) \cup \varepsilon)^*$ and $((\mathrm{ve}) \cup \varepsilon)^*$, respectively. Next we develop the formalism for the operation $\oplus$ between two e-systems.

Let e be a basic edit operation and let θ be an error type. We define

$$\mathrm{pr}_\theta(e) = \begin{cases} e, & \text{if } e \in \varepsilon \cup \theta, \\ \lambda/\lambda, & \text{if } e \notin \theta,\ e \in \iota, \\ x/x, & \text{if } e \notin \theta, e \in (\sigma \cup \delta),\ e = x/y. \end{cases}$$

The above operation is extended naturally to e-strings: if h is the e-string $e_1 \ldots e_n$ then $\mathrm{pr}_\theta(h) = \mathrm{pr}_\theta(e_1) \ldots \mathrm{pr}_\theta(e_n)$. Thus, $\mathrm{pr}_\theta(h)$ results by replacing each basic edit operation of h that is not in θ with an appropriate basic operation in ε, or with λ/λ. Now for any two e-systems D and D' we define the e-system

$$D \oplus D' = \{h \in (\varepsilon \cup \theta_D \cup \theta_{D'})^* \mid \mathrm{pr}_{\theta_D}(h) \in D \text{ and } \mathrm{pr}_{\theta_{D'}}(h) \in D'\}.$$

Thus, h is in $D \oplus D'$ if it involves errors of types θ_D and $\theta_{D'}$ such that the errors of type θ_D satisfy the constraints of D, namely $\mathrm{pr}_{\theta_D}(h) \in D$, and the errors of type $\theta_{D'}$ satisfy the constraints of D'. In practice it appears that the error types θ_D and $\theta_{D'}$ should be disjoint. Our results, however, remain true without this restriction and, therefore, we allow the possibility that $\theta_D \cap \theta_{D'}$ is nonempty.

Theorem 6 *Let D and D' be two regular e-systems. Then, $C_{D \oplus D'} \leq 1 + C_D C_{D'}$.*

Suppose we wish to model, for instance, bursts of transitions and transversions such that the error density of transitions is twice that of transversions. For this we can consider the class $Z = \{Z(m,n) \mid 0 < 2m < n-1\}$ such that $Z(m,n) = [(\text{it})\mathbf{b}](2m,n) \oplus [(\text{ve})\mathbf{b}](m,n)$. Then, by the above theorem, it would follow that $C_Z(m,n) = O(m^4 + m^2n + n^2)$.

Before we prove the above statement we need to establish an auxiliary lemma. Let τ and θ be two *disjoint* error types and let $A = (\varepsilon \cup \theta, Q, s, F, t)$ be a DFA. We can write $\tau = \iota_1 \cup \tau_1$ such that ι_1 is a subset of ι and τ_1 is a subset of $\sigma \cup \delta$. Define the DFA $A^\tau = (\varepsilon \cup \theta \cup \tau, Q, s, F, t^\tau)$ such that t^τ is constructed as follows:

- t^τ includes t; that is, $t^\tau(p,e) = q$ for all states p, q and $e \in \varepsilon \cup \theta$ with $t(p,e) = q$.
- $t^\tau(q, \lambda/x) = q$, for all states q and $\lambda/x \in \iota_1$.
- For all states p, q and $x/x \in \varepsilon$, if $t(p, x/x) = q$, then $t^\tau(p, x/y) = q$ for every x/y in τ_1.

Obviously, A and A^τ have the same state sets.

Lemma 7 *The automaton A^τ accepts the language $\{h \in (\varepsilon \cup \theta \cup \tau)^* \mid pr_\theta(h) \in L(A)\}$.*

Proof. By the construction of t^τ we observe that (i) if $t^\tau(q, \lambda/x) = q$ and λ/x is in ι_1 then $\mathrm{pr}_\theta(\lambda/x) = \lambda/\lambda$, and (ii) if $t^\tau(p, x/y) = q$ and x/y is in τ_1 then $t(p, x/x) = q$ and $\mathrm{pr}_\theta(x/y) = x/x$. First suppose that h is in $(\varepsilon \cup \theta \cup \tau)^*$ such that $\mathrm{pr}_\theta(h) \in L(A)$. Then $h = e_1 \ldots e_n$, with each e_i being in $\varepsilon \cup \theta \cup \tau$, and $\mathrm{pr}_\theta(h) = \mathrm{pr}_\theta(e_1) \ldots \mathrm{pr}_\theta(e_n)$. As some of the elements $\mathrm{pr}_\theta(e_i)$ might be empty, there is an accepting extended computation $p_0 \mathrm{pr}_\theta(e_1) p_1 \ldots \mathrm{pr}_\theta(e_n) p_n$ of the automaton A. If e_i is in τ_1 then e_i is of the form x/y with $x \in \Sigma$ and $y \neq x$, which implies that $\mathrm{pr}_\theta(e_i)$ must be x/x. As $t(p_{i-1}, x/x) = p_i$ one has that $t^\tau(p_{i-1}, x/y') = p_i$ for all x/y' in τ_1. Hence, $t^\tau(p_{i-1}, e_i) = p_i$. In the cases $e_i \in \iota_1$ and $e_i \in \varepsilon \cup \theta$ it follows again that $t^\tau(p_{i-1}, e_i) = p_i$ and, therefore, $p_0 e_1 p_1 \ldots e_n p_n$ is an accepting computation of the automaton A^τ. Hence, h is in $L(A^\tau)$.

Now suppose that h is in $L(A^\tau)$. Then $h = e_1 \ldots e_n$, with each e_i being in $\varepsilon \cup \theta \cup \tau$, and there is an accepting computation $p_0 e_1 p_1 \ldots e_n p_n$ of A^τ. It is sufficient to show that $p_0 \mathrm{pr}_\theta(e_1) p_1 \ldots \mathrm{pr}_\theta(e_n) p_n$ is an extended computation of the automaton A; that is, $t(p_{i-1}, \mathrm{pr}_\theta(e_i)) = p_i$ for all i. This can be shown using the fact that $t^\tau(p_{i-1}, e_i) = p_i$ and by considering the three cases $e_i \in \varepsilon \cup \theta$, $e_i \in \iota_1$, and $e_i \in \tau_1$. □

Proof of Theorem 6. Let A and A' be minimal complete DFAs accepting D and D', respectively. Let $\tau' = \theta_{D'} \setminus \theta_D$ and $\tau = \theta_D \setminus \theta_{D'}$. Then

$$\varepsilon \cup \theta_D \cup \theta_{D'} = \varepsilon \cup \theta_D \cup \tau' = \varepsilon \cup \theta_{D'} \cup \tau.$$

By the above lemma it follows that $D \oplus D'$ is the language accepted by the automaton $A'^\tau \cap A^{\tau'}$, which is of size $|A||A'| = C_D C_{D'}$. In general, it might be necessary to add a sink state to the automaton $A'^\tau \cap A^{\tau'}$, in order to make it complete. Hence, $C_{D \oplus D'} \leq 1 + C_D C_{D'}$. □

We note that, in many cases, Theorem 6 can be improved slightly by considering A and A' to be trim automata; that is, automata whose states can be reached from

the start state and can reach a final state. It follows then that $C_{D\oplus D'} \leq 1 + |A||A'|$. It can be shown for example that, for $D = [\sigma\mathbf{b}](1,3)$ and $D' = [\delta\mathbf{b}](1,4)$, we have that $C_{D\oplus D'} = 11$, and $|A| = 3$ and $|A'| = 4$, where A and A' are trim automata accepting D and D', respectively.

7. Concluding Remarks

We have presented a method for describing error situations as formal languages, which allows one to reason about errors using tools from automata. In particular, we were able to obtain results about the complexity of describing error situations, motivated by the need for evaluating the efficiency of algorithms that compute error-correcting capabilities of languages. In the case of scattered errors, the complexity turns out to be very high. One can still define automata with reasonable size, however, for e-systems with high error density, by choosing small values for the parameters m and n. For example, the descriptional complexity of the e-systems $[\tau\mathbf{s}](3,6)$, $[\tau\mathbf{s}](3,10)$ and $[\tau\mathbf{s}](4,10)$ is no more than 735, 2783, and 18215, respectively, where τ is any error type involving insertions and substitutions/deletions.

We believe that the approach of e-systems can be used to model errors in various domains that involve processing or transmission of information in the presence of errors.

Acknowledgements

The authors would like to thank Jeremy Newton-Smith for creating the electronic form of Figure 1.

References

[1] L. R. BAHL, F. JELINEK, Decoding for channels with insertions, deletions, and substitutions with applications to speech recognition. *IEEE Transactions on Information Theory* **21** (1975), 404–411.

[2] C. CHARALAMBIDES, *Combinatorics.* National University of Athens, Athens, 1984 (in Greek).

[3] D. GUSFIELD, *Algorithms on Strings, Trees, and Sequences.* Cambridge University Press, 1997.

[4] H. JÜRGENSEN, S. KONSTANTINIDIS, Error correction for channels with substitutions, insertions, and deletions. In: J.-Y. CHOUINARD, P. FORTIER, T. A. GULLIVER (eds.), *Information Theory and Applications 2, Fourth Canadian Workshop on Information Theory.* LNCS **1133**, Springer-Verlag, 1996, 149–163.

[5] L. KARI, R. KITTO, G. THIERRIN, Codes, involutions and DNA encodings. In: W. BRAUER, H. EHRIG, J. KARHUMAKI, A. SALOMAA (eds.), *Formal and Natural Computing*, LNCS **2300**, Springer Verlag, 2002, 376–393.

[6] L. KARI, S. KONSTANTINIDIS, E. LOSSEVA, G. WOZNIAK, Sticky-free and overhang-free DNA languages. Submitted for publication.

[7] L. KARI, S. KONSTANTINIDIS, S. PERRON, G. WOZNIAK, Maximal error-correcting capabilities of languages (tentative). Manuscript in preparation.

[8] S. KONSTANTINIDIS, An algebra of discrete channels that involve combinations of three basic error types. *Information and Computation* **167** (2001), 120–131.

[9] S. KONSTANTINIDIS, Transducers and the properties of error-detection, error-correction and finite-delay decodability. *Journal of Universal Computer Science* **8** (2002), 278–291.

[10] K. KUKICH, Techniques for automatically correcting words in text. *ACM Computing Surveys* **24** (1992), 377–439.

[11] A. LEON-GARCIA, I. WIDJAJA, *Communication Networks.* McGraw-Hill Higher Education, 2000.

[12] B. LEWIN, *Genes VII.* Oxford University Press, 2000.

[13] F. J. MACWILLIAMS, N. J. A. SLOANE, *The theory of Error-correcting Codes.* North Holland, Amsterdam, 1977.

[14] A. MARATHE, A. CONDON, R. CORN, On combinatorial DNA word design. In: E. WINFREE, D. GIFFORD (eds.), *DNA Based Computers V*, DIMACS Series, AMS Press, 2000, 75–89.

[15] G. ROZENBERG, A. SALOMAA (eds.), *Handbook of Formal Languages, Vol. I.* Springer-Verlag, Berlin, 1997.

(Received: January 23, 2003; revised: December 12, 2003)