FOR THE FAMILY OF NACOSS

CSC 304

Prefixes, Suffixes and Substrings in a Multilanguage: A Perspective

Dasharath Singh

(Former Professor: Indian Institute of Technology Bombay)
Corresponding author
Mathematics Department, Ahmadu Bello University Zaria, Nigeria

Ahmed Ibrahim Isah

Mathematics Department, Ahmadu Bello University Zaria, Nigeria

Abstract

In this paper, *prefix*, *suffix* and *substring* relations are described in a Multilanguage. It is shown that prefix and suffix relations give rise to a relational structure of a linear multiset order, and substring relations determine only a partial multiset order. The distinctive features of these relations in multilanguage vs. standard formal language are identified by using their graphical and Hasse diagram representations. Finally, certain monoids of these relations and their homomorphisms are described.

Mathematics Subject Classification: 68Q45, 94A45, 03E02

Keywords: String, Prefix, Suffix, Substring, Monoid, Multilanguage

1 Introduction

Gottlob Frege's formulation of the first-order logic in 1879 is known to be the first description of the structure of a formal language. It was immensely augmented by Hilbert, Alan Turing and Emil Post. In the mid-1950s, Noam Chomsky's endeavor to formulate a general theory of the syntax of *natural languages* made a great impact on the development of formal language theory. In the 1960s and 1970s, particularly due to the advent of electronic computers, much of the foundations for the theory of formal languages were erected in order to describe the process linked with the use of computers and communication devices. All forms of information (whether numbers, names, pictures, or sound waves) could be represented as *strings* and a collection of strings could be structured to represent a language became central to diverse applications of formal languages in computer science and linguistics, and more recently, in soft sciences such as biology and economics. In fact, the study of *stringology* has come to occupy the core of the formal language theory.

On the other hand, the fact that many *query* languages and *database* languages do require multiset-based semantics, the study of *stringology* needs to be extended to *multilanguages*.

This paper attempts to describe prefix, suffix and substring relations and their properties in a multilanguage. In particular, monoids of these relations and their homomorphisms are described.

2 Some Basics of Standard Formal Language Theory

Most of the definitions given here are standard and mimicked from various sources, particularly from [1] and those listed in the references.

Basically, a language L is defined as a set of strings over a fixed alphabet Σ. An alphabet is a nonempty set of distinguishable symbols, called *terminal* symbols, sometimes denoted Σ_T. A language deployed to generate strings of another language over an alphabet Σ is called *metalanguage* often denoted Σ_N, and it is supposed to consist of a set of *syntactic classes* or *variables* called *nonterminal* symbols. An element of an alphabet is called a *letter* or a *character*. An alphabet need not be finite or even countably infinite. However, for our purposes here, we shall assume it to be finite. A string is a finite ordered set of symbols from the alphabet. A string is also called a (finite) *sequence*, a *word*, or a *sentence* depending on its nature. A string consisting of k symbols $(k > 0)$ is said to be a string of length k. In the event, an empty string (i.e., a string of length 0), denoted □, □, □, or 1, is admitted, the system can have strings of length $k \geq 0$.

Strings can also be defined as functions [1]. For any integer $n \geq 1$, let $[n] = \{1, 2, \dots, n\}$, and for $n = 0$, let $[0] = \emptyset$.

A string u of length n over a fixed Σ is a function

$$u\colon [n] \longrightarrow \Sigma.$$

When $n = 0$, the special string $u\colon [0] \longrightarrow \Sigma$, of length 0 (having no symbol), is called the *empty* or *null string*.

Given a string $u\colon [n] \longrightarrow \Sigma$ of length $n \geq 1$, let $u(i)$ denote the i^{th} letter in the string u.

A string u of length n can also be represented as

$$u = u_1\, u_2 \ldots u_n,\ u_i \in \Sigma,\ i = \overline{1, n}.$$

For example, let $\Sigma = \{a, b\}$ and $u\colon [5] \longrightarrow \Sigma$ be defined such that $u(1) = a,\ \ u(2) = b,$ $u(3) = b, u(4) = a$ and $u(5) = a$, then

$$u = abbaa.$$

Let $\Sigma^{□}$ denote the set of all strings and $\Sigma^{+} = \Sigma^{□} - \{□\}$, denote the set of all nonempty strings over Σ. A *formal language* (or simply a *language*) L is a subset of $\Sigma^{□}$ i.e., $L \subseteq \Sigma^{□}$.

A language can also be considered as a subset of the free monoid on an alphabet; however, as the language consisting of a free monoid is too large, it is not found suitable for practical applications.

Following [1], we describe below two basic operations which are central to both standard formal languages and multilanguages.

Concatenation

Let $u\colon [m] \longrightarrow \Sigma$ and $v\colon [n] \longrightarrow \Sigma$ be two strings over Σ, then concatenation of u and v, denoted $u.\, v$ or uv, is the string $uv\colon [m + n] \longrightarrow \Sigma$, defined as

$$u\,v\,(i) = \begin{cases} u(i) & \text{if } 1 \leq i \leq m, \\ v(i - m) & \text{if } m + 1 \leq i \leq m + n. \end{cases}$$

It is immediate to see that the following holds:

$u\varepsilon = \varepsilon u = u$ and $u(vw) = (uv)w$.

In other words, concatenation is a binary operation on $\Sigma^{□}$ which is associative and has ε as an identity. Thus, $(\Sigma^{□},\ ., \varepsilon)$ or $(\Sigma^{□},\ .)$ is a monoid, and $(\Sigma^{+},\ \cdot)$ is a (free) semigroup generated by Σ.

Note that concatenation, in general, is not commutative i.e., $uv \neq vu$. For example, let $u = aab$, $v = bb$ over $\Sigma = \{a, b\}$, then $uv \neq vu$.

Powers of strings

Given a string $u \in \Sigma^{\square}$ and $n \geq 0$, u^n is defined as follows:

$$u^n = \begin{cases} \varepsilon & \text{if } n = 0, \\ u^{n-1}u & \text{if } n \geq 1. \end{cases}$$

Clearly, $u^1 = u$, $\varepsilon^n = \varepsilon$, and $u^n u = u u^n$, for all $n \geq 0$.
For example, let $\Sigma = \{a, b, c\}$, $u = abba$, $v = aba$, $w = abca$, $x = a$, $y = \varepsilon$; then $(uv)w = u(vw) = abbaabaabca$, $u^2 = abbaabba$, $x^6 = aaaaaa$, $xy = x = a$, etc.

It is known that, in general, uncountably infinite amount of languages exist over any alphabet. However, in the case of a formal language theory, as it is defined by a *grammar* or by an *automaton*, and as such operating in a natural language, at most countably infinite amount of languages can exist.
In view of enormous applications of formal languages, it is not surprising that a vast amount of literature is available on the topic (see 1, 2, 3, 4, 5, for details).

3 Multilanguages

Knuth, in the summer of 1964, while writing the book entitled *The Art of Computer Programming,* drafted a chapter on *The theory of languages* where he emphasized the need to keep a track of the *multiplicity* of strings in the language so that a string would appear several times if there were several ways to parse it. This was recognized quite natural from a programmer's point of view because *transformations* on *context-free grammar* (its terminal strings form *multisets* if it is non-circular) were found most useful in practice, especially when they yielded isomorphisms between *parse trees*. In fact, Knuth [6] is the first place where an elaborate description of m*ultilanguages* appeared. Knuth defines a multilanguage simply as a multiset of strings.

3.1 Basics of a Multilanguage

In the following, we briefly describe multisets, multiset relations, and lexicographic ordering on multisets. Further, characterization of prefix, suffix and substring relations in a multilanguage is described which would help studying *stringology* in a multisetbased environment.

Multisets

A *Multiset* (mset, for short) is an unordered collection of objects in which, unlike an ordinary set, duplicates or multiples of objects are admitted. For example, a string stripped of its ordering is a multiset.

An mset containing one occurrence of a, two occurrences of b, and three occurrences of c, is variably represented as $[[a, b, b, c, c, c]]$ or $[a, b, b, c, c, c]$ or $[a, b, c]_{1,2,3}$ or $[a, 2b, 3c]$ or $[a.\ 1, b.\ 2, c.\ 3]$ or $[1/a, 2/b, 3/c]$ or $[a^1, b^2, c^3]$ or $[a^1b^2c^3]$. For convenience, the curly brackets are also used in place of the square brackets. In fact the last form of representation as string, even without using any bracket, turn out to be the most compact one, especially in computational parlance. In general, if x_1 appears k_1 times, x_2 appears k_2 times, … , x_n appears k_n times in an mset A, then A is expressed as $A = \{k_1/x_1, k_2/x_2, \dots, k_n/x_n\}$.

Formally, a multiset over Σ is a mapping $A : \Sigma \longrightarrow$ □, where □ is the set of natural numbers, including zero. Let $A(a)$ or $C_A(a)$ denote the number or count of *copies* of the symbol a in the multiset A. The *empty* multiset, denoted □, is defined as □ $(a) = 0$, for all $a \in \Sigma$. Follows that A is a set, whenever $A(a) \leq 1$.

Let A and B be msets over Σ. A is said to be a *submultiset* (*msubset* or *submset*, for short) of B, written $A \subseteq B$ or $B \supseteq A$, if $A(a) \leq B(a)$, $\forall a \in \Sigma$. Thus, $A = B$ iff $A \subseteq B$ and $B \subseteq A$.

Note that an mset need not be finite, in general. However, for our purposes here, we shall consider it finite.

Multiset relations

Let X^n be a cardinality bounded mset space defined on a set X and $A \in X^n$. Any subset R of AxA is said to be an *mset relation* on A, denoted R: $A \longrightarrow A$, where every member of R has a count $C_1(x, y)$ and $C_2(x, y)$. Note that $m/x\ R -$ related to n/y is symbolized as $(m/x, n/y) \in R$ or $m/x\ R\ n/y$. In other words, $R = \{(m/x, n/y)/\ mn$: $(m/x, n/y) \in^{mn} R\}$, where $(m/x, n/y) \in^{mn} R$ symbolizes that the ordered pair (x, y) occurs mn times in R. The domain of the relation R on A ($Dom\ R$, for short) $=$ $\{x \in^s A$: $\exists y \in^t A$ such that $s/x\ R\ t/y\}$ where $C_{dom\ R}(x) = \sup\{C_1(x, y)$: $x \in^s A\}$, and the range of R on A ($Ran\ R$, for short) $= \{y \in^t A$: $\exists x \in^s A$ such that $s/x\ R\ t/y\}$ where $C_{ran\ R}(x) = \sup\{\ C_2(x, y)$: $y \in^t A\}$. Note that these Sups always exist for a cardinality bounded msets.

For example, let $A = [5/x, 7/y, 9/z]$ be an mset. Then, $R = \{(2/x, 3/y)/6, (4/x, 3/$

$z)/12$, $(4/y, 5/z)/20$, $(6/y, 5/x)/30$, $(4/z, 4/z)/16$, $(5/z, 3/x)/15$, $(3/x, 6/y)/18\}$ is an mset relation on A with $Dom\ R = [4/x, 6/y, 5/z]$, and $ran\ R = [5/x, 6/y, 5/z]$.
An mset relation R on an mset A is called *reflexive* iff $\forall\, m/x$ in A, $m/x\ R\ m/x$; *irreflexive* iff $m/x\ R\ m/x$ never holds; *symmetric* iff $m/x\ R\ n/y$ imply $n/y\ R\ m/x$, *antisymmetric* iff $\forall\ m/x, n/y$ in A, $m/x\ R\ n/y$ and $n/y\ R\ m/x$ imply $m/x = n/y$; *transitive* iff $\forall\, m/x, n/y,\ k/z$ in A, $m/x\ R\ n/y$ and $n/y\ R\ k/z$ imply $m/x\ R\ k/z$.

An mset relation R on an mset A is said to be an *mset order relation* (*mset order*, for short) or a *partial mset order* relation, if it is reflexive, antisymmetric and transitive. The pair (A, R) is called an *ordered mset* or a *partially ordered mset (pomset) structure*. A partial mset order is called a *linear mset order* (or a *complete mset order*) on A, if for every two elements $m/x \neq n/y$ of A, either $m/x\ R\ n/y$ or $n/y\ R\ m/x$ (see [7], [8], for further details).

Over an alphabet Σ, let $\Sigma^{\circledast}$ denote the multiset of all strings, including the empty string which is unique, and $\Sigma^{\oplus}$, the multiset of all non-empty strings. A *multilanguage*, denoted M_L, is an msubset of $\Sigma^{\circledast}$.
Let for each $a\ \in \Sigma,\ u \in \Sigma^{\circledast}$, $|u|_a$ denote the number of occurrences of the symbol a in the string u.

Observe that if $\Sigma = \emptyset$ then $\emptyset^{\circledast} = \{\varepsilon\}$, and if $\Sigma\ \neq \emptyset$ then $\Sigma^{\circledast}$ is countably infinite.
Interestingly, even for a tiny alphabet that consists of a single letter e.g., let $\Sigma = \{a\}$, $\Sigma^{\circledast}$ contains infinitely many strings viz., $a^0\ =$□, $a^1\ =\ a,\ a^2\ =\ aa$, etc., and, in general, for every $i > 0$, the string a^i is in $\Sigma^{\circledast}$. Therefore, given any alphabet Σ, there are infinitely many multilanguages over Σ.

Orderings on strings in a Multilanguage
Let $\Sigma = \{a_1, \ldots, a_k\}$ such that $a_1 < a_2 < \cdots a_k,\ \ u, v \in \Sigma^{\circledast}$, then the lexicographic ordering, denoted $\preccurlyeq$, is defined as:

$$u \preccurlyeq \begin{cases} & \text{if } v = u, \text{ or} \\ v & \text{if } v = uy \text{ for some } y \in \Sigma^{\circledast}, \text{ or} \\ & \text{if } u = xa_iy,\ v = xa_jz, \text{ and } a_i < a_j \text{ for some } x, y, z \in \Sigma^{\circledast}. \end{cases}$$

In fact, lexicographic ordering of $\Sigma^{\circledast}$ is a monotonic and linear mset order.
For example, the lexicographic ordering of the strings $abbb, a, b, a, bb, aa, aaa, aa$ over $\Sigma = \{a, b\}$ is $a, a, b, aa, aa, bb, aaa, abbb$ and also, lexicographic ordering of $\Sigma^{\circledast}$ itself is $[\varepsilon, a, b, aa, ab, ba, bb, aaa, aab, \ldots]$, etc.

Prefix, Suffix and Substring in a Multilanguage

Let Σ be an alphabet,$\Sigma^{\circledast}$ be the mset of all strings overΣ, and$u, v \in \Sigma^{\circledast}$.

u is a *substring* of v iff there exist $x, y \in \Sigma^{\circledast}$ such that $v = xuy$. In other words, a substring (also, called *segment*, *subword* or *factor*) of a string is any sequence of consecutive symbols that appear in the string.

The term *substrings of a string* represents the set of all substrings of a string.

u is a *prefix* of v iffthere is some $y \in \Sigma^{\circledast}$ such that $v = uy$. That is, a prefix of a string u is a substring of u that occurs at the beginning of$u$.

The term *prefixes of a string* represents the set of all prefixes of a string.

u is a *suffix* of v iffthere is some $x \in \Sigma^{\circledast}$ such that $v = xu$. That is, a suffix of a string u is a substring that occurs at the end of u.

The term *suffixes of a string* represents the set of all suffixes of a string.

Observe that a prefix is any sequence of leading symbols of the string, and a suffix is any sequence of trailing symbols of the string. Moreover, a substring of a string is a prefix of a suffix of the string, and also, a suffix of a prefix.

u is a proper prefix (suffix or substring) of v iff u is a prefix (respectively, suffix or substring) of v and$u \neq v$.

Note that prefixes and suffixes both are substrings, but the converse need not hold.

For example, let $w = abaa$ be a string over $\Sigma = \{a, b\}$. Its prefixes, suffixes and substrings are as follows:

Prefixes: □$, a, ab, aba, abaa$.

Suffixes: □$, a, aa, baa, abaa$.

Substrings: □,$a, b, ab, a, ba, aba, a, aa, baa, abaa$.

The *border* of a string is both its prefix and suffix, e.g., bab is a border of$babab$.

Remark 3.1.1

It needs to be emphasized that prefixes, suffixes and substrings are *multisets*. The prefixes and suffixes are defined in the same way both in standard formal language and multilanguage. For example, prefixes and suffixes of the string$w$ considered above are the same in both. However, for substrings, this is not the case. The substrings of w in a multilanguage are as given above, where as in a standard formal language, substrings of w are □$, a, b, ab, ba, aba, aa, baa, abaa$.

Prefix, Suffix and Substring Closures in a Multilanguage

Let M_L be a multilanguage. In line with the description of these notions for a standard formal language, we define them for M_L as follows: *Prefix closure* ofM_L:

$pref_{M_L}(v) = \{u \mid v = uy;\ v \in M_L, u, y \in \Sigma^{\circledast}\}$, and

$$pref(M_L) = \biguplus_{v \in M_L} pref_{M_L}(v) = \{u \mid v = uy;\ v \in M_L, u, y \in \Sigma^{\circledast}\}.$$

Suffix closure of M_L:

$suff_{M_L}(v) = \{u |\ v = xu;\ v \in M_L, x, u \in \Sigma^{\circledast}\}$, and

$$suff(M_L) = \uplus_{v \in M_L} suff_{M_L}(v) = \{u |\ v = xu;\ v \in M_L, x, u \in \Sigma^{\circledast}\}.$$

Substring closure of M_L:

$subs_{M_L}(v) = \{u |\ v = xuy;\ v \in M_L, x, u, y \in \Sigma^{\circledast}\}$, and

$$subs(M_L) = \uplus_{v \in M_L} subs_{M_L}(v) = \{u |\ v = xuy;\ v \in M_L, x, u, y \in \Sigma^{\circledast}\}.$$

Example

Let $M_L = \{ab, ab, aba\}$ over $\Sigma = \{a, b\}$. Then

$pref(M_L) = \{\varepsilon, a, ab, a, ab, a, ab, aba\}$,

$suff(M_L) = \{\varepsilon, b, ab, b, ab, a, ba, aba\}$, and

$subs(M_L) = \{\varepsilon, a, b, ab, a, b, ab, a, b, ab, a, ba, aba\}$.

As in a standard formal language L, the following results hold in M_L as well:

A multilanguage is said to be *prefix, suffix or substring closed* if $pref(M_L) = M_L$, $suff(M_L) = M_L$ or $subs(M_L) = M_L$, respectively.

For example, $M_L = [\varepsilon, a, a, b]$ is both prefix, suffix and substring closed.

The prefix, suffix and substring closure operators in a multilanguage M_L are idempotent:

$pref(pref(M_L)) = pref(M_L)$, $suff(suff(M_L)) = suff(M_L)$, $subs(subs(M_L)) = subs(M_L)$.

Remark 3.1.2

It may, however, be noted that the prefix, suffix and substring closures of a multilanguage and that of a standard formal language differ. For example, in the case of a multilanguage $M_L = \{ba, aba\}$, $suff(M_L) = \{\varepsilon, a, ba, a, ba, aba\}$, and in a standard formal language $L = \{ba, aba\}$, $suff(L) = \{\varepsilon, a, ba, aba\}$.

3.2Prefix, Suffix and Substring Relations and Their Characteristics in a Multilanguage

Let X consist of prefixes (suffixes or substrings) of a string $v \in \Sigma^{\circledast}$ and R, be the respective binary relation on X. Hence forth, R shall stand for R *(prefix, suffix or substring)*, unless otherwise stated.

i. R is *reflexive*: $\forall\ m/x \in X, m/x\ R\ m/x$. Since every prefix (suffix or substring) u is a prefix (respectively, suffix or substring) of itself.

ii. R (prefix or suffix) is *complete*: $\forall\ m/x \neq n/y \in X$, either $m/x\ R\ n/y$ or $n/\ y\ R\ m/x$. Since for every pair of prefixes (or suffixes) of a string v, one is a prefix (respectively, suffix) of the other. However, for substrings, this is not licit. For example, let $v = abbc$ and $m/x = 1/a, n/y = 1/c$ then both $m/x, n/y \in X$, but neither $m/x\ R\ n/y$ nor $n/y\ R\ m/x$.

iii. R is *antisymmetric*: $\forall\ m/x, n/y \in X$, if $m/x\ R\ n/y$ and $n/y\ R\ m/x$, then $m/x = n/y$. Since whenever u is a prefix (suffix or substring) of v and v is a prefix (respectively, suffix or substring) of u, then u must be the same as v as duplicates are indistinguishable.

iv. R is *transitive*: $\forall\ m/x,\ n/y,\ k/z \in X$, if $m/x\ R\ n/y$ and $n/y\ R\ k/z$, then $m/x\ R\ k/z$. Since whenever u is a prefix (suffix or substring) of v and v is a prefix (respectively, suffix or substring) of w, clearly u is a prefix (respectively, suffix or substring) of w.

Clearly, R (prefix or suffix) is a linear mset order, where as R (substring) is only a partial mset order.

Summarily, R on X gives rise to a *multiset relational structure* (X, R), hence forth notated $\mathbb{X} = (X, R)$, on X. In general, X is an mset and R is an mset relation. In particular, X may be a list of *attributes* defined on a domain and R, the relation name (see [9], for related descriptions). In the following, we describe some typical features of such an mset structure, which would be exactly the features possessed by its relations.

Maximal, Minimal and Isolated Elements of $\mathbb{X} = (\boldsymbol{X}, \boldsymbol{R})$

Let $\mathbb{X} = (X, R)$, where X consists of all nonempty prefixes (suffixes or substrings) of a string $v \in \Sigma^{\oplus}$ and R, its respective relation on X. Then, m/x is a *maximal* element of $\mathbb{X}$ iff $\forall\ n/y \in X, n/y\ R\ m/x$; for R (prefix or suffix), m/x is a *minimal* element of $\mathbb{X}$ iff $\forall\ n/y \in X, m/x\ R\ n/y$; for R (substring), m/x is a *minimal* element of $\mathbb{X}$ iff $\forall\ n/y \in X$, $m/x\ R\ n/y$ or else they are unrelated; and

m/x is an *isolated* element of $\mathbb{X}$ iff m/x is both a maximal and minimal element of $\mathbb{X}$.

For example, Let $\Sigma = \{a, b\}$, $v = abba \in \Sigma^{\oplus}$, then the non-empty prefixes and substrings of v are $a, ab, abb, abba$ and $a, b, ab, b, bb, abb, a, ba, bba, abba$, respectively.

Now for R (prefix), $1/a$ and $1/abba$ are the minimal and maximal elements, respectively, and for R (substring), $2/a$ and $2/b$ are the minimal elements and $1/abba$ is the maximal element.

Also, let $u = a \in \Sigma^{\oplus}$, then $1/a$ is the isolated element for R.

Proposition 3.2.1

$\mathbb{X} = (X, R)$ always has a unique maximal element and, only under R (prefix or suffix), it has also a unique minimal element.

Proof

Let $p/z \neq t/s \in X$, and let $p/z, t/s$ both be maximal elements of $\mathbb{X}$. Since p/z is a maximal element, $t/s\ R\ p/z$. Similarly, since t/s is a maximal element, $p/z\ R\ t/s$.

Now, by antisymmetry of these relations, $p/z = t/s$. This contradicts the assumption that $p/z \neq t/s$. Hence $\mathbb{X}$ has a unique maximal element. Similarly, the other parts can be proved.

Chains of $\mathbb{X} = (\boldsymbol{X}, \boldsymbol{R})$

A submset structure $\mathbb{C} = (C \subseteq X, R)$ of the prefix (suffix or substring) mset relational structure $\mathbb{X}$ is a chain in $\mathbb{X}$, if distinct points of C are pair wise comparable in X i.e., $\forall$ $m/x \neq n/y \in X$, either $m/x\, R\, n/y$ or $n/y\, R\, m/x$ in X. Also, $\mathbb{X}$ itself is said to be a chain if the distinct points on X are pair wise comparable. It is easy to see that every (C, R) with R (prefix or suffix) is a chain, and the mset of all chains in a prefix (suffix or substring) relational structure is partially ordered by mset *inclusion* relation. The maximal elements in this pomset relational structure are the maximal chains.
For example, let $v = aba \in \Sigma^{\oplus}$, and $X = [2/a, 1/b, 1/ab, 1/ba, 1/aba]$ be the mset of all nonempty substrings of v. Then, submsets $C_1 = [2/a, 1/ab]$, $C_2 = [2/a, 1/ba]$, $C_3 = [2/a, 1/ab, 1/aba]$, $C_4 = [1/b, 1/ab]$, $C_5 = [1/b, 1/ab, 1/aba]$ are some chains in $\mathbb{X}$**,** and C_3, C_5 are the maximal chains among these.

Graphical Representation of $\boldsymbol{R}$ on $\boldsymbol{X}$

Let X consist of all prefixes (suffixes or substrings) of a string $v \in \Sigma^{\oplus}$ and R, its respective relation on X. This relation can be represented by a *directed graph* $G = (V, E)$, where V (the vertex mset) is defined by $V \overset{\text{def}}{=} X$, and E (the directed edges mset) by
$(m/x, n/\, y) \in E$ iff $(m/x \neq n/\, y) \in R$, for $n \geq 2$.

For example, Let $\Sigma = \{a, b\}$, $v = abaa \in \Sigma^{\oplus}$, and let the non-empty prefixes and substrings of v be given as:
Prefixes: $s = a$, $u = ab, w = aba, x = abaa$, and
Substrings: $r = a, s = b, t = ab, u = ba, w = aa, x = aba, y = baa, z = abaa.$

Then the directed graph for the prefix relation R on $X = [1/s, 1/u, 1/w, 1/x]$ is presented in fig A, while that of the substring relation R on $X = [3/r, 1/s, 1/t, 1/u, 1/w, 1/x, 1/y, 1/z]$ is presented in fig B below.

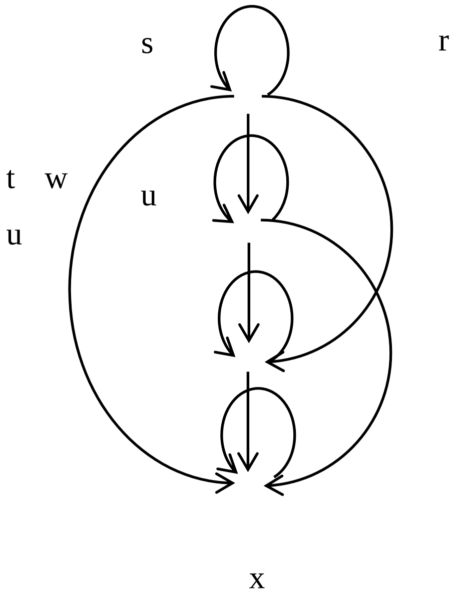

Fig A
Graph of the prefix relation

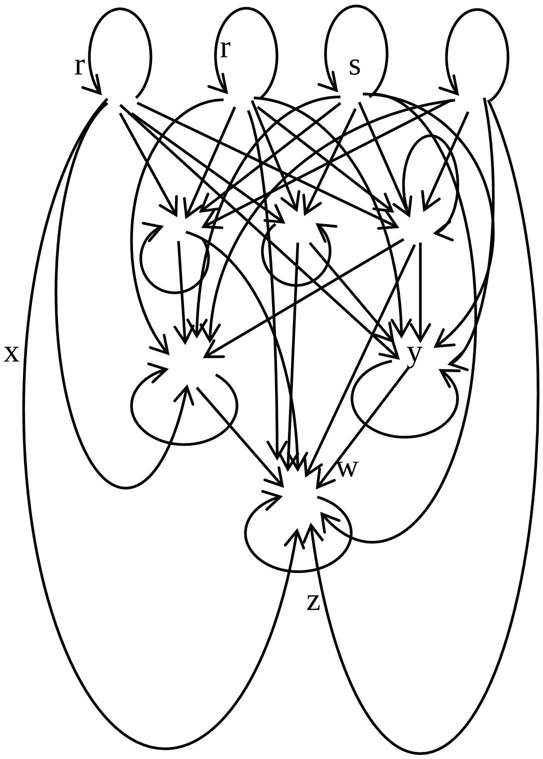

Fig B
Graph of the substring relation

Proposition 3.2.2

A prefix (suffix or substring) relation graph is a directed acyclic graph for $n \geq 2$. **Proof**
Let $G = (V, E)$ denote the graph of a prefix (suffix or substring) relation and let $(m/x, n/y) \in E$ be represented as $m/x\ E\ n/y$.

Let $n \geq 2$, and the graph is cyclic. That is, there exist distinct elements $m_1/x_1, \ldots, m_k/x_k \in V$ such that $m_1/x_1\ E\ \ldots\ E\ m_k/x_k\ E\ m_1/x_1$.

Since these relations are transitive, $m_1/x_1\ E\ m_2/x_2$ and $m_2/x_2\ E\ m_3/x_3$ imply $m_1/x_1\ E\ m_3/x_3$ and, in turn, by repeated application of transitivity, we have $m_1/x_1\ E\ m_k/\ x_k$. Hence, $m_1/x_1 = m_k/x_k$, by antisymmetry of these relations. This contradicts the assumption that $m_1/x_1, \ldots, m_k/x_k$ are distinct and $n \geq 2$.

Hasse diagram of $\boldsymbol{R}$ on $\boldsymbol{X}$

Let $\mathbb{X} = (X, R)$, where X consists of all nonempty prefixes (suffixes or substrings) of a string $v \in \Sigma^{\oplus}$ and R, its respective relation on X. In order to represent the relations by Hasse diagram, we assume that $\forall m/x, n/y \in X, m/x\ R\ n/y$ only if $\nexists\ p/z \in X$ such that $m/x\ R\ p/z$ and $p/z\ R\ n/y$.

For each ordered pair for which this relation holds, draw m/x in a vertical plane below n/y (or n/y below m/x) and connect them with a straight line, i.e., a Hasse diagram for prefix (suffix or substring) relation is a directed graph with vertex mset V and edges set E minus self-loops and edges implied by transitivity.

As an example, the Hasse diagrams for the prefix relation in fig A and, that of the substring relation in fig B, are presented in Fig C and D, respectively.

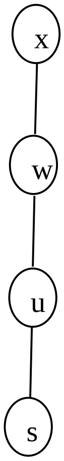

Fig C

Hasse diagram for the prefix relation

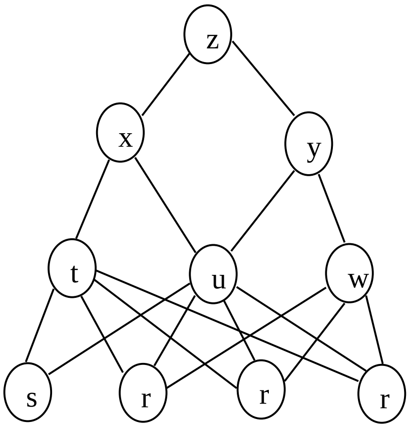

Fig D

Hasse diagram for the substring relation

Remark 3.2.3

Akin to set relations, two or more different prefix (suffix or substring) mset relations may have similar Hasse diagrams. For example, the Hasse diagram of the suffix relation R on $X = [k, l, m, n]$, for the nonempty suffixes of the string $abaa$ with $k = a, l = aa, m = baa, n = abaa$, is similar to that of the prefix relation presented in fig C.

4 Monoids of prefixes, suffixes and substrings of a nonempty string in a multilanguage

In view of the fact that homomorphisms of semigroups and monoids have useful applications in the *economic design of sequential machines* and in *formal languages*, this section introduces these notions in a multilanguage.

Let Σ be an alphabet and $\Sigma^{\oplus}$, the mset of all nonempty strings over Σ. For $v \in \Sigma^{\oplus}$, let $v^{\circledast}$ denote the multiset of all prefixes (suffixes or substrings) of v, and R its respective relation. We define a *binary* operation $*$ on $v^{\circledast}$ as follows:

i. $\forall m/x, n/y \in v^{\circledast}$, if $m/x \; R \; n/y$ and $\nexists \; k/z \in v^{\circledast}$ such that $m/x \; R \; k/z \land k/z \; R \; n/y$, then $m/x * n/y = n/y$, ii. $\forall m/x, n/y \in v^{\circledast}$, if $m/x \; R \; n/y$ and $\exists$ $k_1/z_1, k_2/z_2, \dots, k_n/z_n \in v^{\circledast}$, such that $m/x \; R \; k_1/z_1 \; R \; k_2/z_2 \; R \dots R \; k_n/z_n \; R \; n/y$, then $m/x * n/y = m/x *$
$k_1/z_1 * k_2/z_2 * \dots * k_n/z_n * n/y$.

It is easy to see that $*$ is associative and ε is the identity element. Hence, $(v^{\circledast}, *, \varepsilon)$, usually written as $(v^{\circledast}, *)$, is a monoid. Also, every element $m/x \in v^{\circledast}$ is idempotent with respect to the operation $*$, since $m/x * m/x = m/x, \forall m/x \in v^{\circledast}$. However, $*$ is not commutative as $m/x * n/y \neq n/y * m/x, \forall m/x, n/y \in v^{\circledast}$.

For example, let $\Sigma = \{a, b\}$, and let $v = abaa \in \Sigma^{\circledast}$. Consider the substrings $\varepsilon, a, b, ab, a, ba, aba, a, aa, baa, abaa$ of v and let $p = \varepsilon, r = a, s = b, t = ab, u = ba$, $w = aa, m = aba, n = baa$, and $q = abaa$. Then, $v^{\circledast} = [1/p, 3/r, 1/s, 1/t, 1/\, u, 1/m$, $1/w, 1/n, 1/q]$.
It is straightforward to see that $*$ is associative, since $\forall m/x, n/y, k/z \in v^{\circledast}$, $(m/x * n/y) * k/z = m/x * (n/y * k/z)$, and $1/p$ is the identity element. Therefore, $(v^{\circledast}$, $*)$ is a monoid.

Functions between Monoids of Prefixes, Suffixes and Substrings
Let $M = (v^{\circledast}, *)$ and $N = (u^{\circledast}, *)$ be two monoids of prefixes (suffixes or substrings) of $v, u \in \Sigma^{\circledast}$. A *homomorphism* between M and N is a function $f\colon v^{\circledast} \longrightarrow u^{\circledast}$ such that for every element $m/x \in v^{\circledast}$ there exists exactly one element $n/y \in u^{\circledast}$ such that $(m/\, x$, $n/y) \in f$, the pair (x, y) occurs only $C_1(x, y)$ times and, for every pair of elements m/x, $n/y \in v^{\circledast}$, if $m/x * n/y$, then

(i) $f(m/x * n/y) = f(m/x) * f(n/y)$; and

(ii) $f(\varepsilon_{v^{\circledast}}) = \varepsilon_{u^{\circledast}}$, where $\varepsilon_{v^{\circledast}}$ and $\varepsilon_{u^{\circledast}}$ are the same identity on $v^{\circledast}$ and $u^{\circledast}$ both, as mentioned earlier.

For example, let $\Sigma = \{a, b\}$ and $v = aba$, $u = abaa \in \Sigma^{\circledast}$; then their substrings are $\varepsilon, a, b, ab, a, ba, aba$ and $\varepsilon, a, b, ab, a, ba, aba, a, aa, baa, abaa$, respectively. For the

substrings of v, let $d = \varepsilon, k = a, m = b, n = ab, p = ba, q = aba, v^{\circledast} = [1/d, 2/k, 1/m, 1/n, 1/p, 1/q]$ and, for that of u, let $e = \varepsilon, r = a, s = b, t = ab, h = ba, w = aa, x = aba, l = baa, z = abaa, u^{\circledast} = [1/e, 3/r, 1/s, 1/t, 1/h, 1/w, 1/x, 1/l, 1/z]$. Then,
$f\colon (v^{\circledast}, *) \longrightarrow (u^{\circledast}, *)$, defined by $f\{(1/d, 1/e)/1, (2/k, 2/r)/2, (1/m, 1/w)/1, (1/n, 1/l)/1, (1/p, 1/l)/1, (1/q, 1/z)/1\}$, is a *monoid homomorphism of substrings*,
while $g\colon (v^{\circledast}, *) \longrightarrow (u^{\circledast}, *)$, defined by $g\{(1/d, 1/e)/1, (2/k, 2/r)/2, (1/m, 1/h)/1, (1/n, 1/w)/1, (1/p, 1/x)/1, (1/q, 1/z)/1\}$ is not a homomorphism, since $g(1/m * 1/n)$ does not imply $g(1/m) * g(1/n)$.

Let $f\colon M = (v^{\circledast}, *) \longrightarrow N = (u^{\circledast}, *)$ be a monoid homomorphism. We define
(i) $f\colon M \longrightarrow N$ is an injection iff $f\colon v^{\circledast} \longrightarrow u^{\circledast}$ is an injection, and
$\forall x(x \in v^{\circledast} \Longrightarrow |x| \leq |f(x)|)$,
(ii) $f\colon M \longrightarrow N$ is a surjection iff $f\colon v^{\circledast} \longrightarrow u^{\circledast}$ is a surjection, and
$\forall x(x \in v^{\circledast} \Longrightarrow |x| \geq |f(x)|)$, and
(iii) $f\colon M \longrightarrow N$ is a bijection iff $f\colon v^{\circledast} \longrightarrow u^{\circledast}$ is a bijection, and
$\forall x(x \in v^{\circledast} \Longrightarrow |x| = |f(x)|)$.
Note that $|u|$ denotes the length of the string u.

For example, consider the strings $k = bab, m = aba \in \Sigma^{\circledast}$. Then, $\varepsilon, b, a, ba, b, ab, bab$ and $\varepsilon, a, b, ab, a, ba, aba$, respectively are their substrings.
For k, let $x = \varepsilon, n = a, p = b, q = ba, r = ab, s = bab, k^{\circledast} = [1/x, 2/p, 1/n, 1/q, 1/r, 1/s]$, and for m, let $y = \varepsilon, t = a, u = b, v = ab, w = ba, z = aba, m^{\circledast} = [1/y, 2/t, 1/u, 1/v, 1/w, 1/z]$, then $f\colon (k^{\circledast}, *) \longrightarrow (m^{\circledast}, *)$, defined by $f\{(1/x, 1/y)/1, (2/p, 2/t)/2, (1/n, 1/u)/1, (1/q, 1/v)/1, (1/r, 1/w)/1, (1/s, 1/z)/1\}$, is an isomorphism.

Concluding Remarks

In view of the fact that *strings* are known to be relatively (in comparison with *integers* or even *booleans*, etc.) suitable basic data types in a (querry) language, and as many (database) languages and systems do require multiset-based semantics, formulation of prefix, suffix and substring relational structures in this paper may be found useful in construction of formal power series developed in [10], for example. Monoids and their homomorphisms described in this paper may find useful applications in the economical design of sequential machines and formal languages.

Acknowledgement: We are beholden to Professor D. E. Knuth for making paper [6] available to the first Author of this paper in a personal correspondence.

References

[1] J. Gallier, Introduction to the Theory of Computation, Formal Languages and Automata Models of Computation, Lecture Notes, (2010) 1- 60.

[2] J. Kari, Automata and Formal Languages, University of Turku, Lecture Notes, (2011) 1 – 150.

[3] M. Kudlek, C. Martin-Vide, and G. Pãun, *Toward a Formal Macroset Theory*, In: Multiset Processing, C.S. Calude *et al.* (Eds.), LNCS 2235, Springer – Verlag, (2001) 123-133.

[4] E. Csuhaj–Varjύ, C. Martίn–Vide and V. Mitrana, Multiset Automata*, In:

Multiset Processing, C.S. Calude *et al.* (Eds.), LNCS 2235, Springer – Verlag, (2001) 69 – 83.

[5] K. Ruohonen, Formal Languages, Lecture Notes, (2009) 1-94.

[6] D. E. Knuth, *Context-free Multilanguages*, In: Theoretical Studies in Computer Science, Jeffrey D. Ullman, Symour Ginsburg (Eds), Academic Press, (1992) 1-13. [7] D. Singh, A. M. Ibrahim, Y. Tella, and J. N. Singh, An Overview of the Applications of Multisets, NOVI Sad J. Math, 37 (2) (2007) 73 – 92.

[8] K. P. Girish and J. J. Sunil, General relations between partially ordered multisets and their chains and antichains, Mathematical Communications, 14 (2) (2009) 193-205.

[9] P. W. P. J. Grefen and R. A. de By, A Multiset Extended Relational Algebra: A Formal Approach to a Practical Issue, In: 10th International Conference on Data Engineering (ICDE), Houston, TX, USA, (1994) 80 - 88.

[10] A. Salomaa, Formal Languages, Academic press, New York San Francisco, 1973.

Received: August 17, 2014