Chapter 2

Review of Literature

A language is a system of communication. But a formal language is a rigorous mathematical representation of its alphabet of symbols and formation rules. Formal language is an essential as well as interesting subject of both Linguistics and Computer Science. In Linguistics, formal languages are used for the scientific study of human language. In Computer Science, formal languages are used for the precise definition of programming languages and therefore it is useful in the design of compiler. Though extensive research work are done on formal language but still there remain much scope of research in this field.

This chapter is divided into four sections. The first section is dedicated to general information about formal languages. Formal language is a huge topic. Due to the importance of the formal language for this research problem these are placed at the very first section. In the next section finite state automata are discussed. In the third section Parikh matrices are discussed. The last section deals with the previous works on development of context-free grammars for natural languages .

2.1 Basics of Formal language:

In the very beginning some basic information is discussed about formal language to get an easy entrance to the subject.

- A set is a collection of objects. A set is denoted by uppercase letters. The objects of a set are called elements. The empty set is denoted by Φ.

- The union of two sets A and B is defined and denoted by $A \cup B = \{x | x \in A \text{ or } x \in B\}$ eg. For $A = \{a, b, c, d, e\}$ and $B = \{e, f\}$; $A \cup B = \{a, b, c, d, e\}$.

- The intersection of two sets A and B is defined and denoted by $A \cap B = \{x | x \in A \text{ and } x \in B\}$ eg. For $A = \{a, b, c, d, e\}$ and $B = \{e, f\}$; $A \cap B = \{e\}$.

- The difference of these two sets A and B is defined and denoted by $A - B = \{x | x \in A \text{ and } x \notin B\}$ eg. For $A = \{a, b, c, d, e\}$ and $B = \{e, f\}$; $A - B = \{a, b, c\}$. When two sets have no element in common then these sets are called disjoint sets.

- The cardinality of a set is the number of elements of the set. The cardinality of the set $A = \{a, b, c, d, e\}$ is $|A| = 5$.

- The Cartesian product is defined as $A \times B = \{(x, y) : x \in A \text{ and } y \in B\}$. For $A = \{a, b, c, d, e\}$ and $B = \{e, f\}$;
$A \times B = \{(a, e), (b, e), (c, e), (d, e), (e, e), (a, f), (b, f), (c, f), (d, f), (e, f)\}$.

- Relation is a subset of the Cartesian product. A relation from X to Y is a set of ordered pairs (x, y) with $x \in X$ and $y \in Y$.

- Function $f : A \to B$ is a relation in which every element of the domain A is uniquely assigned to an element of the codomain B.

- One-to-one function or injective function is a function $f : A \to B$ if $f(a) = f(b) \Rightarrow a = b$.
- Onto function or surjective function is a function $f : A \to B$ if $f(A) = B$.
- One-to-one onto function or bijective function is a function which is both one-to-one and onto.
- Invertible function is a function if its inverse relation is a function from B to A.
- Graph consists of a finite set V of vertices and a finite set E of edges and a function γ which assign each edge to a pair of vertices.
- Cartesian coordinate system is a coordinate system where the location of each point is determined in a plane by a pair of coordinates. The coordinates are the perpendicular distances of the point from two perpendicular lines known as axes.
- Three dimensional coordinate system is a coordinate system where the location of a point is determined in a three dimensional space by a triplet of coordinates. Each coordinate gives the distance of that point from the origin measured along the given axis, which is equal to the distance of that point from the plane determined by the other two axes.

The books [72, 73, 93] give the basic ideas related to the subject of formal language. [63] also gives all fundamental mathematical ideas of formal language theory.

Transformation of natural languages to formal language is an interesting area for research. Natural language processing has many hurdles like word sense disambiguation, multi-word expression, machine translation etc.

- The process of finding out the accurate sense of a word corresponding to a context is called word sense disambiguation.
- Multi-word expression is idiomatic expression made of two or more sequences that has properties that are not predictable from the properties of the individual sequence.
- Machine translation is the process of translating a text from one language to another by using machine .

The transformation of natural languages to formal language is not a smooth process. Extensive research work are going on in this field. The paper [54] and the thesis [37] throw some light on the same discipline. A very lucid comparison between natural languages and formal language is done in [35].

There are many scopes of formal language. It has the immense power to contribute and mix up with other branches of science. Like natural languages graph theory can also be associated with formal language. For example the paper [81] can be cited. Formal words are very interesting topic of research. So many experiments are done with formal words . Shuffle product of two words or languages are examples of these types of experiments. Shuffle product is the set of words obtained by interleaving the letters of these words such that the order of appearance of all letters of each word is respected. The shuffle products of two languages is the union of all the shuffle products of two words taken one from each of these two languages. [81, 85] are cited in this context.

2.2 On Finite state automata:

Finite state automata constitute an affluent section of theoretical computer science. Finite state automata are divided into two parts. Deterministic finite state automata (DFSA) and non-deterministic finite state automata (NFSA).

- DFSA is a machine that takes as input a sequence of symbols, and it answers with yes or no according to whether the sequence is accepted or rejected. A deterministic automata is one in which transition from one state to another is determined by the current configuration.
- If the internal state, input and contents of storage are known, but it is not possible to predict the future behaviour of the automaton, it is said to be non-deterministic finite state automaton (NFSA).

Every NFSA can be transformed to DFSA [39]. Many books [48, 95] and research papers [70], [22], [64] are written on the subject. Finite automata are used in text processing, compilers, and hardware design. Finite-state automata are very constructive mathematical models of computation used to design both computer programs and sequential logic circuits. There are numbers of papers which have investigated on the development of algorithms to construct finite state automata. A few examples [2, 30, 47, 33, 34, 91] are cited in this context. In the paper [16] an algorithm is also developed to determine the deterministic FSA corresponding to a regular word. The following literatures are also based on finite state automata [70, 22, 52, 42, 20].

A non-deterministic finite-state automaton where the input is an infinite word is called Büchi automaton. [66] is based on infinite input of FSA.

2.3 On Parikh matrix:

The core research problem of this present study is based on Parikh matrix. In 1966 R.J.Parikh introduced a notion called Parikh Mapping [69]. This notion is an important tool in the theory of formal languages. Using this tool words can be expressed as vectors. If the domain of a Parikh mapping is a context-free language then the range is always a semilinear set. Extensive research work are done on the subject [40], [71], [36]. It is a fact that from the transition of a sequence to a vector much of the information are lost about the sequence. To overcome this difficulty a new type of function based on matrix is introduced by [61]. With the extension of Parikh matrix an interesting interconnection between mirror images of words and inverses of matrices are investigated in [60]. The Parikh matrix mapping is not injective in general; this problem is known as M- ambiguity. The present study is mainly concerned with this M-ambiguity.

Since the introduction of the notion of Parikh vector in 1966 [69] continuous research work are going on in this field. In this introductory paper [69] Parikh vector first came into light. For the word $aabcadcd$ the Parikh vector is $(3, 1, 2, 2)$. From this array it is known that how many a s, b s, c s and d s are there in the word $aabcadcd$. Though it is a long time since it was discovered, Parikh vector has not lost its importance till date. The research work [23] is dealing with Parikh vector. A sharpening of Parikh mapping namely Parikh matrix is introduced in [61] and this matrix representation gives more information than Parikh vector does. Because for the same word $cbbac$ the Parikh matrix is
$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 2 & 2 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 \end{pmatrix}$. From this matrix along with numbers of a s, b s and c s

one comes to know numbers of ab s, bc s and abc s also. So Parikh matrix became more useful than Parikh vector was. Interesting point is that the representation of the formal words by Parikh matrix is not unique. As for example the above matrix has two more corresponding words $cbbca$ and $bccba$. So uniqueness is not attained. Though corresponding to a word there is one and only one matrix but on the other hand corresponding to a matrix there may be more than one word. This property is known as M-ambiguity or amiability. It is the main problem of Parikh matrix tool. Considerable work is done on the M-ambiguity. A few papers [7, 83, 9, 8, 77, 89, 15] are cited over M-ambiguity. The concept of subword is another important concept about Parikh matrix. In [79] subword condition was introduced. As mentioned before a word u is a sub- word of a word w, if there exist words $x_1 \cdots x_n$ and $y_0 \cdots y_n$, (some of them possibly empty), such that $u = x_1 \cdots x_n$ and $w = y_0 x_1 y_1 \cdots x_n y_n$. For example if $w = abaabcac$ is a word over the alphabet $\Sigma = \{a, b, c\}$ then abc is a sub-word of w. Two occurrences of a sub-word are considered different if they differed by at least one position of some letter. In the word $w = abaabcac$, the number of occurrences of the word abc as a sub-word of w is $|w|_{abc} = 8$. Till its introduction much work has been done on the topic. For example [76, 75, 77, 38, 86]. From [61] we have the result that the entries $m_{i,j+1}, 1 \leq i \leq j \leq k$ in a Parikh matrix $\Psi_{M_k}(w)$ satisfy the inequality $m_{i,j+1} \leq m_{i,j} \leq m_{i+1,j+1}$. For ternary sequence, the inequalities $m_{1,3} \leq m_{1,2} \leq m_{2,3}$, $m_{1,4} \leq m_{1,3} \leq m_{2,4}$, $m_{2,4} \leq m_{2,3} \leq m_{3,4}$ are satisfied.

In the paper [32] it is given that for $w \in Z^* = \{a, b, c\}^*$ using the notation

$$\Psi_{M_3}(W) = \begin{pmatrix} 1 & A & E & x \\ 0 & 1 & B & F \\ 0 & 0 & 1 & C \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

it can be obtained that $AF + CE - ABC \leq x \leq \frac{EF}{B}$. In the paper [17], a new approach has been discussed to prove the above theorems.

Adrian Atanasiu proposed in the paper [6] that the Parikh matrix $M_a = \begin{pmatrix} 1 & p & q \\ 0 & 1 & n \\ 0 & 0 & 1 \end{pmatrix}$ corresponds to a binary word $\alpha = a^{x_1}ba^{x_2}b \cdots a^{x_n}ba^{x_{n+1}}$ (by detailing the appearances of letter b) if and only if $(x_1, x_2, x_3, \cdots, x_{n+1}) \in \mathbb{Z}^{n+1}$ (where $\mathbb{Z}$ is the set of Natural numbers including 0) is a solution of the system

$$\begin{cases} x_1 + x_2 + \cdots + x_{n-1} + x_n + x_{n+1} = p \\ nx_1 + (n-1)x_2 + \cdots + 2x_{n-1} + x_n = q. \end{cases}$$

It is a Diophantine equation.

A linear Diophantine equation (in two variables) is an equation of the general form $ax + by = c$, where solutions are sought with a, b and c integers. Such equations can be solved completely. A general quadratic Diophantine equation in two variables x and y is given by $ax^2 + cy^2 = k$, where a, c and k are specified (positive or negative) integers and x and y are unknown integers satisfying the equation whose values are sought. In the published paper [19], similar results are derived for binary words using algebraic equations and it thus suggested that a new set of algebraic equations corresponding to a binary word can also be used for finding amiable words. Similar set of algebraic equations corresponding to a ternary word is proposed in [17]. This conception is extended to tertiary sequences in [18]. Theory used in above papers is Algebraic solvability. When expressed as a polynomial equation of degree n such as $x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1x + a_0 = 0$ it is said to be algebraically solvable if its roots $x_1, ..., x_n$ can all be expressed by algebraic expressions in the coefficients $a_0, \cdots, a_{n-1}$. The roots must be expressible as finite combinations of the coefficients and constants using the five algebraic operations addition, subtraction, multiplication, division, and root extraction.

Ratio property and weak-ratio property introduced in [87] also throws light on solving M-ambiguity. The ratio property is a sufficient condition for the words to have the same Parikh matrix. The weak-ratio property is used in [57] to prove some interesting lemma regarding the commutativity of the Parikh matrix of two words over binary and ternary alphabets. In the paper [89] the weak-ratio property is extended to binary arrays and new extensions of this property namely row-ratio property and column-ratio property are introduced. Various results are also studied in that field. Weak-ratio property is also used in [88] with association with Istrail morphism to explore a new area investigation of morphic images of words. In the paper [15] effort has been given to study ratio property and weak ratio property in a generalized approach . Words over tertiary alphabet are discussed in the light of ratio property. It is seen that two words satisfying ratio property make M-ambiguous words when they are allied in certain ways. In the paper [80] the concept of position indicators for binary alphabets is discussed. Using this concept picture image of a binary word canbe found out. In the paper [89] a chessboard pattern of representation of binary arrays is discussed. The graphical representation of binary words discussed in [19] gives another type of picture representation of binary words. Moreover this type of representation can be extended to ternary words [17].

There are many fields of studies where Parikh matrix can be applied. The following research papers are on application of Parikh matrix on various fields. [10] in coding theory, [3] is on l-morphism of Parikh matrix. [31] is on the Parikh counting functions of sparse context-free languages. [45] is on Parikh slender languages and power series. Subword balance, position indices and power sums are described in [78]. [46] is on slender languages. One application of Parikh matrix is on message authentication code. Languages attached to Parikh Matrices are described in [11].

Parikh matrix, being a field of much interest and vast scope, can be approached and studied from other branches of knowledge also.

2.4 On development of context-free grammar for natural languages:

In this present study some effort is also given to the application of Parikh matrix in the field of natural language processing. Parikh matrix is based on context-free language. There are some methods by which context-free grammar for natural languages can be developed. Some methods are explored and some methods are yet to be explored. Some of the papers which deal with the development of context-free grammar for Bengali are [82, 43, 12, 58, 67, 44, 65, 5]. In [90], probabilistic context-free grammar for Hindi language is developed.

From the very beginning of Linguistic subject, Mathematics is used to solve various problems. Application of Mathematics in Linguistics is a common practice. In this connection the books [49, 50] have important contributions. These two books are very helpful to study the basics of Mathematics necessary for understanding Linguistic. The book [50] is for Linguists. The two books [1, 56] are also on computational linguistics.

As Parikh matrix is only applicable to formal languages so those literatures which developed context-free grammar for natural languages are important for the study. [28, 92, 27] are cited in this context. Formal theory of natural language syntax was first introduced by Noam Chomsky [24, 26]. Chomsky gave the idea of context-free, transformational phrase structure grammar formalisms and their comparison. In this present study it is tried to transfer a tool from formal language to natural languages.

Recognition of the sentence structure is called parsing. Context-free parsing techniques are well suited to be incorporated into real-world natural language processing systems because of their time efficiency and low memory requirements. Though, it is known that some natural language phenomena cannot be handled with the context-free grammar formalism, researchers often use the context-free idea as the core of their grammar formalism and enhance it with context-sensitive feature, for example, the paper [68]. The main contribution of [53] is in a syntactic analyser for Czech language. However, the presented algorithms are language-independent, so other languages with an appropriate grammar can be modelled as well.

Parsing also means taking a sentence of any natural language as an input and producing some sort of linguistic structure for it. Like other language processing, parsing a Bengali sentence is a primary need for Bengali language processing.

To examine a given language is regular or not, it is sufficient to build a regular expression for the language. To prove that a given language is not regular, it is customary to use some tool. A good tool for doing this is the Pumping Lemma. A strong pumping lemma for context-free languages is described in [94]. The systems discussed in [84] can be regarded as context-free grammars with weights.

Assigning mathematical tools to natural languages is a subject of research interest. It is also a challenging area of interest. This is discussed in [13]. Linguistic and logic are studied in [24] side by side. The research work [21] is also on mathematical application on Linguistics. [41] is a very useful book on computational Linguistics. [51, 14, 4] are some useful books on natural language processing. The book [14] is a familiar book on NLP. [4] is a book on automatic ambiguity resolution in Natural language processing.

Formal language is a mandatory subject in the field of Computer Science. The subject is very rich in its contents and continuous pouring of knowledge in this field make the subject an ever expanding one.

*****