# Natural Language Processing Through Different Mathematical and Statistical Tools

## Thesis

submitted to the

Kumaun University, Nainital

By

**Raj Kishor Bisht**

For the award of the degree of

**Doctor of Philosophy**

in

**Mathematics**

Under the supervision of

**Prof. H.S. Dhami**

Department of Mathematics,
Kumaun University, S.S.J. Campus Almora,
Uttarakhand (India)-263601

**2008**

# Certificate

This is to certify that the thesis entitled **"Natural Language Processing Through Different Mathematical and Statistical Tools"** submitted to Kumaun University, Nainital for the award of the degree of **Doctor of Philosophy** in Mathematics is a record of bona fide work carried out by Mr. Raj Kishor Bisht, under my guidance and supervision.

I hereby certify that he has completed the research work for the full period as required in the ordinance 6. He has put in the required attendance in the department and signed in the prescribed register during the period. I also certify that no part of this thesis has been submitted for any other degree or diploma.

( Prof. H. S. Dhami)  
Supervisor,  
Professor, Department of Mathematics  
Kumaun University, S.S.J.Campus Almora  
Uttarakhand (India)

Forwarded by

( Prof. S. B. Pandey )  
Convener,  
RDC and Board of Studies in Mathematics  
Kumaun University, Nainital, Uttarakhand (India)

# Preface

In September 2002, I started my research work after qualifying National Eligibility Test (CSIR) with junior research fellowship under the supervision of Prof. H. S. Dhami at Dept of Mathematics, Kumaun University, S.S.J. Campus Almora. My respectable guide has suggested me to do some application oriented work. At that time I can remember, my guide had received a letter from ISI Kolkata in which different natural language processing tasks have been described for research purpose as my guide had previously done some work in Mathematical Linguistics. He advised me that natural language processing (NLP) is the fast emerging field in India and some remarkable mathematical work can be done in this field. This was the beginning of my research work. I started searching topics in NLP and gone through the literature but could not get the satisfaction. I saw a book “Foundations of statistical natural language processing” on the web. This book is one of the leading books devoted to mathematical and statistical tools applicable in natural language processing. It took almost six months after starting my research work to get this book in my hand from USA. After getting this book, I gone through this book and got some basic idea in NLP. I went to IIT Kanpur and collected some of the research papers in this field and discussed these with my guide. The subject was easy to understand but very tough to do some work in this as it requires a lot of computational work. After casting my nets in all possible directions for one year, I was disappointed and I even requested my guide to change the topic. We discussed all the difficulties in this field. He listened all my problems and encouraged me that I have the capability to do some work in this field in spite of the difficulties associated with it and said that a little achievement in this field can lead to a better future. His words given me a new energy to restart my work on the same topic. Rejuvenated by his encouraging remarks I gathered all momentum and sent research papers to the leading journals of the field but still could not get any success as they demand a lot of practical computationl work which was not possible at our level at that time. I came in contact of Professor M.M. S. Rauthan, Head, Dept. of Computer Science, HNB Garhwal University and after discussions of several rounds could get success in developing computer programs of the mathematical formalism, done by us. My works started paying results when one of my research papers was accepted for publication in Taylor Francis Group and it proved to be a turning point in my research work. After two and half years with the due permission of my guide and CSIR, I resigned the senior research fellowship and started my teaching carrier at Amrapali Institute, Haldwani (Uttarakhand) and continued my research work. Teaching in MCA classes gave me some advantage in this field. During the period I also attended short term courses in Discrete mathematical Structure at MANIT Bhopal, Natural Language Processing and Text Mining at IIT Kharagpur and Applied Numerical methods in IIT Roorkee. My contacts and exposure with workers and authorities of the field made me to realize the potential and applicability of my research work. I have now to my credit substantial in the form of publications of proceedings of national and international conferences. I still think this is my first step in this field and still miles have to be covered.

# Acknowledgement

I would like to express my sincere thanks to my honourable guide Prof. H. S. Dhami, Dept. of Mathematics, Kumaun University S. S. J. Campus Almora for his continuous encouragement and support to me. I have no words to convey my gratitude to him as he had been disturbed by me very often due to my frequent visit to his house. I would also like to thanks Smt. Shanti Dhami, W/o Prof. H.S.Dhami for the affections, I got from her during the period of my research work and Mr. Himanshu Dhami for his help in purchasing books and research papers online.

I am very thankful to Dr. Neeraj Tewari, Reader, Department of Statistics, Kumaun University, S. S. J. Campus Almora for his valuable suggestions and kind support.

I would like to express my gratitude to Prof. M.M.S. Rauthan, Head, Department Computer Science, H.N.B.Garhwal University, Garhwal for having discussion many times in various issues related to computational work.

I convey my gratitude to all faculty members of Department of Mathematics, Kumaun University S.S.J. Campus Almora for their kind cooperation.

I am thankful to my colleague research scholars at Department of Mathematics, Kumaun University S.S.J. campus Almora for their assistance.

I am very thankful to Mr. Shailesh Lohiya, Librarian, Amrapali Institute, Haldwani for helping me in getting various research papers.

I would like to express my sincere gratitude to the management of Amrapali Institute for their support in attending various conferences/ seminars.

I am very thankful to my colleague faculty members of Computer Science Department, Amrapali Institute, Haldwani for having discussion in various problems.

I would like to express my thanks with gratitude for the help received by me from the staff members of various libraries, I have visited during the period.

I would like to acknowledge with thanks the help and affection I got from my mother in law Mrs. Bimala Pant, father in law Mr. H.C.Pant and brother in law Mr. Bhasker Pant.

I am indebted to my father Mr. D.K.Bisht, brother Mr. Yugal Kishore Bisht, sister Mrs. Hema Tewari, brother in law Mr. Yogesh Tewari for their continuous moral support, encouragement and blessing.

For my wife Ila, I have no words to express my feelings as she is the person with whom I have discussed every problem. I am indebted to her for the love, affection, encouragement and suggestions I received from her during the years.

I am thankful to all the people who have helped me directly or indirectly during the years.

Le last but not least, I am very grateful to almighty God for giving me the strength, patient and his blessing to complete the research work.

Date: (Raj Kishor Bisht)

# Research Paper published out of the thesis

Raj Kishor Bisht, H. S. Dhami and Neeraj Tiwari (2006), An Evaluation of Different Statistical Techniques of Collocation Extraction Using a Probability Measure to Word Combinations, Journal of Quantitative Linguistics, Volume 13, Numbers 2 – 3, 161 – 175.

Raj Kishor Bisht, H. S. Dhami (2007) "Extracting collocation from a small sample of text", to be appear in the proceedings of 12th annual conference of Gwalior Academy of Mathematical Science at MANIT Bhopal.

Raj Kishor Bisht, H. S. Dhami (2008), On Some properties of content words in a document, to be appear in the proceedings of 6th international conference of information science technology and management IIT Delhi.

# Research Paper communicated

Raj Kishor Bisht, H. S. Dhami (2008) Fuzzy set theoretic approach to collocation extraction, under review process of the journal 'Fuzzy Decision Making'.

Raj Kishor Bisht, H. S. Dhami (2008) Matrix representation of words, under review process of the journal 'Theory of Computing'.

## Research Papers presented in conferences /Seminars

“Collocations: An application to English-Hindi computational lexicography”, presented in the national seminar on “Era of Globalization: Challenges in Management & IT” held on 25-26 November 05 at AIMCA, Haldwani.

“Distribution of words in a text” presented in the 8th conference of the International Academy of Physical Sciences held on 29-31December 05 at Meerut. (Not the part of the thesis)

# Contents

# Chapter I

# A Brief Historical Survey and Outline of the Program

## 1.1 Introduction

We live in the age of information. It pours upon us from the pages of newspapers and magazines, radio loudspeakers, TV and computer screens. Our ancestors invented natural language many thousands of years ago for the needs of a developing human society. The structure and use of a natural language is based on the assumption that the participants of the conversation share a very similar experience and knowledge, as well as a manner of feeling, reasoning, and acting. The great challenge of the problem of intelligent automatic text processing is to use unrestricted natural language to exchange information with a creature of a totally different nature: the computer. For the last two centuries, humanity has successfully coped with the automation of many tasks using mechanical and electrical devices, and these devices faithfully serve people in their everyday life. In the second half of the twentieth century, human attention has turned to the automation of natural language processing. People now want assistance not only in mechanical, but also in intellectual efforts. They would like the machine to read an unprepared text, to test it for correctness, to execute the instructions contained in the text, or even to comprehend it well enough to produce a reasonable response based on its meaning. Human beings want to keep for themselves only the final decisions. The processing of natural language has become one of the main problems in information exchange. The rapid development of computers in the last two decades has made possible the implementation of many ideas to solve the problems that one could not even imagine being solved automatically, say, 45 years ago, when the first computer appeared.

The vast expressive power of language is made possible by two principles: the arbitrary sound meaning pairing underlying words and the discrete combinatorial system underlying grammar. All kinds of unexpected events can be communicated, because our knowledge of language is couched in abstract symbols that can embrace a vast set of concepts and can be combined freely into an even vaster set. This logic is in conformity with the abstraction principles of mathematics and therefore mathematics plays a vital role in processing of natural languages. When mathematical tools are applied with the motivation for the quest for knowledge and the desire to understand the nature of human language and the common patterns that occur in language use, this very interesting and applied field of interest is known as ***Mathematical Linguistics.*** This branch of applied Mathematics introduces the mathematical foundations of linguistics to computer scientists, engineers, and mathematicians interested in natural language processing. The main thrust in this arena is on the development of a basic mathematical structure of human languages rather than of a particular language· It is also a kind of mathematical modelling in which we try to find out the solutions of different real world problems arising in linguistics by using different mathematical and statistical techniques.

Mathematical linguistics is rooted both in Euclid’s (circa 325–265 BCE) axiomatic method and in Panini’s (circa 520–460 BCE) method of grammatical description. To be sure, both *Euclid* and *Panini* built upon a considerable body of knowledge amassed by their precursors, but the systematicity, thoroughness, and sheer scope of the *Elements* and the *Ashtdhyayı* would place them among the greatest landmarks of all intellectual history even if we disregard the key methodological advance they made. One can observe that the two methods are fundamentally very similar: the axiomatic method starts with a set of statements assumed to be true and transfers truth from the axioms to other statements by means of a fixed set of logical rules, while the method of grammar is to start with a set of expressions assumed to be grammatical both in form and meaning and to transfer grammaticality to other expressions by means of a fixed set of grammatical rules.  The antique works of Euclid and Panini demonstrate that although the intellectual roots of modern linguistics go back thousands of years but considerable interest in applying the then newly developing ideas about finite state machines and other kinds of automata, both deterministic and stochastic, to natural language started after the works of Chomsky (1994) and Zellig Harris (1982).

Mathematical linguistics mainly comprises two areas of research: the study of statistical structure of texts and the construction of mathematical models of phonological and grammatical structure of language. These two branches of mathematical linguistics may be termed as statistical and algebraic linguistics. So far the mathematical concepts used by linguists come mostly from algebra; mapping between sets; relation between mappings; partially ordered sets; semi groups; free groups; free algebra etc.  When we speak of mathematical linguistics, the need of the hour on one side is that mathematicians be familiar with the language technology and make attempts to provide the successful, popular and convincing solutions to linguistics problems while on the other side those recent advances in mathematics are presented before linguistics that may be relevant to language modelling community.

In the space of the last ten years, statistical methods have gone from being virtually unknown in computational linguistics to being a fundamental given. In 1996, no one can profess to be a computational linguist without a passing knowledge of statistical methods. More seriously, statistical techniques have brought significant advances in broad-coverage language processing. Statistical methods have made real progress possible on a number of issues that had previously stymied attempts to liberate systems from toy domains; issues that include disambiguation, error correction, and the induction of the sheer volume of information requisite for handling unrestricted text and the sense of progress has generated a great deal of enthusiasm for statistical methods in computational linguistics. A large part of computational linguistics focuses on practical applications, and is little concerned with human language processing. Nonetheless, at least some computational linguists aim to advance our scientific understanding of the human language faculty by better understanding the computational properties of language. One of the most interesting and challenging questions about human language computation is just how people are able to deal so effortlessly with the very issues that make processing unrestricted text so difficult. Statistical methods provide the most promising current answers, and as a result those in the cognitive reaches of computational linguistics also share the excitement about statistical methods.

Quantitative linguistics (QL) is a branch of science that is not as new as is usually supposed. Its actual beginning was in the early 1930s, and the discipline developed mainly in Eastern Europe. Like computational linguistics, QL deals with linguistic phenomena from a mathematical point of view: This discipline employs mathematical analysis, probability theory and stochastic processes, and differential equations to model and understand phenomena of language and communication. The mathematical theories of QL are more developed as theories than merely tools to compute. The introduction of quantitative models actually gave a new impetus to an in-depth understanding of the nature of linguistic entities. This important branch of studies collected a number of specialists over the decades; but only recently did they find the opportunity to join in the International Quantitative Linguistics Conference, the first of which was held at the University of Trier, Germany, in September 1991, with 120 participants from 16 countries. This volume collects 22 papers accepted and presented, in addition to four general lectures given by invited speakers, and it provides a representative overview of the state of the art. The topical sections were eight in number, and covered a large spectrum of interests: from phonetics to statistics, from modeling to dialectology, as well as reports and projects of different kinds.

Yehoshua Bar-Hillel (1970) was an Israeli logician and philosopher of language who made significant contributions in a number of linguistic fields: formal and algebraic linguistics, logical aspects of natural language, and computational linguistics, in particular machine translation and information retrieval. His principal essays are included in two collections *Language and Information* and *Aspects of Languag*e. In most of his writings, Bar-Hillel's aim was to bridge the "disastrous" gap between logic and linguistics, believing that linguists (particularly semanticists) had ignored logic to their detriment; and that logicians had ignored linguistics by creating a formal system devoid of any relevance to natural language in actual use. His major contribution to algebraic linguistics was categorical grammar, a "decision procedure" for identifying constituents in grammatically well-formed sentences, based on bringing logic and linguistics closer. An account of his works has been presented in the book edited by Strazny Philip (2005).

In the current literature on natural language processing (NLP), a distinction is often made between "rule-based" and "statistical" methods for NLP. However, it is seldom made clear what the terms "rule-based" and "statistical" really refer to in this connection. According to the recently published Handbook of Natural Language processing (2000), NLP is concerned with “the design and implementation of effective natural language input and output components for computational systems”. The most important problems in NLP therefore have to do with natural language input and output. The main branches are being mentioned in the next few lines with their definitions-

- Part-of-speech tagging: Annotating natural language sentences or texts with parts-of-speech.
- Natural language generation: Producing natural language sentences or texts from non-linguistic representations.
- Machine translation: Translating sentences or texts in a source language to sentences or texts in a target language.

A method for solving an NLP problem ‘P’ is typically supposed to consist of two elements:

1. A mathematical model $M$ defining an abstract problem $Q$ that can be used to model $P$.
2. An algorithm $A$ that effectively computes $Q$.

Evaluation of NLP systems can have different purposes and consider many different dimensions of a system. Consequently, there are a wide variety of methods that can be used for evaluation. Many of these methods involve empirical experiments or quasi-experiments in which the system is applied to a representative sample of data in order to provide quantitative measures of aspects such as efficiency, accuracy and robustness. These evaluation methods can make use of statistics in at least three different ways:

- Descriptive statistics
- Estimation
- Hypothesis testing

Before exemplifying the use of descriptive statistics, estimation and hypothesis testing in natural language processing, it is worth pointing out that these methods can be applied to any kind of NLP system, regardless of whether the system itself makes use of statistical methods.

Statistical estimation becomes relevant when we want to generalize the experimental results obtained for a particular test sample. For example, suppose that a particular system '*s*' obtains accuracy rate '*r*' when applied to a particular test corpus. How much confidence should we place on r as an estimate of the true accuracy rate $\rho$ of system $s$ ? According to statistical theory, the answer depends on a number of factors such as the amount of variation and the size of the test sample. The standard method for dealing with this problem is to compute a confidence interval *i,* which allows us to say that the real accuracy rate $\rho$ shall lie in the interval $\left[r-\frac{i}{2}, r+\frac{i}{2}\right]$ with probability p. Commonly used values of $\rho$ are 0.95 and 0.99.

Statistical hypothesis testing is crucial when we want to compare the experimental results of different systems applied to the same test sample. For example, suppose that two systems $s_1$ and $s_2$ obtain an error rate of $r_1$ and $r_2$ , when measured with respect to a particular test corpus, and suppose furthermore that $r_1 < r_2$ .Can we draw the conclusion that $s_1$ has higher accuracy than $s_2$ in general? Again, statistical theory tells us that the answer depends on a number of factors including the size of the difference $r_2 - r_1$ , the amount of variation, and the size of the test sample. And again, there are standard tests available for testing whether a difference is statistically significant. Standard tests of statistical significance for this kind of situation include the paired t-test, Wilcoxon's signed ranks test, and McNemar's test.

Generally for the natural language processing tasks, we use texts and regard the textual context as a surrogate for situating language in a real word context. A body of text is called a corpus and several such collections of texts are known as corpora. For analysis and inference from the collection, Mathematics and Statistics have an important role to play in it as a language is a set of words of that language and every language has an alphabet on which the words are generated.

## 1.2. NLP Problems considered in the present study

There are numerous problems in natural language processing. This section contains information about those problems which have been addressed by us during the course of present study.

### 1.2.1. Weighting of words in a text

Information retrieval is the task of getting most relevant information against a query asked by a user in natural language. The goal of Information Retrieval research is to develop models and algorithms for retrieving information from document repositories. The system returns a list of document against the query entered by the user. There are two main models; exact match (Boolean) and relevance based model. Exact match system returns documents that contain the words of query, relevance based model assigns ranks to the documents according to their relevance to the query. Relevance based model can be defined formally as a quadruple $\lfloor D, Q, F, R(q_i, d_j) \rfloor$ where $D$ is a set of logical views of documents, $Q$ is a set of user queries, $F$ is the framework for modeling documents and queries and $R(q_i, d_j)$ is a ranking function which associates a numeric value to the document $d_j$ according to a system assigned likelihood of relevance to a given user query $q_i$. The quality of a ranking function is an important factor that determines the quality of the

IR system. The vector space model is one of the most widely used models for relevant document retrieval. In this documents and queries are represented in a high dimensional space corresponding to the importance or weight of words in the documents. The most relevant documents for a query are represented by the vectors which are closest to the query vector. Closeness is calculated by looking at the angle between query vector and document vector. Figure 1.1 shows a two dimensional vector space model. The two dimensions represent two words asked by user (For example, Indian economy). Documents are represented as vectors according to their weights with respect to the two words. A document vector which is most relevant to the query, i.e., ‘Indian Economy’ has a lowest angle from query vector.

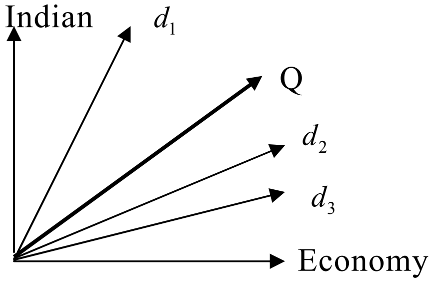

Fig.1.1 Vector space model

Weighting of words in vector space model is an important task. The other main focus of information retrieval research is to know the relevance of a word or a phrase to a particular document or to know how informative a word is, that appears in a document. For this purpose words have been divided in some categories to get some idea about their informative nature. ‘Content words’ are those words which determine the exposition of all ideas and facts pertinent to the document. ‘Function words’ are the determiners, prepositions, auxiliary verbs; etc, which are very often accompanied by content words. Function words are the integral parts of any document. The words or phrases relevant to a document form a set known as index language of the document and all those words or phrases contained in the index language are said to be index terms. Index language may be described as pre-coordinate or post-coordinate; the first indicates that terms are coordinated at the time of indexing and the latter at the time of searching. More specifically, in pre-coordinate indexing a logical combination of any index terms may be used as a label to identify a class of documents, whereas in post coordinate indexing the same class would be identified at search time by combining the classes of documents labeled with the individual index terms.

Mathematical and Statistical models are quite useful for indexing of words. A detailed quantitative model for automatic indexing based on some statistical assumptions about the distribution of words in a text has been worked out by Bookstein, Swanson (1974, 1975). One of the mathematical ways to know the relevance a word in a given document is to assign weights to words or terms. Weight of a term can be seen in two different aspects, local and global. Local weights are functions of frequency of a word in a document and global weights are functions of the frequency of a term in the entire collection. The first local weighting scheme was introduced by Luhn (1958). He proposed the use of a term frequency ($tf$) to measure the terms significance in the document which is defined as follows:

$tf_{i,k}$ = The frequency occurrence of the $i^{th}$ term in $k^{th}$ document

Term frequency provides the information about the importance of a word in a given document. It was considered that the higher frequency of a term indicates the excellence of the word in describing the contents of document however, using within document frequency alone is not enough because it cannot be used to discriminate document in the collection effectively. For example in the case where all the documents are composed of the article related to computer science, the frequency of the word computer will be obviously high. This situation shows that the word computer does not have ability in discriminating the documents. The more documents represented by a particular term, the less important this term is in terms of distinguishing one document from another. A good document representation has to be able to summarize and discriminate the documents at the same time. Term frequency provides a local weight calculation for each term as functions words appear a lot of times in a text. Semantically focused words will often occur several times in a document if it occurs at all and semantically unfocused words are spread out homogeneously over all documents. For example the words 'education' and 'make' both have different properties. 'Education' refers to a narrowly defined concept that is only relevant to a small set of topics while 'make' can be used in almost any topic.

A better version of the above idea was given by Sparck Jones (1972) who added an inverse document frequency $(idf)$ to the weighting as a global weight which can be formulated as follows:

$$idf = \log\left(\frac{N}{df_i}\right),$$

where $N =$ Number of documents in the collection of text.

$df_i$ = Number of documents in which the $i^{th}$ term appears.

The formula $\log(N/df_i) = \log N - \log df_i$, gives full weight to words that occur in one document $(\log N - \log 1 = \log N)$ and a word that occurred in all documents would get zero weight $(\log N - \log N = 0)$. Global weighting is important for discriminating terms because very high frequency words cannot be considered to be good discriminators if they appear in most of the documents in the collection.

Combining the term frequency and inverse document frequency, more accurate weight of a term (known as $tf.idf$ weighting) in the document is defined as:

$$W_{i,k} = tf_{i,k}.idf_i$$

Importance of a word in a document can be checked through different measures of variability such as variance, adaptation, Burstiness, entropy etc. So far these measures are used for global weighting. Statistical definition of these measures is given in section 1.3. Words can be further categorized in two ways. Content words in a text are those words which contain their selves the information about the text. Function words are those words which accompany content words. Articles such as ‘a’, ‘an’, ‘the’ and helping verbs such as ‘is’, ‘am’, ‘are’ etc. are the examples of function words.

Weight of a word in a document is still a research problem. In the present study an effort has been made to check the suitability of these measures for local weight.

### 1.2.2. Collocation extraction from a text

Machine translation is another important task of natural Language processing which deals with the automatic translation of one language into another through computer. It is very useful for a new learner of language. Quality of machine translation depends upon the excellence of the translation of words included in a sentence into another language. Some words when come together form a different meaning from the sum of their individual meanings or they have a unique meaning. ‘Collocations’ are one of such word combinations. A Collocation is an expression consisting of two or more words that correspond to some conventional way of saying things. Natural languages are full of collocations. Since a large number of collocations are present in any language, automatic identification of collocations from a text needs a statistical and computational model. The independence property $P(XY) = P(X)P(Y)$ of two random variables $X$ and $Y$ provides a way to check the dependence of words as collocation can be interpreted as a combination of words in which the words are dependent to each other. There are many statistical techniques of collocation extraction available in the literature. No method is perfect in itself. All methods provide a list of candidate collocations, which are transferred to a lexicographer for final decision. Precision and recall provide the appropriateness of a statistical method of collocation extraction. Precision gives the information about the accuracy of the retrieved list of collocations, while recall provides the percentage of collocations extracted from the text.

$$\Pr ecision = \frac{Actual\ collocations\ in\ the\ retrieved\ list\ of\ collocations}{Total\ number\ of\ retrieved\ word\ combinations}$$

$$\mathrm{Re}\,call = \frac{Actual\ collocations\ in\ the\ retrieved\ list\ of\ collocations}{Total\ number\ of\ collocations\ in\ the\ text}$$

In the present study we have evaluated the different statistical techniques of collocation extraction for our compiled corpus and suggested some other measures of collocations extraction.

### 1.2.3 Numerical and Algebraic properties of words

Words are the important structure of any language. Word structure can be studied in two ways:

(i) **Syntax**: Syntax is a precise rule that tell us the symbols we are allowed to use and how to put them together into legal expressions.

(ii) **Semantics**: It is a precise rule that tell us the meaning of the symbol and legal expression.

Every language is based on a set of letters. An alphabet $A$ is a finite set of symbols called letters of a language. For example $\{0, 1\}$ is an alphabet. The set $\{a, b, c, ......., z\}$ is an alphabet of English language. A syllable is a symbol $a^n$, where $a$ is a letter of the alphabet $A$ and $n$ is an integer. A word over $A$ is a finite sequence of elements of $A$. Words can also be defined as finite ordered sequence of syllables. For example $ab, abb$ etc are words over the alphabet $\{a, b\}$. Words in formal language satisfy some rules and words of a natural language in general do not follow a particular rule. Mathematical analysis of word structure is possible as it is a sequence of letters or syllables and therefore, study of mathematical structure of words is similar in formal and natural languages. Starting from formal languages, the mathematical results obtained for words can be extended to natural languages.

The properties of words in the form of numerical properties of words are firstly introduced in Parikh vector (1966). The Parikh mapping (vector) is an old and important tool in the theory of formal languages. One of the important results concerning the Parikh mapping is that the image by Parikh mapping of a context free language is always a semi linear set but Parikh vector does not provide too much information about a word. Alexandru et al (2001) introduced a sharpening of the Parikh mapping, that is, Parikh matrix, where somewhat more information about word is preserved than in the original Parikh mapping. It was based on a certain type of matrices. The classical Parikh vector appears in such a matrix as the second diagonal. The basic concepts paving the path for Parikh matrix are as follows:

### *1.2.3.1 Subwords*

Let $\sum$ be an alphabet. The set of all words over $\sum$ is denoted by $\sum^*$ and the empty word is $\lambda$. If $w \in \sum^*$, then $|w|$ denotes the length of $w$. Let $u, w \in \sum^*$. The word $u$ is said to be the scattered subword or simply subword of $w$ if $w$, as a sequence of letters, contains $u$ as a subsequence. Formally this implies that there exist words $x_1, x_2, \ldots\ldots, x_n$ and $y_1, y_2, \ldots\ldots, y_n$ in $\sum^*$, some of them possibly empty such that $u = x_1 \ldots\ldots\ldots x_n$ and $w = y_0 x_1 y_1 \ldots\ldots x_n y_n$.

The number of occurrences of a word $u$ as a subword in $w$ is denoted by $|w|_u$. For instance $|abba|_{ba} = 2$ and $|abbcca|_{abc} = 4$.

Parikh matrix mapping uses a special type of matrices, called triangular matrices. A triangular matrix is a square matrix $M = [m_{i\,j}]_{1 \le i, j \le k}$, such that $m_{i\,j}$ is a non negative integer for all $1 \le i, j \le k$, $m_{i\,j} = 0$ for all $1 \le j < i \le k$ and $m_{i\,i} = 1$ for $1 \le i \le k$. Let the set of such triangular matrices is denoted by $M_k$.

An ordered alphabet is an alphabet $\sum = \{a_1, a_2, \ldots\ldots, a_k\}$ with a relation of order $<$ on it. The alphabet can be written as $\sum = \{a_1 < a_2 < \ldots\ldots < a_k\}$.

### *1.2.3.2 Parikh matrix mapping*

Let $\sum = \{a_1 < a_2 < \ldots\ldots < a_k\}$ be an alphabet, where $k \ge 1$. The Parikh matrix mapping denoted by $\psi_{M_k}$ is the morphism:

$$\psi_{M_k} : \sum^* \to M_{k+1},$$

defined by the condition: if $\psi_{M_k}(a_q) = (m_{i,j})_{1 \le i,\, j \le (k+1)}$, then for each

$1 \leq i \leq (k+1)$, $m_{i,i} = 1$, $m_{q,q+1} = 1$ and all other elements of the matrix $\psi_{M_k}(a_q)$ are 0.

**Example 1.1:** Let $\Sigma$ be an ordered alphabet $\{a < b < c\}$. Then the Parikh matrix mapping $\psi_{\Sigma,3}$ represent each word over $\Sigma^*$ as a $4 \times 4$ upper triangular matrix with unit diagonal with non negative integral entries. As an example, we can cite: $\psi_{M_3}(ab^2) = \psi_{M_3}(a)\psi_{M_3}(b)\psi_{M_3}(b)$.

Thus

$$\psi_{M_3}(ab^2) = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 2 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The main property of the Parikh matrix mapping can be expressed in the form of following theorem:

**Theorem 1.2 :** Let $\Sigma = \{a_1 < a_2 < ...... < a_k\}$ be an alphabet, where $k \geq 1$ and assume that $w \in \Sigma^*$. The matrix $\psi_{\Sigma,k}(w) = (m_{i\,j})_{1 \leq i,\, j \leq (k+1)}$ has the following properties:

$$1.\ m_{i\,j} = 0, \ \ \textit{for all}\ 1 \leq j < i \leq (k+1),$$
$$2.\ m_{i\,i} = 1, \ \ \textit{for all}\ 1 \leq i \leq (k+1)$$
$$3.\ m_{i\ j+1} = |w|_{a_{i\,j}}, \ \textit{for all}\ 1 \leq i \leq j \leq k.$$

where $a_{i\,j}$ denotes the word $a_i a_{i+1} .........a_j$.

Parikh matrix mapping itself is not an injective mapping. The aspect of injectivity in Parikh matrix mapping has been discussed by Atanasiu et al (2001). In the present study the focus is on the construction of a word matrix which could have injectivity as an inbuilt facility and could generate Parikh vector.

### *1.2.3.3 Index enumerator of scattered subwords*

Let us consider a collection of polynomials $P_{w,a_{ij}}(q)$ indexed by $w \in \sum^*$ and $a_{ij} = a_i a_{i+1} .........a_j$ with $1 \le i \le j \le k$. These polynomials "q-count" the numbers $|w|_{a_{ij}}$ for general $u$ and $v$. To construct $P_{w,a_{ij}}(q)$, we consider a factorization

$$w = u_i a_i u_{i+1} a_{i+1} ......u_j a_j u_{j+1}$$

………………………..(1.1)

with $u_s \in \sum^*$ for $i \le s \le j+1$.

The numbers

$$I_i = |u_i| + 1,\ I_{i+1} = |u_i| + |u_{i+1}| + 2,\ I_j = |u_i| + |u_{i+1}| + ......... + |u_j| + j - i + 1 \text{…… (1.2)}$$

are simply the indices (positions) in $w$ where the letters $a_i, a_{i+1}, .........., a_j$ appear respectively, in the factorization 1.1. To construct the polynomial $P_{w,a_{ij}}(q)$, a monomial $q^{I_i + I_{i+1} + ......+ I_j}$ for each such factorization can be formed in $N[q]$, where $N[q]$ is the collection of polynomials in the variable q with coefficients from $N$ and all these monomials can be added.

$$P_{w,a_{ij}}(q) = \sum_{w = u_i a_i u_{i+1} a_{i+1} ......u_j a_j u_{j+1}} q^{I_i + I_{i+1} + ......+ I_j} \quad \text{…………………. (1.3)}$$

**Example 1.3:** Suppose $\sum = \{a < b < c\}$ and $w = ab^2ac^2a$. If $i = 1, j = 3$, then $a_{ij} = abc$ .Writing the indices and underlining the positions of $a,b$ and $c$ for each appearance of $abc$ as a scattered subword of $w$, we obtain

|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | *Positions* | *Contribution* |
|---|---|---|---|---|---|---|---|---|---|
| *Factorization* : | $\underline{a}$ | $\underline{b}$ | $b$ | $a$ | $\underline{c}$ | $c$ | $a$ | 1, 2, 5 | $q^8$ |
| *Factorization* : | $\underline{a}$ | $\underline{b}$ | $b$ | $a$ | $c$ | $\underline{c}$ | $a$ | 1, 2, 6 | $q^9$ |
| *Factorization* : | $\underline{a}$ | $b$ | $\underline{b}$ | $a$ | $\underline{c}$ | $c$ | $a$ | 1, 3, 5 | $q^9$ |
| *Factorization* : | $\underline{a}$ | $b$ | $\underline{b}$ | $a$ | $c$ | $\underline{c}$ | $a$ | 1, 3, 6 | $q^{10}$ |

Consequently $$P_{w,abc}(q) = q^8 + 2q^9 + q^{10}.$$

Similarly, we can find

$$P_{w,a}(q) = q + q^4 + q^7 \qquad P_{w,ab}(q) = q^3 + q^4$$

$$P_{w,b}(q) = q^2 + q^3 \qquad P_{w,ac}(q) = q^7 + q^8$$

$$P_{w,c}(q) = q^5 + q^6 \qquad P_{w,bc}(q) = q^7 + 2q^8 + q^9$$

For $q = 1$, it can be observed that $P_{w,a_{i\,j}}(q) = |w|_{a_{i\,j}}$. Thus the polynomial $P_{w,a_{i\,j}}(q)$ “q-count” the number of occurrences of $a_i a_{i+1}.........a_j$ as a scattered subword of $w$.

### *1.2.3.4 The Parikh q-matrix encoding*

Let the collection of k-dimensional upper triangular matrices with entries in $N[q]$ is denoted by $M_k(q)$. Let $I_k$ denote the identity matrix of dimension $k$. A mapping $\overline{\psi} : \Sigma \times N \to M_{k+1}(q)$ is defined as follows:

The matrix $\overline{\psi}(a_l, j)$ corresponding to a pair $a_l \in \Sigma = \{a_1 < a_2 < .... < a_k\}$ and $j \in N$, is defined as the matrix obtained from $I_{k+1}$ by changing the $(l, l+1)^{st}$ entry in $I_{k+1}$ to $q^j$. Thus if $\overline{\psi}(a_l, j) = (m_{i\,j})_{1 \le i,\, j \le k+1}$, then $m_{i,i} = 1$ for $1 \le i \le k+1, m_{l,l+1} = q^j$ and all other entries of the matrix $\overline{\psi}(a_l, j)$ are zero.

**Example 1.4:** Let $\Sigma = \{a < b < c\}$, then

$$\overline{\psi}(a,j) = \begin{bmatrix} 1 & q^j & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \ \overline{\psi}(b,j) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & q^j & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \ \overline{\psi}(a,j) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & q^j \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Parikh q-matrix encoding is defined as the mapping $\Sigma^* \to M_{k+1}(q)$ by setting $\overline{\psi}(\lambda) = I_{k+1}$ and

$$\overline{\psi}(w_1 w_2 .......w_n) = \overline{\psi}(w_1, 1)\overline{\psi}(w_2, 2).....\overline{\psi}(w_n, n)$$

**Example 1.5:** Let $\Sigma = \{a < b < c\}$, then $\overline{\psi}(ab^2) = \overline{\psi}(a,1)\overline{\psi}(b,2)\overline{\psi}(b,3)$. Thus

$$\overline{\psi}(ab^2) = \begin{bmatrix} 1 & q & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & q^2 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & q^3 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & q & q^3 + q^4 & 0 \\ 0 & 1 & q^2 + q^3 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

## 1.3 Mathematical and Statistical definitions/tools used in present study

Mathematical and Statistical tools and models have shown remarkable presence in getting the solution of many real world linguistics and natural language processing problems. This section contains compendium of those Mathematical and Statistical tools/ definitions, which have been utilized in the present study.

### 1.3.1 Random Variable:

A random variable is a function $f : \Omega \to R$, where $\Omega$ is the sample space for outcomes of a certain event and $R$ is the set of real numbers. If a random variable takes at most a countable number of values, it is called a discrete random variable.

### 1.3.2 Expectation or Mean:

The expectation is the mean or average of a random variable. If $X$ is a discrete random variable with probability mass function $p(x)$, then the expectation of $X$ is defined as:

$$E(x) = \sum_{i=1}^{n} p_i x_i, \qquad \sum_{i=1}^{n} p_i = 1 \qquad \text{......................(1.4)}$$

### 1.3.3 Empirical Estimates of variability

Let $X$ be a random variable representing the number of occurrence of a word in a document. Following are some measures to estimates variability of a random variable. Inverse document frequency (IDF), Burstiness and adaptation are generally used for information retrieval purposes.

#### *1.3.3.1 Variance*

The variance of a random variable is a measure of whether the values of random variable tend to be consistent over trials or to vary a lot. Variance is defined as follows:

$$Var(X) = E(X - \overline{X})^2 = \frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i} \qquad \text{.............(1.5)}$$

#### *1.3.3.2 Inverse document frequency*

In information retrieval, inverse document frequency is a useful measure to discriminate between function and content words. Inverse document frequency is defined as follows:

$$Idf = \log\left(\frac{N}{df_i}\right), \qquad \text{........................(1.6)}$$

where $N =$ Number of documents in the collection of text.

$df_i$ = Number of documents in which the $i^{th}$ term appears.

#### *1.3.3.3 Adaptation*

Adaptation shows the chances of appearing a word in more documents when we have already observed one occurrence of the word in the text. Adaptation is defined as follows:

$$\begin{aligned} Adp &= P\left(X \geq 2 \middle/ X \geq 1\right) \\ &= \frac{P(X \geq 2)}{P(X \geq 1)} \end{aligned} \qquad \text{..............................(1.7)}$$

#### *1.3.3.4 Burstiness*

Burstiness shows the average frequency of a term in the documents in which the word appears at least once. Burstiness is defined as follows:

$$B = \frac{\overline{X}}{P(X \geq 1)} = \frac{Tf}{Df} \qquad \text{............................(1.8)}$$

where $Tf$ is the total number of appearance of a word in a text.

#### *1.3.3.5 Entropy*

Let $P(x)$ be the probability mass function of a random variable $X$, over a discrete set of symbols $X$. The entropy (or self information) is the average uncertainty of a single random variable:

$$H(X) = -\sum_{x \in X} P(x) \log_2 P(x) \qquad \text{..............................(1.9)}$$

### 1.3.4 Mutual Information

Mutual information is defined in information theory. It is used to estimate the dependency/independency between two random variables. Mutual information between two random variable sometimes known as

Pointwise mutual information is defined as follows:

$$I(x,y)=\log_2\frac{P(x,y)}{P(x).P(y)}, \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1.10)$$

where $P(x,y)$ is the probability of happening of two events simultaneously. Whenever $P(x,y)=P(x).P(y)$, that is, the two events are independent, $I(x,y)=0$, which indicates that the two events are independent.

### 1.3.5 Coefficient of Correlation:

Karl Pearson's correlation coefficient between two random variables $X$ and $Y$ is defined as follows:

$$r(X,Y)=\frac{Cov(X,Y)}{\sigma_x\sigma_y}, \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1.11)$$

where $Cov(X,Y)$ is the covariance between two random variables which is defined as:

$$\begin{aligned}Cov(X,Y) &= E(X-\overline{X})(Y-\overline{Y})\\ &= \frac{1}{n}\sum(x_i-\overline{x})(y_i-\overline{y})\end{aligned}$$

$\sigma_x$ and $\sigma_y$ are standard deviations of $X$ and $Y$ respectively defined as:

$$\sigma_x=\sqrt{Var(X)}\ ,\ \sigma_y=\sqrt{Var(Y)}\,.$$

### 1.3.6 Rank Correlation coefficient:

Rank correlation coefficient is used to estimates the correlation between the ranks assigned by two characteristics when a group of n individuals is arranged in order of merit or proficiency in possession of two characteristics A and B. Let $(x_i, y_i);\, i=1,2,3,........n$ be the ranks of $i^{th}$ individual in two characteristics A and B respectively. Personian coefficient of correlation between the ranks $x_i$'s and $y_i$'s known as rank correlation coefficient between A and B is defined as follows:

$$r = 1 - \frac{6\sum_{i=1}^{n} d_i^{\,2}}{n(n^2-1)} \qquad \text{..................................(1.12)}$$

where $d_i = x_i - y_i$.

For repeated ranks, common ranks are given to the repeated items. This common rank is the average of the ranks which these items would have assumed if they were slightly different from each other and next item will get the rank next to the ranks already assumed. For repeated ranks rank correlation coefficient is defined as :

$$r = 1 - \frac{6\sum_{i=1}^{n} d_i^{\,2}}{n(n^2-1)} + \sum \frac{m_i(m_i^{\,2}-1)}{12},\ i = 1,2,3,.....,$$ where $m_i$ is the number of times $i^{th}$ item is repeated.

### 1.3.7 Test of significance:

Test of significance is an important aspect of sampling theory, which enables us to decide on the basis of the sample results, if

(i) The deviation between the observed sample statistics and the hypothetical parameter value, or

(ii) The deviation between two sample statistics,

is significant or might be attributed to chance or the fluctuations of sampling.

### 1.3.8 Null hypothesis:

For applying the test of significance we first set up a hypothesis, a definite statement about the population parameter. Such a hypothesis, which is usually a hypothesis of no difference, is called Null hypothesis and is usually denoted by $H_0$.

### 1.3.9 Alternative hypothesis

A hypothesis which is complementary to the null hypothesis is called an alternative hypothesis, usually denoted by $H_1$.

### 1.3.10 The Z statistic:

For large values of n, the number of trials, almost all the distributions, e.g., Binomial, Poisson, Negative Binomial etc are very closely approximated by normal distribution. For a random variable $X \sim N(\mu, \sigma^2)$ the Z statistic is defined as:

$$Z = \frac{X-\mu}{\sigma} = \frac{X-E(X)}{\sqrt{Var(X)}} \sim N(0, 1) \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(1.13)$$

The critical value of $|Z|$ is 1.96 at 5% level of significance and 2.58 for 1% level of significance. On choosing a particular level of significance, we can decide whether the null hypothesis is accepted or rejected on the basis of the obtained value of $|Z|$.

### 1.3.11 Chi square test:

If $O_{i,}(i = 1, 2, \ldots\ldots\ldots n)$ is a set of observed (experimental) frequencies and $E_{i,}(i = 1, 2, \ldots\ldots\ldots n)$ is the corresponding set of expected (theoretical) frequencies, then Karl Pearson's Chi-square given by

$$\chi^2 = \sum_{i=1}^{n} \left[ \frac{(O_i - E_i)^2}{E_i} \right], \qquad \left( \sum_{i=1}^{n} O_i = \sum_{i=1}^{n} E_i \right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.14)$$

follows chi square distribution with $(n-1)$ degree of freedom.

On choosing a particular level of significance, from the tabulated value of $\chi^2$ at a particular degree of freedom, we can decide whether the null hypothesis is accepted or rejected on the basis of the obtained value of. $\chi^2$.

### 1.3.12 The t-test:

The t-statistic is used when sample size is small. Let $x\ x_i, (i = 1, 2, ...n)$ be a random sample of size n from a normal population with mean $\mu$ and variance $\sigma^2$, then the t-statistic is defined as

$$t = \frac{\overline{x} - \mu}{S / \sqrt{n}} \qquad \text{.........................(1.15)}$$

where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$ is an unbiased estimate of the population variance $\sigma^2$.

### 1.3.13 Likelihood ratio test:

Likelihood ratio test is more interpretable than $\chi^2$- test in the sense that it tells us how much more likely one hypothesis is than the other. In this, we test the null hypothesis $H_0$ against the alternative hypothesis $H_1$. Maximum likelihood estimates are calculated for each hypothesis and the log of the likelihood ratio $\lambda$ is calculated as follows:

$$\log \lambda = \log \frac{L(H_0)}{L(H_1)} \qquad \text{.................................. (1.16)}$$

It was shown by Mood, Graybill & Boes (1974, pp. 440) that the quantity $-2\log \lambda$ has an asymptotically $\chi^2$- distribution with r degrees of freedom, where r is the number of parameters under comparison. If the calculated values of $-2\log \lambda$ is less than the tabulated value of $\chi^2$ at given level of significance, we accept the null hypothesis $H_0$, otherwise the hypothesis $H_1$ is accepted.

### 1.3.14 Fuzzy set

Fuzzy Set Theory was formalized by Professor Lofti Zadeh at the University of California in 1965. Classical or crisp sets are those sets which deal with exact properties of elements. In classical or crisp set, an element of universal set either belongs to the set or does not belong to the set. In a classical or crisp set we assign only two values 0 or 1 to different elements depend on their belongingness to the set. If an element is a member of the set then it is 1 otherwise 0. This approach is well defined for exact properties, such as the set of positive real numbers on a set of real numbers. We may assign 1 for every positive real number and 0 for every non positive real number.

The classical approach does not work well for linguistic terms such as good student, hot temperature etc. Consider a classical set of good students and assume that a student is good if he get equal to or more than 70% marks. Then, a student having 69.9% marks will not be included in the set; however who is very close to be considered for good student. Consider a classical set of tall people and suppose that the person with height greater than or equal to 6 feet is considered as tall. Then, a person having height 6 feet and 1 inch will be included in the set of tall people and the person having height 5 feet 11 inches will not be included in the set. In this case such a representation of reality leaves much to be desired. On the other hand, using the fuzzy logic, the person being 6-1 tall can still have a full membership of the set of tall people, but the person that is 5-11 tall, can have 90% membership of the set. The 5-11 people thus can be described as a "quite tall" representation in a model. Hence Fuzzy sets are very useful in these situations.

A Fuzzy set is a set in which all the elements of a universal set are included with a grade of membership which shows their degree of belongingness to the set. Instead of assigning 0 or 1 we use the closed interval [0, 1]. Zero grade of membership indicates that the element does not belong to the set and one grade of membership gives full support to the element for its belongingness to the set.

### 1.3.15 Fuzzy Logic

As the name suggests, the fuzzy logic is a logic, which deals with the values, which are approximate rather than exact. The classical logic relies on something which is either true or false. A True element is usually assigned a value of 1 and false has a value 0. Thus, something either completely belongs to a set or it is completely excluded from the set. The fuzzy logic broadens this definition of classical logic. The basis of the logic is fuzzy sets. Fuzzy logic allows a much easier application of many problems that cannot be easily implemented using classical approach. The importance of fuzzy logic derives from the fact that most modes of human reasoning and especially common sense reasoning are approximate in nature. The fundamental difference between classical propositions and fuzzy propositions is in the range of their truth values. In case of fuzzy propositions the truth or falsity is a matter of degree. Assuming that truth and falsity are expressed by values 1 and 0, respectively, the degree of truth of each fuzzy proposition is expressed by a number in the unit interval [0, 1].

Lofti Zadeh described the essential characteristics of fuzzy logic as follows.

- In fuzzy logic, exact reasoning is viewed as a limiting case of approximate reasoning.
- In fuzzy logic everything is a matter of degree.
- Any logical system can be fuzzified.
- In fuzzy logic, knowledge is interpreted as a collection of elastic or, equivalently , fuzzy constraint on a collection of variables
- Inference is viewed as a process of propagation of elastic constraints.

### 1.3.16 Fuzzy inference system

Fuzzy inference system provides the facility to incorporate all the auxiliary information to draw the conclusions. Matlab Fuzzy logic toolbox provides an opportunity to look at all the components of Fuzzy inference system. Fuzzy inference system consists a baseline model, membership functions for different fuzzy sets, fuzzy rules and outcome of a problem.

#### *1.3.16 .1 Baseline Model:*

To draw any inferences from the Fuzzy inference system, the first step is to define the baseline model. For this one needs to choose the baseline model that is, input variables, output variables, implication method, aggregation method and defuzzification method. We have to define the fuzzy sets for all the input variables and also for the final or output variable. Input variables are those variables from which one has to draw the inferences i.e. the input variables contains some auxiliary information and the output variable is the variable, which defines the final grade of membership for all the elements of the set. To define the fuzzy sets for any input variable, first, we have to choose the range and the membership function for that input variable. Range of the input variable can be defined by taking the minimum and the maximum value of the input variable. There exist many membership functions in fuzzy inference system. The membership function for any input variable can be chosen according to the property of that input variable. These membership functions appear as a graph in fuzzy inference system. After choosing the membership function for the input variable, we define the different fuzzy sets for the input variable, using the selected membership function. An example of the model (consisting of fuzzy sets and range for the input variable) of input variable is given in figure 2. The membership function for different input variables can be different. Actually the membership function depends upon the characteristics of the input variable; hence it may vary from one input variable to another input variable. After defining the range and the fuzzy sets for all the input variables, we define the range and the fuzzy set for the output variable.

### *1.3.16.2 Fuzzy Rules*

Construction of rules in fuzzy inference system is an important step. Rules can be derived from a common knowledge about required inference procedure. In the real life situations, all the human being makes the decision. These decisions are based on rules. For Example: if the weather is fine and today is holiday, then we may decide to go out or if the forecast says that the weather will be bad today, but fine tomorrow, then we make a decision not to go today, and postpone it till tomorrow. Similarly, in fuzzy inference system, we have to define certain rules in order to draw some conclusion from input variables. These rules are based on common sense. The fuzzy rules, are formulated using a series of if-then statements, combined with AND/OR operators. These rules are very useful to find out the final decision.

### 1.3.17 Free Group

An alphabet $\Sigma$ is a set of symbols called letters. A word over $\Sigma$ is a finite sequence of elements of $\Sigma$. Let$\Sigma = \{a,b\}$, then $ab, abb, aab$ etc are words over the alphabet$\Sigma$. By a syllable we mean a symbol$a^n$, where $a$ is a letter of the alphabet $\Sigma$ and $n$ is an integer. Words are also defined as finite ordered sequence of syllables as given in the book of Crowell et al (1997). An empty word is a neutral element denoted by $\lambda$ or 1. The set of all the words over $\Sigma$ is denoted by $\Sigma^*$. In case if $u = w_1 a^0 w_2$ and $v = w_1 w_2$, where $w_1$ and $w_2$are words over $\Sigma$, the word $v$ is obtained from the word $u$ by an elementary contraction of type I or the word $u$ is obtained from the word $v$ by an elementary expansion of type I. When $u = w_1 a^p a^q w_2$ and $v = w_1 a^{p+q} w_2$, where $w_1$ and $w_2$ are words over $\Sigma$, then $v$ is obtained from the word $u$ by an elementary contraction of type II or the word $u$ is obtained from the word $v$ by an elementary expansion of type II.

The juxtaposition or concatenation of two words $w_1$ and $w_2$ is a word $w_1 w_2$ obtained by concatenating the two words $w_1$ and $w_2$. For example if $w_1 = ab$ and $w_2 = aab$, then juxtaposition of the two words $w_1$ and $w_2$ is expressed as $w_1 w_2 = abaab$. The set of all words over alphabet $\Sigma$ with a binary operation 'juxtaposition' and neutral element 1 is a free monoid as given in Lothaire M. (1997). The inverse $w^{-1}$ of a word $w$ is obtained by reversing the order of its syllables and changing the sign of the exponent of each syllable such that $ww^{-1} = 1$. On including the inverses of each element, the monoid $\Sigma^*$ becomes a group known as free group on the alphabet $\Sigma$.

### 1.3.18 Metric Space

Let $X \neq \phi$ be any given space. Let $x, y, z \in X$ be arbitrary. A function $d : X \times X \to R$ is said to be distance function if it has the following properties:

1. $d(x, y) \geq 0$
2. $d(x, y) = d(y, x)$
3. $d(x, y) + d(y, x) \geq d(x, z)$
4. $d(x, y) = 0$ *iff* $x = y$

The non empty set $X$ with a metric $d$ is called a metric space $(X, d)$.

## 1.4. Review of Literature

The first source of literature under the category of reference books in natural language processing for me, has been the book of C. D. Manning and H. Schütze (2002), which provides a detail discussion on different NLP problems and the application of statistical techniques in these problems. Other books in this field are the book of Jurafsky Daniel, Martin James H. (2004) and R. Dale, H. Moisl, and H. Somers (2002). The book of M. Lothaire (1997) provides a detail discussion on words and algebraic properties of words. The formation of free group for words can be studied through the book of R. H. Crowell and R.H.Fox (1997). Klir, George J. and Yuan Bo (2001) have provided a detail knowledge of fuzzy sets, fuzzy logic and their application. Marie (2006) in his book has discussed some algorithms in information extraction in the context of information retrieval.

With an intention of introducing the mathematical foundations of linguistics to computer scientists, engineers and mathematicians in natural language processing, Kornai (2008) in his book entitled "Mathematical Linguistics" has covered an extremely rich array of topics including not only syntax and semantics but also phonology and morphology, probabilistic approaches, complexity, learnability and the analysis of speech and handwriting. Sandor Dominich (2008) has given a detail discussion of different mathematical tools which are basis for information retrieval. This book takes a unique approach to information retrieval by laying down the foundations for a modern algebra of information retrieval based on lattice theory with an aim of demonstrating the advantage of this method in formulation of new retrieval methods. Term weighting, vector space model and other features of information retrieval have been discussed in the book of Manning et al (2008).

Looking at the works in the direction of term weighting, the works of Spark Jones (1972, 1998) can be cited as one of the earliest works devoted to the importance of terms and its application to information retrieval. Bookstein and Wanson (1974, 1975) have suggested some models for automatic indexing. The papers of Kenneth W. Church and William A. Gale (1991) and Slava M. Katz (1996) are important works in the arena of distribution of words in a text. Chisholm, E. and Kolda Tamara G. (1999) have discussed term weighting schemes in vector space model. Losee M. Robert (2001) has provided the basis of Luhn and Zipf's models. Zanette et al (2005) have investigated the origin of Zipf's law for words in written texts by means of a stochastic dynamic model for text generation. Patricia Guilpin and Christian Guilpin (2005) have developed a new method of linguistic and statistical analysis of the frequency of a particular word at different times or in different styles. An important problem of Identification of language from a given small piece of text is analyzed by Murthy et al (2006).

Reviewing literature for collocations and methods of collocation extraction, a lot of papers can be cited. As knowledge regarding collocations is very important for many applications of natural processing tasks, a lot of work in this direction has been done so far. The first effort for collocation extraction was the use of frequency measure utilized by Choueka, Klein and Neuwitz (1983) to identify a particular type of collocations. The utilization of mutual information for collocation extraction was suggested by Church and Hanks (1989). Smadja (1993) has defined a multi stage process for collocation extraction. Church and Gale used the $\chi^2$ - test for the identification of translation pairs in aligned corpora. The application of likelihood ratio test to collocation discovery was suggested by Dunning (1993). Johansson (1996) has presented his paper on bigrams at the 16th International Conference on Computational Linguistics. The use of frequency counts of the collocational relations extracted from the corpus for information extraction was investigated by Dekang Lin (1998). Chen, K. Kishida, H. Jiang, and Q Liang (1999) suggested a two stage method for the extraction of domain specific collocations. The use of Collocation in finding word similarity was suggested by Dekang Lin (1999). Kathleen R. McKeown and Dragomir R. Radev (2000) have discussed in detail about collocations and different methods of collocation extraction. Darren Pearce et al (2001) described a collocation extraction technique based on the restrictions on the possible substitutions for synonyms within candidate phrases and also utilized WordNet in the technique (2001).

Some other works related to collocation are: automatic retrieval of frequent idiomatic and collocational expressions in a large corpus by Y. Choueka, S.T. Klein, and E. Neuwitz (1983), noun classification from predicate-argument structures by Hindle (1990), automatically extracting and representing collocations for language generation by F.A. Smadja and K.R. McKeown (1990), a method for disambiguating word senses in a large corpus by Gale, W. A., Church, K. W. and Yarowsky, D. (1992), a methodology for automatic term recognition by Sophia Ananiadou (1994), parsing, word associations and typical predicate-argument relations by K. Church, W. Gale, P. Hanks, D. Hindle and Moon (1994), identifying and translation technical terminology by Dagan and K. Church (1994), a comparative study of automatic extraction of collocations from corpora by K. Kita, Y. Kato, T. Omoto, and Y. Yano (1994), technical terminology: some linguistic properties and an algorithm for identification in text by John S. Justeson and Slava M. Katz (1995), translating collocations for bilingual lexicons by Smadja, K.R.

McKeown, and V. Hatzivassiloglou (1996), learning bilingual collocations by word-level sorting by Haruno, S. Ikehara and T. Yamazaki (1996), class phrase models for language modeling by K. Ries, F.D. Buo, and A. Waibel (1996), collocations in language learning by K. Kita and H. Ogata (1997), a statistical analysis of morphemes in Japanese terminology by Kyo Kageura (1997), searching corpora for dictionary phrases by Debra S. Baddorf and Martha W. Evens (1998), automatic word sense discrimination by Schütze, H. (1998), empirical methods for MT lexicon development by I.D. Melamed (1998), similarity-based models of word co-occurrence probabilities by Dagan, L. Lee and F. Pereira (1999), automatic identification of non-compositional phrases by Dekang Lin (1999), identifying contextual information for multi-word term extraction by D. Maynard and S. Ananiadou (1999), combining linguistics with statistics for multiword term extraction by Dias, S. Guilloré, J-C. Bassano, and J.G. Pereira Lopes (2000), knowledge-lite extraction of multi-word units with language filters and entropy thresholds by M. Merkel and M. Andersson (2000), methods for the qualitative evaluation of lexical association measures by Evert Stefan and Krenn Brigitte (2001), a case study on extracting PP-verb collocations by Krenn, Brigitte and Evert, Stefan (2001), extracting morpheme pairs from bilingual terminological corpora by Keita Tsuji and Kyo Kageura (2001), collocations and lexical functions by Igor Mel'cuk, extraction of Chinese compound words - an experimental study on a very large corpus by J. Zhang, J. Gao, and M. Zhou (2008).

The work of Parikh (1966) was itself the first kind of work in the direction of finding numerical properties of words. Later on Mateescu A., Salomaa A., Salomaa K. and Sheng Yu (2001) extended this approach by defining Parikh matrix. Atanasiu, Martin-Vide and Mateescu (2001) have discussed codifiable languages in connection with Parikh matrix mapping. Mateescu A., Salomaa A., Salomaa K and Sheng Yu (2002) have discussed subword histories and Parikh matrices. Parikh matrix was generalized by T.-F Serbanuta (2004). In order to make injective Parikh matrix mapping, Egecioglu (2004) has defined a q-matrix encoding of Parikh matrix mapping. Salomaa A. (2004) studied the connections between subwords and certain matrix mappings. Salomaa A., Sheng Yu (2004) have defined relation between subword conditions and subword histories.

## 1.5 Structural outline of the thesis

After presenting historical survey and an outline of the programme in the customary manner in this chapter, I have discussed the properties of content words in second chapter. In this chapter some measures used for global weighting have been examined for local weights. Two texts have been chosen, one as a training data for the sake of experiment and the other for the verification of the results. Ranks have been assigned to non function words based on their frequencies. It was found that Zipf's law is not a good descriptor of the relationship between frequency and rank for non function words. In order to describe this relationship in a better way, some modifications have been implemented in Zipf's law. Finally, weight of a term in a document has been defined using the standard Z score of these measures. The experiment shows that the weighting scheme is quite successful in discriminating content words from non content words in a document by assigning a proper weight to words. This chapter in the form of a research paper has appeared in the proceedings of the 6th international conference of information science technology and management at IIT Delhi.

Third chapter deals with the evaluation of different statistical techniques of collocation extraction like frequency measure, Mutual information score, t-test, Likelihood ratio test, Chi square test etc. All these techniques have been applied to a compiled corpus of about one million words and results have been analyzed. A probability measure has been suggested before applying the above tests to filter out unnecessary free word combinations. This part of chapter third has appeared in the Journal of Quantitative linguistics, a research Journal of Taylor Francis Group. In the remaining part of this chapter, a new collocation extraction technique has been suggested based on the proposed probability measure and Z statistic and this part is likely to appear in the proceedings of 12th annual conference of Gwalior Academy of Mathematical Science at MANIT Bhopal.

In Chapter 4, we have utilized fuzzy approach to collocation extraction. We have defined two different fuzzy sets of word combinations. The collocations have been extracted on the basis of high mutual information and significance t-score. The generated fuzzy sets have been utilized in the construction of fuzzy set theoretic models. Fuzzy inference system of Matlab has been also utilized for collocation extraction.

The fifth chapter is concerned with matrix representation of words. A word matrix mapping has been defined by defining juxtaposition and elementary contractions for matrices of letters of a word. The definitions paving the way for word matrix have been illustrated by pertinent examples. Subword occurrence has been calculated using the word matrix and some algebraic properties of word matrices have been proved.

The concept of word matrix mapping has been extended in sixth chapter. An isomorphic mapping has been defined from the word matrix mapping to q-matrix encoding which provides the q-index generator of a word. A distance function has been defined to examine the similarity of words and metric space has been generated for word matrices. Distance polynomial is defined and has been utilized for study of morphology in natural languages.

# Chapter II

## On Some Properties of Content Words in a Document

*Content words have a great role in an information retrieval system. Generally title of a document is considered for an information retrieval system but every time title does not include all content words. In this chapter an effort has been made in the direction of finding some properties of contents words in a document so that automatic generation of content words can be made possible. Zipf's law has been also rechecked for non-function words. For experiment purpose two texts have been taken. First document "Mr. Bonaparte of Corsica" available at project www.projectgutenburg.org has been chosen for training data and second "Napoleon and His Marshals - Vol. I, chapter I" from http://www.napoleonic-literature.com/Book_15/V1C1.htm has been taken as candidate data to verify the results of the formulae proposed. Effect of length of paragraphs in different measures (like as Mean, Variance, Adaptation, Burstiness, Entropy etc) has been studied by taking paragraphs of different lengths. A weighting scheme has been defined for assigning weights to different words on the basis of standard Z scores.*

### 2.1 Introduction

Performance of an Information retrieval system depends on term weighting schemes. In the book of Manning et al (2002) a detail discussion is given on term weighting schemes. Weight of a term can be seen in two different aspects, local and global. Local weights are functions of frequency of a word in a document and global weights are functions of the frequency of a term in the entire collection (1999). Mean, variance, entropy, burstiness, adaptation are some weights which are used for global weighting. Church et al (1991) have shown that the variability is associated with content. Generally terms appear in the title or terms with high frequency are considered for getting information about the content of a document but these are not necessarily applicable for content words. For improving the quality of information retrieval it is required to find a set of representatives of a document. We have analyzed the behaviour of content words in a document based on the following statistical tools in the context of global weighting schemes.

### 2.1.1 Mean

The basic information used in term weighting is term frequency and document frequency. Term frequency is the number of times a term appears in a text and document frequency is the number of documents in which a term appears. Mean provides the information about the average appearance of the term in a text by utilizing term frequency and document frequency. Let $x_i$ be the random variable which represents the number of times a word appears in a document and $f_i$ be the corresponding frequency of documents.

$$Mean(X) = \frac{\sum f_i x_i}{\sum f_i} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(2.1)}$$

In addition to mean, measures of variability discussed in section 1.3.3 of chapter I, have been utilized for global weighting schemes.

### 2.1.2 Zipf's law

If we count up, how often each word type of a language occurs in a large corpus and then list the words in order of their frequency of occurrence, Zipf's law explores the relationship between the frequency of a word $f$ and its position in the list, known as its rank $r$. According to

Zipf's law $f\,\alpha\,\frac{1}{r}$ or there exists a constant $k$ such that $f.r=k$. The validity and possibilities for the derivation of Zipf's law has been studied extensively by Mandelbrot and has derived the following more general relationship between rank and frequency:

$$f=P(r+\rho)^{-B} \text{ or } \log f=\log P-B\log(r+\rho) \qquad \ldots\ldots\ldots\ldots\ldots\ldots..(2.2)$$

Here $P,B$ and $\rho$ are parameters of a text, that collectively measure the richness of the text.

## 2.2 Methodology

For the purpose of experiment, we have chosen the a text "Mr. Bonaparte of Corsica" Author: John Kendrick Bangs [eText # 3236] from www.gutenberg.org as training data. The whole text has been partitioned into paragraphs of same length (Number of words in a paragraph). Different Lengths of paragraph (100, 300, 500) have been taken to check the effect of length in different measures of variability and values of different measures have been calculated for each paragraph length. Paragraphs of lengths 100, 300 and 500 have been denoted by P1, P3 and P5 respectively. Correlation coefficients between the values of each measure for different paragraph lengths have been calculated. Some words have been chosen from the text in the frequency range 15-300. Simulation of Zipf's law for non function words has yielded a new relation between frequency and rank. The standardization process has been initiated in order to make same scale of each measure and this work has been accomplished by the calculation of standard Z score of each measure and then by defining a weighting scheme of words on the basis of the standard Z scores. The frequencies of different words for three different lengths of paragraph in the training document have been demonstrated in the table 2.1.

Table. 2.1 Words and their corresponding frequencies for three different paragraph lengths in first text.

| **Word** | **x** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Napoleon | f (P1) | 163 | 96 | 42 | 9 | 1 | 0 | 0 | 0 | 0 | 0 |
| | f(P2) | 19 | 29 | 20 | 13 | 11 | 9 | 2 | 0 | 0 | 0 |
| | f(P3) | 8 | 9 | 11 | 6 | 7 | 7 | 6 | 4 | 1 | 0 |
| Paris | f (P1) | 258 | 40 | 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | f(P2) | 57 | 32 | 9 | 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| | f(P3) | 21 | 24 | 11 | 4 | 2 | 0 | 0 | 0 | 0 | 0 |
| Emperor | f (P1) | 260 | 43 | 6 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| | f(P2) | 68 | 19 | 10 | 5 | 0 | 1 | 0 | 0 | 0 | 0 |
| | f(P3) | 32 | 16 | 4 | 7 | 1 | 1 | 1 | 0 | 0 | 0 |
| Joseph | f (P1) | 272 | 25 | 8 | 5 | 1 | 0 | 0 | 0 | 0 | 0 |
| | f(P3) | 76 | 17 | 2 | 2 | 2 | 1 | 1 | 2 | 0 | 0 |
| | f(P5) | 37 | 16 | 2 | 0 | 3 | 1 | 1 | 0 | 1 | 1 |
| France | f (P1) | 270 | 38 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | f(P3) | 72 | 20 | 9 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| | f(P5) | 34 | 17 | 8 | 2 | 0 | 1 | 0 | 0 | 0 | 0 |
| Here | f (P1) | 268 | 40 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | f(P3) | 68 | 27 | 7 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| | f(P5) | 31 | 21 | 7 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| Again | f (P1) | 289 | 21 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | f(P3) | 82 | 19 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | f(P5) | 43 | 15 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| First | f (P1) | 271 | 39 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| | f(P3) | 68 | 29 | 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| | f(P5) | 32 | 22 | 4 | 4 | 0 | 0 | 0 | 0 | 0 | 0 |

| **Word** | **x** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| French | f (P1) | 273 | 35 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | f(P3) | 72 | 22 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | f(P5) | 33 | 19 | 8 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| Make | f (P1) | 269 | 41 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | f(P3) | 67 | 30 | 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| | f(P5) | 29 | 24 | 8 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Italy | f (P1) | 291 | 17 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| | f(P3) | 89 | 7 | 6 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| | f(P5) | 49 | 8 | 3 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| History | f (P1) | 292 | 17 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| | f(P3) | 86 | 14 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| | f(P5) | 46 | 12 | 3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| Look | f (P1) | 293 | 17 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | f(P3) | 87 | 13 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | f(P5) | 45 | 15 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Military | f (P1) | 249 | 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | f(P3) | 88 | 13 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | f(P5) | 50 | 7 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Place | f (P1) | 292 | 17 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | f(P3) | 87 | 12 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| | f(P5) | 45 | 13 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Little | f (P1) | 279 | 29 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | f(P3) | 78 | 19 | 5 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| | f(P5) | 39 | 16 | 4 | 1 | 2 | 0 | 0 | 0 | 0 | 0 |
| Very | f (P1) | 286 | 23 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | f(P3) | 79 | 21 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | f(P5) | 40 | 17 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

For the verification of the proposed weighting scheme and defined relation between frequency and rank, a new document “Napoleon and His Marshals - Vol. I, chapter I” from http://www.napoleonic- literature.com/ Book_15/V1C1.htm has been chosen and it was found that the proposed measures are quite successful for any text.

## 2.3 Different measures of variability for a document

All measures of variability discussed in section 2.1 are quite useful for global weighting. For local weight we shall check the suitability of these measures. In this section, we have analyzed some of the measures of variability.

### 2.3.1 Appropriateness of Mean

Mean is directly proportional to the frequency of a word in a document and the hence frequency of a word in a document can be a good measure for providing weights to terms. Following table exhibits the correlation between mean values of different lengths of paragraphs. Since the correlation is prefect, which evinces the fact that mean is independent of length of paragraphs.

Table 2.2 Correlation between mean values of different paragraph lengths P1, P3, P5

| | Mean (P1) | Mean(P3) | Mean(P5) |
|---|---|---|---|
| Mean(P1) | | 1 | 1 |
| Mean(P3) | 1 | | 1 |
| Mean(P5) | 1 | 1 | |

Following Figure shows the graphical representation of variation in mean of seventeen words for three different lengths of paragraphs.

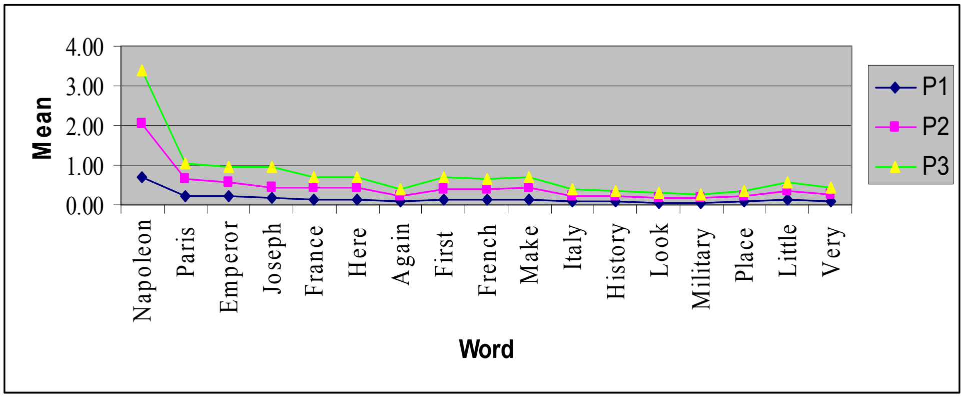

Fig. 2.1 Variation in mean for different lengths of paragraphs

### 2.3.2 Appropriateness of Variance

In any document we have function words as well as content words. The distribution of function words in a document is symmetrical as functions words are always accompanied by content words. High variability can be observed in the distribution of content words and it can provide a chance to look for content. Following table shows the correlation between variances for different paragraph lengths which envisages the fact that all variances are highly correlated and in turn signify that variance is independent of length of paragraphs.

Table 2.3 Correlation between variances of different paragraph lengths

| | Var(P1) | Var(P3) | Var(P5) |
|---|---|---|---|
| Var(P1) | | .95 | .93 |
| Var(P3) | .95 | | .98 |
| Var(P5) | .93 | .98 | |

Pictorial representation of variation in variances can be depicted by following figure.

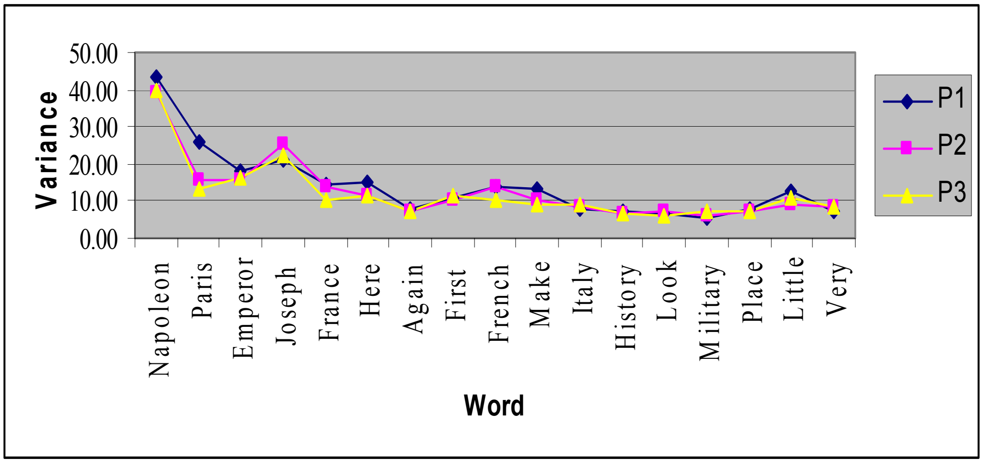

Fig. 2.2 Variation in variances for different lengths of paragraphs

### 2.3.3 Appropriateness of Inverse document frequency (Idf)

The formula for inverse document frequency

$$\log(N / df_i) = \log N - \log df_i,$$

gives full weight to words that occur in one document $(\log N - \log 1 = \log N)$ and a word that occurred in all documents would get zero weight $(\log N - \log N = 0)$ and hence very useful for global weighting. Inverse document frequency can also be used for analyzing the behavior of content words in a document. Function words appear almost in every paragraph, therefore for function words inverse document frequency will be zero. Since the local weight of a term depends on the frequency of a word in a text and a content word will have high frequency in comparison to non-content words, it can be inferred that a content word will appear in more paragraphs in comparison to non-content words and hence inverse document frequency will have lower value than non-content words.

The following table demonstrates that all Idfs are highly correlated and thus signify that inverse document frequency is independent of length of paragraphs.

Table 2.4 Correlation between Idf values of different paragraph lengths.

| | Idf (P1) | Idf (P3) | Idf (P5) |
|---|---|---|---|
| Idf (P1) | | .98 | .94 |
| Idf (P3) | .98 | | .98 |
| Idf (P5) | .94 | .98 | |

Following figure shows the pictorial representation of variation in variances of seventeen words for three different lengths of paragraphs.

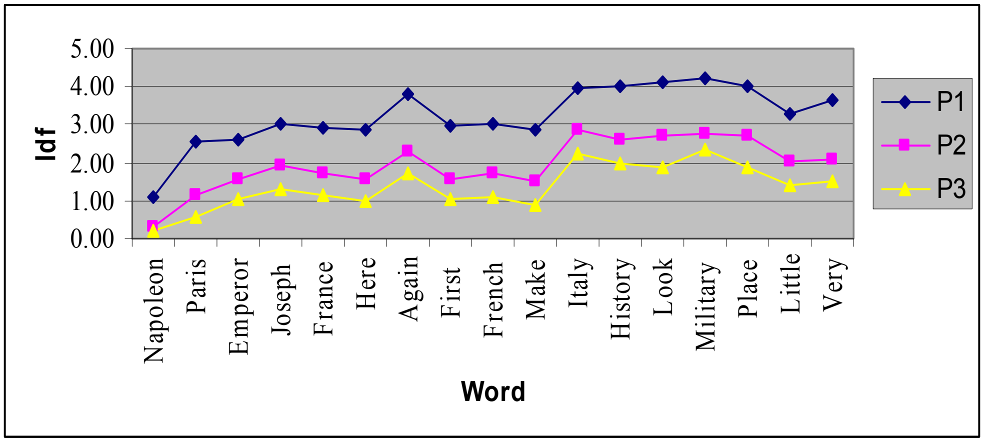

Fig. 2.3 Variation in Idf for different lengths of paragraphs

### 2.3.4 Appropriateness of Adaptation

Adaptation is the measurement of getting the information about the appearance of a word more than once, when it has already appeared in the document. Contents words are more likely to appear again if they emerge once in the text. Thus content words will have high value of adaptation in comparison to non-content words. Following table exhibits that correlation between adaptation values is not so good, which signifies that adaptation is affected by the change in the length of paragraphs. Therefore adaptation can not be a good measure for local weight.

Table 2.5 Correlation between Adaptation of different paragraph lengths.

| | Adp (P1) | Adp (P3) | Adp (P5) |
|---|---|---|---|
| Adp (P1) | | .75 | .55 |
| Adp (P3) | .75 | | .79 |
| Adp (P5) | .55 | .79 | |

Following figure shows the graphical representation of variation in adaptation values of seventeen words for three different lengths of paragraphs.

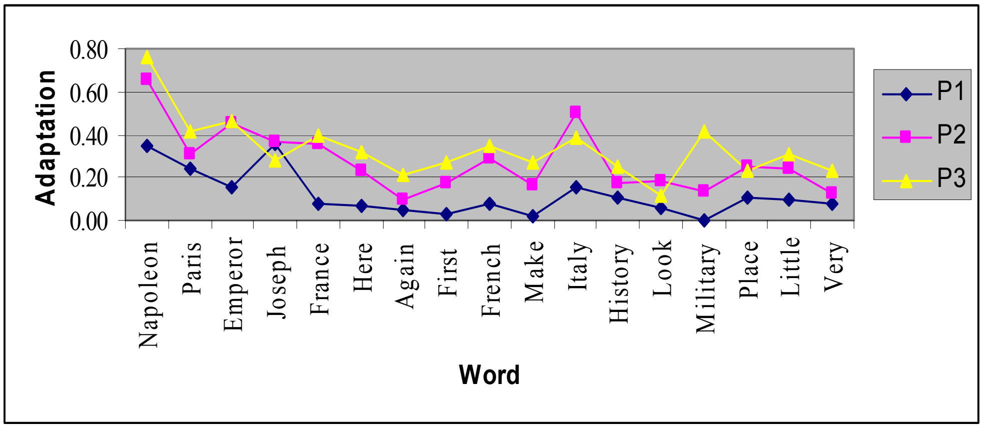

Fig. 2.4 Variation in Adaptation for different lengths of paragraphs

### 2.3.5 Appropriateness of Burstiness

Since content words have higher frequency than non-content words, burstiness will have higher value for content words in comparison to non-content words. Table 2.6 exhibits a good correlation between burstiness values for different paragraph lengths.

Table 2.6 Correlation between Burstiness of different paragraph lengths.

| | Br (P1) | Br (P3) | Br (P5) |
|---|---|---|---|
| Br (P1) | | .79 | .80 |
| Br (P3) | .79 | | .96 |
| Br (P5) | .80 | .96 | |

Fig. 2.5 shows the graphical representation of variation in burstiness values of seventeen words for three different lengths of paragraphs.

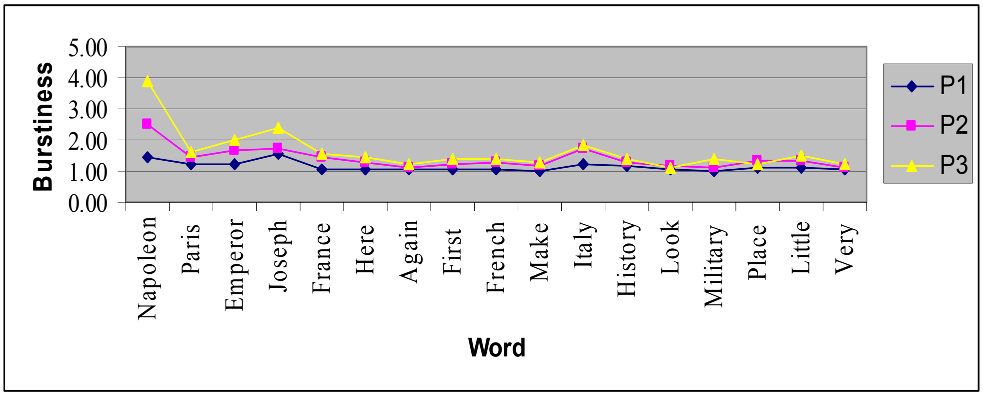

Fig.2.5 Variation in Burstiness for different lengths of paragraphs

### 2.3.6 Appropriateness of Entropy

Content word contains more information in itself, thus leads to high entropy. From table 2.7 it can be observed that all entropy values are almost perfect correlated and hence not affected by the change in the length of paragraphs.

Table 2.7 Correlation between Entropy of different paragraph lengths.

| | H (P1) | H (P3) | H (P5) |
|---|---|---|---|
| H (P1) | | .99 | .99 |
| H (P3) | .99 | | 1.00 |
| H (P5) | .99 | 1.00 | |

The graphical representation of variation in entropy of seventeen words for three different lengths of paragraphs has been demonstrated in Fig.2.6

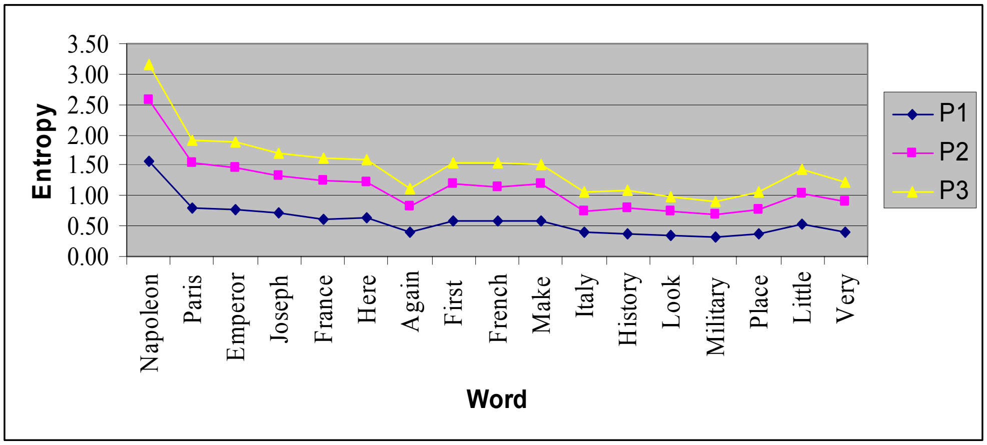

Fig. 2.6 Variation in Entropy for different lengths of paragraphs

Table 2.8 shows the values of different measures and their corresponding mean and standard deviations obtained from paragraph of length 300. Table 2.9 shows the correlation coefficient between different measures, which signifies that, all the measures are highly correlated with each other except inverse document frequency. Its application for local weights can be insignificant as we are not concerned about relevance of a particular paragraph in a document. Our main objective is to find the weight of a term in whole document.

Table 2.8 Values of different measures for different words in document I.

| Word | Mean | Br | Idf | Var | H |
|---|---|---|---|---|---|
| Napoleon | 2.03 | 2.49 | 0.29 | 39.27 | 2.56 |
| Paris | 0.64 | 1.43 | 1.16 | 15.54 | 1.55 |
| Emperor | 0.57 | 1.69 | 1.56 | 15.87 | 1.45 |
| Joseph | 0.45 | 1.70 | 1.93 | 25.12 | 1.32 |
| France | 0.43 | 1.42 | 1.73 | 13.80 | 1.24 |
| Here | 0.43 | 1.26 | 1.56 | 11.30 | 1.23 |
| Again | 0.22 | 1.10 | 2.29 | 7.29 | 0.82 |
| First | 0.41 | 1.20 | 1.56 | 10.22 | 1.19 |
| French | 0.39 | 1.29 | 1.73 | 14.16 | 1.14 |
| Make | 0.42 | 1.19 | 1.52 | 10.26 | 1.20 |
| Italy | 0.23 | 1.71 | 2.88 | 8.40 | 0.75 |
| History | 0.21 | 1.29 | 2.60 | 6.66 | 0.78 |
| Look | 0.18 | 1.19 | 2.69 | 7.17 | 0.73 |
| Military | 0.17 | 1.13 | 2.78 | 6.06 | 0.68 |
| Place | 0.20 | 1.31 | 2.69 | 7.18 | 0.78 |
| Little | 0.33 | 1.36 | 2.04 | 8.80 | 1.03 |
| Very | 0.26 | 1.13 | 2.10 | 8.64 | 0.91 |
| Mean | 0.45 | 1.41 | 1.95 | 12.69 | 1.14 |
| S.D. | 0.42 | 0.33 | 0.66 | 8.11 | 0.44 |

Table 2.9 Correlation coefficient between different measures

|  | Mean | Burstiness | Idf | Variance | Entropy |
|---|---|---|---|---|---|
| **Mean** |  |  |  |  |  |
| **Burstiness** | 0.86 |  |  |  |  |
| **Idf** | -0.82 | -0.56 |  |  |  |
| **Variance** | 0.91 | 0.88 | -0.78 |  |  |
| **Entropy** | 0.95 | 0.80 | -0.94 | 0.93 |  |

From the analysis (performed in 2.3.1 to 2.3.6), it can be concluded that except adaptation and Idf other measures are quite useful for the identification of content words in a document.

## 2.4 Weighting of words:

In this section we shall find the standard Z score for each kind of measure discussed in section 2.3, except adaptation and Idf. It has the advantage of converting all the measures having different scales into normal distribution with mean zero and variance 1 and in replica converts the units of each measure in same scale. For each kind of measure, we shall calculate the mean and standard deviation based on the observation taken from the document while the standard Z score shall be calculated by the formula:

$$Z = \frac{X - mean}{S.D.}, \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.3)$$

where $X$ represent the value of different measures. Table 2.10 represents the standard Z scores for each measure.

Weight of a word in a document shall be obtained by normalizing each Z score to unit length using cosine normalization and finding the average of these values with the help of following formula-

$$W = \frac{1}{4}.\frac{(Z_m + Z_{Br} + Z_{Var} + Z_H)}{\sqrt{Z_m{}^2 + Z_{Br}{}^2 + Z_{Var}{}^2 + Z_H{}^2}} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.4)$$

where $Z_m, Z_{Br}, Z_{Var}, Z_H$ represent the standard Z scores for mean, Burstiness, Variance and Entropy respectively.

Table 2.10 Standard Z scores for different measures

| **Word** | $Z_m$ | $Z_{Br}$ | $Z_{Var}$ | $Z_H$ |
|---|---|---|---|---|
| Napoleon | 3.76 | 3.27 | 3.28 | 3.23 |
| Paris | 0.45 | 0.08 | 0.35 | 0.93 |
| Emperor | 0.29 | 0.84 | 0.39 | 0.70 |
| Joseph | -0.01 | 0.89 | 1.53 | 0.42 |
| France | -0.05 | 0.03 | 0.14 | 0.22 |
| Here | -0.05 | -0.46 | -0.17 | 0.21 |
| Again | -0.54 | -0.95 | -0.67 | -0.72 |
| First | -0.10 | -0.64 | -0.30 | 0.11 |
| French | -0.15 | -0.36 | 0.18 | 0.01 |
| Make | -0.08 | -0.65 | -0.30 | 0.13 |
| Italy | -0.52 | 0.92 | -0.53 | -0.89 |
| History | -0.56 | -0.35 | -0.74 | -0.81 |
| Look | -0.63 | -0.67 | -0.68 | -0.93 |
| Military | -0.68 | -0.84 | -0.82 | -1.04 |
| Place | -0.59 | -0.30 | -0.68 | -0.82 |
| Little | -0.29 | -0.15 | -0.48 | -0.25 |
| Very | -0.45 | -0.86 | -0.50 | -0.52 |

Equation (2.4) gives the weights of words in the interval [-1, 1], as the inverse document frequency is in negative correlation with each of the measure. The above defined weight of words in a document shall separate content words from non content words very well as it assigns positive scores to content words and negative score to non content words as is evident from the table 2.11.

Table 2.11 Weights of different words

| **Word** | **Weight** |
|---|---|
| Napoleon | 0.50 |
| Emperor | 0.46 |
| Paris | 0.41 |
| Joseph | 0.39 |
| France | 0.31 |
| Italy | -0.17 |
| French | -0.19 |
| Here | -0.22 |
| Make | -0.31 |
| First | -0.32 |
| Little | -0.46 |
| Place | -0.48 |
| History | -0.48 |
| Very | -0.48 |
| Again | -0.49 |
| Look | -0.49 |
| Military | -0.49 |

## 2.5 Zipf’s law for non function words

In order to check the validity of Zipf’s law for non function words of a document, we have applied the generalized Zipf’s law given by Mandelbrot to the words of our present study. It has been observed that the law does not provide good results but there is a chance to improve the formula. For this purpose, we have ventured to look the relationship between ranks and mean. As mean is directly related to frequency, we can think of obtaining the relationship between frequency and rank. We have proposed the following relationship between these two measures:

$$Mean = \frac{1}{a+br^c} \quad or \qquad \text{.........................}(2.5)$$
$$Frequency = \frac{N}{a+br^c}$$

where $a,b,c$ are parameters of the text and $N$ is the number of paragraphs in the text. Following table shows the values of mean and corresponding ranks sorted according ranks in ascending order.

Table 2.12 Values of Mean and corresponding Ranks of different words

| **Word** | **Frequency** | **Rank** |
|---|---|---|
| Napoleon | 209 | 1 |
| Paris | 66 | 2 |
| Emperor | 59 | 3 |
| Joseph | 46 | 4 |
| France | 44 | 5 |
| Here | 44 | 5 |
| Make | 43 | 7 |
| First | 42 | 8 |
| French | 40 | 9 |
| Little | 34 | 10 |
| Very | 27 | 11 |
| Italy | 24 | 12 |
| Again | 23 | 13 |
| History | 22 | 14 |
| Place | 21 | 15 |
| Look | 19 | 16 |
| Military | 18 | 17 |

Actual frequencies and expected frequencies against rank for the relation $f = P(r + \rho)^{-B}$ have been plotted in following figure. Curve expert' software has been utilized to fit the curves. For the relation $f = P(r + \rho)^{-B}$, the following values have been obtained: $P = 26176.42,\ \rho = 16.22,\ c = 2.03$. The correlation coefficient between actual frequencies and expected frequencies is .77.

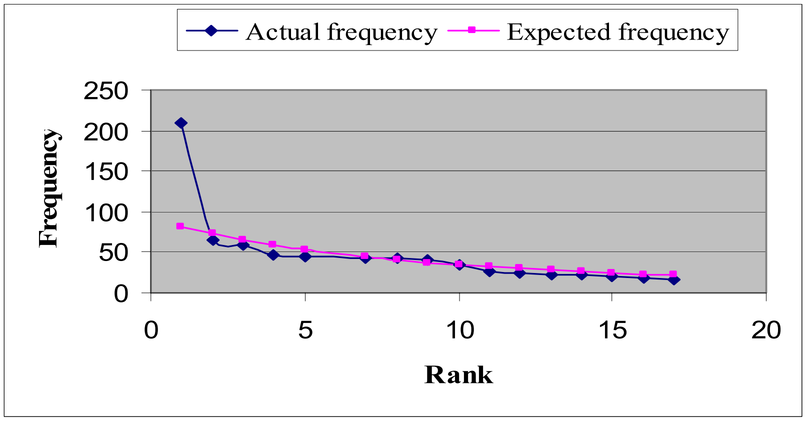

Fig. 2.7 Actual frequency versus expected frequency for $f = P(r + \rho)^{-B}$

Actual frequencies and expected frequencies against rank for the proposed relation have been plotted in the adjoining figure. For the proposed relation $f = N.(a + br^{c})^{-1}$ the following values have been obtained: $a = -57.91,\ b = 58.4,\ c = .021$ and $N = 103$. The correlation coefficient between actual frequencies and expected frequencies is .99.

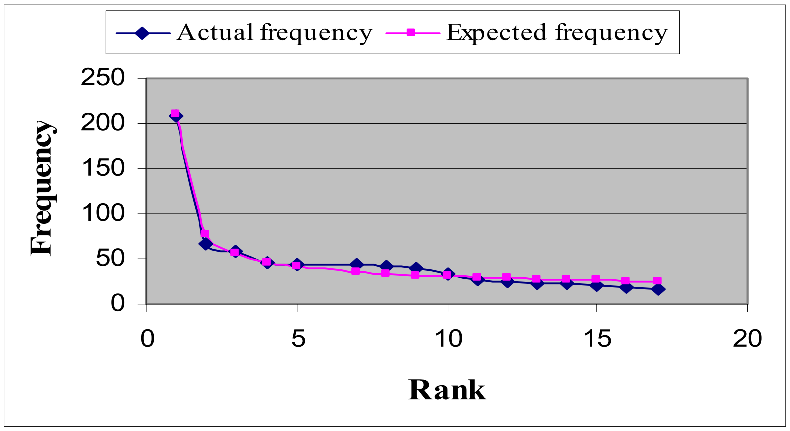

Fig. 2.8 Actual frequency versus expected frequency for $f = N.(a + br^{c})^{-1}$

## 2.6 Verification of the proposed schemes

For the verification of the proposed weighting scheme and redefined Zipf's law, we have taken another text "Napoleon and His Marshals - Vol. I, chapter I". The text has been partitioned into paragraphs of length 300. Some words have been chosen from the text randomly. The proposed relation between rank and frequency has been checked for its suitability for a given text and weights of words have been calculated.

Tables have been generated for different discussed statistical measures. Table 2.13 shows the raw frequencies of different words in the second text while the values of different measures in this text have been shown in table 2.14. Table 2.15 demonstrates the standard Z scores and weights of the words are exhibited in table 2.16. The appropriateness of the proposed relation between rank and frequency has been depicted by Figure 2.9. The parameters of the proposed relation have been obtained as: a= -.01, $b = .73$, c=0.5 and $N = 56$. The correlation coefficient between actual frequencies and expected frequencies being .94.

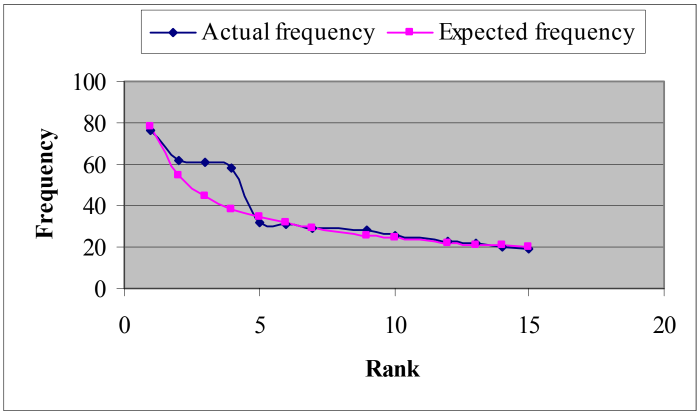

Fig 2.9 Actual frequency versus expected frequency using $f = N.(a+br^{c})^{-1}$ for the second text.

Table 2.13 Words and their corresponding frequencies in second text.

| **Word** | **x** | **0** | **1** | **2** | **3** | **4** | **5** | **6** |
|---|---|---|---|---|---|---|---|---|
| Napoleon | | 26 | 12 | 8 | 6 | 4 | 0 | 0 |
| Bonaparte | | 34 | 14 | 7 | 1 | 0 | 0 | 0 |
| Europe | | 26 | 11 | 11 | 6 | 1 | 0 | 1 |
| France | | 23 | 16 | 5 | 3 | 6 | 1 | 2 |
| England | | 29 | 12 | 7 | 2 | 4 | 2 | 0 |
| Battle | | 39 | 8 | 7 | 1 | 1 | 0 | 0 |
| Mind | **f** | 37 | 14 | 3 | 2 | 0 | 0 | 0 |
| English | | 42 | 9 | 3 | 1 | 0 | 1 | 0 |
| French | | 42 | 10 | 3 | 1 | 0 | 0 | 0 |
| World | | 35 | 17 | 3 | 1 | 0 | 0 | 0 |
| Italy | | 40 | 12 | 2 | 2 | 0 | 0 | 0 |
| Power | | 36 | 16 | 1 | 2 | 1 | 0 | 0 |
| Army | | 36 | 12 | 6 | 0 | 1 | 0 | 0 |
| Moment | | 42 | 9 | 4 | 1 | 0 | 0 | 0 |
| First | | 34 | 15 | 5 | 1 | 5 | 0 | 0 |

Table 2.14 Values of different measures for different words in the document II.

| **Words** | **Mean** | **Br** | **Var** | **H** |
|---|---|---|---|---|
| Napoleon | 1.11 | 2.07 | 18.67 | 2.01 |
| Bonaparte | 0.55 | 1.41 | 8.46 | 1.42 |
| Europe | 1.09 | 2.03 | 13.51 | 1.99 |
| France | 1.36 | 2.30 | 21.84 | 2.20 |
| England | 1.04 | 2.15 | 18.65 | 1.96 |
| Battle | 0.52 | 1.71 | 9.20 | 1.35 |
| Mind | 0.46 | 1.37 | 7.98 | 1.29 |
| English | 0.41 | 1.64 | 7.59 | 1.17 |
| French | 0.34 | 1.36 | 6.14 | 1.09 |
| World | 0.46 | 1.24 | 6.48 | 1.28 |
| Italy | 0.39 | 1.38 | 7.34 | 1.17 |
| Power | 0.50 | 1.40 | 8.00 | 1.31 |
| Army | 0.52 | 1.45 | 7.6 | 1.35 |
| Moment | 0.36 | 1.43 | 6.71 | 1.11 |
| First | 0.57 | 1.45 | 8.34 | 1.46 |
| Mean | 0.65 | 1.63 | 10.43 | 1.48 |
| S.D. | 0.32 | 0.33 | 4.97 | 0.36 |

Table 2.15 Standard Z scores for different measures

| **Word** | $Z_m$ | $Z_{Br}$ | $Z_{Var}$ | $Z_H$ |
|---|---|---|---|---|
| Napoleon | 1.44 | 1.33 | 1.66 | 1.47 |
| Bonaparte | -0.31 | -0.67 | -0.40 | -0.17 |
| Europe | 1.38 | 1.21 | 0.62 | 1.42 |
| France | 2.22 | 2.03 | 2.30 | 2.00 |
| England | 1.22 | 1.58 | 1.65 | 1.33 |
| Battle | -0.41 | 0.24 | -0.25 | -0.36 |
| Mind | -0.59 | -0.79 | -0.49 | -0.53 |

Table 2.15 continue

| **Word** | $Z_m$ | $Z_{Br}$ | $Z_{Var}$ | $Z_H$ |
|---|---|---|---|---|
| French | -0.97 | -0.82 | -0.86 | -1.08 |
| World | -0.59 | -1.18 | -0.79 | -0.56 |
| Italy | -0.81 | -0.76 | -0.62 | -0.86 |
| Power | -0.47 | -0.70 | -0.49 | -0.47 |
| Army | -0.41 | -0.55 | -0.57 | -0.36 |
| Moment | -0.91 | -0.61 | -0.75 | -1.03 |
| First | -0.25 | -0.55 | -0.42 | -0.06 |

Table 2.16 Weight of different words in II document

| **Word** | **Weight** |
|---|---|
| France | 0.50 |
| Napoleon | 0.50 |
| England | 0.50 |
| Europe | 0.48 |
| Battle | -0.30 |
| English | -0.42 |
| First | -0.43 |
| Bonaparte | -0.45 |
| World | -0.48 |
| Moment | -0.49 |
| Mind | -0.49 |
| Army | -0.49 |
| Power | -0.49 |
| Italy | -0.50 |
| French | -0.50 |

## 2.7 Conclusion

On the basis of the analysis of some properties of content words used for global weighting it was found that measures except adaptation and Inverse document frequency can be taken for assigning weights to different words in a document. The proposed measure not only separates content words from non-content words but also assigns weights according to their relevance of the document. For example, the words ‘France’ and ‘Here’ have same frequency in document I but ‘France’ has been selected as content word while ‘Here’ is excluded from the list of content word. It demonstrates that the proposed measure has the capability of differentiating between content words and non content words, no matter whether they have same frequency or not. Two different text related to Napoleon have been selected. The first document is a novel about Napoleon while the second one is concerned with his experiences at the war front.

Zipf’s law has been also restructured for non function words and it has been shown that the proposed relation describes the relation between rank and frequency in better sense.

# Chapter III

# An Evaluation of Different Statistical Techniques of Collocation Extraction Using a Probability Measure to Word Combinations

*A collocation is a recurrent combination of words which appears more often than by chance. Collocations play an important role in many natural language processing tasks such as computational lexicography, word sense disambiguation, machine translation, information retrieval etc. In the present chapter, an attempt has been made to evaluate different collocation extraction techniques for a small corpus. For this purpose, some word pairs have been taken from the corpus and results of each technique have been shown and discussed. Finally, a probability measure has been suggested to filter out some free word combinations before applying these tests, which in turn improves the results obtained from the algorithms. The supremacy of the method designed in the last part of this chapter, over the existing ones has been demonstrated with the help of precision and recall.*

## 3.1 Introduction

A Collocation is an expression consisting of two or more words that correspond to some conventional way of saying things. Natural languages are full of collocations. Classification of words merely on the basis of their meaning is not enough, sometimes their co-occurrence with other words change the meaning dramatically. A word can be better understood by the "company that it keeps". In our daily work we use many idioms and phrases to convey our thoughts to others. Phrases and idioms are the word combinations whose semantics has nothing to do with the meaning of the individual components. In other words we can say that the phrases are the rigid word combinations whose meaning cannot be derived exactly by its word by word meaning. For example 'a white elephant', 'at the eleventh hour', 'flesh and blood' etc. are the word combinations whose word by word meaning is totally different from their actual meanings. On the other hand, free word combinations have the properties that each of the words from the combination can be replaced by another without seriously modifying the overall meaning of the composite unit and if one of the words is omitted, a reader cannot easily infer it from the remaining ones. For example 'end of the lecture', 'buy a house', 'excellent work' etc., are free word combinations. 'Collocations' are a class of word groups which lie between idioms and free word combinations, however it is very typical to draw a line between a phrase and a collocation. Collocations are lexical phenomenon that has linguistics and lexicographic status. Collocations are domain specific, idiosyncratic word combinations with a stable syntactic structure. Manning & Schutze (2002) defined collocation as 'a sequence of two or more consecutive words that has characteristics of a syntactic and semantic unit and whose exact and unambiguous meaning cannot be derived directly from the meaning of its components'.

Natural language expression is an expression which is used by a native speaker. A natural language expression is said to be compositional if the meaning of the expression can be predicted from the meaning of the parts. Collocations show limited compositionality, as either the meaning of the combination is completely different from the free combinations or there is an added element of meaning that cannot be predicted from the parts. For example 'powerful medicine', 'strong coffee', 'white man' etc. have special meanings in themselves. Collocations are also characterized by limited substitutability and limited-modifiability. Limited substitutability simply refers that it is not possible to substitute other words for the constituent word of a collocation. For example, 'powerful coffee' is not appropriate. Similarly, limited modifiability implies that we cannot modify a collocation, such as, change from singular to plural, change in gender etc. According to Smadja (1993), 'collocations are arbitrary, language specific, recurrent in context and common in technical language'.

Collocations are utilized for many natural language applications such as word sense disambiguation, machine translation, computational lexicography, information retrieval, natural language generation etc. Word sense disambiguation is the task of determining the sense of an ambiguous word in a particular use of words. Many words have several meanings or senses. For such words, given out of context, there is ambiguity about their interpretation. For example, the word 'bank' has one meaning with 'river' while different meaning with 'savings'. Translation is one of the important applications of collocations. Since there is no regular syntactic and semantic pattern of collocations, they cannot be translated on word by word basis. Collocation translation improves the quality of machine translation. For example, 'strong tea' in English is '*karak chay'* in Hindi, 'pocket money' in English is '*jeb kharch'* in Hindi. Automatic identification of important collocations to be listed in a dictionary is the task of computational lexicography. Collocational knowledge can improve the performance of information retrieval system. Naturalness of a language like 'strong tea' rather than 'powerful tea' can be well studied through collocations.

Since Collocations occur repeatedly in language, it is essential to look at a large collection of text (corpus) for the occurrence pattern of such type of word pairs so as to get the clear idea of how human acquire, produce and understand language. For that purpose, Mathematics and

Statistics play an important and essential role. Frequency measure and other statistical techniques are very useful for acquiring collocations from a large sample of language. Thus, collocation acquisition comes under the general category of Statistical corpus based approaches to language. Frequency measure was used by Choueka, Klein & Neuwitz (1983) to identify a particular type of collocations. Church & Hanks (1989) used mutual information to extract word pairs that tend to co-occur within a fixed size window (normally 5 words), in which extracted words may not be directly related. The t-test was suggested by Church & Hanks (1989) to find words whose co-occurrence patterns best distinguish between two words. Church & Gale (1991) used the $\chi^2$- test in statistical natural language processing for the identification of translation pairs in aligned corpora. Smadja (1993) extracted collocations through a multi stage process taking the relative positions of co-occurring words into account. Dunning (1993) applied likelihood ratio test to collocation discovery. Collocation extraction and their use in finding word similarity was suggested by Lin (1998).

In Section 3.2 of the present chapter some of the collocation extraction techniques have been described. In order to evaluate these techniques for a small corpus, we have compiled a corpus of about 1.5 million words from different online novels. Details of which can be obtained from www.free-online-novels.com. In Section 3.3, these techniques of collocation extraction have been applied to the compiled corpus and discussion on these results has been made. In Section 3.4, a probability measure to word combinations method has been suggested to filter some word pairs before applying these tests to improve the results of these tests. Finally, Section 3.5 deals with the discussion of the above study.

## 3.2 Different Collocation Extraction Techniques

There are various statistical techniques for collocation extraction discussed by different authors. In this section we discuss these techniques with relative advantages and disadvantages.

### 3.2.1 Frequency Measure

The simplest approach to find collocations in a text is of frequency measure. If two words occur together a lot, then it signifies that this may be a kind of word combination which has some special meaning rather than an ordinary word combination. In their work, Choueka, Klein & Neuwitz (1983) retrieved the word sequences that appear more frequently than a given threshold. In practical they retrieved word sequences consisting of two to six words but theoretically they were interested for any length. They used 11 million word corpus from the New York Times archives. This approach was very sensitive to corpus size as it was based on it.

### 3.2.2 Mutual Information

Mutual information from information theory has been utilized to find the closeness between word pairs by Church & Hanks (1989) and Hindle (1990). Mutual information for two events x and y is defined as:

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x).P(y)} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3.1)$$

If we write $w_1$ and $w_2$ for the first and second word respectively, instead of x and y, then the mutual information for the two words $w_1$ and $w_2$ is given by:

$$I(w_1,w_2) = \log_2 \frac{P(w_1,w_2)}{P(w_1).P(w_2)}, \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots(3.2)$$

where $P(w_1, w_2)$ is the probability of two words $w_1$ and $w_2$ coming together in a certain text and $P(w_1)$ and $P(w_2)$ are the probabilities of $w_1$ and $w_2$ appearing separately in the text, respectively.

If $P(w_1, w_2) = P(w_1) . P(w_2)$, that is, the two words are independent to each other, then $I(w_1, w_2) = 0$, which indicates that these two words are not making a collocation. A high mutual information score indicates a good collocation.

### 3.2.3 Hypothesis testing

To know whether the two words occur together more often than by chance, that is, they form a collocation, hypothesis testing may be utilized. For that purpose, we formulate a null hypothesis $H_0$ that the two words $w_1$ and $w_2$ appear independently in the text. So under the null hypothesis $H_0$, the probability that the words $w_1$ and $w_2$ are coming together is simply given by:

$$P(w_1, w_2) = P(w_1) . P(w_2) .$$

The null hypothesis has been tested by different authors using t-test, $\chi^2$-test and likelihood ratio test. If the null hypothesis is accepted, we conclude that the occurrence of two words is independent of each other. Otherwise, we may conclude that they depend on each other, that is, they form collocations. The different tests used in the literature for testing of null hypothesis are described as follows:

#### 3.2.3.1 The t test

The t-test has been used by Church & Hanks (1989) for collocation discovery to test the validity of a hypothesis. In t-test we use the null hypothesis that the sample is drawn from a distribution with mean $\mu$, taking sample mean and variance into account. t-test considers the difference between the observed and expected mean. The t statistic is defined as:

$$t = \frac{\overline{x} - \mu}{\sqrt{s^2 / N}} \sim t_{n-1}(\alpha) \qquad \text{..........................(3.3)}$$

where $\overline{x}$ is the sample mean, $s^2$ is the sample variance, $N$ is the sample size, $\mu$ is the mean of the distribution and $t_{n-1}(\alpha)$ denotes a t- distribution with (n-1) degrees of freedom at $\alpha$ level of significance. To apply t-test for testing the independence of two words $w_1$ and $w_2$, we assume that $f(w_1)$, $f(w_2)$ and $f(w_1, w_2)$ are the respective frequencies of the word $w_1$, $w_2$ and $w_1\, w_2$ in the corpus and $N$ is the total number of words / bigrams in the corpus. Then, we have,

$$P(w_1) = \frac{f(w_1)}{N} \quad (\text{say } p_1),$$

$$P(w_2) = \frac{f(w_2)}{N} \quad (\text{say } p_2),$$

$$P(w_1, w_2) = \frac{f(w_1, w_2)}{N} \quad (\text{say } p_{12}),$$

The null hypothesis is

$$H_0 : \quad P(w_1, w_2) = P(w_1).P(w_2) = p_1 \,.\, p_2$$

If we select bigrams (word pairs) randomly then the process of randomly generating bigrams of words and assigning 1 to the outcome that the particular word combination for which we are looking for is a collocation and 0 to any other outcome follows a Bernoulli distribution. For the Bernoulli distribution we have

Mean $(\mu) = p$ and Variance $(\sigma^2) = p(1-p)$.

Thus, if the null hypothesis is true, the mean of the distribution is $\mu = p_1.p_2$. Also, for the sample, we have $P(w_1, w_2) = p_{12}$. Therefore, using

Binomial distribution, sample mean $\overline{x} = p_{12}$ and sample variance $s^2 = p_{12}(1 - p_{12})$.

Using (3.3), we calculate the value of t and compare it with the tabulated value at given level of significance. If the value of $|t|$ for a particular bigram is greater than the value obtained from the table, we reject the null hypothesis, which indicates that the bigram may be considered as a collocation.

### 3.2.3.2 Pearson's chi-square test

The t-test assumes that the probabilities are distributed approximately normal, which is not true in general. $\chi^2$- test is an alternative test for dependence which does not assume normally distributed probabilities. To test the independence between the words, the $\chi^2$- test compares the observed frequencies with the expected frequencies. A 2×2 contingency table for testing the independence of the occurrence of two words $w_1$ and $w_2$ is given as:

| | $w_1$ | $w_1^c$ | Total |
|---|---|---|---|
| $w_2$ | $f(w_1, w_2)$ | $f(w_1^c, w_2)$ | $f(w_2)$ |
| $w_2^c$ | $f(w_1, w_2^c)$ | $f(w_1^c, w_2^c)$ | $f(w_2^c)$ |
| Total | $f(w_1)$ | $f(w_1^c)$ | $N$ |

,

where $f(w_1^c, w_2)$ is the number of bigrams with the first word not being $w_1$ but second being $w_2$ = $f(w_2) - f(w_1, w_2)$, $f(w_1, w_2^c)$ is the number of bigrams with the second word not being $w_2$ but first being

$w_1$= $f(w_1)-f(w_1,w_2)$ and $f(w_1{}^c,w_2{}^c)$ is the number of bigrams with neither first word being $w_1$ nor the second word being

$$w_2 = N-[\,f(w_1,w_2)+f(w_1{}^c,w_2)+f(w_1,w_2{}^c)\,]$$

The test statistics for testing of the null hypothesis $H_0$ is given by

$$\chi^2 = \sum_{i,j}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2{}_{(n-1)}(\alpha)\,, \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3.4)$$

where i and j range over rows and columns of the table respectively, $O_{ij}$ and $E_{ij}$ are the observed and expected values for cell $(i,j)$ and $\chi^2{}_{(n-1)}(\alpha)$ denotes a $\chi^2$- distribution for (n-1) degree of freedom and $\alpha$ level of significance. For 2×2 contingency table, $\chi^2$ is calculated directly from observed frequencies using the simplified formula:

$$\chi^2 = \frac{N(O_{11}O_{22}-O_{12}O_{21})^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})} \quad \ldots\ldots\ldots\ldots\ldots(3.5)$$

This follows a $\chi^2$ distribution with 1 degree of freedom. Choosing an appropriate significance level, $\chi^2$ values for one degree of freedom from the table provides us the criteria for testing of null hypothesis. If the calculated value of $\chi^2$ for a certain bigram comes beyond the tabulated value, it indicates that the null hypothesis is false. This provides us a basis for considering the given bigram as a collocation.

### 2.3.3. Likelihood ratio test

Likelihood ratio test is more interpretable than $\chi^2$- test in the sense that it tells us how much more likely one hypothesis is than the other.

Dunning (1993) examined two hypothesis for the occurrence pattern of bigrams. These are:

$H_0$ : $w_1$ and $w_2$ are independent,

$$\Rightarrow P(w_2 / w_1) = P(w_2 / w_1{}^c) = p_0,$$

that is, the probability of appearance of $w_2$ under the restriction that $w_1$ appears just before $w_2$ is same as the probability of appearance of $w_2$ under the condition that a word other than $w_1$ has appeared before $w_2$.

The alternative hypothesis is defined as:

$H_1$ : $w_1$ and $w_2$ are dependent, that is,

$$P(w_2 / w_1) = p_1 \neq P(w_2 / w_1{}^c) = p_2.$$

The maximum likelihood estimates of $p_0$, $p_1$ and $p_2$ are given by

$$P(w_2 / w_1) = \frac{f(w_1, w_2)}{f(w_1)} = p_1,$$

$$P(w_2 / w_1{}^c) = \frac{f(w_1{}^c, w_2)}{f(w_1{}^c)} = p_2 \text{ and}$$

$$P(w_2) = \frac{f(w_2)}{N} = p_0 \ .$$

Assuming that the occurrence of a word $w_i$ in the corpus follows the Binomial distribution, $b(r; n, x) = {}^n c_r x^r (1-x)^{n-r}$, the likelihood of getting the count for $w_1$ , $w_2$ and $w_1\, w_2$ that we actually observed is given by:

$L(H_0) = b(f(w_1, w_2); f(w_1), p) b(f(w_1{}^c, w_2); f(w_1{}^c), p)$, for the null hypothesis $H_0$ and

$L(H_1) = b(f(w_1, w_2); f(w_1), p_1) b(f(w_1{}^c, w_2); f(w_1{}^c), p_2)$, for the alternative hypothesis $H_1$.

The log of the likelihood ratio $\lambda$ will be:

$$\log\lambda = \log\frac{L(H_0)}{L(H_1)} \qquad \text{..............................(3.6)}$$

It was shown by Mood, Graybill & Boes (1974, pp. 440) that the quantity $-2\log\lambda$ has an asymptotically $\chi^2$- distribution with r degrees of freedom, where r is the number of parameters under comparison. If the calculated values of $-2\log\lambda$ is less than the tabulated value of $\chi^2$ at given level of significance, we accept the null hypothesis $H_0$, indicating the independence of the words $w_1$ and $w_2$, otherwise the hypothesis $H_1$ is accepted which may be considered as the presence of collocation between the words $w_1$ and $w_2$.

## 3.3 Evaluation of Collocation Extraction Techniques

We have applied all the techniques described in Section 3.2 to our corpus of about 1.5 million words compiled from different online novels and stories available at internet. The frequency counts of different word combinations for the given corpus are presented in table 3.1 in which the word combinations have been arranged in descending order with respect to the frequency that the two words coming together, that is $f(w_1, w_2)$. The word pairs are selected randomly. The results of the different techniques are as follows:

From the table 3.1, we find that the free word combinations like, ‘time before’, ‘long after’, ‘must take’ etc. are ranked higher while word combinations such as ‘attentive reader’, ‘national anthem’, ‘human being’ which are semantic units and hence form collocations are ranked below. Therefore, we conclude that frequency measure is not sufficient to extract collocations from a small corpus.

Table 3.1 Frequency counts of different word combinations for the corpus

| $w_1$ | $w_2$ | $f(w_1)$ | $f(w_2)$ | $f(w_1, w_2)$ |
|---|---|---:|---:|---:|
| like | that | 3693 | 17807 | 189 |
| last | night | 1180 | 1060 | 124 |
| last | time | 1180 | 3126 | 77 |
| dark | shadow | 543 | 140 | 67 |
| great | deal | 1100 | 198 | 44 |
| time | before | 3126 | 1934 | 25 |
| strong | enough | 289 | 934 | 24 |
| make | sense | 1411 | 298 | 23 |
| long | after | 1422 | 2267 | 22 |
| must | take | 1207 | 1931 | 21 |
| night | air | 1060 | 545 | 13 |
| long | way | 1422 | 2063 | 13 |
| only | because | 2106 | 1132 | 12 |
| human | being | 242 | 33 | 11 |
| national | guard | 68 | 151 | 11 |
| public | opinion | 228 | 92 | 9 |
| make | use | 1411 | 399 | 8 |
| human | nature | 242 | 142 | 7 |
| strong | man | 289 | 2303 | 7 |
| most | powerful | 1024 | 118 | 6 |
| still | force | 1624 | 195 | 6 |
| only | chance | 2106 | 283 | 6 |
| only | after | 2097 | 2268 | 6 |
| little | chap | 2006 | 136 | 2 |
| together | against | 521 | 892 | 2 |
| attentive | reader | 11 | 9 | 1 |
| national | anthem | 68 | 5 | 1 |
| midnight | desert | 53 | 87 | 1 |
| last | among | 1080 | 191 | 1 |
| over | many | 2767 | 753 | 1 |

The scores of mutual information calculated through (3.2) are presented in table 3.2. From Section 3.2, we conclude that for the cases of zero mutual information, the two words will be independent. Mutual information approaching to zero shows a poor dependence between words. However, there is no boundary on the basis of which we can decide whether the word pairs are collocations or not. Only conclusion which we can draw is that the words pairs which have mutual information far from zero have a good tendency of making collocations. So, the technique of mutual information also fails to clearly extract collocations from a given corpus.

Table 3.2 Mutual information scores of different word combinations from the corpus

| $w_1$ | $w_2$ | $M.I.$ | $w_1$ | $w_2$ | $M.I.$ |
|---|---|---|---|---|---|
| attentive | reader | 13.85 | only | chance | 3.88 |
| national | anthem | 12.07 | must | take | 3.72 |
| human | being | 10.98 | little | chap | 3.43 |
| national | guard | 10.62 | long | after | 3.32 |
| dark | shadow | 10.34 | only | because | 2.88 |
| public | opinion | 9.30 | last | among | 2.83 |
| midnight | desert | 8.31 | still | force | 4.80 |
| human | nature | 8.22 | make | use | 4.38 |
| great | deal | 8.21 | strong | man | 3.95 |
| last | night | 7.18 | long | way | 2.70 |
| strong | enough | 7.03 | together | against | 2.66 |
| make | sense | 6.32 | time | before | 2.60 |
| most | powerful | 6.19 | like | that | 2.07 |
| night | air | 5.04 | only | after | 0.89 |
| last | time | 4.94 | over | many | 0.50 |

Table 3.3 demonstrates the values of t-score, chi square and likelihood ratio for different word pairs in the corpus, using (3.3), (3.4) and (3.6), respectively. From table 3.3, we observe that some free word combinations such as, ‘time before’, ‘like that’, ‘long after’ etc. are termed as collocations by t-test, which is a drawback of applying t-test to a small corpus. Looking over to the results of the chi square test, again we find that some free word combinations such as ‘together against’, ‘still force’ etc. are included in the list of collocations. Therefore, we can

say that Chi square is also not a good measure of collocations extraction. Likelihood ratio test gives slightly better result than Chi square test but many free word combinations are still the part of the result.

Table 3.3 Values of t-score, Chi square and likelihood ratio for different bigrams in the corpus.

| $w_1$ | $w_2$ | $t\text{-}score$ | $Chi\ \ Sq.$ | $-2\log\lambda$ |
|---|---|---:|---:|---:|
| last | night | 11.06 | 17858.83 | 1017.54 |
| like | that | 10.50 | 472.46 | 263.26 |
| last | time | 8.49 | 2218.87 | 384.96 |
| dark | shadow | 8.18 | 86749.47 | 873.99 |
| great | deal | 6.61 | 129996.20 | 425.67 |
| strong | enough | 4.86 | 3093.03 | 187.39 |
| make | sense | 4.74 | 1806.62 | 158.54 |
| must | take | 4.24 | 238.6 | 70.21 |
| long | after | 4.22 | 179.45 | 62.31 |
| time | before | 4.18 | 106.51 | 48.85 |
| night | air | 3.50 | 405.06 | 66.21 |
| human | being | 3.31 | 22264.97 | 150.17 |
| national | guard | 3.31 | 17311.17 | 142.69 |
| long | way | 3.05 | 60.86 | 26.86 |
| only | because | 3.00 | 1.1* | 1.21* |
| public | opinion | 3.00 | 5662.30 | 99.35 |
| make | use | 2.69 | 151.75 | 33.59 |
| human | nature | 2.64 | 2083.60 | 66.43 |
| strong | man | 2.47* | 94.89 | 25.41 |
| most | powerful | 2.42* | 426.57 | 39.99 |
| still | force | 2.36* | 155.59 | 28.55 |
| only | chance | 2.28* | 77.36 | 21.28 |
| little | chap | 1.28* | 17.77 | 5.9* |
| together | against | 1.19* | 8.89 | 4.02* |
| only | after | 1.13* | 2.37* | 1.89* |
| attentive | reader | 1.00* | 14853.40 | 17.42 |
| national | anthem | 1.00* | 4323.70 | 14.97 |
| midnight | desert | 1.00* | 316.98 | 9.57 |
| last | among | 0.86* | 5.27* | 2.21* |
| over | many | 0.42* | 0.12* | 0.14* |

* Rejected for collocation at $\alpha$ = .005

The results obtained in tables 3.1, 3.2 and 3.3 reflect that although the existing techniques of collocation extraction are useful for identification of collocations but most of the times they provide misleading results as many free word combinations are also termed as collocations or some actual collocations are not accepted inspite of their high rankings. These results motivated us to search for a filter, before applying these techniques. We shall see that these drawbacks can be removed to a great extent by using the proposed filter which is described in the next section.

## 3.4 Probability Measure for Collocation Extraction

In this section we have given a method to filter some free word combinations before applying the statistical techniques of collocation extraction described in Section 3.2. After using this filter, only selected number of word pairs will be taken for further processing, which in turn improves the results of these techniques.

If $f(w_1)$, $f(w_2)$ and $f(w_1, w_2)$ represent the frequencies of a word $w_1$, $w_2$ and the combination of two words $w_1$ and $w_2$ respectively, then the number of combination per word or proportion of combination due to the words $w_1$ and $w_2$ can be derived from the ratios:

$$\lambda_1 = \frac{f(w_1, w_2)}{f(w_1)} \quad \text{and}$$

$$\lambda_2 = \frac{f(w_1, w_2)}{f(w_2)}, \quad \text{where } 0 < \lambda_i < 1,\ i = 1, 2.$$

We can draw a two dimensional vector expressible in the form:

A=$\lambda_1 \hat{\lambda}_1 + \lambda_2 \hat{\lambda}_2$, where $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are unit vectors.

Since for a perfect closeness,

Maximum $[f(w_1, w_2)]$ = $f(w_1) = f(w_2)$, that is max. $\lambda_i$ = 1 $\forall$ $1 \le i \le 2$. Thus for maximum values of $\lambda_1$ and $\lambda_2$ the vector will be as follows:

$$\mathrm{A}_{\max} = \hat{\lambda}_1 + \hat{\lambda}_2 .$$

We have the lengths of these vectors as follows:

$$\|A\| = \sqrt{\lambda_1^{\,2} + \lambda_2^{\,2}} \quad \text{and} \quad \|A_{\max}\| = \sqrt{2}$$

The probability that the two words $w_1$ and $w_2$ make good combination can be denoted by

$$P = \frac{\|A\|}{\|A_{\max}\|},$$

as the ratio $\frac{\|A\|}{\|A_{\max}\|}$ cannot be greater than one and it always takes non-negative values except perhaps on a set N whose probability measure is zero.

Calculating the probability for each word combination and rejecting the word combinations whose probability is too low will reduce the number of words to be proceeded for further testing. Table 3.4 demonstrates the probability ($P$) for each word combination tested for collocation in Section 3.3. The lower value of $P$ in the table 3.4 ($P < .05$) provides us a criteria for rejecting some word combinations before further processing.

The results of different tests for collocation extraction discussed in Section 3.2 for remaining word pairs after applying the filter discussed above are shown in table 3.5.

Table 3.4.The $P$ values for different word combinations in the corpus.

| $w_1$ | $w_2$ | $P$ | $w_1$ | $w_2$ | $P$ |
|---|---|---|---|---|---|
| dark | shadow | 0.350 | night | air | 0.019* |
| human | being | 0.238 | strong | man | 0.017* |
| great | deal | 0.160 | midnight | desert | 0.016* |
| national | anthem | 0.142 | only | chance | 0.015* |
| national | guard | 0.125 | make | use | 0.015* |
| last | night | 0.111 | must | take | 0.015* |
| attentive | reader | 0.102 | long | after | 0.013* |
| public | opinion | 0.075 | time | before | 0.011* |
| strong | enough | 0.061 | little | chap | 0.010* |
| make | sense | 0.056 | only | because | 0.009* |
| last | time | 0.049* | long | way | 0.008* |
| human | nature | 0.040* | last | among | 0.004* |
| like | that | 0.037* | together | against | 0.003* |
| most | powerful | 0.036* | only | after | 0.003* |
| still | force | 0.022* | over | many | 0.001* |

*Rejected for further processing due to low probability ($P < .05$)

Table 3.5 The values of mutual information, t-score, Chi square and likelihood ratio for the word combinations left after filtration process.

| $w_1$ | $w_2$ | $M.I.$ | $t\ Score$ | $Chi\ Sq.$ | $-2\log\lambda$ |
|---|---|---|---|---|---|
| dark | shadow | 10.34 | 8.18 | 86749.47 | 873.99 |
| human | being | 10.98 | 3.31 | 22264.97 | 150.17 |
| great | deal | 8.21 | 6.61 | 129996.20 | 425.67 |
| national | anthem | 12.08 | 1.00 | 4323.70 | 14.97 |
| national | guard | 10.62 | 3.31 | 17311.17 | 142.69 |
| last | night | 7.19 | 11.06 | 17858.83 | 1017.54 |
| attentive | reader | 13.86 | 1.00* | 14853.40 | 17.42 |
| public | opinion | 9.30 | 3.00 | 5662.30 | 99.35 |
| strong | enough | 7.03 | 4.86 | 3093.03 | 187.39 |
| make | sense | 6.33 | 4.74 | 1806.62 | 158.54 |

* Rejected for collocation at $\alpha = .005$

From table 3.5, we observe that after applying the probability criteria, all the free word combinations which were described as collocations on the basis of techniques discussed in Section 3.2, are filtered out. The word combinations, which are considered as collocations on the basis of different tests demonstrated in table 3.5 do not seem to contain free word combinations.

## 3.5 Collocation extraction

In this section we have suggested the use of Z statistics for extracting collocations from a small sample of text. In the first stage, we have used the probability measure defined in the section 3.4 to filter out some free word combinations and in the second stage we have used the 'Z-statistic' to extract collocations from the corpus.

### 3.5.1 The Z-Statistic

In this section we will utilize the Z-statistic to extract collocation. Let $N$ be the total number of words in the text, then probability of a word appearing in a text can be given as $P(w) = f(w)/N$.

Therefore, $P(w_1/w_2) = P(w_1w_2)/P(w_2) = \dfrac{f(w_1w_2)/N}{f(w_2)/N} = \dfrac{f(w_1w_2)}{f(w_2)} = \lambda_2$

Similarly $\lambda_1 = P(w_2/w_1)$

If the two words are independent, then $P(w_1w_2) = P(w_1)P(w_2)$. Therefore, $\lambda'_2 = P(w_1/w_2) = P(w_1)$, where $\lambda'_2$ is the value of $\lambda_2$, when the two words are independent. Similarly, $\lambda'_1 = P(w_2/w_1) = P(w_2)$.

Let $A_0$ be the length of the vector when two words are independent and $P_0$ be the corresponding probability. Then $A_0 = \sqrt{\lambda'^2_1 + \lambda'^2_2}$ and $P_0 = \dfrac{\|A_0\|}{\|A_{\max.}\|}$.

Our objective is to know whether there is any significant difference between the two probabilities, that is between $P$ and $P_0$. For this we define that let X be the random variable representing the probability of a word combination to be considered for collocation. Choosing the Null hypothesis:

$H_0$: The two words are independent i.e. $P(w_1w_2) = P(w_1)P(w_2)$

If the null hypothesis is true, then in the random process of taking combinations of words if we assign 1 to the outcome of a particular word combination and 0 to any other outcome, then it follows a Bernoulli trial with $P_0 = \frac{\|A_0\|}{\|A_{\max .}\|}$ .

The mean of this distribution is $\mu = P_0$ and the variance is $\sigma^2 = P_0(1-P_0)$.

Under the assumption of standard normal distribution, the test statistic Z = $\frac{X \sim \mu}{\sigma}$ ~ N (0, 1). If the value of Z is significant, it will show that the null hypothesis is wrong, that is, the two word combinations are not independent and can be considered for collocations. We have taken the 1% significance level to draw inference from the 'Z- statistic'. In our present study we have taken bigrams only, that is n = 2.

### 3.5.2 Collocation extraction for the sample corpus

We have applied the proposed method to our compiled corpus of about one million words compiled from some novels of Project Gutenberg (vide appendix) available on www.gutenberg.org/etext/<no.> by randomly selecting the word pairs from the corpus. The values of Z-statistic for some word pairs have been shown in the table 3.6.

Table 3.6: P and Z scores for different word pairs

| $w_1$ | $w_1$ | $f(w_1)$ | $f(w_2)$ | $f(w_1w_2)$ | $P$ | $Z$ |
|---|---|---|---|---|---|---|
| fire | bucket+ | 291 | 15 | 5 | 0.24 | 16.87 |
| christmas | eve+ | 72 | 33 | 9 | 0.21 | 29.09 |
| more | than | 2124 | 1563 | 369 | 0.21 | 4.89 |
| little | episode | 1630 | 16 | 4 | 0.18 | 5.32 |
| young | man+ | 741 | 2138 | 147 | 0.15 | 3.77 |
| base | camp+ | 54 | 55 | 7 | 0.13 | 17.86 |
| great | deal+ | 911 | 118 | 20 | 0.12 | 4.84 |
| human | being+ | 182 | 735 | 30 | 0.12 | 5.31 |
| both | sides | 409 | 69 | 11 | 0.11 | 6.84 |
| away | from | 862 | 3945 | 135 | 0.11 | 2.13* |
| public | opinion+ | 190 | 103 | 10 | 0.08 | 6.47 |
| last | century | 846 | 32 | 3 | 0.07 | 2.76 |
| clumsy | fashion | 11 | 126 | 1 | 0.06 | 7 |
| long | journey | 967 | 100 | 7 | 0.05 | 1.92* |
| cheerful | noise | 48 | 720 | 3 | 0.04 | 1.99* |
| national | guard+ | 39 | 163 | 2 | 0.04 | 3.51 |
| human | nature+ | 182 | 251 | 7 | 0.03 | 2.31* |
| evil | eye+ | 124 | 259 | 4 | 0.03 | 1.81* |
| good | terms | 1299 | 88 | 3 | 0.02 | 0.79* |
| dark | shadow | 279 | 93 | 3 | 0.02 | 1.70* |
| come | over | 1358 | 1394 | 19 | 0.01 | 0.35* |
| time | before | 1463 | 1164 | 11 | 0.01 | 0.21* |
| like | some | 1602 | 1786 | 11 | 0.01 | 0.12* |
| along | over | 367 | 1394 | 3 | 0.01 | 0.16* |
| away | down | 862 | 1517 | 6 | 0.01 | 0.13* |
| might | still | 1143 | 799 | 5 | 0.01 | 0.15* |
| looking | through | 431 | 1010 | 3 | 0.01 | 0.17* |
| very | dark | 1410 | 279 | 2 | 0.01 | 0.14* |

* Rejected for collocation at 1% level of significance ( Z=2.58)
+ Actual collocation

We have checked the validity of the word pairs through www.thefreedictionary.com. Here twenty eight word pairs have been shown on exemplary basis. Out of which only eleven word pairs have been found meaningful. The precision and recall of the proposed method is 61% and 80% respectively. For the same set of word pairs we have found that the precision and recall of t-test is 50% and 70% respectively. From the table it can be observed that word combinations that have a low probability measure score also have low Z score. Probability measure can be used to filter out free word combinations. We can choose the criteria $P < .01$ to reject word pairs before applying Z-test.

## 3.6 Discussion

The present work was carried out with the intention of checking out the suitability of different statistical methods available in the literature for extraction of collocations for a small corpus. In Section 3.3, we have found that the different techniques of collocation extraction are insufficient for identification of collocations from a small corpus. Therefore, in Section 3.4, we have suggested a filter based on the probability measure to word combinations. It was found that if the proposed filter is applied to the word combinations before applying the different statistical techniques of collocation extraction, there is a remarkable improvement in the results. Due to the application of the proposed filter, many word combinations which are actually not collocations were discarded before applying the further statistical techniques for collocation extraction. This process has served two important purposes. Many free word combinations which were included in the list of collocations as the result of various statistical techniques have been filtered out before application of these techniques and Secondly, due to filtering out of many free word combinations at the initial stage, a lot of computation and labour has been saved while applying the different statistical techniques for collocation extraction. For instance, in our study of small corpus, out of 30 word combinations, 20 were filtered out as a result of the proposed filter and the further processing was carried out for only 10 word combinations. Therefore, we conclude that for improving the performance of collocations extraction techniques, the prescribed method of filtering may be applied before applying the statistical techniques for collocation extraction.

In section 3.5, we have a collocation extraction technique which is an effort to look at the word pairs which form collocation especially considering the concept of semantic unit. The suggested method has shown a good result in comparison to t-test. The proposed method is quite useful for the extraction of collocation when the size of the corpus is small. The suggested technique does not depend on the size of the corpus which is an additional advantage of the proposed method over existing methods.

# Chapter IV

# Fuzzy Approach to Collocation Extraction

*Mathematical models of collocation extraction look at the closeness between two words. The term "closeness" is a linguistic term. Hence closeness can be better understood through fuzzy approach. In the present chapter an attempt has been made to utilize Mutual information and t-test, the two existing techniques of collocation extraction, for fuzzification. In the first part two fuzzy set theoretical models have been proposed to identify collocations and second part fuzzy inference system has been utilized for collocation extraction. It has been shown that fuzzy set theoretical approach works very well for collocation extraction. The working data has been based on a corpus of about one million words contained in different novels constituting project Gutenberg available on www.gutenberg.org.*

## 4.1 Introduction

Almost all the techniques of collocation extraction look at whether the probability of seeing a combination differs significantly from what we would expect from their component words, that is, the product of the probabilities of component words. If the probability of a word combination does not differ significantly, then we reject the word combination. To decide whether a word combination makes a collocation or not is a vague measurement. One can not apply a particular rule of collocation extraction for every word combinations; thus fuzzy approach is quite useful for collocation extraction. The classical logic relies on something which is either true or false. A True element is usually assigned a value of 1 and false has a value 0. Thus, something either completely belongs to a set or it is completely excluded from the set. The fuzzy logic broadens this definition of classical logic. The basis of the logic is fuzzy sets. Unlike in classical sets, where membership is full or none, an object is allowed to belong only partly to one set. The membership of an object to a particular set is described by a real value which lies between 0 and 1. Thus, for instance, an element can have a membership value 0.5, which describes a 50% membership in a given set. Such logic allows a much easier application of many problems that cannot be easily implemented using classical approach.

In the first paper on fuzzy decision making, Bellman and Zadeh (1970) suggest a fuzzy model of decision making in which relevant goals and constraints were expressed in terms of fuzzy sets and a decision is determined by an appropriate aggregation of these fuzzy sets. Fuzzy logic provides an easy way to check the possibility whether a word combination can be considered as collocation or not. Detailed discussion on fuzzy sets, fuzzy logic and fuzzy decision making can be had from the works of Klir et al (2001).

In the first part of this chapter, we have expressed relevant goals and constraints in terms of fuzzy sets of high mutual information score and significant t-score. For the decision regarding collocation, two different aggregation models of these fuzzy sets have been defined. In the second part fuzzy inference system has been utilized for collocation extraction. Fuzzy logic allows the formation of a logic based model by utilizing the reasoning behind the existing methods, so we have constructed a model that has the simplicity of the logic based model and performs better than the existing statistical models.

## 4.2 Fuzzy set theoretic models for collocation extraction

In this section, we have mentioned the two techniques of collocation extraction: mutual information and t-score. The two fuzzy sets based on these two methods, have been defined.

### 4.2.1 Fuzzification of Mutual Information

The term ‘high mutual information score’ is quite vague as it does not provide a basis to say which mutual information score can be considered as high. Let us consider a fuzzy set A of collocations, then each word combination will be a member of the set A with a particular grade of membership. Grade of membership can be defined with the help of mutual information scores. To find the grade of membership using the mutual information scores, we have analyzed the pattern of mutual information scores. Instead of looking for every mutual information score, we have classified mutual information scores into small class intervals and assigned ranks to word combinations according to the classes, so that word combinations falling under a class have same grade of membership. Table 4.1 shows the mutual information and their corresponding ranks. In this table, we have selected too small class intervals on the basis of the observation that small class interval leads to accuracy in the prediction of ranks from the distribution. We can define a function to get the rank of a word combination from its mutual information. Let $x \in R^+$ be the mutual information of a word combination and $y \in I^+$ be the corresponding rank, then $f : R^+ \to I^+$ such that $y = f(x) = \left\lceil \frac{x}{0.25} \right\rceil$. $\lceil x \rceil$ is the ceiling function.

Table 4.1 Mutual information scores and corresponding ranks.

| Mutual Information score | Rank |
|---|---|
| 0 - 0.25 | 1 |
| 0.26 - 0.50 | 2 |
| 0.51 - 0.75 | 3 |
| 0.76 - 1.00 | 4 |
| …………….. | … |
| …………….. | … |
| 12.76 - 13.00 | 51 |
| ……………….. | … |

To find the grade of membership from the rank, we can define another function $g : I^+ \to [0,1]$ such that $y = g(x)$, where $x \in I^+$ is the rank and $y \in [0,1]$ is the corresponding grade of membership. For defining the function $g(x)$, we know that it should tend to zero, when $x$ tends to one (lowest mutual information rank) and to one, when $x$ tends to infinity (highest mutual information score) respectively. Therefore we can define $g(x)$ as follows:

$$y = g(x) = \frac{\log x}{\log (x + a)}, \; a > 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4.1)$$

where $a$ is the parameter.

From the function $y = g(x)$, it is clear that $x \to 1, y \to 0$ and $x \to \infty, y \to 1$. From observations, the value of the parameter $a$ is chosen equal to 10.

Finally grade of membership for a word combination based on the mutual information scores can be given as:

$$A_I = g\,of\,(x) = g(f(x)) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4.2)$$

The figurative representation of grade of membership is as under:

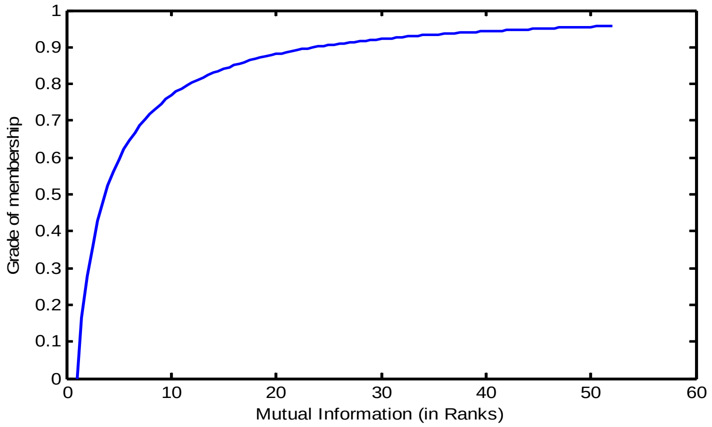

Fig.4.1 Grade of membership for mutual information scores

### 4.2.2 Fuzzification of t-score

To accept only those bigrams for collocations which have $|t|$ score greater than 2.57 is accurate as far as we take the $t$-test into consideration but extraction of collocation is not a pure mathematical job since the decision as to what constitutes a collocation is affected by its linguistic aspect also. This provides us a reason to think about those word combinations whose $t$ scores are less than 2.57 but very close to it. Therefore it opens a way to make a fuzzy set for collocations based on $t$ scores. The membership function for a word combination x can be defined as follows:

$$A_t(x) = \begin{cases} 1 & if \quad |t| \geq 2.57 \\ \dfrac{|t|}{2.57} & otherwise \end{cases} \qquad \text{……………………(4.3)}$$

while the grades after application of t-score can be depicted as demonstrated in following figure:

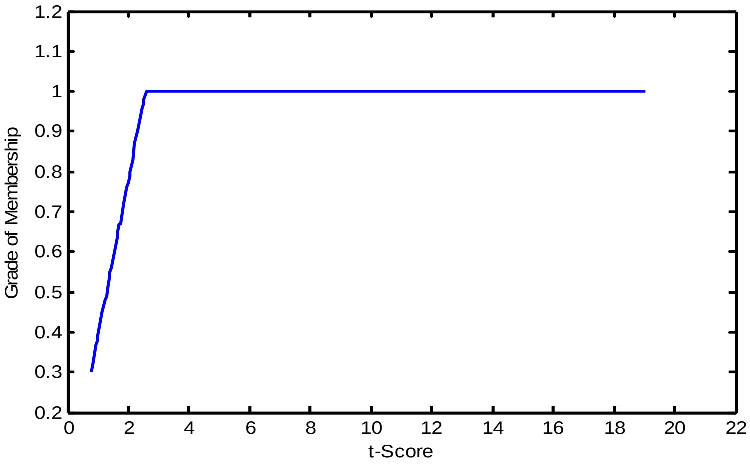

Fig.4.2 Grade of membership for collocation using t-score

We have defined two fuzzy decision models for collocation extraction by aggregating the two defined fuzzy sets in two different ways.

### 4.2.3 Model I

Decision regarding word combination for being collocation can be made possible by using fuzzy decision making. In the first paper on fuzzy decision making, Bellman and Zadeh (1970) suggested a fuzzy model of decision making in which relevant goals and constraints are expressed in terms of fuzzy sets and a decision is determined by an appropriate aggregation of these fuzzy sets.

For collocation extraction we can characterize the decision model as follows:

- A set $A$ of word combinations.
- A fuzzy set $A_I$ describing the grade of membership of each word combination for high mutual information. High mutual information has been taken as the goal.

- A fuzzy set $A_t$ describing the grade of membership of each word combination for a reasonable t-score. The t- score has been taken as constraint.
- A fuzzy decision set $D$ describing the grade of membership of each word combination for being collocation. The fuzzy decision set $D$ is conceived as a fuzzy set on $A$ that simultaneously satisfies the given goal and constraint.
- For a word combination $x$, the fuzzy decision $D(x)$ can be defined as $D(x) = \min[A_I(x), A_t(x)]$

### 4.2.4 Model II

In this model different opinions have been aggregated to find the fuzzy set of collocation. We have used the most common method based on the probabilistic interpretation of membership functions given by multiple experts Klir et al (2001). In this process, different experts are asked to evaluate some $x \in X$ for its belongingness to $A$, where $A$ is a fuzzy set on $X$ that represents a linguistic term associated with a given linguistic variable.

For some $x \in X$, let $A_r(x)$ denote the answer of the $r^{th}$ expert $(r \in N)$ in the term of belongingness of $x$ to $A$, where $A_r(x)$ has only two values 0 and 1 for $x \in A$ and $x \notin A$ respectively. Then the membership function can be defined as follows:

$$A(x) = \frac{\sum_{r=1}^{n} A_r(x)}{n}$$, where n is the number of experts.

The collocation extraction model can be characterized as follows:

- A set $A$ of word combinations.

- A fuzzy set $A_I$ describing the grade of membership of each word combination for high mutual information. Mutual information has been taken as one expert.
- A fuzzy set $A_t$ describing the grade of membership of each word combination for a reasonable t-score. The t- score has been taken as second expert.
- For a word combination $x$, $A_I(x)$ and $A_t(x)$ be the opinions of the mutual information score and t- score respectively, in terms of grade of membership of x to its belongingness to $A$.
- A fuzzy set $A$ describing the grade of membership of each word combination for being collocation.
- For a word combination $x$, The grade of membership can be given as $A(x) = \dfrac{A_I(x) + A_t(x)}{2}$

On using the above fuzzy decision models, we can obtain the grade of membership of each word combination for being collocation. A word combination that attains a maximum grade of membership can be taken as collocation.

## 4.3 Fuzzy Inference System for Collocation extraction

Fuzzy logic is a logical system which is very easy to understand. Word combinations getting scored by mutual information and t-score are evaluated by the rules of fuzzy inference system (FIS). The rules that are used to determine relevance of a word combination, come from the reasoning behind the existing techniques (e.g., if the mutual information score is high, then the word combination is highly relevant to be considered for collocation; if the t-score is significant, then word combination is relevant to be considered for collocation, etc.). Fuzzy logic expresses relevance as degrees of memberships (e.g., word combination could have a relevance measure with the following degrees of membership: 0.9 highly relevant and 0.5 reasonably relevant and 0.1 not relevant).

Matlab Fuzzy Logic Toolbox provides an opportunity to look at all the components of FIS. It allows modifications, examination and verification of the effects of changes. Here we discuss the components of FIS.

### 4.3.1 Baseline model

The collocation extraction Fuzzy inference system (CE-FIS) is based on the two existing techniques of collocation extraction, i.e., Mutual information and t-test which are the input variables for CE-FIS. Grade of membership is the out put variable.

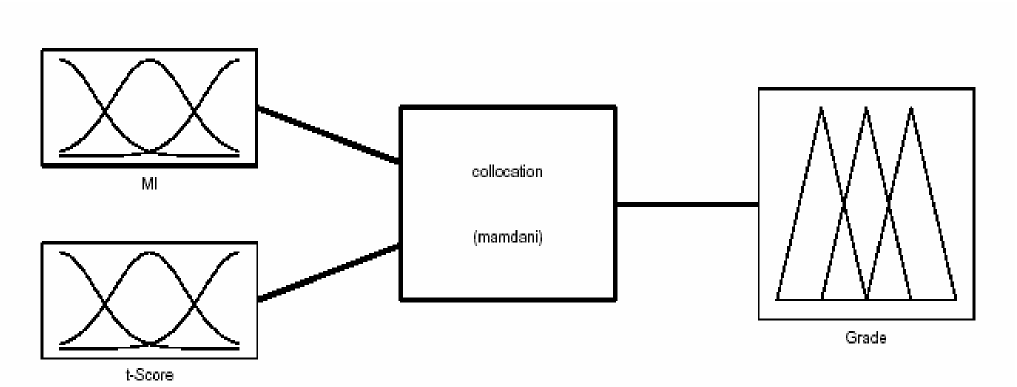

Fig 4.3 Fuzzy inference system for collocation extraction

### 4.3.2 Fuzzy sets / Membership Function

Every input variable can be defined using two or three fuzzy sets. A membership function gives mathematical meaning to the linguistic variable such as high mutual information, low mutual information. Mutual information can be defined using three fuzzy sets associated with each linguistic variable: high, average and low and t-score can be defined using two fuzzy sets associated with each variable: significant and non-significant. Output variable (relevance of a word combination for being collocation) can also be defined through three fuzzy sets namely: high, average and low. A membership function is a curve that defines how each point in the input/output space is mapped to degree of membership of fuzzy set. There are various inbuilt membership functions in fuzzy logic tool box. We have selected Pi shaped built-in membership function with different parameters for each input and output variable on account of its suitability.

Syntax of Pi shaped built-in membership is:

y = pimf (x, [a b c d] )

This spline-based curve is so named because of its Π shape. This membership function is evaluated at the points determined by the vector x. The parameters a and d locate the "feet" of the curve, while b and c locate its "shoulders." Table 4.2 shows the chosen values of parameters 'a', 'b', 'c', 'd' for different membership functions.

Table 4.2 Values of parameters for different membership functions

| Membership function | value pf parameters | | | |
|---|---|---|---|---|
| | a | b | c | d |
| Low mutual information | -5.4 | -0.6 | 3 | 5.889 |
| Average mutual information | 3.002 | 5.922 | 6.175 | 9.092 |
| High mutual information | 6.143 | 8.94 | 13.1 | 17.9 |
| Non-significant t-score | -10.9 | -1.26 | -0.0159 | 2.778 |
| Significant t-score | -0.0476 | 2.68 | 14 | 23.6 |
| Low relevance | -0.45 | -0.05 | 0.184 | 0.5013 |
| Average relevance | 0.192 | 0.491 | 0.515 | 0.7976 |
| High relevance | 0.504 | 0.803 | 1.05 | 1.45 |

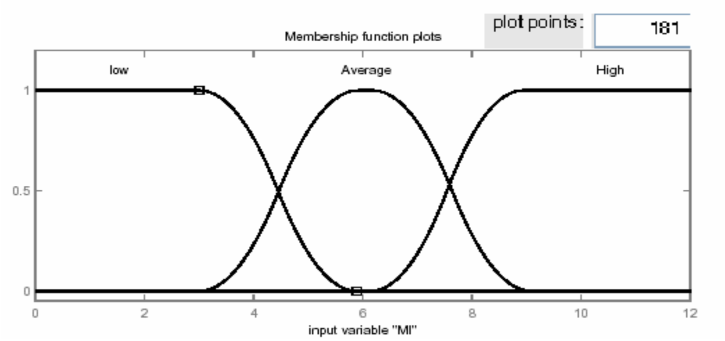

Fig.4.4 Membership functions for Mutual Information scores

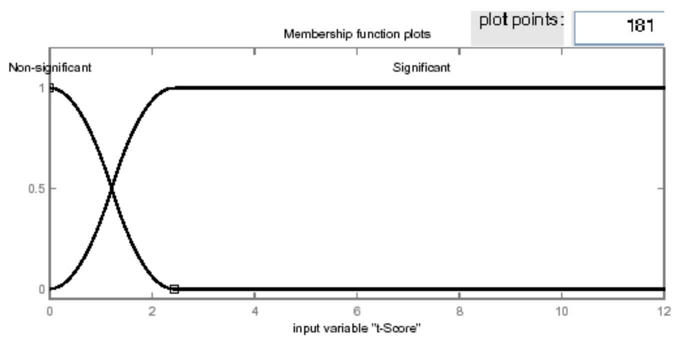

Fig.4.5. Membership functions for t-scores

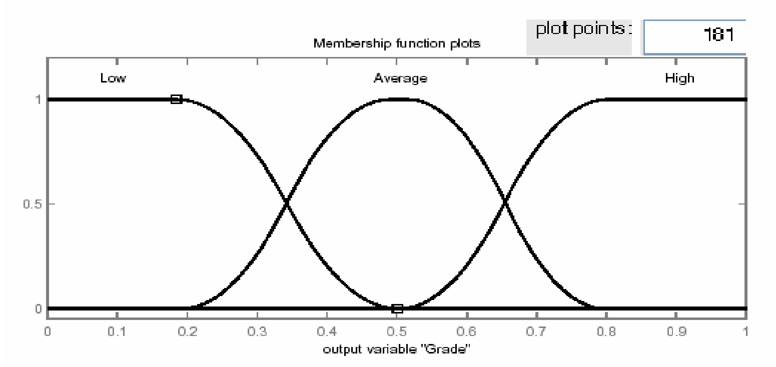

Fig.4.6 Membership functions for relevance

### 4.3.4 Rules

Matlab fuzzy toolbox allows defining rules by taking different fuzzy sets of input and output variables. Rules can be derived by simple reasoning of mutual information and t-score. High mutual information shows presence of a collocation and for a particular significance level, a word combination for which t-score is significant, can be considered for collocation. Following rules have been adopted for the CE-FIS.

- If (MI is low) and (t-score is non-significant), then (Relevance is low)
- If (MI is high) and (t-score is significant), then (Relevance is high)
- If (MI is low) and (t-score is significant), then (Relevance is average)
- If (MI is high) and (t-score is non-significant), then (Relevance is average)

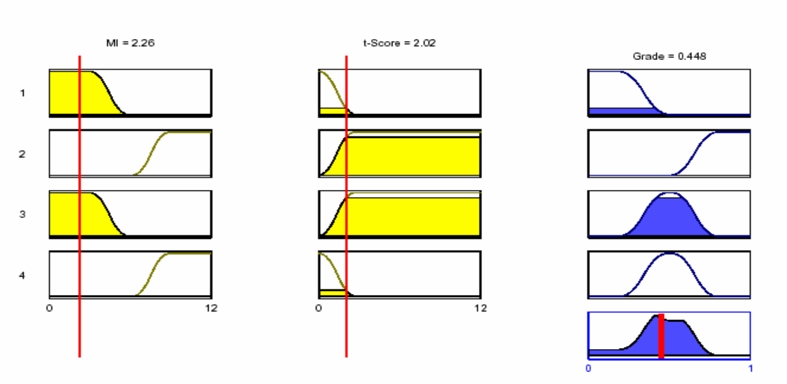

Fig. 4.7 Fuzzy Inference process

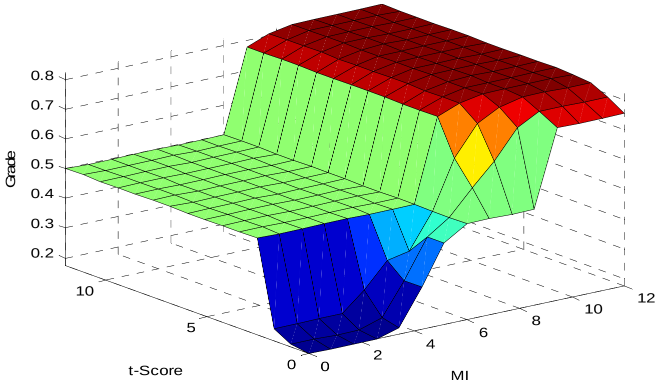

Fig. 4.8 Fuzzy rule surface

We have taken the example of seventy word pairs from the compiled corpus. Table 4.3 shows the mutual information scores and their corresponding grades of membership for different word combinations.

Table 4.3**:** Mutual Information, t-score and the respective grades of membership for different word combinations.

| $W_1$ | $W_2$ | $f(w_1)$ | $f(w_2)$ | $f(w_1 w_2)$ | $MI$ | $A_i$ | $\|t\|$ | $A_t$ |
|---|---|---|---|---|---|---|---|---|
| christmas | eve | 72 | 33 | 9 | 11.96 | 0.95 | 3.00 | 1.00 |
| base | camp | 54 | 55 | 7 | 11.27 | 0.95 | 2.64 | 1.00 |
| public | opinion | 190 | 103 | 10 | 9.07 | 0.94 | 3.16 | 1.00 |
| both | sides | 409 | 69 | 11 | 8.68 | 0.93 | 3.31 | 1.00 |
| human | being | 182 | 735 | 30 | 7.88 | 0.93 | 5.45 | 1.00 |
| great | deal | 911 | 118 | 20 | 7.61 | 0.92 | 4.45 | 1.00 |
| human | nature | 182 | 251 | 7 | 7.33 | 0.92 | 2.63 | 1.00 |
| take | care | 808 | 228 | 20 | 6.83 | 0.91 | 4.43 | 1.00 |
| young | man | 741 | 2138 | 147 | 6.61 | 0.91 | 12.00 | 1.00 |
| early | days | 182 | 497 | 7 | 6.35 | 0.91 | 2.61 | 1.00 |
| long | journey | 967 | 100 | 7 | 6.25 | 0.91 | 2.61 | 1.00 |
| last | night | 846 | 856 | 47 | 6.09 | 0.90 | 6.76 | 1.00 |
| fire | bucket | 291 | 15 | 5 | 10.23 | 0.94 | 2.23* | 0.87 |
| away | from | 862 | 3945 | 135 | 5.38 | 0.89 | 11.34 | 1.00 |
| came | along | 1360 | 367 | 13 | 4.77 | 0.88 | 3.47 | 1.00 |
| every | night | 676 | 856 | 13 | 4.56 | 0.87 | 3.45 | 1.00 |
| trench | life | 99 | 1102 | 6 | 5.85 | 0.90 | 2.41* | 0.94 |
| night | before | 856 | 1164 | 18 | 4.25 | 0.86 | 4.02 | 1.00 |
| must | take | 1144 | 808 | 16 | 4.18 | 0.86 | 3.78 | 1.00 |
| last | time | 846 | 1463 | 20 | 4.09 | 0.85 | 4.21 | 1.00 |
| strong | man | 172 | 2183 | 7 | 4.29 | 0.86 | 2.51* | 0.98 |
| long | after | 967 | 1304 | 18 | 3.91 | 0.85 | 3.96 | 1.00 |
| might | even | 1143 | 768 | 10 | 3.58 | 0.83 | 2.90 | 1.00 |
| come | over | 1358 | 1394 | 19 | 3.40 | 0.83 | 3.95 | 1.00 |
| look | upon | 756 | 1913 | 12 | 3.12 | 0.81 | 3.07 | 1.00 |
| little | episode | 1630 | 16 | 4 | 7.33 | 0.92 | 1.99* | 0.77 |
| almost | every | 518 | 676 | 6 | 4.17 | 0.86 | 2.31* | 0.90 |
| evil | eye | 124 | 259 | 4 | 7.03 | 0.92 | 1.98* | 0.77 |
| time | before | 1463 | 1164 | 11 | 2.76 | 0.79 | 2.83 | 1.00 |
| long | way | 967 | 1084 | 8 | 3.00 | 0.80 | 2.48* | 0.96 |
| painful | experience | 27 | 87 | 3 | 10.39 | 0.95 | 1.73* | 0.67 |
| make | use | 963 | 222 | 5 | 4.62 | 0.87 | 2.15* | 0.83 |
| very | like | 1410 | 1602 | 11 | 2.36 | 0.76 | 2.67 | 1.00 |
| dark | shadow | 279 | 93 | 3 | 6.92 | 0.92 | 1.72* | 0.67 |
| last | century | 846 | 32 | 3 | 6.86 | 0.91 | 1.72* | 0.67 |
| cheerful | noise | 48 | 720 | 3 | 6.51 | 0.91 | 1.71* | 0.67 |

| $W_1$ | $W_2$ | $f(w_1)$ | $f(w_2)$ | $f(w_1w_2)$ | $MI$ | $A_i$ | $\|t\|$ | $A_t$ |
|---|---|---|---|---|---|---|---|---|
| only | because | 1187 | 371 | 5 | 3.58 | 0.83 | 2.05* | 0.80 |
| your | book | 2888 | 153 | 5 | 3.57 | 0.83 | 2.05* | 0.80 |
| another | half | 693 | 696 | 5 | 3.45 | 0.83 | 2.03* | 0.79 |
| welcome | relief | 70 | 71 | 2 | 8.72 | 0.93 | 1.41* | 0.55 |
| national | guard | 39 | 163 | 2 | 8.37 | 0.93 | 1.41* | 0.55 |
| good | terms | 1299 | 88 | 3 | 4.79 | 0.88 | 1.67* | 0.65 |
| horrible | thing | 55 | 580 | 2 | 6.04 | 0.90 | 1.39* | 0.54 |
| stark | madness | 6 | 22 | 1 | 12.96 | 0.96 | 1.00* | 0.39 |
| usual | hour | 115 | 344 | 2 | 5.73 | 0.90 | 1.39* | 0.54 |
| step | towards | 135 | 348 | 2 | 5.48 | 0.89 | 1.38* | 0.54 |
| away | down | 862 | 1517 | 6 | 2.27 | 0.75 | 1.94* | 0.76 |
| valid | reason | 6 | 151 | 1 | 10.18 | 0.94 | 1.00* | 0.39 |
| might | still | 1143 | 799 | 5 | 2.52 | 0.77 | 1.85* | 0.72 |
| spiritual | creature | 15 | 62 | 1 | 10.14 | 0.94 | 1.00* | 0.39 |
| rapid | motion | 32 | 42 | 1 | 9.61 | 0.94 | 1.00* | 0.39 |
| clumsy | fashion | 11 | 126 | 1 | 9.57 | 0.94 | 1.00* | 0.39 |
| like | myself | 1602 | 372 | 4 | 2.82 | 0.79 | 1.72* | 0.67 |
| visible | effort | 27 | 73 | 1 | 9.06 | 0.94 | 1.00* | 0.39 |
| empty | tent | 128 | 16 | 1 | 9.00 | 0.94 | 1.00* | 0.39 |
| huge | space | 35 | 64 | 1 | 8.87 | 0.93 | 1.00* | 0.39 |
| peasant | girl | 10 | 253 | 1 | 8.70 | 0.93 | 1.00* | 0.39 |
| wild | dreams | 136 | 46 | 1 | 7.39 | 0.92 | 0.99* | 0.39 |
| except | myself | 81 | 1602 | 2 | 4.02 | 0.85 | 1.33* | 0.52 |
| wrong | way | 121 | 1084 | 2 | 4.00 | 0.85 | 1.33* | 0.52 |
| round | upon | 387 | 1913 | 4 | 2.51 | 0.77 | 1.65* | 0.64 |
| last | link | 864 | 15 | 1 | 6.34 | 0.91 | 0.99* | 0.38 |
| looking | through | 431 | 1010 | 3 | 2.86 | 0.79 | 1.49* | 0.58 |
| human | affairs | 182 | 86 | 1 | 6.07 | 0.90 | 0.99* | 0.38 |
| water | level | 286 | 65 | 1 | 5.82 | 0.90 | 0.98* | 0.38 |
| most | powerful | 723 | 33 | 1 | 5.46 | 0.89 | 0.98* | 0.38 |
| little | chap | 1630 | 132 | 2 | 3.29 | 0.82 | 1.27* | 0.49 |
| weary | night | 46 | 856 | 1 | 4.74 | 0.87 | 0.96* | 0.37 |
| great | emotions | 911 | 45 | 1 | 4.68 | 0.87 | 0.96* | 0.37 |
| make | sense | 963 | 192 | 1 | 2.51 | 0.77 | 0.82* | 0.32 |

* Rejected for collocation ( $|t| < 2.57$ )

## 4.4 Evaluation of the fuzzy models

For evaluating the fuzzy set theoretic approach, we have taken the example of seventy word pairs from the compiled corpus. Mutual information scores, t-scores and their corresponding grades of membership have been calculated for different word combinations. Table 4.3 shows the frequencies $f(w_1)$, $f(w_2)$, $f(w_1 w_2)$ of words $w_1, w_2$ and their combination $w_1 w_2$ respectively with their respective mutual information score, t-score and grades of membership. Table 4.4 and 4.5 show the grades of membership for different bigrams in the text using model I and II respectively. Different grades of membership have been chosen to compare the results of proposed models with mutual information and t-score.

Assistance from www.thefreedictionary.com and “Cambridge Advanced Lerner’s Dictionary” has been taken to check the validity of word pairs. Only 16 word pairs (star marked in table 4.4 & 4.5) have been found as meaningful collocations. Results of mutual information and t-score have been compared with the results obtained by model I and Model II. Table 4.7 shows the precision and recall for different mutual information scores. For t-score precision is 42% and recall is 63%. Table 4.8 and 4.9 show the precision and recall of the proposed models.

From table 4.7, 4.8, 4.9 and 4.10, we can observe that fuzzy set theoretical models based on mutual information and t-score provide a better opportunity to extract collocations than using mutual information and t-score alone. Particularly at more than or equal to .90 and .95 grade of membership both the models have shown a good result. Model I and II have almost same results except the change of scale.

Table 4.4 Grades of membership for different word combinations using model I.

| $w_1$ | $w_2$ | $D$ | $w_1$ | $w_2$ | $D$ |
|---|---|---|---|---|---|
| *christmas | eve | 0.95 | away | down | 0.75 |
| *base | camp | 0.95 | might | still | 0.72 |
| *public | opinion | 0.94 | painful | experience | 0.67 |
| *both | sides | 0.93 | dark | shadow | 0.67 |
| *human | being | 0.93 | last | century | 0.67 |
| *great | deal | 0.92 | cheerful | noise | 0.67 |
| *human | nature | 0.92 | like | myself | 0.67 |
| *take | care | 0.91 | good | terms | 0.65 |
| *young | man | 0.91 | round | upon | 0.64 |
| *early | days | 0.91 | looking | through | 0.58 |
| long | journey | 0.91 | welcome | relief | 0.55 |
| last | night | 0.90 | *national | guard | 0.55 |
| trench | life | 0.90 | horrible | thing | 0.54 |
| away | from | 0.89 | usual | hour | 0.54 |
| came | along | 0.88 | step | towards | 0.54 |
| *fire | bucket | 0.87 | except | myself | 0.52 |
| every | night | 0.87 | wrong | way | 0.52 |
| night | before | 0.86 | little | chap | 0.49 |
| must | take | 0.86 | stark | madness | 0.39 |
| strong | man | 0.86 | valid | reason | 0.39 |
| almost | every | 0.86 | spiritual | creature | 0.39 |
| last | time | 0.85 | rapid | motion | 0.39 |
| long | after | 0.85 | clumsy | fashion | 0.39 |
| might | even | 0.83 | visible | effort | 0.39 |
| come | over | 0.83 | empty | tent | 0.39 |
| make | use | 0.83 | huge | space | 0.39 |
| *look | upon | 0.81 | peasant | girl | 0.39 |
| long | way | 0.80 | wild | dreams | 0.39 |
| only | because | 0.80 | last | link | 0.38 |
| your | book | 0.80 | human | affairs | 0.38 |
| time | before | 0.79 | *water | level | 0.38 |
| another | half | 0.79 | most | powerful | 0.38 |
| little | episode | 0.77 | weary | night | 0.37 |
| *evil | eye | 0.77 | great | emotions | 0.37 |
| very | like | 0.76 | *make | sense | 0.32 |

*actual collocation

Table 4.5: Grades of membership for different word combinations using model II.

| $w_1$ | $w_2$ | $A$ | $w_1$ | $w_2$ | $A$ |
|---|---|---|---|---|---|
| *christmas | eve | 0.98 | another | half | 0.81 |
| *base | camp | 0.98 | dark | shadow | 0.79 |
| *public | opinion | 0.97 | last | century | 0.79 |
| *both | sides | 0.97 | cheerful | noise | 0.79 |
| *human | being | 0.96 | good | terms | 0.76 |
| *great | deal | 0.96 | away | down | 0.75 |
| *human | nature | 0.96 | might | still | 0.74 |
| *take | care | 0.96 | welcome | relief | 0.74 |
| *young | man | 0.96 | *national | guard | 0.74 |
| *early | days | 0.95 | like | myself | 0.73 |
| long | journey | 0.95 | horrible | thing | 0.72 |
| last | night | 0.95 | usual | hour | 0.72 |
| away | from | 0.94 | step | towards | 0.71 |
| came | along | 0.94 | round | upon | 0.71 |
| every | night | 0.93 | looking | through | 0.69 |
| night | before | 0.93 | except | myself | 0.68 |
| must | take | 0.93 | wrong | way | 0.68 |
| last | time | 0.93 | stark | madness | 0.67 |
| long | after | 0.92 | valid | reason | 0.67 |
| strong | man | 0.92 | spiritual | creature | 0.67 |
| trench | life | 0.92 | rapid | motion | 0.66 |
| might | even | 0.92 | clumsy | fashion | 0.66 |
| come | over | 0.91 | visible | effort | 0.66 |
| *fire | bucket | 0.91 | empty | tent | 0.66 |
| *look | upon | 0.91 | huge | space | 0.66 |
| time | before | 0.89 | peasant | girl | 0.66 |
| long | way | 0.88 | little | chap | 0.66 |
| almost | every | 0.88 | wild | dreams | 0.65 |
| very | like | 0.88 | last | link | 0.65 |
| make | use | 0.85 | human | affairs | 0.64 |
| little | episode | 0.85 | *water | level | 0.64 |
| *evil | eye | 0.84 | most | powerful | 0.64 |
| only | because | 0.82 | weary | night | 0.62 |
| your | book | 0.82 | great | emotions | 0.62 |
| painful | experience | 0.81 | *make | sense | 0.54 |

*actual collocation

Table 4.6: Grades of membership for different word combinations using CE-FIS

| $w_1$ | $w_2$ | $D$ | $w_1$ | $w_2$ | $D$ |
|---|---|---|---|---|---|
| *christmas | eve | 0.82 | come | over | 0.50 |
| *public | opinion | 0.82 | long | after | 0.50 |
| *both | sides | 0.82 | last | time | 0.50 |
| *base | camp | 0.82 | must | take | 0.50 |
| *human | being | 0.81 | night | before | 0.50 |
| *great | deal | 0.80 | every | night | 0.50 |
| *human | nature | 0.79 | came | along | 0.50 |
| *fire | bucket | 0.79 | away | from | 0.50 |
| *take | care | 0.78 | long | way | 0.49 |
| *young | man | 0.77 | strong | man | 0.49 |
| painful | experience | 0.71 | almost | every | 0.47 |
| little | episode | 0.69 | night | air | 0.45 |
| welcome | relief | 0.66 | only | because | 0.45 |
| *national | guard | 0.66 | your | book | 0.45 |
| *evil | eye | 0.65 | another | half | 0.45 |
| *early | days | 0.64 | away | down | 0.44 |
| dark | shadow | 0.63 | make | use | 0.44 |
| last | century | 0.63 | like | myself | 0.41 |
| cheerful | noise | 0.62 | round | upon | 0.40 |
| wild | dreams | 0.62 | trench | life | 0.38 |
| last | link | 0.62 | *water | level | 0.38 |
| long | journey | 0.62 | usual | hour | 0.38 |
| stark | madness | 0.59 | step | towards | 0.37 |
| valid | reason | 0.59 | most | powerful | 0.37 |
| spiritual | creature | 0.59 | looking | through | 0.37 |
| rapid | motion | 0.59 | along | over | 0.37 |
| clumsy | fashion | 0.59 | good | terms | 0.36 |
| visible | effort | 0.59 | weary | night | 0.35 |
| empty | tent | 0.59 | except | myself | 0.34 |
| huge | space | 0.59 | wrong | way | 0.34 |
| peasant | girl | 0.59 | great | emotions | 0.34 |
| last | night | 0.50 | little | chap | 0.33 |
| horrible | thing | 0.50 | things | behind | 0.33 |
| human | affairs | 0.50 | very | dark | 0.31 |
| *look | upon | 0.50 | *make | sense | 0.26 |

*actual collocation

Table 4.7 Precision and recall for Mutual Information.

| Mutual Information<br>( equal to or more than) | Precision<br>(in %) | Recall<br>( in %) |
|---|---|---|
| 10.0 | 43 | 19 |
| 9.0 | 33 | 25 |
| 8.0 | 35 | 38 |
| 7.0 | 43 | 63 |

Table 4.8 Precision and recall for model I

| Grade of membership<br>( equal to or more than) | Precision<br>(in %) | Recall<br>( in %) |
|---|---|---|
| 0.95 | 100 | 13 |
| **0.90** | **77** | **63** |
| 0.85 | 48 | 69 |
| 0.80 | 40 | 75 |

Table 4.9 :Precision and recall for model II

| Grade of membership<br>( equal to or more than) | Precision<br>(in %) | Recall<br>( in %) |
|---|---|---|
| 0.98 | 100 | 12 |
| **0.95** | **83** | **63** |
| 0.90 | 48 | 75 |
| 0.85 | 39 | 75 |

Table 4.10 Precision and recall for CE-FIS

| Grade of membership<br>( equal to or more than) | Precision<br>(in %) | Recall<br>( in %) |
|---|---|---|
| 0.80 | 100 | 38 |
| **0.70** | **91** | **63** |
| 0.60 | 59 | 81 |

## 4.5 Discussion

The present work was carried out to utilize the fuzzy inference system for collocation extraction. The previous two techniques were deterministic crisp formulae and it is difficult to make a decision about something which is vague and uncertain with deterministic crisp formulas. Fuzzy logic is based on the theory of fuzzy set which includes the elements with a grade of membership. Fuzzy logic for collocation extraction provides the benefits of the previous two approaches while overcoming their drawbacks. A close look at tables 4.4 and 4.5, helps us in drawing the inference that some word combinations have high mutual information score (e.g., Stark madness) but have low grade of membership and are therefore rejected by t-test (e.g., fire bucket) but high grade of membership. Table 4.10 shows the precision and recall of CE-FIS and it is very high in comparison to the two tests. Word combinations falling in the category of grade of membership more than 70 show high relevance to the set of collocations.

Finally, it can be said that the previous two methods, the t-test and Mutual information were not showing good results when they were being used separately. Fuzzy approach combines the two methods and the proposed way of the combination of the two methods provides better results than their individual results.

## Chapter V

# Matrix Representation of Words

*In the present chapter we have generated word matrices by defining juxtaposition and elementary contractions for matrices of letters of an alphabet. The definitions paving the way for word matrix have been illustrated by pertinent examples. Subword occurrence has been calculated using the word matrix and some algebraic properties of word matrices have been proved.*

### 5.1 Introduction

In order to facilitate the inroads of vectors and matrices into the arena of words and languages, Parikh matrix mapping was introduced by Mateescu et al (2001) in the spirit of classical Parikh mapping (1966). Extension of Parikh matrix mapping, induced by word over the alphabet can be seen in the work of Serbanuta (2004). The aspect of injectivity in Parikh matrix mapping has been discussed by Atanasiu et al (2001). We are adopting entirely different approach of selecting such a matrix which could have injectivity as an inbuilt facility and at the same time could generate Parikh vector. We have defined word matrix mapping by considering matrices having $k$ $(k \geq 1)$ rows and finite number of columns depending on the number of letters in the alphabet and number of syllables in a word respectively. The problems of subword occurrences have been also derived with the help of assured word matrix mapping. Elementary contractions and juxtaposition of word matrices have been defined to get reduced and expanded word matrix respectively.

### 5.1.1. Preliminaries

We have begun with some basic notations and definitions. An alphabet $\Sigma$ is a set of symbols called letters. A word over $\Sigma$ is a finite sequence of elements of $\Sigma$. Let $\Sigma = \{a, b\}$, then $ab, abb, aab$ etc are words over the alphabet $\Sigma$. By a syllable we mean a symbol $a^n$, where $a$ is a letter of the alphabet $\Sigma$ and $n$ is an integer. Words are also defined as finite ordered sequence of syllables, as mentioned in the book of Crowell et al (1977). An empty word is a neutral element denoted by $\lambda$ or 1. The set of all the words over $\Sigma$ is denoted by $\Sigma^*$. In case if $u = w_1 a^0 w_2$ and $v = w_1 w_2$, where $w_1$ and $w_2$ are words over $\Sigma$, the word $v$ is obtained from the word $u$ by an elementary contraction of type I or the word $u$ is obtained from the word $v$ by an elementary expansion of type I. When $u = w_1 a^p a^q w_2$ and $v = w_1 a^{p+q} w_2$, where $w_1$ and $w_2$ are words over $\Sigma$, then $v$ is obtained from the word $u$ by an elementary contraction of type II or the word $u$ is obtained from the word $v$ by an elementary expansion of type II. The juxtaposition or concatenation of two words $w_1$ and $w_2$ is a word $w_1 w_2$ obtained by concatenating the two words $w_1$ and $w_2$. For example if $w_1 = ab$ and $w_2 = aab$, then juxtaposition of the two words $w_1$ and $w_2$ is expressed as $w_1 w_2 = abaab$. The set of all words over alphabet $\Sigma$ with a binary operation 'juxtaposition' and neutral element 1 is a free monoid as given in Lothaire M. (1997). The inverse $w^{-1}$ of a word $w$ is obtained by reversing the order of its syllables and changing the sign of the exponent of each syllable such that $ww^{-1} = 1$. The monoid $\Sigma^*$ is a free group on the alphabet $\Sigma$ as the inverse of each element exists in it.

In this chapter we have used ordered alphabets. An ordered alphabet is an alphabet $\Sigma = \{a_1 < a_2 < a_3 .......... < a_k\}$ with a relation of order ("<") on it. Let $\Sigma$ be an alphabet and $u, v \in \Sigma^*$. We say that $u$ is scattered subword (or simply subword) of $v$ if $v$, as a sequence of letters, contains $u$ as a subsequence. Formally it means that there exist words $x_1, x_2, ........, x_n$ and $y_0, y_1, ........y_n$ in $\Sigma^*$, some of them possibly empty, such that $u = x_1 x_2 ........ x_n$ and $v = y_0 x_1 y_1 x_2 ........ x_n y_n$. We will denote the number of occurrences of word $u$ as a subword in $v$ by $|w|_u$. For instance,

$$|abb|_{ab} = 2, \ |abbcc|_{abc} = 4 \quad \text{and} \quad |abba|_{ba} = 2.$$

Let $\Sigma = \{a_1 < a_2 < a_3 .......... < a_k\}$ be an ordered alphabet and $w \in \Sigma^*$. The Parikh mapping $\varphi : \Sigma^* \to N^k$, is defined by $\varphi(w) = \left(|w|_{a_1,} |w|_{a_2}, ........, |w|_{a_k}\right)$ and the Parikh vector of $w$ is $\left(|w|_{a_1,} |w|_{a_2}, ........, |w|_{a_k}\right)$. The mirror image of a word $w \in \Sigma^*$, denoted by $mi(w)$, is defined as: $mi(\lambda) = \lambda$ and $mi(b_1 b_2 ...... b_n) = b_n ........ b_2 b_1$ where $b_i \in \Sigma, 1 \le i \le n$.

## 5.2 Generation of Word matrix

In our present study $M_k$ has been used to denote the set of matrices having $k$ rows and finite number of columns, which have at most one non- zero entry in each column and other entries being zero. Word matrix mapping has been defined with the help of the following definitions.

**Definition 5.2.1** Let $\Sigma = \{a_1 < a_2 < a_3 .......... < a_k\}$ be an ordered alphabet, where $k \ge 1$ and $\Sigma' = \{a_i{}^p : \ a_i \in \Sigma \ and \ \ p \in I\}$ be the set of syllables over the alphabet $\Sigma$. The mapping $\xi^{m_k} : \Sigma' \to M_k$ may be defined as:

$$If \quad \xi^{m_k}(a_l{}^p) = [m_{i\,j}]_{k \times 1}, \ where \ \ p \in I, \ \ then$$

$$m_{i\,j} = \begin{cases} p & for \quad i = l \\ 0 & otherwise \end{cases}$$

To illustrate the point we can cite the following examples.

**Example 5.2.2** Let $\Sigma = \{a < b\}$, then $\xi^{m_2}\left(a^2\right) = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$ and $\xi^{m_2}\left(b^2\right) = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$.

From definition 5.2.1, it can be observed that the matrix of a syllable over an alphabet having cardinality $k$ will contain $k$ rows and one column.

**Definition 5.2.3** Let $\Sigma = \{a_1 < a_2 < a_3 .......... < a_k\}$ be an ordered alphabet and $\xi^{m_k}(a_{l_1}{}^{p_1}) = [a_{i\,j}]_{k\times 1}$ and $\xi^{m_k}(a_{l_2}{}^{p_2}) = [b_{i\,j}]_{k\times 1}$ be the matrices of two syllables. The composition of juxtaposition on the matrices of syllables can be defined as:

$$\xi^{m_k}(a_{l_1}{}^{p_1})\ \xi^{m_k}(a_{l_2}{}^{p_2}) = [c_{i\,j}]_{k\times 2}, \quad \text{where}$$

$$c_{i\,j} = \begin{cases} a_{i\,j} & for \quad j = 1 \\ b_{i\,j} & for \quad j = 2 \end{cases}$$

**Example 5.2.4** Let $\Sigma = \{a < b\}$, then $\xi^{m_2}\left(a^2\right)\xi^{m_2}\left(b^3\right)\xi^{m_2}\left(a\right) = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 3 & 0 \end{bmatrix}$

Finally with the help of above two definitions, we can define word matrix mapping.

**Definition 5.2.5** The word matrix mapping shall be the morphism $\xi^{m_k} : \Sigma^* \to M_k$, defined as follows:

$$If \quad \xi^{m_k}(a_l{}^p) = [m_{i\,j}]_{k\times 1},\ where\ \ p \in I,\ \ then$$

$$m_{i\,j} = \begin{cases} p & for \quad i = l \\ 0 & otherwise \end{cases} \qquad \text{and}$$

$$\xi^{m_k}(a_{l_1}{}^{p_1} a_{l_2}{}^{p_2} ................ a_{l_n}{}^{p_n}) = \xi^{m_k}(a_{l_1}{}^{p_1})\xi^{m_k}(a_{l_2}{}^{p_2}) ................ \xi^{m_k}(a_{l_n}{}^{p_n})$$

**Example 5.2.6** Let $\Sigma = \{a < b < c\}$be the ordered alphabet and assume that $w = ab^2c$. Then $\xi^{m_3}(w)$is a matrix of order $3 \times 3$, which can be computed as follows:

$$\xi^{m_3}(ab^2c) = \xi^{m_3}(a)\xi^{m_3}(b^2)\xi^{m_3}(c)$$
$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

**Example 5.2.7** Let $\Sigma = \{a < b\}$**,** be the ordered alphabet and $w = ab^2a^2$, then

$$\xi^{m_k}(ab^2a^2) = \xi^{m_k}(a)\ \xi^{m_k}(b^2)\xi^{m_k}(a^2)$$
$$= \begin{bmatrix} 1 & 0 & 2 \\ 0 & 2 & 0 \end{bmatrix}$$

For defining elementary contractions of type I and II in a matrix of a word, we suppose that let $\Sigma = \{a_1 < a_2 < a_3 .......... < a_k\}$ be an ordered alphabet, where $k \geq 1$ and further assume that $[m_{ij}]_{k \times n}$ be the matrix before the applications of elementary contractions and $[m'_{ij}]_{k \times n-1}$ be the matrix of the same word after applying elementary contractions at $r^{th}$ column.

**Definition 5.2.8** Elementary contraction of type I for matrix of a word can be defined as follows:

$$\begin{array}{lll} \textit{If} & m_{ir} = 0, & \textit{for any } r, \quad 1 \leq r \leq n, \ 1 \leq i \leq k \\ \textit{then}, & m'_{ij} = & \begin{cases} a_{ij} & \textit{for} \quad 1 \leq i \leq k, 1 \leq j < r \\ a_{ij+1} & \textit{for} \quad 1 \leq i \leq k, \ j \geq r \end{cases} \end{array}$$

**Example 5.2.9** Let $w = ab^0c$ be a word generated on alphabet $A = \{a, b, c\}$. Then the order of the matrix of the word $w = ab^0c$ will be $3 \times 3$ and the matrix will be given by $\xi^{m_3}(w)=\begin{bmatrix}1&0&0\\0&0&0\\0&0&1\end{bmatrix}$. On using elementary contraction of type I at 2nd column, the order of the matrix of the reduced word will be $3\times 2$ and $m'_{ij}=\begin{cases}m_{ij} & for\ 1\le j<2,\ 1\le i\le 3\\ m_{i\,j+1} & for\ \ j\ge 2,\ 1\le i\le 3\end{cases}$.

The matrix of the reduced word will be $\xi^{m_3}(w)=\begin{bmatrix}1&0\\0&0\\0&1\end{bmatrix}$, which will generate the word $w=ac$.

**Definition 5.2.10** Elementary contraction of type II for matrix of a word can be defined as follows:

$$If\quad m_{l\,r}\ne 0\ \ and\ \ m_{l\,r+1}\ne 0\ \ for\ any\ r\ and\ l,\quad 1\le l\le k, 1\le r\le n$$

$$then,\quad m'_{i\,j}=\begin{cases}m_{i\,j} & for\quad 1\le i\le k\,, 1\le j<r\\ m_{i\,j}+m_{i\,j+1} & for\quad 1\le i\le k\,,\ j=r\\ m_{i\,j+1} & for\quad 1\le i\le k\,,\ \ j>r\end{cases}$$

**Example 5.2.11** Let the word $w=abba$ be generated on the alphabet $A=\{a, b\}$. Then the order of the matrix of the word $w=abba$ will be $2\times 4$ and the matrix will be given by $\xi^{m_2}(w)=\begin{bmatrix}1&0&0&1\\0&1&1&0\end{bmatrix}$. On using elementary contraction of type II at 2nd column the order of the matrix of the reduced word will be $2\times 3$ and since $m_{22}\ne 0\ and\ m_{23}\ne 0$ $(l=2,\ r=2)$, then

$$m'_{i\,j}=\begin{cases}m_{i\,j} & for\quad 1\le i\le 2,\ 1\le j<2\\ m_{i\,j}+m_{i\,j+1} & for\quad 1\le i\le 2,\quad j=2\\ m_{i\,j+1} & for\quad 1\le i\le 2,\quad j>2\end{cases}$$

and the matrix of the reduced word will be $\xi^{m_2}(w)=\begin{bmatrix}1 & 0 & 1\\ 0 & 2 & 0\end{bmatrix}$, which will generate the word $w=ab^2c$.

**Definition 5.2.12** Let $w_1=x_1x_2..........x_n$ and $w_2=y_1y_2...........y_m$ be the two words on alphabet $\Sigma=\{a_1<a_2<a_3.........<a_k\}$ and the matrices of the two words be denoted by $\xi^{m_k}(w_1)=[a_{ij}]_{k\times n}$ and $\xi^{m_k}(w_2)=[b_{ij}]_{k\times m}$. The matrix of the word obtained by juxtaposition of two words can be defined as $\xi^{m_k}(w_1)\xi^{m_k}(w_2)=\xi^{m_k}(w_1w_2)=[c_{ij}]_{k\times(n+m)}$, where

$$c_{ij}=\begin{cases}a_{i\,j} & for \quad 1\le i\le k,\ \ 1\le j\le n\\ b_{i\,j-n} & for \quad 1\le i\le k,\ \ n+1\le j\le n+m\end{cases}$$

**Example 5.2.13** Let $w_1=ab$ and $w_2=bab$ be the two words on alphabet $A=\{a,\,b\}$. Here $r=2,\,n=2,\,m=3$ and so the matrices of the two words can be denoted by

$$\xi^{m_2}(w_1)=[a_{ij}]_{2\times 2}=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \text{ and } \xi^{m_2}(w_2)=[b_{ij}]_{2\times 3}=\begin{bmatrix}0 & 1 & 0\\ 1 & 0 & 1\end{bmatrix}.$$

The matrix of the word obtained by juxtaposition of two words can be defined as $\xi^{m_2}(w_1\,w_2)=[c_{ij}]_{2\times 5}$, where

$$c_{ij}=\begin{cases}a_{i\,j} & for \quad 1\le i\le 2,\ \ 1\le j\le 2\\ b_{i\,j-2} & for \quad 1\le i\le 2,\ \ 3\le j\le 5\end{cases} \quad \text{and}$$

$$\xi^{m_2}(w_1w_2)=\begin{bmatrix}1 & 0 & 0 & 1 & 0\\ 0 & 1 & 1 & 0 & 1\end{bmatrix}$$

## 5.3 Properties of elements of a word matrix

**Theorem 5.3.1** Let $\sum = \{a_1 < a_2 < a_3 .......... < a_k\}$ be an ordered alphabet, where $k \geq 1$ and let $w = a_{l_1}{}^{p_1} a_{l_2}{}^{p_2} .......... a_{l_n}{}^{p_n} = s_1 s_2 ........ s_n \in \sum^*$, where $s_i (i \geq 1)$ is the $i^{th}$ syllable of the word. The matrix $\xi^{m_k}(w) = [m_{i\,j}]_{k \times n}$ has the following properties:

(i) $m_{i\,j} = \begin{cases} p & if \quad s_j = a_i{}^p \\ 0 & otherwise \end{cases}$

(ii) $\sum_{j=1}^{n} m_{i\,j} = |\, w|_{a_i}$

**Proof:** (i) Since $w = a_{l_1}{}^{p_1} a_{l_2}{}^{p_2} .......... a_{l_n}{}^{p_n} = s_1 s_2 ........ s_n$, then we have

$$s_1 = a_{l_1}{}^{p_1},\ s_2 = a_{l_2}{}^{p_2}, .........s_n = a_{l_n}{}^{p_n}.$$

From definition 2.5,

$$\xi^{m_k}(w) = \xi^{m_k}(a_{l_1}{}^{p_1})\ \xi^{m_k}(a_{l_2}{}^{p_2})\ .........\ \xi^{m_k}(a_{l_n}{}^{p_n}) = [m_{i\,j}]_{k \times n},$$

and $\qquad m_{i\,j} = p_1 \quad for \quad i = l_1 \ \ and \ \ j = 1$

$s_1 = a_{l_1}{}^{p_1} \Rightarrow m_{l_1 1} = p_1,$

$$s_2 = a_{l_2}{}^{p_2} \Rightarrow m_{l_2 2} = p_2,$$

$$..........................$$

$s_n = a_{l_n}{}^{p_n} \Rightarrow m_{l_n n} = p_n$, Other entries being zero in the matrix. Therefore,

$$m_{i\,j} = \begin{cases} p & if \quad s_j = a_i{}^p \\ 0 & otherwise \end{cases}.$$

(ii) $\sum_{j=1}^{n} m_{i\,j} = m_{i1} + m_{i2} + ............ + m_{i\,n}$

Since $m_{i\,j}$ represents the number of times the letter $a_i$ repeated at $j^{th}$ syllable if $a_i$ appears at $j^{th}$ syllable, otherwise zero. Therefore the sum of all the entries of $i^{th}$ row will give the number of times the letter $a_i$ repeated in the word and hence

$$\sum_{j=1}^{n} m_{i\,j} = |\,w\,|_{a_i}$$

**Corollary 5.3.2** For each $a_i \in \Sigma$, if we define $C_i = \{j:\ m_{i\,j} \neq 0\}$, then the number of occurrences of $a_i a_j$, $a_i a_j a_k$ and in generalized way, $a_{i_1} a_{i_2} \ldots\ldots\ldots\ldots a_{i_n}$ in $w$ are given by

$$|w|_{a_i a_j} = \sum_{l_1 \in C_i} m_{i\,l_1} \sum_{l_2 \in C_j\,(l_2 > l_1)} m_{j\,l_2}$$

$$|w|_{a_i a_j a_k} = \sum_{l_1 \in C_i} m_{i\,l_1} \sum_{l_2 \in C_j\,(l_2 > l_1)} m_{j\,l_2} \sum_{l_3 \in C_k\,(l_3 > l_2)} m_{k\,l_3}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$|w|_{a_{r_1} a_{r_2} \ldots\ldots\ldots a_{r_n}} = \sum_{l_1 \in C_i} m_{r_1\,l_1} \sum_{l_2 \in C_j\,(l_2 > l_1)} m_{r_2\,l_2} \ldots\ldots\ldots\ldots\ldots\ldots \sum_{l_n \in C_n\,(l_n > l_{n-1})} m_{r_n\ l_n}$$

**Example 5.3.3** Let $\Sigma = \{a < b < c\}$ and $w = aabbc$**.** Matrix of the word $w = aabbc$ is $\xi^{m_3}(w) = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$.

$|w|_a = \sum_{j=1}^{5} m_{1\,j} = 2$, $|w|_b = \sum_{j=1}^{5} m_{2\,j} = 2$ and similarly $|w|_c = \sum_{j=1}^{5} m_{3\,j} = 1$.

Again we have $C_1 = \{1, 2\}$, $C_2 = \{3,\ 4\}$ and $C_3 = \{5\}$.

$$|w|_{ab} = \sum_{l_1 \in C_1} m_{1\,l_1} \sum_{l_2 \in C_2\,(l_2 > l_1)} m_{2\,l_2} = m_{11}\left(m_{22} + m_{23} + m_{24} + m_{25}\right) + m_{12}\left(m_{23} + m_{24} + m_{25}\right)$$

$$= 1(0 + 1 + 1 + 0) + 1(1 + 1 + 0)$$

$$= 4$$

If we write $w = a^2 b^2 c$, then

$$\xi^{m_3}(w) = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ and}$$

$|w|_a = \sum_{j=1}^{5} m_{1j} = 2$, $|w|_b = \sum_{j=1}^{5} m_{2j} = 2$ and similarly $|w|_c = \sum_{j=1}^{5} m_{3j} = 1$.

$C_1 = \{1\}$, $C_2 = \{2\}$ and $C_3 = \{3\}$.

$$\begin{aligned} |w|_{ab} = \sum_{l_1 \in C_1} m_{1l_1} \sum_{l_2 \in C_2 (l_2 > l_1)} m_{2l_2} &= m_{11}\left(m_{22} + m_{23}\right) \\ &= 2(2+0) \\ &= 4 \end{aligned}$$

**Example 5.3.4** Let $\Sigma = \{a < b < c\}$ and $w = abcabc$. Matrix of the word $w = abcabc$ is $\xi^{m_3}(w) = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$.

We will show that occurrence of a subword can be calculated with the help of the word matrix irrespective of the order of letters in it.

We have $C_1 = \{1, 4\}$, $C_2 = \{2, 5\}$, $C_3 = \{3, 6\}$ and $|w| = 6$

$$\begin{aligned} |w|_{ab} = \sum_{l_1 \in C_1} m_{1l} \sum_{l_2 \in C_2 (l_2 > l_1)} m_{2l_2} &= m_{11}\left(m_{22} + m_{23} + m_{24} + m_{25} + m_{26}\right) + m_{14}\left(m_{25} + m_{26}\right) \\ &= 1(1+0+0+1+0) + 1(1+0) \\ &= 2+1 \\ &= 3 \end{aligned}$$

$$\begin{aligned} |w|_{ba} = \sum_{l_1 \in C_2} m_{2l_1} \sum_{l_2 \in C_2 (l_2 > l_1)} m_{1l_2} &= m_{22}\left(m_{13} + m_{14} + m_{15} + m_{16}\right) + m_{25}\left(m_{16}\right) \\ &= 1(0+1+0+0) + 1(0) \\ &= 1 \end{aligned}$$

$$\begin{aligned}|w|_{abc} &= \sum_{l_1 \in C_1} m_{1l_1} \sum_{l_2 \in C_2 (l_2 > l_1)} m_{2\,l_2} \sum_{l_3 \in C_3 (l_3 > l_2)} m_{3l_3} \\ &= m_{11}\left[m_{22}\left(m_{33} + m_{34} + m_{35} + m_{36}\right) + m_{23}\left(m_{34} + m_{35} + m_{36}\right) + m_{24}\left(m_{35} + m_{36}\right) + m_{25}\left(m_{36}\right)\right] \\ &\quad + m_{14}\left[m_{25}\left(m_{36}\right)\right] \\ &= 1\left[1\ \left(1 + 0 + 0 + 1\right) + 0(0 + 0 + 1) + 0(0 + 1) + 1(1)\right] + 1\left[1(1)\right] \\ &= 4\end{aligned}$$

Similarly $|w|_{ac} = 3, |w|_{cab} = 1$ and $|w|_{cb} = 1$ etc.

The word matrix provides us an opportunity to check whether a word belongs to a language or not.

Considering the language

$$L = \left\{x : x \in \{a, b\}^*,\ \textit{the number of } a's \textit{ in } x \textit{ is a multiple of } 3\right\}.$$

From the (ii) property of theorem 3.1, the following condition holds for the given language.

$$\sum_{j=1}^{n} m_{i\,j} = 3.k\ \ (k \in N)$$

## 5.4 Word matrix for the Inverse and mirror image of a word

**Definition 5.4.1** Let $\Sigma = \{a_1 < a_2 < a_3 .......... < a_k\}$ be an ordered alphabet, where $k \geq 1$ and let $w = {a_{l_1}}^{p_1} {a_{l_2}}^{p_2} .......... {a_{l_n}}^{p_n} = s_1 s_2 ........ s_n \in \Sigma^*$, where $s_i (i \geq 1)$ is the $i^{th}$ syllable of the word and $\xi^{m_k}(w)$ is the matrix of the word $w$. The inverse of the word $w$, is the word $w^{-1} = {a_{l_n}}^{-p_n} {a_{l_{n-1}}}^{-p_{n-1}} ........... {a_{l_1}}^{-p_1}$ and the matrix of $w^{-1}$ shall be $\xi^{m_k}(w^{-1})$, which can be obtained from the $\xi^{m_k}(w)$ in the following manner:

If $\quad \xi^{m_k}(w) = [a_{i\,j}]_{k \times n} \quad$ and $\quad \xi^{m_k}(w^{-1}) = [b_{i\,j}]_{k \times n}$,

Then $\quad b_{i\,j} = -a_{i\ n-j+1}$

Given a word $w = a_{l_1}{}^{p_1} a_{l_2}{}^{p_2} .........a_{l_n}{}^{p_n}$, the mirror image of word $w$ is denoted by $mi(w)$, that is $mi(w) = a_{l_n}{}^{p_n} a_{l_{n-1}}{}^{p_{n-1}} ...........a_{l_1}{}^{p_1}$.

From the above definitions of inverse of a word and mirror image of a word, it can be easily observed that

$$\xi^{m_k}(w^{-1}) = -\xi^{m_k}\big(mi(w)\big).$$

## 5.5 Properties of word matrix

In this section we shall define some properties of word matrix.

**Theorem 5.5.1** Let $\Sigma = \{a_1 < a_2 < a_3 .......... < a_k\}$ be an ordered alphabet, where $k \geq 1$. The set of matrices $M_k$ under the operation of juxtaposition on matrices is a free Group.

**Proof:** Let $w_i = x_1 x_2 ....x_n$, $w_j = y_1 y_2 ....y_m$ and $w_k = z_1 z_2 ...........z_p$ be any three arbitrary words on alphabet $\Sigma = \{a_1 < a_2 < a_3 .......... < a_k\}$, then the matrices of the three words can be denoted by $\xi^{m_k}(w_i) = [a_{ij}]_{k\times n}$, $\xi^{m_k}(w_j) = [b_{ij}]_{k\times m}$ and $\xi^{m_k}(w_k) = [c_{ij}]_{k\times p}$.

**1.Closure property:** Since $\xi^{m_k}(w_1)\xi^{m_k}(w_2) = \xi^{m_k}(w_1 w_2) = [c_{ij}]_{k\times(n+m)}$. Therefore juxtaposition of matrices of words satisfies the closure property.

**2. Associative law:** Since we have

$$\begin{aligned}\xi^{m_k}(w_i)\big(\xi^{m_k}(w_j)\xi^{m_k}(w_k)\big) &= \xi^{m_k}(w_i)\xi^{m_k}(w_j w_k)\\ &= \xi^{m_k}(w_i w_j w_k)\\ &= \xi^{m_k}(w_i w_j)\xi^{m_k}(w_k)\\ &= \big(\xi^{m_k}(w_i)\xi^{m_k}(w_j)\big)\xi^{m_k}(w_k)\end{aligned}$$

Hence juxtaposition of matrices of words is associative.

**3. Identity:** The null matrix of order $k \times 1$ is the identity. Let the identity be denoted by $\xi^{m_k}(w_0)$, then for matrix of any word $\xi^{m_k}(w_i)$, we have $\xi^{m_k}(w_0)\xi^{m_k}(w_i) = \xi^{m_k}(w_i)$ on applying elementary contraction of type I. Thus the null matrix of order $k \times 1$ is the identity element.

**4. Inverse:** Let $\xi^{m_k}(w_i) = [a_{ij}]_{k \times n}$ be the matrix of any word $w_i$ , then there exists $\xi^{m_k}(w_i^{-1}) = [b_{ij}]_{k \times n}$, where $b_{ij} = -a_{i\, n-j+1}$

such that $\xi^{m_k}(w_i)\xi^{m_k}(w_i^{-1}) = \xi^{m_k}(w_0)$ ,using elementary contraction of type II and I.

Therefore $\xi^{m_k}(w_i^{-1})$ is the inverse of $\xi^{m_k}(w_i)$.

Hence $M_k$ is a free group under the binary operation juxtaposition of matrices.

**Theorem 5.5.2** Let $\Sigma = \{a_1 < a_2 < a_3 .......... < a_k\}$ be an ordered alphabet, where $k \geq 1$. The word matrix mapping $\xi^{m_k} : \Sigma^* \to M_k$ is an isomorphism from the group $\Sigma^*$ to $M_k$.

**Proof:** Since every word has a unique matrix and all the matrices of $M_k$ are mapping of some of the words of $\Sigma^*$ , thus the mapping is one - one onto mapping Also from definition 2.12, $\xi^{m_k}(w_1 w_2) = \xi^{m_k}(w_1)\xi^{m_k}(w_2)$. Hence $\xi^{m_k}$ is an isomorphism from $\Sigma^*$ to $M_k$.

**Theorem 5.5.3** Let $\Sigma = \{a_1 < a_2 < a_3 .......... < a_k\}$ be an ordered alphabet, where $k \geq 1$ and assume that $w = s_1 s_2 ........ s_n \in \Sigma^*$, where $s_i = a_j^{\ p} (i, j \geq 1)$ is the $i^{th}$ syllable of the word and $\xi^{m_k}(w)$ is the matrix of the word. If we denote the transpose of the matrix by $\xi^{m_k'}(w)$, then for $p = 1$

$$\xi^{m_k}(w).\xi^{m_k'}(w) = dia\left(|w|_{a_1}, |w|_{a_2}, \ldots\ldots\ldots\ldots\ldots\ldots, |w|_{a_k}\right)$$

**Proof:** Let $\xi^{m_k}(w) = [a_{ij}]_{k\times n}$ and $\xi^{m_k'}(w) = [b_{ij}]_{n\times k}$, then $b_{ij} = a_{ji}$. Further let $\xi^{m_k}(w)\xi^{m_k'}(w) = [c_{ij}]_{k\times k}$, then

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \ldots\ldots\ldots\ldots\ldots\ldots + a_{in}b_{nj}$$

Case 1 : $i = j$

$$c_{ij} = a_{i1}b_{1i} + a_{i2}b_{2i} + \ldots\ldots\ldots\ldots\ldots\ldots + a_{in}b_{ni}$$

$$= a_{i1}a_{i1} + a_{i2}a_{i2} + \ldots\ldots\ldots\ldots\ldots\ldots + a_{in}a_{in}$$

$= a_{i1} + a_{i2} + \ldots\ldots\ldots\ldots\ldots\ldots + a_{in}$ since $a_{ij} = 1 \; or \; 0$

$$= \sum_{j=1}^{n} a_{ij}$$

$$= |w|_{a_i}$$

Case 2: $i \neq j$

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \ldots\ldots\ldots\ldots\ldots\ldots + a_{in}b_{nj}$$

$$= a_{i1}a_{j1} + a_{i2}a_{j2} + \ldots\ldots\ldots\ldots\ldots\ldots + a_{in}a_{jn}$$

If $a_{ij} = 1$, then from definition 2, other entries in the $j^{th}$ column will be zero that is in a column only one entry will be non-zero and other will be zero. Therefore the product of two entries of a column will be zero.

$$c_{ij} = 0$$

Hence we have $\xi^{m_k}(w).\xi^{m_k'}(w) = [c_{ij}]_{k\times k}$, where

$$c_{ij} = \begin{cases} |w|_{a_i} & if \quad i = j \\ 0 & otherwise \end{cases}$$

**Corollary 5.5.4** The diagonal matrix obtained by the product of the matrix of a word $w$, that is, $\xi^{m_k}(w)$ and its transpose matrix $\xi^{m_k'}(w)$ is the Parikh vector of $w$.

## 5.6 Conclusion

In the present chapter an injective word matrix mapping has been defined and elementary contractions and juxtaposition of words have been defined using word matrices. It has been also shown that Parikh vector of a word can be generated with help of the defined word matrix and subword occurrence can be calculated irrespective of the order of the letters in it which is an additional advantage of the proposed word matrix. The algebraic properties of words can be well studied by word matrix mapping. I have tried to deal with many sets of stone tables pertinent with formal language theory, all of which are valid in their own domains and have accentuated those aspects which are relevant in natural language processing.

# Chapter VI

# Extending Matrix Representation of Words

*In the present chapter word matrices have been generalized by defining a word matrix mapping induced by a word. Word matrix mapping is further extended to q-matrix encoding, which takes its values in matrices with polynomial entries. A distance function has been defined to examine the similarity of words and metric space has been generated for word matrices. Distance polynomial is defined and has been utilized for study of morphology in natural languages.*

## 6.1 Introduction

This chapter is the extension of the work carried out in the previous chapter, where we had introduced a word matrix mapping and the matrix of a word was defined with respect to the alphabet. Subword (having no repeating letter) occurrences were calculated with the help of the word matrix mapping. Here we wish to define Generalization of the word matrix mapping in which emphasis shall be on word matrix mapping induced by a word. Examples of words from real world have been cited. It shall provide the way to calculate subword (having repeating letters) occurrence. Word q-matrix encoding has been defined in order to count the subword occurrence by giving a polynomial which describes relative position of subword as the power of q.

## 6.2 Extending Word matrix

In our present study $M_k(q)$ has been used to denote the set of matrices having $k$ rows and finite number of columns which have at most one element in $N[q]$ in each column and other entries being zero. Word matrix mapping has been defined with the help of the following definitions.

**Definition 6.2.1** Let $\Sigma = \{a_1 < a_2 < a_3 .......... < a_k\}$ be an ordered alphabet, where $k \geq 1$. Let $u = b_1 b_2 ........ b_n$ be a word in $\Sigma^*$ $\left(b_i \in \Sigma \;\; for\ all \;\; 1 \leq i \leq n\right)$. The mapping induced by the word $u$ over the alphabet $\Sigma$ shall be the mapping $\xi_u^{m_k} : \Sigma \to M_n$ defined as:

$$If \quad \xi_u^{m_k}(a_l) = [m_{ij}]_{n\times 1}, \quad then$$

$$m_{ij} = \begin{cases} 1 & for \quad b_i = a_l \\ 0 & otherwise \end{cases}$$

To illustrate the point we can cite the following examples.

**Example 6.2.2** Let $\Sigma = \{a < b\}$and $u = aba$, then $\xi_u^{m_2}(a) = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$ and $\xi_u^{m_2}(b) = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$.

From definition 6.2.1, it can be observed that the matrix of a syllable over an alphabet having cardinality $n$ will contain $n$ rows and one column.

**Definition 6.2.3** Let $\Sigma = \{a_1 < a_2 < a_3 .......... < a_k\}$ be an ordered alphabet and $\xi_u^{m_k}(a_{l_1}) = [a_{ij}]_{n\times 1}$ and $\xi_u^{m_k}(a_{l_2}) = [b_{ij}]_{n\times 1}$ be the matrices of two syllables. The composition of juxtaposition on the matrices of syllables can be defined as:

$$\xi_u^{m_k}(a_{l_1})\,\xi_u^{m_k}(a_{l_2}) = [c_{i\,j}]_{n\times 2}, \quad \text{where}$$

$$c_{i\,j} = \begin{cases} a_{i\,j} & for \quad j = 1 \\ b_{i\,j} & for \quad j = 2 \end{cases}$$

**Example 6.2.4** Let $\Sigma = \{a < b\}$ and $u = aba$, then

$$\xi_u^{m_2}(a)\xi_u^{m_2}(b)\xi_u^{m_2}(a) = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

Finally with the help of above two definitions, we can define word matrix mapping.

**Definition 6.2.5** The word matrix mapping induced by the word $u = b_1 b_2 .........b_n$ shall be the morphism $\xi_u^{m_k} : \Sigma^* \to M_n$, defined as follows:

$$If \quad \xi_u^{m_k}(a_l) = [m_{i\,j}]_{n\times 1}, \quad then$$

$$m_{i\,j} = \begin{cases} 1 & for \quad b_i = a_l \\ 0 & otherwise \end{cases}$$

and $$\xi_u^{m_k}(a_{l_1} a_{l_2} ...............a_{l_n}) = \xi_u^{m_k}(a_{l_1})\xi_u^{m_k}(a_{l_2})...............\xi_u^{m_k}(a_{l_n})$$

**Example 6.2.6** Let $\Sigma = \{a < b < c\}$ be the ordered alphabet and $u = bc$. Then for $w = abc$, $\xi^{m_3}(w)$ is a matrix of order $2\times 3$ that can be computed as follows:

$$\xi_u^{m_3}(abc) = \xi_u^{m_3}(a)\,\xi_u^{m_3}(b)\,\xi_u^{m_3}(c)$$

$$= \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

**Example 6.2.7** Let $\Sigma = \{a < b\}$ be the ordered alphabet and $u = bab$, then for $w = abba$, then

$$\begin{aligned}\xi_u^{m_k}(abba) &= \xi_u^{m_k}(a)\ \xi_u^{m_k}(b)\,\xi_u^{m_k}(b)\xi_u^{m_k}(a)\\ &= \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}\end{aligned}$$

## 6.3 Properties of the elements of extended word matrix

**Theorem 6.3.1** Let $\Sigma = \{a_1 < a_2 < a_3 .......... < a_k\}$ be an ordered alphabet and $u = b_1 b_2 .........b_n \ \left(b_i \in \Sigma \ \ \textit{for all} \ \ 1 \le i \le n\right)$, then for a word

$w = a_{l_1} a_{l_2} ..........a_{l_m} \in \Sigma^*$,

matrix $\xi_u^{m_k}(w) = [m_{ij}]_{n\times m}$ has the following properties:

(i) $m_{ij} = \begin{cases} 1 & \textit{if} \quad b_i = a_{l_j}, \quad 1 \le i \le n \ \&\ 1 \le j \le m \\ 0 & \textit{otherwise} \end{cases}$

(ii) $\displaystyle\sum_{j=1}^{m} m_{ij} = |\,w\,|_{b_i}$

**Proof:** (i) Since $w = a_{l_1} a_{l_2} ..........a_{l_m}$, then we have

$$\xi_u^{m_k}(a_{l_1}) = [a_{ij}^1]_{n\times 1}, \text{ where } \ a_{ij}^1 = \begin{cases} 1 & \textit{if} \quad b_i = a_{l_1} \\ 0 & \textit{otherwise} \end{cases}$$

..............................................................

$$\xi_u^{m_k}(a_{l_m}) = [a_{ij}^m]_{n\times 1}, \text{ where } \ a_{ij}^m = \begin{cases} 1 & \textit{if} \quad b_i = a_{l_m} \\ 0 & \textit{otherwise} \end{cases}$$

From definition 6.2.5,

$$\begin{aligned}\xi_u^{m_k}(w) &= \xi_u^{m_k}(a_{l_1})\xi_u^{m_k}(a_{l_2}) \,........\, \xi_u^{m_k}(a_{l_m})\\ &= [m_{ij}]_{k\times n}, \quad \text{where}\end{aligned}$$

$$m_{ij} = \begin{cases} a_{ij}^{1} & for \quad j = 1 \\ a_{ij}^{2} & for \quad j = 2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{ij}^{m} & for \quad j = m \end{cases}$$

Therefore, we have $m_{ij} = \begin{cases} 1 & if \quad b_i = a_{l_j}, \quad 1 \le i \le n \;\&\; 1 \le j \le m \\ 0 & otherwise \end{cases}$ .

(ii) $\sum_{j=1}^{n} m_{ij} = m_{i1} + m_{i2} + \ldots\ldots\ldots\ldots + m_{in}$

Since $m_{ij}$ represents the number of times the letter $b_i$ is repeated at the position of $j^{th}$ syllable, therefore the sum of all the entries of $i^{th}$ row will give the number of times the letter $a_i$ repeated in the word and hence

$$\sum_{j=1}^{n} m_{ij} = |\, w \,|_{b_i}$$

**Corollary 6.3.2** For each $a_{r_i} \in \Sigma$, if we define $C_i = \{j\!: \; m_{ij} \ne 0\}$, then the number of occurrences of $b_i b_j$, $b_i b_j b_k$ and in generalized way $b_{r_1} b_{r_2} \ldots\ldots\ldots. b_{r_n}$ in relation with word $w$ are given by

$$|w|_{b_i b_j} = \sum_{l_1 \in C_i} m_{i\,l_1} \sum_{l_2 \in C_j\,(l_2 > l_1)} m_{j\,l_2}$$

$$|w|_{b_i b_j b_k} = \sum_{l_1 \in C_i} m_{i\,l_1} \sum_{l_2 \in C_j\,(l_2 > l_1)} m_{j\,l_2} \sum_{l_3 \in C_k\,(l_3 > l_2)} m_{k\,l_3}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$|w|_{b_{r_1} b_{r_2} \ldots\ldots\ldots. b_{r_n}} = \sum_{l_1 \in C_i} m_{r_1\,l_1} \sum_{l_2 \in C_j\,(l_2 > l_1)} m_{r_2\,l_2} \ldots\ldots\ldots\ldots\ldots\ldots \sum_{l_n \in C_n\,(l_n > l_{n-1})} m_{r_n\,l_n}$$

**Example 6.3.3** Let $\Sigma = \{a < b < c\}$ and $u = acb$, then for a word $w = abcabc$

$$\xi_u^{m_3}(w) = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

We have $C_1 = \{1, 4\}$, $C_2 = \{3, 6\}$, $C_3 = \{2, 5\}$ and $|w| = 6$

$$|w|_{ab} = \sum_{l_1 \in C_1} m_{1l_1} \sum_{l_2 \in C_2 (l_2 > l_1)} m_{2l_2} = m_{11}(m_{23} + m_{26}) + m_{14}(m_{26})$$

$$= 1(1+1)+1(1)$$

$$= 3$$

$$|w|_{ba} = \sum_{l_1 \in C_3} m_{3l_1} \sum_{l_2 \in C_1 (l_2 > l_1)} m_{1l_2} = m_{32}(m_{14})$$

$$= 1(1)$$

$$= 1$$

**Corollary 6.3.4** Let $u = b_1 b_2 .........b_n$, then for a word $w$,

$$|w|_u = |w|_{b_1 b_2 .........b_n} = \sum_{l_1 \in C_1} m_{1l_1} \sum_{l_2 \in C_2 (l_2 > l_1)} m_{2\,l_2} .................... \sum_{l_n \in C_n (l_n > l_{n-1})} m_{n\,l_n}$$

**Example 6.3.5** In example 6.3.3

$$|w|_u = \sum_{l_1 \in C_1} m_{1l_1} \sum_{l_2 \in C_2 (l_2 > l_1)} m_{2\,l_2} \sum_{l_3 \in C_3 (l_3 > l_2)} m_{3l_3}$$

$$= m_{11} \lfloor m_{23} \{ m_{35} \} \rfloor$$

$$= 1$$

**Example 6.3.6** Let $\Sigma = \{a < b < c < d\}$ and for the real given word $u = bad$, the word $w = cab$ shall have the matrix with respect to $u$:

$$\xi_u^{m_3}(w) = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

**Example 6.3.7** Let $\Sigma = \{a < b < c < ........... < z\}$ be the set of English alphabet, then for a word $u = boat$, the matrix of any other word can be defined with respect to the word *boat*. The elements of matrix induced by the word *boat* will represent the occurrence of the letters of the word *boat* in the given word. Given a word $w = about$, the matrix of the word *about* induced by *boat* shall be

$$\xi_u^{m_{26}}(w) = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

We have, $C_1 = \{2\}$ $C_2 = \{3\}$, $C_3 = \{1\}$, $C_4 = \{5\}$ and $|w| = 5$

$$|w|_{bo} = \sum_{l_1 \in C_1} m_{1l_1} \sum_{l_2 \in C_2(l_2 > l_1)} m_{2l_2} = m_{12}(m_{23}) = 1$$

$$|w|_{at} = \sum_{l_1 \in C_3} m_{3l_1} \sum_{l_2 \in C_4(l_2 > l_1)} m_{4l_2} = m_{31}(m_{45}) = 1$$

$$|w|_{bot} = \sum_{l_1 \in C_1} m_{1l_1} \sum_{l_2 \in C_2(l_2 > l_1)} m_{2l_2} \sum_{l_3 \in C_4(l_3 > l_2)} m_{4l_3} = m_{12}(m_{23}(m_{45})) = 1$$

Above analysis also reflects that $bo, at, bot$ etc. are common subwords in both of the words.

## 6.4 Properties of extended word matrix

**Theorem 6.4.1** Let $\Sigma = \{a_1 < a_2 < a_3 .......... < a_k\}$ be an ordered alphabet and $u = b_1 b_2 ......... b_n$ $(b_i \in \Sigma \ \ for\ all \ \ 1 \le i \le n)$, where $k \ge 1$. The set of matrices $\xi_u^{m_k}(w_i)$ under the operation of juxtaposition on matrices is a free Group.

**Proof:** Proof of the theorem is similar to the proof of theorem 5.5.1 on using extended word matrix in place of usual word matrix.

**Theorem 6.4.2** Let $\Sigma = \{a_1 < a_2 < a_3 .......... < a_k\}$ be an ordered alphabet and $u = b_1 b_2 ......... b_n$ $(b_i \in \Sigma \ \ for\ all \ \ 1 \le i \le n)$. The induced word matrix mapping $\xi_u^{m_k} : \Sigma^* \to M_n$ is an isomorphism from the group $\Sigma^*$ to $M_n$.

**Proof:** Since every word has a unique matrix induced by the word $u = b_1 b_2 ......... b_n$ and all the matrices of $M_n$ are mapping of some of the words of $\Sigma^*$ , thus the mapping is one -one onto mapping Also generalizing definition 5.2.12, we have $\xi_u^{m_k}(w_1 w_2) = \xi_u^{m_k}(w_1)\xi_u^{m_k}(w_2)$. Hence $\xi_u^{m_k}$ is an isomorphism from $\Sigma^*$ to $M_n$.

**Theorem 6.4.3** Let $\Sigma = \{a_1 < a_2 < a_3 .......... < a_k\}$ be an ordered alphabet, where $k \ge 1$ and $u = b_1 b_2 ......... b_n$ $(b_i \in \Sigma \ \ for\ all \ \ 1 \le i \le n)$, further let $\xi_u^{m_k}(w)$ and $\xi_u^{m_k'}(w)$ are the matrices of a word $w = a_{l_1} a_{l_2} .......... a_{l_m} \in \Sigma^*$ induced by $u$ and its transpose respectively. Then we have

$$\xi_u^{m_k}(w).\xi_u^{m_k'}(w) = dia\left( |w|_{b_1}, |w|_{b_2}, .................., |w|_{b_n} \right)$$

**Proof:** Proof of the theorem is similar to the proof of theorem 5.5.3 on using extended word matrix in place of usual word matrix.

## 6.5 Word $q$ matrix encoding

Let $N[q]$ be the set of the elements of the form of $(q^n, n \in N)$ and $M_k(q)$ be the set of matrices having $k$ rows and finite number of columns which have at most one element of $N[q]$ in each column and other entries being zero. The encoding exercise can be performed with the help of following definitions-

**Definition 6.5.1** Let $\Sigma = \{a_1 < a_2 < a_3 .......... < a_k\}$ be an ordered alphabet, where $k \geq 1$ and $\xi^{m_k}(w)$ be the set of word matrices. The q-matrix encoding of a word matrix shall be the mapping $\xi_q^{m_k} : \xi^{m_k}(w) \rightarrow M_k(q)$ defined as:

$$If \quad \xi_q^{m_k}\left([a_{i\,j}]_{k\times n}\right) = [b_{i\,j}]_{k\times n}, \quad then$$

$$b_{i\,j} = \begin{cases} q^j & if \quad a_{i\,j} = 1 \\ 0 & otherwise \end{cases}$$

**Example 6.5.2** Let $\Sigma = \{a < b < c\}$ and $w = abbcc$, then

$$\xi^{m_k}(w) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \quad \text{and} \quad \xi_q^{m_k}(w) = \begin{bmatrix} q^1 & 0 & 0 & 0 & 0 \\ 0 & q^2 & q^3 & 0 & 0 \\ 0 & 0 & 0 & q^4 & q^5 \end{bmatrix}$$

**Example 6.5.3** Let $\Sigma = \{a < b < c < d\}$ and for the real given word $u = bad$,

$$\xi^{m_3}(w) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \xi_q^{m_3}(w) = \begin{bmatrix} 0 & q^2 & 0 \\ q^1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & q^3 \end{bmatrix}$$

## 6.6 Properties of elements of a word q-matrix encoding

**Theorem 6.6.1** Let $\Sigma = \{a_1 < a_2 < a_3 .......... < a_k\}$ be an ordered alphabet, where $k \geq 1$ and let us assume that $w = b_1 b_2 ....... b_n \in \Sigma^*$ $(b_i \in \Sigma)$. Then the matrix $\xi_q^{m_k}(w) = [m_{i\,j}]_{k\times n}$ shall have the following property:

(i) $\sum_{j=1}^{n} m_{i\,j} = P_{w,a}(q)$, where $P_{w,a}(q)$ is the polynomial of q-count of the word $a$ as a subword of $w$.

**Proof:**

Since $\sum_{j=1}^{n} m_{i\,j} = m_{i1} + m_{i2} + \ldots\ldots\ldots\ldots + m_{i\,n}$

Since $m_{i\,j} = q^{j}$, if $a_i$ appears at $j^{th}$ syllable. Therefore the sum of all the entries of $i^{th}$ row will give the q-count of the letter $a_i$ appeared in the word and hence $\sum_{j=1}^{n} m_{i\,j} = P_{w,a}(q)$.

**Example 6.6.2** Let $\Sigma = \{a < b < c\}$ and $w = abbcc$, then

$$\xi_q^{m_k}(w) = \begin{bmatrix} q^1 & 0 & 0 & 0 & 0 \\ 0 & q^2 & q^3 & 0 & 0 \\ 0 & 0 & 0 & q^4 & q^5 \end{bmatrix}$$

$$\begin{aligned} \sum_{j=1}^{5} m_{1\,j} &= m_{11} + m_{12} + m_{13} + m_{14} + m_{15} \\ &= q^1 + 0 + 0 + 0 + 0 \\ &= q^1 \\ &= P_{w,a}(q) \end{aligned}$$

$$\begin{aligned} \sum_{j=1}^{5} m_{2\,j} &= m_{21} + m_{22} + m_{23} + m_{24} + m_{25} \\ &= 0 + q^2 + q^3 + 0 + 0 \\ &= q^2 + q^3 \\ &= P_{w,b}(q) \end{aligned}$$

**Corollary 6.6.3** For each $b_i \in \Sigma$, if we define $C_i = \{j\colon\, m_{i\,j} \neq 0\}$, then the polynomial of q-count of $b_i b_j$, $b_i b_j b_k$ and in generalized way $b_{r_1} b_{r_2} \ldots\ldots\ldots b_{r_n}$ in $w$ are given by

$$P_{w,b_ib_j}(q) = \sum_{l_1 \in C_i} m_{i\,l_1} \sum_{l_2 \in C_j\,(l_2 > l_1)} m_{j\,l_2}$$

$$P_{w,\,b_ib_jb_k}(q) = \sum_{l_1 \in C_i} m_{i\,l_1} \sum_{l_2 \in C_j\,(l_2 > l_1)} m_{j\,l_2} \sum_{l_3 \in C_k\,(l_3 > l_2)} m_{k\,l_3}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$P_{w,b_{r_1}b_{r_2}\ldots\ldots\ldots b_{r_n}}(q) = \sum_{l_1 \in C_i} m_{r_1\,l_1} \sum_{l_2 \in C_j\,(l_2 > l_1)} m_{r_2\,l_2} \ldots\ldots\ldots\ldots\ldots\ldots \sum_{l_n \in C_n\,(l_n > l_{n-1})} m_{r_n\,l_n}$$

**Corollary 6.6.4** On putting $q = 1$, the q-matrix encoding reduces to word matrix mapping.

**Example 6.6.5** Let $\Sigma = \{a < b < c\}$ and $w = abbacca$, then

$$\xi_q^{m_k}(w) = \begin{bmatrix} q^1 & 0 & 0 & q^4 & 0 & 0 & q^7 \\ 0 & q^2 & q^3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & q^5 & q^6 & 0 \end{bmatrix}$$

Thus, $C_1 = \{1, 4, 7\},\ C_2 = \{2, 3\},\ C_3 = \{5, 6\}$

$$\begin{aligned} P_{w,ab}(q) = \sum_{l_1 \in C_1} m_{1l_1} \sum_{l_2 \in C_2\,(l_2 > l_1)} m_{2l_2} &= m_{11}(m_{22} + m_{23}) \\ &= q^1(q^2 + q^3) \\ &= q^3 + q^4 \end{aligned}$$

**Example 6.6.6** Let $\Sigma = \{a < b < c < d\}$ and for the real given word $u = bad$,

$$\xi_q^{\ m_3}(w) = \begin{bmatrix} 0 & q^2 & 0 \\ q^1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & q^3 \end{bmatrix}$$

Thus, $C_1 = \{2\},\ C_2 = \{1\},\ C_4 = \{3\}$

$$P_{w,ad}(q) = \sum_{l_1 \in C_1} m_{1 l_1} \sum_{l_2 \in C_4 (l_2 > l_1)} m_{4 l_2} = m_{12}(m_{43})$$
$$= q^2 (q^3)$$
$$= q^5$$

On putting $q = 1$, we have $P_{w,ad}(1) = |w|_{ad} = 1$

## 6.7 Similarity of words

**Definition 6.7.1** Let $\Sigma = \{a_1 < a_2 < a_3 .......... < a_k\}$ be an ordered alphabet and $w_1 = b_1 b_2 .......b_n$, $w_2 = c_1 c_2 .......c_n$ $(b_i, c_i \in \Sigma \ \ \textit{for all} \ \ 1 \le i \le n)$ be two words on $\Sigma$. Let the matrices of the two words $w_1$ and $w_2$ be $\xi^{m_k}(w_1) = [b_{i\,j}]_{k \times n}$ and $\xi^{m_k}(w_2) = [c_{i\,j}]_{k \times n}$ respectively. Then the difference between any two letters $a_r, a_s$ $(1 \le r, s \le n)$ of the words $w_1 = b_1 b_2 .......b_n$ and $w_2 = c_1 c_2 .......c_n$ can be defined as

$$\delta(b_r, c_s) = \begin{cases} 0 & \textit{if} \quad b_{i\,r} = c_{i\,s} \ \ \textit{for all} \ 1 \le i \le k \\ 1 & \textit{otherwise} \end{cases}$$

**Definition 6.7.2** Let $\Sigma = \{a_1 < a_2 < a_3 .......... < a_k\}$ be an ordered alphabet and $w_1 = b_1 b_2 .......b_n$, $w_2 = c_1 c_2 .......c_m$ $(b_i, c_i \in \Sigma)$ be two words on $\Sigma$. Let the matrices of the two words $w_1$ and $w_2$ be $\xi^{m_k}(w_1) = [b_{i\,j}]_{k \times n}$ and $\xi^{m_k}(w_2) = [c_{i\,j}]_{k \times m}$. Then the distance between two words may be defined as

$$D(w_1, w_2) = \sum_{i=1}^{Max(n,m)} \delta(b_i, c_i)$$

**Example 6.7.3** Let $\Sigma = \{a < b < c < d < e\}$ and for the real given words $w_1 = dead$ and $w_2 = deed$, then matrices of $w_1$ and $w_2$ are

$$w_1 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \text{ and } w_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}.$$

Then, $D(w_1, w_2) = 1$.

**Theorem 6.7.4** Let $\Sigma^*$ be the set of all words on the alphabet $\Sigma$. Then $(\Sigma^*, D)$ shall be a metric space.

**Proof:**

We have the following properties of $D(w_1, w_2)$

(i) Since $\delta(b_k, c_k) = 1 \; or \; 0$ , therefore $D(w_1, w_2) \geq 0$.

(ii) Since $\delta(b_k, c_k) = \delta(c_k, b_k)$ , therefore $D(w_1, w_2) = D(w_2, w_1)$.

(iii) Let $w_1 = b_1 b_2 .......b_n$, $w_2 = c_1 c_2 .......c_m$ and $w_3 = d_1 d_2 .......d_l$ $(b_i, c_i, d_i \in \Sigma)$

Let $D(w_1, w_2) = r, \; 0 \leq r \leq Max(n, m)$ and

$D(w_2, w_3) = r_1, \; 0 \leq r_1 \leq Max(m, l)$

Therefore, $D(w_1, w_2) + D(w_2, w_3) = r + r_1$

Distance between two words shows the number of different letters in two words. $D(w_1, w_2) = r$ implies that there are $r$ different letters in the two words $w_1$ and $w_2$. Similarly $D(w_2, w_3) = r_1$ implies that there are $r_1$ different letters in the two words $w_2$ and $w_3$. There may be some common letters between $r$ and $r_1$ or $r$ and $r_1$ shall be completely different. Therefore,

$$D(w_1, w_3) \leq r + r_1$$

and $$D(w_1, w_3) \leq D(w_1, w_2) + D(w_2, w_3).$$

Hence $(\Sigma^*, D)$ is a metric space.

**Definition 6.7.5** Let $\Sigma = \{a_1 < a_2 < a_3 .........< a_k\}$ be an ordered alphabet and $w_1 = a_1 a_2 .......a_n$, $w_2 = b_1 b_2 .......b_m$ $(b_i \in \Sigma \;\; for \; all \;\; 1 \leq i \leq m)$ be two words on $\Sigma$. Let the matrices of the two words ........be $\xi^{m_k}(w_1) = [a_{ij}]_{k\times n}$ and $\xi^{m_k}(w_2) = [b_{ij}]_{k\times m}$. The distance polynomial between two words may be defined as

$$P(w_1, w_2) = \sum_{i=1}^{Max(n,m)} q^i . \delta(b_i, c_i)$$

**Corollary 6.7.6** On putting $q = 1$ in $P(w_1, w_2)$, we get the distance between two words, that is, $D(w_1, w_2) = P(w_1, w_2)$ for $q = 1$.

**Example 6.7.7** Let $\Sigma = \{a < b < c\}$ be the ordered alphabet. Then for $w_1 = ababc$ and $w_2 = abccab$ we have

$$P(w_1, w_2) = q^3 + q^4 + q^5 + q^6$$

**Definition 6.7.8** Let $\Sigma = \{a < b < c < d < e\}$ and for the real given words $w_1 = dead$ and $w_2 = deed$, we have $D(w_1, w_2) = 1$ and $P(w_1, w_2) = q^3$.

**Definition 6.7.9** Let $\Sigma = \{a_1 < a_2 < a_3 .......... < a_k\}$ be an ordered alphabet and $w = a_1 a_2 .......a_n$. Two words $w_1$ and $w_2$ are said to be $w$-equivalent if both of the words have same distance polynomial with respect to $w$.

**Example 6.7.10** Let $\Sigma = \{a < b < c\}$ be the ordered alphabet. Then for given a word $w = ababc$, the words $w_1 = bbabc$ and $w_2 = cbabc$ are $w$-equivalent as both of the words have same distance polynomial , that is, $P(w, w_1) = q = P(w, w_2)$ with respect to the word $w = ababc$.

**Definition 6.7.11** Let $\Sigma = \{a < b < c < d < e\}$ and for a given real word $w = cede$, the real words $w_1 = dead$ and $w_2 = deed$ are $w$-equivalent as both of the words have same distance polynomial with respect to the word $w$, that is, $P(w,w_1)=q^1+q^3+q^4 = P(w,w_2)$ with respect to the word $w=cede$.

**Theorem 6.7.12** Let $\Sigma=\{a_1<a_2<a_3.........<a_k\}$ be an ordered alphabet. Given a word $w=b_1b_2.......b_n\,(b_i\in\Sigma\ \forall\ 1\le i\le n)$. The relation $R$ of $w$-equivalent defined in the set $\Sigma^*$ is an equivalence relation.

**Proof:**

Since every word is $w$-equivalent to itself, therefore the relation of $w$-equivalent defined in the set $\Sigma^*$ is reflexive.

Let $w_1,w_2\in\Sigma^*$ and $w_1\ R\ w_2$ , then

$w_1\ R\ w_2 \Rightarrow w_1$ is $w$-equivalent to $w_2$

$\Rightarrow P(w,w_1)=P(w,w_2)$

$\Rightarrow P(w,w_2)=P(w,w_1)$

$\Rightarrow w_2$ is $w$-equivalent to $w_1$

$\Rightarrow w_2\ R\ w_1$

The relation $w$-equivalent is symmetric relation.

Let $w_1,w_2,w_3\in\Sigma^*$. Let $w_1\ R\ w_2$ and $w_2\ R\ w_3$, then

$w_1\ R\ w_2$ and $w_2\ R\ w_3 \Rightarrow P(w,w_1)=P(w,w_2)$ and $P(w,w_2)=P(w,w_3)$

$\Rightarrow P(w,w_1)=P(w,w_3)$

$\Rightarrow w_1\ R\ w_3$

The relation $w$-equivalent is transitive relation.

Hence the relation $w$-equivalent defined in the set $\Sigma^*$ is an equivalence relation.

Thus every word $u$ in the set $\Sigma^*$ generates a class of words which is $w$-equivalent to the word $u$. In other words equivalence class generated by the word $u$, that is $[u]_R$, will be the set of all words which are $w$-equivalent to each other having the same distance polynomial $P(q)$

with respect to the word $w$. We shall use $Q_w(P(q))$ for the set of all words which are $w$-equivalent to each other having the same distance polynomial $P(q)$ with respect to the word $w$.

**Theorem 6.7.13** Let $\Sigma = \{a_1 < a_2 < a_3 .......... < a_k\}$ be an ordered alphabet. Given a word $w = a_1 a_2 ....... a_n$, the set $Q_w(P(q))$, where $P(q) = q^{r_1} + q^{r_2} + ....... + q^{r_n}$ $(1 \le r_i \le n \ \ \forall\ 1 \le i \le n)$, has the cardinality $(k-1)^n$ and for $P(q) = q^{r_1} + q^{r_2} + ....... + q^{r_n} + q^{r_{n+1}} + ........ q^{r_{n+m}}$ cardinality is $(k-1)^n k^m$.

**Proof:**

Let $P(q) = q^{r_1}$, which implies that there is only one letter different in all the words of the set $Q_w(P(q))$. So for a given letter there are $k-1$ different letters in $\Sigma$ as we have $k$ letters in the alphabet $\Sigma$. Therefore for a given word there are $k-1$ different possible words which are different in one letter only and the cardinality of the set will be $(k-1)$.

$P(q) = q^{r_1} + q^{r_2}$, which implies that there are two letters different in all the words of the set $Q_w(P(q))$. Two different letters can be placed in $(k-1)(k-1)$ different ways as for each letter $(k-1)$ different letters are possible. Therefore for a given word there are $(k-1)^2$ different words possible which are different in two letters and the cardinality of the set $Q_w(P(q))$ will be $(k-1)^2$.

Similarly for $P(q) = q^{r_1} + q^{r_2} + ....... + q^{r_n}$ there will be $n$ different letters in all the words. As for each letter $(k-1)$ different letters are possible, the total number of ways to replace $n$ different letters are $(k-1)^n$. Therefore for a given word there are $(k-1)^n$ different words possible which are different in $n$ letters and hence the cardinality of the set $Q_w(P(q))$ will be $(k-1)^n$.

For $P(q) = q^{r_1} + q^{r_2} + ....... + q^{r_n} + q^{r_{n+1}} + ........q^{r_{n+m}}$ the set $Q_w(P(q))$ contains words having length more than the length of the word $w$. For each extra letter we have $k$ ways to fill the place. For $m$ different letters we will have $k^m$ ways and total number of ways for assigning $n+m$ different letters are $(k-1)^n k^m$. Therefore the cardinality of the set $Q_w(P(q))$ will be $(k-1)^n k^m$.

**Corollary 6.7.14** Cardinality of the set $Q_w(P(q))$ for different polynomials counts the number of words possible which are different in one letter, two letters, three letters and so on.

## 6.8 Application to Natural Language Processing

Matrix representation of words for formal languages can be extended to natural languages as every natural language has its own set of alphabet. Language like 'English' has 26 letters in its alphabet and hence it is possible to find the matrix of every word. We shall utilize the concept of similarity of words in the study of morphology.

Morphology is the study of the way words are built up from smaller meaning bearing units, morphemes. A morpheme is often defined as the minimal meaning bearing unit. For example the word 'cat' consists of a single morpheme ( the morpheme *cat*)while the word 'cats' consists of two: the morpheme *cat* and the morpheme *-s*. Morphemes can be classified as stems and affixes. The stem is the main morpheme of the word, supplying the main meaning, while the affixes add additional meanings of various kinds

There are two broad classes of ways to form words from morphemes: inflection and derivation. Inflection is the combination of word stem with a grammatical morpheme, usually resulting in a word of the same class as the original stem. For example, English has the inflection morpheme *–s* for making the plural on nouns and the inflectional morpheme *–ed* for making the past tense on verb. Derivation is the combination of a word stem with a grammatical morpheme, usually resulting in a word of a different class, often with a meaning hard to predict exactly. For example the verb 'computerize' can take the derivational suffix –action to produce the noun 'computerization'.

In the present study we are concerned with regularly inflected verbs, the class of main and primary verbs that have same endings marking the same function. These regular verbs have four morphological forms: stem, *-s* form, *-ing* form, *- ed* form (Past participle). For example walk, walks, walking and walked.

**Proposition 6.8.1** Let $w = a_1 a_2 .....a_n \; (a_i \in set\ of\ \ English\ alphabet)$ be a regularly inflected verb over a set of English alphabet. A word $w_1$ is said to be morphological form of $w$ if

$$P(w, w_1) = q^{n+1} \;\; or \;\; q^{n+1} + q^{n+2} \;\; or \quad q^{n+1} + q^{n+2} + q^{n+3}$$

**Proof:** Given a regularly inflected verb, we can generate three morphological form for it by adding *-s*, *-ing*, *- ed*. Hence following distance polynomials may be generated:

$$P(w, w_1) = q^{n+1} \;\; or \;\; q^{n+1} + q^{n+2} + q^{n+3} \;\; or \;\; q^{n+1} + q^{n+2}$$

**Example 6.8.2** Let $w = play$, we have $n = 4$, then

$P(Play, Plays) = q^5$

$P(Play, Playing) = q^5 + q^6 + q^7$

$P(Play, Played) = q^5 + q^6$

Therefore, *plays, playing and played* are the morphological form of the regularly inflected verb *play*. On putting $q = 1$, we get the distance between the two words.

$P(Play, Plays) = 1$

$P(Play, Playing) = 3$

$P(Play, Played) = 2$

## 6.9 Conclusion

In the present chapter a word matrix mapping induced by another word has been defined which can be utilized for natural languages also. Similarity of words provides a computational way to find similar words for a given threshold of distance. Here we have considered morphological form of regularly inflected verbs only but adding certain restrictions, it can be generalized to all forms.

# References

Atanasiu, A., Martin-Vide C. and Mateescu A. (2001) Codifiable Languages and the Parikh Matrix Mapping. Journal of Universal Computer Science 7(8) 783-793.

Atanasiu, A., Martin-Vide C. and Mateescu A. (2001) On the injectivity of the Parikh matrix mapping. Fundamenta Informaticae, 46, 1-11.

Bahl, L. R., Jelinek, F. and Mercer, R. L. (1983) A Maximum Likelihood Approach to Continuous Speech Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 5(2), 179–190.

Bar-Hillel, Yehoshua (1970) Aspects of language: essays in philosophy of language, linguistic philosophy and methodology of linguistics. Jerusalem: Magnes press and Amesterdam: North –Holland.

Baum, L. E. and Petrie, T. (1966) Statistical Inference for Probabilistic Functions of Finite State Markov Chains. Annals of Mathematical Statistics 37, 1559–1563.

Bellman, R.E. & Zadeh, L. A. (1970) Decision making in fuzzy environment, Management Science. 17(4). 141-164.

Bisht R. K., Dhami H. S. and Tiwari Neeraj (2006) An Evaluation of Different Statistical Techniques of Collocation Extraction Using a Probability Measure to Word Combinations, Journal of Quantitative Linguistics, Volume 13, Numbers 2 – 3, 161 – 175.

Bisht R. K. and Dhami H. S. (2007) Collocation extraction from a small sample of text, to be appear in the proceedings of 12th annual conference conference of Gwalior Academy of mathematical Sciences.

Bisht R. K. and Dhami H. S. (2008) On Some properties of content words in a document, to be appear in the proceedings of 6th international conference of information science technology and management.

Black, E., Jelinek, F., Lafferty, J. D., Magerman, D. M., Mercer, R. L. and Roukos, S. (1992) Towards History-Based Grammars: Using Richer Models for Probabilistic Parsing. In Proceedings DARPA Speech and Natural Language Workshop, Harriman, New York, pp. 134–139. Los Altos, CA: Morgan Kaufman.

Bod, R. (1999) Beyond Grammar: An Experience-Based Theory of Language. Cambridge University Press.

Bookstein, D.R. Wanson, (1974) Probabilistic models for automatic indexing. Journal of the American Society for Information Science, 25, 312-318.

Bookstein, D.R. Wanson, (1975) A decision theoretic foundation for indexing. Journal of the American Society for Information Science, 26, 45-50.

Brill, E. (1995) Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. Computational Linguistics, 21(4), 543–566.

Brown, P. F., Della Pietra, V. J., deSouza, P. V. and Mercer, R. L. (1990) Class-Based N-Gram Models of Natural Language. In Proceedings of the IBM Natural Language ITL, pp. 283–298. Paris, France.

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Lafferty, J., Mercer, R. and Rossin, P. (1990) A Statistical Approach to Machine Translation. Computational Linguistics 16(2), 79–85.

Brown, P., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. (1993) The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics 19(2), 263–311.

Cambridge International Dictionary of Idioms. UK. CUP. (1998).

Charniak, E. (1997) Statistical Parsing with a Context-Free Grammar and Word Statistics. In Proceedings of the 14th National Conference on Artificial Intelligence (AAAI-97). Menlo Park: AAAI Press.

Chen, K. Kishida, H. Jiang and Q. Liang (1999) Automatic construction of a Japanese-English lexicon and its application in cross-language information retrieval. In ACM DL/ACM SIGIR Workshop on Multilingual Information Discovery and Access (MIDAS).

Chisholm, E., Kolda Tamara G (March 1999) New term weighting formulas for the vector space method in information retrieval. Technical Report ORNL-TM-13756, Oak Ridge National Laboratory, Oak Ridge, TN.

Chomsky, Noam (1994) In A Companion to the Philosophy of Mind, edited by Samuel Guttenplan, Oxford: Blackwell Publishers 153-167.

Choueka, Y., Klien, T. and Neuwitz, E. (1983). Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. Journal for Literary and Linguistic computing, Vol 4. 34-38.

Church Kenneth W. and Hanks, Patrick. (1989) Word association norms, mutual information and lexicography. In Proceedings of the 27th meeting of the Association of Computational Linguistics 76-83.

Church, K. W. (1988) A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. Second Conference on Applied Natural Language Processing, ACL.

Church, K. W. and Gale, William A. (1991). Concordance for parallel text. In proceedings of the seventh annual conference of the UW centre for new OED and text research, Oxford. 40-62. Computational Linguistics 21(2), 165–202.

Church, K. W. and Mercer, R. L. (1993) Introduction to the Special Issue on Computational Linguistics Using Large Corpora. Computational Linguistics 19, 1–24.

Church, K. W., William A. Gale, (1995) Poisson Mixture. Nature Language Engineering, 1, 163-190.

Church, K.W., P. Hanks, D. Hindle, W. Gale, and R. Moon (1994) Substitutability. In Computational Approaches to the Lexicon, pages 153-180.

Cormen, T. H., Leiserson, C. E. and Rivest, R. L. (1990) Introduction to Algorithms. MIT Press.

Crowell, R.H. and Fox R.H. (1997) Introduction to Knot Theory. Springer-verlag, New York.

Cutting, D., Kupiec, J., Pedersen, J. and Sibun, P. (1992). A Practical Part-of-speech Tagger. In Third Conference on Applied Natural Language Processing, ACL, 133–140.

Dagan L. and K. Church (1994) Termight: Identifying and translation technical terminology. In Proc. of the 4th Conference on Applied Natural Language Processing (ANLP), pages 34-40, Stuttgart, Germany.

Dagan, L. Lee, and F. Pereira (1999) Similarity-based models of word co-occurrence probabilities. Machine Learning, 34(1).

Dale, R., Moisl, H. and Somers, H. (eds.) (2000) Handbook of Natural Language Processing. Marcel Dekker.

Darren, Pearce (2001) Synonymy in collocation extraction. In Proc. of the NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations, CMU.

Darren, Pearce (2001) Using conceptual similarity for collocation extraction. In Proc. of the 4th UK Special Interest Group for Computational Linguistics (CLUK4).

Darren, Pearce and John Carroll (2001) Social butterflies, dark horses and busy bees: Collocation extraction using lexical substitution tests. Ms, Sussex University.

Debra, S. B. and Martha W. Evens (1998) Finding phrases rather than discovering collocations: Searching corpora for dictionary phrases. In Proc. of the 9th Midwest Artificial Intelligence and Cognitive Science Conference (MAICS'98), Dayton, USA.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society 39, 1–38.

Dias, S. Guilloré, J-C. Bassano, and J.G. Pereira Lopes (2000) Combining linguistics with statistics for multiword term extraction: A fruitful association? In Proc. of Recherche d'Informations Assistee par Ordinateur 2000 (RIAO'2000).

Dominich Sandor (2008) The modern Algebra of information retrieval, Springer –Verlag, Berlin, Heidelberg.

Dunning, Ted. (1993). Accurate methods for the statistics of surprise and coincidence. Computational Linguistics. Vol 19. 61-74.

Edmundson, H. P. (1968) Mathematical Models in Linguistics and Language Processing. In Borko, H. (ed.) Automated Language Processing. John Wiley and Sons.

Egecioglu O. (May 2004) A q-matrix encoding extending the Parikh matrix mapping, Proceeding of international conference on Computer and Communication (ICCC 2004), Oradea, Romania.

Evert, Stefan and Krenn, Brigitte (2001) Methods for the qualitative evaluation of lexical association measures. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, Toulouse, France, 188-195.

Evert, Stefan and Krenn, Brigitte (2003) Computational approaches to collocations. Introductory course at the European Summer School on Logic, Language, and Information (ESSLLI 2003), Vienna. (www.collocations.de)

Fung and Yee L.Y. (1998) An IR approach for translating new words from nonparallel, comparable texts. In Proc. of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics: COLING/ACL-98, 414-420.

Fung and. McKeown K.R. (1997) Finding terminology translations from non-parallel corpora. In Proc. of the 5th Annual Workshop on Very Large Corpora, 192-202.

Fung, Kan, M-Y. and Horita Y. (1996) Extracting Japanese domain and technical terms in relatively easy. In Proc. of the 2nd International Conference on New Methods in Natural Language Processing, 148-159.

Gale, W. A., Church, K. W. and Yarowsky, D. (1992) A Method for Disambiguating Word Senses in a Large Corpus. Computers and the Humanities 26, 415–439.

Haruno, S. Ikehara, and T. Yamazaki (1996) Learning bilingual collocations by word-level sorting. In Proc. of the 16th International Conference on Computational Linguistics (COLING '96), Copenhagen, Denmark.

Hindle (1990) Noun classification from predicate-argument structures. In Proc. of the 28th Annual Meeting of the ACL, pages 268-75.

Igor Mel'cuk (1998) Collocations and lexical functions. In Phraseology: Theory, Analysis, and Applications. Oxford: Clarendon Press. 23-54.

Jelinek, F. (1976) Continuous Speech Recognition by Statistical Methods. Proceedings of the IEEE 64(4), 532–557.

Johansson. Good bigrams (1996) In Proc. of the 16th International Conference on Computational Linguistics (COLING '96), Copenhagen, Denmark, 592-597.

John S. Justeson and Slava M. Katz (1995) Technical terminology: Some linguistic properties and an algorithm for identification in text. Natural Language Engineering, 1:9-27.

Jurafsky Daniel, Martin James H. (2004) Speech and Language Processing: An Introduction to natural Language Processing. Computational Linguistics and Speech Processing, Pearson Education Singapore.

Kathleen R. McKeown and Dragomir R. Radev (2000). http://citeseer.ist.psu.edu/mckeown00collocations.html

Katz Slava M., (1996) Distribution of content words and phrases in text and language modeling, Natural Language Engineering, 2(1), 15-59.

Keita Tsuji and Kyo Kageura (2001) Extracting morpheme pairs from bilingual terminological corpora. Terminology, 7(1):101-14.

Kita K. and H. Ogata (1997) Collocations in language learning: Corpus-based automatic compilation of collocations and bilingual collocation concordancer. Computer Assisted Language Learning: An International Journal, 10(3):229-38.

Kita K., Kato Y., Omoto T., and Yano Y. (1994) A comparative study of automatic extraction of collocations from corpora: Mutual information vs. cost criteria. Journal of Natural Language Processing, 1(1):21-33.

Klir, George J. and Yuan Bo. (2001) Fuzzy sets and fuzzy logic theory and application. Prentice Hall of India.

Kornai Andras (2008) Mathematical linguistics: Advanced information and knowledge processing series, Springer XIV 290 pages.

Krenn, Brigitte and Evert, Stefan (2001). Can we do better than frequency? A case study on extracting PP-verb collocations. In Proceedings of the ACL Workshop on Collocations, Toulouse, France, 39-46.

Kyo Kageura (1997) On intra-term relations of complex terms in the description of term formation patterns. In Melanges de Linguistique Offerts a R. Kocourek, Halifax: Les Presses d'ALFA, 105-111.

Kyo Kageura (1998) A statistical analysis of morphemes in Japanese terminology. In Proc. of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics (COLING/ACL-98), pages 638-45, Montreal, Canada.

Lambek J.  (2004) A computational algebraic approach to English grammar, Syntax, 7(2), 128-147.

Lin D. (1998) Using collocation statistics in information extraction. In Proc. of the Seventh Message Understanding Conference (MUC-7).

Lin D. (1999) Automatic identification of non-compositional phrases. In Proc. of the 37th Annual Meeting of the ACL, College Park, USA. 317-324.

Lin, D. (1998) Extracting collocations from text corpora. In first workshop on Computational terminology, Montreal, Canada.

Lothaire M. (1997) Combinatorics on Words, Cambridge University press, Cambridge.

Luhn, H.P. (1958) The automatic creation of literature abstracts. IBM Journal of Research and Development, 2, 159-165.

Manning, C. D. and Schutze H. (2002). Foundations of Statistical Natural Language Processing. MIT Press.

Manning C. D., Raghavan P. and Schutze H. (2008) Introduction to Information retrieval, cambridge university press, New York.

Mateescu A, Salomaa A., Salomaa K and Sheng Yu (February 2002) Subword Histories and Parikh Matrices, TUCS Technical report No. 442.

Mateescu A, Salomaa A., Salomaa K and Sheng Yu, On an extension of the Parikh mapping, TUCS Technical report No. 364.

Mateescu A., Salomaa A., Salomaa K. and Sheng Yu (2001) A sharpening of the Parikh mapping, RAIRO-Theoretical Informatics and Applications 35, 551-564.

Mayfield Tomokiyo and K. Ries (1997) What makes a word: learning base units in Japanese for speech recognition. In Proc. of the Conference on Computational Natural Language Learning (CoNLL-97), 60-69.

Maynard, D. and Ananiadou S. (1999) Identifying contextual information for multi-word term extraction. In 5th International Congress on Terminology and Knowledge Engineering (TKE 99), 212-21.

Melamed, I.D. (1998) Empirical methods for MT lexicon development. In Proc. of AMTA'98: Conference of the Association for Machine Translation in the Americas, pages 18-30.

Melamed, I.D. (2000) Models of translational equivalence among words. Computational Linguistics, 26(2):221-49.

Merialdo, B. (1994) Tagging English Text with a Probabilistic Model. Computational Linguistics 20(2), 155–172.

Merkel M. and Andersson M. (2000) Knowledge-lite extraction of multi-word units with language filters and entropy thresholds. In Proc. of Recherche d'Informations Assistee par Ordinateur 2000 (RIAO'2000).

Merkel M., Nilsson B., and Ahrenberg L. (1994). A phrase-retrieval system based on recurrence. In Proc. of the 2nd Annual Workshop on Very Large Corpora, pages 99-108.

Morton Benson, Evelyn Benson, and Robert Ilson (1986) The BBI Combinatory Dictionary of English: A Guide to Word Combinations. John Benjamins, Amsterdam, Netherlands.

Moens M. F. (2006) Information extraction: Algorithms and prospects in retrieval context, Springer, Nitherlands.

Murthy Kavi Narayana and Kumar G. Bharadwaja (2006) Language Identification from Small Text Samples. Journal of Quantitative Linguistics, Vol.13, Number 1, pp. 57 – 80

Nivre, J. (2000) Sparse Data and Smoothing in Statistical Part-of-Speech Tagging. Journal of Quantitative Linguistics 7(1), 1–18.

Pantel P. and D. Lin. Word-for-word glossing with contextually similar words. In Proc. of the 1st Annual Meeting of the North American Chapter of Association for Computational Linguistics (NAACL2000), pages 78-85, Seattle, USA, 2000.

Parikh R. J., (1966) On context free languages, Journal of the Association for Computing Machinery, 13 , 570-581.

Patricia Guilpin and Christian Guilpin (2005) Linguistic and Statistical Analysis of the Frequency of a Particular Word at Different Times (Diachrony) or in Different Styles (Synchrony), Journal of Quantitative Linguistics, Vol. 12, No. 2-3, pp. 138 – 150

Randolph Quirk (1984) Longman Dictionary of the English Language, Longman, London.

Reinhard Kohler and Burghard B. Rieger (editors)(University of Trier) (1993) Contributions to Quantitative Linguistics, Dordrecht: Kluwer Academic Publishers, pp.436.

Ries K., Buo F.D., and A. Waibel (1996) Class phrase models for language modelling. In Proc. of the 4th International Conference on Spoken Language Processing (ICSLP'96).

Roberto Basili, Maria Teresa Pazienza, and Paola Velardi (1993) Semi-automatic extraction of linguistic information for syntactic disambiguation. Applied Artificial Intelligence, 7, 339-364.

Rorbert M. Losee, (2001) Term Dependence: A Basis for Luhn and Zipf Models, Journal of the American Society for Information Science and Technology, 52(12), 1019-1025.

Salomaa A. (August 2004) Connections between subwords and certain matrix mappings TUCS Technical report No. 620.

Salomaa A. (February 2005) On languages defined by numerical parameters, TUCS Technical report No. 663.

Salomaa A., Ding Cunsheng (August 2005) On some problems of Mateescu concerning subword occurrences, TUCS Technical report No. 701.

Salomaa A., Sheng Yu (November 2004) Subword conditions and subword histories, TUCS Technical report No. 633.

Samuel, K., Carberry, S. and Vijay-Shanker, K. (1998) Dialogue Act Tagging with Transformation-Based Learning. In Proceedings of the 17th International Conference on Computational Linguistics (COLING-14), pp. 1150–1156.

Samuelsson, C., Tapanainen, P. and Voutilainen, A. (1996) Inducing Constraint Grammars. In Miclet, L. and de la Higuera, C. (eds) Grammatical Inference: Learning Syntax from Sentences, Lecture Notes in Artificial Intelligence, Springer. 1147, 146–155.

Schütze, H. (1998) Automatic Word Sense Discrimination. Computational Linguistics 24, 97–237.

Serbanuta T.-F, (2004) Extending Parikh matrices Theoretical Computer science, 310-1, 233-246.

Shannon, C. E. (1948) A Mathematical Theory of Communication. Bell System Technical Journal 27, 379–423, 623–656.

Smadja, F.A. and McKeown, K.R. (1990) Automatically extracting and representing collocations for language generation. In Proc. of the 28th Annual Meeting of the ACL, 252-259.

Smadja, Frank (1993) Retrieving collocations from text: Xtract. Computational Linguistics. Vol 19(1). 143-177.

Smadja, K.R. McKeown, and V. Hatzivassiloglou (1996) Translating collocations for bilingual lexicons: A statistical approach. Computational Linguistics, 22(1):1-38.

Sophia Ananiadou (1994) A methodology for automatic term recognition. In Proc. of the 15th International Conference on Computational Linguistics (COLING '94), pages 1034-8, Kyoto, Japan.

Spark Jones K., (1972) A statistical interpretation of term specificity and its application in retrieval, Journal of Documentation, 28, 111-21.

Spark Jones K., S. Walker and S.E. Robertson (1998) A probabilistic model of information retrieval: Development and status. Technical Report 446, University of Cambridge Computer Laboratory.

Strazny Philip ed. (2005) Encyclopedia of linguistics, New York: Fitzroy Dearborn, vol 1, 124-126,

Tanaka and Y. Matsuo (1999) Extraction of translation equivalents from non-parallel corpora. In Proc. of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99), 109-119.

Tremblay J. P., R.Manohar (1997) Discrete Mathematical Structures with Applications to Computer Science, Tata McGraw-Hill , New Delhi.

Velardi P., Pazienza M.T., and M. Fasolo. How to encode semantic knowledge: A method for meaning representation and computer-aided acquisition. Computational Linguistics, 17(2):153-70, 1991.

Wilks, Y. (ed) Theoretical Issues in Natural Language Processing 3. Hillsdale, NJ: Lawrence, 185–191.

Yamamoto M.and Church K.W. (2001). Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus. Computational Linguistics, 27(1), 1-30.

Yarowsky, D. (1992) Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. In Proceedings of the 14th

International Conference on Computational Linguistics (COLING-14), 454–460.

Yeh, Alexander (2000). More accurate tests for the statistical significance of result differences. In Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000). Saarbrücken, Germany.

Zanette Damia´n H. and Montemurro Marcelo A. (2005) Dynamics of Text Generation with Realistic Zipf's Distribution, Journal of Quantitative Linguistics , Vol. 12, No. 1, pp. 29 – 40

Zellig Harris (1982) A Grammar of English on Mathematical Principles, Wiley, New York.

Zellig harris (1982) A grammar of English on mathematical principles, Wiley, New York.

Zhang, J., Gao, J. and Zhou M. (2000) Extraction of Chinese compound words - an experimental study on a very large corpus. In Proc. of the 2nd Chinese Language Processing Workshop, ACL 2000.

# Appendix

Following novels have been taken for corpus from Project Gutenberg.

1. Title: Bullets & Billets Author: Bruce Bairnsfather [eBook #11232]
2. Title: Radio Boys Cronies, Author: Wayne Whipple and S. F. Aaron [eBook #11861]
3. Title: An Old Maid Author: Honore de Balzac [eBook #1352]
4. Title: Behind the Line Author: Ralph Henry Barbour [eBook #13556]
5. Title: The Light in the Clearing Author: Irving Bacheller [eBook #14150]
6. Title: When William Came Author: Saki [eBook #14540]
7. Title: The Marriage Contract Author: Honore de Balzac [eBook #1556]
8. Title: The Historical Nights Entertainment, Second Series [eBook #7949]
9. Title: The Highwayman, Author: H.C. Bailey [eBook #9749]
10. Title: The Happy Foreigner Author: Enid Bagnold [eBook #9978]
11. Title: Mr. Bonaparte of Corsica Author: John Kendrick Bangs [eText #3236]
12. Title: Love-at-Arms Author: Raphael Sabatini [eText #3530]
13. Title: The Toys of Peace by H.H. Munro ("Saki") [eText #1477]
14. Title: The Lion's Skin Author: Rafael Sabatini [eText #2702]
15. Title: The Lost City by Joseph E. Badger, Jr. [eText #783]
16. Title: The Muse of the Department by Honore de Balzac [eText #1912]
17. Title: The Master of Silence Author: Irving Bacheller [eBook #7486]
18. Title: The Unbearable Bassington Author: Saki [eBook #555]

Note: Footnotes and illustrations have been removed while making corpus.