IDENTIFICATION OF DIFFICULT ENGLISH WORDS FOR ASSISTING HEARING IMPAIRED CHILDREN IN LEARNING LANAGUAGE

Author

Munazza Ansar

00000201975

MS-16(CSE)

Supervisor

Dr. Usman Qamar

Department of Computer Engineering

College of Electrical and Mechanical Engineering

National University of Sciences and Technology

Islamabad

January, 2020

بسم الله الرحمن الرحيم

In the name of Allah most beneficent most merciful

وَلَا يُحِيطُونَ بِشَىۡءٍ مِّنۡ عِلۡمِهِۦٓ إِلَّا بِمَا شَآءَ ۚ

And they can'tencompass any thing from His knowledge, but to extend He wills [2:255]

IDENTIFICATION OF DIFFICULT ENGLISH WORDS FOR ASSISTING HEARING IMPAIRED CHILDREN IN LEARNING LANAGUAGE

Author
Munazza Ansar
00000201975

A thesis submitted in partial fulfillment of the requirements for the degree of
MS Computer Software Engineering

Thesis Supervisor:
DR. USMAN QAMAR
Thesis Supervisor's
Signature:_______________________________________

DEPARTMENT OF COMPUTER ENGINEERING
COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY,
ISLAMABAD
January, 2020

Language Correctness Certificate

This thesis has been read by an English expert and is free of typing, syntax, semantic, grammatical and spelling mistakes. Thesis is also according to the format given by the university.

Signature of Student
Munazza Ansar
Registration Number
00000201975

Signature of Supervisor
Dr. Usman Qamar

Copyright Statement

This page is intentionally left blank

ACKNOWLEDGEMENTS

“This is by the Grace of my Lord to test me whether I am grateful or ungrateful! And whoever is grateful, truly, his gratitude is for (the good of) his own self, and whoever is ungrateful, (he is ungrateful only for the loss of his own self). Certainly! My Lord is Rich (Free of all wants), Bountiful” [An-Naml: 40]

I am indebted to NUST College of Electrical and Mechanical Engineering, for providing me an opportunity for Masters Research. First and foremost I offer my sincerest gratitude to my thesis supervisor, **Dr. Usman Qamar,** who has supported me throughout my thesis, with his patience and knowledge. I would also like to thank **Dr. Muhammad Abbas** and **Dr. Wasi Haider Butt** for being on my thesis guidance and examination committee and also for guidance and cordial support, which helped me in completing this task through various stages. I would also like to thank my family for the support they provided me through my entire life.

Dedicated
to

To my Parents, Brothers, Sister and Advisors

Abstract

Purpose: *This thesis presents the need, purpose, approaches and results for identification of difficult words from English text. With the passage of time, Human Language Technology (HLT) helps disabled, foreigner's and low literate individuals to enhance the communication skills, and learn the language efficiently along with the use of computers and other technologies. Natural language processing part of HLT is an emerging field of computer science that is widely used for processing of unstructured text. In an educational domain, difficult words may not only affect the reading, writing , understandability and interpretation of the text but also results in poor academic achievement of the Hearing Impaired Children when compared to their normal hearing peers. This happens because they lack increased knowledge of vocabulary. Speech language Pathologist/Therapist (SLP), their teachers and parents indulge them in different learning activities to increase their vocabulary. Presenting simple text to such children will help them to use simple words in daily routine to enhance their vocabulary knowledge as they can learn more words in short period of time. It will also help their parents and teachers to prepare reading and writing materials, simpler to learn, for them .It will also help the child to learn language in simpler way. This motivates the need of technique(s) to classify words as difficult or not difficult from the text available in English textbooks, or online study material available for them. So, the prime objective of this research work is to propose and develop a methodology or technique that assist to identify difficult English words from text.* ***Methodology:*** *We proposed a methodology for identification of difficult words from the English text in order to assist hearing impaired children in learning the language in simplest way. After preprocessing of the text and implication of feature extraction technique to extract features based on linguistic rules specific to hearing impaired children, C4.5 decision tree machine learning algorithm is used to classify words as difficult or not difficult from the given text. Proposed technique is applied on different text documents to evaluate its effectiveness.* ***Results:*** *92.5% accuracy is achieved when model is evaluated against annotated tested dataset1 specific to hearing impaired children. Whereas 5-fold cross validation method gives an accuracy of 94.2%. This depicts that for remaining words which are unable to identify by our proposed methodology are considered as errors due to non availability of linguistic rule in training model. It is evaluated that accuracy is strongly dependent on datasets during classification of difficult words.*

Keywords: Classification, Difficult words, Hearing Impairment, Natural Language Processing, Decision Tree

TABLE OF CONTENTS

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

Human language technology is an emerging and challenging field of research that enables the computers to produce process and comprehend the natural language automatically. It constitutes of either text or speech technology or combination of both. It involves sub disciplines of linguistic, natural language processing, computer sciences, artificial intelligence, mathematics, philosophy, and statistics. Whenever, an individual needs to interact with computer or any other technology, he she not only requires the broad knowledge of linguistics but also knows a little bit of information about computer sciences and other fields too. Some of the application of Human Language technology comprises of

- Language translators(English to Chinese or vice versa)
- Device Controllers through speech
- Automatic dictation for hands free texting
- Language learner programs for special person
- Web search engines
- Information management systems
- Spelling and grammar checker
- Lexical Simplification of the text
- Text summarization

HLT provides an access to large amount of information that cannot be accessible by most of the people in past. It also helps disabled, foreigner's and low literate individuals to enhance the communication skills, and learn the language efficiently along with the use of computers and other technologies [1]. With the advancement of such technologies, the focus is towards the automation of system. Manual system of every domain is being automated. Manual system is time taking and hectic as well. Automated system helps in improving performance of any organization working in any domain. Widespread use of internet, computers and modern technologies make it possible to automate whole organization [2]. For example, manual Sale's record storage of any product's store can be difficult to handle huge amount of sales as opposed to the point of sale system (automated system for handling sales) that save, process and store thousands of sales daily efficiently and effectively. Similarly, manual records can be hectic and can contain complex or incorrect data, which leads toward the misguidance or misinterpretation [3].

Natural Language processing is widely used, research oriented and emerging topic for processing of unstructured or free-text record. Research on natural language processing was started in the late 1940s. Russian to English translation was done in 1954. More than 60 sentences from Russian language were converted into English language. Language rules, statistics and computational linguistics were used for conversion. Researchers declared that translation was not efficient as there were problems, so improvement is required [4]. Now a day's research is being done on natural language processing, and many techniques were proposed having different accuracy.

In education domain, it significantly improves the methods of teaching by providing interactive and interesting study materials, generation of different scenarios/activities fulfilling educational needs in different languages. Engineers and scientist with the use of these computing technologies help in assisting the people with hearing impairment by developing different software and techniques. Speech canvases [5], 3D games [6], 3-D talking head [8], Articulation Vacation, Carnival application [7] are some of the computerized systems/applications developed to support speech therapies and language learning activities. All of these are based to enhance the language or vocabulary knowledge of hearing impaired children and fulfills their educational needs to some extent. Through use of language learning software's/expert systems, students can develop better understanding of study materials. Complex study materials can easily be transformed into simple text with the use of such technology. Lexical Text simplification is the vast field of linguistics and has significant importance in education domain. Simple text is not only easy to understand by normal person but also help in enhancing learning skills and fulfills the educational needs of disable person in society. Various techniques or systems are developed for text simplification in different domains and in different languages. It comprises of two important steps, one is identification of difficult words, and other is generating their simpler substitution. During simplification, complexity from the text is removed keeping in view the original meaning of the text. Through text simplification, quality of the text is improved which ultimately results in enhancing comprehensibility of the text. Another advantage of text simplification is that simplified text is available and accessible to large no of people belonging from different domains, cultures and societies. It is specifically useful for the people having language learner disabilities.

1.1 Background and Motivation

Disability in humans refers to mental or physical impairments that gradually restrict one's ability to carry out his/her daily routine activities. It includes disability related to hearing, vision, mobility, Spinal cord, learning, psychological disorders etc.

Hearing Impairment is a type of disability in which an individual is unable to hear and comprehend any sound partially or completely. A person who is suffering from hearing problem may find it difficult to articulate it. It means it is difficult for such individual to produce sounds in syllables as well as using, replicating, and understanding language. Children with hearing impairment often have some degree of speech and/or language delay. They face language impairment. Children development in aspect of learning communication, reading, speaking, listening and writing is greatly affected in earlier hearing loss as compared to hearing loss at later stage in life. Earlier detection of hearing loss in an individual and commencement of intervention by the supporters results in less serious ultimate impact. The language inconsistency causes learning problems that result in reduced academic achievement of such children. Academic achievement depends on parent involvement during their academic sessions and the quantity, quality and timing of the services like speech therapies, activities to enhance learning capabilities etc. in which such children are engaged [10] [67] [68] [69].

As we know language is the only and the best means of communication for human interaction. But it is really difficult for hearing impaired children to learn the language and acquiring it in their daily routine. Language comprises of vocabulary and vocabulary development in hearing impaired children is slow due to less acquisition of knowledge as compared to normal hearing children. They find it hard to hear the individuals and understand their speech. Acquiring less number of words in daily routine hinders hearing impaired children in forming new words on their own. This result in limited knowledge and this is also because of the facts they are unable to learn new words and only remember already known words because of the efforts of their parents and teachers [67] [68] [69].

Hearing impaired children usually show less development in communication, reading and writing skills when compared to the normal hearing children. As we know reading and writing is an important activity which enable individual to recognize, reproduce, learn and acquire language easily and efficiently. The core factors of the text difficulty include pronunciation or phonology of the language along with the semantics for the hearing impaired children [67].

Hearing impaired children are unable to follow deviation from standard English grammar structure known as subject, verb, and object (SVO). So, when parent, teacher or an individual asked them to read or write, it is very difficult and tedious task for them as explained in [12].

1.2 Objective and Contribution

Anything that is made by the combination of characters, words, sentences, and paragraphs is considered as text. Sentences are made by the combination of complex, simple words or both. Usage of more complex words in the text makes it a difficult one which causes problems not only in reading but in writing as well. This ultimately results in poor academic achievement of students. It is more difficult for hearing impaired children to learn and understand the difficult text as compared to their normal hearing peers. This happens because they lack increase knowledge of vocabulary. Speech language Pathologist/Therapist, their teachers and parents indulge them in different learning activities to increase their vocabulary. During this, they usually used flash cards, containing different words, pictures, printed papers, real objects and toys. But if presented words are difficult for them to learn, then it will be very difficult for others to develop understanding of the word in child’s mental lexicon. This will also require a lot of effort and time not only from SLP’s, parents, and teachers but also from hearing impaired students. Efforts are being made to propose methodologies and techniques to detect the difficult/complex words from the free text available in different domains for example for second language learners, for aphasic readers, for multilingual students and in different languages like German, English, Chinese, Swedish, Spanish. Efforts are also being made to propose technologies to increase accuracy of detection of difficult words. But very few efforts are being made relevant to Hearing Impaired Children.

In this thesis, we focused on proposing a methodology for identification of difficult words from the text with significant accuracy for hearing impaired children. Just the words or terminologies from English text are considered for identification, words from other domains are not considered.

Our proposed methodology consists of four parts, collection of dataset, preprocessing of text, feature extraction based on linguistic rules, and identification of difficult English words using machine learning algorithm. In preprocessing part, tagging / categorization of part of speech technique enables to differentiate parts of speech from the written text document. Different linguistic rules explicitly for hearing impaired children are considered for feature extraction. A C4.5 decision tree, a machine learning algorithm is used for classification of difficult words during training and testing phase. Later accuracy of the proposed methodology will be checked. For technique to be implemented, C# will used for the coding purpose.

This thesis will help in identification of difficult words to make the readability of the text easy for such children. It will help child to use simple words in daily routine to enhance their vocabulary knowledge as they can learn more words in short period of time. It will also help their parents and teachers to prepare reading and writing materials simpler to learn for them .It will help the child to learn language in simple way.

In Pakistan, Educational institutes specific for hearing impaired children, adults, children, parents and Speech Language Pathologists (SLP’s) can effectively make use of the proposed methodology in daily routine to identify difficult text from the study materials, newspapers, online on web stores and can prepare activities during speech therapy. It will save time and resources currently utilized manually. It will also enhance child’s vocabulary as they can learn simple words rather than focusing and spending hours on difficult ones.

1.3 Outline

Chapter 2 discusses the work or research done on our research topic. Hearing loss, automatic systems developed for enhancing learning, listening and communication skills of such children, natural language processing used in different systems and difficult word identification systems to provide simplified text to the desired population are discussed in detail. Chapter 3 discusses the proposed methodology is detail. Techniques used during identification process are also described in detail. Chapter 4 discusses the results achieved by applying proposed methodology. In chapter 5 conclusion and future work is discussed.

1.4 Summary

In this chapter, importance of correctness of automated system is described. Natural language processing is widely used and research-oriented topic. In education domain many automated systems are used for assisting person with hearing disability. It is also described that text is more difficult for hearing impaired children due to lack of vocabulary knowledge. Presenting difficult test for reading and writing to hearing impaired children not only hinders their development in learning language but also affect their academic achievement.

Chapter 2

Literature Review

In this chapter, we will discuss in detail about the hearing disability and its overall impact on human development. This chapter is divided into five important sections as follow.

1) Detailed description of Hearing Loss and its impact on child's development
2) Role of Information Technology and development of automated systems for Hearing Impaired Children
3) Spoken language acquisition and vocabulary Knowledge for Hearing impaired children
4) Natural Language Processing and its importance
5) Text Difficulty for hearing impaired children and implication of machine learning for identification of difficult words

2.1 Hearing Loss

Disability in humans refers to mental or physical impairments that gradually restrict one's ability to carry out his/her daily routine activities. It includes disability related to hearing, vision, mobility, Spinal cord, learning, or psychological disorders etc. Hearing impairment is a partial or total inability to hear and comprehend sound. World Health Organization estimates show that almost 466 million people around the globe facing hearing impairment and this gradually increases up to 900 million people by 2025 [13]. Out of these 466 million people, 34 million include children population.

2.1.1 Causes of Hearing Loss

Many factors are involved in causing hearing impairment to an individual. Mostly caused due to frequent exposure of an individual to very loud noise, damage to the middle ear, physical head injury or trauma, diseases like high fever meningitis may damage choclea, use of Oxotoxic medication, Collection of fluid/wax (Otitis media), aging or genetically inherited from the parents [14].

2.1.2 Types of Hearing Loss

2.1.2.1 Mixed hearing loss

Mixed hearing loss is an amalgamation of sensorineural and conductive types of hearing loss. This occurs due to long lasting injury to the outer, middle and inner ear (cochlea) or sometimes due to damage in auditory nerve [14].

2.1.2.2 Conductive hearing loss

In this type of hearing loss, sound is not conducted efficiently through the outer ear canal to the eardrum and the tiny bones of the middle ear. It can often be treated medically or surgically.

Causes of conductive hearing loss includes:-

- Ear infections
- Allergies
- Ear wax
- Infections in the ear canal
- Abnormality of the outer ear, ear canal, or middle ear [14].

2.1.2.3 Sensorineural hearing loss

Damage or injury in the inner ear i.e. cochlea or in auditory nerve pathways that carry sound wave signals from inner ear towards brains causes sensorineural hearing loss. This often leads to permanent type of hearing loss and cannot be treated / corrected medically or surgically. Due to this, ability to hear a faint sound reduces gradually even if sounds in amplified or loud enough for hearing purpose [14].

Factors involve are as follow:

- Illnesses
- Drugs that are toxic to hearing
- Hearing loss that runs in the family (genetic or hereditary)
- Aging
- Exposure to loud noise

2.1.2.4 Central hearing loss

Central hearing loss results from damage or impairment to the nerves or nuclei of the central nervous system, either in the pathways to the brain or in the brain itself [14].

2.1.3 Levels of Hearing Loss

There are four different levels of hearing loss. Ability of an individual to hear distinct type of sound is measured in decibels (dB) [15].Hard of hearing or partial impairment term refer to those who suffers from mild and moderate hearing loss. Sometimes term also used for the ones who face severe hearing loss but not usually. Hearing Impairment persons are able to communicate through spoken language.

2.1.3.1 Mild hearing loss

In case of mild loss, person is unable to hear soft, faded or distant speech. It is difficult for an individual to follow the conversation in a noisy environment. They need to ask for repetition .A person with this loss uses a hearing aid to amplify sounds. Hearing level ranges from 26 to 40 decibels [15].

2.1.3.2 Moderate hearing loss

Person is unable to follow a conversation without any hearing aid in moderate level. They only hear loud or clear speech during conversation in other situation; they need to ask for repetition multiple times. Group activities, classroom discussions present a communicating challenge for them. Vowels and consonants both are difficult to understand by them. Usually the individual’s speech is impaired to some extent. [15].

2.1.3.3 Severe hearing loss

In case severe hearing loss, person is unable to differentiate any sounds without use of hearing aids. Most of the words during communication are unrecognizable to them. Cochlear implants or hearing aids are necessary to enable them to hear sounds [15].

2.1.3.4 Profound hearing loss

In case of profound hearing loss, person is unable to hear any sound or faces difficulty in hearing and comprehending sounds or speech even with the use of hearing aids. Person with this degree of hearing loss depends on lip reading for development of spoken language and/or sign language [15].

2.1.4 Management

Hearing aids and cochlear implants are the devices which help in managing the hearing loss. The choice between two relies on degree and type of hearing loss.

2.1.4.1 Hearing Aids

Hearing aids depicted in figure 1 are the electronic devices constitute of different devices interlinked in a package. It includes a microphone, amplifier, loudspeaker and battery. They are used to amplify sound so that user can listen to sounds/conversation easily and clearly. These are available in different ranges on the basis of size, power level etc. Modern hearing aids are digital which allow different operations to be performed as compared to traditional analog ones [15].

Figure 2.1 : Hearing Aids [15]

2.1.4.2 Cochlear Implant

Sometimes hearing aid does not fulfill the user needs or are not sufficient to help the user. In such cases cochlear implant depicted in figure 2 is useful and helpful. These are the devices used to evade the hair line damaged in cochlea and stimulate the auditory nerve directly. It consists of two components internal and external component. Internal component is under the skin behind the ear and incorporated surgically whereas a narrow wire is threaded into inner ear. With the help of an external magnetic disk internal component is connected to external component. External component is like a hearing aid behind the ear. They used to convert sound waves to electrical impulses and transmit them to inner ear, this property helps the people to hear sounds and understand speech effectively. Now days these are extensively used particularly for people suffer from severe and profound hearing loss [16].

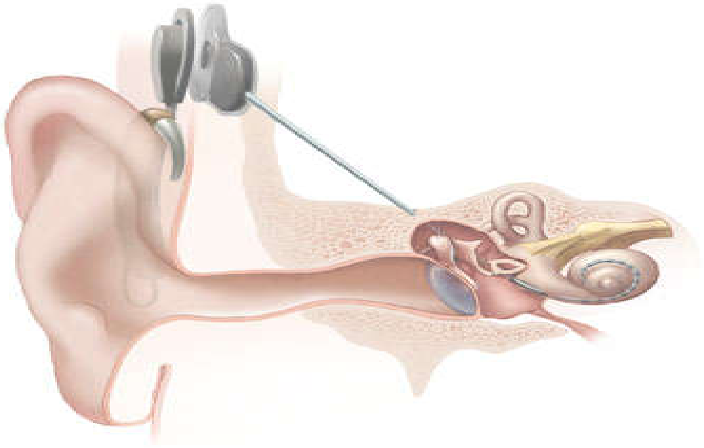


Figure 2.2 : Cochlear Implantation in Human Ear [16]

2.1.5 Hearing Loss and Its Impact on Child's Development

Communication development is a natural process. Human beings develop speech as they hear from outside world. A child starts hearing before birth and speaks at the age of two years. Children with hearing impairment cannot speak since they cannot listen. Therefore, they have speech and/or language delay to some extent. For communication development, language and speech development in an individual is necessary. All these are important for each other. Even as adults we are still learning language. Along with the use of hearing aids and cochlear implant, children also need speech therapy exercises and language learning activities. Speech-language pathologists (SLP)/therapists help to prevent communication disorders. Their task is to ensure that the child must focuses on his hearing ability by practicing different exercises.

Sense of hearing is crucial for communication, speech, and language development. Combining all these, result in developing better learning and articulating skills in an individual. Children's suffering from hearing impairment faces communication and learning difficulties which leads them to social seclusion, poor self-concept and continue to be an unidentified and undeserved population. Development of the child is adversely affected by earlier hearing loss. Earlier detection of hearing loss in an individual and commencement of intervention by the supporters results in less serious ultimate impact. The language inconsistency causes learning problems that result in reduced academic achievement. Academic achievements can be improved through involvement of parents in different activities arranged specifically for them. Quality, and quantity of the all the services necessary for learning language and speech also help in improving the level of academic achievement of such children [10].

Vocabulary development in hearing impaired children is slow. It is difficult for them to grasp their control over vocabulary in no time. Due to this, they are unable to increase their knowledge related to language as language is totally a combination of its vocabulary and grammar. It requires a lot of effort and time not only from the child himself but also from the SLP's, their teachers and parents. Vocabulary development when compared to their hearing peer depicts that gap gradually increases with increase in age. They find difficulty in learning abstract words like ***after, before, jealous, happy*** as compared to concrete words like ***cat, red, five, jump.*** This is due to the fact that concrete words or concepts can easily be visualized by them and they can easily store them in their mental lexicon for longer time. For them complex sentences are difficult. They cannot hear sounds of letters ***th, sh, f, t, or k***. They do not use longer sentences. It is really hard for them to hear words ending with -s or -ed. For example ***boys, girls, booked, cooked*** are harder to hear and understand. This depicts that plural words are harder for them to understand [69] [70].

2.2 Role of Information Technology

Therapist usually used flash cards containing different words, pictures, printed papers, real objects and toys during a typical therapy session carried out for speech and language disorder. Most of the time it is a tedious and difficult task for the therapist to craft a suitable environment keeping in view the needs of the patients for a particular therapy process and achieve the desired goals. Furthermore, formulating a comprehensive and standard report related to each therapy session requires a lot of effort from an individual and also takes a lot of time. So, for the therapist it is hard to prepare the demonstration of the therapy treatment along with its management by keeping its record to be used in future.

In the era of computing and technology, engineers and scientist with the use of new computing technologies help in assisting the people with hearing impairment by developing different software and techniques. Speech canvases [5], 3D games [6], 3-D talking head [8], Articulation Vacation, Carnival application [7] are some of the computerized systems/applications developed to support speech therapies. All of these are based to enhance the language or vocabulary knowledge of children and fulfills their educational needs to some extent. With the help of these technologies person with hearing impairment can participate effectively and can be understandable in the wider world. Some of the work mentioned below in context of IT role to overcome loss in development of a person caused by hearing loss disability.

2.2.1 Automated systems to support Hearing Impaired Children

With the recent technology advancements in field of information technology, computer sciences and artificial intelligence, many automated systems were developed to assist people with special needs. Artificial intelligence is extensively used for this purpose. Artificial intelligence's maturation is signified by ontologies and terminologies [17]. A study was conducted on artificial intelligence used in different domain in last 30 years, in this study it is analyzed that expert systems were developed to solve specific problems. From 1985 to 2013, many researches about artificial intelligence used in medical domain are published each year to support different type of individuals with their specific needs [17]. Specific to hearing impaired individuals, related work is mentioned below in context of automation which assists to overcome loss in development to some extent.

Language disordered children have met their specific requirements by the development of 3-D games environment. According to the specific needs of a user, a therapist could adapt a computerized base game environment which can be beneficial for the treatment procedure of a client.

3D game [6] environment remained helpful in the development of speech disordered children of Turkish language. For the development of 3D games a unity 3D engine was used. Objects that are added in the database, their place need to be fixed in the environment. 25 different objects are being used. Every object that is used by the player has its own functionality such as swing swinging, bouncing and ball rolling etc. These are all those activities provided by them. Therefore, without getting bored, a player can play with these objects. To meet therapy goals, this game has provided various situations in a park environment which helps the therapist in modification of objects and the character that is included in that environment.

Moreover, while playing the game the player can involve himself/herself in many other activities with the help of therapist; meanwhile they respond to the question asked by the therapist. In this way a therapist can build an environment according to the needs of the client such as, their specific languages. Hence traditional therapy sessions are supported by the development of 3D game environment.

Game interface is designed by using different scenarios. These scenarios keep the children busy in doing activities to make them introduced with the environment and help them to get familiar with the names and functions of the objects in that environment. Furthermore, actual therapy sessions are developed through which clients can interact with the environment independently.

Children with speech and language disorder got complete advantages of 3D game environment and environment also provide the opportunities to the therapist to make the treatment process accessible at any time anywhere. The therapist also identified that it was clear that children were more comfortable with their being present setting [6].

Study illustrates the demonstration of a real-time speech visualization application known as Speech Canvas [5]. Accurate pronunciation of less intervals speech sounds is done by prompts and text on screen. Examples include pronunciation of words /t/ and /p/.Speech canvas is tested on Arabic speakers. Therefore, it was concluded that all those Arabic speakers that used speech canvas shows 14.2% enhancement in pronunciation of these alphabets. [5]

Hearing impaired children are trained by the development of training system that is based on Virtual reality articulation. Nearly, twenty meaningful words in mandarin language were accessible to HI children during training process. A 3-D talking head is used for the purpose. Similarly, in order to illustrate the movements of Articulography, and various articulators, the graphic transform technology and Electromagnetic (EMA) is used. Additionally, in the listening and speaking training components of the system, speech corpuses were organized to improve and enhanced the language skills of hearing impaired children. Experiments outcomes presented enhancement in speech production by the Hearing Impaired [8].

Deafness in childhood has great effects on communication, quality of life and education. Language development and delayed speech are the most important outcomes of hearing loss. Computerized system trainings for articulation of speech play a key role in developing and enhancing their speech and language skills. Speech recognition algorithm is used to assimilate it and by providing animated or game like features in the GUI, the software can helped the hearing impaired children in effective training as well as enable them to enhance their communication skills [72].

Therapist has achieved Rehabilitation therapy, but computerized systems allow the self-reliant activities related to these abilities for children with language and speech disabilities in managing the expenses of such sessions. Software application or modules related to video game prototype specific for speech rehabilitation process usually depend on the auditory-verbal theory and educational objectives. It involves accurate evaluation and enhancement of voice production for hearing Impaired Children. [71]

Step by step stages are shown in the above figure. It represents basic production of sound at phonetic level and represents its ending with more multifarious combination of words and speech pattern. Moreover, fourth and fifth level depicts the consonant generation of words. These steps involved sequence of activities, related to combination of vowels and types of articulation. These activities are then recommended to the patients and evaluated them regularly. These exercises are considered as mini games [71].

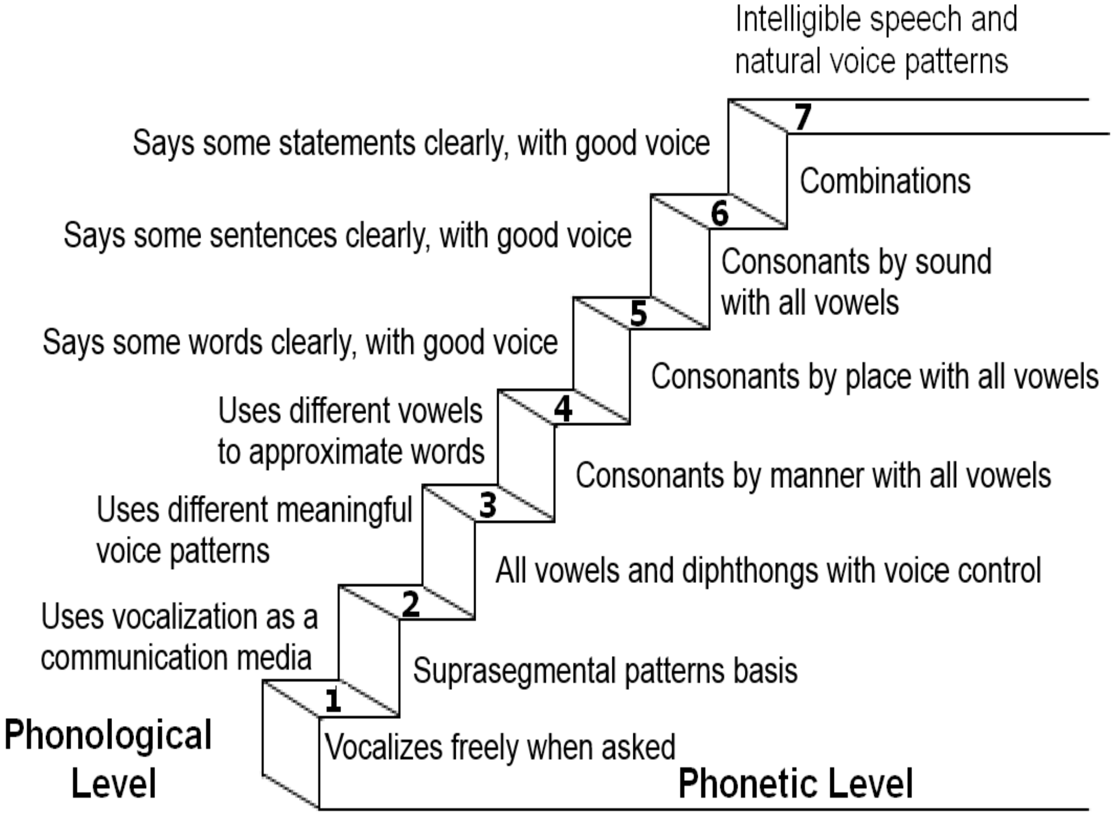

Intelligible speech and natural voice patterns
7
Says some statements clearly, with good voice
Combinations
6
Says some sentences clearly, with good voice
Consonants by sound with all vowels
5
Says some words clearly, with good voice
Consonants by place with all vowels
Uses different vowels to approximate words
4
Consonants by manner with all vowels
Uses different meaningful voice patterns
3
All vowels and diphthongs with voice control
Uses vocalization as a communication media
2
Suprasegmental patterns basis
1
Phonological Level
Vocalizes freely when asked
Phonetic Level

Figure 2.3 : Order of Acquisition of Sounds [71]

Stuttering problems and speech disordered children are diagnosed by the therapist with the help of Malay Speech Therapy Assistance Tools (MSTAT).Speech recognition system uses the Hidden Markov Model (HMM) technique in order to assess speech problem.

To stimulate the therapist questions and answers to the subject intelligent dialogue system is used. In this system the patient will speak the sample words to get to know whether they could speak fluently or not [73].

The training tool was developed for speech therapists in order to help a child in articulation of preliminary sounds like vowels. They are provided training related to these sounds in a speech therapy program by engaging them in a game. With the use of DSP-Based techniques, the speaker's utterance is compared with a template utterance. Accurate pronunciation was identified through identification of possible area for each level. By enhancing the possible area for each vowel in formants plot, accomplishment rates attained for the normal speakers is of 90-98% [74].

Articulation vacation application is developed with which children practiced the articulation of several phonemes during vacation time and play built in summer games. The Articulation Carnival application is a widespread and flexible articulation program to learn pronunciation of ALL English phonemes at the word, phrase, and sentence level [7].

All these above mentioned systems helped hearing impaired children in developing listening and articulating speech to accomplish their routine activities in daily life. With the help of training sessions for speech and pronunciation through activities not only enhance their knowledge related skills but also motivate them to participate in these activities actively.

2.3 Spoken Language Acquisition and Vocabulary Knowledge for Hearing Impaired children

Field of linguistics is not a content based field like science, social studies, economics etc. It is a skill which cannot be adopted by humans rapidly rather it is an ongoing learning process. By the passage of time human's polish their skills and enhance their language by learning it. It is a process of expressing feeling, and thoughts during communication in all over the world [9].

Humans communicate with each other mainly through spoken language. For efficient and effective communication there is an usher need to develop four skills related to language [9].These skills include listening, speaking, reading and writing shown in figure [3].

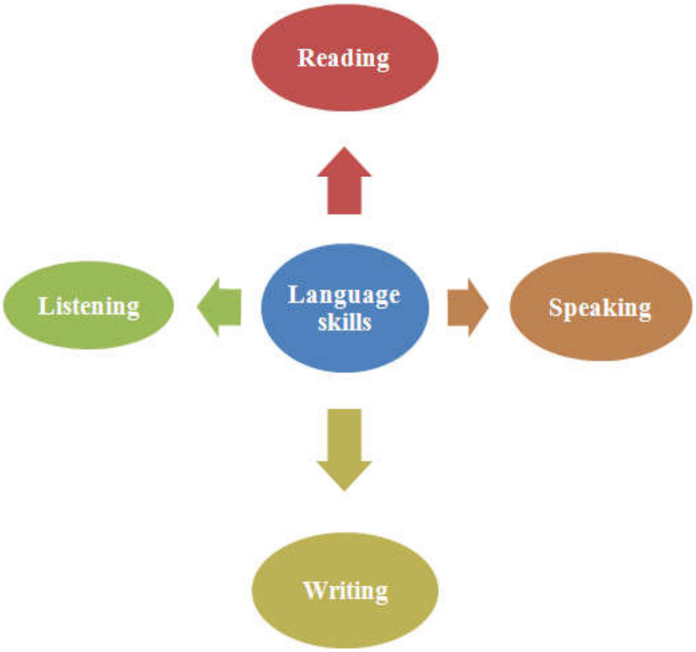


Figure 2.4 : Language Skills

All of them are interlinked to each other in such a way that if an individual is unable to develop any of these skills, then they face difficulties during communication and developing their language skills. These skills help individual to understand speech of others, successfully engage other in conversations, able to read newspapers, books etc. and write emails, letters etc to others. In case of hearing loss due to less perception of sound from the outer world, these skills are not developed up to the mark when compared to the skills of their hearing peers. They face speech and/or language impairments [18]. Due to this their academic achievements are also affected.

Along with the development of these four skills related to language, another important aspect in language development is vocabulary. Vocabulary is defined as increase in number of words known by the person in the mental lexicon and then again utilizing knowledge about these known words to form more new words. This results in enhancing knowledge about specific language. Enhanced vocabulary knowledge directly enhances learning skills of an individual related to particular language. Rich vocabulary enhances the reading, speaking, listening and writing skills of an individual which ultimately improves their communication skills. Thus, It is consider as one of the important tool not only learning language but also for successful communication to be carried out between individuals [19].

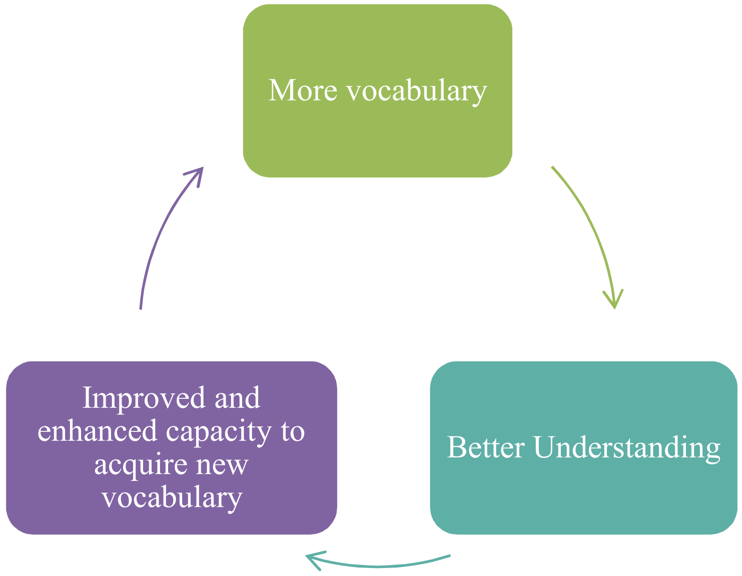


Figure 2.5 : Vocabulary Development Cycle

Figure [4] depicts that vocabulary development is a continuous process.This reflects that an individual who knows more words are capable enough to learn more new words.Learning vocabulary is an ongoing process with the increase in age of an individual and enhancement of knowledge.

Without developing sufficient English vocabulary, person is unable to comprehend conversations and clearly present their own ideas. Vocabulary is all about knowledge related to words. During teaching vocabulary to children, teachers usually focus on the form, meaning and use of word in particular scenario. Form of word includes its pronunciation, spelling, and addition of suffixes and prefixes to generate new words. [Vocabulary and Its Importance in Language Learning]

For example by learning word “impossible “ it means one should know about its form, meaning and use. Teacher will explain child about following important features related to word to the child.

- Its pronunciation by speaking the word to them
- Spelling will be learnt through writing it several times on board.
- Tell the child about addition of any other part to the root word or brief them about the knowledge of prefixes and suffixes.
- Tell them about the meaning of the word
- How a child can use it in their daily routine.

2.3.1 Research Related to Acquiring Language in Hearing Impaired Children

Studied literature suggests that hearing impaired students are supposed to be delayed in the method of acquisition of knowledge.

2.3.1.1 Impact of previous Research with hard of hearing / Hearing Impaired students

Review paper [20] present the vocabulary related specific research on students who faces hard of hearing problem. It [20] includes all the published material from year 1967 to in a peer reviewed journal. Literature review results depicts that majority about 76% studies did not show intervention indicating most of the research as descriptive.. To examine the effect of activities and methods only 10% studies were conducted. The starting points for children are words. Because with help of words children can talk about people thing or place. The educators are recommended to provide child with continual exposure towards new words and develop strategies for the students to make them independent vocabulary learners. Thus, by learning new words one can develop vocabulary.

2.3.1.2 Spoken Language Acquisition/Development in Hearing Impaired Children

Multifarious factors have been observed associated with performance of deaf children. Some children showed strangely high performance while, other indicated comparatively low. Various methods of communication like, dependence on spoken language, highly educated mothers, personalized instructions, ear level aids and verbal instructions remained very supportive in producing great performance by hearing impaired children. Similarly, parents were highly committed for the development of spoken language by concentrating on family resources. Early intervention of such children in education gave fruitful results of spoken language as compared to those who are treated later. Moreover, direct instructions given by the parent are very essential for successful involvement of children at early stages; many studies have shown that hearing loss children having an integrated environment had better spoken language then that of living in isolation. Hearing impaired children who received Auditory/Oral (A/O) mode of communication when integrated had much good spoken language then those of having separated lives. Therefore, A/O programs have great advantages in the development of spoken language for hearing disable children. Different variables are used to select the sample of varied age groups of deaf children in order to develop spoken language for them, like **Hearing Thresh Hold Level, Spoken Language** and **Intelligence Age.** Thresh hold differed at various frequency levels, this was the specific measure of hearing and spoken language. Labeling of pictures of representing different objects was done by deaf children to access spoken language. Furthermore, regression analysis was also done on the basis of their hearing thresh hold levels, age and intelligence. Parent's attitude towards deafness is very important as they are concerned with their children's educational background and communication.

Demographic characters like age, intelligence, parental education, gender and family size are very significant factors for the High Speech **HS** and Low Speech **LS** children. Thus, parents are responsible for the early educational intervention of their deaf children in their learning. The Auditory/Verbal methods deliver individualized instructions that help the children to use speech to communicate with their parents. Hence A/V program remained very successful in the development of spoken language. Greater level of performance in spoken language is connected with four elements; A/O communication, classroom placement, directs instructions given by parents and individualized instructions [21]

2.3.1.3 Vocabulary Comprehension for Deaf or Hard Of Hearing

Vocabulary, an integral part of language used as communication tool all around the globe. If an individual develops and learn vocabulary then he/she is able to communicate in a specific language easily and efficiently. Vocabulary development has become a challenge for the hearing loss students because of their disability. Rapid vocabulary development for the deaf students has done through effective methods like visual representation, vocabulary training, re-teach, meanings of the basic vocabulary and reinforcement. Moreover, students can recall their previous knowledge by playing games and by reviewing the vocabulary words around them. Hearing impaired students need to do much effort to comprehend vocabulary words, therefore these students have been trained by continuous use of sign language as a culture of deaf communication language and by visual representation. Similarly, social and verbal interaction of children with their elders plays a crucial role in building of different and meaningful words but, the children living in the backward areas extremely lack this ability to comprehend difficult words. Vocabulary can be retained in the mental lexicon of individuals using three different stages. In first stage is retaining vocabulary through slow learning. In this stage, person needs constant repetition of a word several times. At second stage, process of direct rapid learning is used. In this, vocabulary can be retained rapidly after limited exposure. It takes less time as compared to slow learning. Third, but not last through indirect learning process, in which no repetition, prompting or modeling is required and vocabulary is obtained through environment using visual aids or other means. Children, who are not able to make proper sound or cannot make words, use sign language as a learning tool to understand and get words for communication with others. However, learning of sign language at early stages has become very crucial for the deaf students to get better understanding of language. Additionally, another method of using sign language in different areas like at grocery stores, shoes shop or in parks helps to enhance vocabulary in hearing loss children. The use of assistive learning activities, physical strategies, combinations of visual representation and mode of reading with sign language are the various components to develop vocabulary and boost reading skills in disable children. Furthermore, Early communication with these children progressively enhance and strengthened their thinking and social skills. In the schools, class room setting is a significant element for the successful achievement of goals by beautifying the walls with different pictures or by introducing images at story telling time. Hence sign language and pictorial representation is very beneficial for hearing impaired students; that will help them in development and comprehension of words with remarkable literacy skills [22].

2.3.1.4 Depth of reading vocabulary

Paper [23] presents a comparative study related to the vocabulary knowledge of children with and without hearing loss. Two vocabulary assessment tasks are conducted to determine the deep knowledge of the words. One task is named as lexical decision task and second is known as use decision task. First assessment is use to know that whether child encounter the word before doing assessment. In first task letters of words are arranged randomly and word is presented to the child. Now, child has to arrange those letters to make a new or complete the incomplete word by memorizing or recalling the vocabulary knowledge. Purpose of second assessment is to check and analyze whether children possess the knowledge about use of any particular word or not. In this, words that are completed by the child during first assessment task are included. Now not words but sentences of those words are presented to the child and child has to tell that in which sentence word is used in its correct form or meaning. These two task scores are analyzed and determine the depth and size of the word for hearing impaired children. Authors in this study imagine that children with hearing impairment possess knowledge about fewer words. Because, they lack behind their hearing peers related to communication input with respect to spoken language that is the reason they encounter fewer words as compare to the hearing children. Second purpose is to determine either the two tasks are similar for both hearing impaired and non hearing impaired. The result showed that hearing impaired score poorly on use decision task and children with hearing impaired children were weak in recognition of words in lexical decision task and incorrect recognition of words in use decision task as compared to hearing children. They can deepen their word knowledge by repeating the word may times in different forms like teachers present or explain the word through flash cards, visual or sign representation [23].

2.3.1.5 Proficiency in English Language Learner

It is essential for an education system to provide such a platform to learner to meet all their needs related to learning. Due to globalization as the language learner increases, the demand for availability of language teacher is also increased. Keeping in view the complexity of language, it is really difficult to find errors. For this, learning proficiency can be measured to find errors using different error models. Paper [24] proposed a measure based on entropy of occurrence of words in sentence. To analyze the stronger correlation of learner proficiency proposed technique is compared with the vocabulary size and frequency of word occurrence with language learner proficiency. The data set is divided into two sets such as learner original data set and learner error data set. In original data set the error data is included of the same learner to avoid duplication of data. Result showed the proposed entropy measure has powerful correlation with learner proficiency when compare to both vocabulary size and entropy. In a majority of subsets related to part of speech of a particular word, the proposed technique has a strong association in context of proficiency.

2.3.1.6 Teaching Vocabulary

All languages are considered as building block of words. Generation of new words neither stops nor does the acquisition of words. Even, in case of our mother language, we are in continuous process of learning new words, and meanings of the already known words. Researcher's most of the investigation includes how the mental lexicon is organized in an individual by doing comparative analysis of the speed at which an individual is able to recall items. If some of the prompts are answered in short span of time than other, then this will reflect the lexical system. Most of the researcher's believed that in human memory all items are arranged in a sequence of their associations and construct a network type of structure. Every item is in one file known as master file and there are various peripheral access file known as detail files which consist of information related to spellings, meanings, syntax and phonology of that particular item. Items in master file are referenced in detail file for relationships. In this way mental lexicon is an overlapping system. Searching a word is like following a constructed network path in processing.

The successful literacy intervention for hearing impaired children was the introduction of reading recovery. For the teaching purpose the teachers of hearing impaired children use different supportive techniques and strategic activities for construction of message. This [25] study examine quantitatively reading lesson of hearing impaired children and compare with non hearing impaired children. In this study mix methodology of quantitative and qualitative method by applying interpretive research is used. The research compare the analysis of 12 hearing impaired and 12 non hearing impaired children through interaction of their reading recovery teachers. The research involved the examination of the children literacy achievement and teacher role in support treated as coded variables. Codes develop to provide facility to teacher that represents events during reading and interaction of teachers. The result showed that focus of teacher enhances the reading and writing ability of hearing impaired children. The limitation of this study was that code used was not based on previous research. Further more research with larger group of students is recommended as a future work.

2.3.1.7 Learning word

In study [26] learning ability of hearing impaired children was examined with help of direct reference and novel mapping. For enhancing children learning abilities the researcher presented various mechanism based on linguistic and cognitive approaches. These children are able to learn new words and the learning of words depends upon vocabulary knowledge. Total of 98 children were participated from twelve different schools. Mode of education in these schools is both sign and oral environment. Deaf children and children with cochlear implant were taking part. Participated children age is ranging from 27 to 82 months. The background of children varies from each other. Language delayed is shown in most of the children. Participated children are divided into two groups such as children with hearing loss, whose teachers believe they understand objects name and those whose parents consent is there. The parental consent rate was high in the study. The language researcher assessed the children individually. The assessment includes direct reference, language measures and novel mapping. After word learning task teacher questionnaires were collected. Result shows that deaf children will acquire vocabulary knowledge through interaction with nature. More the interaction more will be the vocabulary knowledge.

2.3.1.8 Word-formation

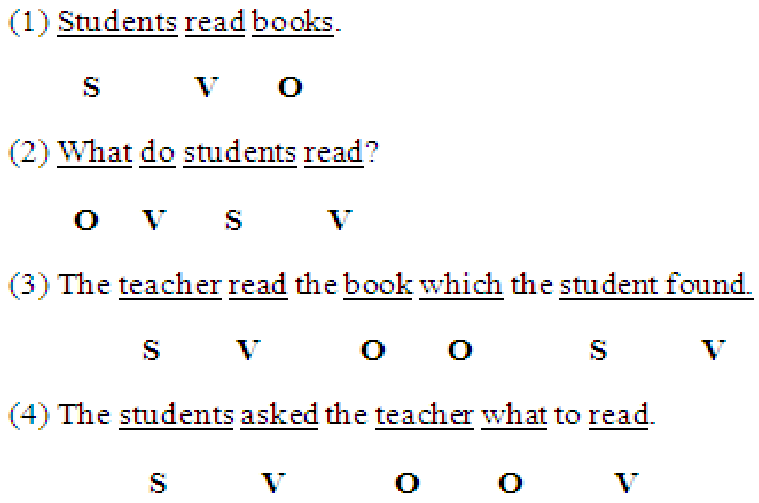
Human language is in a process of continuous evolution. Generating new sounds, words, and sentences is possible through combination of different building bricks of language. The word-formation has not received more attention. The words in English may be simple, compound and complex. The complex word may be free forms and may be bound forms combining prefixes and suffixes. The suffix added to the verbal stem such as -er make the noun eater. Prefixes play the role of no class changing such as do and undo both are verbs. The suffixes in English are mostly class-changing and forms words which act differently from basis. Word formation acts like a bridge between different classes that is a link between lexicon and syntax. The phrasal verbs in English are combinations of verb and adverb. The combination of verb and preposition is some time called semi compounds. The different classes are no means to rigidly compartment but shades are combining to forms varying extents. It is necessary to have a look at the nature of words in order to get clearer idea about the word formation and also understand the features that differentiate the word formation from one to another [27].

2.3.1.9 Processing Written word

Reading is considered as one of the difficult activity and recognizing a word is considered as constituent part of reading process. Language is presented in the form of printed text through writing. Word recognition acts as a bridge between reading and writing a language. An individual who is unable to read a word is also not able to write. This depicts that a poor is unable to process a word. Paper [28] shows comparison between normal and hearing impaired children abilities of writing and reading process. The written alphabetic word has two component systems. One is called direct system and other is called indirect system. In direct system written word is matched with the lexical and with phonological lexicon. On the other hand in indirect system phonological is build. The hearing impaired children reading methods are not consider as valid argument in word recognition. To produce speech and to participate in cognitive activities hearing impaired children must use the phonological information. They use visual clues, lip reading etc to recognize word. It has been observed that certain children access phonological representation methods [28]

2.3.1.10 Grammar Development

Tremendous difficulty is experienced by hearing impaired children in learning spoken languages as compare to children without hearing loss. The hearing loss children had no access or minimal access to the sounds of spoken words due to which the learning phase of hearing loss children is quiet slow. Some hearing impaired children in some way gain the native-like knowledge for spoken language. Many hearing impaired children experience difficulties in reading and writing in expression. In accordance to the deaf children knowledge about English grammar, paper [12] present the research related to the specific sentence on which hearing impaired children face difficulties while reading. For expressing the grammatical relation English language has word order known as (SVO) Subject Verb Order. For simple sentences this is followed order but for complex one, order might get changed In case of hearing impaired children as English learner if their exhibit a non SVO order for sentence then it is a problematic for them to read.



(1) Students read books.

S V O

(2) What do students read?

O V S V

(3) The teacher read the book which the student found.

S V O O S V

(4) The students asked the teacher what to read.

S V O O V

In reading and writing comprehension when SVO order disturb in complex sentence for example sentences no 2,3, and 4, the hearing impaired children face challenges. In this scenario teacher of hearing loss children have unique responsibility to contribute in students learning process. Teachers of such students should participate in professional learning activities to help learning needs of children.

2.3.1.11 Complex Syntactic Construction in Children with Hearing Impairment

Auditory function in hearing impaired children is poor and depends on their ears status and efficacy of hearing aids. Hearing loss can cause different effects on linguistic perception of individuals. It is observe that hearing impaired children undergo from speech and language impairments. Study [29] focuses the syntactic construction in children affected from hearing loss. Syntactic construction abilities of deaf children are poor as compare to the normal hearing children. These children face lot of problems while understanding the complex syntactic structure because they incorrectly used the passive and auxiliary sentences. Lack of exposure to natural language cause the difficulty of understanding syntactic structure in hearing impaired children. This study takes twenty hearing impaired and twenty normal hearing children. . Syntactic construction task are classified into two categories that is sentence with canonical and non-canonical order of words. The tester repeated the sentence many times and there was no time limit for the test Result obtained from the result showed that the hearing impaired children get high score in canonical sentence and gain low score in non-canonical sentences. During the period of language acquisition it is necessary to input rich linguistic to the child.

2.3.1.12 Why Words are difficult for Adults with Developmental Language Impairments (LI)

Word learning problem is diagnosed to the wide extent. Children with Language Impairment LI face more complications then unaffected peers. Sometimes it becomes difficult for language impaired individuals to learn complex words and their meanings. Therefore, such individuals are trained by presenting them new words with meanings. The training studies have provided evidences that word learning is improved by the acquaintances of more words to disable children. Similarly, Language Impaired participants were trained, when they were exposed to new words either with 3 or 10 times each; these observations showed that LI children indicated better understanding of words. Encoding or remembering of words is examined by the 12 sleeping hours of affected and unaffected children. Furthermore, many experiments have been carried out that showed that Language Impaired children got low score due to attention deficit or hyperactivity disorder. Before training sessions groups of participants were made that consist of number of Language Impaired individuals LI and another group of ND. During training visit every participant was interviewed about the sleeping functions, trained on the new words and then tested instantly by free recalls. In the training period participants were presented with training script consisted of number of blocks that they viewed and listened, every block displayed novel referents on the screen for some seconds. In this way the LI participants were actively engaged in learning new words. The

2AFC recognition task was also taken place in which individuals heard the words and recognized whether it is named according to the pictures or not. Free recall is one of the ways of learning through which hearing loss children were taught by recalling orally trained words. Moreover, learning of meanings of words is done through word association method; this method also used trained words in the form of audio recording. Therefore, the above mentioned procedures remained very productive in learning new words and their meanings among hearing impaired children. Hearing loss children are not able to learn words because their working memory is shortage of phonological loops that help in memorizing things for long term. Hence, the 2AFC test minimized the demand of phonological retrieval. Repeated exposure of words assist adults with developmental Language Impairment and consolidation of declarative memory is a great strength for adults with LI. The encoding deficit in hearing disable children does not allow the addition of more words in their long term memory; that can be overcome through proper training channels [11]

2.3.1.13 Complex Sentences and their Punctuation in English Texts Composed by Latvian Students

Complex sentences in Latvian and English language differ significantly due to the use of comma. For Latvian users of English this causes major problems. Students mostly believe in the punctuation rules as discussed in the Latvian classes. The research is not yet taken upon the overuse of comma in English sentence written by Latvian students. Study [30] presents research on Latvian student's punctuation of English complex sentences. Strong linguistics relation has been recognized between the clauses in English Language rather than simply separating dependent from the independent clause as in Latvian language. For analysis the material is extracted from the Latvian composed text corpus written by students. It contains 21,319 words approximately. The result depicts that Latvian student's uses unnecessary separation between nominal and relative clauses through comma from independent clause. This highlights the importance of identification of linguistics relationship between two clauses in complex sentences during punctuation process. Emphasis should be laid that text is different for Latvian in punctuation process. In future various types of sentences and related punctuation in text composed by students for different purpose of communication are still required.

2.3.1.14 Factors related to difficult of words

Lot of studies has been carried out in the domain of vocabulary. This deals with handling specific learners, reducing large amount of vocabulary and adopted teaching methods for improving vocabulary skills. Very little research has been conducted on the topic of learning new words, or why words are difficult to learn for language learners. Based on the studies related to vocabulary acquisition, study highlights some of the factors related to difficulty of words. Linguistic analysis depict that word is constitute of some different set of properties, or what is known as features. It includes pitch, stress, sound and morphological units. Pronunciation of unknown words makes it a difficult word for the learners. Similarly, length of the word is also considered as one of the factors. Shorter words are easy to understand. Part of speech of the word also plays a vital role. Noun are most easy to learn and adverb the most difficult ones to understand. Morphological and semantic features are also considered to enhance complexity of the word [31].

2.4 Natural Language Processing and its Role

Natural language processing (NLP) is broad domain of computer science. It is used to process the unstructured data. Structured data is somehow easy to process but unstructured data is difficult to process. Many tools and techniques have been developed for NLP. Different programming languages also support the NLP, built in libraries or name spaces are used to support natural language processing. In structured data storage, data is stored in the form of tables or relations in database. Relations have different attribute, primary key and other features as well. To process this structured data simple SQL query can be used and many techniques are also proposed such as SVM classifier, decision tree, naïve Bayesian ……. Data or information stored in MS WORD or PDF etc. contain unstructured format. This unstructured data is written in free-text format. To process unstructured data NLP is used. Part of speech tagging is part of NLP in which words from the sentence are placed in the different group constituted by part of speech [32].

Natural language processing is used in implementation of different systems which are discussed below:

2.4.1 Natural Language Specification for generation of UML diagrams:

UML diagrams are generated in design phase of software development. UML diagrams are created from requirement of user. Requirements are written in natural language in free-text format which are gathered from system's user by using different techniques such as questionnaire, interview, story boarding and prototyping as well. It's challenging to create the UML diagrams from text written in natural language, because it is difficult to process unstructured text and requirements written in free text format may be incomplete, ambiguous and inconsistent and one requirement can conflict with another requirement as well. Moreover, understanding of requirement is affected by social, psychological and geographical factor. In software industry it is job of requirement analyst to find and fix ambiguities from requirements. A method and tool are proposed to create UML diagrams from requirements written in free-text format. Proposed method involves the natural language processing to process the requirements written in free-text format. The proposed tool is referred as **Requirement analysis to Provide Instant Diagrams (RAPID)**. RAPID tool is desktop tool that helps software designer to create the UML diagrams from text [33].

There are different linguistic analysis levels used by natural language processing such as.

- Lexical level
- Discourse level
- Syntactic level
- Semantic Level
- Phonetic level
- Morphological level
- Pragmatic level [34][35]

Following steps are involved to convert natural language requirements into UML Diagrams.

- **Normalizing requirements component**:

  The aim of this component or step of RAPID is identification of incomplete requirements and removal of ambiguous requirements.

a. **Syntactic Reconstruction:** In this step, requirements are taken from stakeholder as a user as input. On taken input syntactic reconstruction is applied to divide the complex sentence into smaller and simple sentence for extracting possible information from requirements document. Every Requirement is scanned to test whether sentence structure of requirements is satisfied. If sentence structure of requirement is not satisfied it gives error message and prompt the user to change the sentence to satisfy the rule [33][36].

- **NLP Technologies Used:**

For successful implementation of RAPID tool following Natural Language Processing technologies are used.

a. Open NLP Parser: Requirements which are written in natural language contains different parts of speech. Different parts of speech combine to form to complete sentence. In RAPID implementation, open NLP parser used for POS tagging.

b. **Rapid Stemming Algorithm:** In this step, the words in requirements are converted into their base word by removing its suffixes and affixes [33][37].

c. Word Net: In natural language there can be many words that have same meaning. There can be multiple synonym or similar terms for a word. Word net is used to display synonym or related terms for word**.**

d. Concept Extraction Engine: In this step or module, concept related to requirement document is extracted. Race stemming algorithm, OpenNLP parser and word used for concept extraction engine.

e. Domain ontology: For identification and clarification of concept domain ontology is used.

f. Class Extraction Engine: Output from notion extraction engine module is used and different heuristic rules are applied for class diagram extraction. Domain ontology is also used for refinement of class diagram. Rules are used for identification of classes, identification of attributes of classes and relationship between classes [33].

2.4.2 Natural Language Specification for automated use case generation:

Numbers of activities are involved in software development life cycle (SDLC). The first phase or first step of SDLC is requirement engineering. Successful development of software is strongly dependent on this phase. To represent requirements into diagrammatic notation, use case modeling is used. Use cases are created at requirement engineering phase of SDLC. Requirements are written in natural language. To generate use case model from text written in natural language is challenging and time taking task, detailed understanding of system is required to create use case diagram. An automated system is proposed to generate use case model from requirements in natural language. Proposed system contains the natural language processing (NLP) as one of the major features [38]. Automating system by extracting information from free-text record using natural language processing is an innovation and new domain of computer science. Natural language processing contains both computational linguistic and artificial intelligence. Both these fields facilitate the interaction between human language and computer language [39]. Before creating the use, case diagram there is need to process the text written in natural language. Natural language processing consists of different task such as part of speech tagging, stemming of words and removing stop words. In natural language processing, the text or information written in natural language is extracted and converted into form that is understandable by machine [40]. The importance of natural language processing is increasing in applications that require the interaction between human and computer.

Following steps are involved in creation of use case diagram in proposed system.

- **Acquisition of Text:** This step allows the natural language text to be written in provided text format of system. Text written in other files such as pdf, MS WORD or text file can be imported in system.
- **Segmentation of Text**: In this step, morphological analysis is applied on text written in natural language. Morphological analysis is applied to separate different sentences from documents. The full stop (.) is used to identify end of sentence and separate one sentence from another. If a statement is like " Dr. Mohsin …………….. ", in this case full stop after Dr is not considered as terminating point of sentence.

- **Tokenization of Text:** In the previous steps, different sentences of a document are separated by using morphological analysis. In this step, different words are separated from a sentence. The text is split into tokens.
- **Parts of Speech Tagging:** After separation of different words from sentence, there is need to identify parts of speech. Parts of speech tagging is applied to identify the parts of speech for each word. Parts of speech tagging categorize each word as:

    - Noun
    - Pronoun
    - Proper Noun
    - Verb
    - Helping verb
    - Modal Verb
    - Adverb
    - Verb phrase
    - Adjective
    - Preposition
    - Prepositional Phrase
    - Conjunction

Word net was used for categorization of words.

Knowledge Extraction:

Rules used for identification of actors, use cases and relationship between them.

- **Actors:** Rules used for identification of actors are:

    - Actor type is indicated by common noun.
    - Actor is indicated by proper noun.
    - Proper noun is ignored

- **Use cases:** Rules used for identification of use cases are:

  - Use case type is indicated by main verb.
  - Use case type is indicated by transitive verb.
  - Use case type is indicated by noun following the verb (verb + noun).

- **Relationships:**

  - Sentence boundaries are used to identify relationship between actors and use cases.

Experiments are applied on 50 different documents containing the text written in natural language. The number of words in each document is in the range of 50-300. Proposed methodology achieves the recall and precision of 96% and 84% respectively [38].

Dataset for Natural Language Requirements Processing:

Requirements of users/customers/stakeholders are written in natural language that is easy to understand by human. NLP domain is progressing rapidly. Different related techniques are used to perform different tasks of requirement engineering such as:

- Analysis of requirements.
- Traceability of requirements.
- Classification of requirements as functional or non-functional.
- Ambiguity detection.

But most of the natural language processing techniques are statistical and are using methods of machine learning. Large natural language requirement dataset is needed to facilitate the testing, validation and training of technique. Efforts have been made in past but dataset of natural language requirement was limited.

There were following two issues:

- It is hard to reproduce experiments
- It is hard to generalized results

To overcome these issues, requirements are extracted from web and presented in a XML format along with formatting, annotation and extension of dataset.

Following are challenges for dataset for natural language requirement processing.

- **Extraction of text:** As the requirements are written in free-text format, so the first step for dataset is to extract information from document. The text is extracted in uniform format of XML. The text is extracted from .pdf, .doc, .txt and other formats. Extraction is not fully automated some manual tasks are also involved. Both the manual processing of text and automated processing are used for dataset of high quality.
- **Annotation of dataset:** Organizing documents requirements into uniform format of xml is not sufficient [41]. Testing and validation of algorithms by using supervised or unsupervised machine learning algorithm, manual annotation is required for each task of requirement engineering [42]. For ambiguity detection manual tasks are performed. Domain knowledge is required for annotation of requirements. One person is not sufficient for annotation of requirements, multiple researchers of specific domain are required.
- **Updation and extension of dataset:** By analysis of dataset it is observed that the dataset is partially balanced. Additional documents from web are added and literature related to requirement engineering is also added as well.
- **Definition of API:** API (Application Program Interface) is needed for researchers to access dataset and XML files (meta-data and text) easily. Java Architecture for XML Binding (JAXB) is used to create XSD file automatically. This Work is limited to APIs of high-level, for example computing statistics from the dataset that work on top of Java Architecture for XML Binding.

Data and algorithm are most important ingredients of natural language processing. For successful implementation of tool excessive amount of data is required. Contribution of whole requirement engineering community is also required for annotating the requirements and defining reusable annotation schemes [41].

2.4.3 A Tool for processing of Natural Language in Clinical Text:

Electronic health care records are used for storage and collection of patient's data. Electronic record can be shared easily. Electronic records of patients have great importance in decision support system [37]. Extraction and analysis of information which is stored in free-text record in a large quantity is challenge for achieving decision support system. Machine learning techniques are used in processing of natural language. End user is unaware of machine learning techniques and is just interested in analysis of clinical text. Natural language processing is used for extraction of information from clinical record. A web-based tool is proposed to facilitate the clinical researchers in error identification in model prediction and reviewing output of natural language processing. Proposed tool enables the end users to review the clinical record written in natural language. Multiple visualization is provided so user can easily understand the result, perform any correction in case of error and provide feedback for improving natural language processing models. Case study is conducted to validate the tool. Researchers and clinicians are involved in testing of tool [44].

Following are the main parts of proposed tools:

- **Sensemaking and visualization:** Visual tools are used for displaying summary of text data that is present in large amount. Visual tools used are WordTree [44] and Tiara [45]. Evolution of content for each of topic over time is focused by Tiara, key words are provided by WordTree for exploring the text. Visualization purpose is to provide dataset level overviews and document level views as well.
- **Interactive machine learning:** Efforts has been made to develop a tool which assist the end users in building models by using machine learning and natural language processing techniques.
- **Requirements Designing:** It is assumed that end users are familiar with context of documents that are being reviewed but they are not familiar with machine learning. By keeping in mind this assumption following requirements are outlined.
    - R1: The tool must be able to provide interactive review of clinical free text record and make it possible for end users to work with natural language processing models who are not expert of machine learning.

- R2: Visual presentation must be able to facilitate the creation of accurate natural language processing models.
- R3: The interactive components must be able to support the complete interactive machine learning loop such as, feedback, review and retrain cycle that can be used to build or revise natural language processing models iteratively.

Interface Design: Boolean values (true, false) are used against the values extracted from natural language processing system. A subset of fourteen variables of each of patient is used for demonstration of proposed methodology [43]. Components of user interface of tool are following:

- Review
    - Grid View: 14 variables are shown in columns and individual documents in rows of table.
    - View for displaying statistics and keywords.
    - Document View: To show the report of patient in a full text format. For example, patient's report of endoscopy or pathology.
    - WordTree View: To provide visualization of complete document of health record.
- Feedback: To facilitate user to provide feedback for improving the accuracy of machine learning models.
- Retrain: Last step of cycle of machine learning. It keeps the track of number of times feedback is sent by user.

Implementation and deployment of system: Client-server architecture is implemented for proposed tool.

Results revealed that the interface is user friendly; the end users can easily interpret the output and can use the system having little or no knowledge of natural language processing and machine learning techniques. Results show that it is impossible for practitioners and clinicians to completely access the benefits of natural language processing. Without using the usable tools there is difficulty in applying the techniques to review and revise the findings of natural language processing [43].

2.4.4 Software Engineering in natural language processing:

Although, software engineering is comprised of standard processes, tools, methods and techniques that can be used in development of natural language processing software. Development of software in context of natural language processing can be done in following steps.

- **Open Source Development:** In natural language processing context the developers are freed from legal fringes which occurs as a result of proprietary licensed software. Most of the times software's are not developed for commercial instead they are developed for purpose of research [48].
- **Closed Source Development:** Under the closed source model of development after payment of developed software it is delivered to customers [46].
- **Attributes of Software Quality:** In natural language processing systems, quality of software is one of the important concerns. Following quality attributes are important for natural language processing software [47].
    - Reliability
    - Perception
    - Feature
    - Aesthetics
    - Performance

Justifying the use of text: Following are the benefits of textual documents.

- Document or artifact that can be directly automated.
- Human can easily understand the textual information.
- It is easy to make the textual information.

If we have textual information tools and techniques to support natural language processing can be automated. Text written in natural language can be converted into any other language by using different techniques such as machine learning algorithms. Human can easily understand and interpret the text written in natural language. A holistic view can be generated by conducting research on inter-disciplinary or multidisciplinary areas. By combining two area such as software engineering and natural language processing, benefits of textual information can be achieved. Universal programmability can also be achieved by conducting research on inter-disciplinary areas. Software engineering and natural language processing are divergent fields, but they can be combined to produce better software [46].

2.5 Systems Developed for Identification of Difficult Words

Automatic text simplification is a process which is used to convert complex or difficult text into its simple form. Due to this not only the quality of the text improves but text is also easy to understand. There exist huge amounts of data available to number of people around the globe. People access the data through books, newspapers, or online websites. But what if the text is difficult to understand for them .Due to which people spend amount of time on understanding it which makes them tired and they take it as a tedious task to do rather enjoying getting information. It is especially useful for person facing some difficulty or for foreigners to learn a non-native language. It also helps native language learners to learn the language. Text simplification process constitute of different steps. It includes identification of difficult text, finding simple substitution of the text, ranking the substitutions. Lexical complexity deals with the difficulty of the word and vocabulary of the language. Different systems are developed during recent years to assist people in learning languages. Approaches used both lexicon based and machine learning approaches.

Mining useful information from text a large number of techniques have been used in order to extract useful pattern. In text mining the frequent pattern approach showed better result but to some extend it does not work properly. For frequent pattern mining this paper [49] proposed a text mining approach. A process KDD proposed that basis on taxonomy pattern. Sequential pattern and closed sequential pattern both used pattern based taxonomy. A FP tree is used that arranged all the items in descending order according to the threshold. The items that are frequent in the text are mined using FP-growth. Dataset is categorized in two parts that is list of item and transaction id. The items meet the threshold are kept and items do not meet the threshold are eliminated. The advantage of FP tree is that candidate key is not generated and disadvantage is that face the problem locality issues. The result showed that the problem of low frequency in pattern discovery is solved with help of pattern evolving and pattern deploying techniques. The proposed .technique outperform as compare to the SVM model.

2.5.1 Simplified text for non-native speakers

Main objective of presented study [50] is to substitute complex words with the simple one's. Simple words are the main target of the audience as they are easy to learn and understand. For finding complex words, lexical simplification techniques are used. The presented study focuses on complex words for those who are unable to learn and understand the foreign language. Target audience of the study is non-native speakers. Substitution is generated with the help of lexical technique. Through part of speech tagging word sense label is obtained. As sometimes same word possess different part of speech due to its grammar context ambiguity, so same information is not produced. Training data is obtained after annotation process and is done with respect to POS tagging. Substitution selection is a process in which a replacement for the particular word is obtained. During this process focus is on capturing the substitution which does not change the meaning and form of grammar and the original context meaning is preserved. During substitution ranking process, all captured substitution are ranked according to the level of their simplicity. In this paper, new dataset is formulated in order to evaluate the simplification process specific to needs of non native English speakers. Presented results depicts that with the help of methodology significant, efficient and effective results can be achieved in Selection, Substitution Generation and Ranking.

2.5.2 Aphasic Readers and Text Simplification

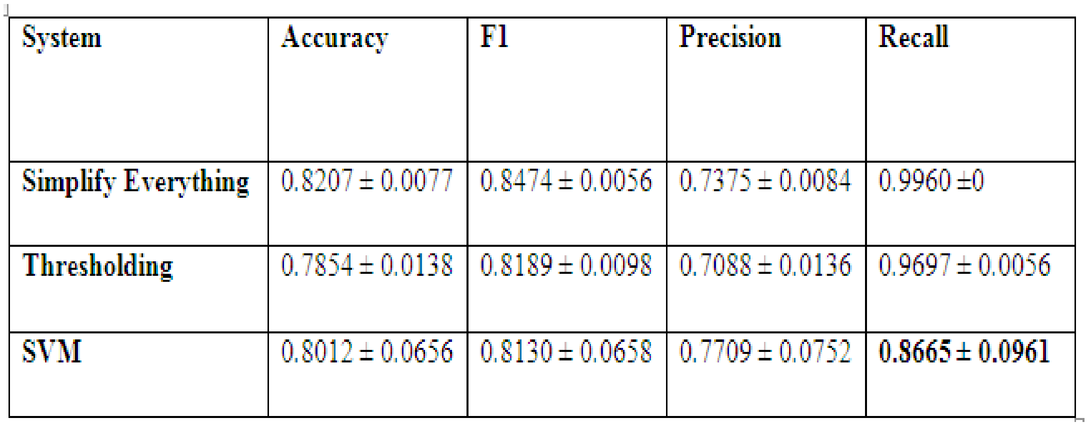
In this paper proposed by john et. al [51] a system is proposed and evaluated to assist aphasic readers during reading newspaper text. As aphasic readers have some degree of language impairment that is because of head injury or head stroke. It is hard for them to read the written text thoroughly, fluently and correctly. Developed methodology constitutes of two steps an analyzer of the English newspaper text presented for reading to aphasic readers and a simplifier that is used for simplifying the text accordingly. Analyzer is further divided into lexical tagger used for POS tagging of the text , morphological analyzer is used for lemmatizing the respective tagged text and a parser used feature based unification of grammar to parse the POS tagged text and punctuation into shallow phrase tagging further. Syntactic and lexical simplifier is used to generate the simplified text of the news presented in the news paper. Word net database is used to generate the synonyms of the word. Most frequent word is selected as simplified one and presented as an output of the simplifier. The system is experimentally evaluated in the field of aphasic readers. System was only tested against aphasic people who have reading fluency score between 3 to 8 on sentence reading tests. The assessment of the readability of the text and usability of the system depends on observations and experiments carried out with aphasic readers. In this paper no machine learning method is utilized. All the tasks are performed using the basic techniques of natural language processing.

2.5.3 Techniques to Identify Difficult Words

This paper by Shardlow [52] present and evaluate different methods to identify complex words from the text. Data set used for the purpose is an annotated corpus of sentences collected and generated from simple Wikipedia histories. Techniques used for identification of complex words constitutes of simplify everything, frequency thresh holding and training and testing a support vector machine. In simplify everything technique, word net database is used to generate most frequent synonym of the word as the appropriate substitution of the complex word. Frequency threshold technique uses SUBTLEX to provide familiarity association. Furthermore, a threshold frequency of all the words is determined by ordering the training data according to the frequency. Accuracy of the system is calculating by placing threshold frequency between frequencies of every pair of words present in the corpus. 5 cross validation method is used to generate mean threshold frequency.

Third technique presented in the paper is support Vector machine that uses many features includes frequency (how many times a word occur in the corpus) , cd count (In how many films a word appears). Length, sense, and synonym count of the word. Syllable count is also used as one of the feature.SVM was trained in Matlab and 5 cross validation was performed on held out training dataset. Results of the paper indicate that first two techniques show high recall whereas SVM shows high precision. This is due to fact that SVM uses no of features to classify word as complex or not complex. Discussion depicts that new techniques can be used to improved the results presented.

Table 2-1: Achieved Accuracy [52]



System	Accuracy	F1	Precision	Recall
Simplify Everything	0.8207 ± 0.0077	0.8474 ± 0.0056	0.7375 ± 0.0084	0.9960 ±0
Thresholding	0.7854 ± 0.0138	0.8189 ± 0.0098	0.7088 ± 0.0136	0.9697 ± 0.0056
SVM	0.8012 ± 0.0656	0.8130 ± 0.0658	0.7709 ± 0.0752	**0.8665 ± 0.0961**

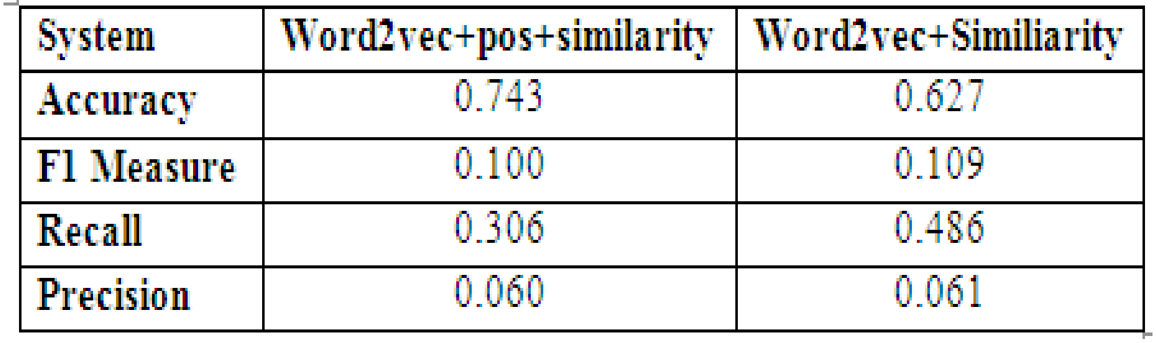
2.5.4 Context-Aware Approach

This paper [53] present and evaluate a context based approach of the word to identify whether it is complex or not. Two types of features are considered in this approach. One is base features that include linguistic and psycholinguistic features of the word. It includes POS Tags, frequency count, synonym count, length of the word. Psycholinguistic features obtained from medical research council databases include concreteness, imagery, familiarity and age of acquisition count of the word. Contextual feature is calculated using the above mentioned base features of surrounding words into consideration. When the value of n=0 it means the word is not considered on the basis of its context. Window size is generated in which all surrounding words of the target word is considered in the window. Furthermore, during this same weight is assigned to each features.. Supervised machine learning process is used for evaluating the technique. Naive Bayes and random forest algorithm is used to train and test the dataset of semi Eval -2016.Results indicate that system performs better when the context of just 5 preceding and following words are taken into consideration. Recall, precision and f-measure achieved from Naïve Bayes learning model is 0.391, 0.118, 0.181 respectively and for random forest learning model 0.548, 0.170, 0.260 respectively for context size=5. When the size increases both the model shows gradual decrease in f-measure.

2.5.5 Word Embedding

The paper presents the technique as a result of author's participation in semiEval Task 2016. Paper [54] highlights the complex word identification for non-native speakers who do not have fluency in English language. Dataset used is provided by the organizers of the semiEval task. Training data is annotated from the 20 authors. If one among twenty authors marks the word as complex then word is marked as complex. This paper focuses on word embedding feature in which unique words from the data is extracted and converted into vectors using neural model. Furthermore, other Word features include length, number of syllables, ambiguity and frequency. After that SVM is used in python to classify the test dataset as complex or not complex. Paper presented two systems with SVM as a classifier one including POS tag as a feature and second without POS tags. Accuracy, precision, recall and f-measure for the system with POS tag as a feature are 0.743, 0.060, 0.306, and 0.100 respectively. For system without POS tag measured accuracy is 0.627, precision 0.061, recall 0.486 and f-measure of 0.109. Results indicated that including POS tags as a feature generate better results. Discussion highlights that the frequency counts and embedding features are reasonable for a given resource check task, if we have provided for English language in general it might have provide with better accuracy.

Table 2-2: System Participated in semi Eval Task 2016Accuracy Achieved [54]



System	Word2vec+pos+similarity	Word2vec+Similiarity
Accuracy	0.743	0.627
F1 Measure	0.100	0.109
Recall	0.306	0.486
Precision	0.060	0.061

2.5.6 Automatic Prediction of Vocabulary Knowledge for Learners of Chinese as a Foreign Language

The paper [55] describes the system that can identify complex words from the Chinese language Presented model detect whether the respective learner knows a Chinese word or not ultimately assist the foreign learners of the Chinese language. Dataset includes 9000 Chinese words used by Chinese learners in Hong Kong. The list is constituted by the Hong Kong Education Bureau. List was annotated by the users after constructing training datasets into five different groups. This is done just to not burden the users with number of vocabulary words. Users annotated it on the basis of provided five points including whether never seen, probably seen, absolutely know but do not know the meaning, probably know the word and able to guess the meaning and in last know both the meaning and the word itself. Test Dataset constitute of 550 words that is taken from the training data keeping in view that they do not occur in both the training and testing dataset.

For classification of text as complex or not, Logistic Regression, Label Propagation and SVM machine learning models are used. Models are trained and tested on small no of dataset. Features include frequency of the character and the word. Minimum and maximum counts are taken into consideration.HSK (Hanyu Shuiping Kaoshi) and TOCFL (Test of Chinese as a Foreign Language) guidelines are used which includes 9600 and 8000 respectively different vocabulary items. These guidelines provide six levels to assess the proficiency of the foreign language. Models are implemented in scikit learn. Results indicate that SVM yields best performance with an accuracy. Logistic Regression and label propagation yields 76.3% and 72% of accuracy respectively. Limitation of the paper depict that it only focuses frequency count of the vocabulary item.

2.5.7 Sem Eval Task 2016

SemEval is a platform provided to linguistic and computing professionals to carry out linguistic tasks. It is related to semantic evaluation of computational linguistics. One of the task performed using this platform is identification of complex words. In year 2016, investigation is carried out to identify complex words from the text available on web. 42 teams take participation. Out of these 20 teams presented their systems. Performance evaluation for all the systems is based on g-score which is known as geometric mean score. System [62] which achieves highest rank in the assigned task is SVG000GG.Technique used to develop the system is ensemble based classifier. It covers total of 69 word features. This is the only system which highlights number of features. After prediction of label from classifiers used in ensemble learning, system generate the final resulted label based on hard voting technique. Results are validated through five cross validation process on dataset. Authors participated in the task highlights following important challenges to achieve accurate accuracy during classification of difficult words; Identification of ambiguity, avoiding jargons and dyslexic sentences. Substitution must be context aware that is original meaning of the sentence is not removed. Various parameters used for lexical simplification of the word must be investigated.

[63] Presented a system for complex word identification from sentences for English readers. Authors presented two systems. One of their systems uses Naive Bayes as classifier learning algorithm. Second system uses Random Forest techniques for the task. Classification is done after extracting eight different features related to lexical and semantic features of the tokenized words. These include frequency count, named entity, stop words, syllable count, hypernym and hyponym, and synset size. Post processing technique is also applied based on rules to achieve enhanced performance. Naïve Bayes classifier achieves an accuracy of 76.7%.

For identification of complex words, paper [64] is submitted to semEval 2016.Two systems were proposed and developed using scikit-learn. After preprocessing technique, feature extraction method is applied. Extracted linguistic features constitutes of word count, average age of acquisition of a lemma, pronunciation count and synset. One of the proposed systems uses decision tree and other uses regression tree learning model with depth of three. Results depicted that system developed using decision model achieves an accuracy of 84.6% and with that of regression, accuracy obtained is 83.8%.Precision and recall of regression tree model is 0.182 and 0.752 respectively. With that of decision tree, recall and precision is 0.698 and 0.189. Limitation includes low precision due to the availability of one annotator as a judgment on tested data.

2.5.8 CWI Shared Tasks 2018

BEA workshop included the report conducted on findings related to second complex word identification shared task. Multilingual and multi-genre datasets are featured in it. Further, these datasets are divided into four tracks i-e English monolingual, Spanish monolingual, German monolingual and multilingual with French test set. Two types of task are included; binary classification and probabilistic classification. 12 teams submitted the results using one of the two tasks and tracks. While, 11 of those teams wrote system description papers which are reported as well as made the part of BEA workshop.

Baseline system is a simple basic system that only uses frequency and length as basic features. It was built for both binary and probabilistic classification. Shared task system included the Unibuc Kernel [56] team who participated on monolingual CWI shared task. The dataset include specific domains i-e News, Wikinews and Wikipedia. Two types of setups were used. One was binary classification setup that used SVM classifiers. Another was regression setup that used V-SVR.

SB@GU system [57] is adapted from shared task system. In this system, experiments are conducted for various datasets. For English dataset, features to be focused were context-free, context-only and context-sensitive. The resultant feature described word sense, topic, frequency and language model probabilities. Hu-berlin is a system design specifically for multilingual binary classification task. It is mainly used for exploration of the n-gram features and its usage [58].

CFILT-IITB developed system in [58] that focused on English monolingual binary classification task. Lexical features of the target word were extracted. They used ensemble classifier, combination of 8 classifier including J48, Random forest, PART, SVM, logistic model tree technique for prediction purpose.

NLP-CIC presented a system for English and Spanish multilingual binary classification tasks. The morphological features of the target word were tested [58].

LaSTUS/TALN presented a system for English monolingual binary classification tasks. For this purpose, two systems were designed. One system based on set of lexical, semantic and contextual features whereas, the second system focused on word embedded features [59].

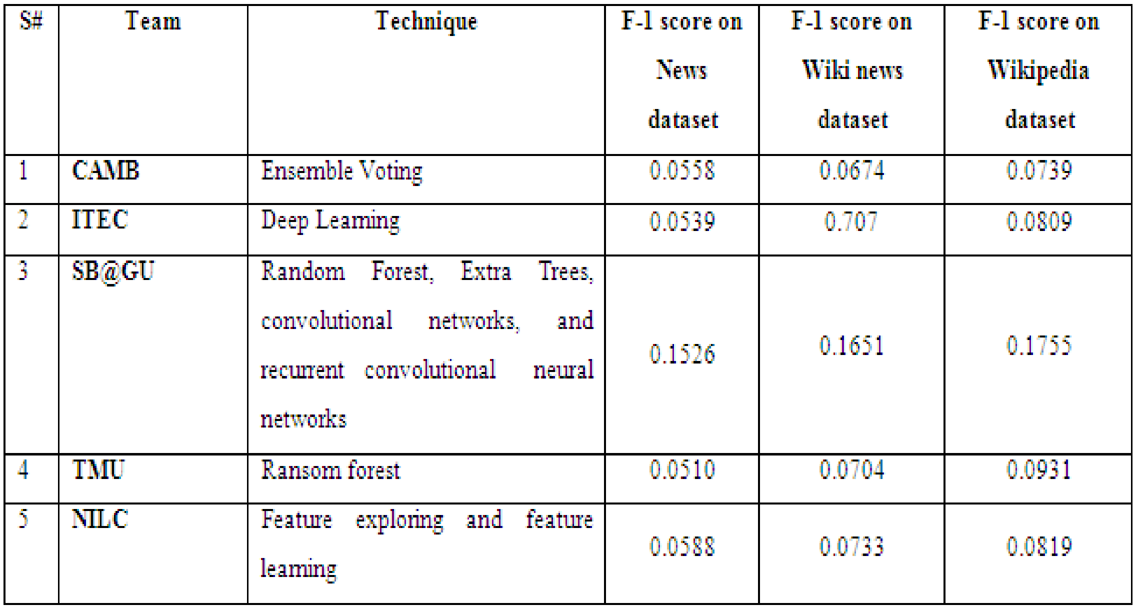
CoastalCPH explains a system that is use to investigate whether multitask learning could be applied to cross-lingual CWI task. It was specifically designed for multilingual and cross lingual domains [58].

Camb described various systems specially developed for English monolinguals focusing on both binary and probability classification tasks. They used lexical features, word n-gram, POS tags and dependency parse relations [60].

System [61] submitted by TMU includes features related to number of characters in the word, number of words in the phrase, and frequency of the word present in learner corpus. Random forest is used in case of binary classification, whereas random forest regressor technique is used for probabilistic classification. They used scikit–learn library for implementation purpose. They achieved the f-1 score of 0.863.

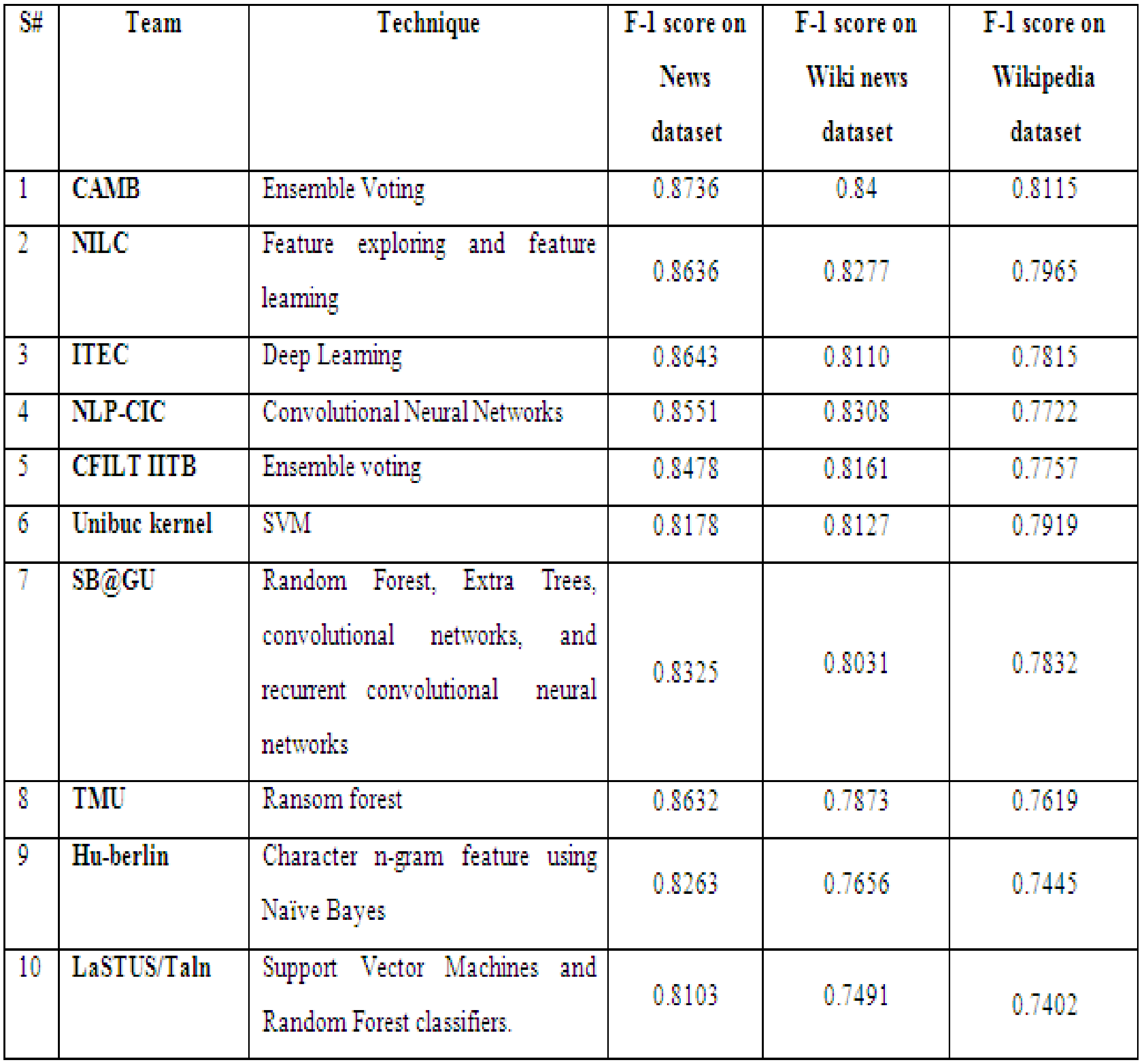
Out of these above mentioned task, system developed by the camb achieved the highest F-1 rank with score of 0.8736 using dataset of news, in case of wiki news F-1 score is 0.84 and in case of Wikipedia corpus, rank is 0.8111 for English language.

Table 2-3: Results of Probabilistic Classification in CWI Shared Task 2018



S#	Team	Technique	F-1 score on News dataset	F-1 score on Wiki news dataset	F-1 score on Wikipedia dataset
1	**CAMB**	Ensemble Voting	0.0558	0.0674	0.0739
2	**ITEC**	Deep Learning	0.0539	0.707	0.0809
3	**SB@GU**	Random Forest, Extra Trees, convolutional networks, and recurrent convolutional neural networks	0.1526	0.1651	0.1755
4	**TMU**	Ransom forest	0.0510	0.0704	0.0931
5	**NILC**	Feature exploring and feature learning	0.0588	0.0733	0.0819

Table 2-4: Binary Classification Accuracy achieved in CWI Shared Task 2018



S#	Team	Technique	F-1 score on News dataset	F-1 score on Wiki news dataset	F-1 score on Wikipedia dataset
1	**CAMB**	Ensemble Voting	0.8736	0.84	0.8115
2	**NILC**	Feature exploring and feature learning	0.8636	0.8277	0.7965
3	**ITEC**	Deep Learning	0.8643	0.8110	0.7815
4	**NLP-CIC**	Convolutional Neural Networks	0.8551	0.8308	0.7722
5	**CFILT IITB**	Ensemble voting	0.8478	0.8161	0.7757
6	**Unibuc kernel**	SVM	0.8178	0.8127	0.7919
7	**SB@GU**	Random Forest, Extra Trees, convolutional networks, and recurrent convolutional neural networks	0.8325	0.8031	0.7832
8	**TMU**	Ransom forest	0.8632	0.7873	0.7619
9	**Hu-berlin**	Character n-gram feature using Naïve Bayes	0.8263	0.7656	0.7445
10	**LaSTUS/Taln**	Support Vector Machines and Random Forest classifiers.	0.8103	0.7491	0.7402

2.5.9 CWIG3G2

Study [65] depicts those previous studies only focuses on the dataset annotated from the non native speakers, considering only the needs of such speakers. Formulated datasets are based on the data collected from Wikipedia sources only. Authors [65] presented CWIG3G2 which is a complex word identification of three genres and two user groups. A new dataset composed from three sources including wiki news, Wikipedia, and news. Annotators are not provided with pre selected tokens but a passage is provided to them consist of three to four sentences unlike of the previous studies. A technique of nearest centroid classifier is also used to evaluate the performance along with other binary classification algorithms. Nearest centroid achieves the highest accuracy. Features consist of POS tagging, character count, frequency count, vowel count, word2vec features and a topic features. Extraction of topic feature includes topic relatedness of a word within context by computing cosine of similarity value. System achieves an f-score of 35.44 which is better than best f-score of semiEval 2016 task that is 35.30.Achievement of better f-score as compared to shared task 2016 is due to the fact that presented dataset is annotated from two different annotators belonging to two different domains. And even distribution of difficult words in test and training data exist.

2.5.10 Complex Scientific Terms

[66] Highlights the research associated with the lexical simplification for converting complex scientific terms used in different domains of medicine and sciences into their simpler forms. Dataset is consists of tokens generated from 2971 articles obtained from domains of medicine, genetics, and pathogens. Complex word is identified after extracting word length, pos, frequency count, proper nouns, and synonym count as a feature of the generated tokens. Conditional random forest learning model is used for this task. It is an iterative model which achieves an overall accuracy of 90% with recall of 0.83 and precision rate is 0.74. Substitution of complex terms system achieve an overall accuracy of 71%. This is due to the fact that system does not find a suitable substitution for the term.

2.6 Summary

After going through the studied literature, the importance of automated systems is analyzed especially for hearing impaired children. Systems are developed to assist hearing impaired students to enhance their learning skills and vocabulary knowledge. Different systems are developed to identify complex words from the text. Studied literature usually focuses on needs of non-native speakers of different languages. They focus word frequency, word length, part of speech as their main features along with other features related to hypernym, hyponym, syllable count etc. Studies do not focus hearing impaired children for identification of difficult words.

Chapter 3

Methodology

3.1 Introduction

Research methodology is the systematic, theoretical analysis of the procedures applied to a field of study. Methodology involves procedures of describing, explaining and predicting phenomena so as to solve a problem; it is the 'how'; the process, or techniques of conducting research.

(Kothari, 2004)

In this chapter, we discussed in detail the proposed methodology used for identification of difficult English words in order to assist the Hearing Impaired Children in learning Language. Proposed methodology not only assists the children but also improved their learning, reading, speaking and writing skills through enhancement in their vocabulary knowledge. Most of the previous studies mainly focus non-native speakers. Rare research is done related to identification of difficult words related to Hearing Impaired Children. So, this research is done to highlight natural language processing through implication of machine learning in the domain of teaching and assisting hearing impaired children in learning English language. As we know learning language is a tedious and tough for hearing impaired children due to less knowledge of vocabulary. And moreover, if we present them with the difficult text to build the vocabulary knowledge then it is a much more complex for them to acknowledge that text and learn it. This happens due to the fact that first of all they do not understand it all and secondly, it becomes a tedious task and makes the child stubborn to learn about it despite of several rehearsals and repetition of learning activities. So, it is the need of hour to provide child with the text that is easy for them to learn, and not so difficult for them. Identification of difficult words also helps teachers, speech language pathologists to devise such learning activities that depict these difficult vocabularies in their simplest form. So, that child learns them in short of span of time. It also lessens the task of the teachers.

Overall contribution of this research methodology is to

- Developing a full fledge data set
- Highlighting the feature set which is specific to hearing impaired children
- Enable child to classify the words as difficult or not difficult from the text.
- To enhance child's reading and writing skills through increase vocabulary knowledge.

Step by step complete methodology is discussed in detail. We proposed a methodology and developed a tool for detection of difficult words through classification process from the unstructured text which is in the form of word documents, PDFs, available on online websites, blogs or in textbooks of English language specific for hearing impaired children. System also generates a synonym list for words to be substituted in replacement of difficult words.

Proposed methodology is composed of following parts

- Collection of Dataset
- Pre Processing
- Feature Extraction
- Classification of Text
- Evaluation of the proposed Methodology

Steps involved in pre-processing of English text written in free-text format, technique(s) used for detection of difficult words are explained in detail. Tools and programming languages used for implementation of proposed tool are also discussed. Architecture diagram of developed tool is also added to show that how identified difficult word(s) are displayed to user. Detection is strongly dependent on the dataset. Data set and linguistic rules specific for hearing impaired children plays an important role for efficient classification of difficult words. Good results are produced by good data and bad data can mislead the researcher. Especially for tool, the quality and reliability of detection is dependent on data set. Formulated dataset is discussed in detail.

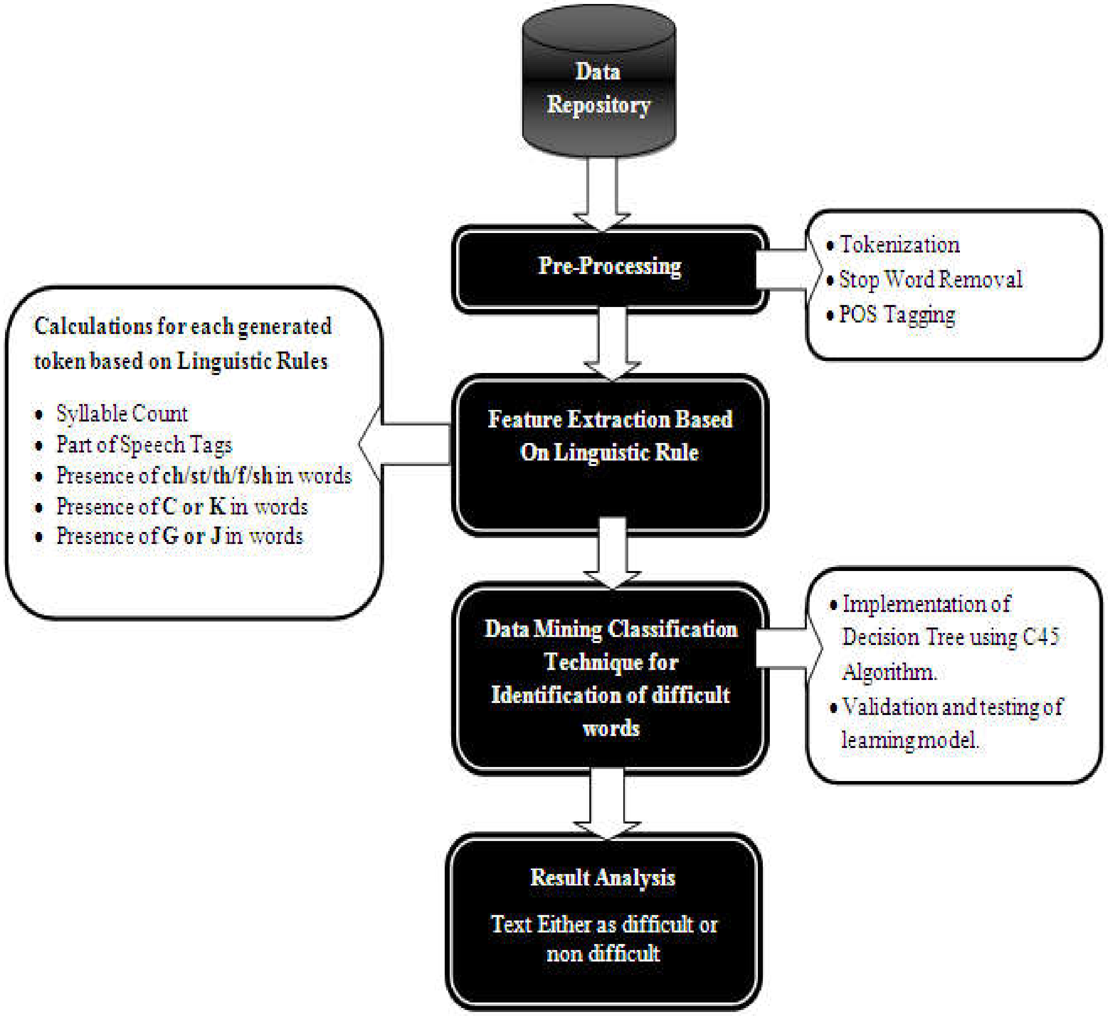

Data Repository
Pre-Processing
• Tokenization
• Stop Word Removal
• POS Tagging
Calculations for each generated token based on Linguistic Rules
• Syllable Count
• Part of Speech Tags
• Presence of ch/st/th/f/sh in words
• Presence of C or K in words
• Presence of G or J in words
Feature Extraction Based On Linguistic Rule
Data Mining Classification Technique for Identification of difficult words
• Implementation of Decision Tree using C45 Algorithm.
• Validation and testing of learning model.
Result Analysis
Text Either as difficult or non difficult

Figure3.1: Proposed Methodology

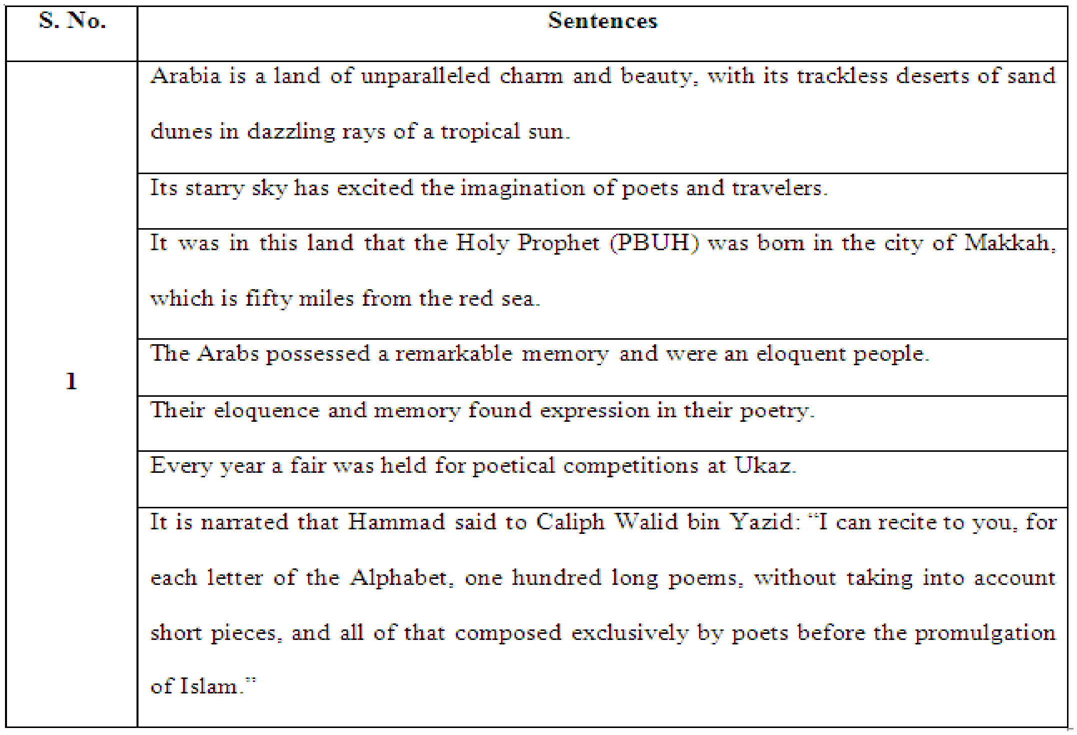
3.2 Collection of Dataset

Collection of dataset for the respective research is a complex task in itself because of targeted audience of hearing impaired children irrespective to the previous studied literature. We use two different datasets. Dataset 1 which is developed manually based on the evaluation of the real time annotators and used during training phase. Main purpose is to label the word as difficult or easy word. Dataset 2 is used in testing Phase. Both Dataset constitutes of word list of English Language.

3.2.1 Description of Data Set 1

Data Set1 comprised of the 100 different paragraphs constitute of minimum of six sentences each. Out of these 100 paragraphs, nearly 50 are collected from the syllabus of hearing impaired children being taught at Secondary school level in hearing impaired schools of Pakistan. And, 50 paragraphs are collected from online source after consulting teachers, parents and SLP's. Sample sentence used in dataset is represented in table 3-1 and sources used to construct dataset are shown in figure 3.2.

Table 3-1: Sample Sentences used in Dataset



S. No.	Sentences
1	Arabia is a land of unparalleled charm and beauty, with its trackless deserts of sand dunes in dazzling rays of a tropical sun.
	Its starry sky has excited the imagination of poets and travelers.
	It was in this land that the Holy Prophet (PBUH) was born in the city of Makkah, which is fifty miles from the red sea.
	The Arabs possessed a remarkable memory and were an eloquent people.
	Their eloquence and memory found expression in their poetry.
	Every year a fair was held for poetical competitions at Ukaz.
	It is narrated that Hammad said to Caliph Walid bin Yazid: "I can recite to you, for each letter of the Alphabet, one hundred long poems, without taking into account short pieces, and all of that composed exclusively by poets before the promulgation of Islam."

3.2.2 Annotation process

These sentences, in the form of paragraphs are presented to teachers of hearing impaired children for annotation process. They are considered as important observers and mentors for hearing impaired children during their learning process. Teachers know their needs, perception and thinking ability better than anyone else because they spent much time with them. They keenly observed them during learning activities being conducted in class. Five teachers participated in the activity and highlight the words from these paragraphs that they find difficult to understand for hearing impaired children during reading and writing activities performed in class. During this process, almost 1000 distinct words are marked as difficult (1) or easy (0) by the teachers. A word is difficult if any of the two teachers marked it as difficult. After getting collection of annotated data, we preprocessed it. Annotated data is in the form.

<S. No.><Sentences><words><label>

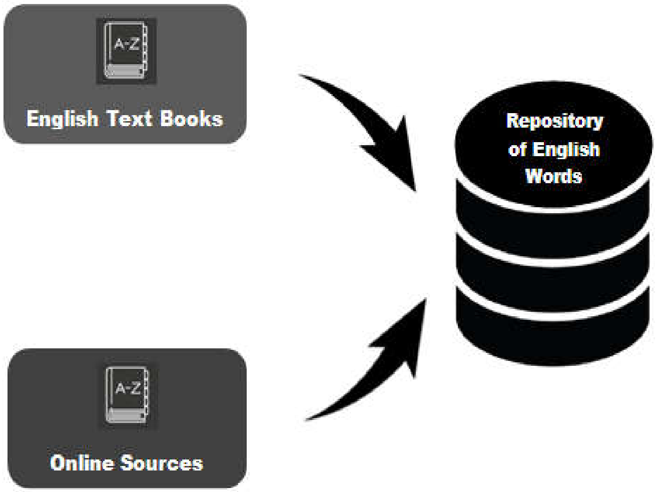


Figure 3.2: Sources used in Dataset

Brief description of dataset is shown in table 3-2.

Table 3-2: Brief Description of Dataset

Total Paragraphs	**100**
Online Sources	25
Textbook Collection	75
Total Words	1000
Words used to train data	600
Words used to test data	400

3.2.3 Description of Data Set 2

To maintain the state of the art, the technique is applied on another dataset as well to check accuracy and for evaluating the proposed methodology. A single paragraph is selected from the dataset which is composed of sentences. Words in these sentences are annotated by native and non native speakers

3.3 Pre-Processing

As English text in textbooks or available online in blogs, websites contains unstructured data written in natural language, so pre-processing of text is necessary. It is a combination of activities in which written text documents in unstructured form are pre-processed. Main objective of this process is to improve the efficiency and effectiveness of the system by eliminating data redundancy. Stop Word removal helps to reduce the data size as they constitute 20 to 30 % of the total words in any English text document. Part of speech help in generating POS tags of the word and contribute towards data analysis. Other techniques like lemmatization or stemming help in reducing the data size by removing majority of the inconsistent data.In our methodology, we use three basic steps of Text preprocessing.

- **Tokenization**

Tokenization is a process in which paragraphs are broken down into sentences and sentences are further divided into single stream of words, symbols or other elements known as tokens. Main objective is to find words in a sentence. In further processing, these tokens are used as input. Most of the words in English language are separated through white spaces. But sometimes other delimiters like punctuation marks can be used to break down the words further if required. Tokenization can be done at paragraph, sentence level and at word level.

For example, In order to identify which words are difficult or not from the sentence mentioned below, first of all we have to tokenize it.

Sentence: Newspaper is considered as one of the best means of communication.

After Tokenization: Newspaper , is, considered , as , one , of , the , best , means , of , communication, . ,

- **Stop Word Removal**

In a text documents, many words occur frequently to fulfill rules of sentence structure. They themselves usually play no key role other than joining sentences together. They are common words and not useful for classification of the text documents. So there is need to remove these words from the text to save computational and processing time and space. Stop words generally constitutes 'the', 'an', 'and', 'or', 'This', 'that', 'these', 'those' and so forth. NLTK build in English stop word list is used in the tool to detect and remove the stop words from the text. List is also modified accordingly to enhance the quality of stop word removal. Punctuation marks or delimiters like comma, brackets, period, colon and semi colon can also be removed by adding them to the respective list. Otherwise they can also be removed using regular expressions.

For example: Consider a generated token list for which we are identifying difficult words

Newspaper , is, considered , as , one , of , the , best , means , of , communication, . ,

In above mentioned example, stop words can be removed using the following approach.

If token 'a' exists in the stop word list, then removed. Otherwise retain the token in the array list.

So, resulted array list after removal of stop words includes [Newspaper, Considered, Best, Means, and Communication].This is the final list generated out of which we will be identifying which one are marked as difficult and which will be marked easy by the classifier.

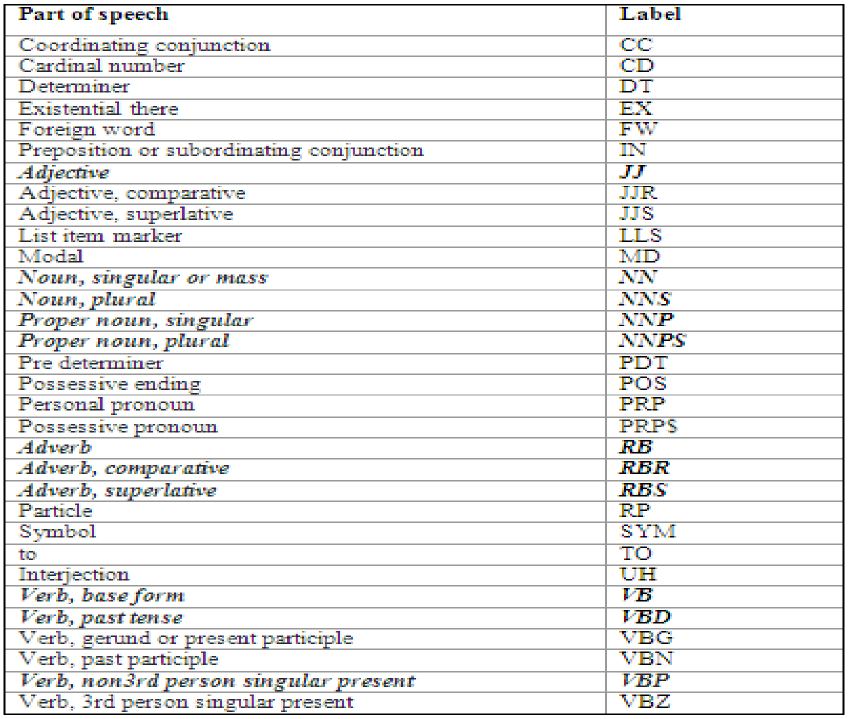
- **Categorizing words into Part of Speech**

Categorizing words/tokens into different part of speech is done through natural language processing technique known as tagging part of speech. It is used to break sentences or words into different Part of speech.

- Noun
- Verb
- Adjective
- Adverb
- Pronoun
- Preposition
- Conjunction
- Interjection

For detection of difficult word from the tokenized words, it is essential to divide words into different parts of speech. Label of part of speech tagging categorizes each word. Parts of speech and its corresponding label are shown in table 3-1. We have considered four major categories of part of speech that is noun, verb, adjective and adverbs. Types of this part of speeches are considered under the umbrella of their parent category. For example proper nouns are considered as noun, comparative and superlative adjectives are considered as adjectives and soon. Most common parts of speech are highlighted in Italic bold in table 3-1.

Table 3-3: Representation of Part of Speech Tagging



Part of speech	**Label**
Coordinating conjunction	CC
Cardinal number	CD
Determiner	DT
Existential there	EX
Foreign word	FW
Preposition or subordinating conjunction	IN
Adjective	***JJ***
Adjective, comparative	JJR
Adjective, superlative	JJS
List item marker	LLS
Modal	MD
Noun, singular or mass	***NN***
Noun, plural	***NNS***
Proper noun, singular	***NNP***
Proper noun, plural	***NNPS***
Pre determiner	PDT
Possessive ending	POS
Personal pronoun	PRP
Possessive pronoun	PRP$
Adverb	***RB***
Adverb, comparative	***RBR***
Adverb, superlative	***RBS***
Particle	RP
Symbol	SYM
to	TO
Interjection	UH
Verb, base form	***VB***
Verb, past tense	***VBD***
Verb, gerund or present participle	VBG
Verb, past participle	VBN
Verb, non3rd person singular present	***VBP***
Verb, 3rd person singular present	VBZ

- **Stemming**

As we know main part of word is stem, base or root of word. Actual meaning is conveyed by stem or root of word. A base word is changed in following ways:

- **Inflection:** For syntactic roles word is modified. Suffix is added. Few examples of inflection are shown in table 3-4
- **Derivation:** Existing words are used to derive new words. Adjective is added as a suffix. Few examples of derivation are shown in table 3-5
- **Compounding:** Existing words are grouped to form new words. Few examples of compounding are shown in table 3-6

Table 3-4: Suffix Types

Original Words	Suffix Type	Modified Word
Doctor	Adding *'s'* to make it plural	• Doctors
small	Adding *'er'* and *'est'* to modified word in a comparative and superlative form	• Smallest • Smaller

Table 3-5: Adjective Added as Suffix

Original word	**Adjective**	**Modified word**
Sad	ness	• Sadness
End	less	• Endless

Table 3-6: Compounding Words

1st word	**2nd word**	**Combined Form**
fare	well	• Farewell
Cup	Board	• Cupboard

We do not apply stemming or lemmatization processing on a text and words are not in its base form, because when words are modified their meaning also changes in the context. Adding suffixes and prefixes to the base word reflected in table results in formation of new word which results in increase knowledge of vocabulary. Affixes constitute of both prefixes and suffixes help the vocabulary learner about learning unknown words by relating it to known words and secondly to build the knowledge related to the affixes (i.e. both prefixes and suffixes).So, affixes and prefixes plays a key role in enhancing vocabulary of hearing impaired children. Formation of new words enhances the complexity of the text and posses as a hurdle for hearing impaired children during their learning process in early years.

3.4 Feature Extraction

Feature extraction is considered as an important and critical step while developing machine learning applications. Algorithms used for machine learning automatically generate output labels through statistical models that map the input data to output label. For example during spam classification, email is taken as input and whether it is a spam or not is considering as an output. Output label is either a discrete label or depicts the probability of email being spam or not. Machine learning algorithms require inputs represented as set of features along with the relevant output labels. It is necessary to map the raw input into features. Features are actually the characteristics to shows that if a given input possesses these characteristics then respective output label is assigned to it otherwise not. The process of selection of features is the most important and technical part which effects the working of whole model so, if the features are properly extracted the complete pipeline will work in a very effective manner. The technique of feature extraction is useful in terms that it minimizes the cost of computation and also increases the overall performance of the classifier being used in the technique.

Studied literature mostly focuses on word frequency, length, part of speech, stop words, syllable count, hypernym, and hyponym and synset size as features of the words. They have developed systems keeping in view a general public mostly for non native speakers. In our proposed methodology, we have use and focused linguistic rules to extract different features. These linguistic rules depict relevant and specific information about why words are difficult for Hearing impaired children.

These linguistic rules are described briefly with respect to characteristics as follow:

3.4.1 No Stemming

We do not used word stemming technique in which words are converted into their base form This is because of the fact that hearing impaired children face difficulties when a suffix or prefix is added to the base word. They usually possess little knowledge about the plurality of the word and tenses. Adding –s, -ed, -ing to the word enhances the trouble for them to understand the word and then interpret it.

3.4.2 Tagging / Categorization of words into Part of Speech

It is one of the steps that are done during preprocessing of the text. Grammar is the key part of any language as it constitute of rules that deal with the morphology, syntax and classification of the word depending on their part of speech. Generally, all words in English language are divided two main classes known as content based word class and function based word class, which is further divided into sub categorized into eight different parts of speech. Content based word class constitutes of all the words that have some concrete meaning and can be modified to form new words. All new words of the English language come under the umbrella of this class. Word frequency in a text is low due to which they are relatively difficult to understand compared otherwise by the people. Function based word class comprised of shorter length words, have abstract meaning or no meaning at all and occur frequently in the text document. Noun, verb, adjective and adverb belong to content based word class whereas pronoun, conjunction, interjection, and preposition belong to function based word class. It is a general perception depicted through studied literature that hearing impaired children find concrete words representing a definite meaning easy to learn as compared to abstract words that does not have any specific meaning. For example, words like cat, jump, run, eat can easily be learnt whereas words like after, before, frequently are difficult for them. So concrete words are easy and abstract words are difficult.

In our proposed methodology we have considered only the content based words categories that are nouns, verbs, adjectives and adverbs. In case of hearing impaired children noun and verbs are much easier for them to understand when compared to adjectives and adverbs.

POS Tagging approach is used to find the difficult words. This feature is not discrete rather it represent categorical approach consider to mark the word as difficult or not. Words are marked with four distinct categories. All nouns present in the dataset are marked with 1; verbs are represented with 2, adjective as 3 and adverbs as 4.

3.4.3 Character count / Word Length

Words that are longer are difficult to understand compared to shorter length words. Hearing impaired children usually focuses on sorter length words as they are easy to remember after few rehearsals.

Length based approach is used to find whether the word is difficult or not.

- If ‘n’ represent a word in a dataset then if length (n) >7, then word is difficult (1) otherwise it is easy (0) for learner to learn.

3.4.4 Syllable Count

All words with greater syllable count are difficult to understand as it is very difficult to pronounce such words. Syllable count approach is used

Algorithm to calculate Syllable Count

1. Find the total number of vowels in the word.

2. Subtract 1 every time from the total count if a word n ends with ‘-e’, ‘-es’, or ‘-ed’.

3. Add 1 in total count if word n ends with ‘-le’ and have constant at the end of the word after trimming ‘–le’.

3.4.5 Presence of -ch/-st/-th/-f/-sh in words

High frequency sound words are difficult to hear, understand and interpret by the hearing impaired children. As their frequency range is nearly or more than 2000 KHZ so they are unable to hear it clearly. **-Ch, -st, -th** and **-sh** diphthongs and letter f, t and k occur in high frequency sound range. Presence of these in words makes them difficult to understand.

For example, the word ‘taste’ is difficult to understand and causes trouble for hearing impaired children during writing as he/she is unable to interpret it properly. This is because of presence of t in the start of the word which produces high frequency sound and -st in the middle make it a more challenging task. Similarly, among word such and very, ‘very is easy for the child because pronunciation of this word falls in the category of low frequency as compared to the such which is a high frequency sound word due to presence of ‘-ch’.

Algorithm

- Check whether the word n contains ‘-st’, ‘-ch’, ‘-sh’, ‘-f’, or ‘-th’.
- If True then assign ‘1’ otherwise return ‘0’.
- End.

Algorithm result depicts that word is considered as difficult when it is true otherwise it is easy.

3.4.6 Presence of C or K in words

Pronunciation of words plays an important role during interpretation of the word in both reading and writing process. If an individual cannot hear or understand the pronunciation clearly then he/she is unable to do interpret it. Pronunciation of ‘C’ or ‘K’ sometime intermingled with each other. For example, in a word cake pronunciation of letter ‘C’ is similar to that of ‘K’. It is not pronounced as ‘see’. Similarly the word communicate in which c produces the sound of ‘kay’ whereas in ‘process’ it is pronounced as ‘see’.

So, it is depicted that words containing either c or k makes the word difficult otherwise it is easy.

3.4.7 Presence of G or J in words

Similarly, presence of G or J in words makes them harder to understand. For example pronunciation of jug, guest, gel, gym always confused hearing impaired children what to write and how to spell after hearing the word with the help of hearing aids.

3.5 Text Classification

After feature extraction, classification of the text is the most important and crucial step. Text Mining is considered as one of the broad field in the domain of computer sciences and artificial intelligence. It gained a lot of consideration due to large amount text data exist in the form of medical records, news, blogs, reports etc. Text data is in the form of unstructured text which is easily perceived by humans but harder to understand by the machines. So there is an usher need to process, organize and retrieve information from this unstructured data and effectively use this in different applications. This can be done effectively through implication of different algorithms design for data / text mining during machine learning. Machine learning deals with the learning of the system, from given examples and experiences and providing step by step solution to a given problem. There are two basic methods of machine learning that is supervised and unsupervised machine learning. In unsupervised machine learning process, deals with unlabeled structured data. There is no need to train the data. Clustering is one of the techniques based on unsupervised learning method in which same collection is clustered together while the different grouped together to form a new cluster. In supervised machine learning, there exist an N number of input variables and M number of output variables. Labeled structured data is used for training the learning model and then predicted the unseen data based on the training data. Training data plays the role of teacher which is supervising the overall classification process. Classification is applied when categories exist as values of output variable and when there exist real time values like weights etc then it is advisable to use regression method. Support vector machine, decision tree, neural networks, naive bayes are some of the supervised machine learning algorithms.

Classification of difficult English words is achieved through implementation of C4.5 decision tree learning algorithm. For this purpose other classifiers studied in literature are used to identify complex words from the text. Classifier's like SVM (Support Vector Machine), Artificial Neural Networks, Ensemble learning models are different techniques used to classify the text as difficult or not [59]. Main purpose of using these approaches is to enhance accuracy of the classification model.

3.5.1 Why use C4.5 decision tree learning algorithm?

For detection of difficult words from the text document, we used decision tree (C45 Algorithm) classifier. Decision tree look likes a flow chart or depicts a tree in its structure constitute of internal or external nodes. External nodes are also known as leaf nodes which in particular represent a label. The top node is a root node. Braches in a tree like structure represents the outcome of an instance. During classification process, path of the unknown input is traced and tested starting from the root node and ends at the leaf node representing finally the output label for that input. For implementation, they do not require any prior domain knowledge rather due to its tree like or hierarchal structure acquiring knowledge and interpreting it is really easy for humans. They are used for classification in application developed in the domain of artificial intelligence, manufacturing, medicines, expert systems, or for financial analysis. Rules can easily be induced from the outcome of the decision tree as it is considered as an inductive learning task. In this, one is using particular facts to make more generalized conclusions

There are different algorithms developed from time to time to implement decision tree. J. Ross Quinlan in 1980's developed ID3 (Iterative Dichotomiser3) algorithm which uses information gain to split the dataset at any node and select the specific attribute.ID3 splits attributes based on their entropy. It does not work best when there exists a noisy data or missing value in the data sets. Entropy is the measure of disinformation.

In 1993, J. Ross Quinlan developed another algorithm named as C4.5 which is an improved version of ID3.It uses gain ratio rather than only information gain to split the data and select the attribute at any specific node. It is capable enough to handle both discrete and categorical data. For continuous variable it automatically calculates the specific threshold which is an average value for that attribute to it. Like ID3 it also calculates the entropy and information gain of specific attribute.

$$Gain\ Rat\text{io}(S, X) = Information\ Gain\ (X)/\ Split\ Information\ (S,X)\ (3)$$

Where Split Information can be calculated as:

$$Split\ Information\ (S, X) = - |Si\ |\ |S|\ y\ i\text{=1 log2 } |Si\ |\ |S|\ (4)$$

Split information is the information generated by dividing the training set into y partitions. Attribute having highest gain ratio is selected as a splitting attribute.

In order to identify difficult words from the text we used C4.5 decision tree algorithm. This is due to the fact that C4.5 has following advantages over ID3 learning algorithm.

- It can easily handle discrete and categorical based variables/attributes during splitting process.
- It can handle noise or missing value attributes in present in the dataset. It simply do not consider missing values for the attributes in gain or entropy calculations.
- Thresholds can easily be used for continuous variables if exist.
- Rules can easily be generated to be used in future.

3.6 Tools and programming languages used:

Tools and programming languages used are listed below.

- **Visual Studio 2017:** It is integrated development environment (IDE) of Microsoft. Visual studio 2017 is latest version of visual studio with user friendly interface. For development of desktop applications, web applications, websites and mobile applications it is used. Windows form, windows store, windows presentation format, Microsoft Silverlight and windows API are platforms of Microsoft development which are used by visual studio for development of different applications. Both the managed code and native code are developed by visual studio. It includes code editor and code refactoring tool. Different languages of Microsoft such as C++, C# and ASP.net are used in visual studio for development of applications.

- **C#**: It is language of .net framework. It is object-oriented programming language. It contains the extensive libraries for development of desktop applications. It encompasses following programming languages disciplines:

    - String typing

- Imperative
- Declarative
- Functional
- Generic
- Object-oriented (class-based)
- Component-oriented

- **Sharp nlp:** Sharpnlp is tool written in C# used for processing of natural language text. Following natural language processing tools are provided by sharp nlp

    - Part of speech tagger
    - Sentence splitter
    - Chunker
    - Tokenizer
    - Parser
    - Name Finder
    - Coreference tool
    - Interface to the WordNet lexical database

- **Accord.net:** The Accord.NET framework is a full fledge machine learning framework written in C#. It is used for processing of statistics, image and signal processing, artificial neural networks and support libraries used for plotting graphs and visualization. Framework provides following complete package of supportive libraries to fulfill functionalities.
    - Accord.Math,
    - Accord.Statistics,
    - Accord.Machinelearning,
    - Accord.Imaging,
    - Audio. Vision,
    - Accord.Controls,
    - Accord.Controls.
    - Accord.Imaging,

- Accord.Controls. Audio,
- Accord.Controls. Vision

Different classes are also used to implement the decision tree including

- Regular Expressions for extracting features
- Machine learning decision tree class
- Codification class for handling numerical and categorical data
- For Generation of decision rules
- For Predicting new output labels
- For Pictorial representation of tree

See Appendix-A for coding of each steps in detail.

3.7 Summary

After getting the data related to difficult words from different sources, datasets are developed and annotated using domain expert's knowledge. For detection of difficult words from the text, after implication of preprocessing and feature extraction techniques, we used C4.5 Decision tree Machine Learning Algorithm during classification module. Dataset is used for training and cross validation purpose whereas dataset 2 is used for testing purpose. Effectiveness of proposed methodology is discussed in next chapter.

Chapter 4

Results and Discussion

4.1 Introduction

We have discussed the proposed methodology in detail step by step. This chapter focuses on evaluating the effectiveness of methodology. Comparison with existing techniques or methodology is also discussed.

4.1.1 Overview:

After, going through, all the steps of proposed methodology by using text available in textbooks, websites or study material used during learning activities. Paragraphs are selected and then provided to experts for annotation process. After attaining the annotated data, we have done preprocessing of the data. Outcome of each step is discussed in detail.

We have 1000 samples of English word are collected after preprocessing in our data set1 which is an annotated data for the classification of difficult and easy words. We have labeled the words as 1 and 0 which shows whether the word belongs to difficult word class or non difficult word class for hearing impaired children respectively. Out of total 1000 words, 600 word samples are used for training purposes while rests of the 400 words are used for testing. Dataset2 which constitutes of 50 samples in total and is only applicable for testing purposes.

5-fold Cross Validation on complete dataset1 is used to evaluate the accuracy of the model. Preprocessing and feature extraction steps are same for both the datasets whereas during classification technique only tested datasets are different. For testing dataset1 and testing dataset2, training dataset is same which constitutes of 600 words whereas for cross validation 1000 words, a complete dataset is divided into folds and then evaluated.

4.2 Evaluation of Results:

Results of all three steps of proposed methodology are discussed below.

4.2.1 Preprocessing of text

4.2.1.1 Tokenization and stop word removal

This step is applied on the 100 paragraphs of dataset1 and a single paragraph which constitutes our second dataset. Tokenized words are then passed to stop word removal in which all words that exist in NLTK Stop word list are removed from the tokenized list. Along with this, words containing numbers and special words are also removed. Sample of obtained results are shown in table 4-1.

Table 4-1: Tokenized List of English Words

Results after tokenization and stop words
Adequate
Perilous
Delicate
Calamity
Mild
Care
Wish

4.2.1.2 Part of speech tagging

Part of speech tagging is applied on tokenized words. Different parts of speeches are identified for tokenized terms. Part of speech tagging is done in broader aspect of four main categories. Obtained results are shown in table 4-2.

Table 4-2: English Words and their Part of Speech Label

Tokenized Terms	Parts of speech Label
Adequate	JJ
Perilous	JJ
Delicate	JJ
Calamity	NN
Seek	VB
Care	NNS
Wish	NN

4.2.2 Feature Extraction:

After preprocessing steps, feature extraction method is applied on the final generated list of words.

Words along with its features are shown in table 4.3. This represents the resulted values of each feature after calculations.1 is for difficult and 0 is for easy whereas distinct 1,2,3,4 in part of speech column represent four categories. All other features are binomial in nature and converted to this form during their respective calculation.

Table 4-3: Extracted Features

words	no_of_char	syllable_count	presence_of_ch,sh,th,st,f	part_of_speech	Pronounce_c_k	pronounce_g_j
Adequate	1	1	0	3	0	0
Perilous	1	1	0	3	0	0
Delicate	1	1	0	3	1	0
Calamity	1	1	0	1	1	0
Care	0	0	0	1	1	0
Seek	0	0	0	2	1	0
Wish	0	0	1	1	0	0

4.2.3 Classification

4.2.3.1 Training and testing Phase

With the help of accord.net and c# language, we have use Decision tree c4.5 model for training and testing of the model. First of all using excels reader, read the training data into the grid view. After importing training data create the C4.5 decision tree and learn the model. As we are using both categorized and discrete features, so it is necessary to use the codification class. This class is used to convert the categorized value into its codeword through mapping technique. Learned tree is then converted into rules and presented graphically on the screen. After training data; test the data using testing module. During testing, output labels of the respective inputs are predicted based on the tree generated during training phase using tree.decide function. In this function, entropy and gain ratio is calculated also and output label is assigned.

4.2.3.2 Evaluation

For evaluation of system model, model is trained by the training data and is then tested for two different datasets. Dataset1 constitutes of 400 words and dataset2 which is an unknown dataset constitute of 50 words. Cross Validation is also done to evaluate the model. Precision, Recall and F-measure for respective test datasets are depicted.

Confusion matrix is used to represent the complete pictorial view of the attained accuracy. N in confusion matrix represents total number of difficult words (correctly identified as difficult or not difficult words) in selected dataset2. Difficult words which are correctly identified as difficult in a text are represented as actual true. Difficult words which are predicted as correctly identified as difficult by our proposed methodology in a text are represented as predicted true. English words which are easy in a text are represented as actual false. English words which are predicted as easy are represented as predicted false.

TP--True positive: Actual True and Predicted as True.

TN-- True negative: Actual False and Predicted as False.

FP--False positive: Actual False and Predicted as True.

FN-- False negative: Actual True and Predicted as False.

4.2.3.2.1 Detection Accuracy:

Detection Accuracy = ((n1+n2)/N) 100 ---------- (4.1)*

n1: Total number of words that are true and detected true.

n2: Total number of words that are false and detected false.

N: total number of words in a document.

4.2.3.2.1.1 For Test Dataset 2

As a result of testing phase, labels to the words as difficult or not difficult are assigned.

Total English Words terms = 50

Number of Difficult Words = 32

Number of Non Difficult Words= 18

By using formula (4.1), detection accuracy is measured as 92%. As the detection accuracy is dependent on the presence of difficult words in a dataset.

n1 = 30

n2=16

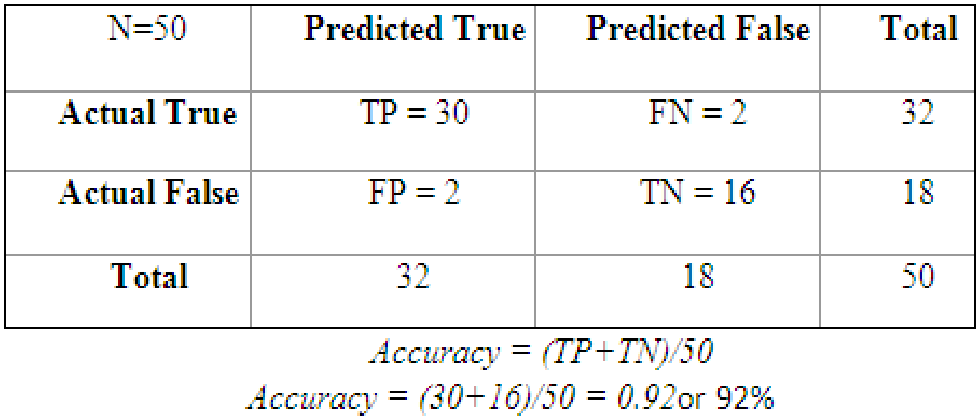
N = 50

Detection Accuracy = ((n1+n2)/N) 100*

Detection Accuracy = ((30+16)/50) 100 = 92%*

To explain the detection accuracy confusion matrix is shown in Table. As mentioned above, total number of words in selected tested dataset2 is 50. Out of 50 words, difficult words are 32 and non-medical terms are 18. Out of 32 difficult words, 30 words are correctly identified by our proposed methodology and 2 words are misidentified as difficult.

Table 4-4: Confusion Matrix



N=50	**Predicted True**	**Predicted False**	**Total**
Actual True	TP = 30	FN = 2	32
Actual False	FP = 2	TN = 16	18
Total	32	18	50

Accuracy = (TP+TN)/50

*Accuracy = (30+16)/50 = 0.92*or 92%

4.2.3.2.1.2 Cross Validation Results of training data

Using Formula Accuracy = (TP+TN)/600---------- (4.2)

*Accuracy = (299+278)/600 = 0.961*or 96.1%

Confusion matrix of training dataset is shown in table 4-5.

Table 4-5: Cross Validation of training dataset

N=600	**Predicted True**	**Predicted False**	**Total**
Actual True	TP =299	FN = 7	306
Actual False	FP = 16	TN = 278	294
Total	315	285	600

5-fold cross validation is also applied on complete dataset1. Dataset composed of 1000 different samples. This dataset is spit into 5 different groups based on 5 folds. Each fold is then tested against remaining folds as training data. Confusion matrix for different folds are depicted in table 4-6,4-7,4-8,4-9,4-10 below.

For Partition 1

Table 4-6: Confusion Matrix Partition 1

N=200	**Predicted True**	**Predicted False**	**Total**
Actual True	TP = 117	FN = 3	120
Actual False	FP = 9	TN = 71	80
Total	126	74	200

Accuracy = (TP+TN)/N---------- (4.2)
*Accuracy = (117+71)/200 = 0.94*or 94%

For Partition 2

Table 4-7: Confusion Matrix Partition 2

N=200	**Predicted True**	**Predicted False**	**Total**
Actual True	TP = 93	FN = 3	96
Actual False	FP = 5	TN = 99	104
Total	98	102	200

Accuracy = (TP+TN)/N---------- (4.2)
Accuracy = (93+99)/200 = 0.96 or 96%

For Partition 3

Table 4-8: Confusion Matrix Partition 3

N=200	**Predicted True**	**Predicted False**	**Total**
Actual True	TP = 89	FN = 0	89
Actual False	FP = 10	TN = 101	111
Total	97	103	200

Accuracy = (TP+TN)/N---------- (4.2)
Accuracy = (89+101)/200 = 0.95 or 95%

For Partition 4

Table 4-9: Confusion Matrix Partition 4

N=200	**Predicted True**	**Predicted False**	**Total**
Actual True	TP = 97	FN = 10	107
Actual False	FP = 2	TN = 91	93
Total	101	99	200

Accuracy = (TP+TN)/N---------- (4.2)
Accuracy = (97+91)/200 = 0.94 or 94%

For Partition 5

Table 4-10: Confusion Matrix Partition 5

N=200	**Predicted True**	**Predicted False**	**Total**
Actual True	TP = 53	FN = 14	67
Actual False	FP =3	TN = 131	133
Total	56	145	200

Accuracy = (TP+TN)/N---------- (4.2)
Accuracy = (53+131)/200 = 0.920 or 92%

On average Cross validation provides an accuracy of 94.2% approximately due to which our proposed methodology outperforms.

4.2.3.2.1.3 For Test Dataset

As a result of testing phase, labels to the words as difficult or not difficult are assigned.

Total English Words terms = 400

Number of Difficult Words = 174

Number of Non Difficult Words= 226

By using formula (4.1), detection accuracy is measured as 92.5%. As the detection accuracy is dependent on the presence of difficult words in a dataset.

n1 = 152

n2=218

N = 400

Detection Accuracy = ((n1+n2)/N) 100*

Detection Accuracy = ((152+218)/400) 100 = 92.5%*

Table 4-11: Confusion Matrix Testing Dataset 1

N=400	**Predicted True**	**Predicted False**	**Total**
Actual True	TP = 152	FN = 22	174
Actual False	FP = 8	TN = 218	226
Total	160	240	400

4.3 Comparison with Existing Techniques:

Research paper [60] developed system to identify difficult words from the text. This paper is attained rank 1 in CWI Shared Task 2018. In this papers author uses ensemble based approach learning to achieve desired results. These papers focuses on word frequency, length, part of speech, stop words, syllable count, hypernym, and hyponym and synset size as features of the words. They have developed systems keeping in view a general public mostly for non native speakers. We considered this paper as state of the art work. Authors use's the accuracy as performance evaluator.

In our proposed methodology, we have use and focused linguistic rules to extract different features. These linguistic rules depict relevant and specific information about why words are difficult for Hearing impaired children. Table 4-11 depicts that when compare to the existing technique our proposed methodology outperformed.

Table 4-12: Comparison with existing techniques

Methodology	Technique used	Target Domain	Accuracy
[60]	Ensemble Voting (combination of Adaboost and random forest).	Non Native speaker	87.36%
Our proposed Methodology	C4.5 Decision Tree Learning	Hearing Impaired Children	92.5% using dataset

4.4 Summary:

In this chapter, results of proposed methodology are discussed in detail. Results of pre-processing, feature extraction and classification are shown separately. Accuracy is calculated to evaluate the identification of difficult words in a text document. 94% aggregate accuracy is achieved according to the results which are generated for evaluation as a result of 5 folds cross validation. On testing dataset, achieves 92.5% accuracy. For rest of the cases, words are not identified by our proposed methodology. This is due to the fact that these may not exist in our training data and computed as an error.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

After discussion of related work, proposed methodology, and results of proposed methodology in detail, we are able to conclude our research work. In this chapter, overall research done in this thesis is concluded and future work is discussed.

In this research work, we dedicated ourselves to propose and implement a technique to identify difficult words from English text for assisting hearing impaired children in learning language. The aim of our proposed technique is to identify difficult words from the text documents available online or in English textbooks in the form of sentences, paragraphs or short stories. In Pakistan, Educational institutes specific for hearing impaired children, adults, children, parents and Speech Language Pathologists (SLP's) can effectively make use of the proposed methodology in daily routine to identify difficult text from the study materials, newspapers, online web stores and can prepare activities during speech therapy. It will save time and resources, currently utilized manually. It will also enhance child's vocabulary as they can learn simple words rather than focusing and spending hours on difficult ones. In short, it will help child to learn and attain knowledge of English language in a simpler way .It will also possess positive impact on their reading, writing and learning skills in a broader aspect. Proposed methodology directly benefits the children with hearing impairment and their parents and indirectly has a definite impact on society.

This research describes the systematic way of developing a proposed methodology. It describes each step in detail involved in implementation. It also describes how the natural language processing, feature extraction and identification of difficult words through classification is done to achieve the required output. The aim of our research was to implement a methodology/technique for hearing impaired children in educational domain.

Main features of our technique were pre-processing of English text, feature extraction based on English linguistic rules specific to hearing impaired children through regular expressions and classification implication using C4.5 decision tree machine learning algorithm. English text document is unstructured data available online or in text documents, which need to be processed before feature extraction and classification process of English words. Natural language processing was used to process the text written in natural language. Tokenization, stop word removal and Part of speech tagging are the techniques of natural language processing which were required in our research. Sharpnlp version of open NLP in c# language was used in implementation of preprocessing step, for natural language processing of written text. Features were extracted which are purely based on English linguistic rules specified for hearing impaired children. For this, we used regular expression method to calculate and extract the features. After that, machine learning classification approach is used. C4.5 decision tree is used for training and testing model.

An efficient technique is achieved with significant increase in accuracy of difficult word identification. 92.5% accuracy is attained using annotated dataset which constitutes of 400 words as tested data used during testing phase.

5.2 Future work

Further research needs to find more elaborated techniques, methods, architectures and its appropriate training algorithms. Decision tree is considered as highly effective and efficient because they can compute with a high numbers of features. There are two main avenues for the extension of this research: application will be performed on more datasets, and other techniques will be applied for the purpose of feature extraction and classification.

5.2.1 Enhanced feature sets

Currently, we have used features based on linguistic rules specific to hearing children. In future, enhancement to feature set can also be applied to enhance the overall performance of the technique. Noun, adjective phrases, 2 or 3 gram words, and total number of synonyms can be used as features to identify difficult words for hearing impaired children.

5.2.2 Improved NLP

English text is available in the form of free-text or unstructured formats. Before implementation of feature extraction and classification steps it is necessary to process unstructured data. Natural language processing tasks can be improved in future work for better results.

5.2.3 More Datasets

Currently, these techniques were only applied on two datasets. In order to show generalizability, it is desirable to replicate this research on further datasets with different properties and features. The dataset in this thesis involved the classification of difficult words for hearing impaired children. This work could be extended to datasets of any sort of language impairment patients facing language learning problems.

5.2.4 Classification technique

Other classification techniques can also be applied and tested for difficult word identification. Currently, binary classification approached was used. Words were classified as either difficult or easy. In future, multi classification approach can be also be used to classify words in levels such as highly difficult, difficult, moderate, or easy. Ensemble approach can also be applied to further improve the accuracy.

5.2.5 Online Application

Knowledge, information or data is of no use if it is not shared in an efficient and quality way. Sharing and exchange of knowledge helps in the advancement of any field. As the future work, it is proposed that develop an online application for identification of difficult words for hearing impaired children, shift the desktop application into web application so users will be able to use it easily without installing the desktop application into their own system.

Appendix A

Coding

Pre-Processing

```
      private void btnPOSTag_Click_1(object sender, EventArgs e)
      {
         var output = new StringBuilder();

         string[] all_sentences = Sentences_split(text_Inn.Text);

         foreach (string sent in all_sentences)
         {
            string[] generate_tokens = Tokenize_Sentence(sent.ToLower());

            string[] a = generate_tokens;
            string[] tags = PosTagTokens(a);

            for (int currentTag = 0; currentTag < tags.Length; currentTag++)
            {if (tags[currentTag]=="NN"|| tags[currentTag]=="NNS"|| tags[currentTag]
=="NNP"|| tags[currentTag] =="NNPS")
               {
                  tags[currentTag] = "1";
}
            else if (tags[currentTag] == "VB" || tags[currentTag] == "VBD" || tags[currentTag]
== "VBG" || tags[currentTag] == "VBN" || tags[currentTag] == "VBP" || tags[currentTag] ==
"VBZ")
               {
                  tags[currentTag] = "2";
               }
            else if (tags[currentTag] == "IN" || tags[currentTag] == "JJ" || tags[currentTag] ==
"JJR" || tags[currentTag] == "JJS")
               {
                  tags[currentTag] = "3";
               }
               else if (tags[currentTag] == "RB" || tags[currentTag] == "RBR" ||
tags[currentTag] == "WRB" || tags[currentTag] == "RBS")
               {
                  tags[currentTag] = "4";
               }
output.Append(a[currentTag]).Append("/").Append(tags[currentTag]).Append("\r\n");
            }  }
         txt_Out.Text = output.ToString();
         string distinct = txt_Out.Text.ToLower();
```

```
        string unique = string.Join(Environment.NewLine, distinct.ToLower().Split(' ', '.',
',','\n').Distinct());
        pos.Text=unique.ToString();
        var Final_Text = new StringBuilder();
        string[] stopWords = File.ReadAllLines(@"C:\NLTK's list of english stopwords.txt");
        stopWords = stopWords.OrderByDescending(w => w.Length).ToArray();
        string inputText = string.Join(Environment.NewLine, pos.Text.ToLower().Split('\n'));
        string resulted_text = Regex.Replace(inputText.ToLower().ToString(), "[^A-Za-z ]", "
");
        var result = string.Join(Environment.NewLine, inputText.ToLower().Split('
').Distinct());
        string outputText = Regex.Replace(result, "\\b" + string.Join("\\b|\\b", stopWords) +
"\\b", " ", RegexOptions.IgnoreCase);
        string Text = Regex.Replace(outputText, @"\s+", "\r\n");
        Final_Text.Append(string.Join(Environment.NewLine, Text)).Append("\r\n\r\n");
        txt_Out.Text = Final_Text.ToString();
        char searchFor = '/';
        var lines = txt_Out.Text.Split(new string[] { System.Environment.NewLine },
StringSplitOptions.None).ToList();
        for (int i = lines.Count - 1; i >= 0; i--)
        {
           char c = lines[i].FirstOrDefault();
           if (c == searchFor || c == '.' || c == ')' || c == '(' || c == ','||c==' ') { lines.RemoveAt(i);
}}
        string textCleaned = string.Join(System.Environment.NewLine, lines);
        string new_lines = Regex.Replace(textCleaned, @"\s+", "\r\n");
        pos.Text = new_lines; }
```

Feature Extraction

```
      private void Feature_extraction_Click(object sender, EventArgs e)
      {StringBuilder output = new StringBuilder();
        StringBuilder pos1 = new StringBuilder();
        StringBuilder feature = new StringBuilder();
        text_Inn.Text = Regex.Replace(text_Inn.Text, "[^A-Za-z ]", " ");
         string sent = text_Inn.Text.ToLower();
       var lines = pos.Text.Split(new string[] { System.Environment.NewLine },
StringSplitOptions.None).ToList();
         for (int i = readlines.Count - 1; i >= 0; i--)
        { int input = readlines[i].IndexOf("/");
           int count = 0;
           if (input >= 0)
           {
            readlines[i] = readlines[i].Substring(count, input);
           }
           else
              readlines[i] =string.Empty;
       }
```

```
        var endline = pos.Text.Split(new string[] { System.Environment.NewLine },
StringSplitOptions.None).ToList();
        for (int i = endline.Count - 1; i >= 0; i--)
        {
]           int input = endline[i].IndexOf("/");
int count =(endline[i].Length)-1;
          if (input >= 0)
          {endline[i] = endline[i].Substring(endline[i].LastIndexOf('/') + 1);      }
          else
            endline[i] = string.Empty; }
        string cc = string.Join(System.Environment.NewLine, endline);
        string Cleaned = Regex.Replace(cc, @"^\s*$\n|\r", string.Empty,
RegexOptions.Multiline).TrimEnd();
        text_Inn.Text = Cleaned;
        string[] po = text_Inn.Text.Split('\n');
        string v= string.Join(System.Environment.NewLine, lines);
        string textCleaned = Regex.Replace(v, @"^\s*$\n|\r", string.Empty,
RegexOptions.Multiline).TrimEnd();
        txt_Out.Text = textCleaned;
 string[] tokens = txt_Out.Text.Split('\n');
         string[] a = tokens;
          string[] b = pos.Text.Split('\n');
        int charac_len ;
        for (int currentTag = 0; currentTag < a.Length; currentTag++)
        {  var regex = new Regex(string.Format(@"\b{0}\b", Regex.Escape(a[currentTag])),
                RegexOptions.IgnoreCase);
          int ba = regex.Matches(sent).Count;
          if (ba >= 4) { ba = 0; }
          else
          { ba = 1; }
          string outputText = Regex.Replace(a[currentTag], @"\s+", "");
          if (outputText.Trim().Length >= 7)
          {
            charac_len = 1;
          }
          else
          {charac_len=0;   }
          string pos = Regex.Replace(b[currentTag], @"\s+", "");
          pos1.Append(string.Join(Environment.NewLine,
a[currentTag])).Append("/").Append(a[currentTag]).Append("\n ");
          int match_c_k, match_g_j, match_ch_f;//= new Regex(tokens[currentTag],
RegexOptions.Compiled | RegexOptions.IgnoreCase).Matches("[CK]+").Count;
          if ((a[currentTag].ToString().Contains("C")) ||
(a[currentTag].ToString().Contains("c")) || (a[currentTag].ToString().Contains("k")) ||
(a[currentTag].ToString().Contains("K")))//|| tokens[currentTag].ToString().EndsWith("ed")))
&& !tokens[currentTag].ToString().EndsWith("le"))
```

```
            {match_c_k = 1; }
            else
            {match_c_k = 0;}
            if ((a[currentTag].ToString().Contains("g")) ||
(a[currentTag].ToString().Contains("G")) || (a[currentTag].ToString().Contains("j")) ||
(a[currentTag].ToString().Contains("J")))//|| tokens[currentTag].ToString().EndsWith("ed")))
&& !tokens[currentTag].ToString().EndsWith("le"))
            { match_g_j = 1; }
            else
            {  match_g_j = 0;    }
            if ((a[currentTag].ToString().ToLower().Contains("st")) ||
(a[currentTag].ToString().ToLower().Contains("ch")) ||
(a[currentTag].ToString().ToLower().Contains("th")) ||
(a[currentTag].ToString().ToLower().Contains("f")) ||
(a[currentTag].ToString().ToLower().Contains("sh")))//||
tokens[currentTag].ToString().EndsWith("ed"))) &&
!tokens[currentTag].ToString().EndsWith("le"))
            {  match_ch_f = 1;   }
            else
            { match_ch_f = 0; }
            int total = Regex.Matches(a[currentTag], @"[AEIOUYaeiouy]+").Count;
            if ((a[currentTag].ToString().ToLower().Trim().EndsWith("e") ||
(a[currentTag].ToString().ToLower().Trim().EndsWith("es") ||
a[currentTag].ToString().ToLower().Trim().EndsWith("ed"))) &&
!a[currentTag].ToString().ToLower().Trim().EndsWith("le"))
            {total--;  }
             if(total>=3)
            {     total = 1; }
             else
               {          total = 0;      }
feature.Append(pos).Append(",").Append(charac_len).Append(",").Append(total).Append(","
).Append(ba).Append(",").Append(match_ch_f).Append(",").Append(po[currentTag]).Appen
d(",").Append(match_g_j).Append(",").Append(match_c_k).Append(Environment.NewLine);
          }
        feature_txt.Text = feature.ToString();
            TextWriter txt = new StreamWriter(@"C:\Users\malik
computer\source\repos\thesis_final_code\thesis_final_code\Data\text_file.txt");
            txt.Write(feature_txt.Text);
            txt.Close();       }
       private void view_data_Click(object sender, EventArgs e)
       { DataTable dt = new DataTable();
  System.IO.StreamReader file = new System.IO.StreamReader(@"C:\Users\malik
computer\source\repos\thesis_final_code\thesis_final_code\Data\text_file.txt");
          string[] columnnames = { "words", "no_of_char", "syllable_count",
"Frequency_of_occurrence", "presence_of_ch,sh,th,st,f", "part_of_speech", "pronounce_g_j",
"Pronounce_c_k" };
```

```
foreach (string c in columnnames)
         {
            dt.Columns.Add(c);
         }
         string newline;
         while ((newline = file.ReadLine()) != null)
         {
            DataRow dr = dt.NewRow();
            string[] values = newline.Split(',');
            for (int i = 0; i < values.Length; i++)
            {
               dr[i] = values[i];
            }
            dt.Rows.Add(dr);
         }
         file.Close();
         f1.Show();
}}}
```

Classification and its evaluation

```
namespace tree
{
   public partial class Decision_tree : Form
   {
      public Decision_tree(object v)
      {
         InitializeComponent();
         openFileDialog.InitialDirectory = Path.Combine(Application.StartupPath,
"Data_files");
dgvtesting.DataSource = v;
      }

      DecisionTree tree;
      C45Learning xx = new C45Learning();
      Func<double[], int> func;
      Codification categorized = null;
      string[] columnNames;
  private void button1_Click(object sender, EventArgs e)
      {
         if (openFileDialog.ShowDialog(this) == DialogResult.OK)
         {
            string filename = openFileDialog.FileName;
            string extension = Path.GetExtension(filename);
            if (extension == ".xls" || extension == ".xlsx")
            {
               ExcelReader db = new ExcelReader(filename, true, false);
               TableSelectDialog t = new TableSelectDialog(db.GetWorksheetList());
```

```
                    if (t.ShowDialog(this) == DialogResult.OK)
                    {
                        DataTable table = db.GetWorksheet(t.Selection);
                       categorized = new Codification(table);
                        DataTable symbols = categorized.Apply(table);
                        double[,] sourceMatrix = symbols.ToMatrix();
                        if (sourceMatrix.GetLength(1) == 8)
                        {
                            MessageBox.Show("Missing class column.");
                        }
                        else
                        {
                            this.dgv.DataSource = table;          }}}}}
        private void test_btn_Click(object sender, EventArgs e)
        {
            if (tree == null || dgvtesting.DataSource == null)
            {
                MessageBox.Show("Please create a machine first.");
                return;
            }
            DataTable sourcetable = dgvtesting.DataSource as DataTable;

   DataTable symbolstest = categorized.Apply(sourcetable);
            double[,] Matrix = symbolstest.ToMatrix(out columnNames);
            double[][] inputs = Matrix.GetColumns(1, 2, 3, 4, 5, 6, 7).ToJagged();

  int[] expected = Matrix.GetColumn(8).ToInt32();

            int[] actual = new int[inputs.Length];
            for (int i = 0; i < inputs.Length; i++)
            {
                actual[i] = tree.Decide(inputs[i]);
                output_column_test.Text += "\n" + actual[i];
            }


ConfusionMatrix confusionMatrix = new ConfusionMatrix(actual, expected, 1, 0);
            dgvperformance.DataSource = new[] { confusionMatrix };

            string answer = actual.ToString();
        }
        double DefaultMissingValueReplacement;

        private void crossvalidation_Click(object sender, EventArgs e)
        {
            DataTable sourcetable = dgv.DataSource as DataTable;
            var categorized = new Codification(sourcetable);
            {
            DefaultMissingValueReplacement = Double.NaN;
             };
            categorized.Learn(sourcetable);
            DataTable symbolstest = categorized.Apply(sourcetable);
            double[,] Matrix = symbolstest.ToMatrix(out columnNames);
            double[][] input = Matrix.GetColumns(1, 2, 3, 4, 5, 6, 7).ToJagged();
int[] output = Matrix.GetColumn(8).ToInt32();
var cv = CrossValidation.Create(

       k: 5,       learner: (p) => new C45Learning(tree),
loss: (actual, expected, p) =>
{
    var cm = new GeneralConfusionMatrix(expected, actual);
    p.Tag = cm;               textBox5.Text = p.Tag.ToString();
    return cm.Accuracy;

},
      fit: (tree, x, y, w) => tree.Learn(x, y, w),x: input, y: output   );

            var result = cv.Learn(input, output);
            textBox5.Text = result.ToString();
                     GeneralConfusionMatrix gcm = result.ToConfusionMatrix(input, output);
            dgvperformance.DataSource = new[] { gcm};
            double accuracy = gcm.Accuracy;
        }

        private void test_data_entry_Click(object sender, EventArgs e)
        {
            if (openFileDialog.ShowDialog(this) == DialogResult.OK)
            {
                string filename = openFileDialog.FileName;
                string extension = Path.GetExtension(filename);
                if (extension == ".xls" || extension == ".xlsx")
                {
                    ExcelReader db = new ExcelReader(filename, true, false);
                    TableSelectDialog t = new TableSelectDialog(db.GetWorksheetList());

                    if (t.ShowDialog(this) == DialogResult.OK)
                    {
                        DataTable table = db.GetWorksheet(t.Selection);


  Codification categorized = new Codification(table);
}


                DataTable symbols = categorized.Apply(table);
                double[,] sourceMatrix = symbols.ToMatrix();


              }          }        }     }
private void create_btn_Click(object sender, EventArgs e)
     {
        if (dgv.DataSource == null)
        {
           MessageBox.Show("Please load some data first.");
           return;
        }
        dgv.EndEdit();
        DataTable table = dgv.DataSource as DataTable;
        DataTable symbols = categorized.Apply(table);
        double[,] sourceMatrix = symbols.ToMatrix(out columnNames);
        double[][] inputs = sourceMatrix.GetColumns(1, 2, 3, 4, 5, 6, 7).ToJagged();
        int[] outputs = sourceMatrix.GetColumn(8).ToInt32();
        dgv.DataSource = symbols;
          DecisionVariable[] variables =
             {
new DecisionVariable("no_of_char", DecisionVariableKind.Discrete),
                new DecisionVariable("syllable_count",DecisionVariableKind.Discrete),
new DecisionVariable("Frequency_of_occurrence",DecisionVariableKind.Discrete),
               new DecisionVariable("presence_of_ch,sh,th,st,f",
DecisionVariableKind.Discrete),//codebook["presence_of_ch,sh,th,st,f"].NumberOfSymbols),
               new DecisionVariable("part_of_speech",5),                  new
DecisionVariable("pronounce_g_j",DecisionVariableKind.Discrete),
   new DecisionVariable("Pronounce_c_k",DecisionVariableKind.Discrete),


             };
  int classCount =
          tree = new DecisionTree(variables, classCount);
               C45Learning id3 = new C45Learning(tree);
        tree = id3.Learn(inputs, outputs);
          decisionTreeView1.TreeSource= tree;
   int[] predicted = tree.Decide(inputs);
 DecisionSet rules = tree.ToRules();
        int y = tree.Decide(inputs[25]);
        textBox1.Text = y.ToString();
 string ruleText = rules.ToString();
        richTextBox1.Text = ruleText.ToString();
int[] query = { 1, 0, 1, 0, 3, 0, 1 };
        int predicted11 = tree.Decide(query);  // result will be 0
        string answer = predicted11.ToString();
}
```

```
      private void generate_labels_Click(object sender, EventArgs e)
          {

         if (tree == null || dgvtesting.DataSource == null)
         {
            MessageBox.Show("Please create a machine first.");
            return;
         }
         dgvtesting.EndEdit();

        DataTable sourcetable = dgvtesting.DataSource as DataTable;
         sourcetable.Locale = CultureInfo.InvariantCulture;
    categorized.Learn(sourcetable);
         DataTable symbolstest = categorized.Apply(sourcetable);
         double[,] Matrix = symbolstest.ToMatrix(out columnNames);
         double[][] inputs = Matrix.GetColumns(1, 2, 3, 4, 5, 6, 7).ToJagged();
         int[] actual = new int[inputs.Length];
         for (int i = 0; i < inputs.Length; i++)
         {
            actual[i] = tree.Decide(inputs[i]);
            output_column_test.Text += "\n" + actual[i];
            sourcetable.Rows[i]["output"] = actual[i];
         }
      }
      private void Get_Synonym_Click(object sender, EventArgs e)
      {
         dgvtesting.AllowUserToAddRows = false;
         if (tree == null || dgvtesting.DataSource == null)
         {
            MessageBox.Show("Please create a machine first.");
            return;
         }
         DataTable sourcetable = dgvtesting.DataSource as DataTable;
         var categorized = new Codification(sourcetable);
categorized.Learn(sourcetable);
         DataTable symbolstest = categorized.Apply(sourcetable);
         double[,] Matrix = symbolstest.ToMatrix(out columnNames);
         double[][] inputs = Matrix.GetColumns(1, 2, 3, 4, 5, 6, 7).ToJagged();
 (sourcetable).DefaultView.RowFilter =
                  string.Format("output LIKE '{0}%' OR output LIKE '% {0}%'", 1);
         string[,] value = new string[dgvtesting.Rows.Count, dgvtesting.Columns.Count];
         foreach (DataGridViewRow row in dgvtesting.Rows)
         {
            foreach (DataGridViewColumn col in dgvtesting.Columns)
            {
               string a = dgvtesting.Rows[row.Index].Cells[8].Value.ToString();
```

```
                    if (a == "1")
                    {
                        value[row.Index, col.Index] =
dgvtesting.Rows[row.Index].Cells[col.Index].Value.ToString();
                    }
                }        }
            int h = 0;
            string strval = "";
            foreach (string ss in value)
            {
                strval += ss + ",";
                if (h == 8)
                {
                    listBox1.Items.Add(strval.TrimEnd(','));
                    strval = "\n";
                    h = -1;
                }
                h++;
            }
            Academia form_acad = new Academia(this.listBox1);
            form_acad.Show();       }
```

Finding synonym

```
private void Get_synonymns_Click(object sender, EventArgs e)
        {
            difficult_words = diff_words.GetItemText(diff_words.SelectedItem);

            foreach (var val in GetSynonyms(difficult_words))
                {
synonym_list.Items.Add(difficult_words + " : " + val);
         }
        }

        public IEnumerable<string> GetSynonyms(string term)
        {
            var appWord = new Microsoft.Office.Interop.Word.Application();
            object objLanguage = Microsoft.Office.Interop.Word.WdLanguageID.wdEnglishUS;
            Microsoft.Office.Interop.Word.SynonymInfo si = appWord.get_SynonymInfo(term,
ref (objLanguage));
            foreach (var meaning in (si.MeaningList as Array))
            {
                yield return meaning.ToString();  }
            appWord.Quit(); //include this to ensure the related process (winword.exe) is correctly
closed.
            System.Runtime.InteropServices.Marshal.ReleaseComObject(appWord);
            objLanguage = null;
```

```
      appWord = null;      }
    private void Academia_Load(object sender, EventArgs e)
    {
      foreach (string item in difficult_word_list.Items)
        textBox1.Text += item.ToString().TrimStart('/');
       var lines = textBox1.Text.Split('\n').ToList();//new string[] {
System.Environment.NewLine }, StringSplitOptions.None).ToList();
      for (int i = lines.Count - 1; i >= 0; i--)
      {
        string c = string.Empty;
        int input = lines[i].IndexOf("/");
        int count = 0;
        if (input >= 0)
        {
          lines[i] = lines[i].Substring(count, input);
        }
        else
          lines[i] = string.Empty;
      }
      string v = string.Join(System.Environment.NewLine, lines);
      string textCleaned = Regex.Replace(v, @"^\s*$\n|\r", string.Empty,
RegexOptions.Multiline).TrimEnd();

      textBox1.Text = textCleaned;
      string[] strval = textBox1.Text.Split('\n');
      foreach (string wor in strval)
      {
        diff_words.Items.Add(wor);
      }  }

    private void diff_words_SelectedIndexChanged(object sender, EventArgs e)
    {
      difficult_words = diff_words.GetItemText(diff_words.SelectedItem) ;

    }
  }
```

References

[1] Knight, Kevin. Human Language Technology: What Machines Do With Text and Speech. (2013).

[2] Muir, Bonnie M. "Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems." *Ergonomics* 37.11 (1994): 1905-1922.

[3] Basma, Sarah, et al. "Error rates in breast imaging reports: comparison of automatic speech recognition and dictation transcription." American Journal of Roentgenology 197.4 (2011): 923-927

[4] Jones, Karen Sparck. "Natural language processing: a historical review." *Current issues in computational linguistics: in honour of Don Walker*. Springer Netherlands, 1994. 3-16.

[5] J.Carmichael , Using Articulatory Visualization Techniques to Improve Pronunication, Current Trends in Information Technology (CTIT), 2013 International Conference on,2013 ,54 – 59 pages

[6] M.Cagatay, P. Ege , G. Tokdemir, N.E. Cagiltay, A Serious Game for Speech Disorder Children Therapy, Health Informatics and Bioinformatics (HIBIT), 2012 7th International Symposium on,2012, 18 – 23 pages

[7] ”Speech and Language Therapy Apps”,https://www.virtualspeechcenter.com/MobileApps.aspx , [last date retrieved: 17th January, 2018]

[8] X. Liu, N. Yan, L. Wang , X. Wu, An Interactive Speech Training System With Virtual reality Articulation For Mandarin-Speaking Hearing Impaired Children,Information and Automation (ICIA), 2013 IEEE International Conference on,2013,191 – 196 pages

[9] Husain, Noushad. (2015). Language and Language Skills.

[10] “Delayed speech or language delayed”, http://kidshealth. org/parent/emotions/behavior/not _talk.html?tracking=PRelatedArticle# ,[date last retrieved: 21st June, 2015].

[11] McGregor, K., Licandro, U., Arenas, R., Eden, N., Stiles, D., Bean, A., & Walker, E. (2013). Why words are hard for adults with developmental language impairments. Journal of Speech, Language, and Hearing Research, 56(6), 1845-1856. https://doi.org/10.1044/1092-4388(2013/12-0233)

[12] Berent, Gerald. (2019). English for Deaf Students: Assessing and Addressing Learners' Grammar Development.

[13] “Deafness and hearing loss”, http://www.who.int/mediacentre/factsh eets/fs300/en/, [last retrieved: 21st June, 2015]

[14] “Types of Hearing Loss:”,http://www.asha.org/public/hearing/Mixed-Hearing-Loss/, [last retrieved: 21st June, 2015]

[15] “What is deafness? What is hearing loss?” , http://www.medicalnews today.com / articles/249285.php, [last date retrieved: 21st June, 2015]

[16] “Cochlear Implants”, http://www.hearingloss.org/content/cochlear-implants, [last date retrieved: 21st June, 2015]

[17] Peek, Niels, et al. "Thirty years of artificial intelligence in medicine (AIME) conferences: A review of research themes." *Artificial intelligence in medicine* 65.1 (2015): 61-73.

[18] Morehouse, K. (2019). What Are the Four Language Skills? • LinguaCore. Retrieved 3 August 2018, from https://www.linguacore.com/blog/the-four-skills/

[19] ALQAHTANI, MOFAREH. (2015). The importance of vocabulary in language learning and how to be taught. International Journal of Teaching and Education. III. 21-34. 10.20472/TE.2015.3.3.002.

[20] John L. Luckner, & Christine Cooke. (2010). A Summary of the Vocabulary Research With Students Who Are Deaf or Hard of Hearing. American Annals Of The Deaf, 155(1), 38-67. doi: 10.1353/aad.0.0129

[21] Musselman, C & Kircaali Iftar, Gonul. (1996). The Development of Spoken Language in Deaf Children: Explaining the Unexplained Variance. Journal of deaf studies and deaf education. 1. 108-21. 10.1093/oxfordjournals.deafed.a014285.

[22] Gallion, Tammy. (2016). Improving Vocabulary Comprehension for Deaf or Hard of Hearing Students. Theses, Dissertations and Capstones. Paper 989

[23] Coppens, Karien & Tellings, Agnes & Verhoeven, Ludo & Schreuder, Robert. (2011). Depth of reading vocabulary in hearing and hearing-impaired children. Reading and writing. 24. 463-477. 10.1007/s11145-010-9237-z.

[24] Flanagan, Brendan & Hirokawa, Sachio. (2015). Correlation between an Entropy Based Measure and English Language Learner Proficiency. 349-353. 10.1109/IIAI-AAI.2015.288.

[25] Charlesworth, Ann & Charlesworth, Robert & Bridie, Raban & Rickards, Field. (2006). Teaching children with hearing loss in reading recovery.

[26] Charlesworth, Ann & Charlesworth, Robert & Bridie, Raban & Rickards, Field. (2006). Teaching children with hearing loss in reading recovery.

[27] Adams, V. (1973). An introduction to modern English word formation. London: Longman.

[28] Leybaert, Jacqueline & Content, Alain & Alegria, Jésus. (2008). The development of written word processing: the case of deaf children The development of written word processing: the case of deaf children. Ilha do Desterro.

[29] Teymouri, Robab & Daneshmandan, Naeimeh & Hemmati, Sahel & Soleimani, Farin. (2014). Perception Development of Complex Syntactic Construction in Children with Hearing Impairment. Iranian Rehabilitation Journal. 12.

[30] Vincela, Zigrida. (2016). Complex Sentences and their Punctuation in English Texts Composed by Latvian Students. Žmogus ir žodis. 18. 96-105. 10.15823/zz.2016.5.

[31] Laufer, Batia. (1990). Why are Some Words More Difficult than Others? Some Intralexical Factors that Affect the Learning of Words. Iral-international Review of Applied Linguistics in Language Teaching - IRAL-INT REV APPL LINGUIST. 28. 293-308. 10.1515/iral.1990.28.4.293.

[32] Manning, Christopher D., and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[33] Priyanka, and Rashmi Phalnikar. "Generating UML Diagrams from natural language specifications." *International Journal of Applied Information Systems, Foundation of Computer Science* 1.8 (2012).

[34] Liddy, Elizabeth D., and Jennifer H. Liddy. "An NLP approach for improving access to statistical information for the masses." (2001)

[35] Kaiya, Haruhiko, and Motoshi Saeki. "Ontology based requirements analysis: lightweight semantic processing approach." *Quality Software, 2005.(QSIC 2005). Fifth international conference on*. IEEE, 2005.

[36] Deeptimahanti, Deva Kumar, and Muhammad Ali Babar. "An automated tool for generating UML models from natural language requirements." *Proceedings of the 2009 IEEE/ACM International Conference on Automated Software Engineering*. IEEE Computer Society, 2009.

[37] Karlsson, Tobias. "Managing large amounts of natural language requirements through natural language processing and information retrieval support." *Master's Thesis, Department of Communication Systems, Lund Institute of Technology* (2004).

[38] Alksasbeh, Malek Zakarya, et al. "AN AUTOMATED USE CASE DIAGRAMS GENERATOR FROM NATURAL LANGUAGE REQUIREMENTS." *Journal of Theoretical and Applied Information Technology* 95.5 (2017): 1182.

[39] Geetha, S., and G. A. Mala. "Automatic Relational Schema Extraction from Natural Language Requirements Specification Text." *Middle-East Journal of Scientific Research* 21.3 (2014): 525-532.

[40] Btoush, Eman S., and Mustafa M. Hammad. "Generating ER diagrams from requirement specifications based on natural language processing." *International Journal of Database Theory and Application* 8.2 (2015): 61-70

[41] Ferrari, Alessio, Giorgio Oronzo Spagnolo, and Stefania Gnesi. "Towards a Dataset for Natural Language Requirements Processing." *REFSQ Workshops*. 2017.

[42] Manning, Christopher D., and Hinrich Schütze. "Foundations of statistical natural language processing, Vol. 999." (1999).

[43] Trivedi, Gaurav, et al. "An Interactive Tool for Natural Language Processing on Clinical Text." *arXiv preprint arXiv:1707.01890* (2017).

[44] Wattenberg, Martin, and Fernanda B. Viégas. "The word tree, an interactive visual concordance." *IEEE transactions on visualization and computer graphics* 14.6 (2008).

[45] Wei, Furu, et al. "Tiara: a visual exploratory text analytic system." *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010.

[46] Yalla, Prasanth, and Nakul Sharma. "Integrating natural language processing and software engineering." *International Journal of Software Engineering and Its Applications* 9.11 (2015): 127-136.

[47] Pressman, Roger S. *Software engineering: a practitioner's approach*. Palgrave Macmillan, 2005.

[48] Leidner, Jochen L. "Current issues in software engineering for natural language processing." Proceedings of the HLT-NAACL 2003 workshop on Software engineering and architecture of language technology systems-Volume 8. Association for Computational Linguistics, 2003.

[49] Tripathi, S., & Sharma, T.G. (2015). A Survey Paper For Finding Frequent Pattern In Text Mining. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET),Volume 4 Issue 3,993-997

[50] Specia, L. (2016). Unsupervised Lexical Simplification for non-native speakers. AAAI'16 Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, pp.3761-3767.

[51] Carroll, John & Minnen, Guido & Canning, Yvonne & Devlin, Siobhan & Tait, John. (1998). Practical Simplification of English Newspaper Text to Assist Aphasic Readers. Proc. of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology.

[52] Shardlow, M. (2013). A Comparison of Techniques to Automatically Identify Complex Words. In *Proceedings of the Student Research Workshop at the 51st Annual Meeting of the Association for Computational Linguistics*Association for Computational Linguistics.

[53] Davoodi, Elnaz & Kosseim, Leila & Mongrain, Matthew. (2017). A Context-Aware Approach for the Identification of Complex Words in Natural Language Texts. 97-100. 10.1109/ICSC.2017.9.

[54] sp, sanjay & Kumar, M & Kp, Soman. (2016). AmritaCEN at SemEval-2016 Task 11: Complex Word Identification using Word Embedding. 1022-1027. 10.18653/v1/S16-1159

[55] Lee, John & Yan Yeung, Chak. (2018). Automatic prediction of vocabulary knowledge for learners of Chinese as a foreign language. 1-4. 10.1109/ICNLSP.2018.8374392.

[56] Butnaru, Andrei & Ionescu, Radu Tudor. (2018). UnibucKernel: A kernel-based learning method for complex word identification. 175-183. 10.18653/v1/W18-0519.

[57] Alfter, David & Pilán, Ildikó. (2018). SB@GU at the Complex Word Identification 2018 Shared Task. 315-321. 10.18653/v1/W18-0537.

[58] Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo H Paetzold, Lucia Specia, Sanja Štajner, Ana{\"ıs Tack, and Marcos Zampieri. 2018. “A Report on the Complex Word Identification Shared Task 2018.”

[59] AbuRa'ed, Ahmed & Saggion, Horacio. (2018). LaSTUS/TALN at Complex Word Identification (CWI) 2018 Shared Task. 159-165. 10.18653/v1/W18-0517.

[60] Gooding, Sian & Kochmar, Ekaterina. (2018). CAMB at CWI Shared Task 2018: Complex Word Identification with Ensemble-Based Voting. 184-194. 10.18653/v1/W18-0520.

[61] Kajiwara, Tomoyuki & Komachi, Mamoru. (2018). Complex Word Identification Based on Frequency in a Learner Corpus. 195-199. 10.18653/v1/W18-0521.

[62] Singh, J., Singh, G. and Virk, R. (2017). An Automated Complex Word Identification from Text: A Survey. Oriental journal of computer science and technology, 10(3), pp.612-617.

[63] Mukherjee, Niloy & Patra, Braja & Das, Dipankar & Bandyopadhyay, Sivaji. (2016). JU NLP at SemEval-2016 Task 11: Identifying Complex Words in a Sentence. 10.18653/v1/S16-1152.

[64] Quijada, Maury & Medero, Julie. (2016). HMC at SemEval-2016 Task 11: Identifying Complex Words Using Depth-limited Decision Trees. 1034-1037. 10.18653/v1/S16-1161.

[65] Yimam, S., Stajner, S., Ried, M., & Biemann, C. (2017). Proceedings of the The 8th International Joint Conference on Natural Language Processing, (pp. 401–407). Taipei, Taiwan, November 27 – December 1, 2017 c 2017 AFNLP.

[66] S Silpa, K & Irshad, M. (2018). Lexical Simplification of Complex Scientific Terms. 1-5. 10.1109/ICETIETR.2018.8529069.

[67] Guzin, Karasu & Umit, Girgin & Uzuner, Yildiz & Zehranur, Kaya. (2016). Vocabulary developing strategies applied to individuals with hearing impairments. Educational Research and Reviews. 11. 1402-1414. 10.5897/ERR2016.2835.

[68] Shojaei, E., Jafari, Z., & Gholami, M. (2016). Effect of Early Intervention on Language Development in Hearing-Impaired Children. Iranian journal of otorhinolaryngology, 28(84), 13–21.

[69] Martineau, G., Lamarche, P., Marcoux, S., & Bernard, P. (2001). The Effect of Early Intervention on Academic Achievement of Hearing-Impaired Children. Early Education & Development, 12(2), 275-289. doi: 10.1207/s15566935eed1202_7

[70] Effects of Hearing Loss on Development. (2015). Retrieved 1 September 2018, from https://www.asha.org/uploadedFiles/AIS-Hearing-Loss-Development-Effects.pdf

[71] D. Loaiza, C. Oviedo, A. Castillo , A.Portilla, G.Alvarez,D. Linares, A. Navarro, G.A. Ivarez, A video game prototype for speech rehabilitation,Games and Virtual Worlds for Serious Applications (VS-GAMES), 2013 5th International Conference on,2013, 1 – 4 pages,

[72] S. Kumar, H.k.Sardana, R. Chhabra, R, K. Resmi, IEEE Graphical Speech Training System for Hearing Impaired, Image Information Processing (ICIIP), 2011 International Conference on 2011, 1 – 6 pages.

[73] T. Tan, H. Liboh, A. K. Ariff, C. Ting, S. Salleh, Application of Malay speech technology in Malay Speech Therapy Assistance Tools, Intelligent and Advanced Systems (ICIAS) , 2007 International Conference on, 2007, 330 – 334 pages

[74] Weerasinghe,DSP-based techniques for speech training of hearing impaired children, Communications,Computers and signal Processing(PACRIM), 2001IEEE Pacific Rim Conference on ,2001, 51 - 54 vol.1 pages

[75] Srivastava, Ashok & Sahami, Mehran. (2009). Text mining. Classification, clustering, and applications. Boca Raton. 10.1201/9781420059458.