Chapter 1

Introduction

“You don't write because you want to say something, you write because you have something to say.”
—F. Scott Fitzgerald

1.0 Introduction
Language is one of the finest inventions of mankind. It is the creative tool for communication. Communicating in one’s own language is natural and effortless task. But when it has to be explained logically; it is indeed a great deal of efforts. As we study, we discover an exciting sphere of human knowledge.

1.1 Language
Language can be defined as a system of systems. Linguists, who study language, have formed a pattern for analyzing it. The grammar of any language is a network of interactive structures and categories. The most smallest and lucid unit for communication is a word. Words alone cannot communicate in an effective way. Words together arranged in a systematic manner make a sentence. Sentences are capable of successful and effective communication.

1.1.1
Words have been divided in to certain grammatical categories for the analysis of the language. No two languages of the world are similar, in any feature. The same labels are used for describing grammatical categories in different languages and they have little different significance for each language. These labels are unique and distinct from one other and still they are interwoven. They are better known as Parts of Speech. Almost every language makes a distinction between open and closed part of speech classes. Robins\(^1\) (1964) describes open classes as those “whose membership is in principle unlimited, varying from time to time and between one speaker and another” and closed
classes are those that “contain a fixed and usually small number of member words, which are the same for all the speakers of the language, or the dialect”.

1.2 Parts of Speech
Gujarati is an Indo-Aryan language. It is a verb-final language and has a relatively free word order; it is an inflectional language. Words are formed by successfully adding suffixes to the root word in series. When suffixes are attached to the root several morphophonemic changes take place. It identifies nine parts of speech in its grammar.

1.2.1 Noun
The label ‘noun’ is assigned to the class of words which has the names of most persons, places and things. Certainly it is an open class. It can have subclasses as proper, common, abstract (which are described in detail in second chapter).
Examples- India, Ganga, cow, Meera, book, picture, strength, heat, etc.

1.2.2 Verb
Verb is the name given to the parts of speech class that has most words describing activity, action, process and even state. It gives the idea of the situation; which is being talked about in the sentence. The verb dominates the sentence in Gujarati. It is an open class and does have subclasses like action, process (which are described in detail in second chapter).
Examples- dance, laugh, run etc

bhArata, ga~gA, gAya, chopaDi, shakti, garami

nAchavu~, gAvu~, doDavu~
1.2.3 Adjective
It is a class of words that denote quality or attributes. It is an open class, but it is dependent on other class; specifically noun. They are the words that modify noun. And many a time they denote the state. Example- *the girl is beautiful*. When they qualify some noun, they precede that noun, as for example, *the beautiful girl* ...
Examples- good, tall, beautiful.
सुनदार, लंबू, सरस
sun~dara, lA~bu~, sarasa

1.2.4 Adverb
Unfortunately there is no perfect and universally approved definition of adverb. Noun, verb and adjective are the basic open classes of grammar. Adverb is certainly not different from them, but does not have a specific function or identification. They act as modifiers, specifically to verbs. Example to walk is the verb and in *Mona walks slowly*, slowly is the adverb.
Adverbs modify verbs, adjective, other adverbs and even whole sentences.
Examples- today, slowly (second chapter has a detailed description).
आज, धीमे
*Aje, dhime*

1.2.5 Pronoun
It is a closed class of parts of speech. They are the class of words which, under certain circumstances, are used as substitutes for the words belonging to open classes, or for larger constitutes. Distinctively they are the substitutes for nouns.
Examples- he, she, it, they, this, that, who, some
हूँ, तूँ, तामे, ते, पेलो, फोल, कोई
*hu~, tu~, tame, te, pelo, koNa, koi*
1.2.6 Conjunctions
Conjunctions are the words used to connect words, phrases, or clauses. They have two subclasses coordinating and subordinating. Coordinating conjunctions assign equal rank to the conjoined elements and subordinating ones mark unequal rank to the conjoined. It is a closed class.
Examples- and, or, that etc
अने, अथवा, के
ane, athavA, ke

1.2.7 Post-Position
It is a small closed class. Basically they are nouns, but their function is to join noun with the verb. They are different from case suffixes in a way that they also can take case suffixes like nouns. Case suffixes are written together with the noun whereas post positions are written separately in Gujarati. English has prepositions whereas Gujarati has post-positions, as for example, on is a preposition in English, where as ‘पर’/para/ in Gujarati is a post position.
‘मा’/mA~ in Gujarati is a case suffix and it is written together with the noun, as ‘कप मा~/kap mA~) (in the cup), but ‘पर’/para/ is a post position and is written with a space, as for example, डप पर. /kapa para/ (on the cup)

1.2.8 Particle
It is a tiny closed class. They add some meaning to both nouns as well as verbs. Sometimes they do not denote any particular lexical meaning, but they denote speakers’ intention and approach towards the situation. As for example, ठुं पाNa अविशा. hu~paNa Avisha (I will also come).

1.2.9 Interjections
They are often exclamatory characters, they have no syntactic connection. They are simply utterances that indicate the speaker’s attitude or that solicits the hearer’s attitude toward the event or condition expressed in a sentence.
Examples- huh, ahh, humm, ahem etc
हें, आह, हुम्म, अहें

he~, hAsha, oha, arere

1.3. Sentence
The major parts of speech have been described in detail in the coming chapters.
All this various parts of speech, each with unique function, combine to form a sentence.

1.3.1
A sentence can be divided into two parts. One part discusses about the activity going on and the other part talks about the objects (animate and inanimate) involved in the activity. The activity part is known as predicate and performer part is the subject. This is a semantic division.

1.3.2
This thesis studies the morohosyntantic structure of the language; hence I will be focusing on the syntactic part. Syntactically the subject part is known as Noun Phrase and the predicate part is known as Verb Phrase.
So we can say sentence is made up of Noun Phrase and Verb Phrase.
Sentence = NP + VP

![Sentence diagram](fig. 1.1)
1.3.3
As said earlier, sentence is made up of words; hence noun phrases and verb phrases are made up of words i.e. various parts of speech.

1.3.4
Noun Phrase has nouns, adjectives, pronouns, conjunctions, post-positions, particles and sometime adverbs too.
Verb Phrase mainly consists of verbs and adverbs.
Verb phrase do include NPs, PPs, Adjective Phrases, nouns, pronouns, adjectives and Post positions. All these are all optional. I have not discussed these categories. Verbs and adverbs are exclusively discussed here at great length.
Interjections, is such a category that can appear in noun and verb phrase.

1.4 Natural Language Processing
The division of language into grammatical labels is done for the better understanding. A child does not learn its first language; it is acquired from the surroundings. But when the language is to be learnt or studied, it needs to be dissected. This goes true for humans and computers too. The computer, understanding the human language is the one of the goal of the computing world and the process/research done for it is known as Natural Language Processing (NLP).

1.4.1
Natural Language Processing is a field of computer science, artificial intelligence and linguistics concerned with the interactions between computers and language of human beings. Machine understanding of natural language would obviously offer numerous business opportunities such as truly conversational human-robot interfaces, conversational information systems, automated text understanding and summarization, information search, junk mail filtering and eventually true thinking machines.
1.4.2
NLP analyzes natural language for studying it, understanding it. The understanding of natural language will help machines to produce such software that humans use naturally, so eventually humans will be able to address the computer as though they are addressing another person.

1.4.3
Though it seems that natural language processing is a very attractive method of human–computer interaction, it is highly difficult to understand the ambiguous nature of natural language.

1.4.4
Understanding and analyzing language means, among other things, knowing what concepts a word or phrase stands for and knowing how to link those concepts together in a meaningful way. It's ironic that natural language, the symbol system that is easiest for humans to learn and use, is hardest for a computer to master. Long after machines have proven capable of inverting large matrices with speed and grace, they still fail to master the basics of our spoken and written languages. To acquire such results, natural language has to undergo various processes and steps. The first step of NLP is analyzing the natural language, by analyzing; here I mean is to dissect the language in such a way that a computer can understand it. In the thesis I have analyzed Gujarati language, which is my first language. And specifically the verb phrase. The verb is the king of the sentence and it dominates other components.

1.5 Technique and Resources

1.5.1
Started by studying the current descriptive Gujarati grammars, I referred the grammars of Dr. Prabodh Pandit, Dr. Vyas, Shri Jayant Kothari, Dr. Urmi Desai, and the latest of Dr. Bhandari. To understand the grammar of the language, one needs to study the language. I had started it from the sentence level. I had randomly collected a huge numbers of sentences from government textbooks (secondary and higher-secondary level), children story
books (Mahabharata, Panchtantra etc of renowned publisher), magazines (articles from famous writers Shri Gunvant Shah and Jay Vasavada), and from day to day conversation of the native speakers.

1.5.2
Gujarat is a vast state. There is a dialectal difference between central Gujarat, north Gujarat, south Gujarat and Saurashtra region. The dialectal variation is phonetic and up to a certain extent lexical. The grammar pattern remains the same. The examples taken form the native speakers; belonged to the central region of Gujarat, mainly from the cities of Amdavad and Baroda. My data had lots of variations in the construction pattern.

1.5.3
Each sentence from the collected data was analysed with the tree-diagram method. This method helped me to understand the basic structure, the components and the various possible constructions in Gujarati. The sentences were divided into clauses, then into phrases and then into tokens. Each token was tagged with parts of speech and also with the theta roles. The syntactic as well as the semantic level were examined. The concentration was exclusively shifted on the verb phrases as it is the topic for research. The study of the patterns of verbs, revealed how it represents tense, aspect, mood and voice in the verb phrase and also how the verb agrees with the components of the noun phrase. This process opened hidden layers of language, it was very strenuous though interesting.

1.6 Description of the chapters
The following portion of the chapter gives the little introduction of each chapter and its components.

1.6.1
The basics of grammar; in the form of parts of speech has been introduced in first chapter. Along with it, NLP is also introduced.
1.6.2
The second chapter includes the sentence pattern, NP and VP. NP and its all components have been talked about. It is necessary to talk about NP and its components because; NP represents the situation described by the verb. It is better to have a whole picture of the sentence; this makes the understanding of VP more clear and lucid. Further it has detailed notes on adverbs and auxiliaries. They are the important components of VP. Though they are obligatory, Gujarati language has a tendency of intensive use of auxiliaries. Tense and explicators are its two types. Tense auxiliaries show tense, Mood and Voice; and explicators show aspect and have a semantic shade. Adverbs are considered as a miscellaneous class, world wide, this is true for Gujarati too. In Gujarati; only adverbs of manner are considered as pure adverbs. Other adverbs are formed from other word classes. Adverbs have been examined from all possible perspectives.

1.6.3
The third chapter is fully dedicated to verbs and the verb phrase. It has the categorization of verbs and syntactic structure of the verb phrase. Verbs have been classified at syntactic and semantic level. All the possible constructions of verb phrase have also been discussed.

1.6.4
Tense, aspect, mood and voice are the features represented in the verb phrase. The fourth chapter deals with the attributes attached to the verb. All the basic attributes along with negation, person-number and gender-number have been discussed at length. Their semantic and morphosyntactic view has been represented.

1.6.5
The computational part is summed up in the fifth chapter. It has topics like NLP, natural and machine language, machine translation its approaches. And finally the analyzer for Gujarati Verb Phrase, its technique and the steps followed. And it also has the verb phrase samples from the analyzer in the simple format.
1.6.6
The sixth chapter contains the conclusion, the methodology used to categorize the verbs and the whole process of making the analyzer. It also has the explanation about the appendix.

1.6.7
Along with the thesis a CD of Gujarati Verb Phrase Analyzer has also been provided. It is a sample analyzer, as verb is an open ended category; new verbs keep on flowing in. It is impossible to include all the verbs.
The CD has the setup of the analyzer, it is compatible on Windows. It is highly recommended to read the ‘read me’ file before installing. Pramukh and Google fonts have also been provided for those who can not type Gujarati.
The result can be viewed in two different formats – tabular and regular (simple). The verb phrase has to be kept in parenthesis; the software will analyze only the phrase in parenthesis.

References-