Original Paper

ISSN: 2321-1520

POS Tagging and Parsing of Gujarati Language

Dr. Nilotpala Gandhi

Professor & Head
Department of Linguistics
Gujarat University
Ahmedabad

1.0. Having its place in top 25 amongst 6,700 Languages of the world, GUJARATI has its own importance due to Linguistic and extra-linguistic factors. Language, being a unique communicative system is one of the major causes of the superiority of Homo-Sapiens to all other living beings. It is a product of human brain, proving it as a Super Computer. 'Language' is more complex than any other programming language of Computers.

1.1. Human Societies could exist only because of its unique communicative system, called 'LANGUAGE'. During evolution, man enabled to convert some of his/her organs of respiration and digestion into vocal apparatus and mother NATURE facilitated him by certain physical adaptations during evolution. So, every human child is now born with a potentiality to acquire 'Language'. According to Chomsky, it is LAD, i.e. Language Acquisition Device.

1.2 But, Computers are not facilitated with this LAD. They have Artificial Intelligence. A human child has worldly knowledge and his/her own intelligences. He can assume and presume and draw the rules of his own language, inductively or deductively. It is very easy for him to learn a language but it is very difficult to teach Machines a language.

2.0. Every Language is different. All languages have some universal features and some unique features. Linguists have to prepare different models for every language. Indian languages are extremely rich in Morphology. Models for English or any other European or American languages cannot be suitable for the computation of Indian Languages. We have to look towards its own tradition, which is the oldest and richest as far as linguistic studies are concerned. So, Paninian Model can be considered as the best suitable Model, but some changes are required as per the different languages.

2.1. Paninian grammar is written in the form of Algebraic formula or Aphorisms, which is very suitable for the computation of any language. It matches the goal of NLP of extracting meaning

from an utterance. That is the reason why Bloomfield called it as 'the greatest monument of human intelligence!

3.0. Processing can be done for all the three forms of Languages- Spoken Written and Sign language. But the Written form is considered as a standard form and much work has been done for written form, we will take only the written form for this research paper. So, for the processing of the written form of the natural languages, five steps are required- Corpus Creation, Analysis, Parsing, Chunking and Generation. We will focus on analysis, i.e. POS Tagging and Parsing for this paper.

3.1. POS- Every language differs in number of Parts of Speech. There are three universal POS as Nouns, verbs and Affixes. Other POS can be some of these as – Pronouns, Adjectives, Adverbs, Demonstratives, Pre or Post Positions, Conjuncts, Particles, Interjections, Articles, and Residuals. All languages do not have all of them. Arabic has only three POS, Classical Sanskrit had 4, English has 8 and Gujarati has 11! So, after collecting a huge corpus, the difficult task of tagging begins.

3.2. POS of Gujarati- As mentioned above, Gujarati has 11 POS. They are – Nouns, Pronouns, Adjectives, Verbs, Adverbs, Post Positions, Particles, Conjuncts, Demonstratives, Quantifiers and Residuals. A BIS Tag-set is created for Indian Languages. So, the BIS Tags are also given along with their descriptions.

3.2.1 Nouns- Nouns are one of the universal POSs. Gujarati Nouns are made up of one or more than one Morphemes. They can be simple as well as derived. But, once they are formed they are known as a single unit. When they become a part of any utterance, they have to pass through the process of inflection. They either take case suffixes or take post positions. Unlike Hindi, case suffixes are written together with the Nouns, so they are not considered as a separate POS. Post positions are considered as a separate POS. The BIS Tag of Noun is 'N'.

3.2.1.1 Types of Nouns- There are different types of Nouns in Gujarati Language. For the purpose of computation it can be taken from two approaches. Computer is not able to recognize from meaning perspective, so it cannot be taught as common noun or abstract noun or so, as we teach to a human child. So for the purpose of Tagging in computation, there are three types of nouns, Common Nouns (BIS Tag is N_NN), Proper Nouns (BIS Tag is N_NNP) and NLoc Nouns (BIS Tag is N_NST), as far as their functions are considered. But from Morphological point of view, i.e. to form a Morphological Analyzer, it can be divided into Masculine, Feminine and Neuter Nouns. Gujarati has three genders and the nouns of all the three genders take different types of suffixes. So, according to the structure, they can be divided according to their gender. Even in gender, they have different types according to the suffixes they take - Masculine Nouns are of three types, Feminine nouns are of two types and Neuter Nouns are of three types. Masculine as ending in 'O' (ઓ) like 'rastO' (રસ્તો), 'ChhokrO' (છોકરો) , 'VaDkO' (વાડકો) etc.; ending in 'a' (અ) as 'mANasa' (માણસ), 'wAgha' (વાઘ), 'kAna' (કાન) etc. and ending in any vowel other than 'a' (અ), as 'hAthI' (હાથી), 'rAjA' (રાજા), etc. Feminine nouns as ending in 'a' (અ) as 'jIbha'

(જીભ), 'A~kha' (આંખ), etc, and ending in any other vowel, like 'chhokrI' (છોકરી), 'mAlA' (માલા), 'bhakti' (ભક્તિ), 'bhAnu' (ભાનુ), etc. Neuter nouns as ending in 'u~' (ું) like 'chhokru~' (છોકરું), 'vandru~'(વાંદરું), etc, ending in 'a' (અ) like 'nAka' (નાક), etc, and ending in any other vowel like 'rU' (ૂ), pANI (પાણી) etc.

3.2.2 Pronouns- Gujarati has six types of pronouns. The BIS Tag for pronoun is 'PR' They are Personal Pronouns (BIS Tag is PR_PRP), Reflexive pronouns (PR-PRF), Relative Pronouns (PR_PRL), Reciprocal Pronouns (PR_PRC), Wh word Pronouns (PR_PRQ) and Indefinite Pronouns(PR_PRI). All Pronouns take case suffixes like Nouns. For Morphological Analyzer, they can also be classified according to the suffixes they take.

3.2.3 Verbs- Verbs have to play a pivotal role in any sentence. They are the Universal and most important POS for all languages. Its BIS Tag is 'V'. There is two types of verbs as far as structure is considered, Main (V_VM) and Auxiliary (V_VAUX). Verb Roots can also be classified according to their meaning and structure. Main Verbs can be classified into Finite (V_VM_VF) and Non-finite (V_VM_VNF) verbs. The Gerund form (V_VM_VNG) and the Infinitive forms (V_VM_VINF) have also special Tags. The Auxiliary verbs (V_VAUX) have typical function in Indian Languages. There are two types of Auxiliaries, (1) denoting Tense and Mood and (2) Explicators. For Tagging Main and Auxiliary classification is enough, but for the Morphological Analyzer Gujarati Verbs can be classified into 8 (Eight) groups. These groups are based on the Morphological Operations. They are – 'cAla' Group (ચાલ), 'Arambha' Group (આરંભ જૂથ), 'bes' Group (બેસ જૂથ), 'gA Group' (ગા જૂથ), 'khA' Group (ખા જૂથ), 'jA' Group (જા જૂથ), 'sU' Group (સૂ જૂથ) and 'bI' Group (બી જૂથ).

3.2.4 Adjectives- There are different types of Adjectives as far as meaning is concerned e.g. Adjectives denoting Shape, Size, Color, Qualities, etc. But for tagging they can come under one umbrella. The BIS Tag for Adjectives is 'JJ'. For Analyzer, they take different forms like Variable and Non-variable Adjectives.

3.2.5 Adverbs- There are very few words which can be considered as pure Adverbs. Only those words denoting Manner are Adverbs of Manner. The BIS Tag of Adverb is 'RB'. Other Nouns along with some suffixes function as Adverbs. So they are not taken as Adverbs as far as Tagging is concerned.

3.2.6 Postpositions-There are two types of units which denote Karaka Relations in Gujarati- Case suffixes (વિભક્તિ પ્રત્યયો) and Postpositions (અનુગો કે નામયોગિઓ). Case Suffixes are written together with the words so they are not having separate position as POS. But, certain words which denote Karaka Relations can be considered as a seperate POS. Words like 'par'(૫૨), 'mATe' (માટે), 'thakI' (થકી) etc are Postpositions. The BIS Tag for post-position is 'PSP'.

3.2.7 Conjunctions- There are two types of Conjunctions (BIS Tag is 'CC') in Gujarati- Conjunctions denoting Co-ordination (CC_CCD) and Conjunctions denoting Sub-ordination (CC_CCS). Conjunctions like 'ane' (અને), 'ke'(કે) etc. denote Co-ordination and the Conjunctions like 'tethi' (તેથી), 'evu~' (એવું) denote Sub-ordination.

3.2.8 Particles- Certain words are particles by default. The BIS Tag is 'RP' As for example, pan (પણ), jo (જો), to (તો), etc. are particles by default (RP_RPD). Some Interjections (RP_INJ) like 'he' (હે), 'arre' (અરે) etc are also particles. Certain Intensifiers (RP_INTF) like 'bahu' (બહુ), 'thoDu~' (થોડું), 'ghaNu~' (ઘણું) are also Particles. And there are certain Particles which denote Negation (RP_NEG) like, 'na' (ન), 'nahi' (નહિ) etc.

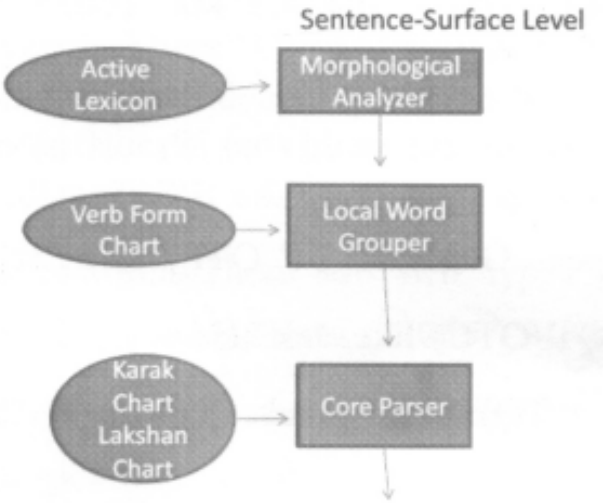
3.2.9 Demonstratives – There are certain words which function as Demonstratives. The BIS Tag of Demonstratives is 'DM'. Many of them are common with Pronouns, but they can be identified as Demonstratives from the context. They are Deictic (DM_DMD) like 'A'(આ), 'pelu~' (પેલું) etc., Relative (DM_DMR) like 'je' (જે), 'te' (તે) etc., 'Wh-word' (DM_DMQ) like 'kon' (કોણ), 'shu~' (શું) etc and Indefinite (DM_DMI) like 'koi' (કોઈ), 'ka~ik' (કંઈક) etc.

3.2.10 Quantifiers – There are certain words which act as adjectives and Quantifies the nouns to whom they are attached. The BIS Tag for Quanifiers is 'QT'. They can be of three types- General (QT_QTF) like 'thoDu~' (થોડું), 'ghaNu~' (ઘણું) etc., Cardinals (QT_QTC) like, 'ek' (એક-**1**), 'be' (બે-**2**), 'traN' (ત્રણ-**3**) etc., and Ordinals (QT_QTO) like 'pahelu~' (પહેલું), 'biju~' (બીજું), 'triju~'(ત્રીજું) etc.

3.2.11 Residuals – There are certain words which do not fall in any of these categories, they come into Residuals. The BIS Tag for Residuals is 'RD' There can be Foreign words (RD_RDF) like TV, CD written in another script, which is foreign to Gujarati. Even the Symbols (RD_SYM) like $, & are not known to Gujarati script, they also fall into this category. The Punctuation marks (RD_PUNC) are also under Residuals. The echo words (RD_ECH) like 'bAm' (બામ) in 'kAm bAm' (કામ-બામ) also come into this category and some other unknown words (RD_UNK) also come into this category.

3.3 Tagging is not so easy. It is context bound. One word can have more than one tag. In that case the context should be taken into consideration. Eg. in the sentence 'All writers should know this strategy', the word 'Writers' can be tagged as Noun, to be more precise, common Noun, so it should be tagged as N_NN. But in the sentence 'the writer Mr. Smith should know this strategy', 'Writer' is an adjective, so it should be tagged as JJ. So, the Tagging depends on the context. The role of the word plays in the sentence decides its category and should be tagged accordingly.

4.0 PARSING- Parsing requires a lot of preparations. After creating a good Corpus, analyzing and tagging is not enough. The process of Parsing is very complex. The process of tagging and parsing both requires an appropriate understanding of grammar as well as psychological processes behind the formation. When an utterance appears at the surface level, it has to pass through many processes. It can be seen from the following diagram.



4.1 Morphological Analyzer – Morphological Analyzer needs Paradigms to analyze the words. As mentioned above, Indian languages are very rich as far as Morphological operations are concerned. Nouns, Pronouns and Verbs have many variants. So we have to prepare Nominal, Pronominal and Verbal Paradigms for Morphological Analyzer. Gujarati Nouns can be classified into eight groups according to their Morphological operations. They can be classified as – (1) Masculine ending in 'O' (2) Masculine ending in 'a' (3) Masculine ending in any vowel other than 'a' (4) Feminine ending in 'a' (5) Feminine ending in any vowel other than 'a' (6) Neuter ending in 'u~' (7) Neuter ending in 'a' (8) Neuter ending in any other vowel except 'a' and 'u~'. All these nouns take different case suffixes for both the numbers- singular and plural.

4.1.1 Nominal Paradigms - We will take an example of different paradigms for 'a'-ending Neuter and Masculine Nouns like 'bALaka' (બાળક) and 'deva' (દેવ) as follows –

એકવચન	**બહુવચન**	**એકવચન**	**બહુવચન**
Singular	Plural	Singular	Plural
બાળક bALaka	બાળકો bALako	દેવ deva	દેવો devo
બાળકે bALake	બાળકોએbALakoe	દેવે deve	દેવોએ devoe
બાળકને bALakane	બાળકોને bALakone	દેવને devane	દેવોને devone
બાળકથી bALakathI	બાળકોથી bALakothI	દેવથી devathI	દેવોથી devothI
બાળકમાં bALakamA~	બાળકોમાં bALakomA~	દેવમાં devamA~	દેવોમાંdevomA~
બાળકનો bALakano	બાળકોનો bALakono	દેવનો devano	દેવોનો devono
બાળકની bALakanI	બાળકોની bALakonI	દેવની devanI	દેવોની dvonI
બાળકનું bALakanu~	બાળકોનું bALakonu~	દેવનું devanu~	દેવોનું devonu~
બાળકના bALakanA	બાળકોના bALakonA	દેવના devanA	દેવોના devonA

બાળકનાં bALaknA~ બાળકોનાં bALakonA~ દેવનાં devanA~ દેવોનાંdevonA~
બાળકમાંથી bALakamA~thI બાળકોમાંથી bALakomA~thI
દેવમાંથી devamAthI દેવોમાંથી devomA~thI
બાળકમાંનો bALakamA~no બાળકોમાંનો bALakomA~no
દેવમાંનો devamA~no દેવમાંનો devomA~no
બાળકમાંની bALakmA~nI બાળકોમાંની bALakomA~nI
દેવમાંની devamA~nI દેવોમાંની devomA~nI
બાળકમાંનું bALakomA~nu~ બાળકોમાંનું bALakomA~nu~
દેવમાંનું devam~Anu~ દેવોમાંનું devomA~nu~
બાળકમાંના bALakamA~nA બાળકોમાંના bALakomA~nA
દેવમાંના devamA~nA દેવોમાંના devomA~nA
બાળકમાંનાં bALakmA~nA~ બાળકોમાંનાં bALakomA~nA~
દેવમાંનાં devamA~nA~ દેવોમાંનાં devomAnA~

4.1.2 Pronominal Paradigms- Similarly, Gujarati Pronouns also have many variants. We can frame paradigms of personal Pronouns as follows-

હું, મેં, મને, મારો, મારી, મારું, મારા, મારાં, મારાથી, મારામાં,
મારામાંથી, મારામાંનો, મારામાંની, મારામાંનું, મારામાંના, મારામાંનાં

અમે, અમે, અમને, અમારો, અમારી, અમારું, અમારા, અમારાં,
અમારાથી, અમારામાં, અમારામાં, અમારામાંથી, અમારામાંનો,
અમારામાંની, અમારામાંનું, અમારામાંના, અમારામાંનાં

તું, તેં, તને, તારો, તારી, તારું, તારા, તારાં, તારાથી, તારામાં,
તારામાંથી, તારામાંનો, તારામાંની, તારામાંનું, તારામાંના, તારામાંનાં

તમે, તમે, તમને, તમારો, તમારી, તમારું, તમારા, તમારાં, તમારાથી,
તમારામાં, તમારામાંથી, તમારામાંનો, તમારામાંની, તમારામાંનું,
તમારામાંના, તમારામાંનાં

તે, તેણે, તેને, તેનો, તેની તેનું, તેના, તેનાં, તેનાથી, તેનામાં,
તેનામાંથી, તેનામાંનો, તેનામાંની, તેનામાંનું, તેનામાંના, તેનામાંનાં

તેઓ, તેણે, તેમને, તેમનો તેમની, તેમનું, તેમના, તેમનાં, તેમનાથી,
તેમનામાં, તેમનામાંથી, તેમનામાંનો, તેમનામાંની, તેમનામાંનું,
તેમનામાંના, તેમનામાંનાં

The paradigms of all other pronouns can be prepared on the basis of above diagrams.

4.1.3 Verbal Paradigms- From Morphological point of view, words can be classified into eight groups. There are approximately 2400 verb roots in Gujarati. They can be classified in to 8 groups. They are

cAla Group (ચાલ જૂથ) (250 Verb Roots)
Arambha Group (આરંભ જૂથ) (725 Verb Roots)
Besa Group (બેસ જૂથ) (5 Verb Roots)
Kara Group (કર જૂથ) (1300 Verb Roots)
khA Group (ખા જૂથ) (6 Verb Roots)
gA Group (ગા જૂથ) (44 Verb Roots)
jA Group (જા જૂથ) (1 Verb Root)
sU Group (સૂ જૂથ) (1 Verb Root)
bI Group (બી જૂથ) (1 Verb Root)

The verbal Suffixes they take for Verb root cAl () is as follows-

• Present Tense (વર્તમાન કાળ) - 4	(ચાલું, ચાલીએ, ચાલો, ચાલે)
	(cAlu~, cAlie, cAlo, cAle)
• Future Tense (ભવિષ્ય કાળ) - 4	(ચાલીશ, ચાલશું, ચાલશો, ચાલશે.)
	(cAlish, cAlshu~, cAlsho, cAlshe)
• Past Tense (ભૂતકાળ) - 5	(ચાલ્યો, ચાલી, ચાલ્યું, ચાલ્યા, ચાલ્યાં)
	(cAlyo. cAlI, cAlyu~, cAlyA, calyA~)
• Continuous Aspest (અપૂર્ણ અવસ્થા)- 5	(ચાલતો, ચાલતી, ચાલતું, ચાલતા, ચાલતાં)
	(cAlto, cAltI, cAltu~, cAltA, cAltA~)
• Expected Aspect(આગામી અવસ્થા)- 5	(ચાલવાનો, ચાલવાની, ચાલવાનું, ચાલવાના, ચાલવાનાં)
	(cAlvAno, cAlvAnI, cAlvAnu~, cAlvAnA, cAlvAnA~)
• Past participle (ભૂત કૃદન્ત) - 6	(ચાલેલ, ચાલેલો, ચાલેલી, ચાલેલું, ચાલેલા, ચાલેલાં,)
	(Calel, cAlelo, cAlelI, cAlelu~, clelA, cAlelA~)
• Future Participle(ભવિષ્ય કૃદન્ત) - 6	(ચાલનાર, ચાલનારો, ચાલનારી, ચાલનારું, ચાલનારા, ચાલનારાં)
	(cAlnAr, cAlnAro, cAlnArI, cAlnAru~, cAlnArA, cAlnArA~)
• Potential Participle(વિધ્યર્થ કૃદન્ત) - 5	(ચાલવું, ચાલવો, ચાલવી, ચાલવા, ચાલવાં)
	(cAlvu~, cAlvo, cAlvI, cAlvu~, cAlvA, cAlvA~)
• Imperative (આજ્ઞાર્થ) - 2	(ચાલ, ચાલો)
	(cAla, cAlo)
• Future Imperative(ભાવિ આજ્ઞાર્થ) - 2	(ચાલજે, ચાલજો,)
	(cAlje, cAljo)

• Conditional (શરતી ભૂત) - 1 (ચાલત)
(cAlata)

45

Thus there are 45 verbal suffixes which make 270 verbal forms (45 forms of Active, 45 forms of Passive, 45 forms of Causal, 45 forms of Causal Passive, 45 forms of double causal and 45 forms of double causal passive). There are other forms of verbal Nouns and Gerund which make 400 plus forms.POS Tagging and Parsing Nilotpala Gandhi Page 12

4.2 Local Word Grouper- A Sentence is made up of Phrases. There are two types of phrases, Noun Phrase and Verb Phrase; Phrases are made up of words and words are made up of Root and suffixes. In case of Noun Phrase, Adjectives, Nouns or Pronouns and Post positions make a phrase; and in case of Verb Phrase, Adverb, Main verb, and Auxiliary make a phrase. So, using the paradigms of Nouns, Pronouns and verbs, Local Word Grouper is formed. The lists of Adjectives, Adverbs, Conjuncts and Interjections should also be provided.

4.3 Core Parser – Karaka Relations are very important for Parsing. Core Parser needs to analyze Karaka relations. Gujarati denotes the karaka relations either through case suffixes or through Post positions. Case suffixes are written together with the Nouns to whom they are attached to, but the post positions are written separately. So case suffixes are not being considered as a separate 'Token', but Post Positions are considered as separate tokens. It also needs to correlate Karaka and Vibhakti along with post positions. These post positions are denoted either by nAm yogi or by NLoc Nouns.

4.4 Pre-requirements- As mentioned above, Machines are not able to draw grammar rules, a special grammar for machines should be prepared. A special dictionary along with the paradigms is also needed for Local word Grouper.

5.0 Summing up- Gujarati, genealogically belonging to Indo-Aryan Language family is extremely rich in Morphology. Paninian Model can be suitable for its computation, but it must be described and analyzed from its own approach. Paninian Model for Sanskrit describes eight cases and six karakas where as Gujarati has 13 karakas and five Vibhaktis. Sanskrit has six tenses and four moods and no aspect, where as Guajarati has three tenses, nine Aspects and Five moods. So, it's a time to prepare its new model for computation.

5.1 POS Tagging and Parsing are very important stages for the computation of any language. Richer the language tougher Morphology. But much work has been done for Gujarati in this direction. Like in other fields, it can lead other vernacular Languages in the field of computation too.

❑