[VNSGU JOURNAL OF SCIENCE AND TECHNOLOGY] Vol.4. No. 1, July, 2015
127 - 133,ISSN : 0975-5446

Morphological Rule Set and Lexicon of Gujarati Grammar: A Linguistics Approach

KAPADIA Utkarsh N.
Department of Computer Science
Veer Narmad South Gujarat University, Surat
utkarsh_kapadia@yahoo.com

DESAI Apurva A.
Department of Computer Science
Veer Narmad South Gujarat University, Surat
aadesai@vnsgu.ac.in

Abstract

Morphological analysis is at the core in the field of linguistics. For morphological analysis there are various methods like rule based and machine learning. For performing analysis via any approach, it is inevitable to validate words against morphological rules and lexicon. This paper presents morphological rules for Gujarati language classes and lexicon database. Gujarati lexicon based on UNICODE system is developed containing above 15000 words. All words are categorized in various grammar classes, and affixes have been separated, inflection and derivation rules are defined.

Keywords: Gujarati morphology, morphological rules, Gujarati lexicon

1. Introduction

The Gujarati language is native to the Indian state of Gujarat. Some estimates suggest that there are about 46.1 million people speaking the Gujarati language all over the world. Similar to the Hindi language, it also belongs to the Indo-Aryan family of languages.

Indian languages like Gujarati, Hindi etc are morphologically and inflectionally rich. Gujarati is also rich in morphology but poor in resources. First step in linguistic is to have word analysis of language. For developing grammar inference system for Indian language like Gujarati, we must necessarily have lexical analysis at its base. It is found that there are very less resources and mainly database available for language processing. It was decided to develop lexicon and rule database for Gujarati that can be used subsequently. It is a first step to progress towards the goal of developing sharable database having morphological grammar rules, and lexicons.

2. Related Work

Beth Levin [1] has classified 3000 English verbs according to shared meaning and behavior. Levin [1] starts with hypothesis that a verb's meaning influences its syntactic behavior and develops it into powerful tool for studying English verb lexicon. They [1] have isolated three classes by examining verb behavior with respect to wide range of syntactic alterations that reflect verb meaning.

Tagging of various Gujarati part of speech like nouns, verbs is done by Babu Suthar [2]. It has been demonstrated in detail how each grammar class alters with respect to number, gender and tense. Gujarati words can be classified in two classes, namely Open class and closed classes [2]. There are few resources in form of printed media like "Practical Gujarati Grmmar" by Arunoday Foundaiton[3]. It discusses Gujarati Grammar classes and word inflections. Unfortunately, most of them are not developed with linguistic perspective in mind and follows dictionary-based approach that merely lists all words and their vocal information. English morphology is discussed in Cross-linguistic typology [6] which is important for linguistic study of English.

Some foundations have put in their efforts to develop Gujarati Lexicon that can be used for Linguistics study of Gujarati script. Some of them are also available online mainly Gujarati Laxicon[4]. Ratilal Chandaria foundation have developed Gujarati database named Bhagwadgomandal [5]. It also follows dictionary approach mainly and all words are tagged according to their grammar classes.

Many resource based on Hindi is available in area of linguistic study of the language. Debasari Chakrabarti and Pushpak Bhattacharya present verb alteration in Hindi language and their syntactic alteration rules in their study [7]. Study made by them confined to mainly verbs and their syntactic alternations. Combination of Inflectional and derivational approach in morphological analysis is applied to design derivational morphological analyzer. Nikhil, Abhilash and Dipti have presented the same in their study [8]. Derivational rules designed by studying properties of word suffixes manually in their work. Classes should be merged into a single class as they take similar inflections [9]. Smriti Singh and Vijayanthi Sarma discuss inflection categories and inflection classes in detail in their study.

The morphological analyzer developed by Vishal and Gurpreet stores all the commonly used word forms for all Hindi root words in its database [10]. Thus, space is a constraint for this analyzer but the search time is quite low. The morph analyzer developed by Niraj and Robert extracts a set of suffix replacement rules from a corpus and a dictionary [11]. The rules are applied to an inflected word to obtain the root word. They show that the process of developing such rules sets is simple and it can be applied to develop morphological analyzers of other Indian languages.

Inflection rules are derived for making rule based stemmers for Gujarati [12]. Some of rules are presented for stemming of gujarati sentence in Gujarati Stemmer [13]. POS tagging is done by rule based approach in their study [15]. We have developed Gujarati lexicon and grammar rules in form of UNICODE database which can be utilized for NLP.

3. Gujarati Morphology

3.1 Language Basics

Gujarati language belongs, like the Marathi, Hindi, Punjabi, Oriya and many other Indian dialects, to the Aryan family, being a daughter of the language of Indian origin i.e. Sanskrit. There are eight parts of speech as in English, mainly four, that are Noun (નામ), the Pronoun (સર્વનામ), the Adjective (વિશેષણ), and the Verb (ક્રીયાપદ).

Gujarati has three genders (masculine, feminine and neutral), two number types (singular and plural) and three cases i.e. nominative, vocative/oblique and locative. Gujarati language is different from English, which follows S-V-O while Gujarati language follows S-O-V.

Gujarati words can be divided into Open class and Closed class. Open class consists of nouns, verbs, adjectives and adverbs, while closed class consists of pronouns and other pro-forms, noun adjuncts, verb adjuncts, conjunctions and others[2].

A natural language can be viewed as a set of sentences; a sentence as a string of one or more words from vocabulary of language and grammar as a finite set of rules for the set of sentences (infinite) comprising language under study.

Typical reproduction rules of natural language are:

<Sentence> -> <Noun phrase> <Verb phrase> '.'
<Noun phrase> -> <Adjective> <Noun Phrase>
<Noun phrase> -> <Article> <Noun>
<Noun> -> <Noun> | <Proper Noun>
<Noun> -> <Pronoun> | <Noun> <Relative Cl.>
<Verb phrase> -> <Verb> (<Noun phrase>)

3.2 Morphology

Not all Gujarati words in frequent use are stored in the dictionary. For example, for a single noun in Gujarati, over 200 forms that are either adjectives or adverbs may be possible. Similarly, a verb may exhibit over 450 forms. At the same time, the language is expected to include over 10,000 nouns and over 1,900 verbs. Over 175 postpositions can be attached to nominal and verbal entities. Some postpositions can occur in compound forms with most other postpositions.

We present here approach to morphological rule grammar development from the stand-point of Gujarati morphology.

Morphology is the study of the form of words:

- The word is basic element of speech
- The smallest unit which can be used to form words are morphemes
- Words can be formed by process of Derivation and inflection
- In syntactic analysis, Word is used to mean a basic functional unit.

Few examples of rules, which results in words, formed by roots and morphemes, are shown in Table1.

Table1: Example of words and its morphemes

Type	**Root + Morpheme**	**Word**
Inflectional	રમ + શે ((to play) (tense maker)	□□□□ (verb) (will play)
Derivational	□□ + □□□ ((to play) (case maker)	□□□□□(noun) (player)

Left side of the Morphemes columns has morphological roots, and affixes are next to them in Table1. Word column shows the word formed by combining morphemes to the left. Morphemes are smallest units, which can be combined with other morphemes or root to form the word.

Derivation is the process of obtaining new morphemes from the source morphemes. Inflection is a process of modification of a root to express grammatical relationship.

We can classify Gujarati morphemes into following groups:

Free form or bound form morphemes, called (ધાતુ /મૂળરૂપ), they represent lexical meanings. They are similar to stem or root but not equal.

Affixes called (પ્રત્યય / પુર્વગ). They are also called inflectional/derivational prefixes(પુર્વગ) and suffixes (પ્રત્યય)

Affixes of type (વિભક્તિ પ્રત્યય). They are mostly derivational in nature

There are two types of Gujarati words. One type is made up of combination of at least one free-form morphemes and zero or more affixes. The other type consists of one bound-form morpheme and more than one affix.

4. Rules of Morphology

Morphological analysis is applied to the categories of nouns, pronouns, adjectives, verbs, adverbs, postpositions, conjunctions and interjections. In Gujarati, it is convenient to use rules of replacement to capture all types of morphological behavior. Rules of replacement are generic enough to also cover all possibilities of additions and deletions of consonants and vowels.

4.1 Pronoun Morphology

Grammatical definition of pronoun says that they are used in place of noun. Gujarati personal pronouns differentiate three persons (first, second and third) and two numbers (singular, plural). They have also inclusive and exclusive contrast in third person plural. In addition, their second person plural form is also used as honorific. Personal pronouns take various cases, which include nominative, ergative, accusative/dative, genitive and locative and instrumental. From these genitive forms distinguish among three genders and two numbers[2]. E.g. હું , મેં, અમે, આપણે, મને, આપણને etc.

Exhaustive list of all possible (over 200) inflections of all pronouns is prepared because pronouns show very irregular behavior. The ratio of inflectional rules to actual forms in the case of pronouns is close to one. A pronoun has a specific single oblique form to which all shabdayogi avyays are attached.

4.2 Noun morphology

Major types of nouns in Gujarati are: Proper Noun, Abstract Noun, Common Noun etc. Gujarati has three genders, two numbers and three cases they are nominative, vocative/oblique and locative.

Some common noun inflections rules for Nominative case are given in Table2:

Table2: Common Nouns with thier formation rules

Stem + <Affixes>	Noun Form	Formation Rules	Class
છોકર + ઓ →	છોકરો	chhokr-o	Masc-Sg
છોકર + આ →	છોકરા	chhokr-aa	Masc-Plural
છોકર + ઈ →	છોકરી	chhokr-i-∅	Fem-Sg
છોકર + ઈ+ ઓ →	છોકરીઓ	chhokr-i-o	Fem-Pl
છોકર + ઉં →	છોકરું	chhokr-u	Neu-Sg
છોકર + આ+ ઓ (	છોકરા ઓ	chhokr-aa-o	Neu-Pl/Masc-Pl

We can generalize the noun structure as follow: Noun stem + Gender marker + Number marker.

Gujarati has many nouns which end in -o and -i, but they do not fall into the category of masculine or feminine respectively. Consider the nouns ઘો (reptile) and પાણી (water). So it would be a wrong analysis as -o and -i are not the gender markers. They are a part of the stem for these two words and therefore we can not consider them gender marker.

So based on above facts we have separated nouns, which are having affixes as part of their stem itself.

4.3 Adjective Morphology

Adjectives may be divided into declinable and indeclinable (વિકારી અને અવિકારી) categories. Declinable adjectives changes as per gender of noun while indeclinable adjectives doe not vary as per noun's gender. Consider the examples of સારુ(good) and સુંદર(beautiful), variable and invariable adjectives respectively.

Table3: Adjective Formations

		Nominative		Locative
	Rules	Sing	Plu	
Masc	sar-o	સારો [છોકરો]	સારા [છોકરા]	સારે [મહિને]
Fem	sar-i	સારી [છોકરી]	સારી [છોકરીઓ]	સારી [છોકરીને]
Neut	sar-uN	સારું [કારખાનું]	સારા [કારખાનાં]	સારે [કારખાને]

Based on above facts rules more than 1000 indeclinable adjectives were separated and stored separately.

We can classify Adjectives based on grammatical sub classes in to following sub classes: Adjective of Quality, Quantity, and Relation.

Generation of Adjectives from Nouns (સંજ્ઞા પરથી વિશેષણ):

Nouns + Affix	Adjectives
સમાજ + ઈક → e.g ધન, અઠવાડિયુ	સામાજિક
રાષ્ટ્ર + ઈય → e.g. રાષ્ટ્ર, માનવ, પૂજન	રાષ્ટ્રીય
વાસ + ઈ (e.g □□□□□, □□□□	□□□□

Table4: Adjective Formations

There are other affixes like that were identified mainly they are (ળુ, આળ, સુ, વી, વાન, માન, વાળ, વાળું, અક,ટ,ય etc.).

4.4 Verb Morphology

Gujarati inflected verbs have the following pattern: verb stem + inflectional material. Inflectional material may consists of various features such as tense, person, gender. In Gujarati, verbs also inflect for imperative (present and future), desiderative, obligatory, conditional, and infinitive constructions.

A regular verb root generates over 80 forms. In addition to regular verbs, there are over 35 irregular verbs. The rules are represented in the form of tables.

5. Lexical Dictionary

These lexicons were developed using Gujarati grammar book [2] and Gujarati dictionary from Gujarati Lexicon [3]. 15664 entries in dictionary were stored for testing; summary is shown in Table 5:

Table5: Gujarati Lexical Dictionary

File Name	Number of Lexicons (Entries)	Contents
tbl_Pronoun	210	Pronouns
tbl_ProperNouns	8500	Proper Nouns
tbl_CommonNoun	4780	Common Nouns
tbl_Adverb	60	Adverbs
tbl_VerbStems	707	Verb Roots
tbl_VerbAffix	60	Verb Affixes
tbl_Adjectives	1352	Adjectives
tbl_Terminals	5	Terminals
TOTAL	**15664**	

6. Conclusion

In this paper we have presented useful package composed of morphological grammar rules, dictionary, test data, and a set of API. The rules are implemented in database for further processing and development of morphological analyzer for Gujarati language.

Acknowledgement

We acknowledge UGC for Special Assistance Program (SAP) for Natural Language Processing and Data Mining (file number is F.3-48/2011) under which this work has been done.

References

1. Levin Beth, 1993, English Verb Classes and Alternations A Preliminary Investigation, The University of Chicago Press, Chicago.
2. "Brief outline of Gujarati Parts of Speech" – Babu Suthar – University of Pennsylvania, PA
3. "Practical Gujarati Grammar" – August 2010, 2nd Edition , Dr. Arvin kothari, Arunoday Publication, Ahmedabad, India
4. Gujarati Lexicon – Ratilal Chandria Foundation – Ahmeadabad, Gujarat, India
5. Gujarati Bhagwadgo mandal - http://www.bagwadgomandal.org
6. R.M.W. Dixon, 2003. Word: A Cross – Linguistic Typology, Press Syndicate of Cambridge, Cambridge, U. K. 2004.
7. Syntactic Alteration to Hindi Verbs with Reference to the Morphological Paradigm – Debasari Chkraborti & Pushpak Bhattacharya
8. Hindi Derivational Morphological Analyzer – Nikhil Kanuparthi, Abhliash Inumella, Dipti Misra Sharma, IIT Hydrabad, India
9. Hindi Noun Inflection and Distributed Morphology – Smriti Singh, Vijayanthi M Sharma
10. Hindi Morphological Analyzer and Generator – Vishal Goyal and G S Lehal, Punjabi University
11. Developing Morphological Analyzer for South Asian Languages: Experimenting with Hindi and Gujarati Languages – Niraj Aswani, Robert Gaizauskas, University of Sheffield, UK
12. A light weight Stemmer for Gujarati – Juhi Ameta, Nisheeth Joshi, Iti Mathur, Banasthali University, Rajasthan
13. Hybrid Inflectional Stemmer and Rule-Based Derivational Stemmer for Gujarati – Kartik Suba, Dipti Jiandani, (Dept Of Comuter Science, DDIT),Pushpak Bhattacharyya, IIT, Bombay