RESEARCH ARTICLE OPEN ACCESS

Phonetic Dictionary for Natural Language Processing: Kannada

Mallamma V. Reddy*, Hanumanthappa M. **, Jyothi N.M***, Rashmi S.
*(Department of Computer Science, Rani Channamma University, Vidyasangam, Belgaum-591156, India)
** (Department of Computer Science, Bangalore University, Jnanabharathi Campus, Bangalore-561156, India)
***(Department of Master of Computer Applications, Bapuji Institute of Engineering and Technology, Davangere-577004, India)

ABSTRACT
India has 22 officially recognized languages: Assamese, Bengali, English, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Tamil, Telugu, and Urdu. Clearly, India owns the language diversity problem. In the age of Internet, the multiplicity of languages makes it even more necessary to have sophisticated Systems for Natural Language Process. In this paper we are developing the phonetic dictionary for natural language processing particularly for Kannada. Phonetics is the scientific study of speech sounds. Acoustic phonetics studies the physical properties of sounds and provides a language to distinguish one sound from another in quality and quantity. Kannada language is one of the major Dravidian languages of India. The language uses forty nine phonemic letters, divided into three groups: Swaragalu (thirteen letters); Yogavaahakagalu (two letters); and Vyanjanagalu (thirty-four letters), similar to the vowels and consonants of English, respectively.

Keywords **-** Information Retrieval, Natural Language processing (NLP), Phonetics

I. INTRODUCTION

Kannada (ಕನ್ನಡ) or Canarese, the official language of the southern Indian state of Karnataka. Kannada is a Dravidian language [1] spoken by about 44 million people in the Indian states of Karnataka, Andhra Pradesh, Tamil Nadu and Maharashtra. The Kannada alphabet (ಕನ್ನಡ ಲಿಪಿ) developed from the Kadamba and Cālukya scripts, descendents of Brahmi [2] which were used between the 5th and 7th centuries AD. These scripts developed into the Old Kannada script, which by about 1500 had morphed into the Kannada and Telugu scripts. Under the influence of Christian missionary organizations, Kannada and Telugu scripts were standardized at the beginning of the 19th century. Kannada has 44 speech sounds. Among them 35 are consonants and 9 are vowels. The vowels are further classified into short vowels, long vowels and diphthongs [3]. Notable features of Kannada language are:

- *Type of writing system:* alphasyllabary in which all consonants have an inherent vowel. Other vowels are indicated with diacritics, which can appear above, below, before or after the consonants.
- When they appear the beginning of a syllable, vowels are written as independent letters.
- When consonants appear together without intervening vowels, the second consonant is written as a special conjunct symbol, usually below the first.
- Direction of writing: left to right in horizontal lines

The phonetic can be defined as:

- pho·net·i·cal. of or pertaining to speech sounds, their production, or their transcription in written symbols.
- Corresponding to pronunciation: phonetic transcription.
- Agreeing with pronunciation: phonetic spelling.
- Concerning or involving the discrimination of nondistinctive elements of a language. In English, certain phonological features, as length and aspiration, are phonetic but not phonemic.

A phonetic dictionary is a dictionary that allows you to locate words by the "way they sound", i.e. a dictionary that matches common or phonetic misspellings with the correct spelling of a word. Such a dictionary uses pronunciation respelling to aid the search for or recognition of a word.

II. HEADINGS

The Phonetics was studied as early as 500 BC in the Indian subcontinent, with Pāṇ ini's account of the place and manner of articulation of consonants in his 5th century BC treatise on Sanskrit. The major Indic alphabets today order their consonants according to Pāṇ ini's classification. The Phoenicians are credited as the first to create a phonetic writing system, from which all major modern phonetic alphabets are now derived. Modern phonetics [4] begins with attempts

to introduce systems of precise notation for speech sounds.

Stephen Hawking's Speech synthesiser: Stephen Hawking, an English theoretical physicist, cosmologist, author and Director of Research at the Centre for Theoretical Cosmology within the University of Cambridge uses speech synthesiser when he communicates. Using this he is able to convert the text into speech. A program called EZ keys has been written by world plus Inc. whatever he types on the keyboard the cursor will automatically scan each word row and column wise. Then the system produces the respective sounds. There is also word prediction option.

NSA [5] has proposed a method of using the Soundex and Asoundex approach. Here the author has carefully addressed the issues of homophones. 'New' and 'Knew'; 'no' and 'know'; to, two, too; meat, meet: are some of the examples of homophones. The work is mainly concentrated on generating the "name code". Code generated by the program is then compared with the names in the test data. It has been shown for "Malay" language (Language spoken by Malaysian people)

The phonetic segmentation [6] and considerably deals with articulation phonetics. The work is concentrated on cross language phonetic segmentation using Hidden Markov Models (HMMs). It also provides extensive models that are applicable across languages. The efficiency was tested on Appen Spanish speech corpus and the efficiency of about 61.5% was achieved.

The speech recognition system uses the knowledge obtained from phonotactics, phonology, and acoustic phonetics to apprehend admissible phonetic information. Hence the system can make broad classifications and also detailed phonetic distinctions. Espy-Wilson et al has proposed a paper [7] which discusses a framework for developing a phonetically based recognition system. The recognition task is the class of sounds known as the semivowels. Though the results obtained from the study were incomplete, it stimulated the bedrock for further studies.

A phonetic segmentation method based on speech analysis under Microcanonical Multiscale Formalism (MMF). The MMF [6] depends on computation of local geometrical parameters, singularity exponents (SE). The SE conveys the phoneme boundaries. The author has proposed the 2-steps technique, which exploits the SE to upgrade the segmentation accuracy. The first step concentrates on detecting the boundaries of original signals and a low-pass filtered version and the union of all the detected boundaries are considered as the candidates. In the second step the hypothesis test was used over the original signal on the local SE distribution to determine the final boundaries. The authors have also done detailed evaluation and comparison test on TIMIT (acoustic-phonetic continuous speech corpus) which aids the other researchers to accomplish the collation task.

English language is homophonic in nature. The same words have multiple spelling orders. For example, 'Soumya' and 'Sowmya'. Hence the author, proposed a algorithm [8] based on English phonetic spelling correction, has shown the mechanisms to solve common spelling errors such as missing letters, extra letters, disordered letters, as well as phonetic spelling errors in the perspective of from the same pronunciation, similar pronunciation. Algorithms such as phonetic spelling correction, phonetic spelling regulations, edit-distance and habit-distance is put forward carefully by the authors, thereby the overall exposition about spelling correction system is vigilantly proposed.

III. METHODOLOGY

Kannada script has forty-nine characters in its alphasyllabary and is phonemic. The Kannada character set is almost identical to that of other Indian languages. The number of written symbols, however, is far more than the 49 characters in the alphasyllabary, because different characters can be combined to form compound characters (ottaksharas). Each written symbol in the Kannada script corresponds with one syllable, as opposed to one phoneme in languages like English. The Kannada writing system is an abugida, with consonants appearing with an inherent vowel.

The characters are classified into three categories [3] swaras (vowels), vyanjanas (consonants) and Yogavaahakas (part vowel, part consonants– two letters: anusvara and visarga).

The name given for a pure, true letter is akshara, akkara or varna. Each letter has its own form (ākāra) and sound (shabda); providing the visible and audible representations, respectively. Kannada is written from left to right.Kannada alphabet (aksharamale or varnamale) now consists of 49 letters.

Each sound has its own distinct letter, and therefore every word is pronounced exactly as it is spelt; so the ear is a sufficient guide. After the exact sounds of the letters have been once gained, every word can be pronounced with perfect accuracy. The accent falls on the first syllable. Each consonant sound has two distinct pronunciations. The sound with normal pronunciation (deergha) is generally used in the varnamala or aksharamala:

1. short/brief one (ಕ್, also known as hrasva, ಹ್ರಸ್ವ), without any vowel.
2. long/normal one (ಕ, also known as deergha, ದೀರ್ಘ), in union with the first vowel (ಅ).

swaras (Vowels): A letter of the alphabet shown in Table I. that represents a speech sound created by the relatively free passage of breath through the larynx and oral cavity. There are thirteen vowels (swaras).

Kannada	ಅ	ಆ	ಇ	ಈ	ಉ	ಊ	ಋ	ೠ
ISO notation	a	ā	i	ī	ou	oū	ru	rū
Kannada	ಎ	ಏ	ಐ	ಒ	ಓ	ಔ		
ISO notation	ye	yē	ai	o	ō	au		

Table I. List of Vowels in Kannada

vyanjanas (Consonats): a letter of the alphabet that represents a speech sound produced by a partial or complete obstruction of the air stream by a constriction of the speech organs. There are two types of consonants (vyanjana) are identified in Kannada: the structured consonants and the unstructured consonants. The structured consonants are classified according to where the tongue touches the palate of the mouth and are classified accordingly into five structured groups.

Structured consonants: These consonants are shown in Table II with their IAST transcriptions.

Unstructured consonants: The unstructured consonants are consonants that do not fall into any of the above structures are shown in Table III:

	voiceles s	**voiceles s aspirate**	**voice d**	**voiced aspirat e**	**nasa l**
Velars	ಕ (ka)	ಖ (kha)	ಗ (ga)	ಘ (gha)	ಙ (nga)
Palatals	ಚ (cha)	ಛ (chha)	ಜ (ja)	ಝ (jha)	ಞ (ña)
Retrofle x	ಟ (ṭ a)	ಠ (ṭ ha)	ಡ (ḍ a)	ಢ (ḍ ha)	ಣ (ṇ a)
Dentals	ತ (ta)	ಥ (tha)	ದ (da)	ಧ (dha)	ನ (na)
Labials	ಪ (pa)	ಫ (pha)	ಬ (ba)	ಭ (bha)	ಮ (ma)

Table II. List of Structured consonants in Kannada

ಯ (ya)	ರ (ra)	ಱ (ṟ a)	ಲ (la)	ವ (va)	ಶ (śa)
ಷ (ṣ a)	ಸ (sa)	ಹ (ha)	ಳ (ḷ a)	ೞ (ḻ)	ಷ (ṣ a)

Table III. List of Unstructured consonants in Kannada

Consonant Conjuncts: The Kannada script is rich in conjunct consonant clusters, with most consonants having a standard subjoined form and few true ligature clusters. Here Table IV. Shown some of consonant conjuncts follow, although the forms of individual conjuncts may differ according to font.

ಕ	ಕ್ಕ	ಕ್ಖ	ಕ್ಗ	ಕ್ಘ	ಕ್ಙ	ಕ್ಚ	ಕ್ಛ	ಕ್ಷ
ಕ್ಥ	ಕ್‌ದ	ಕ್ಧ	ಕ್ನ	ಕ್ಪ	ಕ್ಫ	ಕ್ಬ	ಕ್ಭ	
ಕ್ಠ	ಕ್ಡ	ಕ್ಢ	ಕ್ಣ	ಕ್ತ	ಕ್ವ	ಕ್ಶ	ಕ್ಮ	ಕ್ಸ
ಕ್ಜ	ಕ್ಝ	ಕ್ಞ	ಕ್ಟ	ಕ್ಯ	ಕ್ರ	ಕ್ಲ	ಕ್ಢ	

Table IV. Some of consonant conjuncts in Kannada: Ka

The Yogavaahakas (part-vowel, part consonant) include two letters:

1. The anusvara: ಅಂ (aom)
2. The visarga: ಅಃ (ahā)

Another two Yogavaahakas used in Sanskrit, but present in Kannada script, are known as Ardhavisarga:

1. The Jihvamuliya: ೱ
2. The Upadhmaniya: ೲ

IV. IMPLEMENTATION

Phonetics (pronounced /fəˈnɛtɪks/, from the Greek: φωνή, phōnē, 'sound, voice') is a branch of linguistics that comprises the study of the sounds of human speech, or—in the case of sign languages—the equivalent aspects of sign. It is concerned with the physical properties of speech sounds or signs (phones): their physiological production, acoustic properties, auditory perception, and neurophysiologic status. Phonology, on the other hand, is concerned with the abstract, grammatical characterization of systems of sounds or signs.

The field of phonetics is a multilayered subject of linguistics [2] that focuses on speech. In the case of oral languages (phonetics) as a research discipline has three main branches or areas of study to implement phonetic dictionary for Kannada:

1). *Articulatory phonetics*: is concerned with the articulation of speech: The position, shape, and movement of articulators or speech organs, such as the lips, tongue, and vocal folds as shown in Fig.1.

2). *Acoustic phonetics:* is concerned with acoustics of speech i.e. the study of the physical transmission of speech sounds from the speaker to the listener: The spectro-temporal properties of the sound waves produced by speech, such as their frequency, amplitude, and harmonic structure.

3). *Auditory phonetics:* is concerned with speech perception: the perception, categorization, and recognition of speech sounds and the role of the auditory system and the brain in the same. In other words it is the study of the reception and perception of speech sounds by the listener.

These areas are inter-connected through the common mechanism of sound, such as wavelength (pitch), amplitude, and harmonics. Based on places of articulation we develop the phonetic dictionary for Kannada.

In the production of vowel sounds the air-current coming from lungs is allowed to go out without any obstruction in the mouth. However, there is a definite pattern in the production of different vowels sounds. Direction in which the tongue moves and variation in the shape of lips result in the change in the shape of the air chamber.

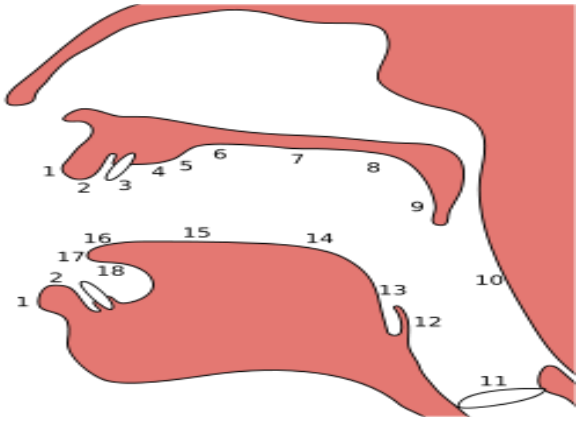

Fig.1: Places of articulation (passive & active) [9]:
1. Exo-labial, 2. Endo-labial, 3. Dental, 4. Alveolar,5. Post-alveolar, 6. Pre-palatal, 7. Palatal,
8. Velar, 9. Uvular, 10. Pharyngeal, 11. Glottal, 12. Epiglottal, 13. Radical, 14. Postero-dorsal,
15. Antero-dorsal, 16. Laminal,
17. Apical, 18. Sub-apical

It is this particular change which is responsible for the above mention production. In the production of Kannada vowels the vocal cords are vibrated and the nasal passage is closed.

V. CONCLUSION

This paper includes the basic information about the phonetics of the Kannada syllables and place of articulation; this will help to build the phonetic dictionary. The phonetic dictionary is useful to learn natural languages which are the media for transforming and communicating thoughts between the individuals. This phonetic dictionary helps even foreigners to learn and spell Indian languages fluently. The future work associated with dictionary-based phonetics are, (1) the processing of inflected words, (2) phrase identification and translation, and (4) lexical ambiguity.

REFERENCES

[1] http://en.wikipedia.org/wiki/Kannada_language
[2] "*BhashaIndia.com :: Kannada*". Archived from the original on 2008-04-14. Retrieved 2014-05-01.
[3] http://en.wikipedia.org/wiki/Kannada_alphabet
[4] http://en.wikipedia.org/wiki/Phonetics
[5] *Phonetic coding methods for Malay names retrieval*, Semantic Technology and Information Retrieval (STAIR),2011,Mutalib, IEEE International Conference on 28-29 June 2011 Page(s):125 - 129 E-ISBN :978-1-61284-353-7 Print ISBN:978-1-61284-354-4 INSPEC Accession Number:12192197
[6] *Improving text-independent phonetic segmentation based on the Microcanonical Multiscale Formalism*. Khanagha, V, Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on 22-27 May 2011 Page(s): 4484 – 4487 ISSN:1520-6149 E-ISBN:978-1-4577-0537-3 Print ISBN:978-1-4577-0538-0
[7] *A phonetically based semivowel recognition system*, Espy-Wilson, Carol Y, Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86. (Volume: 11) Digital Object Identifier: 10.1109/ICASSP.1986.1168807
[8] *Based on the Phonetic Spelling Correction System Research and Implementation*, Li Zhao, Computational Intelligence and Software Engineering, 2009. CiSE 2009. IEEE International Conference on 11-13 Dec. 2009 Page(s):1 – 4 E-ISBN : 978-1-4244-4507-3 Print ISBN: 978-1-4244-4507-3
[9] http://en.wikipedia.org/wiki/Place_of_articulation