IDENTIFICATION OF PROSODIC FEATURES OF PUNJABI FOR ENHANCING THE PRONUNCIATION LEXICON SPECIFICATION (PLS) FOR VOICE BROWSING

Swaran Lata[1], Prashant Verma[2] and Swati Arora[3]

[1]Centre of Linguistics, JNU, New Mehrauli Road, New Delhi
[2]Web Standardization Initiative, DeitY, New Delhi
[3]Web Standardization Initiative, DeitY, New Delhi

ABSTRACT

Voice browsing requires speech interface framework. Pronunciation Lexicon Specification (PLS) 1.0 is a recommendation of Voice Browser Working Group of W3C (World-Wide Web Consortium), a machine-readable specification of pronunciation information which can be used for speech technology development. This global PLS standard is applicable across European and Asian languages and this specification is extendable to all human languages. However, it currently does not cover morphological, syntactic and semantic information associated with pronunciations. In Indian languages, grammatical information is relatively encoded in its morphology, than syntax unlike English where the grammatical information is an integral part of syntax. In this paper, PLS 1.0 has been examined from the perspective of augmentation of prosodic features of Punjabi such as tone, germination etc.

Keywords:

PLS, W3C, POS, TTS, XML, Punjabi, Tone, Prosody, Morphology, Phonology, Phonetic, Geminations

1.INTRODUCTION

Pronunciation Lexicon Specification (PLS) is a recommendation of World Wide Web Consortium (W3C) and its current version is PLS 1.0 (2008) (http://www.w3.org/TR/pronunciation-lexicon/) produced by Voice Browser Working Group of W3C. PLS is designed to enable interoperable specification of pronunciation information for both speech recognition and speech synthesis engines within voice browsing applications. It helps developers in supporting the accurate specification of pronunciation information for international use through the use of language tag as provisioned. The current version of PLS may be referred as base line specification as it addresses the requirements of Latin script based languages only however few examples have been cited for Japanese and Chinese, thus keeping the specification very broad based. The specification covers the multiple pronunciations and multiple orthography in the XML structure at the lexicon level thus providing the flexibility of creating language specific PLS documents. The Meta tags feature is available for describing the domain and end use. Thus the PLS data can be prepared in the XML format for specific language using the base line PLS specification of W3C. The pronunciation lexicon markup language enables consistent platform for independent control of pronunciations for use by voice browsing applications. Thus this specification can be extended to all other human languages by examining the language-specific requirements.

PLS 1.0 recommendation currently does not cover morphological, syntactic and semantic information associated with pronunciations (such as word stems, inter-word semantic links, pronunciation statistics, prosody etc.), hence it will be appropriate to research the additional language specific requirements in this context. Some of the prosodic features in a speech signal are reflected as pitch variations, shortening or lengthening of sound units, prom

nence of certain syllables etc. The stress, rhythm, tone, accent etc. are some of the parameters which can be used to define Prosody. Hence it is appropriate to examine such features to work towards enhancement of PLS.

All Indo-Aryan languages share common phonetic features however Punjabi in this family is highly tonal as discussed by Swaran Lata et. al. in her paper on " Exploratory Analysis of Punjabi Tones in relation to orthographic characters: A Case Study". Therefore Phonological features of Punjabi language such as stress, tone, gemination, nasalization etc. will be used to examine the extra XML elements required to be incorporated in PLS. Thus augmented PLS as proposed for Punjabi will become applicable to all indo-Aryan languages other than the tonal and Gemination aspects however an incremental effort will be required to map these concepts for a specific language.

1.1 Related work

PLS work for European languages, SI-PRON, a comprehensive pronunciation lexicon of 1.4 m words for Slovenian language has been prepared. Swedish Pronunciation lexicon consisting of 8529 words has been developed. Similar work has been reported for Turkish, named as Finite State Pronunciation Lexicon. Turkish being an agglutinating language with extremely productive inflectional and derivational morphology, It has an essentially infinite lexicon. It takes word form as an input and produces all possible pronunciations.

PLS work for some Indian languages has been initiated using the available base specification. PLS data of 3 lakh words in Bangla and Hindi has already been developed. Similar data for Marathi, Punjabi, Assamese, Manipuri, Bodo is under preparation.

2. PHONETIC AND PHONOLOGICAL ANALYSIS OF PUNJABI

POS Tag set has been standardised for use in PLS as discussed in Paper “Standardization of POS Tag Set for Indian Languages based on XML Internationalization best practices guidelines “by Swaran Lata et al. presented at WILDRE, 2012. The following sections illustrate the prosodic features of Punjabi with the help of examples using IPA for transcription and above referred POS Tag set.

It is proposed to use IPA for presenting the analysis. The IPA charts for Punjabi as drawn from the International IPA used for transcribing the data in this paper is given below:

Consonants

	Bilabial	**Labio-dental**	**Dental**	**Alveolar**	**Post-Alveolar**	**Retroflex**	**Palatal**	**Velar**	**Uvular**	**Glottal**
Plosive	p b p^h		t̪ d̪ $t̪^h$			ʈ ɖ $ʈ^h$		k g k^h	q	
Nasal	m		n			ɳ	ɲ	ŋ		
Trill				r						
Flap						ɽ				
Fricative		v		s z	ʃ			x ɣ		h
Approximant		ʋ								
Lateral Approximant				l		ɭ				
Affricate							tʃ dʒ $tʃ^h$			

Figure 1. Punjabi consonants IPA chart

Vowels

	Front		**Central**		**Back**	
	Short	**Long**	**Short**	**Long**	**Short**	**Long**
Close	I	i			ʊ	u
Close-Mid		e				o
Open-Mid		æ	ə			ɔ
Open				a		

Figure 2. Punjabi vowels IPA chart

2.1 POS based inflection

POS is an important feature of Punjabi. Major parts of speech in Punjabi are ਨਾਂਵ /nãv/ (noun), ਪੜਨਾਂਵ /pƏɽnãv/ (pronoun), ਕਿਰਿਆ /kIrja/ (verb), ਵਿਸ਼ੇਸ਼ਣ /viʃeʃƏɳ/ (adjective), ਕਿਰਿਆ ਵਿਸ਼ੇਸ਼ਣ /kIrja viʃeʃƏɳ/ (adverb), ਸੰਬੰਧਕ /sƏmbƏndƏk/ (preposition), ਯੋਜਕ /jodʒƏk/(conjunction) and ਵਿਸਮਿਕ /vIsmIk/(interjection) etc. Punjabi has a rich base of POS based inflections such as

Word	IPA	POS	Gloss
ਉੱਕਰਵਾਂ	/ʊkkərʋã/	JJ,M,S	engraved, etched
ਉਕਰਵਾਉਣਾ	/ʊkərʋaʊɳa/	VM,M,S	to get engraved, inscribed
ਉਕਰਵਾਈ	/ʊkərʋai/	N,F,S	wages for

2.2 Gemination

Punjabi has an abundance of geminates. As gemination is phonemic in Punjabi and it results in distinctive words with/without gemination, for e.g.

Word	IPA	POS	Gloss
ਦਸ	/dəs/	JJ	Digit Ten
ਦੱਸ	/dəss/	V	To Tell

The words borrowed from English like net, set are pronounced with stress, hence the orthographic representation in Punjabi is done using the germination. E.g.

Word	IPA	POS	Gloss
ਨੈੱਟ	/ nætʈ /	N	material with knotted strings or wire put Table
ਸੈੱਟ	/ sætʈ /	N/V	Group of things/to put something in order

2.3 Tone

Punjabi is highly tonal (Haudricourt, 1971) and three types of tone is found i.e. high-tone /Ó/, low-tone /Ò/ and mid-tone /ō/). There are five tonal consonants discussed in “Challenges for Design of Pronunciation Lexicon Specification (PLS) for Punjabi Language," by Swaran Lata et al.

Word	IPA	POS	TONE	Gloss
ਸਨ	/sən/	V,Aux	Nil	Were
ਸੰਨ	/sə̃n/	N,M	Nil	Year
ਸੰਨ੍ਹ	/sə́n/	N,M	HighTone	Hole made in wall by thieves

2.4 Prolative Vowel

The variation in vowel length in some cases leads to distinct words having different POS.

Word	IPA	POS	Gloss
ਲਮਕਾ	/ləməka/	N,M	Delay
ਲਮਕਾੱ	/ləməkaː/	V	To hang

2.5 Nasalization

Nasalisation is phonemic in Punjabi. Tippi and Bindi are used to represent nasalisation. Functionally both are same however there are some rules in orthography with regard to use of *tippi* and *bindi*. *Tippi* is used only in conjunction with the vowels and *matras* i.e.**[**/ਅ, ਇ, ਿ, ੁ, ੂ] / ə, I, I, U, u / and rest of all full vowels and *matras* uses *bind e.g*:

Word	**IPA**	**Gloss**
ਘਟਾ	/kə́ʈa/	To subtract/decrease
ਘੰਟਾ	/kə̃́ʈa/	Large Bell

3. FRAMEWORK FOR PROSODIC DATA REPRESENTATION IN PLS

The XML schema needs to be evolved which will help in capturing the language specific morphological features in PLS. The proposed XML design will also be targetted towards search optimization of PLS data leveraging the morphological features. To capture the inflections, a new POS attribute is proposed to be added within the <Lexicon> element. POS as an attribute is placed into root element <category> and all the sub-categories and its inflected words of particular pos are placed inside. The attribute prefer is used for most frequently used words. All the inflected variations of the root word are placed together for ease of access by speech systems. The proposed XML will help in data optimization and in enhancing the search optimization. All these word inflections can be factored out into an external PLS document also which is referenced by the <lexicon> element of SSML. Noun, Adjective, Verb, Adverb are taken as lexicon element and the requisite Number, Gender, Person inflection. In some cases same words have different POS attributes; all these types of entries are captured in the proposed XML based on the attributes. In the proposed XML we captured four levels of the POS such as Gender, Number, and Case etc. A User can define more levels according to their requirements. The new categories are shown in shaded/highlighted portion in table1.

Table 1. list of elements used for augmented PLS for Punjabi language

Elements	Attributes	Description
<lexicon>	version xml:base xmlns xml:lang alphabet xml:script	root element for PLS
<meta>	name http-equiv content	element containing meta data
<metadata>		element containing meta data
<lexeme>	xml:id role	the container element for a single lexical entry
<category> for POS Information	POS	Element containing first level of POS eq. Noun, Verb, Adjective etc.
<sub category> for POS Information	Gender, case, number etc.	Element containing second and third level of POS eq. Singular, masculine etc
<grapheme>		contains orthographic information for a lexeme
<phoneme>	prefer alphabet	contains pronunciation information for a lexeme
<alias>	Prefer	contains acronym expansions and orthographic substitutions
<example>		contains an example of the usage for a lexeme

Standard POS tag-set proposed to represent data in PLS format using XML specification will enable a reusable and extendable architecture that would be useful for development of Web based Indian language technologies such as machine translation, cross-lingual information access, Pronunciation Lexicons and other natural language processing technologies.

4. SAMPLE XML REPRESENTATION OF PROSODIC PLS

The following XML examples will serve as a guideline for developing large vocabulary Punjabi PLS data incorporating prosodic features of the language.

4.1 POS Inflections in Punjabi

a) Word Inflection for Number, Gender and Person

i. Inflection for Number change

Word	IPA	POS	Gloss
ਮੁੰਡਾ	/mũda/	N,M,S	Boy
ਮੁੰਡੇ	/mũde/	N,M,Pl	Boys

XML example :

```
<?xml version="1.0" encoding="UTF-8"?>
<lexicon version="1.0"
    xmlns="http://www.w3.org/2005/01/pronunciation-lexicon"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.w3.org/2005/01/pronunciation-lexicon
      http://www.w3.org/TR/2007/CR-pronunciation-lexicon-0071212/pls.xsd"
    alphabet="ipa" xml:lang="pan" xml:script=”guru”>// script tag add here
 <lexeme>
<category pos="N">// verb starts here
<content gender="M”>
<content1 number="S" >
<graphme> ਮੁੰਡਾ </grapheme>
<phoneme prefer=”true’> /mũda/</phoneme>
</content1>
<content2 number="Pl" >
<graphme> ਮੁੰਡੇ </grapheme>
<phoneme prefer=”true’> mũde</phoneme>
</content2>
</content>
<category>
</lexeme>
</lexicon>
```

ii. Inflection for gender change

Word	IPA	POS	Gloss
ਘੋੜਾ	/koɽa/	N,M	Horse
ਘੋੜੀ	/koɽi/	N,F,S	Mare

XML example:

```
<?xml version="1.0" encoding="UTF-8"?>
<lexicon version="1.0" xmlns=http://www.w3.org/2005/01/pronunciation-lexicon
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation=http://www.w3.org/2005/01/pronunciation-lexicon
http://www.w3.org/TR/2007/CR-pronunciation-lexicon-0071212/pls.xsd alphabet="ipa"
xml:lang="pan" xml:script=”guru”>// script tag add here

<lexeme>
<category pos="N">
<content gender="M”>
<graphme> ਘੋੜਾ</grapheme>
<phoneme prefer=”true’> koɽa </phoneme>
</content>
<content gender="F”>
<content1 number="S" >
<graphme> ਘੋੜੀ </grapheme>
<phoneme prefer=”true’> koɽi </phoneme>
</content1>
</content>
<category>
</lexeme>
</lexicon>
```

b) Inflection for person change

Word	IPA	POS	Gloss
ਮੁੰਡੇ	/mŨɖe/	N,M,P	boys
ਮੁੰਡਿਆਂ	/mŨɖIã/	N,M,P	boys
ਮੁੰਡਿਓ	/mŨɖIo/	N,M,P	boys

4.2 Inflection leading to change in POS due to addition of prefixes or suffixes

There are words which change their POS, pronunciation and meaning due to inflections (i.e. addition of prefixes or suffixes). Such inflections with their POS variations will be captured in PLS as given in following example for suffix based inflection :

Word	IPA	POS	Gloss
ਪਹਿਲ	/pél/	N,F	Priority, first step/initiative
ਪਹਿਲਾ	/péla/	JJ,M	First/Primary
ਪਹਿਲੇ	/péle /	JJ,M	First/Foremost
ਪਹਿਲੂ	/pélu/	N,M	Aspect/point of view
ਪਹਿਲਾਂ	/pélã/	RB	Formerly/Before hand
ਪਹਿਲਣ	/péləɳ/	JJ	Calved for the first time

XML example :

```
  <?xml version="1.0" encoding="UTF-8"?>
<lexicon version="1.0"
     xmlns="http://www.w3.org/2005/01/pronunciation-lexicon"
     xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
     xsi:schemaLocation="http://www.w3.org/2005/01/pronunciation-lexicon
       http://www.w3.org/TR/2007/CR-pronunciation-lexicon-0071212/pls.xsd"
     alphabet="ipa" xml:lang="pan" xml:script=”guru”>// script tag add here
  <lexeme>

<category pos="N">
<content gender="F”>
<graphme> ਪਹਿਲ</grapheme>
<phoneme prefer=”true’> pél </phoneme>
</content>
<content gender="M”>
<graphme> ਪਹਿਲੂ </grapheme>
<phoneme prefer=”true’> pélu </phoneme>
</content>
<category>
<category pos="JJ">
<graphme> ਪਹਿਲਣ </grapheme>
<phoneme prefer=”true’> pélən̪ </phoneme>
<content gender="M”>
<graphme> ਪਹਿਲਾ </grapheme>
<phoneme prefer=”true’> péla </phoneme>
<graphme> ਪਹਿਲੇ </grapheme>
<phoneme prefer=”true’> péle </phoneme>
</content>
<category>
<category pos="RB">
<graphme> ਪਹਿਲਾਂ </grapheme>
<phoneme prefer=”true’> pélã̃ </phoneme>
<category>
</lexeme>
</lexicon>
```

4.3 Vowel Lengthening

Graphemes to Phoneme conversion are relatively direct for some languages, while it can be highly unpredictable for others, like English and some of the Indian Languages. In Punjabi same grapheme information may have different pronunciation based on its Part of Speech information and its semantic e.g.

Word	IPA	POS	Gloss
ਹਰਾ	/həra/	N	Green
ਹਰਾ	/həra:/	V	To defeat

```
<?xml version="1.0" encoding="UTF-8"?>
<lexicon version="1.0"
    xmlns="http://www.w3.org/2005/01/pronunciation-lexicon"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.w3.org/2005/01/pronunciation-lexicon
      http://www.w3.org/TR/2007/CR-pronunciation-lexicon-0071212/pls.xsd"
    alphabet="ipa" xml:lang="pan" xml:script=”guru”>// script tag add here
 <lexeme>
<category pos="N">
<graphme> ਹਰਾ </grapheme>
<phoneme prefer=”true’> həra </phoneme>
</category>
<category pos="V">
<graphme> ਹਰਾ </grapheme>
<phoneme prefer=”true’> həra: </phoneme>
</category>
</lexeme>
</lexicon>
```

4.4 Nasalization

If a nasal consonant or a nasal sound occurs at the end of the word then it affects the previous vowel i.e. the previous vowel gets nasalized. This phenomenon is predominant in Punjabi and needs to be captured in PLS e.g.:

Word	IPA	IPA (PLS)	POS	Gloss
ਆਇਆਂ	/aiã/	/ ãiã/	V	Welcome
ਗਮ	/gəm/	/gə̃m/	N	Sorrow
ਜਾਣਾ	/ʤaɳa/	/ʤãɳa/	VM	To go

```
<?xml version="1.0" encoding="UTF-8"?>
<lexicon version="1.0"
    xmlns="http://www.w3.org/2005/01/pronunciation-lexicon"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.w3.org/2005/01/pronunciation-lexicon
      http://www.w3.org/TR/2007/CR-pronunciation-lexicon-0071212/pls.xsd"
    alphabet="ipa" xml:lang="pan" xml:script=”guru”>// script tag add here
 <lexeme>
<category pos="N">
<graphme> ਗਮ </grapheme>
<phoneme prefer=”true’> gə̃m </phoneme>
</category>
```

<category pos="V">
<graphme> ਆਇਆਂ </grapheme>
<phoneme prefer=”true’> ãiã </phoneme>
<category1 pos="VM">\
<graphme> ਜਾਣਾ</grapheme>
<phoneme prefer=”true’> ʤãɳa </phoneme>
</category1>
</category>
</lexeme>
</lexicon>

4.5. Homographs

a) Same spelling and pronunciation but different POS

Word	IPA	POS	Gloss
ਕੰਨੀ	/kə̃ni/	N,F	Edge/Border
ਕੰਨੀ	/kə̃ni/	RB	On the side of/By or with ear

b) Same spelling but different pronunciation and POS

Word	IPA	POS	Gloss
ਹਰਾ	/hərа/	N	Green
ਹਰਾ	/hərа:/	V	To defeat.

4.6. Multiple Spellings and Pronunciation

There are words which have more than one spelling and thus different pronunciation. When both or all the varieties are equally and frequently used, then we have to decide whether to keep both or all the forms or a single standard one in the PLS data. It is important to capture such variations especially the frequently used ones. In case of Punjabi PLS we have kept all the forms and the standard one is marked as “true”. The standard pronunciation can be used by language learners.

<?xml version="1.0" encoding="UTF-8"?>
<lexicon version="1.0" xmlns="http://www.w3.org/2005/01/pronunciation-lexicon"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.w3.org/2005/01/pronunciation-lexicon
http://www.w3.org/TR/2007/CR-pronunciation-lexicon-20071212/pls.xsd"
alphabet="ipa" xml:lang="en-US">
<lexeme>
<graphme>ਗੁਰਦਵਾਰਾ</grapheme>
<phoneme prefer=”true’>gUrdwara</phoneme>
<graphme>ਗੁਰਦੁਆਰਾ</grapheme>
<phoneme> gUrdUara</phoneme>
</lexeme>
</lexicon>

4.7 Borrowed Words

Native speakers are not phonetically trained, so they cannot speak borrowed words properly. They assimilate and variations occur while borrowing words from different languages. Like when Punjabi borrows word from other languages it changes its gender or other categories according to its nature or behaviour. Punjabi language has borrowed extensively from other languages, including Sanskrit, Hindi, Urdu, Persian and English.

Words with Nukta borrowed from Urdu

Punjabi speakers find it difficult to pronounce words with Nukta, which are borrowed from Urdu like [زمانہ] /zəmana/ is spoken as [ਜਮਾਨਾ] (jamaanaa)/ʤəmana/ in Punjabi. These words poses a challenge in building PLS for Punjabi language, in deciding which pronunciation should be kept in the database, either or both.

Word	IPA	Gloss
زمانہ	zəmana	Specific Period referred in a context
ਜਮਾਨਾ	ʤəmana	Specific Period referred in a context

5. XML BASED PLS DATA IMPLEMENTATION

POS is a readily available source for feature extraction as is evident from above examples. POS based PLS with overriding phonological features of a language such as stress, tone, gemination, nasalization etc. can be used for machine learning of prosodic features. The large PLS data of phonetically rich words can be a useful resource for training of speech systems.

```
<? xml version="1.0" encoding="UTF-8"?>
<lexicon version="1.0"
    xmlns="http://www.w3.org/2005/01/pronunciation-lexicon"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.w3.org/2005/01/pronunciation-lexicon
      http://www.w3.org/TR/2007/CR-pronunciation-lexicon-0071212/pls.xsd"
    alphabet="ipa" xml:lang="pan" xml:script=”guru”>// script tag add here
 <lexeme>

<category pos="verb">// verb starts here
<content gender="Masculine” Transitivity=”Intransitive”>
<content1 number="Singular" > // Singular starts here
<content2 phase="non-perfect"> //non perfect phase starts here
 <grapheme> ਊਂਘ  </grapheme>
<phoneme> ũg </phoneme>
<grapheme> ਊਂਘਣਾ  </grapheme>
<phoneme> ũgŋa </phoneme>
<grapheme> ਊਂਘਣੀ </grapheme>
<phoneme> ũgŋi </phoneme>
```

</content2> // **non perfect phase close here**

<content2 case="perfect"> // **perfect phase starts here**
<grapheme> ਊਂਘਿਆ </grapheme>
<phoneme> ũgIa</phoneme>
</content2> // **perfect phase close here**
</content1> // **singular close here**

<content1 number="Plural"> // **plural starts here**
<content2 phase="non-perfect">// **non perfect phase starts here**
<grapheme> ਊਂਘਣ </grapheme>
<phoneme> ũgŋ</phoneme>
<grapheme> ਊਂਘਦ </grapheme>
<phoneme> ũgd</phoneme>
<grapheme> ਊਂਘਦਿਓ </grapheme>
<phoneme> ũgdIo</phoneme>
<grapheme> ਊਂਘਾਂਗੇ </grapheme>
<phoneme>ũgãge </phoneme>
</content2>// **non perfect phase close here**
<content1> // **plural close here**
</content> // **Masculine close here**

<content gender="Feminine" Transitivity="Intransitive">

<content1 number="Singular" > // **Singular starts here**
<content2 phase="non-perfect"> // **non-perfect starts here**
<grapheme> ਊਂਘਣੀ </grapheme>
<phoneme>ũgŋi</phoneme>
<grapheme> ਊਂਘਾਂਗੀ </grapheme>
<phoneme>ũgãgi</phoneme>
</content2> // **non-perfect close here**

<content2 phase="perfect"> // **perfect starts here**
<grapheme> ਊਂਘੀ </grapheme>
<phoneme> ũgi </phoneme>
</content2> // **perfect close here**
</content1> // **singular close here**
<content1 number="Plural" > // **plural starts here**
<content2 phase="non-perfect"> //**non-perfect starts here**
<grapheme> ਊਂਘਣੀਆਂ </grapheme>
<phoneme> ũgŋiã </phoneme>
<grapheme> ਊਂਘਦੀਓ </grapheme>
<phoneme> ũgdio </phoneme>
</content2> //**non-perfect close here**

<content2 phase="perfect"> // **perfect starts here**

<grapheme> ਉਂਘਿਆਂ </grapheme>
<phoneme> ũgiã </phoneme>
</content2> //**perfect close here**
</content1>
</content> **// Feminine close here**
</category> **// Verb Pos category close here**

<category pos="noun">// **Noun Pos category starts here**
<content gender="Masculine”>
<content1 number="Singular" > **// Singular starts here**
<content2 case="oblique"> **//oblique case starts here**
<grapheme> ਅਮਲ਼ </grapheme>
<phoneme>əməl </phoneme>//**frequently used**
</content2>

<content2 case="direct"> **//direct case starts here**
<grapheme> ਅਮਲ਼ਾ </grapheme>
<phoneme>əmələ </phoneme>//**frequently used**
</content2>

<content2 case="ablative"> **//ablative case starts here**
<grapheme> ਅਮਲ਼ਿਓ </grapheme>// **not frequently used**
<phoneme> əməlIo </phoneme>
</content2>
</content1>

<content1 number="plural" > **// Plural starts here**
<content2 case="direct"> **//direct case starts here**
<grapheme> ਅਮਲ਼ </grapheme>
<phoneme prefer=1> əməl </phoneme>//**frequently used**
</content2>
<content2 case="oblique"> **//oblique case starts here**
<grapheme> ਅਮਲ਼ਿਆਂ </grapheme>
<phoneme> əməlIã </phoneme>//**not frequently used**
</content2>
</content1>
</content>
</category> // Noun Pos category close here

<category pos="Adjective"> // Adjective Pos category close here
<content gender="both”>
<content1 number="Singular" > **// Singular starts here**
<content2 case="direct"> **//direct case starts here**
<grapheme> ਉਜੱਡ </grapheme>
<phoneme> ʊdʒdʒəd̪ </phoneme>//**frequently used**
</content2>

<content2 case="vocative"> **//vocative case starts here**

<grapheme> ਉਜੱਡਾ </grapheme>
<phoneme> ʊdʒdʒəɖa </phoneme>
</content2>
</content1>

<content1 number="Plural" > // **Plural starts here**
<content2 case="oblique"> //**Oblique case starts here**
<grapheme> ਉਜੱਡਾਂ </grapheme>
<phoneme> ʊdʒdʒəɖã </phoneme>

<content2 case="vocative"> //**Vocative case starts here**
<grapheme> ਉਜੱਡੇ </grapheme>
<phoneme>ʊdʒdʒəɖo </phoneme>
</content2>
</content1>
</content>
</category>// Adjective POS ends here
<category pos="Adverb">// adverb POS starts here
<content gender="both”>
<content1 number="Singular" > // **Singular starts here**
<content2 case="direct"> //**direct case starts here**
<grapheme> ਅੰਦਰ </grapheme>
<phoneme>ə̃dər</phoneme>
</content2>

<content2 case=" Ablative"> // **Ablative case starts here**
<grapheme> ਅੰਦਰੋਂ </grapheme>
<phoneme>ə̃dərõ </phoneme>
</content2>

<content2 case="Locative"> // **Locative case starts here**
<grapheme> ਅੰਦਰੇ </grapheme>
<phoneme> ə̃dəre</phoneme>
</content2>
</content1>

<content1 number="Plural" > // **Plural starts here**
<content2 case="Oblique"> //**Oblique case starts here**
<grapheme> ਅੰਦਰਾਂ </grapheme>
<phoneme> ə̃dərã</phoneme>
</content2>

<content2 case="Locative"> //**Locative case starts here**
<grapheme> ਅੰਦਰੀਂ </grapheme>
<phoneme>ə̃dərĩ </phoneme>
</content2>
</content1>
</content>

</category>// **Adverb POS ends here**
</lexeme>
</lexicon>

6. CONCLUSION & FUTURE WORK

A unified augmented PLS framework for Indo Aryan languages has been proposed in this paper, which provides broad guidelines and criterion for generating prosodically rich PLS. The Punjabi specific features can be substituted for a specific language of this family by incorporating phonetic and phonological nuances of that language.

Prosodically rich PLS data can be created by taking due care in capturing all possible POS based variations. The data thus generated using the proposed Standards POS Tag-set will be interoperable across various versions of PLS. This data can be used by language researchers and can also be utilised for defining Grapheme to Phoneme conversion rules in addition to the voice browsing applications. The detailed analysis can be used for incorporating morphological features in the future W3C PLS recommendations which would aid building Multilingual voice based search systems in Indian Languages in the near future.

The paper provides broad guidelines and criterion for generating Prosodically rich Punjabi PLS data by taking due care in capturing all possible POS based variations and implementation of PLS data by introducing new elements in XML format. The word list of 3000 to 5000 root words for Punjabi from major POS categories such as Noun, Verb, Adjective, Adverb and other granular features may be collated along with their POS variations and a PLS document of 10,000 words can be created which can serve as a useful resource for TTS developers for improving the naturalness of TTS output and also for building automatic speech recognition engines. This data can also be used for machine learning and voice based search systems and browsers in Indian languages.

7. ACKNOWLEDGEMENT

The authors would like to thank Dept of Linguistics at Jawaharlal Nehru University, New Delhi for useful technical feedback during the work. The authors also would like to thank Department of Electronics & Information Technology, Govt. of India for providing infrastructure support.

REFERENCES

[1] Bailey T Grahme (1914), A Punjabi Phonetic Reader, London.
[2] Banerjee Esha, Kaushik Shiv, Nainwani Pinkey, Bansal Akanksha, Jha, Girish Nath(2013), Linking and Referencing Multi-lingual corpora in Indian languages, in proceedings of the 6th LTC, Zygmunt Vetulani & Hans Uszkoreit (ed), pp 65-68, Fundacja, Uniwersytetu im. A. Mickiewicza, Poznan, Poland, 2013.
[3] Gros J.Z (2006), SI-PRON Pronuntiation Lexicon : A new language resource for Slovenia, Informatica
[4] Das Mandal Shayamal, Chandra Somnath, Lata Swaran. (2010), Use of Parts of Speech (POS and morphological information for resolving multiple PLS in Indian languages- Bengali as a case study, USA: W3C workshop on conversational applications, USA.

[5] Oflazer, The Architecture and the Implementation of a Finite State Pronunciation Lexicon for Turkish.
[6] BIS Standard IS 13194: 1991 (http://varamozhi.sourceforge.net/iscii91.pdf)
[7] Gill Harjeet Singh and Henry A. Gleason (1969), A Reference Grammar of Punjabi, Department of Linguistics, Punjabi University, Patiala.
[8] Haudricourt A.G. (1971), Tones in Punjabi. Paris: C.N.R.S.
[9] Hirschberg Julia (2000), The complexity of predicting prosodic boundary locations with the help of dependency trees.
[10] Lata Swaran (2012), Somnath Chandra and Swati Arora, Standardization of POS Tag Set for Indian Languages based on XML Internationalization best practices guidelines, WILDRE 2012.
[11] Lata Swaran (2011), Challenges for Design of Pronunciation Lexicon Specification (PLS) for Punjabi Language, LTC 2011.
[12] Lata Swaran (2012), Exploratory Analysis of Punjabi Tones in relation to orthographic characters: A Case Study, WILDRE 2012.
[13] Pandey Pramod(2014), Sounds & their Patterns in Indic Languages.
[14] POS tag set:http://tdil-dc.in/tdildcMain/articles/780732Draft%20POS%20Tag%20standard.pdf
[15] Singh Brar Buta (2008), Punjabi Vyakaran, Siddhant Ate Vihar, Punjabi University, Regional Center, Bhatinda.
[16] Singh Khaira Surinder (2011), Punjabi Bhasha: Viyakaran and Banter, Punjabi University Patiala.
[17] Singh Puar Joginder (1990), The Punjabi Verb Form and Function, Publication Bureau, Punjabi University, Patiala.
[18] Singh Sandhu Balbir (1986), The Articulatory & acoustic structure of the Punjabi consonants, Punjabi University Patiala.
[19] Singh Chander Shekhar (2001), Punjabi Prosody: The old Tradition & The new Paradigm.
[20] Singh Harkirat (1991), Punjabi Diyan bhashai Visheshtawan, Publication Bureau, Punjabi University, Patiala.
[21] Singh Harkeerat (1988), Punjabi Baare, Punjabi University, Patiala.
[22] Singh Joginder (2010), Bhashavigian: Sankalp Ate Dishavan, Punjabi Bhasha Akademi, Jalandhar.
[23] Singh Dr. Atam (1993), Linguistics, Punjab State University, Chandigarh.
[24] Singh Dr. Premprakash (2010), Sidhantik Bhasha Vigyan.
[25] Talukdar Pran Hari 2010, Machine learning methods, Probabilistic methods or manually written rules.
[26] Talukdar Pran Hari. (2010), Speech Production, Analysis and Coding: Introduction to Speech Processing, LAP LAMBERT Academic Publishing.
[27] Hockett F. Charles, A course in Modern Linguistics , The macmillan company
[28] Punjabi Morphological analyzer and generator, Advanced Centre for Technical Development of Punjabi Language, Literature and Culture http://www.learnpunjabi.org/punjabi_mor_ana.asp

Author

Swaran Lata : Centre of Linguistics, JNU, New Mehrauli Road, New Delhi,India.