

# Vision and Roadmap for PLS Standardization in Indic Languages

**By :**

**Swaran Lata, Country Manager , W3C India & Head , TDIL Programme**

**Dept of Information Technology , Govt of India**

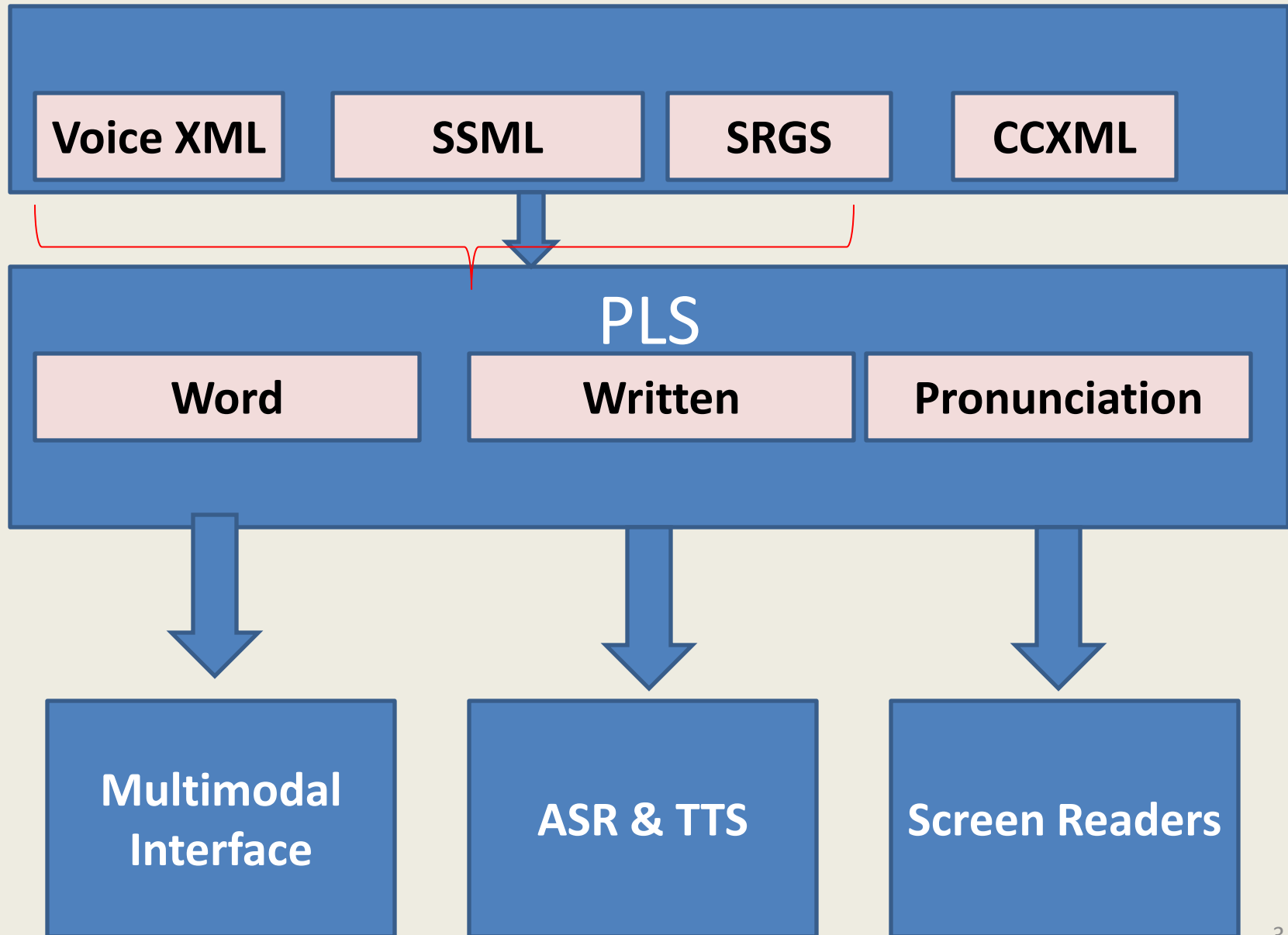
**E-mail : [swaran@w3.org](mailto:swaran@w3.org)**

**[slata@mit.gov.in](mailto:slata@mit.gov.in)**

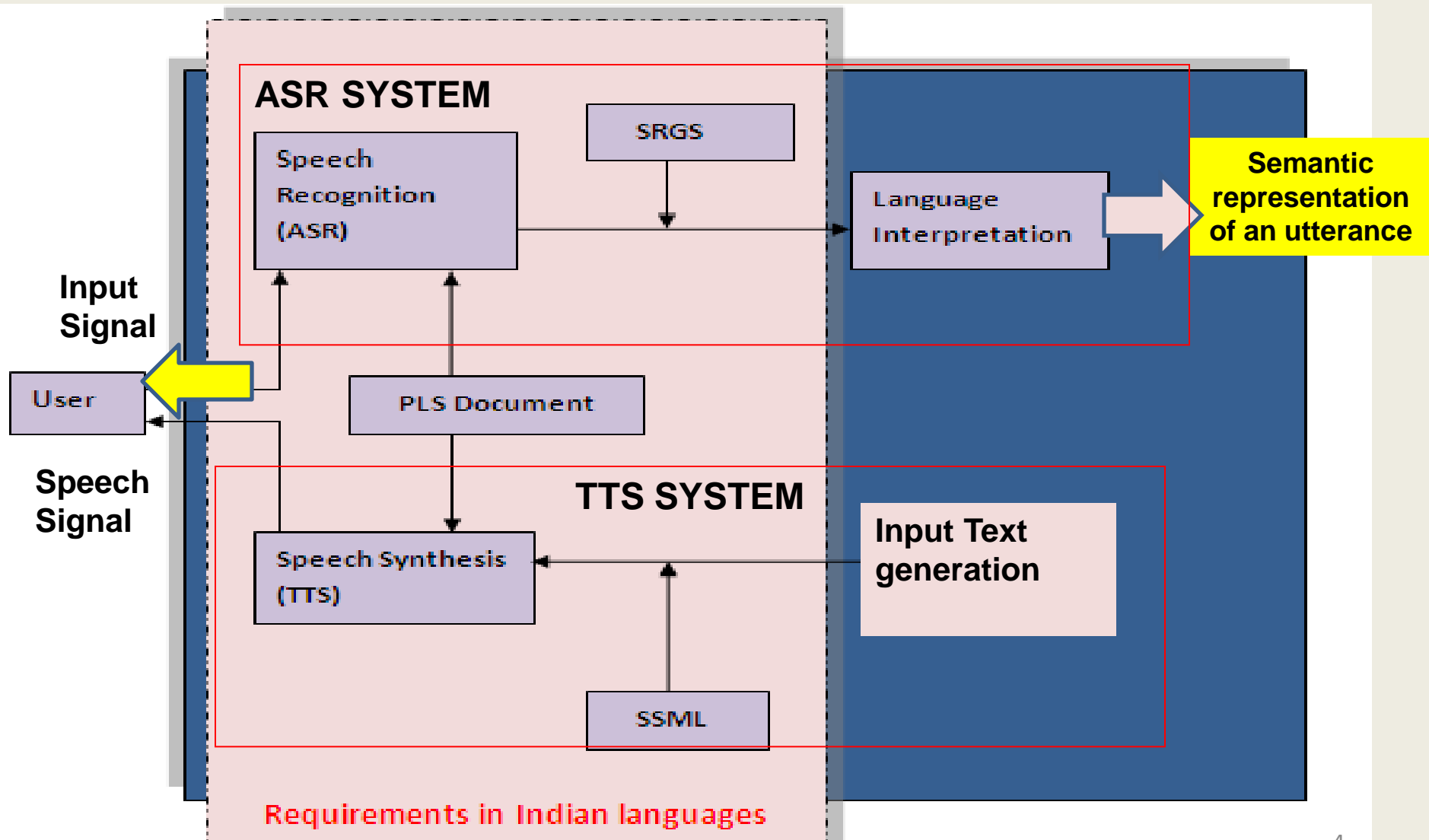
# What is PLS ?

- Interoperable specification of pronunciation information for ASR & TTS engines
- Provides mapping between alphabets , words/phrases , their written representation and pronunciation
- It can also cover pronunciation/spelling variations based on orthography/ Country/region/person etc.

# Speech Interface Framework



# Indian Language requirements



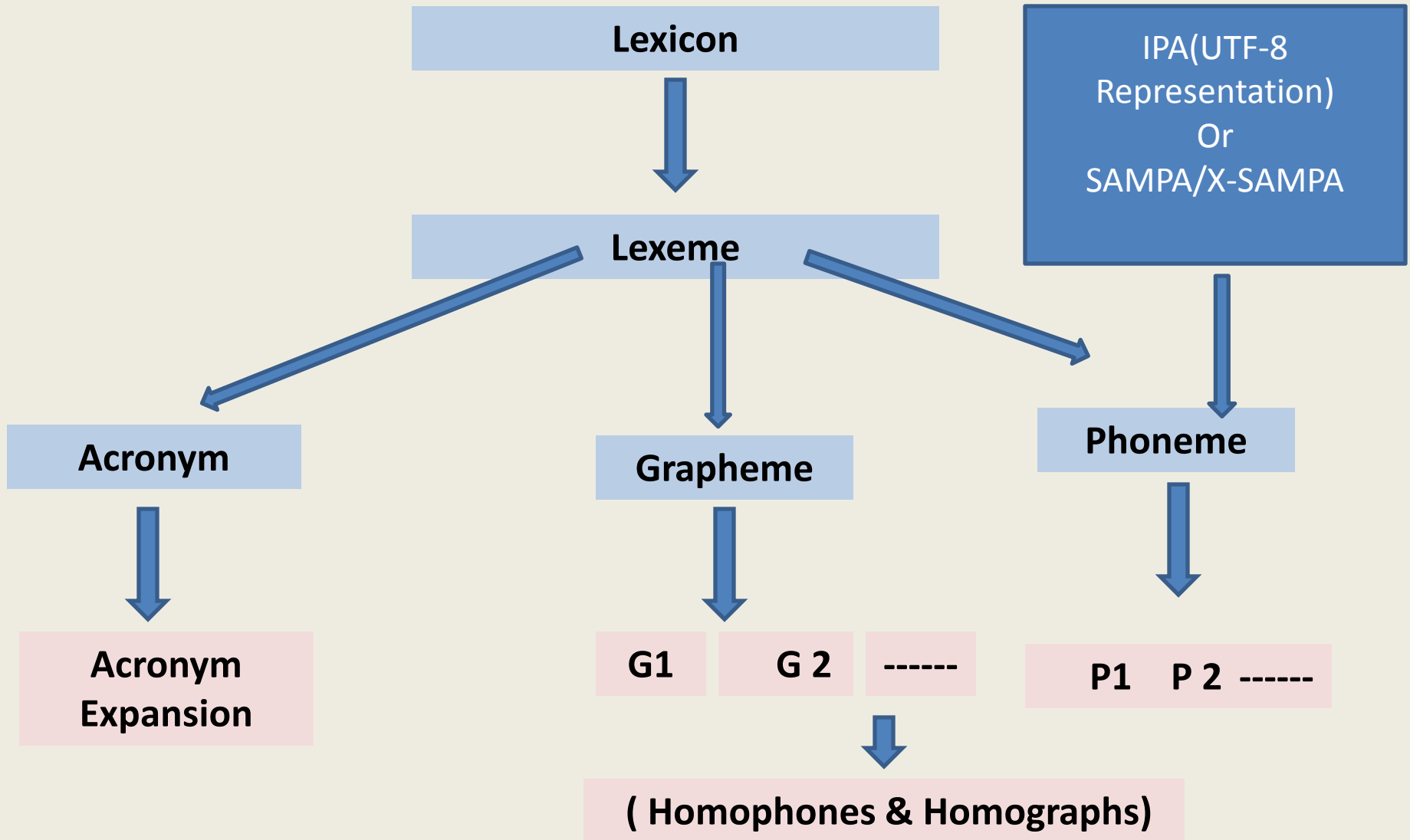
# PLS Definition

Elements	Attributes	Description
*<lexicon>	<ul style="list-style-type: none"><li>•xml:lang</li><li>•Role</li></ul>	Root element of PLS(container element)  <b>Role</b> : describes additional information to help the selection of the most relevant pronunciation for a given <a href="#">orthography</a> .
*<meta>	Name http equiv content	Element containing metadata
*<lexeme>	Xml:id	Contains element for single lexical entry
*<morpheme>		Smallest component of a word that has semantic meaning
*<grapheme>		Describing orthographic of the : 1) Homophones  2) Homographes  3) Acronyms expansion

# PLS Definition

Elements	Attributes	Description
*<alias>	Prefer	Contain acronym expansions and orthographic substitution
*<phoneme>	Prefer alphabet	Contain pronunciation information for a lexeme
<u>&lt;example&gt;</u>		contains an example of the usage for a <u>lexeme</u>

# PLS Structure



# PLS Document

<?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="hin">

PLS  
Header

<metadata>

-title

-Creator

-date

-Description

-format

-subject

</metadata>

Dublin Core Metadata  
element set(DS-ES)

←-----lexeme entry-----→

</lexicon>

# Definition– <lexeme>

- The **<lexeme>** element is the container of a lexicon entry. It is composed of:
  - One or more **<grapheme>** elements that indicate the words/phrases to be matched in the input
  - One or more either **<phoneme>** or **<alias>** elements that indicate the possible pronunciations or expansions respectively
- First considerations:
  - More **<grapheme>** elements may be present  
→ this means that all of them will match the pronunciations
  - More **<phoneme>** elements may be present  
→ this means that several pronunciations are in alternative
  - A mixture of **<alias>** and **<phoneme>** elements may be present  
→ there is a preference mechanism to choose the single one for TTS

# Definition– <grapheme>

The <grapheme> element contains DATA that represents orthographies:

- Regional spelling variations
- Free spelling variations e.g. "हिंदी" and "हिन्दी"
- Homographs :  
words with different meanings but the same spelling (and sometimes different pronunciations), called [homographs](#)
- Homophones  
words with the same pronunciation but different meanings (and possibly different spellings)

# Definition– <alias>

- The <alias> elements are contained inside <lexeme>
- <alias> is used to indicate the pronunciation of an acronym or an abbreviated term in the form of other orthographies.
- <alias> may contain
  - A “prefer” attribute to indicate precedence among pronunciations
- Both <phoneme> and <alias> may occur in a <lexeme>

## Example of lexeme with both <phoneme> and <alias>:

```
<?xml version="1.0" encoding="UTF-8"?>
<lexicon version="1.0" xmlns="http://www.w3.org/2005/pronunciation-lexicon"
  alphabet="ipa" xml:lang="en">
  <lexeme>
    <grapheme>W3C</grapheme>
    <alias>World Wide Web Consortium</alias>
  </lexeme>
</lexicon>
```

# Example- Homophones

Example 1 :

<lexicon>

**Homophones**

<lexeme> (same pronunciation but different meanings )

<grapheme meaning="mare"> घोड़ी </grapheme>

<grapheme meaning="wooden stand"> घोड़ी</grapheme>

<phoneme>IPA Representation घोड़ी </phoneme>

<phoneme>IPA Representation घोड़ी</phoneme>

</lexeme>

</lexicon>

# Example- Homographs

Example 2 :

<lexicon>

<lexeme> (different meanings but the same spelling )

<grapheme meaning =“सोना”> कनक </grapheme> Homographs

<grapheme meaning =“गेहूँ”> कनक </grapheme>

<phoneme>IPA Representation for कनक </phoneme>

<phoneme>IPA Representation for कनक </phoneme>

</lexeme>

</lexicon>

# An SSML with PLS

```
<?xml version= "1.0" encoding = "utf-8" >
```

```
<speak version= "1.0" xmlns=
```

```
http://www.w3.org/2001/10/synthesis xml:lang= "hi">
```

## PLS document

```
<lexicon url= http://www.example.com/pl\_lexicon.pls/>
```

```
</speak>
```

- PLS factorize all the changes in external document.
- TTS engine loads the PLS documents and applies it transparently to the SSML document

# Future updations of PLS using RDF

- Morphological –word stems
- Syntactic and Semantic Information- Inter-word semantic links
- Pronunciation statistics

# Pros of using IPA

Pros :

- Vowels & Consonants symbols
- Syllable delimiter
  - Diacritics
  - stress symbols
  - Lexical tone symbols, International markers
- Access white space characters to enhance readability
- IPA to Unicode mapping

# Issues

- Which IPA/SAMPA/X-SAMPA?
- Reference phoneme set
- Problems of allophonic symbols and IPA is used
  - Inconsistencies between lexicons for identical language
- Standardize metadata
- Collect list of Acronyms
- Decision on Most Representation Word List
- Orthographic variation coverage
- How to port existing data to PLS?

Thanks