3.4 Indian Standard Lexware Format

Preamble

It is today an accepted fact that the lexicon is the source of power in any language processing system. The more well structured and knowledge rich the lexicon, the better is the quality of the analysis and the generation of linguistic data. The coverage of the whole gamut of phenomena in any natural language needs constructing accurate and vast lexicons of fine quality. This is often a very large task needing quality manpower in large number.

In a country like India with multiple languages, the need for the standardization of the Lexware format was long felt so that this resource can be developed in a collaborative manner. Enabling of Indian languages on the internet and the localization of Information Technology has been taken up as a priority task by the Ministry of Communications & Information Technology (MCIT) in India. The idea is to be able to pull the resources of lexicon building activity going on at various places and in various languages. If all these efforts follow a simple and standard format, the collaborative development of this very important linguistic resource will proceed efficiently. Also the future linkages of various language lexicons will be easier.

Basic Concepts

For any language processing system, the three most important Lexical resources are

1. Multilingual Lexicon
2. Language Word Net
3. Ontology

The multilingual lexicon defines the mapping of words from one language to another. The WordNet is a massive network which connects words with various semantic relations. The Ontology is a categorization of concepts, which is language independent, and guides the various decisions and choices in lexicon and WordNet building.

In this document we primarily discuss the standardization of the lexicon—the entities in it, the format of the entries and the linkage of the entries with the WordNet and the Ontology. The diagram in Lex. 1 illustrates the various entities that must be present in a lexicon.

Lex.1: Essential Entries in the Lexicon

Here the language specific strings map to ‘concepts’ which are language independent. Thus dog(a-kind-of animal) along with other word-net attributes represents a language independent concept which is linked to kuttaa in Hindi, kukur in Bengali, and so on. Here dog(a-kind-of animal) along with other word-net attributes belongs to the space of concepts while kuttaa or kukur belongs to the space of language specific strings.

The space of attributes is of great importance since the flags in the space guide the analysis and generation processes. The attributes are of two kinds—syntactic and semantic. The part of speech and the morphological behavior belong to the set of syntactic attributes, while the information like animate/inanimate, perishable, eatable etc. belongs to the set of semantic attributes. While the former is a well-defined set, the latter is quite open-ended and depends on the domain and the application envisaged. However, some semantic attributes seem to be frequently used across the domains and they also can looked upon as a standard set. Lex.2 depicts the inter-relationships.

It is reemphasized in this document that our standardization effort recognizes the supreme importance of the semantic attributes. Though open ended, the lexicon building activity can be accelerated by setting up the ontology. Thus semantic attributes connect the lexicon space with the ontology space.

The disambiguation constructs like a-kind-of, part-of, has-part etc. are semantic relations with other words. Thus disambiguation set connects the lexicon space with the WordNet space.
Having described these basic ideas we now proceed to the recommendation for standardization.

**Standardization**

The nomenclature used in the lexicon is the following:

- **Root Word**: Word from which the various forms are generated.
- **Morphology Paradigm Number**: An integer specifying the table in which the morphological transformation rule of the root are specified. The table specifies the strings that must be attached depending on the subject object and/or other case related information. For example, one such paradigmatic information could be ladakA (boy) in Hindi takes the string oM ne after deleting A, in plural number and past tense for a transitive verb.
- **Domain**: Used to indicate the domain to which a particular meaning of the word belongs. For example – G is used for general, –IT for information technology, –H for health and so on.
- **English Meaning**: English meaning of the word followed by an example sentence that implicates the meaning.
- **Syntactic Information**: This is part of speech, morphology paradigm and such other information. Also are included disambiguation rules for verbs. See below (verb pattern scheme).

**Semantic Tag**: Tags or disambiguation rules for each English sense of the word. These tags are of two kinds:
The above syntactic information are stored along with the verb entries in the lexicon to facilitate disambiguation.

**Semantic Information used for the Verbs**

(Used only when there are multiple meanings)

- **sth**(to sb) .......... something(to somebody)
- **sb**(with sth) .......... somebody(with something)
- **sth**(from sb) .......... something(from somebody)
- **sth**(for sth) .......... something(for something)
- **sb/sth**(with sth) ...... somebody/something (with something)

**Examples**

**For Noun**

**Line 1:** Root Word

**Line 2:** Morphology-Paradigm

**Line 3:** ~Domain1 English Meaning1 [Example Sentence]; ~Domain2 English Meaning2 [Example sentence]; ....

**Line 4:** hypernymy1[meronymy1]; hypernymy2[meronymy2]; ....

**Line 5:** ~Hindi Meaning1: m/f/n paradigm no, ontology tags1; Hindi Meaning2: m/f/n paradigm no, ontology tags2; ....

**Line 6:** ~AS (Assamese meaning with other informations)

**Line 7:** ~( Similarly meanings for other Indian Languages can be entered in separate lines one after another)

**Example for noun**

**Line 1:** head /English root word/

**Line 2:** 65 noun/Table number with morphological and categorical information/

---

**Examples**

**For Noun**

- The court of his innocence.
- The nurse shook the medicine up.
- Officials believe that a settlement is possible.
- We had not decided what we ought to do next/what to do next.
- Mary hates to drive in the rush-hour.
- We dread Mary/Mary's taking over the business.
- She spotted a man waving in the crowd.
- We watched the men unpack the china.
- The fridge keeps the beer cool.
- The court considered Smith a trustworthy witness.
- The police didn't accept the story as true (or as the fact).
- The thief forced Sita to hand over the money.
- The policeman got the traffic moving.
- Mother won't let the children play in the road.
- Henri taught the children French.
- Henri taught French to the Children children.
- Colleagues told Paul that the job would not be easy.
- Employers announced to journalists that the dispute had been settled.
~G part of the body containing the eyes, nose, mouth and brain
[He fell and hit his head];
~ADM chief person of a group or organisation
[Report to the Head immediately];
~FIN accounts head
[Into which head should I put the given expenditure];
~G head of the coin
[We tossed a coin and it came down heads]

Different English senses of the root word/

~HN sira:m 6, inanimate, concrete;
aXyakRa:m 6, animate, concrete;
SIrRaka:m 6, inanimate, abstract;
ciwwa:m 8, inanimate, concrete
/Hindi meanings with gender information, paradigm type and ontology/

Note: It may be noted that for machine translation task, we avoid putting too many meanings in the lexical data-base unless these can be disambiguated. In this sense, it may be advisable to keep only two meanings as 'sira' and 'hedh' (since the English word 'head' in all other contexts have been frequently used as it is in Hindi).

For Verb

Line 1: Root Word
Line 2: Morphology-Paradigm No verb
Line 3: ~Domain1 English Meaning1 [Example Sentence];
~Domain2 English Meaning2 [Example Sentence];

Line 4: hypernymy1[, Syntactic Info1][, Semantic Tag1 for Subject][, Semantic Tag1 for Object1][, Semantic Tag2 for Subject][, Semantic Tag2 for Object2];

Line 5: ~HN Hindi Meaning1: Paradigm-No,
[Vibhakti Parasarg Info for Subject][,Vibhakti Prasarg Info for Object], [ontology tags];
(Similarly meanings in other Indian Languages can be entered in lines below)

Example for verb

Line 1: murder /English root word/
Line 2: 16 verb /Table of morphological and categorical information/
Line 3: ~G kill somebody unlawfully and intentionally
[He murdered her with a knife];....
/Different English senses of the root word/
Line 4: kill, T n, T n. pr, l, sb, sb- with- sth, [ ];[ ];
/Disambiguation rules for disambiguating different senses of the verb/
Line 5: ~HN hawtyaA kara:257 ne, ko, VoA; /Hindi meaning with subject and object
Vibhakti parasarga and ontology tags/

For Adjective

Line 1: Root Word
Line 2: Paradigm-No adjective
Line 3: ~Domain1 English Meaning1 [Example Sentence];
~Domain2 English Meaning2 [Example Sentence];

Line 4: hypernymy1[, antonymy1]; hypernymy2[, antonymy2];

Line 5: ~HN Hindi Meaning1, ontology tags1;
Hindi Meaning2, ontology tags2;...

Line 6: ~ (Similarly meanings in other Indian Languages can be entered in lines below)

Example for adjective

Line 1: mysterious / English root word /
Line 2: 12 adjective /Table of morphological and categorical information/
Line 3: ~G difficult to understand or explain [She gave me a mysterious look];/Different English senses of the root word/
Line 4: state, clear /Wordnet Tag/
Line 5: ~HN hrahasamaya, Hindi Meaning/

For Adverb

Line 1: Root Word
Line 2: Morphology-Paradigm No adverb
Line 3: ~Domain English Meaning1[Example Sentence];
~Domain English Meaning2 [Example Sentence];

Line 4: hypernymy1[, synonymy1]; hypernymy2[, synonymy2];

Line 5: ~HN Hindi Meaning1[, ontology tags1];
Example for adverb
Line 1: mysteriously /English root word/
Line 2: 13 adverb/Table of morphological and
categorical information /
Line 3: ~G mysteriously [The main witness had
mysteriously disappeared];
/Different English senses of the root word/
Line 4: manner /WordNet Hypernymy/
Line 5: ~H N rahasyamaya DaM ga se, /Hindi
Meaning/

Semantic Attributes For Lexware Standard

Noun
- Proper Noun (PROP, eg, rAma)
- Common Noun (If a noun is neither Proper nor
Collective then it is taken for granted that it is
Common Noun and hence, no symbol is kept
for it.)
- Collective Noun (COLCT, eg, BidZa)
- Animate (ANIM T)
  - Flora (FLORA)
    - Shrubs (FLORA-SHRB , eg, tulasI)*
    - Aquatic plants (FLORA-AQTC, eg kamala)
    - Climbers (FLORA-CLMB , eg , aMgura kI
bea)
  - Trees (FLORA-TREE, eg, Ama)
- Fauna (FAUNA)
  - Mammals
    - Person (ANIM T-FAUNA-M M L-PRSN, eg
ladZakI)
    - Ape (ANIM T-FAUNA-M M L-APE ,eg, laM
gura)
    - Lesser Mammals (ANIM T-FAUNA-
M M L-LSM M L, eg dAzlPZina)
  - Reptiles (ANIM T-FAUNA-RPTL, eg,
sAM zpa)
  - Amphibians (ANIM T-FAUNA-AM PH B,
eg, kaCuA)
  - Aquatic Animals (ANIM T-FAUNA-AQAN ,
kekadZa)
- Inanimate (INANI)
  - Object
    - Artifact (INANI-ARTFCT, eg, cammaca )
    - Natural Object (INANI-NAT-OBJCT,
pahAdZa)
  - Edible (INANI-EDL-OBJCT, eg,miTAI)
  - Anatomical (INANI-ANTM-OBJCT, eg
uM gall, bAla)
  - Chemical (INANI-CHM-OBJCT, eg,
amI)
  - Physical (INANI-PH SCL-OBJCT, eg,
kalama, maMca)
  - Imaginary (INANI-IM GN-OBJCT, eg
amQwa)
- Place
  - Imaginary Place (INANI-IMGY-PLC,
eg,sVarga)
  - Physical Place (INANI-PH SCL-PLC, eg,
pATaSAlA)
- Event
  - Natural Event (INANI-NAT-EVENT, eg,
BU kaM pa)
  - Historical Event (INANI-HIST-EVENT, eg,
praWama viSvayuxXa)
  - Planned Event (INANI-PLND-EVENT,
eg,bama-visPota)
  - Social Event (INANI-SCL-EVENT, eg,
janma)
  - Fateful Event (INANI-FTFL-EVENT, eg,
LAztarI nikalanA)
  - Fatal Event (INANI-FTL-EVENT, eg,
xurGatanA)
- Abstract (ABS)
  - Quality (INANI-ABS-QUAL, eg,acCAI)
  - Perception (INANI-ABS-PRCP, eg,sl canA)
Cognition (INANI-ABS-COGN, eg kalpanA’)

Colour (INANI-ABS-COLR, eg, lAla)

Title (INANI-ABS-TITL, eg, propZesara)

Measurement (INANI-ABS-MSRMNT, eg, lambAI)

Time (TIME)
1. Period (INANI-ABS-TIME-PRD, eg, GaMtA,)
2. Season (INANI-ABS-TIME-SSN, eg, garmI)
3. Historical ages (INANI-ABS-TIME-HIST, eg, pARANa yuga)

Action
1. Social (INANI-ABS-ACT-SCL, eg, vivAha)
2. Anti-social (INANI-ABS-ACT-ANTISCL, eg, corI)
3. Occupation (INANI-ABS-ACT-OCP, eg, axyApana)
4. Communication (INANI-ABS-ACT-COMM, eg, parAmarSa)
5. Physical Action (INANI-ABS-ACT-PHSLACT, eg, dubakI)

Object (INANI-ABS-OBJCT, eg SabXa)

Logos
1. Religion (RLGN)
2. Philosophy (PHIL) (Metaphysics, Epistemology, Logic, Ethics)
3. Social Sciences (SCLSC) : political science, economics, commerce, law, public administration, social services, education, anthropology, psychology, sociology, folklore.
4. Language (LNG) (Grammar)
5. The Arts (ARTS) : Fine and Performing Arts (Music, Literature, Painting, Sculpture, Film, Drama), Aesthetics and Rhetoric, Useful arts & Crafts
6. Geography (GEOG)
7. History (HIST)
8. Natural Sciences (NATSC)(Physics, Chemistry, Bio-sciences, Astronomy, Geology)
9. Mathematics (MATHS)
11. Sports (SPRT) (Indoor games, Outdoor games)
12. Transport (TRN SPT)
13. Home Science (HSC) (Food & Nutrition)
15. Fashion Designing (FSH D) (Dresses, Textile)
16. Miscellaneous (MSL)

State (STE)

Physical State (INANI-STE-PHSL)
1. Solid (INANI-STE-PHSL-SLD, eg, pawWara)
2. Liquid (INANI-STE-PHSL-LQD, eg, xUXa)
3. Gas (INANI-STE-PHSL-GAS, eg, AzksIjana)

Biological State (INANI-STE-BIO, eg, bacapana)

Mental State (INANI-STE-MNTL, eg, avasAxa)

Social State (INANI-STE-SCL eg, haMgAmA)

Process (PRCS)

Physical Process (INANI-PHYS-PRCS eg KAnA banAne kI viXi)

Mental Process (INANI-MNTL-PRCS eg yojanA)

Miscellaneous Noun Attributes:
- Abbreviation (ABRV, eg, UNL)
- Acronym (ACRNM, eg, UNESCO)
- Heading (HEAD, eg, THE INTERNATIONAL TELECOMMUNICATION UNION)

Verb

State (STE)

Physical State (INANI-STE-PHSL)

Verb of Action (VOA)
- Change (VOA-CHNG, eg, baxalanA)
- Cognition (VOA-COGN, eg, vicAra karanA, nirNaya lenA)
Commencement (VOA-CMNCT, eg, barasane laganA)
Communication (VOA-COMM, eg, liKanA,bolanA)
Competition (VOA-COMPT, eg, ladZanA)
Completion (VOA-CMPLT, eg, KA cukAna)
Consumption (VOA-CNSMP, eg, KanA, plnA)
Contact (VOA-CNTCT, eg, CU nA)
Creation (VOA-CRTE, eg, banAnA)
Destruction (VOA-DSTN, eg, cUra-cUra karanA)
Emotion (VOA-EMOT, eg, haMzsanA)
Event (VOA-EVNT, eg, barasanA, himapAwa honA)
Grooming(VOA-GROOM, eg, saMzvaranA)
Maintenance (VOA-MNTC, eg, yaWA-sWiwi banAe raKanA)
Motion (VOA-MOTN, eg, calanA)
Perception (VOA-PRCP, eg, xeKanA, sunanA)
Performance (VOA-PRFM, eg, nQwyA karanA)
Possession (VOA-POSS, eg, kabjZA karanA)
Social (VOA-SCL, eg, cunAva ladZanA, SAxI karanA)

Verb of State (VOS)

Physical State (VOS-PHY-ST, eg, KadZA rahanA)
Mental State (VOS-MNT-ST, eg, ciDZanA)
Temporal Verbs (TMP, eg calanA)
Verbs of Continuity (CONT, eg bahanA)
Verbs of Volition (VLTN, eg KAna)
Verbs of Non-volition (NVLTN, eg vivaSa honA)

Special Verb Attributes :
Idiom (V-IDM, eg, cU he billi kA Kela karanA)

Adjective

Descriptive
Weight (ADJ-DES-WT, eg, BAri puswaka)
Shape (ADJ-DES-SHP, eg, laM ba rAswa)
Colour (ADJ-DES-CLR, eg, laLa kapadZA)
Strength (ADJ-DES-ST R NGTH, eg, kamajZora kadZI)
Qualitative (ADJ-DES-QUAL, eg, acCA ladZakA)

Special Adjective Attribute :
Nouns used as Adjective (N-ADJ, eg, kaM pyUtara Kela, majaxUrA saM Ga)

Adverb
Time (ADV-TIME, eg, ke pahale)
Frequency (ADV-FREQ, eg, bAra-bAra, xo bAra)
Place (ADV-PLC, eg, hara jagaha)
Manner (ADV-MAN, eg, wejZI se)
Quantity (ADV-QUAN, eg, bahuwa)
Reason (ADV-RSN, eg, isalie)
Interrogative (ADV-INTRO, eg, kaba)
Affirmative (ADV-AFRM, eg, niSciwa hI)
Negative (ADV-NGTV, eg, nahI, SAyaxa)

Conclusions
The above discussion forms the basis of the standardization work and is called the Lexware Standard D o cument: (Foundations). This has been evolved from the experiences of the various lexical resources development activity going on at IIT Kanpur, IIT Bombay and such other places. This document will be followed by:
The main aim for all this effort is to ensure that a massive multilingual lexical resource is built for the languages of India. This will ultimately lead to (semi) automatic MT systems with capability for NL understanding. As mentioned before, this is a manpower intensive activity requiring well trained personnel in large number. The task has to be executed in a distributed manner. It is envisaged that each resource center will establish linkages to the language of the center as shown in the examples above.

Transliteration scheme followed in the above description is given below:

- A character with nukta is represented by corresponding Devanagari symbol followed by 'Z' e.g. ज - > jZ. 'Z' is "nukta operator".
- For other non-devanagari symbols we use 'V' & 'Y' as "previous" & "following" operators respectively e.g. ऐ - > eV, ए - > eY.

It is expected that this document will be carefully read, discussed and reflected upon. Also some languages may have their special requirements. Any feedback or comment pertaining to this may be sent to one of the following:

(Prof. R.M.K. Sinha, rmk@cs.iitk.ac.in, Tel: 0512-597174
Prof. Pushpak Bhattacharya, pb@cs.iitb.ac.in, Tel: 022-5767718
Prof. Uma Maheswar Rao, guraoch@uohyd.ernet.in)

Proposed Standard for Indian Script to Roman Transliteration Table (IN SROT)

Indian Scripts are phonetic and have similarity in alphabetic correspondence. Transliteration between Indian languages is simple, unambiguous and phonetically similar. However, there is need for a Scheme for transliteration from Indian Scripts to Roman Script. There have been schemes but they lack in readability of the transliterated text.

IN SROT (Indian Script Roman Transliteration Table) is proposed as a standard transliteration scheme considering the need for readability, and ease in decoding the romanized text unambiguously. This is orthographic representation, Transliteration table, rules of syllable formation & disambiguation/decoding and examples are given below:

### Vowels

<table>
<thead>
<tr>
<th>Syllabic Form</th>
<th>Intra Syllabic Form</th>
<th>IN SROT</th>
</tr>
</thead>
<tbody>
<tr>
<td>अ</td>
<td>-</td>
<td>a</td>
</tr>
<tr>
<td>आ</td>
<td>ए</td>
<td>A</td>
</tr>
<tr>
<td>इ</td>
<td>i</td>
<td>I</td>
</tr>
<tr>
<td>ई</td>
<td>u</td>
<td>U</td>
</tr>
<tr>
<td>उ</td>
<td>e</td>
<td>E</td>
</tr>
<tr>
<td>ऊ</td>
<td>o</td>
<td>O</td>
</tr>
<tr>
<td>ऋ</td>
<td>ah</td>
<td>Ah</td>
</tr>
<tr>
<td>ऌ</td>
<td>र</td>
<td>Ri</td>
</tr>
<tr>
<td>ऍ</td>
<td>ल</td>
<td>RI</td>
</tr>
<tr>
<td>ऎ</td>
<td>व</td>
<td>M</td>
</tr>
<tr>
<td>ए</td>
<td>म</td>
<td>Mh</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The character ओ (Ri) / ओ (IRi) does not represent a single vocalic sound in Hindi, but is vocalic in terms of the Script, having separate syllabic and intra syllabic forms.

Example: hiM x1 akRaramAIA = हिंदी अकरामाया
Capital letters denote ‘dirgh/prolonged vowel/mAtrA. e.g. [ a (अ), A (आ)], [ i (इ), I (ई)], [ ku (कू), kU (कू)], [ kE (के), kE (के)], ...

h is added to consonants/vowels to denote aspirated or closer sounds, eg. kha (ख), cha (च), jha (ञ), ...

N ha (ण), nha (ञ), ah (आ), Ah (ओ), aM h (अ), AM h (ओ), kaM h (के)

M denotes anusvAra/bindu and combines with previous vowel as in

aM (अ), kaM (क), kiM (कि), suM ga (सिंह)

[Note: Graphemic code joining as अ + ए = ओ (ahM) is not permitted.]

e.g. kh (क), Nh (ण), Ah (आ), (ो)M h, (ो)M h

Examples of word/phrase level transliteration

कट (kaTa), खाट (khATa), घूट (ghRita), छिप (chipa)

Intra Syllabic Vowels Forms combined with consonant character

k (ka) क (ka) क (kI) क (kI)
k (ku) क (kU) क (kRi) क (kE)

Rule for decoding from Romanized text

‘a’ is integral part of consonant. Omission of ‘a’ results into ‘pure’ consonants e.g. ka (क), k (क / a).

Vowel combined with ‘pure’ consonant transforms into mAtrA. e.g. khi (क), ghU (गु), kAh (को).

Full stop “.” should be followed by a space.