8.11 Preparation Problems in Developing Lexical Resources for Computing

Rita Mathur

Abstract—The paper focuses on the various problems that occur inevitably during the process of developing a lexical resource for computing. Since lexical resource is vital for computing, developmental process involves a multifaceted task. It requires at least three kinds of activities pertaining to linguistic, grammatical, and lexicographical areas. Linguistic activity includes morphological and semantic information. Word formation rules and word sense disambiguation are the core of linguistic activity. Tagging for parts of speech requires grammatical activity and selection of canonical forms or lemma is integral part of the lexicographical activity. Thus, development of a lexical resource is a kind of amalgamation of linguistic theory and practice.

Index Terms—Basic form, Morphological form, Tagging

I. INTRODUCTION

MOST of the languages consist of vocabulary or lexical items belonging to different sources. Diachronically they undergo many changes. Numbers of lexical items are borrowed. They become part of the language, the native speakers accept them. Interestingly only basic forms are borrowed, not the inflectional and derivational forms, which means the basic forms loose their morphological forms of the source language and adopt the morphology of the target language, in which they are merged. For instance, Hindi word [jangal] 'forest' is borrowed in English, but the oblique plural [jangalo~] is neither borrowed nor it is a possible form in English. The word has to follow English word formation rules. Similarly English words [rail, doctor] are borrowed into Hindi and are nativized. They inflect for plural and oblique forms as per Hindi morphological rules.

In Hindi the vocabulary can be classified into various strata. Initially classification is possible into *native* and *borrowed* words. *Native* words can be defined as inherited by the language from its earlier stages of the development, through phonological change. On the other hand, loan or borrowed words remain same. They retain their form of their source language. These are known as *tadbhava* (originated from) and *tatsama* (same as that) respectively. However, these terms are used for borrowed words from Sanskrit. In Hindi these *tatsama* and *tadbhava* words have different morphophonemic patterns.

Another classification may be according to the internal structure, grammatical category, and inflectional and derivational behavior of the lexical items. For instance, in Hindi a noun inflects for gender and number, verb inflects for tense, aspect, and mood and also for gender. Whereas in English verbs do not inflect for gender. Besides inflections, derivation and concatenation of suffixes are to be sorted out.

Lexical resources, therefore, are not simply copying or collecting lexical items. It involves the expertise of the compiler who should be well acquainted with linguistic principles and also well versed with the language. There are various kinds of lexical resources, for Instance, word list, machine-readable dictionary, text corpora, thesaurus, word net so on so forth. The list is not exhaustive. This paper would concentrate on the building of a word list and discuss about salient features essential for developing. Nevertheless these are not universal; one has to consider the idiosyncratic featuresof the language that have to be sorted out before taking up such an activity.

II. WORDLIST

Word list may consist of basic forms. Morphological forms may be generated through a programme. This saves space and word list may contain more lexical items. On the other hand, base words as well as morphological words can be stored in the database. In the latter case words need not to be generated; the search engine would be able to search and produce the correct output. Nevertheless, in both the cases annotation of the words is very significant. Words are annotated for canonical forms, morphological forms, parts of speech, sense and grammar. Since languages are dynamic and native speakers' pragmatic skills enable them to use words in variety of contexts, lexical items have multiple senses and usages. This is the problematic area if the word list is for machine translation. In that case, mere word list is not sufficient. It has to be annotated text corpora, machine-readable dictionary or word net. Care has to be taken for ontology, concordances, etc. Semantic disambiguation and various semantic relations have to be carefully marked.

A: Problems with Morphology:

There is no universal morphology. Each language has its own closed set of morphological rules. The generalizations are possible only within the language and they do not apply onto various other languages. Thus, in order to develop any lexical resource, morphology of that particular language has to be explored. Morphology of agglutinative language will be quite different from that of an inflectional language. Moreover, this will be different for highly inflectional languages. Morphology could be simple and complex.

Rita Mathur is with Aliyawar Jung National Institute for the Hearing Handicapped, Mumbai (e-mail: ritamathur@yahoo.com).

A complex morphology would be more challenging for computing.

Cross-linguistic examinations reveal that there are languages in which grammatical information is presented by the length of a vowel, within the morpheme. In New Zealand Maori (as reported in McCarthy: 1992) Nouns denote plurality by lengthening the vowel of the first syllable. For instance,

Singular: tangata 'person' wahine 'woman'

Plural: t**aa**ngata 'persons' w**aa**hine 'women'

Above examples suggest that word formation rules of the morphology are paired with phonological rules grouped together at various levels. The output of each morphological rule is cycled through the phonological rules of that level. In turn, phonology of that level triggers the next level of morphological rules, paring with phonological rules of that level. In this sense, the rules of morphology and phonology are cyclical. They are made to apply in a cycle, first to the root then out ward to the affixes nearest to the root and then to the outer layer of the affixes. Thus, the "lexical rules" apply within the words whereas "post lexical" rules apply across the boundary.

Hindi Morphological patterns are seen on these distinct levels. They can be defined as *non-concatenative* (level 0, where vowel reduction is seen) and *concatenative* (levels for affixation) of morphology. The influx of phonology is inevitable on all the levels. In classical linguistics this phenomenon is described as 'vowel alternant'. Vowel alternation is a kind of idiosyncratic feature of Hindi. This can be seen in *causativization* of a verb stems. This involves the alternation between short and long vowel as well as the concatenation of the affixes. The formation of the verbal sets contains phonological and semantic information. Morphologically they are partially similar. Syntactically causative forms reveal direct and indirect causation. Kellog [1887:252] notes this distinction as 'primitive', 'causal' and 'second causal' verbs. Saksena [1982] digresses from the traditional notion, which considers an intransitive verb form to be a base for all the formations and argues that the transitive verb forms, having a long vowel, should be considered as a basic form. The reason being, with intransitive verb form as basic, the morphological generalizations are not possible as the causativization by vowel lengthening is syntactically restricted. For instance, transitive form /parh/ 'to read' might become /paarh/* which is not possible. Phonologically it is ambiguous to know that short vowel will alternate with long /ii/, /uu/ or with long /E/ and /O/. While via reduction rule long /E/ and /O/ alters with /i/ and /u/, respectively. Pray [1976] considers that the majority of verb stems are *'tadbhava'* forms. '*tatsama'* forms are mostly borrowed as nouns or adjectives, not as verbs. Hence, the alternation of tense and lax vowels characterizes many kinds of phonological forms in addition to a sequence, which constitute a stem but not the entire word [Pray: 1976: 96].

The discussion above endorses vowel reduction before concatenation of affixes. It is observed that the causative formation on the Hindi verbs is a systematic process. Following are a few examples:

Level 0: [chiil] vt > [chil] vi 'to peel'
Level 1: [[chil]aa]] c [[chil]vaa]] c 'to cause to peel'
L 0: [naac] vi > [nac] 'to dance'
L 1 : [[nac]aa]]c > [[nac]vaa]] c 'to cause to dance'
L 0 : [TaTol] > [TaTul] 'to feel'
L 1 : [[TaTul]vaa]] c 'to cause to feel'

Another systematic process of non-concatenative morphology is seen in the derivation of nominal forms, diminutive forms, and compounding. For example:

lohaa 'Iron' > luhaar 'black smith'
sonaa 'gold' > sunaar 'gold smith'
lotaa 'tumbler' > lutiya 'small tumbler'
khaat 'bed' > khatiya 'small bed'
ghoDa + dOD > ghUD dOD 'horse race'
raajaa + vaaDaa > rajvaaDaa 'royal dwelling'
paanii + chakkii > panchakki 'flour mill, run by hydro energy'

Besides these alternations, morphological and word boundaries should be considered. For instance, in Marathi *khaaNyaasaaThi* 'for eating' is one morphological word, whereas, in Hindi it is a phrase, not a morphological word (*i.e., khaane ke liye*). Another interesting example of word and morpheme boundary can be seen with the pair of words : *premlata* 'a name' and *komalata* 'tenderness'. Though there is partial phonetic similarity but in earlier example there is a word boundary *prem+lata* and in latter case there is a morpheme boundary *komala#ta*. In English, words like *'blackberry', 'cranberry'* and *'raspberry'* put forth the same problem. Thus the concept of word boundary and morpheme boundary must be accounted for. This elaborates the indigenous nature of morphological profile of a language, which is inclusive of morphemic and allomorphic generalizations along with segmentation, productive rules, and typology; if ignored the computing outputs are going to suffer.

B. *Problems with Semantics:*

Meaning is not a single unitary concept, rather, it is a group of components or complexes. These are perceived as a *gestalt*. It appears in the minds of the speakers with *associative bond*. These semantic components are associated with logically implicational constraints. For Instance, [animate] vs. [non animate], [human] vs. [non human], [male] vs. [female]. At the same time words are related with hyponymy,

homonymy, synonymy, and antonymy relations. Thus one has to consider all the important factors for the selection of lemma.

These semantic values are based on ontology, argument structure along with selection restrictions, and semantic relations.

C: Problems with Orthography:

Pronunciation affects the writing. In English, the gap between spoken and writing system is much more in comparison to that of Hindi, Marathi, Spanish, etc. Spelling variation is almost nil in English. But in Hindi there are a lot of possible spelling variations. This makes the task tougher. As the pronunciation is variable, people do not speak in the same way. The knowledge about various dialectal and regional variations becomes a necessity. Moreover, a possible disparity is seen between *'in vogue usage'* and in *'traditional usage'*. This is even found in leading media expressions. Hindi is a potential language for such kind of tiff. The issues involve use of *maatraa, nuktaa, halant, anusvaar, and chandrabindu in* Hindi. Many linguists and language experts opine to keep the variations, since people accept them. On the other hand, purists debate for the correctness. Nevertheless, Hindi orthography is very much affected by modern trends and mixing attitude of the people. Ultimately the compiler has to decide with due considerations to traditional usage, printing and writing conveniences. Secondly, he may be motivated by his own individual preferences. Whatever is selected consistency with pattern remains most important factor.

III. STANDARDISATION

To develop lexical resources methodology, software tools, and standards are to be followed meticulously. Selection of the lexical entry should be in accordance to the needs of computing. There are many standardized formats and software tools that make the task simpler and enable a linguist and/or a lexicographer to work on a lexical resource. Nevertheless, compiling of the data should be according to the purpose of the software. Grammatical and orthographical tags in a word list would suffice the needs of spelling checker, parser and so on. But for machine translation semantic disambiguation tags are more important. For that purpose lexical resource must be in the form of annotated text corpora or a word net. Apart from 'purpose' the 'ultimate users' are important. Therefore, word list or any other kind of lexical data must be specialized. A general lexicon may not be sufficient. Keeping purpose and user in mind, generalized rules must be formulated. Again these rules have to have their base in idiosyncratic nature of the language.

Consistency: In order to standardize the word list or any other type of lexical resource, consistency is necessary. Work has been done for the standards of description and creation of the lexical resources, particularly, to facilitate engineering applications (Peters: HTML page) to name a few, as mentioned by Peters, TEI, EAGLES, ISLE, GENELEX, etc. They have defined large numbers of lexical features and tag sets, which can be used to create a meta data for computing. Consistency should be observed at all levels. At the level of orthography spelling standardization, regional variation, and style should be selected consistently. For Instance, In Hindi, the forms of *khaDI boli* are different from that of other dialects, namely, *purvi* or *bihari*. One must adhere to the variety, which is selected for developing a resource. Standard guidelines are available for the spelling, innovative changes. For instance, in Hindi, care must be taken for problematic issues mentioned above and a consistent pattern according to the standard guidelines should be followed. Personal preferences may be utilized sometimes if the compiler or lexicographer is a native speaker of the language.

IV. CONCLUSION

To conclude I would like to emphasize that building of a lexical resource is a joint effort of the programmer and the linguist. Linguist's responsibilities are more because a programmer would bank upon a linguist for the resource. In nutshell following are the area pertaining to various linguistic levels.

- *Phonology* : How words or lexical items are syllabified. In case of stress languages, stress markers should be analyzed.
- *Morphology*: Morpheme and word boundaries must be analyzed. Allomorphic rules should be formulated. Stem, word formation rules are to be looked for.
- *Syntax*: Parts of speech and other syntactic devices viz. inflection and derivation, mass noun vs. count noun, modifiers, their gradable nature, gender, and agreement must be explored.
- *Semantics*: Meaning may be analyzed as per lexical meaning, connotational meaning and collocational meaning. Ontology or related concepts must be explored. Homonymy, polysemy, antonymy are to be carefully marked.

REFERENCES

[1] Kellog, S.H. (1965): A Grammar Of Hindi Language: Routledge and Kegan Paul, London

[2] McCarthy,J.J (1981): A Prosodic Theory Of Non-Concatenative Morphology: Linguistic Inquiry , Vol. 12 no.3:373-417

[3] McCarthy, A.C. (1992): Current Morphology: Routledge , New York.

[4] Pray, B.R. (1970): Topics In Hindi Urdu Grammar, Research Monograph series No. 1, University of California.

[5] Saksena,A. (1982): Topics In The Analysis Of Causative With An Account Of Hindi Paradigms: University College Press , London.

[6] Singh, A.B. (1971): On Echo Words In Hindi: Indian Linguistics, Katre Felicitation Vol.30 Deccan College, Poona.

[7] Wim Peters: Lexical Resources: HTML page: www.dcs.shef.ac.uk/~wim/BALRIC_introduction_language_resource.pdf

This paper was presented at LRIL-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt. of India.