MULTILINGUAL DICTIONARY DEVELOPMENT WORKSHOP
Jan 31-Feb 1, 2008
CFILT, IIT Bombay

Date: Jan 31 – Feb 1, 2008
Venue: Conference Room, Old Guest House (Jal Vihar), IIT Bombay
Institutes Participated: AU-KBC Chennai, Banasthali Vidyapith, CDAC Mumbai, CDAC Noida, CDAC Pune, IIIT Hyderabad, IIT Bombay, IIT Kharagpur, HCU (Hyderabad Central University) Hyderabad, Utkal University, Jadavpur University
Start Time: 10 a.m. Jan 31
End Time: 5.30 p.m. Feb 1

Presentations:

1. Multilingual Dictionary Framework, and Hindi and Marathi Experience: Pushpak Bhattacharyya, Prabhakar Pande and Gauri Dabke (IITB)
2. Dictionary database, tools and interface: Mitesh Khapra (IITB)
3. Dictionary Development Experience by various participating language groups starting with Umamaheswar Rao (Central Univerity, Hyderabad)
4. Function word dictionary: Dipti Misra Sharma (IIITH)

KEY PRINCIPLES FOR DEVELOPING SYNSETS

1. Try to transfer the meaning from one language to another- DO NOT TRANSLATE WORDS.
2. Produce a synset and put the list of words that come most naturally to a native speaker of the language (which essentially is the list of most frequent words representing a concept.
3. Skip a synset if it appears to be very rare or Language-Specific (or culture-specific).
4. Create a set, **not necessarily in the same POS or category as provided by the source language synset** that refers to the same concept that is defined by the source language synset.
5. Proper nouns/Named-entities/MWEs not be stored- Should appear in a separate list called Gazetteer.

IMPORTANT DISCUSSION POINTS

1. **Coverage** (Source language-Hindi): Domain specific highly frequent synsets should be provided.
2. **Coverage of words in the Target Language**: Highly frequent words referring to the most common set of concepts in a language should be same across other languages – according to the hypothesis proposed.
3. **Granularity**: Rare and language/culture specific sysnsets should be ignored
4. **Source Bias**: To avoid source Bias- DO NOT TRANSLATE WORDS. HIT THE MEANING INSTEAD.
5. **Bilingual Dictionary Vs Feature based Multilingual Dictionary for Function words**: Developing Bilingual dictionaries is a simple and not very time taking task. On, the other hand it faces a number of problems like- handling one to many mappings and resolving ambiguities. Feature based dictionary take time and effort to build but is always a better option in order to deal with the problems mentioned above.

6. **Function words-** Bilingual dictionary or Feature based dictionary: HCU, Hyd will work out the function word matrix that will be filled up by other groups. Function Words dictionary includes-

- Pronoun
- Conjunction
- Particle (emphatic, inclusive)
- Dubitative
- Quotative
- Intensifier
- Interrogative
- Tense-Aspect-Mood (TAM) string
- Postposition

7. **Morphological Analyzer's output format** – The analyzer provided by IIT Hyd can output an analysis along with 17 attributes out of which 8 are always present and the rest 9 can be put on or off depending upon the features marked in a language morphologically. The groups using their own Morphological Analyzer need to convert the output in the same format.
8. **Named Entities/MWEs/Proper Nouns**: Should be stored in a Gazetteer

SUGGESTIONS FROM VARIOUS GROUPS

1. Choice of synsets: Highly frequent synsets should be chosen and circulated to various groups
2. Need to have more coordination between E-ILMT and IL-ILMT activities.
3. Multilingual Dictionary Development Tool kit
    a. Provision to search using an ID, a word, or a part of a word.
    b. Provision to modify the font size
    c. Support for Urdu fonts
    d. Provision to extract out a list of words of a particular POS category
4. To hold focused NER and MWE workshops

SOME COMMON PROBLEMS IN HINDI SYNSETS POINTED OUT BY VARIOUS GROUPS

- Difficult to understand the concept and the concept definition in few cases
- Conflict within a concept, synset members, gloss and examples
- Naturalization of the examples is not given much importance
- Coverage of the domain should be of high priority
- Words belonging to a synset are not equivalent in meaning at times (example: teen and anal).

TO DO TASKS WITH DEADLINES

- <u>**FOR IL-ILMT**</u>

1. Hindi Unique root words list from IIIT Hyd to IIT Bombay**- Feb 15**
2. 2000 Hindi synsets from Tourism domain from IIT Bombay to all the groups**- Feb 20**
3. 2000 synsets (Hindi to Target Language) creation by each group**- April 15**
4. Hindi-Telugu Function word matrix with IDs by HCU**- Feb 29**
5. Function word matrix filling by each group for other languages**- March 31**
6. Detailed guidelines for developing synsets using Hindi synsets as the base to be put up by IIT Bombay**.**
7. Detailed guidelines along with the Multilingual Dictionary Development tool for both online as well as offline use.

8. Tool for Hindi to English cross linkage by IIT Bombay that will be used by all the groups

- **FOR E-ILMT, CLIA**

1. 2500 synsets for Hindi-English Cross linkage to various groups by IIT Bombay**- Feb 15**
2. Hindi-English Cross Linkage of 2500 synsets by each group**-** finish by **Feb 29**

SYNSET/WORD LIST DISTRIBUTION

	E-ILMT	**IL-ILMT**
General	----	***5760 already circulated***
Tourism	***9800 already circulated***	***2000 to be circulated***
Health	***None***	***2000 to be circulated***
Function word	----	***Matrix to be circulated***

***Total number of synsets from General+tourism+Health to be covered – 15K**
***Average number of synsets currently being developed per person - 35**