Tamil WordNet

S. Rajendran
Department of Linguistics
Tamil University, Thanjavur 613010
raj_ushush@yahoo.com

Abstract

Wordnets for Indian Languages have been getting built on a massive scale under the stewardship of Pushpak Bhattacharyya, who is the torch bearer of the wordnet development in India. Wordnet preparation for Indian languages is in full swing. Wordnet building for Tamil has been started quite earlier parallel to Hindi wordnet. A Tamil wordnet model has been prepared and put as an open source in a website. Now the Tamil team has joined hands with the Indo-wordnet teams to build Tamil wordnet as one of the components of the bigger Dravidian wordnet. The paper presents the experience of the Tamil wordnet team.

1 Introduction

A wordnet plays an important role both in the development of NLP applications such as a machine translation system and a question-answering system as well as for lexical studies of a language. While wordnets have been already compiled for most of the European languages, these resources are under preparation for Indian languages. This paper presents the lexicographic and computational issues faced in an attempt to build a Tamil wordnet.

2 Characteristics of Tamil

Tamil is a Dravidian language. It is a verb final, relatively free-word order and morphologically rich language. Like other Dravidian languages, Tamil is agglutinative. Computationally, each root word can take a few thousand inflected word-forms, out of which only a few hundred will exist in a typical corpus. Subject-verb agreement is required for the grammaticality of a Tamil sentence. Tamil allows subject and object drop as well as verbless sentences. In addition, the subject of a sentence or a clause can be a possessive Noun Phrase (NP) or an NP in nominative or dative case. As Tamil is an agglutinative language, several suffixes can be added to the root word, thus forming thousands of different word forms.

The verb is the chief constituent of a sentence. Verbs take different argument structures based on their semantic nature. These argument structures define the case markers of noun phrases, which are, for instance, direct and indirect objects to the verb. In addition, the predicate which is in finite form has to agree with the subject in terms of person-number-gender (e.g. *avan vandtaan* ‘he came_3PMS’). In Tamil, two noun phrases can constitute a sentence (*avan maaNavan* ‘he student’ ‘He is a student’). As Tamil has relatively free word-order, the constituents of a sentence can be shuffled retaining the same meaning (*avan pazham caappiTTaan* ‘he fruit ate_3PMS’, *pazham avan caappiTTaan* ‘fruit he ate_3PMS’ ‘He ate a fruit’).

Tamil nouns are inflected for case and number (plural). The morphotactics of nominal forms of Tamil is as follows:

Noun + (Plural maker) + Case marker

Consider the morphological decomposition of *kaalkaLai* ‘legs_ACC’:

kaalkaLai => *kaal* <N> + *kaL* <plural> + *ai* <accusative case>

Verbs are inflected for tense and finite and non-finite markers. A finite verb shows subject-agreement marker.

Verb + Tense + Subject agreement marker (person-number-gender)

Consider the morphological decomposition of the word *paTittaan* ‘he read’:
paTittaan => *paTi*<V> + *tt* <past tense> + *aan* <3rd person, singular, masculine>

The agreement marker can simultaneously represent three distinct grammatical features: person, number and gender. For example, the morpheme *–aan* itself indicates the third person, singular number and masculine gender.

3 Challenges of representing lexical know ledge in Tamil

There is often a clash between scientific taxonomy and folk taxonomy, particularly in such areas as those of flora and fauna. As Tamil is aiming to equip itself for scientific use, preference may be felt for a scientific classification over that enshrined in folk taxonomy which reflect culture. Modern scientific taxonomy excludes *iraal* 'shrimp' from its traditional place in the class of fish, and classifies *vauvaal* 'bat' not as a bird but as an animal. We would wish to place *puumi* 'earth' and *cantiran* 'moon' separately from *naTcattirankaL* 'stars' and *cuuriyan* 'sun' along with 'stars', and to eliminate *iraaku* 'a mythical serpent believed to consume the sun and the moon and cause eclipses' and *keetu* 'a mythical celestial dragon believed to cause eclipses' from the class of astronomical phenomena. Preference for astronomy over astrology reflects a contemporary scientific approach to ontology.

Tamil being a Dravidian language holds some unique features which cannot be represented in the “expansion approach” of finding equivalents to Hindi synsets.

Tamil forms pronominalized nouns from verb stems by a productive derivative process:

va-ndt-a-van ‘come_PAST_RP_3PMS’ ‘he who came’
paTi-tt-a-van ‘study_PAST_RP_3PMS’ ‘he who studied/educated male person’

Adjective also forms such pronominalized forms:
nalla-van ‘good_3PMS’ ‘good male person’

Such derivative forms are not found in Hindi. Similarly syntax of adjective is different form that of Hindi. In Hindi, like English, adjective can occur as a complement to the ‘be’ verb occupying predicate position. In Tamil adjective cannot occur independently in the predicate position. Only the pronominalized forms of the adjectives can occur in the predicate position. For example, *vah acchaa hai* in Hindi has been translated into Tamil as *avan naalavan* ‘he good_3PMS’ ‘He is good’.

4 Challenges to creating synsets

Creating synset for a concept, though appears simple at the outset, the complexity arises when synsets are created based on Hindi/English glosses. The untrained lexicographers are tempted to find out equivalents for the set of words grouped under a synset. Mostly they get biased by the words listed under each Hindi synset. The co-synonyms given under a synset are mostly misleading; words which are different in meaning are put under the same synset.

The arrangement of co-synonyms given under each synset need to be considered seriously. Though a number of words are given as synonyms, they may vary by style or register. The order of precedence of co-synonyms in a synset can be based on the frequency of occurrence in a corpus. But in the absence of a corpus with the information on frequency of occurrence of words, this is not feasible. One has to depend on ones own intuition. But in Tamil context, this too creates problem. For example, though *cuuriyan* ‘sun’ and *candtiran* ‘moon’ are the frequently used words, because of their Sanskrit origin, they are not be given precedence over other items.

There is a unique set of verbal nouns in Tamil formed by the suffixation of *–kai*, *-tal* and *–al* to verb stems (e.g. *varukai*, *varutal, varal;* they all mean ‘coming’). Though the corresponding three members can be considered as synonyms belonging to a synset, they differ in their syntactic distribution.

5 Challenges to linking with English and Hindi synsets

Linking English and Hindi synsets with their Tamil equivalents throw challenges in a number of cases. The following problems are noted by the lexicographers:

ID 591: *maarnaa* ‘beat’ is given a wrong gloss ‘land’

ID 1392: Hindi synset of the concept *bhaims* which means ‘buffalo’ is given gloss in English as ‘buffalo’. But in ID 1411 for Hindi *bhaimsa*

which means ‘female buffalo’, English gloss is given as ‘buffalo’ instead of ‘female buffalo’.
`

ID 1427: For the Hindi synset of the concept *duniyaa* which means ‘world’, the words *mrutyulook* which means ‘dead world’ and *jiivlook* which means ‘living world’ are given as co-synonyms.

ID 1492 : For Hindi concept which means ‘balcony’, the example is given to mean ‘shed’.

ID 1683: For the Hindi concept which means ‘lion’, the synset includes words denoting tiger. In Tamil separate words are used for ‘lion’ and ‘tiger’.

ID 9199: The Hindi word *sambandhini* which means ‘lady member of the person related by a marriage alliance’, Tamil equivalent is not available. Tamil *campati* means ‘person related by a marriage alliance’.

ID 6672: For Hindi word *hiraN* ‘male dear’, English gloss is given wrongly as “sheet, flat_ solid”.

ID 4226: For Hindi *daND* ‘punishment’, English gloss is given wrongly as ‘music, medicine’.
ID 120, 121: The synonyms given for the Hindi concepts *preem* ‘love’ and *sneeh* ‘affectionateness’ are misleading.

ID 24: For the Hindi *sheerni* which means “female lion”, the words denoting tiger are given as co-synonyms. This causes confusion.

ID 172: For the Hindi *aparaadhi* which means ‘accused’, Tamil has an equivalent only in nominal form. The word which appears at the modifier position of a compound word is given adjective status in Hindi. If all the nouns at the modifier position in compounds are given adjective status, then there will be multiple categorical status for a single lexical item.

The senses relation metonymy sometimes creates problem. For example, the names of trees and fruits which are metonymically related have same forms in Hindi. This creates confusion in Tamil. For example, *aam* in Hindi means both tree as well as fruit, where as in Tamil separate words are available to denote mango tree and mango fruit: *maamaram* ‘mango tree’ and *maampazham* ‘mango fruit’.

6 Computational environment and tasks in linking

The synset correspondences are made for individual words. The working tool goes by synsets. There is always a chance of misinterpreting the meaning even though meanings are provided in Hindi and English with examples. (Some of the examples and descriptions are not free from ambiguity.) If the synsets are given keeping mind the hierarchical relations (or thesaurus-classification), the lexicographers can do the job of finding equivalents in their languages for Hindi concepts efficiently.

7 Interfacing issues

The working tool in which the synsets corresponding to Hindi have to be filled up does not contain information on semantic relations such as hyponymy-hypernymy, meronymy-holonymy, antonymy, troponymy, etc. Such information may help the lexicographers to find correct correspondences.

8 No. of synsets linked, difficult synsets etc.

2000 synsets have been linked. 2851 unique words are found for 1969 synsets. The following table gives the number of synsets in each grammatical category based on the number of co-synonyms found.

	Nouns	**Verbs**	**Adj**	**Adv**
Synsets	**2403**	**223**	**172**	**53**
1	1938	153	105	30
2	291	47	38	15
3	89	18	20	5
4	28	4	7	3
5	29	1	2	--
6	15	--	--	--
7	7	--	--	--
8	3	--	--	--
9	1	--	--	--
10	1	--	--	--
11	1	--	--	--

As verbs are more polysemous than other categories and their meaning depends on the words associated with them, deciding on the co-synonyms needs to be done with great care.

9 Conclusion

Presently only the concepts, synsets and examples have been dealt with. The semantic relations such like antonymy, hypernymy, hyponymy, meronymy, holonymy, troponymy, entailment etc. are ignored. If the individual wordnets are built based on the underlying ontology (at least on a thesaurus-classification) and semantic relations linking concepts, the lexicographers will find it easy to build their wordnet with their native intuition. But at present we are asked to give equivalents for Hindi concepts. Such approach is unnatural and ad hoc. That makes the job of building wordnet uninteresting and unscientific. This approach leads to committing mistakes. What we are engaged to make is creating a Hindi-Tamil bilingual dictionary, rather than a wordnet.

The expansion approach, though economical, will be biased over Hindi. For a language like Tamil which has rich lexical resources already available in electronic and paper forms, it is advisable to build its own wordnet independently and then link it with Indo-wordnet.

References

English WordNet - A Lexical Database for English, http://wordnet.princeton.edu/, 2009.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1993. *Five Papers on WordNet*, MIT press. www.mit.edu/~6.863/spring2009/readings/5papers.pdf

Hindi WordNet - A Lexical Database for Hindi, http://www.cfilt.iitb.ac.in/wordnet/webhwn/, 2009.

Hindi WordNet Documentation, http://www.cfilt.iitb.ac.in/wordnet/webhwn/other/hwn_docs_2.doc, 2009.

Manish Sinha, Mahesh Reddy, Pushpak Bhattacharyya. 2006. *An Approach towards Construction and Application of Multilingual Indo-WordNet*, Third International Conference on Global WordNet, Jeju Island, Korea.

Rajendran, S. 2001. *taRkaalat tamizc coRkaLanjciyam* [Thesaurus for Modern Tamil]. Tamil University, Thanjavur.

Rajendran, S. 2002. '*Preliminaries to the preparation of a Word Net for Tamil*.' Language in India 2:1, www.languageinindia.com

Rajendran, S., S. Arulmozi, B. Kumara Shanmugam, S. Baskaran, and S. Thiagarajan. 2002. "*Tamil WordNet*." Proceedings of the First International Global WordNet Conference. CIIL, Mysore, 271-274.

Vossen P. (eds.) 1998. EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers**,** Dordrecht.