WordNets for Indian Languages: Some Issues

Panchanan Mohanty
Centre for ALTS
University of Hyderabad
Hyderabad 500046, India
panchanan_mohanty@yahoo.com

Abstract

The WordNet experiment still remains an experiment even after two decades though a lot of progress has been made and a number of new horizons in this area unknown at that time have been discovered since then. It is, in fact, a good sign that indicates that the field has been growing gradually. The present paper aims to discuss some specific issues in building WordNets for Indian languages and argue that inclusion of corpus analysis data is necessary for creating better synsets.

1 Introduction

Creation of the Hindi WordNet is a pioneering effort by the Indian Institute of Technology Bombay NLP group (Chakrabarti et al. 2002), and it has become the model for building WordNets in other Indian languages. In fact, the WordNet groups of various Indian languages are following the Hindi WordNet, mostly translating the Hindi synsets with additions and deletions, wherever necessary.

With the sole exception of the Tamil, no other Indian language seems to have a large-sized corpus-based dictionary. Hence, conventional dictionaries have been used to find synsets in Hindi and the strengths and weaknesses of the Hindi WordNet are being transferred to other Indian languages. This paper looks at the synonyms of selected Hindi synsets along with entries in the dictionaries of certain Indian languages, i.e. Hindi, Oriya, Bangla, and discusses the problems involved in depending on such dictionaries.

2 Hindi, Oriya and Bangla : A Brief Note

Hindi, Oriya and Bangla- all these languages belong to the Indo-Aryan group of the Indo-European family of languages and have SOV as the unmarked word order along with other common morpho-syntactic characteristics of the verb-final languages. But Oriya and Bangla have assimilated a significant amount of Dravidian vocabulary as well as grammatical features due to their close contact with Dravidan for millennia. (Chatterji 1970, Mohanty 2008). Hindi shows less of these features for obvious reasons.

3 Semantics of Simple and Conjunct Verbs in Indian Languages

Verbs form more or less a closed set in every language and therefore, unlike nouns, their number is always limited. For this reason, verbs are more polysemous than nouns in almost all languages. Fellbaum (1998b:99) mentions: “A relational analysis of English verbs has revealed some striking ways in which verbs differ from nouns. The semantics of verbs are generally more complex.” It means the same principles should not govern the creation of NounNets and VerbNets. In other words, creation of a VerbNet in a language should follow certain principles that may be different from the ones followed in the creation of a NounNet. But it does not seem to be practised in the Indian language WordNets because of the dependence on the available conventional dictionaries.

Verbs can be divided into three categories, and let me illustrate these with example from Hindi:

a) Simple verbs: These consist of single verb roots, e.g. /ja:/ ‘to go’, /kha:/ ‘to eat’.

b) Compound verbs: These consist of two or more verb roots, e.g. /kha: le-/ ‘to eat up’, /gir paR-/ ‘to fall down’.

c) Conjunct verbs: These consist of a noun or an adjective followed by a verb root. Though a

group of verb roots can be used in this slot, /kar-/ ‘to do’ and /ho-/ ‘to be’ are the most common ones, e.g. /sapha: kar-/ 'to clean’, /band he-/ ‘to be closed’in Oriya.

It is worthwhile to note that Sanskrit did not have compound verbs and it was quite possible in it to substitute a simple verb for a conjunct verb in a sentence without significantly affecting its meaning. In fact, creation of conjunct verbs by adding the verb root √kṛ ‘to do’ to a noun is a very productive process in Sanskrit. The following examples are illustrative:

1) ra:mah ka:vyam paThati
Ram poetry reads

2) ra:mah ka:vyam paThanam karoti
Ram poetry reading does
‘Ram reads poetry.’

Most probably under the influence of Sanskrit, the same trend has been followed by the lexicographers of Neo Indo-Aryan (henceforth NIA) languages. I must point out here that there are many simple verbs in Indian languages, especially the so-called *deshaja* or native ones, for which conjunct equivalents are not available. The following Oriya examples will drive home the point:

3) dhaka:iba: ‘to gasp’
4) ba:Rhiba: ‘to serve (food)'

We cannot substituts conjunct verbs for these simple ones in Oriya. There are also concepts, though very few in number, which are expressed through only conjunct verbs and simple verbs are non-existent in such cases. Let us take the concepts ‘to work’ and ‘to love’ in English. A conjunct verb is used to express each of these concepts in all the major Indian languages. For example, for ‘to work’ we have /ka:m kar-/ in Hindi, /ka:ma kar-/ in Oriya, /ka:j kar-/ in Bangla, /pani cey-/ in Telugu. It can be glossed as ‘to work-do’. Then ‘to love’ can be rendered as /pya:r kar-/ in Hindi, /bhala pa:-/ in Oriya, /prem kar-/ in Bangla. This concept can be glossed as ‘to love-do’. None of these Indian languages uses a single verb to express any of these concepts.

However, if we look at the head entries of simple verbs in various NIA language dictionaries, we find that conjunct verbs are the preferred synonyms in most cases. Consider the following examples:

Oriya

5) paRiba: = patita heba ‘to fall’
6) dhariba: = dha:raNa kariba: ‘to hold’

Bangla

7) kha:wa: = bhakkhaN kara:, pa:n kara:
8) dhara: = dha:raN kara:

Hindi

9) kha:na: = bhojan karna:, bhakSan karna:
10) ja:na: = prastha:n karna:, gaman karna:

Before discussing these, I should metion that conjunct verbs can be of two types: synthetic and analytic. A synthetic conjunct verb is one in which both the constituents form an inseparable whole from the semantic point of view. In other words, it is almost like a frozen expression that is semantically non-compositional in nature. On the other hand, an analytic conjunct verb is semantically compositional, and it is the result of a productive process. But, interestingly and surprisingly, these analytic conjunct verb synonyms are not appropriate in meaning to their simple counterparts in many cases. For example, consider the following Oriya and Hindi simple verbs along with their conjunct synonyms in order to see how they are inappropriate in various contexts:

Oriya

11) dhariba: = dha:raNa kariba: ‘to catch, hold’
11a) se bahut ma:cha dharila:/*dha:raNa kala:
he/she many fish caught
‘He/she caught a lot of fish.’
11b) ta:ku thaNDa: dharichi/*dha:raNa karichi
to him/her cold has caught
‘He/she has caught cold.’

Hindi

12) kha:na: = bhojan karna:/bhakhshaN karna:
12a)usne ma:r kha:ya:/*bhojan kiya:/*bhakhshaN kiya:
he/she beating ate
'He was beaten up.’
12b)usne samosa: kha:ya:/ ??bhojan kiya:/ ??bhakhshaN kiya:
he/she *samosa:* ate

'He/she ate a *samosa:*.'

All these examples clearly show that a conjunct verb need not as a rule be a synonym of a simple verb. But this is what is found in the Hindi synsets. For example, /rona:, rudan karna:, krandan karna:/ 'to shed tears from eyes' (ID No. 235), /pi:na:, pa:n karna:/ 'to drink' (ID No. 6855), /joDna:, jama: karna:, sancit karna:, ikaT-Tha: karna:, ekatrit karna:/ 'to put together' (ID No. 7674), /sunna:, shravaN karna:/, 'to hear' (ID No. 8334), /paDhna:, adhyayan karna:/ 'to read' (ID No. 11727), /kha:na:, ji:mna:, bhojan karna:/ 'to eat' (ID No. 13868), etc.

Use of a simple verb in the place of a conjunct verb also poses similar problems. Consider the following examples from Oriya in which substitution of a conjunct verb, i.e. /dha:raNa kari/ by the corresponding a simple verb, i.e. /dhari/ makes the sentence unacceptable:

13a) hari kappa:Lare tiLaka dha:raNa kari mandiraku gala:
Hari forehead-on sacred mark having put temple-to went
'Hari went to the temple with the sacred mark on his forehead.'

13b) * hari kappa:Lare tiLaka dhari mandiraku gala:
Hari forehead-on sacred mark having put temple-to went
'Hari went to the temple with the sacred mark on his forehead.'

These examples demonstrate that the collocational or selectional restrictions of simple and conjunct verbs are different. Fellbaum (1998b:73) states: "We have generally avoided placing verbs that differ significantly with respect to their selectional restrictions into the same synset." For these reasons, the Indo-WordNet community has to examine the facts thoroughly and decide whether a conjunct verb should usually be given as a synonym for a simple verb or vice-versa.

There is another issue that demands our attention. According to Fellbaum (1998a:2), "Linguistic theories attempt to model human grammar, or linguistic competence, but often these theories rely on data that are not well documented in actual use." WordNet is no exception to it. If a WordNet has to be an analogue of the lexical knowledge of its speakers, then it must be organised accordingly. I should mention here Lakoff's (1987) emphasis on prototypes for deciding the most unmarked representative of a concept. If we agree with Lakoff (1987), then a prototype is the best example of a concept. It follows from it that ordering of the responses obtained form the speakers of a language with reference to a concept is indicative of their cognitive organisation of different signifiers of a signified. Let us again consider the synset /kha:na:, ji:mna:, bhojan karna:/ 'to eat' (ID No. 13868). It has /ji:mna:/ listed in the second position; but its occurrence is extremely rare. On the other hand, /bhojan karna:/ that is listed last is more commonly used than /ji:mna:/ and this is reflected in the Hindi corpus also.

All these clearly indicate that conventional dictionaries are not the best resources for determining synsets and therefore, corpus data may be considered as an aditional and reliable resource for this purpose.

4 Adjectives in the Hindi WordNet

The other concern is efficacy of the Hindi synsets. We will concentrate only on three issues which are demonstrated below by taking the following examples from the category of adjectives. The first concept under consideration is /jo a:ya: hua: ho/ (ID No. 7) 'one who has come', and there are only two synsets given in Hindi, i.e. /a:gat, sama:gat/. But a careful scrutiny shows that /a:ya: hua:, a:ye hue/ are very well used in Hindi and hence, should be listed as equivalents in the synsets. These two synonyms also occur in the DoE Hindi corpus. For example:

13) citpa:wan bra:hmaN a:rya nahi:M haiM, mishr se a:e hue yahu:di: haiM.
'The Chitpawan Brahmins are not Aryans, they are Jews who came from Egypt.'

The second concept is /jo yogya na ho/ (ID No. 23) 'one that is not fit' for which ten synonyms have been given, i.e./ayogya, anupayukta, na:qa:bil, na:ka:bil, anarha, na:la:yak, na:-la:yak, anal:yak, apa:rag, aprabhu/. Among these, the occurrence of /aprabhu/ is extremely rare and it is not found in even in some celebrated Hindi Dictionaries, like Ram Chandra Varma's (n.d.) *ma:nak hindi: kosh*. Then, /anal:yak/ is used only in poetry (ibid.:93) and the occurrence of /anarha/ is found mostly in formal Hindi.

The third concept under consideration is /bina: na:mka: ya: jiska: koi: na:m na ho/ 'without a name or one who does not have a name.' The three synonyms given for it are listed in this order: /ana:m, bena:m, na:mhi:n/. When I looked at the frequency of occurrences of these three words in the Hindi corpus, I found that each one of these occurred only once. Therefore, I asked some native speakers of Hindi who preferred the following order based on their intuition: /na:mhi:n, bena:m, ana:m/. It is just the reverse of the order given in the Hindi WordNet. Therefore, a detailed corpus analysis is necessary. Not only that, such an analysis will also reveal very interesting aspects of the meaning of a word. Let us take a fascinating example. Earlier, the meaning of the English phrasal verb 'to set in' was given as follows: 'to begin, to become prevalent, to run landwards' (Macdonald 1972:1239) and 'appear and gradually increase, flow, become settled' (Garmonsway 1987:661). But after a careful scrutiny of The Bank of English corpus, Sinclair (1991:73-75) has discovered its hidden meaning: "The most striking feature of this phrasal verb is the nature of the subjects. In general, they refer to unpleasant states of affairs." Since then its meaning has been changed to refer to something unpleasant to begin and continue *CCELD:*1323).

Again, a corpus analysis of the use of 'got' and 'gotten', the two irregular forms of the verb 'to get' in American English, shows that the former does not usually express a perfective meaning (e.g. He hasn't got an examination tomorrow) whereas the latter almost always does that (e.g. He has gotten his salary today) (Biber et al. 1995:398-399). These examples make it evident that inclusion of corpus analysis in the development of a WordNet will bring about a qualitative change in it.

5 Conclusion

To sum up, the following points have been highlighted in this paper:

(i) Though simple verbs and their conjunct counterparts are synonymous in Sanskrit, the same is not always true in the NIA languages. That is why, Indo WordNet developers must be very careful while giving a conjunct verb as a synonym for a simple verb.

(ii) The Hindi WordNet, which is quite elaborate and exhaustive, is based primarily on conventional Hindi dictionaries. Though these lexicographers were great scholars, they did not have a corpus to fall back on. Therefore, along with these dictionaries corpus data have to be taken into account while creating synsets in Hindi as well as in other Indian languages.

(iii) Some of the given synonyms are either very rare or register-specific. Therefore, trying to find equivalents for such words in other Indian languages is not only very difficult, but also not desirable.

(iv) Some synonyms, thought quite commonly used, do not find a place in certain synsets.

(v) Synonyms in the Hindi synsets may be arranged according to their frequency of occurrence in the corpus and the same should be followed in the case of other Indian languages. It will facilitate a cross-lingual comparison of the semantics of a lexical item in the languages concerned whose outcome can be used for various purposes.

Before closing let me quote from Miller (1998: xix), who states: "We have always considered WordNet to be an experiment, not a product." Let us listen to the leader.

References

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan 1995. *Longman Grammar of Spoken and Written English.* Harlow: Pearson Education Limited.

CCELD = Collins Cobuild English Language Dictionary 1987. London, Glasgow: Collins.

Chakrabarti, Debasri, Dipak Kumar Narayan, Prabhakar Pandey, & Pushpak Bhattacharyya 2002. Experiences in building the Indo WordNet: a WordNet for Hindi. In, *Proceedings of the First Global WordNet Conference.* Central Institute of Indian Languages, Mysore, pp. 57-64.

Chatterji, Suniti Kumar 1970. *The Origin an Development of the Bengali Language.* George Allen and Unwin, London. (first published in 1926)

Das, Gyanendramohan 1986. *ba:ngla: bha:Sa:r abhidha:n.* Sahitya Sansad, Calcutta. (2nd edition)

Fellbaum, Christiane. (ed.) 1998. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, Mass.

Fellbaum, Christiane. 1998a. Introduction. In, Fellbaum (ed.), pp. 3-19.

Fellbaum, Christiane. 1998b. A semantic network of English verbs. In, Fellbaum (ed.), pp.69-104.
Garmonsway, G. N. 1987. *The Modern English Dictionary*. Leicester: Galley Press.
Hindi WordNet. Indian Institute of Technology Bombay, Mumbai. (available at CALTS, University of Hyderabad, Hydrabad).
Lakoff, George 1987. *Women, Fire and Dangerous Things: What Categories Reveal about the Mind.* University of Chicago Press, Chicago.
Macdonald, A. M. 1972. *Chambers Twentieth Century Dictionary*. Bombay: Allied Publishers. (new edition)
Mohanty, Panchanan 2008. Dravidian Substratum and Indo-Aryan Languages. *International Journal of Dravidian Linguistics.* Vol. XXXVII, No. 1, pp.1-20.
Miller, George A. 1998. Foreword. In, Fellbaum (ed.), pp. xv-xxii.
Nanda Sharma, Gopinath. 2008. *shabdatattwabodha abhidha:na*. Friends' Publishers, Cuttack. (reprint)
Sinclair, John 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.
Varma, Ram Chandra n.d. *ma:nak hindi: kosh*, Vols. 1 & 2. Hindi Sahitya Sammelan, Prayag.