Mapping and Structural Analysis of Multi-lingual Wordnets

J. Ramanand, Akshay Ukey, Brahm Kiran Singh, Pushpak Bhattacharyya
ramanand@it.iitb.ac.in, {akshayu,brahm,pb}@cse.iitb.ac.in
Indian Institute of Technology, Bombay

Abstract

In this paper, we present observations on structural properties of wordnets of three languages: English, Hindi, and Marathi. Hindi and Marathi, spoken widely in India, rank 5th and 14th respectively in the world in terms of the number of people speaking these languages. The observations suggest the existence of the 'small world' property in wordnets and also lend credence to the belief that the world of concepts, which the words are manifestations of, share common properties across languages. These concepts are represented by synsets (sets of synonymous words) in wordnets. Therefore, it makes sense to link the synsets of wordnets of different languages to create a global wordnet grid. In fact, the EuroWordnet project is already doing so for a number of European languages. We too report our work on linking English, Hindi and Marathi synsets. The first task – linking of English and Hindi wordnets – requires clever ideas on mapping accurately the synsets of the two languages. The second task – relatively easier due to the close correspondence between Hindi and Marathi – reduces to programmatically borrowing lexico-semantic relations from the Hindi wordnet into the Marathi wordnet, since the Marathi wordnet is largely aligned with the Hindi wordnet. To the best of our knowledge, ours is the first in-depth investigation into lexical knowledge networks of multiple languages with a view to evaluating them, and also the first attempt to create a multiwordnet involving two major Indian languages and English.

1 Introduction

Princeton wordnet ([MB90]) is a lexical knowledge network of English words. Its growing popularity as a useful resource for English and its incorporation in natural language tasks has prompted the creation of similar wordnets in other languages as well. The English wordnet is maintained manually by a team of lexicographers and computer scientists. This manual method sacrifices speed for quality and may be impractical for efforts to bootstrap new wordnets in other languages, especially if there is a lack of linguistic support. This has therefore motivated research into using existing wordnets in Indian languages like Hindi to create data for languages like Marathi which are part of the same language family. Also, matching synsets from wordnets in unrelated languages such as English and Hindi provide benefits in multi-lingual contexts.

The increasing number of wordnets has also sparked off interest in understanding the structure and properties of these wordnets with a view to characterising and comparing them. A starting point is to study some statistical properties by considering a wordnet as a graph. This yields interesting observations on degree distribution,

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

nature of clustering, and shortest path lengths of the graphs. These help understand the nature of relatedness of words in languages, enhance the usefulness of wordnets and could also help design these databases better.

The roadmap of this paper is as follows: Section 2 discusses the related work. Section 3 introduces the basics of wordnet relevant to this discussion. Section 4 is on the structural properties of wordnets in English, Hindi, and Marathi. Section 5 outlines an automatic algorithm for bridging synsets of the Princeton English wordnet (EWN) and the Hindi wordnet (HWN). Section 6 gives a programmatic way of establishing relations in the Marathi wordnet (MWN) using HWN. Section 7 discusses the results and observations and suggests directions for future work.

2 Related work

Our study of the structure of wordnets has been motivated by the desire to compare and evaluate many different lexical knowledge networks, *viz.* wordnets (Princeton wordnet [MB90], EuroWordnet [VR98], Hindi wordnet [NC02] *etc.*), Conceptnet [LS04], Hownet [ZQ00], Mindnet [VR98], VerbNet [KS05], IEEE SUMO [NP01] and ontologies [GU04]. Another source of motivation has been the 'small world' property observed in complex networks [WA06, WC03] (do wordnets possess such properties?).

Mapping/Linking/Bridging of knowledge networks has always been a problem of great interest. Quite a few wordnets in Eurowordnet [VO98] have been created by leveraging existing wordnets. [PB02] gives details of methodologies used for this purpose. [SR06] is interesting from the point of view of linking wordnets of two sister languages (Hindi and Marathi).

Mapping synsets of one wordnet to the synsets of another involves word sense disambiguation (WSD) [YA92] as an important sub-task. [LL00] discusses a set of automatic WSD techniques for linking Korean words collected from a bilingual machine readable dictionary (MRD) to English WordNet synsets. An example of work on aligning one wordnet (Italian) with another (Princeton) is in [PB02]. Many of these mapping efforts have used the idea of the Lesk algorithm [LE86].

The mapping of Princeton wordnet to other knowledge networks has also received attention in the lexical knowledge network community [NP03].

A multi-lingual linked structure of wordnets is the goal of the Global Wordnet effort. Wordnets of many languages of the world can be found at the website (http://www.globalwordnet.org) of this endeavour.

3 A primer on wordnet

A wordnet for a language is a linked structure of concept nodes represented by sets of synonymous words called *synsets*, which are connected through lexico-semantic relations. For the user, the wordnet is a rich lexicon-like database that is queried using an API or a browser to obtain information about words. The basic idea of a wordnet can be presented through the lexical matrix. The rows in the matrix represent concepts, while the columns stand for words. Thus, a particular column represents the polysemy of a word, while a particular row depicts synonymy. The entries of a row define a synset. An example is shown in Table 1. Here, *board* is a polysemous words with several meanings. The synset for the sense *'a stout length of sawn timber'* consists of the words *board* and *plank*.

For the discussions that follow, we give examples from Hindi and Marathi wordnets – with adequate translations in English – to keep the multi-lingual flavour.

Synsets are the building blocks of wordnets. The principles of *minimality, coverage* and *replaceability* govern the creation of the synsets:

1. **Minimality**: Only the minimal set that uniquely identifies the concept is used to create the synset, *e.g.*

S.No.	(Senses, Words)	**Board**	**Plank**	**Table**	**Card**	**Gameboard**
1.	A committee having supervisory powers	x				
2.	A stout length of sawn timber	x	x			
3.	A surface for board games	x				x
4.	A table for meals	x		x		
5.	An endorsed policy of a political party		x			
6.	A set of data arranged in rows and columns			x		
7.	A printed circuit board	x			x	

Table 1: Example of entries in the English wordnet lexical matrix

{ghar, kamaraa, kaksh}[1] (*room*).
The Hindi word ghar is ambiguous and cannot by itself uniquely denote the concept of a *room*. For instance, it could also mean *house*, *native country*, or *family*. The addition of kamaraa and kaksh (also meaning *room*) to the synset brings out this unique sense.

2. **Coverage**: The synset should contain all the words denoting a concept. The words are listed in order of (decreasing) frequency of their occurrence in the corpus. *e.g.* {ghar, kamaraa, kaksh} (*room*)

3. **Replaceability**: The words forming the synset should be mutually replaceable in a specific context. Two synonyms may mutually replace each other in a context C, if the substitution of the one for the other in C does not alter the meaning of the sentence. Consider,
   {svadesh, ghar} (*motherland*)
   amerikaa meN do saal bitaane ke baad shyaam svadesh/ghar lauTaa
   Literal translation: America in two years stay after Shyam motherland returned
   'Shyam returned to his motherland after spending two years in America'
   The replaceability criterion is observed with respect to synonymy (semantic properties) and not with respect to the syntactic properties (such as subcategorization).

To explicate the meaning, a synset is associated with a gloss of definition and an example sentence.

3.1 Lexico-Semantic relations

A wordnet incorporates semantic and lexical relationships among synsets.

3.1.1 Semantic Relations

Semantic relations link two synsets. Examples of these are:

1. **Hypernymy and Hyponymy** encode semantic relations between a more general term and specific instances of it.
   {belpatra, belpattii, bilvapatra} *'a leaf of a tree named bel'* → {pattaa, paat, parN, patra, dal} *'leaf'*
   Here, belpatra (*a leaf of a tree named bel*) is a kind of pattaa (*leaf*). pattaa (*leaf*) is the hypernym of belpatra (*a leaf of a tree named bel*), and belpatra (*a leaf of a tree named bel*) is a hyponym of pattaa (*leaf*).

[1]For transliterations of the Devanagari script used in Hindi and Marathi, refer to http://www.aczoom.com/itrans/ and http://en.wikipedia.org/wiki/ITRANS

2. **Meronymy and Holonymy** express the part-of relationship and its inverse.
   {jaR, muul, sor} *'root'* → {peR, vriksh, paadap, drum} *'tree'*
   Here, jaR (*root*) is the part of peR (*tree*), implies jaR (*root*) is the meronym of peR (*tree*) and peR (*tree*) is the holonym of jaR (*root*).

3. **Entailment** is a semantic relationship between two verbs. Any verb *A* entails a verb *B*, if the meaning of *B* follows logically and is strictly included in the meaning of *A*. This relation is unidirectional. For instance, *snoring* entails *sleeping*, but *sleeping* does not entail *snoring*.
   {kharraaTaa lenaa, naak bajaanaa} *'snore'* → {sonaa} *'sleep'*

4. **Troponymy** is a semantic relation between two verbs when one is a specific 'manner' elaboration of another. For instance,
   {dahaaRanaa} *'to roar'* is the troponym of {bolanaa} *'to speak'*.

5. **Cross-linkage between different parts of speech**: Some wordnets like the HWN also link synsets across different parts of speech.

3.1.2 Lexical Relations

Lexical relations link two specific words in two different synsets. Examples of these are:

1. **Antonymy** is a lexical relation indicating 'opposites'. For instance,
   {**moTaa**, sthuulkaay} (*'fat'*) → {**patlaa**, dublaa} *'thin'*.
   patlaa (*thin*) is the antonym of moTaa (*fat*) and vice versa.

2. **Gradation** is a lexical relation that represents possible intermediate states between two antonyms. *e.g.*, {jawaanii} *'youth'* between {shaishav} *'childhood'* and {buDaapaa} *'old age'*.

4 Structural properties of wordnets

Wordnets can be represented as graphs and studied for graphical properties like *average shortest path*. 'Small World' properties ([WA06]), which have been observed in complex and large networks ranging from the likes of citation graphs to the web graph to biological oscillators, are also seen to occur in wordnets.

The *Small World* nature of these graphs means that despite the formidable size of the graphs, the average shortest path between nodes is small. Another statistic is that of *cluster coefficient* which measures whether "friends" of a node are also "friends" of each other. Results show that this is indeed true of wordnets, indicating a grouping of concepts that could suggest the presence of "clouds" or "cores" in wordnets. The last measure is that of the shape of the degree distribution graph of these nodes. Each wordnet shows a characteristic power-law distribution in its degree distribution, which indicates the presence of a few highly-connected hubs and a majority of nodes with much lesser connectivity. This also explains why the average path length is small – this can be attributed to the 'popular' hub concepts.

The observations were carried out on EWN, HWN, and MWN. Only semantic relations were considered for the small world properties as these relations link one synset to another. In contrast, lexical relations such as antonymy link two specific words within different synsets and do not apply to the other words in those synsets. All the links in the graphs are directed. For EWN, nouns and verbs were studied individually, as each set is a large graph in its own right and is separately provided in the wordnet database. The data is summarised in Table 2.

Wordnet	No. of Nodes	No. of Edges	Edges per Node
English WN v2.1 (Nouns)	84709	226483	2.674
English WN v2.1 (Verbs)	13769	29883	2.167
Hindi WN v1.0	24041	80138	3.333
Marathi WN v1.0	21116	39136	1.853

Table 2: Surface level statistics of wordnets

4.1 Degree Distribution

We compute a distribution function $P(k)$ which is the proportion of total number of nodes that have exactly k edges emanating from them ([WC03]). The function was calculated as follows:

1. Get degree k for each node
2. For each unique k, count the total number of nodes whose degree is k
3. For each unique k, $P(k)$ = (degree occurrences/total number of synsets)

Plotting $P(k)$ vs. k shows a power-law characterised by an exponent γ. A log-log plot shows a straight-line, indicating the scale-free nature of the graph. This shape was seen repeated for all wordnet graphs. A sample graph (for HWN) is shown in Figure 1.

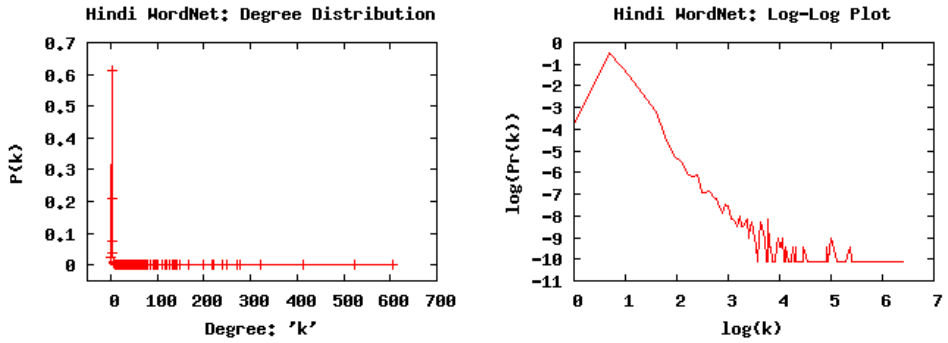

Figure 1: Degree Distribution and Log-Log plot for HWN

By measuring the slope of the line in the log-log plot, we obtained exponents γ as shown in Table 3.

Wordnet	English WN (Nouns)	English WN (Verbs)	Hindi WN	Marathi WN
Exponent(γ)	-2.063	-2.224	-2.592	-2.841

Table 3: Exponents for the Degree Distributions

These results show that while most of the nodes in the graphs have very low degree, a few nodes with very high connectivity exist. These concepts are abstract or common concepts that tend to have many specific instantiations and are richly connected to other concepts. For instance, in EWN, the synset for the concept {*person, individual, someone, somebody, mortal, soul*} has 403 relations, the synset for {*city*} has 666 relations, while nodes such as {*tour-de-force*} or {*oversight, inadvertence*} have just one relation (to their parent). Similarly,

the synsets equivalent to {*person, individual, ...*} in HWN and MWN have 607 and 626 relations respectively, the highest in each wordnet. The lower exponents (by absolute value) in the English wordnet show that the gap between proportions of degree-poor nodes and degree-rich nodes is lower than in the newer wordnets. This is possibly due to the relative maturity of EWN. Wordnet building involves identifying new synsets and creating appropriate links among synsets, which remains an ongoing task. A new synset will at least be linked to its parent hypernym. A synset in a more mature database is likely to have greater 'richness' by being linked to more synsets, whereas in a new wordnet, a greater proportion of synsets will only have the parental link. In the newer wordnets, the hubs are much more important and vital to the network than in the older database. This can be one indication of the maturity of a wordnet.

4.2 Cluster Coefficient

Cluster Coefficient C_i for a node i (with degree k_i) of a directed graph is defined as follows [WA06]:

$$C_i = \frac{|E(\Gamma_i)|}{2 \times \binom{k_i}{2}}$$

where Γ_i is the subgraph made of the neighbours of i, $|E(\Gamma_i)|$ is the number of edges of the subgraph, and $2 \times \binom{k_i}{2}$ is the total number of possible edges in Γ_i.

One extreme is where no neighbour of a node is connected to other neighbours of that node giving $C_i = 0$, whereas at the other end, each neighbour is adjacent to every other neighbour, thus forming a clique and giving $C_i = 1$. The cluster coefficient for the entire graph is found by averaging cluster coefficients for its nodes.

For the wordnets, the results are shown in Table 4. The results show that the coefficient is much higher than would be possible for a random graph, where it would be closer to $1/N$ (where N is the number of nodes). In EWN, the nodes with smaller degrees (usually ≤ 5) tend to have a higher C_i, while the degree-rich hubs have very low C_i as it is very unlikely that many of their neighbours will be related to each other. In fact, diverse groups connect to each other via these hubs. It is also seen that synsets pertaining to a specific domain such as the synset for {*American_football*} tend to have greater C_i. The newer wordnets have lower clustering coefficients as the relations structure among synsets is not very rich.

Wordnet	English WN (Nouns)	English WN (Verbs)	Hindi WN	Marathi WN
Cluster Coefficient	0.526	0.632	0.268	0.358

Table 4: Cluster Coefficients

4.3 Shortest Path

The shortest path length between two vertices i and j in a graph is the smallest number of edges required to traverse from i to j. Further, the shortest lengths between all pairs in the graph are averaged to produce the average length for the graph. The results are summarised in Table 5. The average length is fairly small for graphs of these sizes. The hubs of high degree are responsible for these short distances by being well-connected. The length in HWN and MWN is smaller, primarily because of the relatively smaller size of the graphs.

4.4 Link Distribution in wordnets

Wordnets have a collection of relations of different types that connect synsets. These are not standardised, but the nature of relations is fairly similar across different wordnets as the expectations from them are common. From the results (Figure 2), it can be seen that the taxonomic relations *i.e.* hypernymy/hyponymy usually dominate.

Wordnet	Average Shortest Path	Median Avg. Shortest Path	Std. Dev. Shortest Path	Maximum Shortest Path
English WN (Nouns)*	8.878	8.779	7.174	20
English WN (Verbs)	9.611	9.399	7.997	27
Hindi WN	4.378	4.339	2.639	15
Marathi WN	4.255	4.132	0.187	20

(*A 10 % sample was used for calculation)

Table 5: Average Shortest Path for the wordnets

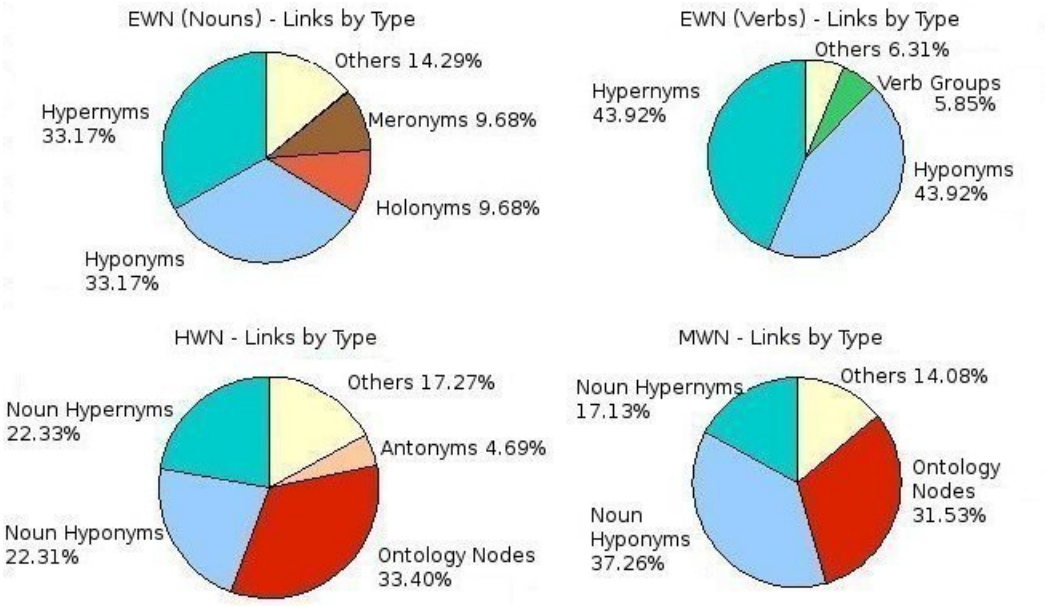

Figure 2: Distribution of Link Types for wordnets

4.5 Synset Sizes

Each synset entry has one or more words in it, which are (near) synonyms for each other and have that specific concept as their meaning. Figure 3 shows the distribution of the number of nodes with different synset sizes. This can be taken to be a distribution of synonymy in the languages. We see that all the wordnets show very similar graphs (the English wordnet has a greater number of short synsets just because of the size of the database), indicating that there seems to be no significant linguistic differences in this distribution.

5 Mapping English and Hindi wordnet synsets

5.1 Intuition

The algorithm takes as input an English synset and produces as output the best matching Hindi synset. First, a set of *candidate synsets* is obtained by finding the Hindi translations of the first word in the input synset and then finding the Hindi synsets that contain one or more of these translations in them. The first word in an English synset best represents the sense of the synset ([MB90]). Hence, the Hindi synsets which denote senses closely related to that denoted by the input synset are likely to contain some translation of this first word.

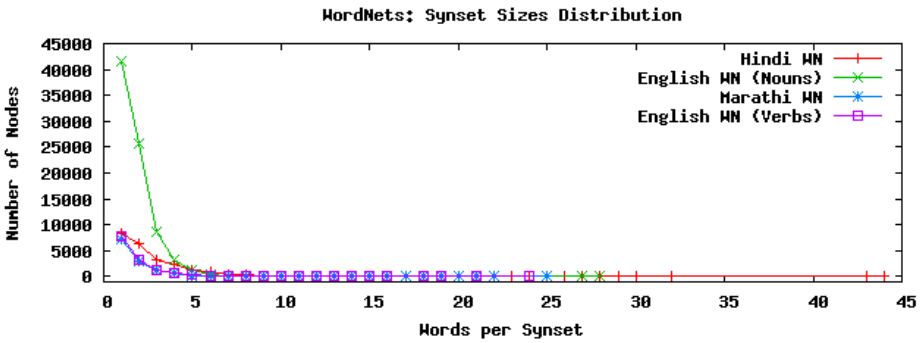

Figure 3: Synset Sizes Distributions for wordnets

After finding the *candidate synsets*, the actual weighting procedure starts. This involves the following steps:

1. Find the hypernymy hierarchies of the *candidate synsets* and call them the *candidate hierarchies*.
2. Traverse the hypernymy hierarchy of the input synset and for each synset in the hierarchy, find the Hindi translations of the words occurring in the synset.
3. Find the synsets occurring in the *candidate hierarchies*, which contain any of these translations. Increase the weights of these hierarchies if such a match is found.

At the end of the weighting procedure, the *candidate synset* corresponding to the hierarchy with the maximum weight is returned as output. In case there is no *candidate synset*, the string "No match found" is returned as output.

The hypernymy hierarchy is employed for many word sense disambiguation endeavours ([LE86]). Hence it was the natural choice for this algorithm too.

5.2 Algorithm

Following are the steps to find a match for a given English synset:

1. The first word of the given English synset is extracted and all the Hindi translations of this word are found out from a bilingual English-Hindi dictionary.
2. All the Hindi synsets which contain any of these translations are determined. They are called the *candidate synsets*.
3. The hypernymy hierarchies of these *candidate synsets* are obtained. They are called the *candidate hierarchies*.
4. The hypernymy hierarchy of the given English synset is obtained.
5. For each synset occurring in the English hypernymy hierarchy, the Hindi translations of all the words occurring in it are found out.
6. These resulting Hindi words are then searched for matches in the *candidate hierarchies*. If a match is found, the weight of the *candidate hierarchy* is increased. Initially, the weights of all the *candidate hierarchies* are set to zero. The increment is a function of:

(a) The level of the Hindi synset in the *candidate hierarchy*, where the match is found. The level of a synset in its hierarchy is defined as the number of synsets by which the synset is separated from the *candidate synset* in the hypernymy hierarchy. The level of the original *candidate synset* is set to 1, that of its hypernym is 2 and so on.

(b) The level of the English synset being considered in the English hierarchy. The level is defined in a similar manner as above: it is the number of synsets by which the present synset is separated from the original synset. The level of the original synset is set to 1.

7. The total weight of each *candidate hierarchy* is computed depending on the number of matches thus found.

8. The *candidate synset* whose *candidate hierarchy* has the maximum weight is mapped to the input English synset.

The following points are to be kept in mind:

1. The increment awarded for a match at lower levels in the hypernymy hierarchy is more than that awarded when the synsets involved are at higher levels. This is because the synsets having higher levels are farther off – in terms of the *sense denoted by them* – from the original synset.

2. The increment should depend on the levels of the English and Hindi synsets in consideration, and that too in a symmetric manner.

3. Due to the limitations of the lexical resources, the algorithm employs substring matching technique.

The function used to increment the weightage of the *candidate hierarchy* is:

$$\textit{Increment} = \frac{[(15 - m) + (15 - n)]}{2}$$

where,

m: The level of English synset in its hierarchy, whose translation has found a match in some *candidate hierarchy*.

n: The level of the Hindi synset in the *candidate hierarchy*, where the match is found.

Justification: As described above, a symmetric function was required, which decreased in value as the levels of the synsets increased. The above menioned function was heuristically chosen. The number 15 was chosen since of all the English synsets whose matches were found, the maximum depth of the hypernymy hierarchy was found to be 15. Also, the maximum shortest path between any 2 synsets in the Hindi wordnet is 15 (Table 5).

Figure 4 illustrates the working of the algorithm. The input synset is {*substance*, *matter*}. Two of the *candidate synsets* are shown: {arth, abhipraay, aashay, matalab, bhaav, maane, taatpary} and {padaarth}, along with their *candidate hierarchies*. The *candidate hierarchy* corresponding to the synset {padaarth} has more number of matches as compared to the number of matches for the other *candidate hierarchy*. The final weight of the *candidate hierarchy 1* is 40.5 and that of the *candidate hierarchy 2* is 108.5. Hence, the synset {padaarth} is mapped to the given English synset. Please note that for the purpose of matching, substring matching technique is employed.

5.3 Results

We have mapped the complete EWN V2.0 to the HWN. Approximately 6500 mappings have been checked manually. If we insist on exact matches, we get about 10% accuracy[2], whereas matching with near synonyms

[2]The accuracy is based on the judgement of a human expert, who was involved in building the Hindi WordNet itself.

(i.e. match with immediate hypernyms are accepted) yield about 25% accuracy. The reasons for the low accuracy are (i) a much larger number of English synsets, (ii) the relative immaturity of the Hindi wordnet, and (ii) the deficiencies in the English to Hindi dictionaries. We believe our approach is powerful and interesting, and with the improvement and enrichment of Hindi lexical resources, it will yield much better results.

6 Indo Wordnet Creation - Linking of Wordnets in different Indian Languages

We have, for long, been engaged in building lexical resources for Indian languages ([NC02]) with focus on Hindi and Marathi (http://www.cfilt.iitb.ac.in). The HWN and MWN more or less follow the design principles of the Princeton wordnet for English, paying particular attention to language specific phenomena (such as complex predicates meaning verbs with incorporated nouns and compound verbs) whenever they arise.

While HWN had been created from first principles by looking up listed meanings of words from different dictionaries, MWN has been created derivatively from HWN. That is, the synsets of HWN are adapted to MWN via addition or deletion of synonyms in the synset. For example, the synset in HWN for the word `peR` (meaning 'tree') is {`peR, vriksh, paadap, drum, taru, viTap, ruuksh, ruukh, adhrip, taru-var`}. MWN deletes {`peR, viTap, ruuksh, ruukh, adhrip`} and adds `jhaaR` to it. Thus, the synset for 'tree' in MWN is {`jhaaR, vriksh, taruvar, drum, taru, paadap`}. Hindi and Marathi being close members of the same language family, many Hindi words have the same meaning in Marathi – especially the `tatsam` words (directly borrowed from Sanskrit).

6.1 Relation Borrowing in Marathi wordnet

The process of setting up lexico-semantic relations in one wordnet using the corresponding information from another wordnet is called *Relation Borrowing* ([SR06]). Described below are the different kinds of *Relation Borrowing* from HWN to MWN:

1. **When the meaning is found in both Hindi and Marathi**: This is the most common case, since Hindi and Marathi are sister languages and exist in almost identical cultural settings. The relations are established in MWN for that meaning, using the procedure explained in Figure 5.

2. **When the meaning is found in Hindi but not in Marathi**: Relation borrowing is not possible. For instance, {`daadaa, baabaa, aajaa, daddaa, pitaamaha, prapitaa`} is a synset in Hindi for 'paternal grandfather'. There are no equivalents in Marathi.

3. **When the meaning is found in Marathi, but not in Hindi**: The relations must be set up manually. For example, {`gudhipaadvaa, varshpratipadaa`} is a synset in Marathi for 'New Year' which does not have any equivalent in Hindi.

The algorithm for *Relation Borrowing* is given in Figure 5.
Following data structures are used for the linking purpose:

1. A table called *tbl_all_words*, which, for each word, stores the part of speech (PoS) and an array of ids for synsets in which the word participates. Table 6 illustrates this for the Hindi word *kara* (meaning *to do*).

2. A table called *tbl_all_synsets* which stores the synset ids, the synsets, and the glosses for the various meanings.

3. A table *tbl_<PoS>_<Relation>* for each PoS and Relation combination. For example, *tbl_noun_hypernymy* is the table for the hypernymy semantic relation.

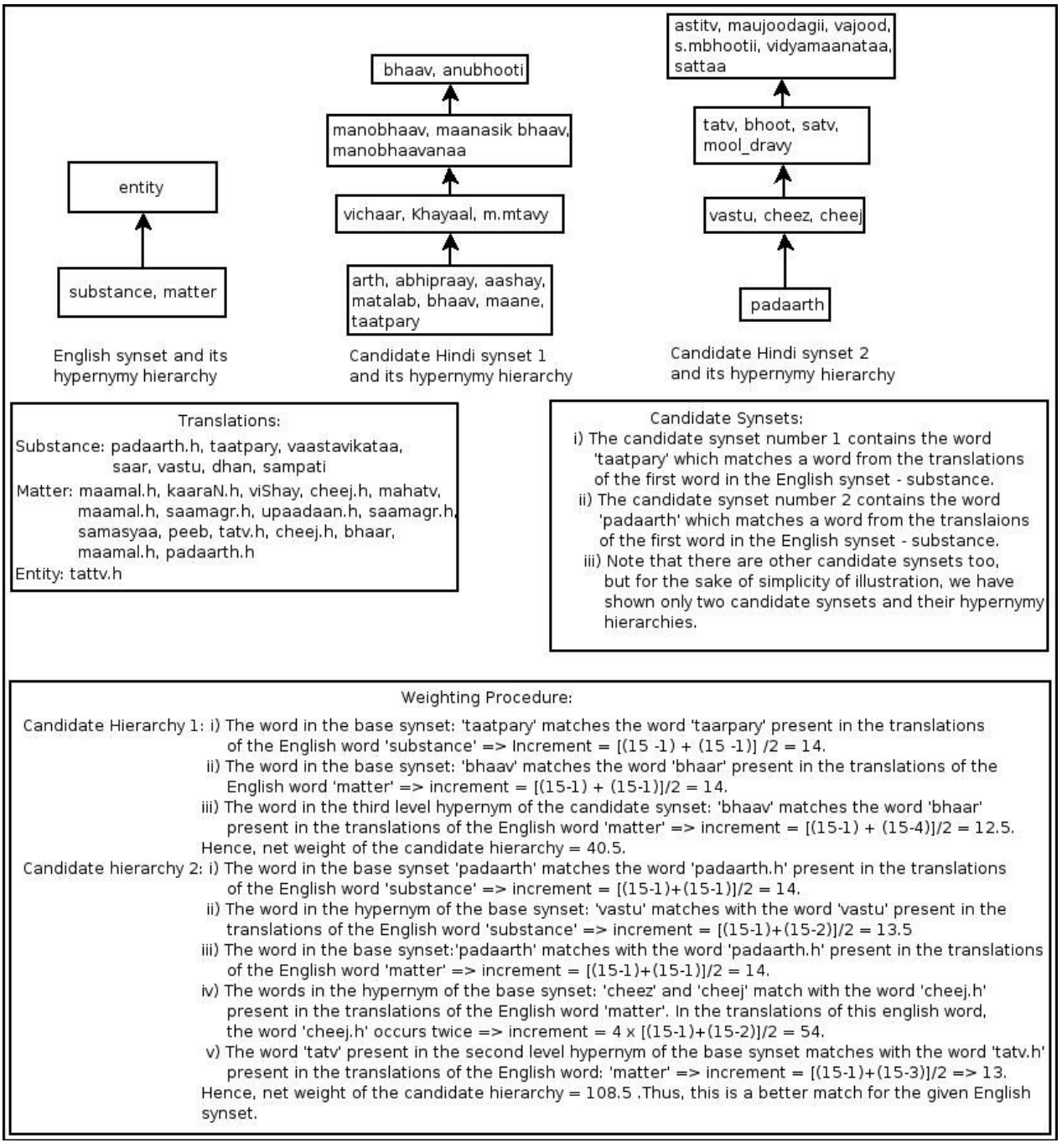

Figure 4: Block Diagram for English to Hindi synset Mapping

```
for each synset identity marathi_synset_id in Marathi WordNet do
  if (marathi_synset_id == hindi_synset_id) do
    for each relation r pointed by hindi_synset_id do
      if (relation type of r is semantic) do
        clamp the synset identity linked by relation r into marathi_synset_id
      end if
      else
        clamp the synset identity linked by relation r in hindi_synset_id to marathi_synset_id
        AND manually insert corresponding lexical elements
      end else
    end for
  end if
end for
```

Figure 5: Algorithm for *Relation Borrowing* between HWN and MWN

hindi_synset_id	*word*	*category*
491	*kara*	noun
3295	*kara*	verb
3529	*kara*	noun

Table 6: Example of *tbl_all_words* entries

Using the basic ideas outlined above, the synsets of MWN are completely linked with semantic and lexical relations. This saves a lot of manual labour. An interface has been designed to facilitate the simultaneous browsing of HWN and MWN. The input to this browser is a search string in any of the two languages. The search results for both the languages are displayed simultaneously.

6.2 Results on HWN and MWN linking

The Marathi wordnet obtained after establishment of relations using the above methodology was evaluated manually by lexicographers. Out of more than 12500 synsets created for Marathi wordnet, the program established relations for around 9000 synsets. The number of synsets for which relations were established is less because of the third case explained in section 6.1. Out of this number, the lexicographers qualitatively evaluated sets of 15% synsets sampled from each part of speech. We find that on an average, about 75% of the synsets of the Marathi wordnet are linked correctly. Considering only the case where synsets are aligned in both the wordnets, the average accuracy is 93%. The inaccuracy is a reflection of the incorrect links between synsets in the Hindi wordnet which is induced in the Marathi wordnet. The incorrect links are due to human error in the development of the wordnet. A huge dataset of more than 25,000 synsets cannot be checked manually. Only an application or a tool scanning the whole wordnet for a specific task reveals such errors. The error-cum-incomplete cases are mainly due to the absence of a complete repository of synsets in the Marathi wordnet which is still growing.

7 Conclusions and Future Work

We have presented our work on statistical measurements on wordnet structures in a trilingual setting – English, Hindi and Marathi. Many kinds of observations on EWN, HWN and MWN have been tabulated, graphically depicted and interpreted. These observations draw attention to the close correspondence and largely similar

structural properties of different wordnets. It is interesting to note that just like the English wordnet, wordnets in Hindi and Marathi also exhibit the small-world nature. We plan to use this kind of information for evaluating wordnets. This can have implications for design choices in database implementations of such lexicons since an overwhelming majority of nodes have very low out-degrees. For instance, [KH05] use degree distribution data to optimise the schema of a multi-lingual database to improve performance.

Addressing the problem of matching of English and Hindi synsets revealed the challenges in such linking, the most significant of which is the disparity in the sizes of the wordnets. We would like to improve upon the algorithm presented for the task. The heuristic function used produces promising but not very good results. We would like to improve upon the matching quality by keeping the basic approach same and changing the heuristic function to not just depend on the levels of the synsets where the match was found, but also on the difference between the levels. This might lead to a better heuristic since the closely related synsets across the languages should have matches between the synsets which are at nearly the same levels.

The creation of the Marathi wordnet is being led by the wordnet of its sister language, *viz.* Hindi. The programmatic borrowing of semantic relations has dramatically cut down on the manual efforts of constructing this valuable resource. The method will prove more and more effective as the coverage of the Marathi wordnet increases.

All these issues as described above are – to our mind – valuable steps towards (i) wordnet evaluation and (ii) creation of the multi-lingual Indo wordnet linked with the global wordnet grid.

Acknowledgements: We thank the lexicographers at the Centre for Indian Language Technology (CFILT), IIT Bombay, for their evaluation of results and for their valuable suggestions.

References

[FR98] X. Farreres, G. Rigau, H. Rodriguez. *Using WordNet for Building WordNets*. Proceedings of the COLING-ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal, Canada, 1998.

[GU04] N. Guarino. *Towards a Formal Evaluation of Ontology Quality – (Why Evaluate Ontology Technologies? Because It Works!)*. IEEE Intelligent Systems, Vol. 19, No. 4:74-81, 2004.

[KH05] A. Kumaran, J. Haritsa. *SemEQUAL: Multilingual Semantic Matching in Relational Systems*. Proceedings of the 10th International Conference on Database Systems for Advanced Applications (DASFAA), Beijing, China, 2005

[KS05] K. Kipper-Schuler. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph. D. Thesis. University of Pennsylvania, 2005.

[LE86] M. Lesk. *Automatic sense disambiguation using machine readable dictionaries: How to tell a pinecone from a ice cream cone*. Proceedings of the SIGDOC '86, 1986.

[LL00] C. Lee, G. Lee, S. JungYun. *Automatic WordNet Mapping using Word Sense Disambiguation*. Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC 2000), 2000.

[LS04] H. Liu, P. Singh. *Commonsense Reasoning in and over Natural Language*. Proceedings of the 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems, 2004.

[MB90] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller. *Introduction to Wordnet: an on-line lexical database*. International Journal of Lexicography Vol. 3, No. 4, pages 235-244, 1990.

[NC02] D. Narayan, D. Chakrabarty, P. Pande, P. Bhattacharyya. *An Experience in Building the Indo-WordNet – A WordNet for Hindi*. Proceedings of the First International Conference on Global WordNet (GWC 02), Mysore, India, 2002.

[NP01] I. Niles, A. Pease. *Towards a Standard Upper Ontology*. Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001). Chris Welty and Barry Smith, eds., Ogunquit, Maine, 2001.

[NP03] I. Niles, A. Pease. *Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology*. Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE 03), Las Vegas, Nevada, 2003.

[PB02] E. Pianta, L. Bentivogli, C. Girardi. *MultiWordNet: Developing an Aligned Multilingual Database*. Proceedings of the First International Conference on Global WordNet (GWC 02), Mysore, India, 2002.

[SR06] M. Sinha, M. Reddy, P. Bhattacharyya. *An Approach Towards Construction and Application of Multilingual Indo-WordNet*. Proceedings of the 3rd Global Wordnet Conference (GWC 05), Jeju Island, Korea, 2006.

[VO98] P. Vossen. *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, 1998.

[VR98] L. Vanderwende, S. Richardson, W. Dolan. *MindNet: acquiring and structuring semantic information from text*. Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics, 1998.

[WA06] D. Watts. *Small Worlds – The Dynamics of Networks between Order and Randomness*. Princeton University Press, 2006.

[WC03] X. F. Wang, G. Chen. *Complex Networks: Small-World, Scale-Free and Beyond*. IEEE Circuits and Systems Magazine, 2003.

[YA92] D. Yarowsky. *Word Sense Disambiguation using statistical model of Roget's categories trained on large corpora*. Proceedings of the 14th International Conference on Computational Linguistics (COLING-92), pages 454-460, Nantes, France, 1992.

[ZQ00] Zhendong Dong, Qiang Dong. *An Introduction to HowNet*. Available from http://www.keenage.com, 2000.