

Towards Evaluating Lexico-Semantic Networks

M.Tech Project - Stage Two Report

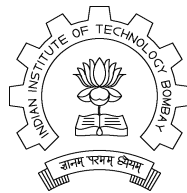
Submitted in partial fulfillment of the requirements
for the degree of

Master of Technology
by

J. Ramanand, KReSIT
Roll No: 05329402

under the guidance of

Prof. Pushpak Bhattacharyya
Computer Science and Engineering
IIT-Bombay



Kanwal Rekhi School of Information Technology
Indian Institute of Technology, Bombay
Mumbai

Abstract

This second stage report describes some interesting structural properties of wordnets in English, Hindi, and Marathi. Three of these properties show that wordnets exhibit a 'small world' and 'scale-free' nature, just like many other complex real-life networks. These observations may be used in the quest for evaluating lexico-semantic networks. The report ends by outlining the further directions for this project in the next stage.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Outline	2
2	Small World Properties	3
2.1	Introduction	3
2.2	Degree Distribution	3
2.2.1	Results	4
2.2.2	Comments	6
2.3	Cluster Coefficient	6
2.3.1	Results	7
2.3.2	Comments	7
2.4	Shortest Path	9
3	Other Statistics	10
3.1	Links Distribution in wordnets	10
3.2	Synset Sizes Distribution	10
4	Conclusion and Future Work	13

Chapter 1

Introduction

1.1 Motivation

Lexico-semantic networks such as WordNet [1] are widely used in natural language processing tasks. Starting with English, wordnets are also now available in other languages such as several European languages and Indian languages like Hindi and Marathi. Other knowledge networks such as ConceptNet have also been proposed and created. This begs the question: what is the quality of such knowledge networks?

This question calls for methods of evaluating lexico-semantic networks based on established criteria. A lexico-semantic network provides access to words, concepts, glosses, and relations between concepts and words, usually by way of APIs. Also, wordnets are being created, either from first principles or by bootstrapping from related language wordnets, which raises questions about the maturity of a wordnet at a given stage. These are some of the aspects this project wishes to explore and answer.

During the first stage of this project [3], we concluded that there were no available approaches to evaluating lexico-semantic networks in the literature. So to build a relevant background, we looked at various approaches to evaluations in the area of ontologies, and outlined some of the challenges for rating lexical knowledge networks.

In this second stage, we studied some structural properties of different wordnets to inspect the possibility that these properties could reflect the intrinsic nature of the wordnets. The experiments were carried out on the following: Princeton English Wordnet (EWN), Hindi Wordnet (HWN) [2], and Marathi Wordnet (MWN). This exercise also helped to verify whether the wordnets exhibit the ‘Small World’ phenomenon that is seen in many complex networks. Some of these properties seem to be useful in coming up with an evaluation scheme for wordnets and other lexico-semantic networks.

1.2 Outline

The roadmap for this project is as follows: We begin by briefly discussing the ‘Small World’ phenomenon and properties used to describe it. We then provide results of studying these properties on EWN, HWN, and MWN. In EWN, we studied nouns and verbs separately. This is followed by a brief summary of results of measuring the synset size and links distribution in these wordnets. We conclude by summing up the work done in the second stage and outlining the possible future approaches in devising an evaluation scheme for lexico-semantic networks.

Chapter 2

Small World Properties

2.1 Introduction

The intuition behind the “Small World” property of graphs is that despite very large graph sizes and possible unrelatedness, the average shortest path between nodes is fairly small. In the real world, this is often manifested as two strangers finding a short link of common acquaintances between them [4]. This model was created [5] to explain how graphs had average shortest paths similar to the model of random graphs and also had high clustering tendencies as in regular lattices. Many complex networks such as the Web graph, biological oscillators, citation graphs *etc.* exhibit these properties. Additionally, their connectivity distributions are ‘scale-free’. This means their connectivity distributions are in a power-law form that is independent of the size of the network [4]. In such graphs, most nodes have very few link connections, but there exist a few dominant nodes that are very rich in their degrees.

The motivation for studying such properties of complex networks is that they may help characterising these graphs which are often difficult to comprehend. These properties are also commonly seen across different networks, including languages [6]. Specifically, the small world and scale-free nature of networks has been observed for a knowledge network like the English wordnet [7]. This throws up several interesting questions: do these properties hold for different languages and language families? Can they provide insights into languages? Can they point to the richness and maturity of languages? More pertinently in this project’s context, can they help rate language knowledge networks on maturity and quality?

2.2 Degree Distribution

The degree k of each synset in a wordnet graph is defined as the number of semantic relations emanating from that synset *i.e.* here we restrict ourselves

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

Table 2.1: Exponents for the Degree Distributions

to outbound degree. We compute a distribution function $P(k)$ which is the proportion of total number of nodes that have exactly k edges [4]. 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 , $P(k) = (\text{degree occurrences}/\text{total number of synsets})$

2.2.1 Results

Plotting $P(k)$ vs. k shows a power-law characterised by an exponent γ . A log-log plot shows a significant straight-line. This shape was seen repeated for all wordnet graphs. A sample graph (for HWN) is shown in Figure 2.1.

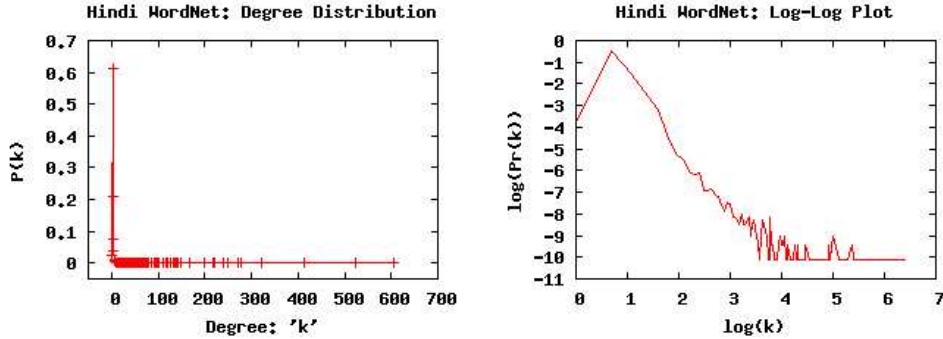


Figure 2.1: Degree Distribution and Log-Log plot for HWN

By measuring the slope of the line in the log-log plot, we obtained exponents γ as shown in Table 2.1.

The top ten nodes in terms of out-degree for the wordnets are summarised in Tables 2.2, 2.3, 2.4, and 2.5.

No.	Degree	Synset
1.	664	city,metropolis,urban_center
2.	611	law,jurisprudence
3.	526	United_Kingdom,UK,...
4.	400	person,individual,someone,somebody,mortal,soul
5.	400	bird_genus
6.	375	military,armed_forces,armed_services,military_machine,war_machine
7.	372	writer,author
8.	361	mammal_genus
9.	359	herb,herbaceous_plant
10.	322	asterid_dicot_genus

Table 2.2: EWN Nouns - Top Ten by degree

No.	Degree	Synset
1.	397	change,alter,modify
2.	188	change
3.	132	be
4.	123	travel,go,move,locomote
5.	112	remove,take,take_away,withdraw
6.	98	supply,provide,render,furnish
7.	96	move
8.	93	move, displace
9.	85	cover
10.	76	put,set,place,pose,position,lay

Table 2.3: EWN Verbs - Top Ten by degree

No.	Degree	Synset
1.	607	vyaktii, maanas, shaks, shakhs, ba.ndaa (person)
2.	524	karm, karanii, kaam, kaarya, krtya, kaarvaaii, kaarvaahii (action)
3.	414	avasthaa, dashaa, haalaat, sthithii, vrttii, suurat, haal, gatii (state)
4.	323	maanav_krti, maanavkriti, maanav_nirmiti_vastu, maanav_krt_vastu (human action)
5.	277	kriyaa (act)
6.	273	pauraaNik_purush, puraaNiiya_purush (historical/classical/mythological male)
7.	249	peD, vrksh, paadap, taru, viTap, ruksh, ruukh, adhrup, aga, anokaha (tree)
8.	240	rog, biimaarii, vyaadhii,marz, ajaar, apaaTav (disease)
9.	222	aujaar, upakaraN, karaN, saadhan, hathiyaar (instrument)
10.	218	sthaan, jagah, sthal, pradesh (place)

Table 2.4: HWN - Top Ten by degree

No.	Degree	Synset
1.	626	vyaktii, maaNus, isama, manushya, paTThaa, paThyaa (person)
2.	546	karm, krtii, kriyaa, kaam, kaarya, krtya (action)
3.	428	avasthaa, sthithii, dashaa, gat (state)
4.	397	nagar, shehar (city)
5.	377	jilhaa, pargaNaa (district)
6.	253	jhaaD, vrksh, taruvar, drum, taruu, paadap (tree)
7.	244	rog, aajaar, dukhaNe, vikaar, vyaadhii (disease)
8.	227	upakaraN, saadhan, avajaar (instrument)
9.	224	ThikaaN, jaagaa, sthaL, sthaan (place)
10.	218	rahivaasii (dweller)

Table 2.5: MWN Nouns - Top Ten by degree

2.2.2 Comments

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, 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). The lower exponents (by absolute value) in 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.

2.3 Cluster Coefficient

Cluster Coefficient C_i for a node i (with degree k_i) of a directed graph is defined as follows [5]:

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

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

Table 2.6: Cluster Coefficients

Synset	Degree	C_i
Hamas, Islamic_Resistance_Movement	3	0.667
air_unit	10	0.044
thing	22	0.082
cell	36	0.007
New_Testament	51	0.012
England	85	0.004
baseball	98	0.002
animal_order	102	0.010
military, armed_forces, armed_services, military_machine, war_machine	224	0.002
law, jurisprudence	626	0.000

Table 2.7: EWN Nouns: Sample Cluster Coefficients

where Γ_i is the subgraph made of i and its neighbours, $|E(\Gamma_i)|$ 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.

2.3.1 Results

For the wordnets, the results are shown in Table 2.6. Also for each wordnet, an illustrative list of individual cluster coefficients of ten nodes with different degrees are also summarised in Tables 2.7, 2.8, 2.9, and 2.10.

2.3.2 Comments

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

Synset	Degree	C_i
oxidise, oxidize, oxidate	3	0.667
ride_horseback	8	0.179
write, compose, pen, indite	22	0.032
kill	38	0.003
dance, trip_the_light_fantastic, trip_the_light_fantastic_toe	41	0.005
make, create	55	0.001
act, move	73	0.001
remove, take, take_away, withdraw	112	0.001
go_around, spread, circulate	123	0.001
change, alter, modify	397	0.000

Table 2.8: EWN Verbs: Sample Cluster Coefficients

Synset	Degree	C_i
kuvyasanii, kutebii, durvyasanii, vyasanii (one who has bad habits)	2	0.500
gaayak_pakshii (song bird)	8	0.125
chiinii, shakkar, sharkaraa, shakar (sugar)	10	0.133
sariisrp_jiiv, sarisrp_ja.ntuu, sariisrp (reptile)	15	0.057
ghaTanaa, baat, daastaa.n (event)	18	0.000
uttariiamariikii_desh, ... (North American country)	26	0.037
padaarth, vastu, chiiij, dravya (thing)	97	0.006
mahilaa, strii, aurat, naarii, ... (woman)	148	0.007
peD, vrksh, paadap, taru, ... (tree)	249	0.004
vyaktii, maanas, shaks, shakhs, ba.ndaa (person)	607	0.002

Table 2.9: HWN: Sample Cluster Coefficients

Synset	Degree	C_i
pohe (a flattened type of rice)	3	0.667
tiiL (mustard)	10	0.044
masaale (spice)	22	0.082
praaNii, jiiv (being)	36	0.007
pashuu, praaNii, janaavar, jitaraap, jitaraab (four-legged animal)	51	0.012
paatr (container)	85	0.004
padhaarth, dravya, vastuu (matter)	98	0.002
shaariirik_bhaag, shaariirik_avayav (body part)	102	0.010
ThikaaN, jaagaa, sthaL, sthaan (place)	224	0.002
vyaktii, maaNus, isama, manushya, paTThaa, paThyaa (person)	626	0.000

Table 2.10: MWN: Sample Cluster Coefficients

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

(*A 10 % sample was used for calculation)

Table 2.11: Average Shortest Path for the wordnets

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.

2.4 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 shortest path length was computed using Dijkstra's single-source shortest path algorithm. The results are summarised in Table 2.11. 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.

Chapter 3

Other Statistics

3.1 Links 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 among different wordnets as the expectations from them are common. From the results (Figs. 3.1, 3.2, 3.3, and 3.4), it can be seen that the taxonomic relations *i.e.* hypernymy/hyponymy usually dominate wordnets.

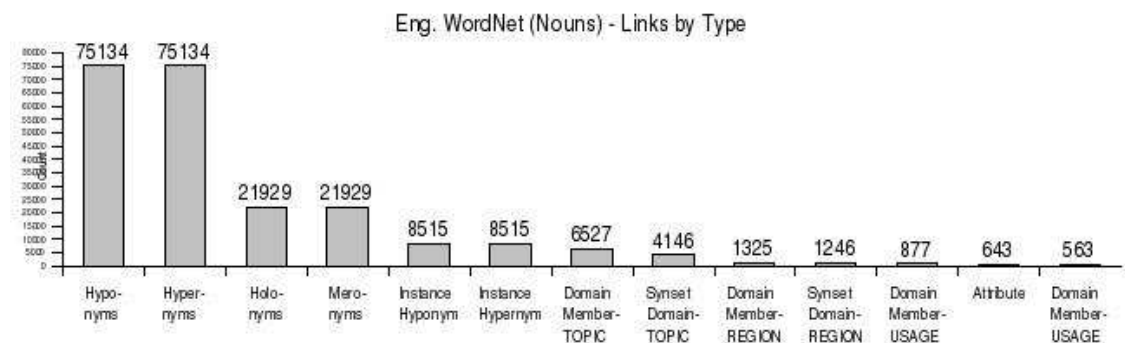


Figure 3.1: Link Types for English wordnet (Nouns)

3.2 Synset Sizes Distribution

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 3.5 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 just has a greater number of short synsets because of the size of the database),

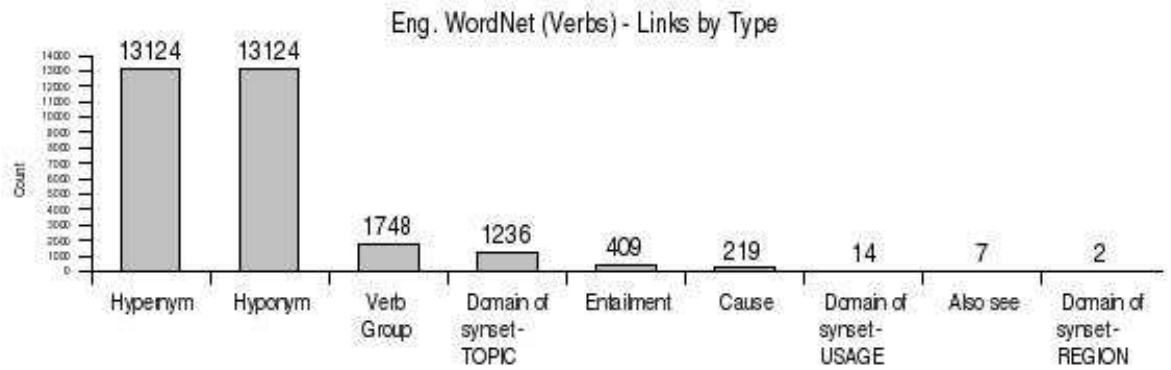


Figure 3.2: Link Types for English wordnet (Verbs)

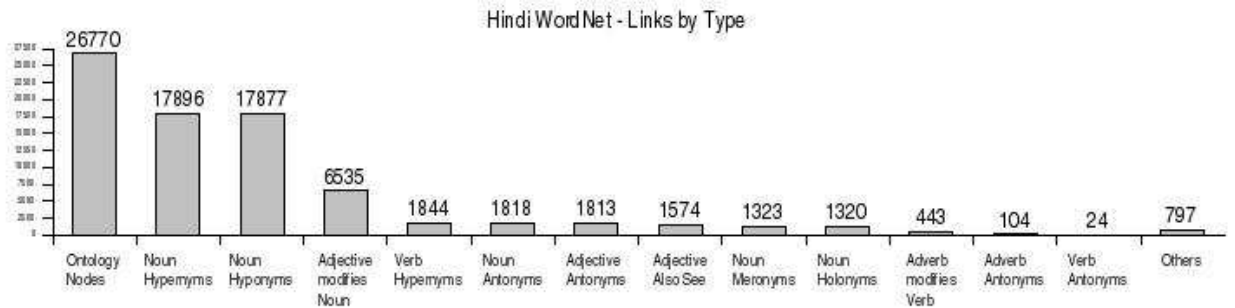


Figure 3.3: Link Types for Hindi wordnet

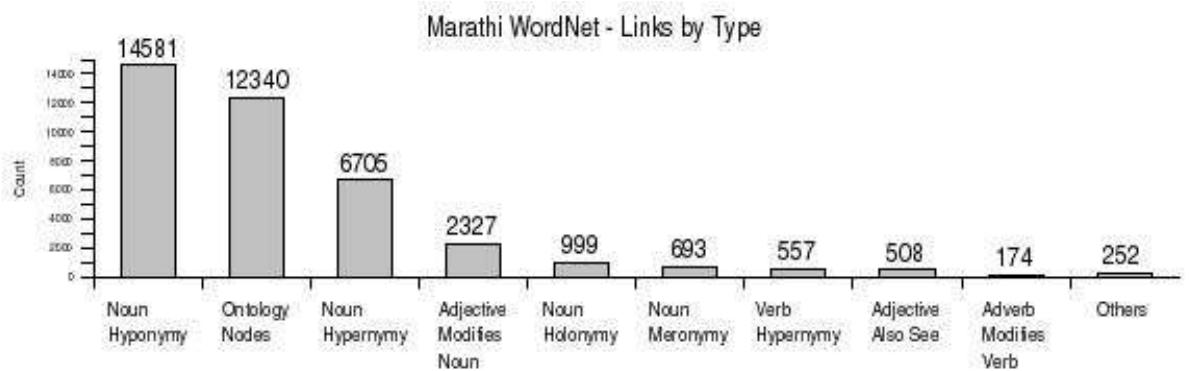


Figure 3.4: Link Types for Marathi wordnet

indicating that there seems to be no significant linguistic differences in this distribution.

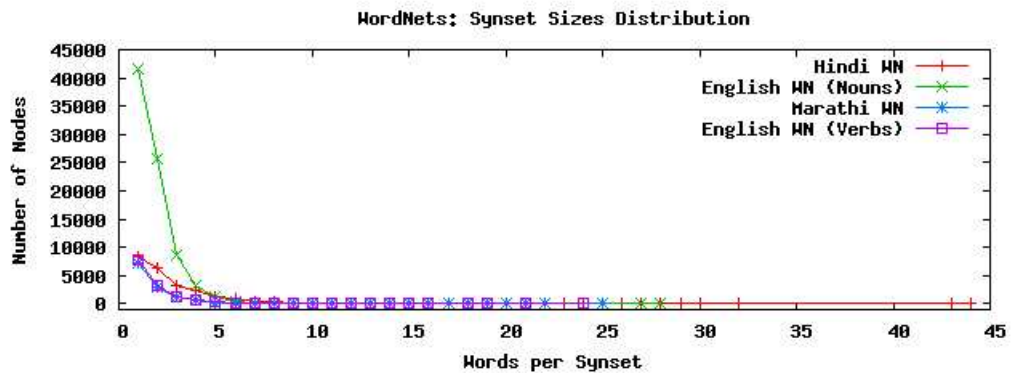


Figure 3.5: Synset Sizes Distributions for wordnets

Chapter 4

Conclusion and Future Work

The significant results obtained in the second stage are:

1. From the observations involving degree distribution, average shortest path, and cluster coefficient, it is clear that EWN, HWN, and MWN display small-world and scale-free properties.
2. These properties are fairly similar across wordnets and so variations in their values may be useful in evaluation of qualities like maturity of a wordnet. These observations lend some credence to the belief that there may exist a common core of concepts for all languages which are being captured by the wordnets (if culture-specific synsets are excluded).
3. Clearly, a few nodes are much more degree-rich than the rest of the nodes, suggesting that these constitute dominant hubs in the network. A significant reason for such richness is when nodes can have many hyponyms, such as in the case for the synset for *city*.
4. The exponent of the degree distribution graph could be a potential indicator of maturity of a graph. EWN results showed lower absolute values than for the newer wordnets.
5. Cluster coefficient is in the range of 0.3 to 0.5 mainly because most nodes have very few neighbours, and would be likely to 'know' each other. Degree-rich nodes have very low cluster coefficient.
6. Average shortest path is low, but greater in the case of a larger sized wordnet graph (such as EWN).
7. The links distribution shows that the taxonomic relations (hypernym, hyponym) are dominant in a wordnet.

For the next stage, the following lines of investigation present themselves:

1. The notion of ‘important’ synsets needs further enquiry. A possible evaluation scheme could involve ensuring that important synsets are correctly described and linked in the network *i.e.*, an error in recording a more important synset is more costly than in other synsets.
2. An evaluation based on using text corpora and other lexical resources to validate wordnets. However, this is easier for English than languages like Hindi.
3. We have not considered parts of speech separately so far for Hindi and Marathi. It may be more useful to consider subgraphs of wordnets based on different parts of speech (this is already easy for English).
4. Currently, comparing wordnets is easier due to a shared set of creation principles and structure. We would also like to bring other lexico-semantic networks into the picture.

Bibliography

- [1] George Miller, Richard Beckwith, Christiane Felbaum, Derek Gross, Katherine J. Miller. *Introduction to Wordnet: an on-line lexical database*. pg 235-244 fo the International Journal of Lexicography 3(4), 1990.
- [2] Narayan D., Chakrabarty D., Pande P., Bhattacharyya P. *An Experience in building the Indo-Wordnet - A Wordnet for Hindi*. International Conference on Global Wordnet (GWC '02), Mysore, India, 2002.
- [3] J. Ramanand. *Towards Evaluating Lexico-Semantic Networks*. First Stage Project Report, 2006.
- [4] Xiao Fan Wang, Guanrong Chen. *Complex Networks: Small-World, Scale-Free and Beyond*. IEEE Circuits and Systems Magazine, First Quarter, 2003.
- [5] Duncan Watts. *The Dynamics of Networks between Order and Randomness*. Princeton University Press, 2006.
- [6] Ricard V. Sole. *Language Networks: their structure, function and evolution*.
- [7] Mariano Sigman, Guillermo A. Cecchi. *Global organization of the Wordnet lexicon*. Proceedings of the National Academy of Sciences of the USA. Vol. 99. Feb 2002.