Journal of Quantitative Linguistics, 2015
Vol. 22, No. 2, 157–176, http://dx.doi.org/10.1080/09296174.2014.1001638

Routledge
Taylor & Francis Group

Relational Model of Conceptual Distance between Bangla Words*

Sibansu Mukhopadhyay[1], Sreerupa Das[2] and Rajkumar Roychoudhury[2]
[1]Department of IT&E, Govt. of WB, Society for Natural Language Technology Research, Kolkata, India; [2]Indian Statistical Institute, Kolkata, India

ABSTRACT

Words in a language are related to each other. This relation is based on their conceptual properties. (This paper avoids using the term "semantic property", generally used by the contemporary NLP workers for measuring distance between words, the reason being that we employ different orientations behind the measurement of relatedness). Essentially, this work considers the psycho-sociological facts in the experiments, where a number of native speakers of Bangla manually suggests distance measurement between any two words. This work presents a statistical approach with a psycho-analytical elaboration for measuring the conceptual distance between words in terms of Bangla language. To be precise it calculates co-relations of the assessments collected through a survey among different individuals. A conceptual distance is used to suggest the implicit pragmatic nature of the Bangla words and it also implies an elementary taxonomy for Bangla words. As a result, the conceptual distance between Bangla words in the semantic field can very usefully be quantified and thus can be a crucial factor for a computational application like Bangla word net. Incidentally we find that there is a very high correlation ($r = 0.95$) between two different sets of human judgments and at the same time an assuringly high correlation ($r = 0.95$ being the upper limit) is observed when the respondents duplicated the same task with the same pairs of words at different points of times. This is a pioneering study in Bangla.

1. INTRODUCTION

Words in a language carry a number of conceptual properties. These properties are not necessarily physical. Ordinarily, from the grammarian's point of view, a word seems to have some intrinsic semantic properties generated through out the word formation process. However, native speakers of a

*Address correspondence to: Rajkumar Roychoudhury, Indian Statistical Institute, Kolkata, India. E-mail: rajdaju@rediffmail.com

language assign some values on a particular word. The values depend on the social concept, the context in which the word is used and the speakers' upbringing etc.

Treating word as a sign was started after Saussure's diction on signifier–signified relation in language. According to Saussure, sign consists of two parts, i.e. a "signifier" (signifiant) and a "signified" (signifié). Sign takes the form as signifier and the sign represents the concept as signified (De Saussure, 1983, p. 67; De Saussure, 1974, p. 67). The relationship between the signifier and the signified is referred to as "signification". Consider a word, for example, manuS "human". manuS is a sign, which takes a linguistic form (word) "manuS" as a signifier and the concept of manuS as it represents the signified of the signifier. Thus we follow this model of "signification" proposed by De Saussure (1983) and consider words as arbitrary signifiers loaded with social values.

Most of the modern trends in linguistics consider language studies objectively. These trends pre-suppose language as a body. Semantics find the relationship between the signifier and the signified. For example, word as a composition of speech sounds means an object or a concept or any kind of entity that may or may not exist. This is the process of signification, which is a social phenomenon. We also consider this process as a social phenomenon, oriented to the social structure and thus signifiers (words) have some common, socially co-related properties underlying ontological design among them. For example, if we consider x as a word we can certainly relate it to a signified denoted by X here. We believe x has a set of properties acquired by the social processes, some of which intersect with the set of properties y where y signifies Y. Consider Table 1 given below to illustrate the model we are discussing. Suppose we take three signs: x, y and z. We agree to take x for manuS "human", y for pakhi "bird" and z for rickshaw "man-driven vehicle". Now follow the table.

In column 1, x is a signifier and X is signified. Similarly y and z are signifiers and Y and Z are signified in column 2 and 3 respectively. We

Table 1 Words and their conceptual ideals.

1	2	3
manuS 'human' (x) (X)	pakhi 'bird' (y) (Y)	rickshaw 'man driven vehicle' (z) (Z)

draw an inference that establishes that *x*, *y* and *z* are equated with *X*, *Y* and *Z* respectively. We propose in this paper that every word (signifier) is in some way connected with each other. Therefore, *x*, *y* and *z* are interconnected to each other in terms of conceptualization or thought process: human and bird are animals, both have two legs and have power to produce sound, etc. and the third sign *z* stands for a type of vehicle which is one of the means of transport used by humans. However, apparently it is not connected semantically with a bird. We propose that a bird and a rickshaw are possibly interconnected in terms of the human thinking process. For example, a human may see a bird passing on a tree when she is riding a rickshaw or there may be other means by which a bird can be associated with a rickshaw.

Linguistic entities like word, in a semantic space, share several common possessions through which they are inter-related, though their conceptual meanings may differ. This commonness or relatedness can be measured in many ways. Quantification, deploying network representation for semantic similarity between words in a language is now a popular trend in the studies of natural language processing, cognitive science as well as psycholinguistics. This paper tries to quantify relatedness between Bangla words with the help of the experiments duly conducted for the relatedness between a set of selected pairs of Bangla words. In this paper a different coinage is chosen to convey our presupposition to the question of subjectivity in the data collection procedure. We use Conceptual Relatedness (henceforth we shall use the abbreviation CR) instead of semantic distance or similarity as a central theme of this study.

CR or more traditionally semantic similarity between words is a context dependent phenomenon. It is necessary to trace the conceptual relation between words to calculate semantic similarity. Properties in the words should be common even if we conclude that these words are semantically not very similar. For example, the words like *Cow* and *Horse* are not semantically similar to each other, but they are conceptually related in terms of the social contexts. Everyone knows that both are considered as domestic animals and historically both cows and horses were used in agriculture and transportation (Bollegala, Matsuo, & Ishizuka, 2009).

It is very crucial to distinguish our aim in this project from the other approaches and to show why we concentrate on the terminology *Conceptual Similarity* using adjective “conceptual” instead of “semantic”. In this paper we have focused at the lexical units, as these are essentially social

phenomena manufactured and value added by the speakers. We believe there are some idiosyncracies as well as some objective perspectives in the language. The idiosyncrasy can be traced, at least qualitatively, by the measurement of CR among different speakers. We also believe that a word is subjectified by a speaker"s social and psychological preferences. In our experiment among the Bangla speakers, we have done a projection which effectively reviews psycho-sociological status of the individuals, albeit through quantified results. We saw that this basically depends on the various facets of human mind. Therefore, it is more effective if one interacts with a social being who points out the distance between two or more words looking around his/her habitation than looking into a pair of words as the dictionary-entities with their fixed semantics.

2. RELATED WORK

We found (Resnik, 1995) that the occurrence of measuring semantic similarity in taxonomy happens among the lexical existences while reviewing the existing literature associated with the study of semantic similarity. We found that the research in this area started with Quillian (1968) and later Collins and Loftus (1975).

The motivation for calculating semantic similarity arises from the fact that although words like *bird* and *airplane* are apparently closer to each other than the pair *bird* and *pond*, the data collected from a set of people with a certain social background may suggest the opposite. Strong correlation between the opinions collected from two different sets of individuals through a psychological test gives us an idea about why the words *bird* and *pond* are perceived as closer to each other than *bird* and *airplan*e. The psychological test makes generalizations (Goldstone, 1994). To determine the probability of relatedness (the concept will be explained later on) between two words we assume that the units of conceptual properties depend on the common conceptual properties of these words. This kind of statistical measurement helps to extract synonyms (Lin, 1998) and retrieve lexical information (Sahami & Heilman, 2006), which is necessary for making an ontological network like WordNet.

To compute semantic similarity one can use the web-based metrics. Some works try to explore cognitive process using the model of priming. In these works it is believed that a speaker has an implicit memory

which by certain psychological behavioural exercises links to the stimuli basically appraised by his or her experience. It is also believed that priming cannot be pre-supposed; it is considered as an automatic process (Sánchez-Casas et al., 2006) The relatedness between words is supposed to reflect the concept of words rather than the anticipation of a formal meaning (Thompson-Schill et al., 1998). In a recent work on priming of Bangla words (Dasgupta et al., 2010) a cross-modal priming experiment was conducted to identify the mental representation and to get access strategies for morphologically derived words in Bangla. It is observed that morphologically transparent words do prime each other despite their phonological associations. However, morphologically opaque but phonologically transparent Bangla words do not show any priming effect.

3. CONCEPTUAL DISTANCE

There are many ways through which the relationship between words can be established, although all kinds of relationship are not to be considered as conceptual relations. We found a strong correlation among the respondents relating pairs of words. The following parameters can be considered in this regard.

3.1 Phonetic Similarity

Two or more words can be related if they are phonetically similar. There is no reason for taking formal semantics as a tool to extract relatedness between these types of words. And it is not the case in conceptual relationships. We have seen that phonetic similarity also triggers a human's cognitive ability to read or to hear words close. For example, in poetry the rhyming words and alliteration relate words within a closed frame. Let us consider the following example:

gOgone gOroje megh, ghOno bOroSa.
Kule Eka bose achi, nahi bhOroSa
"*Clouds rumbling in the sky; teeming rain.*
I sit on the river bank, sad and alone."

(the literal meaning of the words "nahi bhOroSa" would be "without hope/trust").[1]

Note: this poem shows that the final words ("bOroSa" and "bhOroSa") of the two sentences are phonetically similar. If we now ask some native speakers to calculate the distance between such phonetically similar words, the anticipated average figure would be very low, though the words are semantically very distant.

3.2 Relationship Between Synonymous Words

Synonymous words are indeed semantically similar. And this type relatedness is as usual very conceptual. Language has now, as an objective of a modern enterprise (like linguistics), a symbolic dichotomy to be claimed as an established relation between signifier and signified. But it would be fallacious to presuppose that a signified really exists and the chain between signifier and signified is viable. In other words, synonyms disprove the presupposed mono-typical existence of a signified. For example, two or more synonyms essentially are used for separate purposes of speech. Let us consider some real life examples from Bangla.

In Bangla there are many words signifying configuration of feminine as a standard cover term. But, native speakers use many synonymous words, elaborated below. To distinguish conceptual differences between them, we have collected data from the dictionary of Samsad digitized by the University of Chicago.[2] A few examples are given below.

- stri: a wife; a married woman; a woman
- nari: a woman; womankind; a wife
- lOlona: a woman; a gentlewoman, a lady; a wife. (Rarely used in daily speech, specially used in archaic poetry, or for making fun or punning.)
- mohila: a lady, a gentlewoman; a woman.
- bodhu: a wife; a newly married woman; a bride; a married woman; a daughter-in-law

[1]This is a piece from "Sonar Tari" of Rabindranath Tagore and is translated by William Radice, (freely available on the internet).

[2]We have used data from Samsad Bengali-English dictionary digitised by the University of Chicago. In a personal correspondence, James Nye, the Bibliographer for Southern Asia, and Director, South Asia Language and Area Centre, The University of Chicago, informed us that there are approximately 20,950 headwords in this dictionary. See http://dsal.uchicago.edu/dictionaries/biswas-bengali/.

The five words listed earlier, chosen from various terms, are used as synonyms for woman in Bangla. However, such dictionary entries cannot clearly distinguish any useful differences between these terms.

Although these five words are semantically related to each other, we can paradigmatically change these words, as will be clear from the following sentences and thus we can test whether these words are replaceable or not. As we stated earlier we use different synonyms for certain purposes.

(1) stri jatir unnoti-i deSer unnotir prothom dhap
"Advancement of Women is the first step towards the welfare of the state."
(2) nari jatir unnoti-i deSer unnotir prothom dhap
"Advancement of Women is the first step towards the welfare of the state."
(3) narirai puruSer calika Sokti
"Women are men"s guiding force." (Here "narira" is plural of "nari" and "narijati" means only Women.)
(4) lOlonarai puruSer calika Sokti
"Women are men"s guiding force."
(5) je mohilaTike apni dekhchen tini aSole ei ONcOler netri
"The lady you see is, in fact, a leader of this area."
(6) *je bodhuTike apni dekhchen tini aSole ei ONcOler netri
(7) "The bride you see is, in fact, a leader of this area."

Among the examples, (a) and (b), (c) and (d), (e) and (f) are pairs of sentences in which we have just replaced words with their near synonyms. Though these all are synonyms to each other, we cannot put a word arbitrarily without contextualizing the discursive information. For this reason, (d) and (f) and the following sentences are to be considered as culturally un-authentic or inconvenient, although they are not grammatically unacceptable sentences in terms of Bangla. Now consider the following examples.

(1) je nariTike apni dekhchen tini aSole ei ONcOler netri (The lady you see is a local leader.)
(2) baRite to moTe dujon thaki, ami ar amar nari (Only two people live in this house: myself and my wife.)

Here "nari" is used for "lady" and wife respectively.

We cannot say that these two sentences are grammatically unacceptable, but we must say that a native speaker of Bangla cannot agree with this expression. We cannot even say, "bou bOr bhed na dekhe manuS khuMje dEkho" instead of "stri puruS bhed na dekhe manuS khuMje dEkho", though the words "*stri*" and "*bou*" are semantically related. The sentence translated into English is "look for a human being without discriminating between a man and a woman". If we use "bou" (bride) for woman and "bOr" (husband) for man, the sentence loses its universality.

There is also a chain of words, around the abstract concept of "female" in Bangla, which are co-related. For example, as Noun, there are many, such as; nari <> stri <> rOmoni <> lOlona <> ONgona <> kamini <> bonita <> bodhu (> bou) <> mohila <> ghoSit <> konna <> magi (a plang word). There are more words indicating females which are used normally as adjectives, for example balika (girl) <> toruni (young woman) <> briddha (old woman) <> prouRa (old woman but not as old as briddha) <> kumari (unmarried woman) <> bibahita (married woman). Each of the above examples has a distinct set of semantic properties, but has also some commonness. All sets have intersection points, through which one may establish their ontological relationship. Another thing to be noted is that there is a nucleus (abstract/physical/biological/objective) meaning in the words listed above, which substantially correlates the terms.

3.3 Conceptual Relation between Words

The main purpose of this paper is to propose a hypothesis that every word is conceptually linked up with other words, though they are not necessarily synonyms. Moreover, whenever we look at the semantic network between the words, we see that there are certain attributes socially assigned to each of the words at any level of the concept that make a word relate to other words, as discussed in the introduction. Let us now show how the words are termed as CR, are linked to each other conceptually.

The term Semantic Distance is mostly standard dictionary oriented. In this work, we leave the options to the speakers to calculate distance between words instantaneously. Therefore, this assessment procedure becomes more intuitive than consciously determined. To summarize our point of view, the use of words or use of such units of language is flexible in the sense that it depends on the respective psycho-social factors in generating conceptual properties of word. This is the reason for avoiding term like *semantic property*.

Let us consider here two pairs of words between which we have to establish CR. These are; bhaSa ("language") – bakko ("sentence") and paHar ("mountain") – nowka ("boat"). Before we describe the survey methodology let us first try to anticipate the possible relationship between the pair of words just mentioned viz. "bhaSa-bakko" and "paHaR-nowka". A native speaker (in Bangla) is expected to find "bhaSa" and "bakko" to be very close together. In fact "bakko" is considered as a part of "bhaSa" or both may be just part of "ukti" (utterance). However, the same cannot be said of the words "paHaR" and "nowka". "paHaR" or mountain is a part of nature existing from geological times and "nowka" is man-made vehicle for river or sea transport or transportation for any water body. However, for a person who loves nature and spends his/her holidays at mountain resorts, seaside or riverside may consider these two words not too distantly related as they are part of what may be called tourism.

We have surveyed 30 native speakers and found that the speakers varied in their opinions and marked arbitrarily, following a scale of measurement for CD. The details are given in Section 4. The average opinion says "language" and "sentence" are more closely related to each other than "*mountain*" and "*boat*". Now we draw two trees, nodal distances of which indicate the CD between the words.

As argued before, from the first tree (Figure 1) for the pair bhaSa "language" – bakko "sentence" it is easy to understand that as the nodal distance between language and sentence is very short, one concludes that the conceptual distance between these two words is small. On the other hand, sentence is considered as a part of language, i.e. sentence is one of the members of the set called "language".

The second case which is also discussed in some detail (Figure 2) is more complex than the first one. What we see in the second tree is that mountain may be anticipated under two different nodes, but for establishing

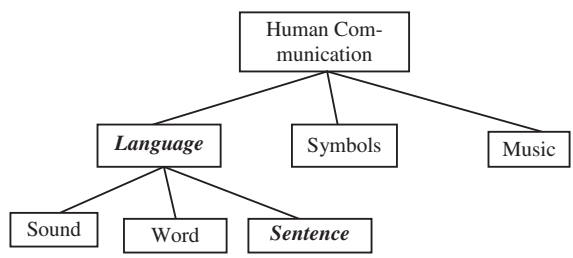

Fig. 1. Conceptual tree for "Language" and "Sentence".

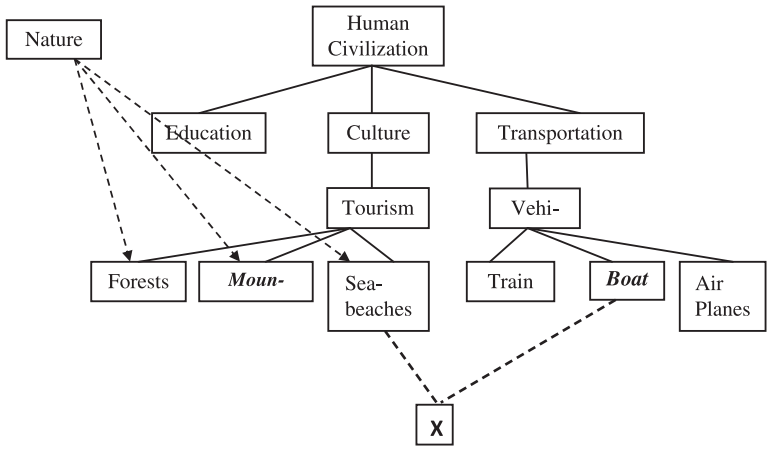

Fig. 2. Conceptual tree in large scale.

the relationship we have considered it under the node tourism as we can connect it to the boat more closely.

It should be noted that this map is developed on the basis of a re-construction mode. But people generally hold the conceptual ideas according to some predictive social pre-consciousness. For example, one surveyee thought that the distance between pakhi "bird" and rikSa "rickshaw" (a manual vehicle) is small. When a surveyee was asked how she designated such closeness, she stated that on that particular day when she was coming to her office, she saw a bird, and she was in a rickshaw. Or it may be that she found a sticker with a picture of a bird pasted on the back of a rickshaw. It happens quite often in Kolkata.

Now consider two more examples to examine this issue with a stretched explanation. Consider the pairs "boat-ship" and "president-king". Conceptually, "boat" is nearer to "ship" than "president" to "king" is. On the other hand "president" is a word nearer to "king" than a "ship". But "king" and "ship" are also related in terms of common conceptual properties like "big" and "large" (i.e. two sets of conceptual properties of these two words intersect in terms of the common member of the sets). For example, we can establish the CR between "king" and "ship" on the basis of, for instance, $P1$ and $Q1$, therefore, $P1 \approx Q1$. This relation is an ontological relationship. Ontology classifies exhaustively entities of a being or a domain or a conceptualization. Depending on some informative features we can map the relationship between the entities and among those vast features we

invariably find the commons between any two entities. This point leads to the idea of an abstract process of explanation for the relations and classifications of the entities or the events in a domain. Thus we can consider this explanation as methodological motivation to describe a domain.

4. CURRENT METHOD

Here, we discuss some methodological parts that track the trajectory of our work. We basically depend on the ontological relationship (as we discussed in the previous section) between two words (as compositions of certain concepts). Ontological mapping is the best way to find out the CR between the words in a language or across different languages. CR is such a function as described in introduction where a set of conceptual properties within the term of a word are assigned to another set of conceptual properties of a different word based on their psycho-sociological expressions.

Let us consider X as a family or a domain, where X_1, X_2, X_3 and X_x are the members (Figure 3). There are also other members like X_4, X_5 whom we do not consider for conceptualization. But we know that they exist. More over it is possible that there are other members of whose existence we are not aware of. If we are going to present ontology of the family, we have to describe the individual members, their positions, the relationship between the members and the possible members through a diagram. The arrows show the directions by which a human switches her thought from one object to another.

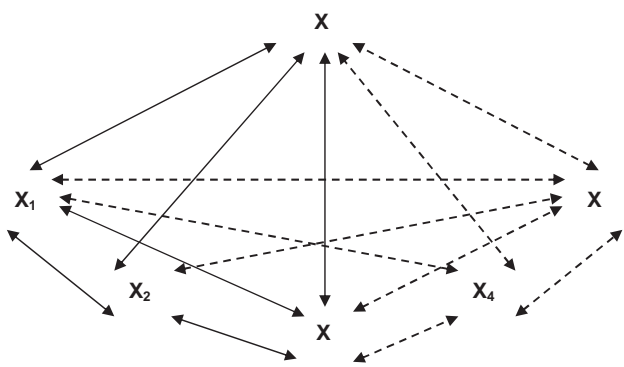

Fig. 3. Ontological pattern of conceptual relation.

Best examples for such categorization schemes are periodic tables or library catalogues. Browsing a library catalogue one gets to know of high-order categorization. The cataloguing systems develop all kinds of mapping between the events and the conceptualization they specify. A word itself is a result of conceptualization that evolved socially. We have kept this in mind while arranging the data sheet.

5. EXPERIMENT

Presently, it is clear that if we have to realistically calculate the distance between the words we have chosen, we cannot go with any pre-supposed quantitative status for the words in a language. We have discussed earlier that such type of pre-supposition is highly dependent on dictionary meaning of a word. Although there are several individuals who set their apparently random responses in manual calculation for measuring the distance between two different words, we find that the correlation between those different individuals is very strong. Moreover, a high correlation is observed when the respondents duplicated the same task at a different point of time. It implies that the association between two words is not really random and may not be a function of time. In the following we discuss this in some detail.

Survey Procedure: Twenty pairs of Bangla words were given to 30 individuals with similar academic and linguistic background. The 30 respondents were divided into two groups, each comprising 15 members, to find out correlation on bias between the groups. It was assumed that the distance between two words can be mapped on to a set $S \in R_+$, where R_+ is the set of real positive numbers including zero. If $x \in S$, then we fixed the domain of x as $0 \leq x \leq 4$. For example, if two words are just synonyms then ideally the distance between them will be zero, whereas two words apparently not semantically related to each other in anyway will be given a score 4 or thereabouts. The respondents were asked to assign the value of the distance by any rational number (up to one decimal point). To reduce the time dependence of CR the respondents are asked to duplicate the task at different points of time. The average score by the respondents is given in Appendix I.

Figure 4 depicts fragments of possible tree diagrams where we associate the words “crow” and “cuckoo” and “rickshaw” and “bird”. The diagram supports the survey data which indicated that “crow” and “cuckoo” are much closer than “rickshaw” and “bird”. In the diagram, against each nodal joint (from where words branch out), the corresponding IC is given. Now it

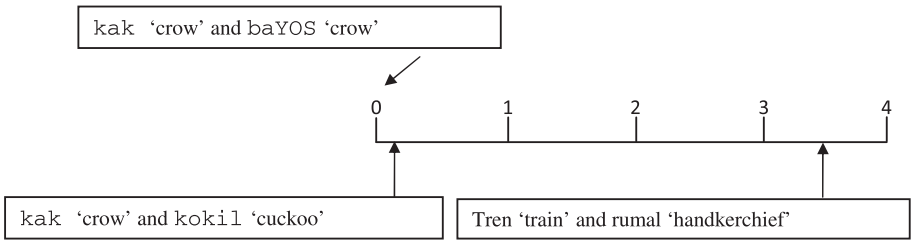

Fig. 4. Scaling the Conceptual Distance between Words.

must be mentioned here that there is no word net available in Bengali. For corpus of words we relied on the University of Chicago document.

The diagram implies that mentally one can distinguish between two close rational numbers; for example, 2.1 and 2.2 while putting a number against a pair. To illustrate the scheme further let us consider the pair of words kak and bayOS (cf. Figure 4, the English meaning for both words is "crow"). These words are synonymous and it is expected that the respondent will measure the distance as zero or very near to zero. Now consider the words "kak" and "kokil" – "cuckoo", though these birds don"t belong to the same species both these birds are common in Bengal. However "kokil" is generally seen or heard only in the spring time.

Also almost all residents of Bengal know that the female cuckoo lays its eggs in a crow"s nest and as the eggs are similar to those of a crow the cuckoos" off-springs are brought up by a crow family. This is another relationship between these two birds. Often peoples" voices are compared to either to that of a crow or to that of a cuckoo. They lie on the extreme of the spectrum of voices because while the crow"s voice is harsh and dissonant; the cuckoo"s voice is melodious. So it is expected that the respondents brought up in a typical Bengal culture would make these two words very close to each other.

Let us now consider the following pairs of words (also cited in Figure 4) having apparently no immediate relationship: Tren "train" and rumal "handkerchief".

It may be that in any Western country people will measure the distance between them almost near to 4. But it may happen that people in India who frequently travel by train find, more often than not, hawkers selling handkerchief in trains. So they may associate these two words together and may not give the maximum score.

Though the three pairs of words mentioned earlier were not included in the sample given to the respondents there was one pair of words for which we found interesting responses. These words were discussed earlier. The words are rikSa "rickshaw" and pakhi "bird" (see Figure 5). Some respondents thought that the distance between them was small. The reason behind this mental process has been speculated earlier. However, there is a statistic which can measure whether the random deviation from the mean is statistically significant. This is the standard deviation, denoted by σ, which is defined by $\sigma^2 = (1/n)\ \Sigma\ (x_i–\mu)$ ^2, where, $i = (i = 1, 2, \ldots, n)$, x_i $(i = 1, 2, 3, \ldots, n)$ are the scores by the individuals and μ is the mean given by Σx_i, $\mu = 1/n$ while n is the total number of participants (which, in the present case, is 30).

For the pair of words "rikSa" (rickshaw) and "pakhi" (bird) the mean distance is 3.25 and $\sigma = 0.22$. The measure of deviation given by CV = standard deviation/mean = 0.07, which is really small. It signifies that there is strong consistency among individuals in measuring distance between a pair of words and this encourages us to speculate that the cognitive process is rather uniform for this group of participants, though there may be variation from individual to individual. We also found strong correlation ($r = 0.95$) between the two groups as far as mean distance is concerned. Moreover, an equally strong correlation (upper limit being 0.95) is observed within a group when the participants were asked to duplicate the same task at different points of time. This means the score by the participants can be considered to be almost time independent.

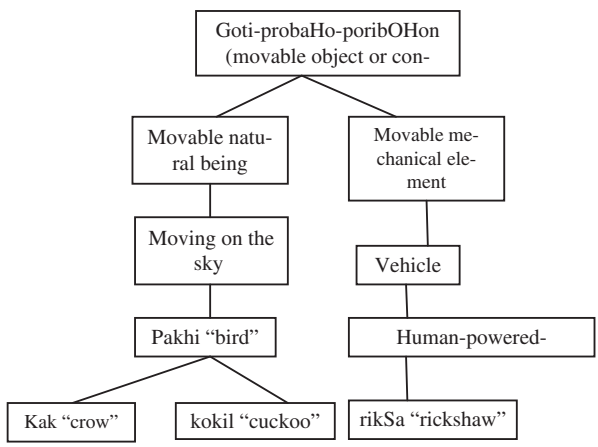

Fig. 5. Conceptual tree: "rickshaw" and "bird".

5.1 Semantic Similarity Based on Corpus Linguistics

To correlate the survey data with the theoretical distance (semantic similarity) we follow the statistical analysis of previous authors, who worked on this topic (Jiang & Conrath, 1997 and references there in). Here we introduce the idea of information content (IC). IC of a concept c can be quantified as IC (c) = $-\log p(c)$ where $p(c)$ is the probability of finding an instant of concept c. It is actually what is called entropy in information and natural sciences. Following Richardson and Smeaton (1995), we shall define $p(c)$ as $p(c) = \text{freq}(c)/N$, where, freq(c) is the number of total entries under the concept c from which the pair of word under consideration are derived (here we consider the nearest node). Concrete numerical examples are given below.

Figure 5 depicts fragments of possible tree diagrams where we associate the words "crow" and "cuckoo" and "rickshaw" and "bird". The diagram supports the survey data which indicated that "crow" and "cuckoo" are much closer than "rickshaw" and "bird". Now it must be mentioned here that there is no WordNet available in Bengali. For corpus of words we relied on the Chicago University document and Samsad Samarthasabda Kosh.

The total number of words in Samsad Samarthasabda Kosh is 62,500. Now, for example, we can measure frequencies of some pairs of words:

(1) *bakko "sentence" and bhaSa "language".*

We have shown a relationship between bakko "sentence" and bhaSa "language". These two words, as we have calculated a tree diagram (Figure 1) depending upon the information of Samsad samarthasabda kosh, are derived from a node called "ukti" "utterance". The total entry under "ukti" from which "bakko" and "bhaSa" are being derived is 45. Therefore, the corresponding IC = $-\log 45/62{,}500 = 7.24$.

The high value of IC indicates that these words are highly correlated which agrees with the survey results.

(2) *paHaR "Mountain" and Nouka "Boat".*

"paHaR" and "nouka" can be anticipated as the derivations from a common node, bhromon (tourism). Total number of entry words under "bhromon" is 162. Then, the related IC is IC = $-\log 162/62{,}500 = 5.95$

This pair of words also gives a high value. However. the participants found these words are not close to each other.

(3) *Pakhi "bird"- rikSa "rikshaw" and kak "crow"- kokil "cuckoo".*

In Samsad samarthasabda kosh rikSa (রিকসা) is accumulated under the heading of *Goti-probaho-poribahan* "speed-wave-transportation". "Goti-probaho-poribahan" consists of 2160 entries. But there is no entry for Pakhi.

Samsad Samarthasabda Kosh does not give any clue for drawing any tree diagram for pakhi and rikSa also there is no WordNet in Bangla. If we can assume a node such as "Goti-probaho-poribahan", under which pakhi and rikSa could possibly be undertaken, we can assumingly calculate some results. Then for "Goti-probaHo-poribOHon" (movable object or concept), the IC is IC = −log 2160/ 62,500 = 3.36 (medium index).

The participants did not find these words close to each other. In a same way, we have calculated Kak and kokil, which very obviously branched out from a common node, pakhi (পাখি) "bird". Entry under "pakhi" in Samsad samarthasabda kosh is 340. Here the IC is IC = −log 340/62,500 = 5.21. This indicates that the words are close but this pair of words was not in our list.

6. CONCLUSION

In this paper an attempt has been made to shed some light on the cognitive process by which a subject relates a pair of words in a particular sociolinguistic context. To set a quantitative measure of the distance between two words a survey was conducted. Thirty people with similar socio cultural background were given 20 pairs of Bangla words each and were asked to find distance between them within a particular range. Some of the pair of words seem to be immediately related to each other but there were others which were not apparently related as for as semantic aspect is concerned. Take, for example, the word /puruS/ "male" and SiNHo) "lion". If one makes a tree, as shown earlier, one has to go to the word "mammal" perhaps, which includes a large number of creatures in the animal kingdom. However, some of the respondents found these words closer than what suggested by the tree diagram (Figure 6). This may be owing to the fact that the compound word *puruS-SiNHo* "male-lion" is of quite common usage in Bengali language. Compounds words like *puruS-SiNHo* "male-lion" and *puruS-bEghro* "male-tiger" belong to the class of compound words (samasa in Sanskrit) which is called (upameya karmadharaya) "puruS-SiNHo" means a man who is as brave as a lion.

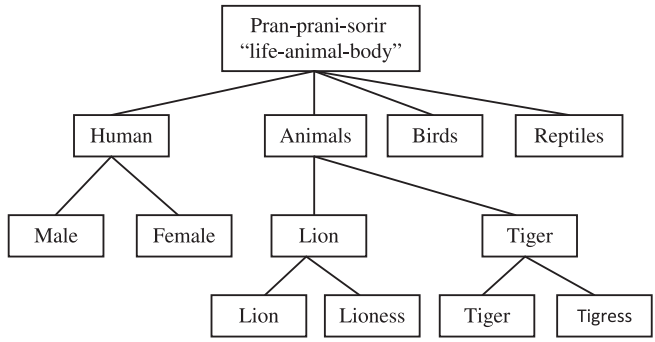

Fig. 6. Conceptual tree for "male" and "lion".

Hence it was not surprising that the words puruS and SiNHo were thought to be close to each other and the mean distance from the respondents" scores turned out to be 1.33 out of 4. The words like baHon "vehicle" and iNdur "rat" (these words are not the part of the survey samples) would seem further away than the words "puruS" (male) and "siNHo" (lion). However, any Indian brought up in Hindu culture or acquainted with Hindu mythology would immediately establish a relation between the two words as Indian mythology "iNdur" (rat) is the vehicle of the god lord Ganesha who is widely known outside India as the elephant god. In the Appendix the average scores assigned against the pair of sample words by 30 individual are given.

$-\log(p)$ calculation for puruS "man" – SiNHo "lion" is given below. These two words have derived from a common higher node, i.e. Pran-prani-sorir "life-animal-body", entry under which is 1115. $-\log 1115/62{,}500 = 4.03$.

Here the index suggests that the words are related to a certain extent. But the survey suggests that these words are pretty close.

One can find many instances where semantically distance words may seem closer to a particular set of individuals brought up in any specific socio-cultural background. The study made here may be considered as a pilot study on a subject on which, to the best of our knowledge, no work has been done so for in Bangla and very little in other Indian languages. We hope our study will stimulate similar studies in Bangla in particular and Indian languages in general and would be helpful towards building up a proper Bangla WordNet.

ACKNOWLEDGEMENT

The survey used in this study was conducted with the help of Baidehi Sengupta and enormous supports for some other issues came from Rimi Ghosh Dastidar. We are thankful to them. We are also grateful to all the people who participated in the survey. We are grateful to the referee for constructive suggestions.

REFERENCES

Bollegala, D., Matsuo, Y., & Ishizuka, M. (2009). A relational model of semantic similarity between words using automatically extracted lexical pattern clusters from the web. *Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2009)*, pp. 803–812 http://www.iba.t.u-tokyo.ac.jp/~danushka/papers/danushka-EMNLP2009.pdf

Collins, A., & Loftus, E. (1975). A spreading activation theory of semantic processing. *Psychological Review*, *82*, 407–428. http://homepage.psy.utexas.edu/homepage/faculty/Markman/PSY394/CollinsLoftus.pdf

Dasgupta, T., Choudhury, M., Bali, K., & Basu, A. (2010). Mental representation and access of polymorphemic words in Bangla: Evidence from cross-modal priming experiments. *International Conference on Natural Language Processing (ICON)*, 58–67.

De Saussure, F. (1916 – 1974). *Course in General Linguistics*. Tr. by Wade Baskin. London: Fontana/Collins.

De Saussure, F. (1916 – 1983). *Course in General Linguistics*. Tr. by Roy Harris. London: Duckworth.

Goldstone, R. L. (1994). Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Memory and Cognition*, *20*, 3–28. http://cognitrn.psych.indiana.edu/rgoldsto/pdfs/siam.pdf

Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. International Conference Research on Computational Linguistics (ROCLING X) (September 1997)

Lin, D. (1998). Automatic Retrieval and Clustering of Similar Words. *COLING-ACL98*, Montreal, Canada, August 1998. http://webdocs.cs.ualberta.ca/~lindek/papers/acl98.pdf

Quillian, M. R. (1968). Semantic Memory. In M. Minsky (Ed), *Semantic Information Processing* (pp. 216–270). Cambridge, MA: MIT Press.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a Taxonomy. Proceedings of the 14th International Joint Conference on Artificial Intelligence, *1*, pp. 448–453, Montreal, August 1995.

Richardson, R., & A. F. Smeaton. (1995). "Using WordNet in a Knowledge-Based Approach to Information Retrieval", Working Paper, CA-0395, School of Computer Applications, Dublin City University, Ireland.

Sahami, M., & Heilman, T. D. (2006). A web-based kernel function for measuring the similarity of short text snippets. *Proceedings of the 15th International World Wide Web Conference (WWW)*, Retrieved 2006, from http://robotics.stanford.edu/users/sahami/papers-dir/www2006.pdf

Sánchez-Casas, R., Ferré, P., García-Albea, J. E., & Guasch, M. (2006). The nature of semantic priming: Effects of the degree of semantic similarity between primes and targets in Spanish. *The European Journal of Cognitive Psychology*, *18*, 161–184. http://psico.fcep.urv.es/projectes/gip/papers/sc_f_ga_g_2006.pdf

Thompson-Schill, S. L., Swick, D., Farah, M. J., D'Esposito, M., Kan, I. P., & Knight, R. T. (1998). Verb generation in patients with focal frontal lesions: A neuropsychological test of neuroimaging findings. *Proceedings of the National Academy of Sciences*, *95*, 15855–15860.

APPENDIX I

Average scores given against each word. 1st Phase and 2nd denote different points of time.		1st phase sample			2nd phase samples		
		1 to 15	16 to 30	1st phase mean	1 to 15	16 to 30	2nd phase mean
Word pairs no.							
টবেলি (Table)	জামা (Shirt)	3.27	3.51	3.39	3.24	3.54	3.38
কম্পডিটার (Computer)	গাছ (Tree)	3.17	3.78	3.47	3.43	3.70	3.56
দোকান (Shop)	বই (Book)	1.31	1.45	1.38	1.41	1.50	1.45
সংিহ (Lion)	পুরুষ (Male)	1.44	1.50	1.47	1.62	1.00	1.31
বিকেল (Evening)	মাস (Month)	2.78	3.08	2.93	2.22	2.20	2.21
অনাহার (Starvation)	বিশ্বাস (Belief)	3.29	3.12	3.20	3.38	3.10	3.24
পাখি (Bird)	রিকশা (Rickshaw)	3.25	3.75	3.5	3.55	3.85	3.70
চোখ (Eye)	অট্টালিকা	3.10	3.18	3.14	2.93	2.71	2.82
জল (Water)	রক্ত (Blood)	2.07	1.58	1.83	2.09	1.55	1.82
লীলা (Significant but unintelligible work or sport)	কেলি (Amorous sport)	1.30	1.28	1.29	1.03	1.33	1.18
বাসনা (Hope)	সংকতে (Signal)	2.27	2.85	2.56	2.78	3.18	2.98
মন্দরি (Temple)	মুখ (Face)	2.46	2.51	2.48	2.38	2.80	2.59
বাক্য (Sentence)	ভাষা (Language)	0.83	0.56	0.70	1.11	0.81	0.96
সাহত্যি (Literature)	জঙ্গল (Forest)	1.93	2.69	2.31	2.34	2.88	2.61
ছেলে (Boy)	ভোজন (Eating)	2.48	2.37	2.43	2.53	2.35	2.44
পাহাড় (Mountain)	নৌকা (Boat)	3.05	3.61	3.33	2.65	3.19	2.92
মন (Mind)	দেখা (Seeing)	2.02	1.80	1.91	1.23	1.23	1.23
তরঙ্গ (Wave)	ফুল (Flower)	3.13	3.05	3.09	3.49	2.75	3.12
ঘাতক (Slaughterer)	খুনী (Murderer)	0.29	0.47	0.38	0.40	0.72	0.56
চুল (Hair)	চশমা (Spectacles)	2.92	2.96	2.94	2.97	2.81	2.89