Graph theoretic interpretation of *Bangla* traditional grammar

Samir Karmakar
School of Languages and Linguistics
Jadavpur University
samirkrmkr@yahoo.co.in

Sayantani Banerjee
School of Languages and Linguistics
Jadavpur University
banerjeesayantni@gmail.com

Soumya Ghosh
School of Languages and Linguistics
Jadavpur University
ghosh.soumya73@yahoo.com

Abstract

The paper is an investigation into the graph theoretic interpretation of the *Bangla* traditional grammar to understand the way grammatical information is structurally encoded in language. The hierarchical and the linear structural principles of grammatical compositionality is discussed in terms of certain graph theoretic concepts like tree, subtree, inverse tree etc.

Translating linguistic structure into the tree structure is not new. In fact, the Transformational-Generative grammar, Tree adjoining grammar etc. have shown quite successfully how syntacto-semantic principles can be talked about in terms of tree structures. The present work differs in certain respects from the assumptions of TG grammarians, primarily because of the type of grammar and language it is dealing with.

1 Introduction

This paper seeks to investigate how substantially the structure of a language can be dealt with the aid of graph theory. Because of being qualified with the discrete structure, natural language can be represented and processed with graph-theoretic methods. Compositional nature of language makes it more viable to the concept of constituent hierarchies within the scope of which syntactic and semantic principles are operating. Under this situation, the graph theoretic interpretation provides excellent opportunities to explore the issues pertinent in structural composition. Certain contemporary models like head-driven phrase structure grammar, semantic networks etc. make good use of graph theoretic methods. WordNet, VerbNet etc. have also their deep connections with this branch of discrete mathematics. Under the influence of these approaches, the current paper seeks to investigate how grammatical regulations are crucial in imposing constraint on the structure with a special reference to the traditional *Bangla* grammar within the framework of graph theoretic enquiry.

Translating linguistic structure into the tree structure is not new. In fact, the Transformational-Generative (hereafter, TG) grammar has shown quite successfully how syntacto-semantic principles can be talked about in terms of tree structures. The present work differs in certain respects from the assumptions of TG grammarians, primarily because of the type of grammar and language it is dealing with. Not only the TG grammar, tree adjoining (hereafter, TA) grammar has also made a good use of the graph theory. Though both TG and TA have made use of the graph theory but definitely from the two different perspectives.

The current proposal resembles TA grammar more closely than the TG grammar; and as a result, the proposed model of language representation and processing can probably be classified in terms of weak generative capacity, a position between context free grammars and indexed grammars.

2 Research Objectives

Within the broader theoretical background as is discussed in Section (1), this paper will investigate the way grammatical structures of *Bangla* as is described in traditional *Bangla* grammar (Chatterji 1939) can be talked about in terms of graph theoretic assumptions. This is possibly the most salient point where the paper does differ even from its nearest kin TA grammar. Under this situation, following two questions will be investigated in this paper: (a) how the structural

D S Sharma, R Sangal and A K Singh. Proc. of the 13th Intl. Conference on Natural Language Processing, pages 129–136,
Varanasi, India. December 2016.

complexity of the *Bangla* sentence can be talked about with respect to graph theoretic assumptions; and, (b) if it is possible to develop a theoretical scheme to capture the syntacto-semantic peculiarities of the individual constituents of a sentence.

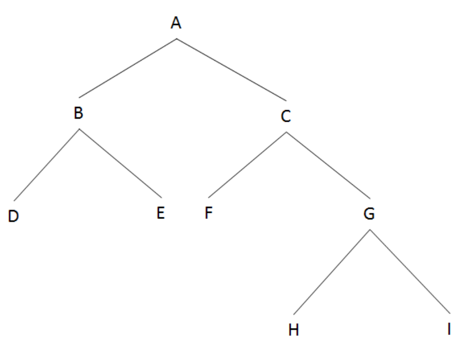
3 Theoretical Background

Approaching the above mentioned research objectives seeks the setting of the theoretical framework which presumes some fundamental understanding of the graph theory and the knowledge of traditional *Bangla* grammar.

3.1 Tree as a Graph

Graphs are not used for some sheer illustrative purposes; In fact, they reveal hidden intricacies involved in complex structures. As a consequence, it is quite essential to concentrate on some basic concepts which are crucial in explaining the construal of graph in general: Mathematically, graph is defined as a set of sets – one of which contains the nodes and other, a set of ordered pairs or edges. For the purpose of this paper, a particular type of graph will be discussed, namely 'tree'. A tree is defined as a graph in which each pair of vertices is connected with a unique path. Tree is characterized as acyclic and directed (Liu, 2001). An example of tree is given in the figure below:

1.

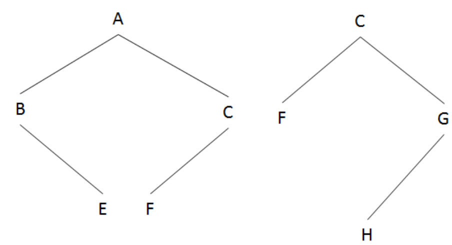


An alternative way to represent the graph of a tree is to follow the technique of embedded bracketing as is shown in (2). In this framework, the notion of hierarchical depth of a particular node can be interpreted in terms of its interiority in the bracketed representation. Furthermore, the concept of edge here in this case is inferred from an understanding of 'who embeds whom'.

2. [A [B [D] [E]] [C [F] [G [H] [I]]]]

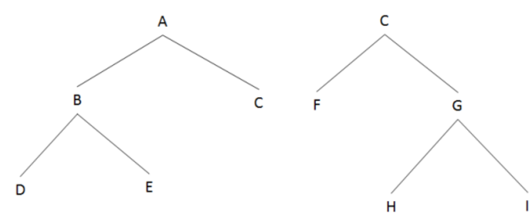
According to (1) and also (2), then, a graph of tree (= G_T) could be represented as {V_T, E_T}, where V_T is the set of vertices/nodes {A, B, C, D, E, F, G, H, I} and E_T is a set of edges {AB, AC, BD, BE, CF, CG, GH, GI}. Each of the members of E_T can be defined as a relation defined over the set of vertices. As per this definition, then, an edge is an ordered pair. In other words, an edge could be conceived as a relation from one vertex to another non-identical vertex. A tree (= G'_T) will be called as the subtree of G_T, iff and only if $V'_T \subset V_T$ and $E'_T \subset E_T$. For example, the trees drawn in (3) are the subtrees of (1):

3.



Two subtrees of a tree will be considered as inverse to each other if and only if their union can result into a tree of which they are the subtrees. Therefore, the following two subtrees are the inverse of each other because their union will produce the graph shown in (1):

4.

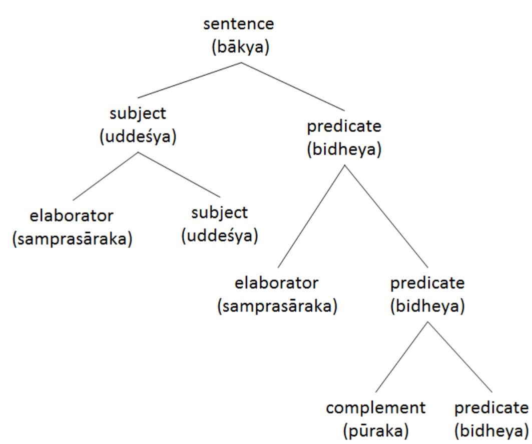


Just like the graph theory, traditional grammar of *Bangla* can also generate the trees. Therefore, it becomes quintessential to explore how and to what extent trees generated by the grammar of a language (here, *Bangla*) resemble the structural aspect of tree as the mathematical object.

3.2 Traditional *Bangla* Grammar

Likewise the western tradition, *Bangla* sentence (*bākya*) also has two distinct parts namely subject (*uddeśya*) and predicate (*bidheya*). Both subject and predicate can be modified; and, modifiers will be classified as elaborator (*samprasāraka*). As a result of elaboration, whatever is produced is classified again either as subject or as predicate. Elaborators are optional and they can appear in any numbers for finer grain specifications of the communicative intent. Unlike elaborators, predicate could have complement (*pūraka*). Complements are not optional; rather, they are the obligatory components of the predicates, and their numbers are often fixed by the semantic expectancy of the verbs. To capture the structural peculiarities of a sentence the notion of subject, predicate, elaborator and complement are extremely useful. In addition to this, the notion of *maximality* is proposed to incorporate the idea of the scope of complete interpretation for a particular structure. Within the maximal scope of a structure, syntactic and semantic necessities of a particular constituent are satisfied and beyond that scope these necessities have no role to play. Accordingly, the addition of complement to the verb in (5) makes their immediate dominator maximal in case of transitive *Bangla* verbs. Note this understanding of maximality is quite different than the one talked about by Chomsky. This issue will be picked up again in our successive discussions to exemplify the way maximality is instrumental in graph theoretic framework of language interpretation.

5.

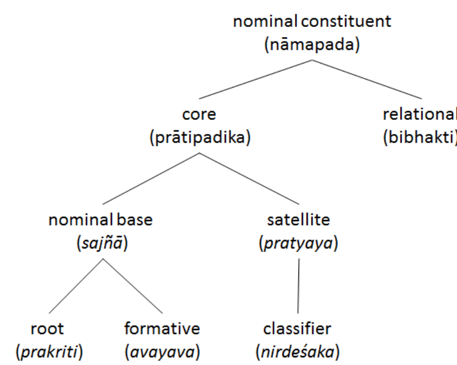


Note: A = sentence; B = subject; C = predicate; D = elaborator; E = subject; F = elaborator; G = predicate; H = complement; I = predicate

In addition to what is discussed earlier, we would like to incorporate the word-internal composition to understand the reach of the graph theoretic methods of language description. More importantly, studies solely depended on the morphology of English often fails to appreciate the linguistic intricacies involved in other vernaculars. Word internal morphological complexity is one such intricacy which demands our attention.

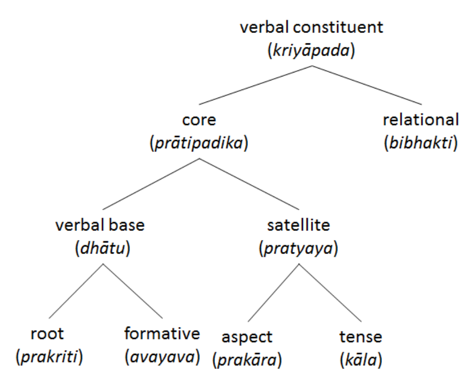
Morphological complexities of words in *Bangla* are more in case of nominal and verbal constituents of a sentence. Nominal and verbal – both types of constituents can be decomposed in two major parts: (a) a part constitutive of core meaning, traditionally termed as *prātipadika*; (b) another part constitutive of relational meaning, traditionally classified as *bibhakti*. In case of nominal, relational meaning denotes case relations and has direct relevance with the core meaning of the verbal constituents; in case of verbal, the relational meaning denotes agreement with the nominal and also contain information about the time. The core meaning is constitutive of base form and the satellite. Depending on how the constituent is classified, base could be of two types: namely, (a) nominal base (*sajñā*) and (b) verbal base (*dhātu*). Both of these two bases are further decomposed into (a) root (*prakriti*) and (b) formative (*avayava*). According to the nature of the base, the formative could be of two types namely gender signifying and causative. In case of verbal constituent, satellite is constitutive of grammatical aspect and tense, whereas classifier has the status of formative in case of nominal.

6.



The terminal vertices of (5) – excepting one marked as predicate (= I) – will be further augmented by (6); whereas, terminal position 'I' will be augmented with (7):

7.

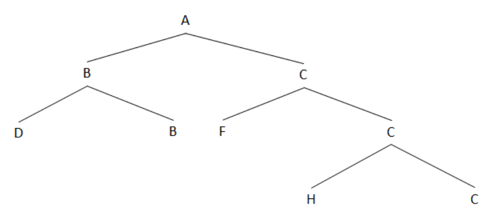


This will then lead us to a fully specified tree capable of representing sentences; however, without postposition (*anusarga*). Sentences containing postposition are kept aside keeping the scope of the paper in mind.

4 Discussion

After having the brief introductions of different theoretical tools, we will now investigate how graph theory and the traditional *Bangla* grammar interact with each other. In continuation to the above discussion, one can now suggest that instead of having a set of distinct vertices/nodes one can simply distinguish the non-identical vertices of (5) in terms of their respective 'maximality' – discussed earlier. Developing such mechanism can be done in two distinct stages: Firstly, to substitute the identical vertices with a single alphabet to indicate the similarities among them; and secondly, to capture the dissimilarities in terms of their respective syntactic and semantic properties certain conceptual measures have to be thought of. With the initiation of the first, (5) is simply transformed into (8):

8.



Similar vertices/nodes are reflecting their endocentric nature. In virtue of being endocentric, they are category preserving. Though the similarities of certain vertices are well-represented in (8), the dissimilarities are hardly traceable from this representation until the seemingly similar vertices are interpreted in terms of maximality. To understand this problem, consider the case of C: Are the three instances of C in (8) same? Are they similar type of 'predicate'? – A little attention will reveal the fact that they are not. Non-branching C can license the verbal constituent only, when the intermediating one has the provisions for the complement(s) (i.e. the terminal vertex H) and the verbal constituents (i.e. the terminal vertex C). Finally, the topmost C is projected due to the addition of elaborator with the complement(s) and the verbal constituent. This simply compels us to import one another concept, namely *hierarchical depth* or *relative interiority*. Hierarchical depth can be translated in terms of certain constitutional scopes. Within a scope different types of syntactic and semantic expectancies are satiated. Under this situation then it becomes essential to distinguish these three instances of Cs in terms of their respective scopes. This leads us to the following proposal: In (8), C of the non-branching vertex is the head. Being non-branching, the hierarchical depth or relative interiority of head C is more than the other Cs above it. Due to the successive projections, this head results into the appearances of C_{MAX} as maximal projection (hence, subscripted with 'MAX') and C_E as elaborated projection (hence, subscripted with 'E'). While getting projected maximally, complements are accommodated; whereas to get elaborated projection, elaborators are accommodated. The distinction between these two types of projections namely maximal and elaborated is instrumental in distinguishing complements from the adjunct. This solution is restricted not only to any specific subtree – rather it has some general appeal: Consider the case of multiple appearances of B. Following the general strategy, outlined above, non-branching B will be labeled as head. This is projected into a higher non-branching vertex B which has the status of elaborator (= D) and therefore must be represented as B_E. Since the elaborators are not the essential part of a lexeme, they are kept out side of the maximal projection in this proposal which is contrary to the basic claim of the X-bar theory where adjuncts are kept within the scope of the maximal projection (Chomsky, 1970). Therefore, (8) is further modified into (9):

9. [A [B_E [D] [B]] [C_E [F] [C_{MAX} [H] [C]]]]

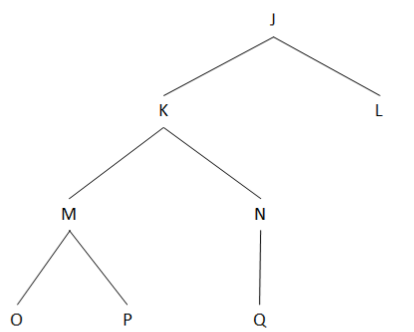
What is of worth mentioning is that the identification of a vertex either as maximal or as elaborator has to do nothing with their relative hierarchies in a tree. Also, it does not mean that the maximal projection will be always embedded within the elaborated projection. Relative saliencies in terms of hierarchical depth or interiority of maximal and elaborated projections may vary in a language like *Bangla* on the basis of how they do appear in the body of a sentence: (9) can simply be rewritten as (10) if F and H changes their respective positions:

10. [A [B_E [D] [B]] [C_{MAX} [H] [C_E [F] [C]]]]

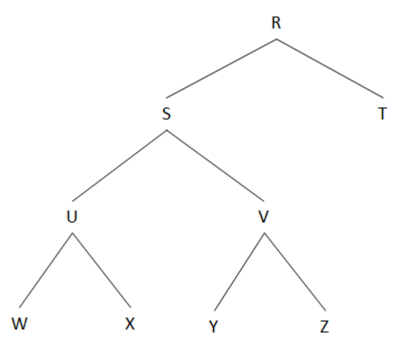
The specification of maximal will only serve the purpose of stating the fact that beyond it appearance of further complement is simply impossible. That means the verb – whose maximal projection it is – is completely saturated once it is maximally projected. This would simply exclude the provision for any intermediating projection to attain the status of maximal in virtue of not getting completely saturated.

Let us consider the case of (6) and (7) now. For the sake of brevity and abstraction, the technical terms of these two trees are replaced with the alphabets in (11) and (12). For the vertices of the tree representing nominal constituent, J-Q alphabets are used. Alphabets from R-Z are allotted to the vertices of the tree representing the verbal constituencies.

11.



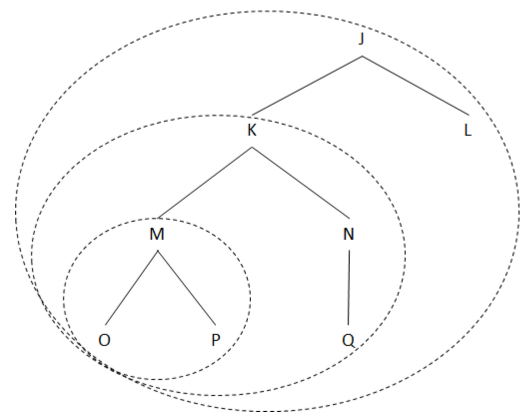
12.



Unlike the structure of the sentence as is represented in (5), the structures of nominal and verbal constituents are well-demarcated in the sense that recurring use of same term is not noticeable. That does not mean that the intermediate projections are not there. Rather, it confirms the intermediating projections by adopting a strict demarcating policy.

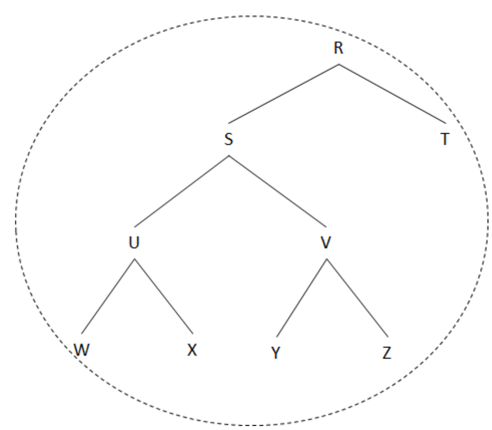
Notion of subtree can be exploited in favor of syntactically and semantically independent expressions that the native speakers of *Bangla* encounter often in their linguistic world. Consider the following example: *bṛddh-ā-ṭi-ke* means 'to the old lady', where *bṛddh-* = O, *-ā-* = P, *-ṭi-* = Q, and *–ke* = L of (11). In case of (11), following expressions will be syntactically and semantically well-formed: (i) *bṛddh-ā* (= M), (ii) *bṛddh-ā-ṭi* (= K), and finally (iii) *bṛddh-ā-ṭi-ke* (= J). This leads to the classification of M, K and J as maximal however with reference to different types of grammatical compulsions: M is maximal because no formative can be added outside its scope; K would not permit the addition of any classifier beyond it; and, no more relational markers are licensed beyond the scope of J. This understanding then in turn justifies why certain trees with the roots M, K, and J are in the subtree relations with (11). Other subtrees of graph (11) other than the ones mentioned above are mathematically possible subtrees but linguistically are not plausible. Grammatically significant trees are represented in (13) with the marking of the concentric circles:

13.



Contrariwise, in case of (12), except the entire tree (which is an improper subtree of itself), no proper (stand-alone) subtree can be identified in spite of the presence of different maximal nodes in different layers of structural hierarchies or interiorities as is evidenced in (14):

14.

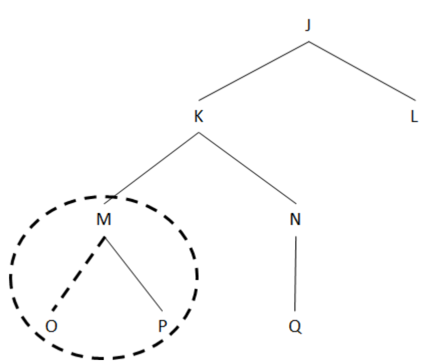


Therefore, in the context of linguistic structures, those subtrees are of immense significance which consist of maximally projected nodes as their roots – with the statutory precaution that all maximally projected nodes are not capable to perform the task of the root of a (sub)tree. It is also worthwhile to note that subtree without a root marked as maximal are optional in nature. Extending this discussion beyond the level of the words will also unveil similar type of findings. For example (9) contains subtrees with the root C_{MAX}, C_E, B_E, and A.

Having discussed this, time now is to look into the issues of the semantic import of the nodes constituting the tree. In other words, the question which will be delved into is 'what constitutes a node?' – As per the principle of compositionality, "the meaning of a complex expression is a function of the meanings of its parts and of the syntactic rules by which they are combined" (Partee, Meulen, and Wall 1990). Under the immediate influence of the compositionality principle then the task will be to explain what governs the combinatorial behavior of neighboring nodes of same hierarchy. Answer to this question must have some provisions for what we will call selectional restrictions. Selectional restriction which is a semantic function associated with each of the nodes remains instrumental in determining 'who can go with whom'.

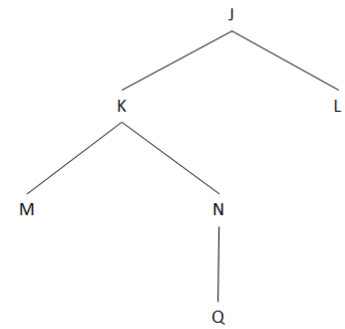
As per the standard practices of lexical semantics like the one proposed by Jackendoff (1990), constitutive nodes should contain information about itself with a special reference to the structure which it is a part of. Since the constituent node has its connotations both at the levels of local and global structures the semantics of a node must contain information about (a) the subtree reflecting its immediate scope and (b) the subtree with a root node marked as maximal where the immediate scope is embedded. As per this proposal then the meaning of an expression, say for example O, will be the following:

15.



A subtree – like the one circled with dashed line in (15) marked with the presence of maximal root – i.e. M, and is in inverse relation with the following tree:

16.



The traditional grammar of *Bangla*, then, can be conceptualized in terms of the following concepts:
i) a set of terminal symbols (= S_T); Here, terminal symbols are the representatives of different word classes;
ii) a set of non-terminal symbols (= S_{NT}); such that, $S_T \cap S_{NT} = \Phi$;
iii) Non-terminal symbols has a set of distinguished members (= D) capable of being marked either with MAX or with E resulting into two distinct partitions, namely D_{MAX} and D_E. Point to be noted $(D_{MAX} \cap D_E = D) \subset S_{NT}$;
iv) a highly distinguished member A which belongs to D;
v) A set of initial trees (T_I) whose interior nodes are labeled by nonterminal symbols. Nonterminal nodes of a particular tree at the bottom can be replaced by the distinguished members of it for further derivation. However, if the bottom consists of terminal symbols, no more replacement as well as further derivation is permitted;
vi) the appearance of lexical element in a tree will indicate the end of a particular path;

The graph theoretic approach to the traditional grammar of *Bangla* will be classified as lexical because each lexical item, in virtue of being the member of word classes belonging to the set of terminal symbols is associated with structural description which is functional is nature. In other words, the lexical constituents of a language must contain the functional descriptions for their respective distribution in a structure. In order to meet this structural goal, then, one needs to identify the information in two broad categories: (a) the information which is locally relevant and constitutes the core of the lexical properties, and (b) the information which is globally relevant and indicates the relation of the lexical core to its global sentential context within which it is grounded. This in one way address the classical problem of symbol grounding which is addressed by Harnard (1990).

As per the theoretical framework described above the structural meaning of classifier (CLS) with respect to (13) will be as follows:

17. CLS
Q
[N[Q]]
[K[M][N]]

The first line of this representation is the grammatical classification of the node Q as classifier. Third line of the representation is a non-maximal subtree which is in inverse relation with the subtree mentioned in the fourth line. Subtree mentioned in the fourth line contains an intervening maximal projection K – which is already discussed in the context of (11) and (13), previously. Following the similar graph theoretic approach, a lexical grammar of simple *Bangla* sentence can be written with the help of graph-theoretic assumptions discussed earlier. In fact the graphical representations that we have illustrated in (5), (6), and (7) can produce the sentences which can be interpreted following the techniques of lexical grammar discussed above. One such sentence is represented in the Appendix of this article to show how rich the structural aspect of *Bangla* is. This paper will be concluded with an observation which we are interested of: Since CLS is a word class, the function mentioned in (17) will be finitely ambiguous in virtue of containing expressions like -*ṭi*, -*ṭā*, -*khāni*, -*khānā* etc. In order to exclude this type of problem, further investigation has to be initiated towards the finer grain analysis of different categories.

References

Barbara H. Partee, Alice ter Meulen, and Robert E. Wall. 1990. *Mathematical Methods in Linguistics*. Kluwer Academic Publishers, Dordrecht.

C.L. Lieu. 1985. *Elements of Discrete Mathematics*. Tata MCgraw-Hill Publishing Company Limited, New Belhi.

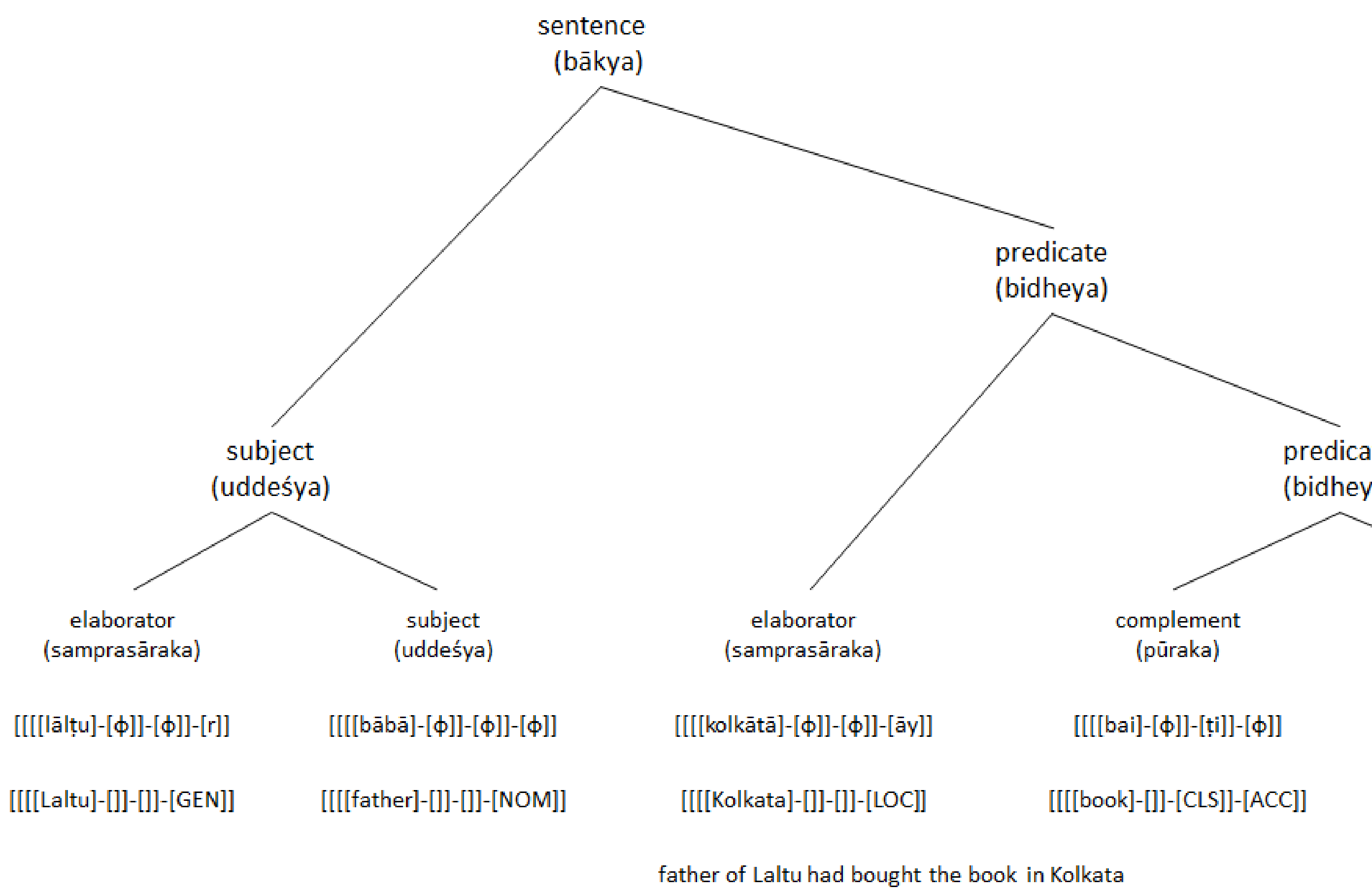
Harnard, S. (1990). The Symbol Grounding Problem, *Physica D* 42, 335-346.

Noam Chomsky. 1970. Remarks on nominalization. In: R. Jacobs and P. Rosenbaum (eds.) *Reading in English Transformational Grammar*, 184-221. Waltham, Ginn.

Ray Jackendoff. 1990. *Semantic Structure*. The MIT Press, Massachussets.

Suniti Kr. Chatterji. 1939. *Bhasha-Prakash Bangla Vyakarana*. Edition 2014. Rupa Publication India Ltd, New Delhi.

Appendix 1


sentence
(bākya)
predicate
(bidheya)
subject
(uddeśya)
predica
(bidhey
elaborator
(samprasāraka)
subject
(uddeśya)
elaborator
(samprasāraka)
complement
(pūraka)
[[[[lālṭu]-[ϕ]]-[ϕ]]-[r]]
[[[[bābā]-[ϕ]]-[ϕ]]-[ϕ]]
[[[[kolkātā]-[ϕ]]-[ϕ]]-[āy]]
[[[[bai]-[ϕ]]-[ṭi]]-[ϕ]]
[[[[Laltu]-[]]-[]]-[GEN]]
[[[[father]-[]]-[]]-[NOM]]
[[[[Kolkata]-[]]-[]]-[LOC]]
[[[[book]-[]]-[CLS]]-[ACC]]
father of Laltu had bought the book in Kolkata