Lang Resources & Evaluation
DOI 10.1007/s10579-014-9266-3

ORIGINAL PAPER

A dependency annotation scheme for Bangla treebank

Sanjay Chatterji · Tanaya Mukherjee Sarkar · Pragati Dhang · Samhita Deb · Sudeshna Sarkar · Jayshree Chakraborty · Anupam Basu

Abstract Dependency grammar is considered appropriate for many Indian languages. In this paper, we present a study of the dependency relations in Bangla language. We have categorized these relations in three different levels, namely intrachunk relations, interchunk relations and interclause relations. Each of these levels is further categorized and an annotation scheme has been developed. Both syntactic and semantic features have been taken into consideration for describing the relations. In our scheme, there are 63 such syntactico–semantic relations. We have verified the scheme by tagging a corpus of 4167 Bangla sentences to create a treebank (KGPBenTreebank).

Keywords Dependency structure · Syntactico–semantic relation · Paninian karak · Modern Bangla grammar and language

1 Introduction

In this paper, we present a dependency annotation scheme for Bangla language. The relations are prepared taking into account modern grammatical and language structures of Bangla mostly studied from Chatterji (2003) and the dependency relations used in other Indian languages like Hindi and Urdu (Sharma et al. 2007; Bhatt et al. 2009). Most of these existing Indian language treebanks follow Paninian grammatical model which is discussed in Bharati et al. (1999).

S. Chatterji (✉) · T. M. Sarkar · P. Dhang · SamhitaDeb · S. Sarkar · A. Basu
Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, India
e-mail: sanjaychatter@gmail.com; sudeshna@cse.iitkgp.ernet.in

J. Chakraborty
Humanities and Social Sciences, Indian Institute of Technology, Kharagpur, India

Published online: 26 March 2014

By observing these grammatical models and dependency relations, we have tried to get a complete list of dependency relations that captures the Bangla language. The relations are described by the syntactic and semantic features occurring between the words.

In Bangla, syntax is not a strong aspect in identifying the relations. This is because the word order in Bangla sentences are relatively less rigid and suffixes and postpositions have different role in different contexts. Therefore, we have considered a balance of syntactic and semantic features to define the Bangla dependency grammar.

Computational processing for Bangla is challenging because of the scarcity of annotated resources. In the absence of treebank of Bangla the work of parsing and some other studies which could have helped in machine translation, question answering, etc. have been hindered.

Different grammars have been advocated for building treebanks in different languages like phrase structure grammar [in Penn Treebank Marcus et al. (1993)] and dependency grammar [in Prague Dependency Treebank Hajič et al. (1996)]. We have used a dependency grammar based scheme to build a treebank of Bangla sentences. The scheme has been evolved by studying carefully the corpus considered for annotation (4167 Bangla sentences) as well as other corpus. The treebank created for these 4167 Bangla sentences is referred to as KGPBenTreebank[1].

The rest of the paper is organized as follows. Some related work has been discussed in Sect. 2. In Sect. 3, we have discussed some assumptions used to define the scheme and in building the KGPBenTreebank. The relation set of the scheme have been categorized and defined in Sects. 4 and 5, respectively. In Sect. 6, we have analyzed the scheme as well as the annotation process. In Sect. 7, we have compared the relations in KGPBenTreebank and Anncorra. Finally, Sect. 8 contains the conclusion. See 'Appendix 1' for the comprehensive dependency relationship set of the scheme.

In the paper, examples are written in "Indian languages TRANSliteration" (*itrans*) Chopde (2000) encoding and the mappings of the *itrans* encoding with glyphs in the Bangla and Hindi scripts are shown in 'Appendix 2'. Each example contains a Bangla sentence followed by the gloss and translation in English. Glosses contain root and features separated by—(dash) symbol. Here, gen (genitive), loc (locative), acc (accusative), pl (plural), and sp (specifier) are the nominal features and past, pre (present), fut (future), prog (progressive), per (perfect), par (participle), inf (infinite), nf (nonfinite) and neg (negative) are the verbal features.

2 Related work

Major treebanks are created on the basis of phrase structure or dependency structure of the language. The phrase structure grammar also known as context free grammar

[1] The dependency grammar for Bangla language and the Bangla treebank is created under the project "The Bangla Treebank". This project is supported by Linguistic Data Consortium for Indian Languages (LDC-IL) built by MHRD, Govt. of India under the aegis of the Central Institute of Indian Languages, Mysore, India. See the link for details. http://www.cel.iitkgp.ernet.in/~oldtools/kgpbentreebank.html.

was used to build some English treebanks. In these treebanks, the intermediate nodes are phrasal nodes and the leaf nodes indicate the words.

One of the earliest such attempt was the ATR/Lancaster Treebank project by Black et al. (1993, 1996) for the American English corpus of 730,000 words. Another contemporary attempt was made in Penn Treebank Marcus et al. (1993) with the same concept and tags to annotate the spoken and written American English corpus of 4.5 million words. The tags used in these two treebanks include 14 phrase structures: Adjective phrase, Adverb phrase, Noun phrase, Prepositional phrase, Simple declarative clause, Clause introduced by subordinating conjunction or 0, Direct question introduced by wh-word or wh-phrase, Declarative sentence with subject-aux inversion, subconstituent of SBARQ excluding wh-word or wh-phrase, Verb phrase, wh-adverb phrase, wh-noun phrase, wh-prepositional phrase, constituent of unknown or uncertain category. The Penn Treebank includes four NULL elements namely *, T, 0(zero) and (rarely) NIL for four empty subject positions. The uses of these NULL elements are described in Santorini and Marcinkiewicz (1991).

The annotation tagset of the above phrase structure treebanks have been widely used by others to annotate other treebanks. For example, a Chinese treebank was created in Xue et al. (2005) based on the Penn Treebank annotation scheme. Three year Wall Street Journal (WSJ) collection of approximately 30 million words was annotated in Charniak et al. (2000) with the same structure as Penn Treebank, except for some additional co-reference marking.

Besides phrase structure treebanks, attempts have been made to build treebanks by collecting the framesets for each lexeme of the sentence. The English Proposition Treebank (Propbank) (Palmer et al. 2005) used the semantic roles of the verbs and analyzed the frequency of syntactic or semantic alternations in the annotation of the Penn Treebank comprehensive corpus.

However, phrase structure grammar is not so appropriate for annotating the treebanks for all languages. Certain Indian languages, some Roman languages (Italian and Spanish) etc., where word structure is not so rigid are some examples of them. For these languages, dependency grammar has been considered as an alternative to phrase structure grammar. Nevertheless, dependency grammar has also been successfully used in English treebanks by Karlsson et al. (1995) and McCord (1990).

The Prague Dependency Treebank Hajič et al. (1996, 2000, 2001) is one of the pioneering work in this direction. In this treebank, 40 syntactic dependency functions has been defined between governor and its dependent nodes, such as: actor/bearer, addressee, patient, origin, effect, cause, regard, concession, aim, manner, extent, substitution, accompaniment, locative, means, temporal, attitude, cause, regard, directional, benefactive, comparison; there are also specific functions for dependents on nouns, for example material, appurtenance, restrictive and descriptive adjunct, the relation of identity, etc. Here, each word and each punctuation mark has been considered as one node. No extra node has been inserted in the tree except root node.

A Quranic Arabic corpus was annotated (Karlsson et al. 1995) with the help of dependency relations of traditional historical Arabic grammar known as iráb. Here,

45 dependency relations[2] are categorized into 5 top categories of dependencies: nominal, verbal, phrases/clauses, adverbial and particle.

Bhatt et al. (2009) have annotated a multilayered treebank for Hindi and Urdu based on the dependency relations of Paninian grammatical model Bharati et al. (1999). In the model, the dependency relations in the sentence were defined between the modifier and the modified words. In the treebank named, Bharati et al. (2002) and Sharma et al. (2007) have considered the chunk as the basic unit of a dependency structure. They have also annotated the corpus using phrase structure grammar with a limited set of category labels: NP, AP, AdvP, VP and CP.

The treebanks are created both manually and semi-automatically. For example, to build the Penn Treebank, the POS tagged sentences were automatically parsed to yield a skeleton syntactic structure and then corrected by human annotators. The aim was to develop a large sized annotated corpus with minimum human effort. The Prague Dependency Treebank was created using 3 level annotation process: morphological annotation, syntactic annotation and linguistic meaning annotation. Here, the trees contain three parts of each node: original word form, morphological information and syntactic tag. The online annotation of Quranic Arabic corpus was done in a multistage approach: automatic rule-based tagging, initial manual verification, and online supervised collaborative proofreading. There were approximately 100 unpaid volunteer annotators and a small number of expert annotators (or supervisors or reviewers) annotating through a popular attracting website.

3 Categorization of the relation set

A clause is a group of words with a subject and a predicate. A clause giving a complete proposition is an independent clause, whereas a dependent clause depends on another clause for making the proposition complete.

A sentence can be either a single independent clause or it may contain, along with an independent clause, one or more dependent clauses. We have considered both the shallow level intrachunk relations and deep interchunk and interclause relations to give a complete analysis of the Bangla sentence.

A word-chunk (chunk) in a sentence is a syntactically correlated non-overlapping group of words. In Bangla sentences two chunks can interchange their places by keeping the meaning of the sentence unaltered. Within a chunk some words are content words while some are function words. A content word is a word which carries meaning independently. One of the content words of a chunk is the head of the chunk.

We have categorized the relation set into three levels of relations. The relations inside a chunk are tagged as intrachunk relations at Level 1. At Level 2, we have tagged the relations between the chunks within a clause. There are three such relation types: Case/Karak relations and Modifier relations and a few other interchunk relations like conjunct, particle, symbol, etc. Finally, the interclause relations are tagged at Level 3.

[2] The list with detailed description of the dependency relations can be seen at http://corpus.quran.com/documentation/syntaxrelation.jsp

Each level of relation is further classified based on syntactico-semantic features. We have used syntactic features like suffix, postposition, and morphological features and semantic features like the type of action, obligation in doing action, and animacy. We use syntactic features till syntax is able to analyze. In case of ambiguity, it goes for semantic features.

The categorization of relations are mentioned below along with the relation tags of the classifications. Each class of relation is defined in Sect. 5 starting from Level 1 relations to Level 3 relations.

Level 1: Intrachunk Relations: Intrachunk relations are ppl, stc, vx, pof, redup, and frag.

Level 2: Interchunk Relations: karta (k1d, k1e, k1p, k1s, and k1g), karma (k2t, k2m, k2g, k2u, and k2s), karan (k3), apadan (k5p, k5s, k5t, and k5d), adhikaran (k7p, k7t, k7d, and k7s), and Other case-alike relations (rh, ru, des, r6v, compr, and sym) are categorized as Case r6, ras, rasneg, nnmod, jnmod, dnmod, pronmod, pnmod, anmod, adv, vmod, neg, and acomp are categorized as Modifier. ccof, pcc, rad, par, qs, end, and sym are categorized as Other interchunk relations.

Level 3: Interclause Relations: Interclause relations are ref, clausal*, clausalcomp, and comp.

A dependent clause in a sentence is either a complement of the main clause or may modify the main clause. Interclause relations indicate the role of dependent clauses. Independent clauses are joined by a conjunct word and no interclause relation is used in such cases.

4 Some salient features of the relationset and the KGPBenTreebank

In the process of building the KGPBenTreebank, we have defined the relations of the scheme and used them to annotate Bangla sentences. The annotated sentences are represented in two ways. In the textual representation, the dependency relation between the pair of words has been shown in the format DEP(CHILD, PARENT) meaning that there is a dependency relation DEP, where the dependent contains the word CHILD and the head contains the word PARENT. In the graphical representation, each sentence is represented in a rooted directed tree structure where each edge and node are labelled. In the structure, head words are used as label of the parent nodes, dependents are used as label of the child nodes and the dependency relations are used as label of the edges. The directions are from parent to child nodes.

During this exercise, we have considered certain features to simplify the task. These are discussed in the following subsections.

4.1 Syntactico-semantic relation set

In the scheme, the head word can have many dependents, but a dependent is related only to one head. A word which is not dependent to any head word is called the root

of the sentence. Dependency relations between the head word and its dependents can be explained syntactically as well as semantically. We have considered the defined syntactico-semantic relation set [incorporating thematic and functional semantic relations occurring at the sentence level] with the major inclinations towards logical relations to capture the dependency relations.

4.2 NULL node insertion

Bangla, like most of the Indian languages, has in general a subject-object-verb sentence structure with verb being the root of the sentence. However, a word or a symbol in a Bangla sentence may be dropped at the surface level. Some of such cases are the copula verb in present tense when it takes attributes, the verbs that can be recovered by the presence of another verb, and the conjunts connecting two words, phrases or clauses. For these cases it is necessary to define a placeholder so that the relations can be shown. We put a <NULL> as a placeholder of the dropped words as discussed below.

a) Bangla copula verbs are dropped in the present tense for the positive polarity instances. In interchunk relations, many of the relations are verb centric. A <NULL> verb is inserted in place of both attributive and existential copula when it is dropped in the surface.
   (1) rAma bhAla chhele <NULL>.
   [Ram good boy]
   Ram is a good boy.
b) A <NULL> verb is inserted on recoverability condition. Here, recoverability is possible from the occurrence of another verb in the sentence.
   (2) Ami Aja dillIte <NULL> Ara kAla kolakAtAYa yAba.
   [I today Delhi-loc and tomorrow Kolkata-loc go-fut]
   Today I will go to Delhi and tomorrow to Kolkata.
c) A <NULL> conjunction is inserted to connect two words, phrases or clauses.
   (3) pena <NULL> penasila <NULL> khAtA eba.n ba;i.
   [pen pencil notebook and book]
   Pen pencil notebook and book.

4.3 Same tree for different sentences

As Bangla is a relatively free word order language, the words of the sentence can change their places. Some of the representations are unmarked (most fluent) and some are marked (rare). An unmarked simple Bangla sentence is shown in Example (4).

(4) mohana Ama khAchchhe.
[Mohan mango eat-pre,prog]
Mohan is eating mango.

This Bangla sentence is a simple sentence containing two nouns and a verb. “mohana” is the agent of this sentence, “Ama” is the object and both are dependent

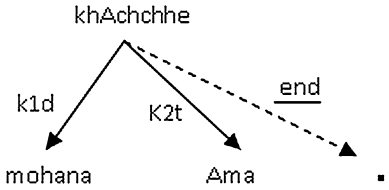

Fig. 1 A dependency tree

of the verb "khAchchhe". However, according to the word being stressed and the focus of discourse level semantics, the words of this sentence can be placed in different orders. The trees of all sentences containing these three words will remain the same as Fig. 1 (*k*1*d*, *k*2*t* and *end* labels are defined in Sect. 5).

4.4 Ambiguity in the sentence

An ambiguous sentence is one having multiple syntactic structures. Like in many other languages Bangla also include ambiguous sentences. For computational simplicity, for each ambiguous sentence we have chosen one interpretation which seems to be the most appropriate according to the context. We have selected the most likely interpretation of ambiguous sentence in the current context.

5 Definition of Bangla dependency relations

In this section, we have defined the dependency relations starting from Level 1 to Level 3. Intrachunk relations (Level 1) of the bottom level are defined in Sect. 5.1. Among interchunk relations of Level 2, Case, Modifier and few other interchunk relations are defined in Sects. 5.2, 5.3 and 5.4, respectively. Finally, interclause relations of the top level are defined in Sect. 5.5.

Each definition of dependency relation contains English and Bangla name of the relation separated by forward slash (/) and an acronym of the relation used for tagging KGPBenTreebank. Each definition is explained with an example Bangla sentence where a pair of words has the corresponding dependency relation. The corresponding dependency relations are shown in textual representation.

5.1 Intrachunk relations

In a chunk, the meaning of a function word is expressed through its relation with the content word. Prepositions and postpositions in the noun chunk and the auxiliary verbs in the verb chunk are examples of function words. Some Bangla postpositions are generated from some verb roots which are otherwise used as main verbs. Some main verbs are also used as auxiliary verbs. The relation between two words within a chunk is identified as intrachunk relation. The modifier intrachunk relations are

not referred here and will be discussed later (Sect. 5.3). Some of the intrachunk relations are discussed below.

Postposition/Anusarga (ppl): Postpositions like 'theke' (from), 'diYe' (by), etc. and prepositions like 'binA' (except) in the noun chunks (NP) are related to head words by Postposition/Anusarga (ppl) dependency relation.

(5) se skula theke phirachhe.
[he school from return-pre,prog]
He is returning from school.
ppl(theke, skula)

Spatio-temporal connection/Sthan-samaygata samparka (stc): Bangla nouns indicating space or time (space-time nouns) like 'bhitara' (inside), 'bAire' (outside), 'upara' (above), 'nicha' (below), 'Age' (before), 'pare' (after), etc. may follow a noun or a pronoun with genitive marker. These space-time nouns may be followed by postposition. We relate the preceding noun or pronoun with space-time noun by Spatio-temporal connection/Sthan-samaygata samparka (stc) dependency relation.

(6) se bA.Dira bhitara theke DAkachhe.
[he home-gen inside from call-pre,prog]
He is calling from the inside of the house.
stc(bA.Dira, bhitara)

Auxiliary verb/Sahayak kriya (vx): Auxiliary verbs in the verb chunks (VP) are related to main verbs by Auxiliary verb/Sahayak kriya (vx) dependency relation.

(7) bAchchArA hese uThala.
[child-pl laugh-nf rise-past]
The children began to laugh.
vx(uThala, hese)

Part of/Kriya antargata bisheshya (pof): In conjunct verb chunks, the main verbs contain a nominal or adjectival part followed by a verb part. The first part is related to the second part by Part of/Kriya antargata bisheshya (pof) dependency relation.

(8) ba;i melA bhAlabhAbe anuShThita haYechhe.
[book fair well happen-par have-pre,per]
The book fair has been executed well.
pof(anuShThita, haYechhe)

Reduplication/Shabda dbaita (redup): Reduplication/Shabda dbaita (redup) dependency relations include nominal, pronominal and adjectival reduplications indicating plurality; verbal, adverbial, postpositional reduplications indicating continuity; reduplicated words indicating accuracy; onomatopoetic words, hedged expressions and also echo words. We consider the first occurrence of such pair of words as the head word and the second word as the dependent. In the following example, we have assigned two serial numbers (1 and 2) to disambiguate the head and the dependent.

(9) bAchchArA galpa karate(1) karate(2) skula theke phirachhe.
[child-pl story do-inf do-inf school from return-pre,prog]

The children are returning from school chatting.
redup(karate(2), karate(1))

Fragment/Bhagnamsha (frag): Suffixes can be written either attached to the word or independently. The suffix occurring independently is related to the word by Fragment/Bhagnamsha (frag) dependency relation.

(10) manamohana si.nha (bhAratera pradhAnamantrI) ra bidesha saphara Achhe.
[Manmohan Singh (India-gen Prime–Minister) -gen foreign trip be-pre]
Manmohan Singh (the Prime minister of India) has a foreign trip.
frag(si.nha, ra)

5.2 Case/Karak

The basic karak relations and their categories, as discussed in the Paninian framework (Bharati et al. 1999), from the perspective of natural language processing (NLP) and also as they are defined in Bangla traditional grammar are of 6 types: Karta, Karma, Karan, Sampradan, Apadan and Adhikaran. In the scheme used by us, the Paninian karak relations, in general, are accepted with some changes as suggested in modern Bangla grammars Chatterji (2003), Chakravarty (2010). Some Bangla noun verb relations have been discussed in Chatterji et al. (2009).

It is difficult to consider sampradan karak relation in Bangla from syntax except in the instances where the words 'dAna' (donate) and 'sampradAna' (donate) are explicitly mentioned. This is because in Bangla sampradan are considered when the nature of the verb indicated that something is given selflessly.

Some Bangla grammarians have advocated that sampradan karak may not be considered separately in Bangla. In this context we quote Suniti Kumar Chatterji from Chatterji (2003). He says that while Sanskrit has special bibhakti for sampradan; Bangla does not have this. Some people use sampradan in Bangla for compatibility with Sanskrit; while others merge sampradan with karma. He considers the second approach reasonable (Page 247) and merged the sampradan with gauna karma (Page 241 and 299).

5.2.1 Subject/Karta

Subject/Karta is the one who does, experiences or exists. It can also be or become something. However, it is referred to as Passive Karta when it acts as the doer of an action in a passive sentence. In Bangla, the karta may take a wide range of suffixes, like genitive, nominative, accusative, locative, etc. The verbs act in different ways and accordingly their subjects are defined. Because of the subcategorization of the subject, instead of including subject in the relationset we have included its subcategories.

Doer subject/Kriya sampadak karta (k1d): When the verb indicates some mental or physical exercise by an animate karta then its subject is defined as Doer subject/Kriya sampadak karta (k1d).

(11) bulabulite dhAna kheYechhe.
[Indian–nightingale-loc paddy eat-pre,per]
Indian nightingale has eaten paddy.
k1d(bulabulite, kheYechhe)

Experiencer subject/Anubhab karta (k1e): When the verb expresses mental state, emotion or event without the subject's conscious effort then its subject is defined as Experiencer subject/Anubhab karta (k1e).

(12) AmAra shIta karachhe.
[I-gen cold do-pre,prog]
I am feeling cold.
k1e(AmAra, karachhe)

Passive subject/Paroksha karta (k1p): When the verb indicates an action in passive construction then its subject is defined as Passive subject/Paroksha karta (k1p). This subject acts as logical subject of the sentence or the verb. It is followed by postpositions 'dbArA' (by), 'diYe' (by) and 'karttRRika' (by) or may take the suffix 'ra'.

(13) AnandamaTha ba ~ Nkimachandra kartRRika rachita <NULL>.
[Anandamath Bankimchandra by write-par]
Anandamath is written by Bankimchandra.
k1p(ba ~ Nkimachandra, <NULL>)

Even though this is the usual Bangla device for passivization, there may be confusion in certain cases. For example, 'ra' suffix is also attached with the karta of active voice verbs (See example 12) and 'dbArA', 'diYe' and 'karttRRika' postpositions are also attached with karan. Again same word forms of some verbs are used both in active and passive voice. For example, 'haYechhe' (has been done) in the following first example is in passive voice and in the second one 'haYechhe' (has given birth) is in active voice. Therefore, we assigned separate tag for the karta of passive voice verbs though it conveys redundant information in many cases.

- AmAra dbArA ei kAja **haYechhe**.
  [I-gen by this work have-pre,per]
  This work is done by me.
- tAra jbara **haYechhe**.
  [he-gen fever have-pre,per]
  He has caught fever.

Noun of proposition/Bidheya karta or Samanadhikaran (k1s): Noun of proposition/ Bidheya karta or Samanadhikaran (k1s) is the complement of the karta in the sentence.

(14) tini bhAla shikShaka chhilena.
[he good teacher be-past]
He was a good teacher.
k1g(tini, chhilena) k1s(shikShaka, chhilena)

General subject/Sadharan karta (k1g): The subject is defined as General subject/ Sadharan karta (k1g) in the following contexts. The subjects which do not belong to any of the subcategories mentioned above (i.e., k1d, k1e, k1p, and k1s) are also tagged as k1g.

1. Subject of copula or be verb.
2. Subject of a verb in which the agent of the action is not specified, though it may be implied.
3. The subject which is in r6 relation with another noun or pronoun.

(15) AmAdera siTi kaleja khuba bhAla <NULL>.
[I-pl,gen city college very good]
Our city college is very good.
k1g (kaleja, <NULL>)

5.2.2 Object/Karma

Object/Karma refers to an object undergoing the action or a person or an object being affected by the action. It also includes some things or positions being achieved or attained through the action. A Bangla sentence with ditransitive verb or causative verb may have two karma. Similarly, transitive verbs and non-transitive verbs may have one and zero karma, respectively. In both active and passive constructions, the karma is tagged in the same way. Because of the subcategorization of the object, instead of including object in the relationset we have included its subcategories.

Transitive object/Sakarmak karma (k2t): Transitive object/Sakarmak karma (k2t) is the karma of a transitive verb.

(16) bhUmikampa sAjAno gochhAno shaharaTA dhba.nsa karala.
[earthquake decorating sorting city-sp destroy do-past]
Earthquake destroyed the beautiful city.
k2t(shaharaTA, karala)

Direct object/Mukhya karma (k2m) & Indirect object/Gauna karma (k2g): The two objects of a ditransitive verb of both passive and active sentences may be tagged as follows. Direct object/Mukhya karma (k2m) is the karma which undergoes the action. Indirect object/Gauna karma (k2g) is the one which is affected by the action, or a recipient or a beneficiary of the action. These two karma generally co-occur.

(17) Ami mAke chiThi likhachhi.
[I mother-acc letter write-pre,prog]
I am writing a letter to my mother.
k2g(mAke, likhachhi) k2m(chiThi, likhachhi)

Purposive object/Uddyeshya karma (k2u) & Predicative object/Bidheya karma (k2s): In the case of active as well as passive sentences with ditransitive verb, if the two objects are in complementary relation, the one which takes the complement is

defined as Purposive object/Uddyeshya karma (k2u). This karma is attached with the suffix 'ke'. Another one which stands as a complement is defined as Predicative object/Bidheya karma (k2s).

(18) tini buddhadebake parameshbarera abatAra balena.
[he Buddhadeb-acc parameswar-gen incarnation tell-pre]
He regards Buddhadeb as the incarnation of God.
k2u(buddhadebake, balena) k2s(abatAra, balena)

A ditransitive verb may take a pair of arguments which refer to the same thing or person. We tag them as k2u and k2s. A ditransitive verb may also take a pair of arguments which refer to two different things or persons. We tag them as k2m and k2g. The nature of such pairs of arguments are different. For example, in the following sentence (i), 'gAndhike' (Gandhi-to) and 'bApu' (Bapu) are two karma of 'balA haYa' (called). These two karma refer to same person. Therefore, they are tagged as k2u and k2s, respectively. In the following sentence (ii), 'bApu' (Bapu) is the agent which is tagged as k1d and 'gAndhike' (Gandhi-to) and 'kathATA' (the words) are two karma of 'balala' (told) and they are tagged as k2g and k2m, respectively.

(i.) **gAndhike bApu** balA haYa.
[Gandhi-acc Bapu say be-pre]
Gandhi is called Bapu.
(ii.) **bApu gAndhike kathATA** balala.
[Bapu Gandhi-acc word-sp say-past]
Bapu told the words to Gandhi.

5.2.3 Instrumental/Karan

*Instrumental/Karan(k3):*Instrumental/Karan (k3) refers to a thing or object which acts as an instrument or means for performing an action or the occurrence of an action.

(19) gAdhAke chAbuka mAro.
[donkey-acc whip beat-pre]
Whip the donkey.
k3(chAbuka, mAro)

5.2.4 Ablative/Apadan

Ablative/Apadan is the source or origin of an action or the point of time which indicates the source of an action or the distance between two places. Because of the subcategorization of the ablative, instead of including ablative in the relationset we have included its subcategories.

Place related ablative/Sthanbachak apadan (k5p): Place related Ablative/ Sthanbachak apadan (k5p) refers to the source or origin of an act.

(20) nadIra ghATa theke ghaTa bhese ela.
[river-gen bank from pot float-nf come-past]
The pot came floating from the river bank.
k5p(ghATa, bhese)

State related ablative/Abasthabachak apadan (k5s): State related ablative/ Abasthabachak apadan (k5s) is the state from where the action takes place. Here karta is not displaced, only object is displaced away from the state of karta.

(21) AmAra ghara theke mandirera chU.DA dekhA yAYa.
[I-gen house from temple-gen pinnacle see go-pre]
The pinnacle of the temple can be seen from my house.
k5s(ghara, dekhA)

Time related ablative/Kalbachak apadan (k5t): Time related ablative/Kalbachak apadan (k5t) refers to the point of time which indicates the source of an action.

(22) sakAla theke bRRiShTi nemechhe.
[morning from rain get–down-pre,per]
It is raining since morning.
k5t(sakAla, nemechhe)

Distance related ablative/Duratbabachak apadan (k5d): Distance related ablative/ Duratbabachak apadan (k5d) indicates the distance between two places one of which is the starting point and the other one is the ending point. The starting point place is tagged as k5d. The corresponding sentence contains two place names, a distance measure (either real or abstract) and a copula.

(23) dilli theke kolakAtA bahu dUre <NULL>.
[Delhi from Kolkata too far-loc]
Delhi is too far from Kolkata.
k5d(dilli, <NULL>)

5.2.5 Locative/Adhikaran

Locative/Adhikaran may be either the place where the action takes place, or the time when it takes place, or the domain about which it takes place or the state in which it takes place. Because of the subcategorization of the locative, instead of including locative in the relationset we have included its subcategories.

Place related locative/Deshadhikaran (k7p): Place related locative/Deshadhikaran (k7p) refers to the place that indicates the occurrence of the act.

(24) bA.Dite phulera gAchha Achhe.
[house-loc flower-gen tree be-pre]
There are flowering trees in the house.
k7p(bA.Dite, Achhe)

Time related locative/Kaladhikaran (k7t): Time related locative/Kaladhikaran (k7t) is that point or duration of time which indicates the occurrence of the act.

(25) bhore sUrya oThe.
[dawn-loc sun rise-pre]
The Sun rises in the dawn.
k7t(bhore, oThe)

Domain related locative/Bishayadhikaran (k7d): Domain related locative/ Bishayadhikaran (k7d) is the thing or area, or things constituting a domain which can be pursued for study, or can be pursued as a profession.

(26) tArA sAhitye paNDita <NULL>.
[he-pl literature-loc expert]
They are expert in literature.
k7d(sAhitye, <NULL>)

State related locative/Bhabadhikaran (k7s): State related locative/Bhabadhikaran (k7s) refers to the state of being for something or someone at a particular time.

(27) tArA khuba sukhe Achhe.
[he-pl very happy-loc be-pre]
They are living with happiness.
k7s(sukhe, Achhe)

5.2.6 Other case-alike/Anyanya karak-sama

Other than the five karak relations mentioned above, there are some more relations between a noun and a verb and between two nouns which are often confused with the karak relations.

Reason/Hetu (rh): Reason/Hetu (rh) represents the reason of the action.

(28) bhaYe bhule yAYa debatAra nAma.
[fear-loc forget-nf go-pre God-gen name]
The name of God is forgotten out of fear.
rh(bhaYe, bhule)

*Purpose/Uddeshya (ru):*Purpose/Uddeshya (ru) represents the purpose of the action.

(29) ratana unnatira janya kaThora parishrama kare.
[Ratan promotion-gen for hard work do-pre]
Ratan works hard for his promotion.
ru(unnatira, kare)

Destination/Gantabyasthal (des): Destination/Gantabyasthal (des) is the place related argument of the nontransitive moving verbs 'yAoYA' (go), 'bhramana karA' (travel), etc. Instead of locative ('e', 'te', etc. bibhakti) markers they may take nominative marker (0 bibhakti). It is a karma but it behaves like an adhikaran.

(30) rAma bA.Di giYechhila.
[Ram home go-past,per]

Ram went home.
des(bA.Di, giYechhila)

Possession/Dakhal (r6v): Possession/Dakhal (r6v) is the relation between a noun and a verb where the noun acts as an owner. The owned part is considered as karma. This owner noun takes genitive marker and the corresponding verb is a 'be' verb (existential).

(31) rAmera ekaTA meYe Achhe.
[Ram-gen one-sp daughter be-pre]
Ram has a daughter.
r6v(rAmera, Achhe)

Comparison/Taratamya (Compr): Comparison/Taratamya (Compr) indicates a comparison between two noun phrases or between their attributes. For example, the phrase, "mitAra theke sundarI meYe" [The girl more beautiful than Mita] indicates the comparison between the 'sundarI' attribute of 'mitA' and 'meYe'.

(32) ekhAne aneke phuTabalera theke krikeTa pachhanda karena.
[here-loc many-loc football-gen from cricket like do-pre]
Here many people like cricket more than football.
compr(phuTabalera, krikeTa)

Similarity/Sadrishya (sim): Similarity/Sadrishya (sim) describes the similarity between two noun phrases or between the attributes of two noun phrases.

(33) se AmAra bonera mata <NULL>.
[he I-gen sister-gen like]
She is like my sister.
sim(bonera, se)

5.3 Modifier relations

Most of the modifier relations can also be considered as intrachunk relations. However, sometimes, it is observed that the modifier relations exist between two different chunks.

Genitive/Sambandha (r6): Genitive/Sambandha (r6) refers to the relation between a noun or a pronoun with genitive marker and another noun. The r6 dependency relation relates certain pairs of nouns. Some possible relations exists between such noun pairs are shown in Table 1 with an example of each.

For differentiating the r6 relation from other relations of the noun with genitive marker in a sentence, the following points may be considered.

1. A noun or pronoun with genitive marker, preceding a postposition is not a candidate of r6 relation.
2. A noun or pronoun with genitive marker that precedes a noun of a complex predicate is not a candidate of r6 relation.

Table 1 Relations between the noun pairs which are included in r6

Relation	Bangla Example	English Translation
Possession	rAjAra rAjya	king's kingdom
Part	shishura mukha	face of a baby
Location	jalera mAchha	fish in water
Function	khAoYAra thAlA	plate for eating
Source	sApera bhaYa	fear from snake
Material	sonAra gaYanA	ornaments made of gold
Measurement	du ghanTAra patha	two hour's journey
Cause effect	sUryera Alo	sunlight
Attribute	premera galpa	love story
Sequence	pA.Nchera pRRiShThA	page number five
Simile	j∼nAnera Alo	light of wisdom
Object	Ishbarera sAdhanA	worship of God
Progeny	gAchhera phala	fruits of tree
Adjectival modifier	guNera chhele	boy in good quality

3. A noun or pronoun with genitive marker, which are related to a mental verb group is not a candidate of r6 relation.
4. A noun or pronoun with genitive marker, which is a karta of a passive verb is not a candidate of r6 relation.
5. A noun or pronoun with genitive marker, which is related with a verb by *r6v* is not a candidate of r6 relation.

Associative relation/Saharthak sambandha (ras) & Non-associative relation/ Namarthak sambandha (rasneg): Associative relation/Saharthak sambandha (ras) occurs with a noun or pronoun which accompanies another noun or pronoun in karta or karma position and it is followed by the postpositions 'sa∼Nge' (with), 'sAthe' (with), or 'diYe' (by). The role of saharthak noun is same as the role of the noun with which it is attached. If it is attached with a karta, its role is a karta and if it is attached with a karma, its role is a karma. When the occurrence of such sambandha is negated, it is defined as Non-associative relation/Namarthak sambandha (rasneg). Namarthak sambandha takes postpositions 'chhA.DA', or 'binA'.

(34) simA nIrAra sAthe melAte gela.
[Sima Nira-gen with fair-loc go-past]
Sima went to the fair with Nira.
ras(nIrAra, simA)

(35) se chini chhA.DA chA khete pachhanda kare.
[he sugar without tea eat-inf like do-pre]
He likes to drink tea without sugar.
rasneg(chini, chA)

Noun noun modifier/Sanyogmulak bisheshya (nnmod): Noun noun modifier/ Sanyogmulak bisheshya (nnmod) refers to the relation between two nouns. The two nouns, however, should not be in r6 relation.

(36) rAjaputa jAti khuba yoddhA jAti <NULL>.
[Rajput cast very warrior cast]
The Rajputs are great warriors.
nnmod(rAjaputa, jAti) nnmod(yoddhA, jAti)

Adjective noun modifier/Bisheshyer bisheshan (jnmod): Adjective noun modifier/ Bisheshyer bisheshan (jnmod) is used to relate an adjective which modifies the meaning of a noun or an adjective. For the purpose of simplification, quantifier, numerical modifier etc. have been included into this relation.

(37) se uchcha phalanashIla bIja diYe chASha kare .
[he high productive seed by cultivation do-pre]
He cultivates using high productive seeds.
jnmod(uchcha, phalanashIla) jnmod(phalanashIla, bIja)

Demonstrative noun modifier/Nirnay suchak sarbanam (dnmod): Demonstrative noun modifier/Nirnay suchak sarbanam (dnmod) is the relation between the head of a noun phrase and its demonstrative.

(38) sei meYeTA nAcha karate pAre nA.
[that girl dance do-inf can-pre no]
That girl can not dance.
dnmod(sei, meYeTA)

Pronominal noun modifier/Sarbanamjata bisheshan (pronmod): Pronominal noun modifier/Sarbanamjata bisheshan (pronmod) accounts for the relation between a noun or a personal pronoun and a reflexive pronoun which functions as an emphatic pronoun.

(39) sbaYa.n sbAmIji ei kathA bishbAsa karena.
[himself Swamiji this word belief do-pre]
Swamiji himself believes this fact.
pronmod(sbaYa.n, sbAmIji)

Participial noun modifier/Kridanta bisheshan (pnmod): Participial noun modifier/ Kridanta bisheshan (pnmod) is the relation between a participial verb and a noun. The participial verb, here, modifies the noun.

(40) bAire rAkhA kApa.Dagulo bRRiShTite bhije gela.
[outside-loc keep-par cloth-pl rainloc wet-nf go-past]
The clothes kept outside became wet in rain.
pnmod(rAkhA, kApa.Dagulo)

Appositional noun modifier/Tulyarupe sthapita bisheshan (anmod): Appositional noun modifier/Tulyarupe sthapita bisheshan (anmod) relates a noun phrase with another immediately following noun phrase, both indicating the same person or thing.

(41) manamohana si.nha, bhAratera pradhAnamantrI, bidesha yAchchhena.
[Manmohan Singh , India-gen prime–minister abroad go-pre,prog]
Manmohan Singh, the Prime minister of India, is going to abroad.
anmod(pradhAnamantrI, manamohana)

Adverbial modifier/Kriya bisheshan jatiya bisheshan (adv): Adverbial modifier/ Kriya bisheshan jatiya bisheshan (adv) relates an adverb with a verb.

(42) rimi shAntabhAbe galpa balala.
[Rimi quietly story tell-past]
Rimi told the story quietly.
adv(shAntabhAbe, balala)

Verb-verb modifier/Kriya jatiya bisheshan (vmod): Verb-verb modifier/Kriya jatiya bisheshan (vmod) is the relation between two verbs indicating two sequential or parallel actions.

(43) bRRiShTite pukurera mAchha nadIte giYe pa.Dala.
[rain-loc pond-gen fish river-loc go-nf fall-past]
Due to rain, the pond fishes got into the river.
vmod(giYe, pa.Dala)

Negation modifier/Namarthak abyay (neg): Negation modifier/Namarthak abyay (neg) is the relation between a negation word 'nA' and 'ni' and the verb it modifies.

(44) TinA bA.Di yAbe nA.
[Tina home go-fut no]
Tina will not go home.
neg(nA, yAbe)

Adjectival complement/Bidheya bisheshan (acomp): Adjectival complement/ Bidheya bisheshan (acomp) is the relation between an adjective and a verb where the adjective is a complement to the verb.

(45) ei mandiraTi khuba prAchIna <NULL>.
[this temple-sp very old]
This temple is very ancient.
acomp(prAchIna, <NULL>)

Question words are tagged according to their usage in an interrogative sentence. How question words act in sentences are shown below.

- 'ke' (who-singular), 'ki' (what), 'kArA' (who-plural), 'konaTA' (which) etc. act as karta.
- 'ki' (what), 'kAke' (whom), 'kAdera' (whose-plural), 'konaTA' (which) etc. as karma.
- 'kena' (why) is used as reason (rh) or purpose (ru).
- 'kakhana' (when), 'kothAYa' (where) etc. act as adhikaran.
- 'kata' (how much) is tagged as acomp or jnmod.
- 'ki' (what) is sometimes added with yes/no type question and is tagged as particle (par).

- When the question words 'ki' (what) and 'kata' (how much) are related with the noun then the dnmod and jnmod tags are used, respectively.

5.4 Few other interchunk relations

Some other interchunk relations are defined below.

Conjunct/Samyojak abyay (ccof): Conjunct/Samyojak abyay (ccof) conjoins between two independent words, phrases or clauses using the conjunct words 'eba.n' (and), 'kintu' (but), 'o' (and), etc. or using the conjunct symbols ',', '-' etc.

(46) darshaka upache pa.Dla nA kintu udaghATita hala eka natuna dika.
[audience overflow-nf fall-past no but reveal-par be-past one new direction]
The audience was not overflowing, but a new direction was revealed.
ccof(upache, kintu) ccof(hala, kintu)

Preconjunct/Abasthatmak abyay (pcc): Preconjunct/Abasthatmak abyay (pcc) relation is tagged between a subordinating conjunction and a verb. The subordinating conjunction may be in the preconjunct or in the postconjunct construction.

(47) yadio Ami bAraNa karalAma tabuo se gela.
[though I forbid do-past still he go-past]
Though I forbade him, still he went.
pcc(gela, tabuo) pcc(karalAma, yadio)

Address word/Sambodhan sabda (rad): Address word/Sambodhan sabda (rad) is the relation between the address word and the verb. Address word is used when a person is addressed to.

(48) sImA, tomAra yAoYA chalabe nA.
[Sima, you-gen go move-fut no]
Sima, you ought not to go.
rad(sImA, chalabe)

Particle/Bakyalankar abyay (par): Particle/Bakyalankar abyay (par) is the relation between the particle and the verb.

(49) nyAyya kAraNa kintu nei.
[logical reason but be-pre,neg]
But there is no logical reason.
par(kintu, nei)

Question mark/Prashnabodhak chihna (qs): Question mark/Prashnabodhak chihna (qs) relation is used between the question mark (?) of the interrogative sentence and the verb.

(50) tomAra kAke beshi pachhanda <NULL>?
[you-gen who-acc more like]

Whom do you like most?
qs(?, <NULL>)

End/Samapti (end): End/Samapti (end) is used to indicate the relation between non-interrogative sentence markers (| and !) and the verb of the main clause.

(51) ete jami kama naShTa haYa.
[it-loc land less spoil be-pre]
For this, misuse of land is minimized.
end(., haYa)

Symbol/Chihna (sym): Symbol/Chihna (sym) is the relation between a symbol except the sentence end markers and those which are used to join two independent clauses and the verb.

(52) rohana, tui chA khete yA.
[Rohan, you tea eat-inf go-pre]
Rohan, go and have tea.
sym(,, yA)

Dependent clauses are attached to the main clause with subordinating conjunctions. When the subordinating conjunction comes before the dependent clause, then it is called preconjunct. When the subordinating conjunction comes after the dependent clause, then it is called postconjunct. If a single clause sentence contains a subordinating conjunction, then it is tagged as particle. It is generally placed at the beginning of the sentence.

5.5 Interclause relations

Interclause relations include relative and complement clause constructions where there is a main clause and at least one dependent clause. In relative clause construction the dependent clause is connected with the main clause by a referent. In some cases a coreferent (in accordance with the referent) may also occur in the initial position of the main clause.

Referent/Nirdesak (ref): Referent/Nirdesak (ref) relation is found between two clauses when a coreferent also surfaces. The role of the dependent clause is the same as the coreferent in the main clause. The ref relation is assigned between the verb of the dependent clause and the coreferent.

(53) yakhana bRRiShTi pa.Dachhila takhana se chhAtA niYe gela.
[when rain fall-past then he umbrella take-nf go-past]
He took an umbrella when it was raining.
ref(pa.Dachhila, takhana)

Clausal Star/Bakyamsha samagra (clausal):* Clausal Star/Bakyamsha samagra (clausal*) relation is found between two clauses when the coreferent does not surface, and the relation is assigned between the verb of the dependent clause and the verb of the main clause. Here, '*' is a variable indicating different dependency relations such as k1 (karta), k2 (karma), k3 (karan), k5 (apadan), rh(hetu), k7p

(deshadhikaran), k7t (kaladhikaran), etc. The '*' indicates the dependency relation between the two clauses. For simplicity, we have not considered subdivisions of karaks (except adhikaran) in this variable position.

(54) yakhana bRRiShTi pa.Dachhila se chhAtA niYe gela.
[when rain fall-past he umbrella take-nf go-past]
He took an umbrella when it was raining.
clausalk7t(pa.Dachhila, niYe)

(55) ye ba;iTA Ami niYechhilAma Ami pherata diYe diYechhi.
[which book-sp I take-past I return give-nf give-pres,per]
I have returned the book which I took.
clausalk2(niYechhilAma, diYe)

Clausal complement/Bakyamsha sampurak (clausalcomp): In a complement clause construction, the dependent clause is connected with the main clause by a complementizer. The dependent clause is considered as clausal complement of the main clause. The verb of the main clause and the verb of the complement clause are connected by Clausal complement/Bakyamsha sampurak (clausalcomp) dependency relation.

(56) se balala ye se kAla kolakAtA yAbe.
[he tell-past that he tomorrow Kolkata go-fut]
He said that he will go to kolkata tomorrow.
clausalcomp(yAbe, balala)

Complementizer/Sampurak (comp): In the complement clause construction, Complementizer/ Sampurak (comp) relation is between the complementizer and the verb of the complement clause. The complementizer which is the introducer of the complement clause acts as a connector between the main clause and the dependent clause. It is attached to the verb of the complement clause. In our analysis, Bangla complementizers are 'ye' (that), 'bale' (that) or a comma. In example (56) 'ye' is related with the verb 'yAbe' by comp dependency relation.

Apart from being a complementizer, 'bale' (due to) can also indicate a reason clause (clausalrh) in multiclause Bangla sentences. It can also be used as a particle, or a verb. Similarly, 'ye' (that or who) can also be a demonstrative pronoun (as shown in example (55)) or a personal pronoun which acts as referent in the subordinate clause. 'ye' can also be used as a particle.

6 Analysis of the scheme and the annotation process

We have used the proposed dependency annotation scheme to build KGPBenTreebank[3]. Then we have analyzed both the annotation scheme and the resource using

[3] The annotation has been done using the Sanchay annotation tool of Singh (2011)

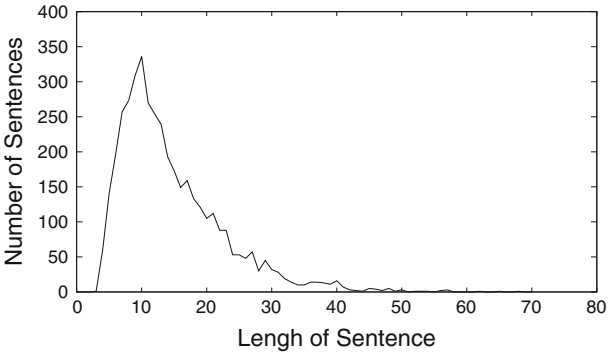

Fig. 2 Number of sentences with their length

standard analysis methods used in many other Treebanks. These analyses show the usability of the resource for other research purposes.

6.1 Corpus statistics

We have used the dependency relations to annotate a treebank of 4167 sentences (56,514 words). The sentences are taken from Blogs, Multikulti (www.multikulti.org.uk), Wikipedia and a portion of CIIL corpus. Each lexical item is annotated manually in two different levels namely lexical category and morphological values as a part of the Part-of-Speech Tagset (IL-POST) project undertaken by MSR India. This two level manually annotated sentences are used as input to our annotation task. The distribution of the length (number of words) of the sentences are plotted in Fig. 2. The length varies from 3 to 67 words, while 93.95 % sentences (3913 sentences) have length between 4 and 29.

6.2 Tagset overview

We have sorted the relations based on their number of occurrences in the KGPBenTreebank. Some relations from the top and bottom positions of the list and their number of occurrences are shown in Table 2. Maximum number of usage is found for ccof relation as there are two ccof tags from the conjunct to each of the parts which are joined by the conjunct. Next highest number of occurrence is found for jnmod due to its usage for different types of modifiers like quantifier, numerical modifier, etc. Number of occurrences of pof indicates the number of complex predicates in the corpus. There are 1039 occurrences of clausalk2. k5d, clausalk7p and clausalk7d have limited usage in Bangla.

The distribution of the occurrence of dependency relations in the annotated corpus is shown in Fig. 3. Among 63 dependency relations 46 relations have been used less than 1,000 times and 10 relations have been used more than 2000 times. So, relations of the tagset have been used to capture the relations of small number of word pairs.

Table 2 Top members with highest and lowest number of occurrences

Top members	No occurrences	Bottom members	No occurrences
ccof	5025	k5d, clausalk7p	1
jnmod	4952	clausalk7d	3
k1g	4312	anmod	14
end	4006	rasneg	15
r6	3397	k1p	26
nnmod	3143	k5p	27
k2t	3098	k5t	35
pof	2562	compr	38

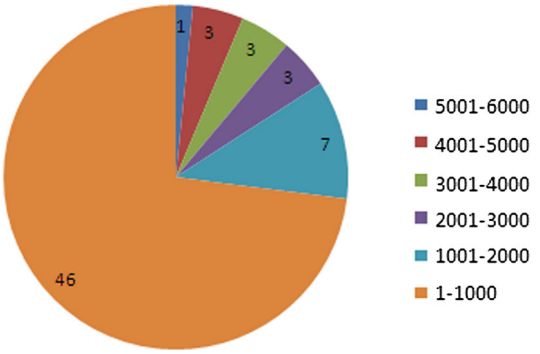

Fig. 3 Distribution of occurrences of the dependency relations

6.3 Inter-annotator agreement

Three annotators have annotated KGPBenTreebank for 3 years. To calculate the agreement between the annotators, a part of KGPBenTreebank containing 485 sentences have been annotated independently by all of them. We calculate the inter-annotators agreement using Fleiss' Kappa as discussed in Fleiss (1971).

Fleiss' Kappa gives a measure (between 0 and 1) of agreement for more than 2 annotators. It is a generalization of the Scott's pi statistics of Scott (1955) for inter-annotator readability. Fleiss' Kappa is calculated as follows.

Suppose N is the total number of words, n is the number of annotators assigning dependency relations to the words and K is the number of dependency relations used for assignment. $N \times n$ is the total number of assignments of relations made by the annotators. Let the subscript i, where $i = 1, \ldots, N$, represent the words and the subscript j, where $j = 1, \ldots, K$ represents the dependency relations. So, n_{ij} is the number of annotators assigning *ith* word to *jth* dependency relation. The proportion of all assignments used for assigning *jth* dependency relation may be defined using Eq. (1). The mean proportion of assignments on all dependency relations may be defined using Eq. (2).

$$P_j = \frac{1}{N \times n} \sum_{i=1}^{N} n_{ij} \tag{1}$$

$$\overline{P_e} = \sum_{j=1}^{K} P_j^2 \tag{2}$$

Among $\frac{n(n-1)}{2}$ pairs of annotators, the extent to which the annotator pairs are agreed for the *ith* word is defined using Eq. (3). The mean of agreements for all words may be defined using Eq. (4).

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{K} n_{ij}(n_{ij} - 1) \tag{3}$$

$$\overline{P} = \frac{1}{N} \sum_{i=1}^{N} P_i \tag{4}$$

The degree of agreement between n annotators is computed in terms of Fleiss' Kappa (κ) using Eq. 5.

$$\kappa = \frac{\overline{P} - \overline{P_e}}{1 - \overline{P_e}} \tag{5}$$

When the annotators are agreed on all assignments, then $\kappa = 1$. The calculated Fleiss' Kappa for the three annotators on 485 sentences containing 6356 words is calculated to be 0.9288. So, the annotators have almost perfect agreement while annotating KGPBenTreebank.

6.4 Length of dependencies

The average length of dependency relations is 2.82 while the longest and shortest dependency relations are *clausalcondition* and *qsk2* and their average lengths are 8.35 and 1, respectively.

In KGPBenTreebank, among 56,524 words, 4167 words (one word from each sentence) are the roots of the sentences. Root words of the sentences are not attached to any other word. The remaining 52,357 words of KGPBenTreebank are attached to some other words of the sentence. Among them, 28876 words are related to adjacent words (previous or next word), 8290 words are related to the words with distance 2 (previous to previous or next to next word), and so on. The increasing length of the dependency relations is plotted against the log of the number of words which are attached to correspondingly distant words in Fig. 4. First two points of the plot are (X=1, Y=log_{10} (28876)) and (X=2, Y=log_{10} (8290)).

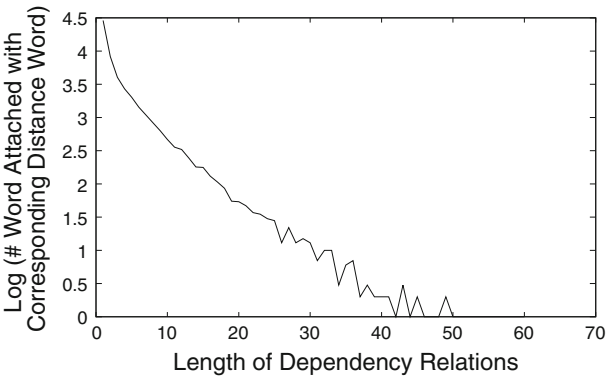

Fig. 4 Distribution of Words based on the length of their head dependents

From this plot we observe that relations in KGPBenTreebank are mostly short distance relations and as the length of the relations increases the log of number of words related to the word with that distance decreases almost exponentially.

7 Comparison between Hindi and Bangla schemes

The dependency annotation model discussed in this paper borrows from the Paninian grammatical model of Bharati et al. (1999), Hindi dependency scheme of Sharma et al. (2007) used in Anncorra, the typed dependency relation model de Marneffe and Manning (2008) followed in Stanford Parser v1.6.9 and the modern Bangla grammatical model discussed by Chatterji 2003 and Bamandev Chakravarty 2010.

In both KGPBenTreebank and Anncorra, Karak relations are divided into subcategories (finer classes). The differences of these division schemes are mentioned in Sect. 7.1. A detailed explanation on the differences between them is given in Sect. 7.2.

7.1 Differences of Karak division schemes

In KGPBenTreebank, five finer divisions of subjective case (karta karak) are used on the basis of the type of the activity of the subject.

1. doer subject (Kriya sampadak karta)
2. experiencer subject (Anubhab karta)
3. passive subject (Paroksha karta)
4. noun of proposition (Bidheya karta or Samanadhikaran)
5. general subject (Sadharan karta)

In Anncorra, six finer divisions of subjective case (karta karak) are used.

1. doer (karta)
2. causer subject (prayojaka karta)

3. causee subject (prayojya karta)
4. mediator causer (madhyastha karta)
5. noun complement of subject (karta samanadhikarana)
6. clausal subject

In KGPBenTreebank, five finer divisions of objective case (karma karak) are used.

1. transitive object (sakarmak karma)
2. direct object (mukhya karma)
3. indirect object (gauna karma)
4. purposive object (uddyeshya karma)
5. predicative object (bidheya karma)

In Anncorra, four finer divisions of objective case (karma karak) are used.

1. direct object (mukhya karma)
2. indirect object (gauna karma)
3. object complement (karma samanadhikaran) [In KGPBenTreebank it occurs as purposive object, predicative object and complement clause.]
4. goal, destination (k2p) [In KGPBenTreebank we have used it as other cases-alike relation.]

In KGPBenTreebank, sampradan karak is not considered and is merged with object/karma.

There are two finer divisions of sampradan (k4) in Anncorra.

1. recipient (sampradan)
2. experiencer karta (anubhava karta)

In KGPBenTreebank, place and time are tagged differently. Therefore, here, there are four finer divisions of ablative case (apadan karak).

1. place related ablative (sthanbachak apadan)
2. state related ablative (abasthabachak apadan)
3. time related ablative (kalbachak apadan)
4. distance related ablative (duratbabachak apadan)

In Anncorra, ablative case (apadan karak) has two finer divisions.

1. apadan karak/source which is related to both place and time.
2. prakriti apadan 'source material' in verbs denoting change of state.

In KGPBenTreebank, locative case (adhikaran karak) has four finer divisions as domain and state or condition are treated in different way.

1. place related locative (deshadhikaran)
2. time related locative (kaladhikaran)
3. domain related locative (bishayadhikaran)
4. state related locative (bhabadhikaran)

In Anncorra, locative case (adhikaran karak) has three finer divisions.

1. place related locative (deshadhikaran)
2. time related locative (kaladhikaran)
3. domain related locative (vishayadhikaran)

The relations used for connecting two clauses in KGPBenTreebank are given below. Here, '*' can be replaced by the relation of the two clauses.

1. referent (nirdesak)
2. clausal* (bakyamsha samagra)
3. clausal complement (bakyamsha sampurak)
4. complementizer (sampurak)

Anncorra has three types of relative clauses.

1. nmod_relc (relative clause constructions modifying a noun)
2. rbmod_relc (jo-vo construction modifying an adverb)
3. jjmod_relc (jo-vo clause construction modifying an adjective)

7.2 Overall differences of annotation schemes

Extensive work has been done on developing Hindi treebank. The dependency annotations scheme of Sharma et al. (2007), Begum et al. (2008) are used for annotating Anncorra. We have borrowed some features of these annotation schemes in the proposed Bangla dependency annotation scheme. Therefore, we discuss the differences between the existing Hindi annotation schemes and the scheme proposed in this paper.

Each difference is explained with an example and a dependency structure. The attachments and tags used by Anncorra which are different from KGPBenTreebank, are shown using thick lines and boldface in the dependency structures. The intrachunk relations of KGPBenTreebank are not included in Anncorra chunk level dependency treebank. The intrachunk attachments are shown using dotted lines and intrachunk relations are underlined.

- Unlike Hindi treebank, causative subject [prayojaka karta] (pk1), mediator causer [madhyastha karta] (mk1), causee subject [prayojo karta] (jk1) are not used in KGPBenTreebank. Rather we considered pk1 as karta, mk1 as karan (as both of them have similar syntactic structure), and jk1 as karma. If there is another karma in the sentence then jk1 has been considered as mukhya karma and the other one as gauna karma. Suniti Kumar Chatterji Chatterji (2003) and Bamandev Chakravarty (2010) have given similar explanation in this concept. The difference is shown with the dependency structure of example 57 in Fig. 5.

  (57) sitA AYA dbArA bAchchAke khAbAra khAoYAchchhe. (Bangla)
  [Sita nurse by child-acc food feed-pre,prog]
  sitA AyA se bachche ko khAnA khilA rahI hai. (Hindi)
  Sita is feeding food to child by the nurse.

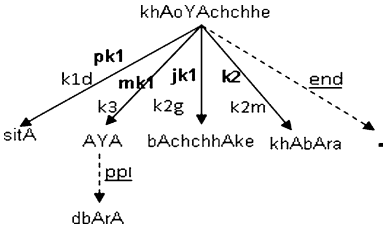

Fig. 5 Dependency Tree of example 57

- Following the opinion in Chatterji (2003) we have merged recipient (sampradan karak) relation with the object (karma karak) relation as explained in Sect. 5.2. See dependency structure of example 58 in Fig. 6a.

  (58) rAma mohanake kShIra dila. (Bangla)
  [Ram Mohan-acc rice-pudding give-past]
  rAma mohana ko kShIra dI. (Hindi)
  Ram gave rice-pudding to Mohan.

- Prati upapad 'direction' (*rd*) relation of Hindi treebank is not used in the KGPBenTreebank. We have considered that the direction towards it is moving is the place where the subject wants to go or reach. Therefore, the corresponding relation is treated as destination/gantabyasthal (des). See dependency structure of example 59 in Fig. 6b.

  (59) sitA grAmera dike yAchchhila. (Bangla)
  [Sita village-gen direction-loc go-past,prog]
  sitA gA.Nba kI aura jA rahI thI. (Hindi)
  Sita was going towards her village.

- In respect of locative case, Anncorra treats both domain and state (or condition of thing or mind) as vishayadhikaran/location. However, KGPBenTreebank has two different tags for these two different cases. See dependency structures of examples 60 and 61 in Fig. 7a, b, respectively.

  (60) tArA sukhe Achhe. (Bangla)
  [he-pl happy-loc be-pre]
  be khusha hai. (Hindi)
  They are living with happiness.

  (61) se sa.ngIte pAradarshI <NULL>. (Bangla)
  [he song-loc expert]
  bo gAne me mAhira hai. (Hindi)
  He is an expert in song.

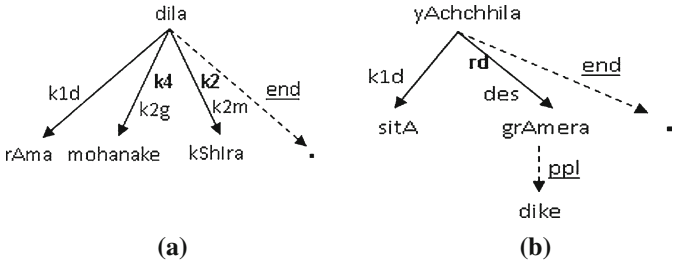

Fig. 6 **a** Dependency Tree of example 58. **b** Dependency Tree of example 59

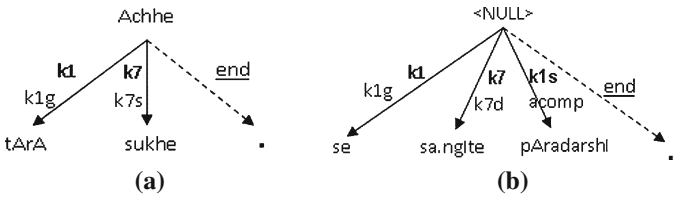

Fig. 7 **a** Dependency Tree of example 60. **b** Dependency Tree of example 61

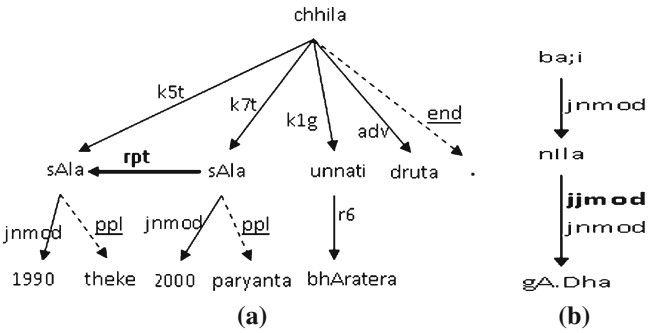

Fig. 8 **a** Dependency Tree of example 62. **b** Dependency Tree of example 63

- Relation point of time (*rpt*) relation is used in Anncorra to connect the starting point of the time or place with the the ending point of the time or place. In KGPBenTreebank, the starting and ending point of times are considered as two different karaks. The starting point is tagged as *k*5*t* and the ending point as *k*7*t*. Similarly, the starting and ending point of places are tagged as *k*5*p* and *k*7*p*,

respectively. This difference is shown using the dependency structures of example 62 in Fig. 8a.

(62) 1990 sAla theke 2000 sAla paryanta bhAratera unnati druta chhila. (Bangla)
[1990 year from 2000 year till India-gen development fast be-past]
sana 1990 se 2000 taka bhArata kI pragati teja rahI. (Hindi)
During the period from 1990 to 2000 Indias development was rapid.

- In Anncorra, the intensifiers modifying adjectives are tagged as *jjmod*. Whereas in KGPBenTreebank, both the relations between the two adjectives and between the adjective and noun are denoted by the same tag *jnmod*. See dependency structures of example 63 in Fig. 8b.

(63) gA.Dha nIla ba;i (Bangla)
[deep blue book]
gaharA nIlI kitAba (Hindi)
Deep blue book.

- In Anncorra, karta is inserted in recoverable condition i.e., when it is not present in the surface but is recoverable from other parts of the sentence. A <NULL> node is created to represent the absent karta. But, in KGPBenTreebank, karta is not inserted in any condition. See dependency structures of example 64 in Fig. 9.

(64) rAma o shyAma hoTele khAbAra khela Ara sinemA dekhala.(Bangla)
[Ram and Shyam hotel-loc food eat-past and movie see-past]
rAma aura shyAma hoTela para khAnA khAyA aura sinemA dekhA. (Hindi)
Ram and Shyam had food in the hotel and watched a movie.

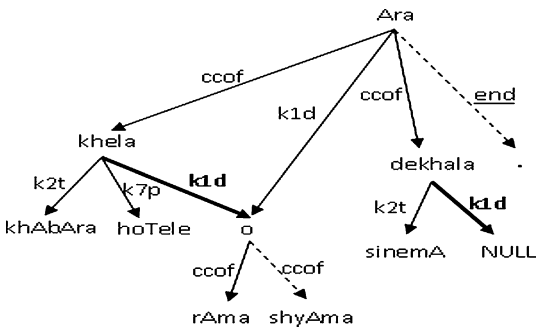

Fig. 9 Dependency Tree of example 64

8 Conclusion

The present paper is on building a dependency annotation scheme for modern Bangla language. The relationship set in the proposed scheme is categorized into 3 levels: intrachunk relations, interchunk relations and interclause relations. Though most of the relations in the treebank have syntactic orientation, semantic relations also have been considered in some cases where the need is felt.

The scheme has been created with the help of traditional Bangla grammar books, the existing schemes for Indian languages and the modern Bangla grammar books. Further, the scheme is corrected and enriched during the annotation process with the help of the annotators and some Bangla language experts. Then the annotated corpus is corrected based on the modified tagset. Further study would be more effective for the improvement or enrichment of the scheme.

Three annotators have electronically annotated 4167 Bangla sentences to test the proposed scheme. The inter-annotator agreement value on 485 Bangla sentences is found to be 0.9288 in terms of Fleiss' Kappa which indicates the consistency of the treebank. However, analysis is required to find and correct the mistakes in the annotation.

Appendix 1: The Relation set of the Bangla Treebank

Intrachunk relations		
ppl	Postposition/Anusarga	Rel. with noun/pron.
stc	Spatio-temp. con./Sthan-samay. samp.	Re. with space-time noun
vx	Auxiliary verb/Sahayak kriya	Related with verb
pof	Part of/Kriya antargata bisheshya	”
redup	Reduplication/Shabda dbaita	Rel. with same rhym.
frag	Fragment/Bhagnamsha	Related with suffix
Karak relations		
k1d	Doer subject/Kriya sampadak karta	Related with verb
k1e	Experiencer subject/Anubhab karta	Related with verb
k1p	Passive subject/Paroksha karta	Related with verb
k1s	Noun of proposition/Samanadhikaran	Related with verb
k1g	General subject/Sadharan karta	Related with verb
k2t	Transitive object/sakarmak karma	Related with verb
k2m	Direct object/Mukhya karma	Related with verb
k2g	Indirect object/Gauna karma	Related with verb
k2u	Purposive object/Uddyeshya karma	Related with verb
k2s	Predicative object/Bidheya karma	Related with verb
k3	Instrumental/Karan	Related with verb
k5p	Place rel. ablative/Sthanbachak apadan	Related with verb
k5s	State rel. ablative/Abasthabachak apadan	Related with verb
k5t	Time rel. ablative/Kalbachak apadan	Related with verb
k5d	Dist. rel. ablative/Duratbabachak apadan	Related with verb

k7p	Place rel. locative/Deshadhikaran	Related with verb
k7t	Time rel. locative/Kaladhikaran	Related with verb
k7d	Domain rel. locative/Bishayadhikaran	Related with verb
k7s	State rel. locative/Bhabadhikaran	Related with verb
rh	Reason/Hetu	Related with verb
ru	Purpose/Uddeshya	Related with verb
des	Destination/Gantabyasthal	Related with verb
r6v	Possession/Dakhal	Related with verb
compr	Comparison/Taratamya	Related with any
sim	Similarity/Sadrishya	Related with any
Modifier Relations		
r6	Genitive/Sambandha	Related with noun
ras	Associative relation/Saharthak sambandha	Related with noun
rasneg	Non-associative relation/Namarthak sambandha	Related with noun
nnmod	Noun noun modifier/Sanyogmulak bisheshya	Related with noun
jnmod	Adj. noun mod./Bisheshyer bisheshan	Related with noun
dnmod	Dem. noun mod./Nirnay suchak sarbanam	Related with noun
pronmod	Pron. noun mod./Sarbanamjata bisheshan	Related with noun
pnmod	Participial noun mod./Kridanta bisheshan	Related with noun
anmod	App. noun mod./Tulyarupe sthapita bisheshan	Related with noun
adv	Adv. mod./Kriya bisheshan jatiya bisheshan	Related with verb
vmod	Verb-verb modifier/Kriya jatiya bisheshan	Related with verb
neg	Negation modifier/Namarthak abyay	Related with verb
acomp	Adjectival Complement/Bidheya bisheshan	Related with verb
Few other interchunk relations		
ccof	Conjunct/Samyojak abyay	Rel. with conjunct
pcc	Preconjunct/Abasthatmak abyay	Related with SC.
rad	Address word/Sambodhan sabda	Related with verb
par	Particle/Bakyalankar abyay	Related with verb
qs	Question mark/Prashnabodhak chihna	Related with verb
end	End/Samapti	Related with verb
sym	Symbol/Chihna	Related with verb
Interclause relation		
ref	Referent/Nirdesak	Rel. with noun/pron
clausal*	Clausal star/Bakyamsha samagra	Related with verb
clausalcomp	Clausal complement/Bakyamsha sampurak	Related with verb
comp	Complementizer/Sampurak	Related with verb

rel.-related, pron.-pronoun, rhym.-rhyming word, mod.-modifier, adj.-adjectival, dem.-demonstrative, app.-appositional, adv.-adverbial, nom.-nominal, bish.-bisheshan, comp.-comparison, sim.-similarity, SC.-subordinating conjunction, temp.-Temporal, con.-Connection, samay.-Samaygata, samp.-Samparka

Appendix 2: Itrans to glyphs in Bangla and Hindi scripts mapping

a অ **अ**	A আ **आ**	i ই **इ**	I ঈ **ई**	u উ **उ**	U ঊ **ऊ**
RRi ঋ **ऋ**	e এ **ए**	ai ঐ **ऐ**	o ও **ओ**	au ঔ **औ**	
k ক **क**	kh খ **ख**	g গ **ग**	gh ঘ **घ**	~N ঙ **ङ**	
ch চ **च**	chh ছ **छ**	j জ **ज**	jh ঝ **झ**	~n ঞ **ञ**	
T ট **ट**	Th ঠ **ठ**	D ড **ड**	Dh ঢ **ढ**	N ণ **ण**	
t ত **त**	th থ **थ**	d দ **द**	dh ধ **ध**	n ন **न**	
p প **प**	ph ফ **फ**	b ব **ब**	bh ভ **भ**	m ম **म**	
y য **य**	Y য় **य़**	r র **र**	l ল **ल**		
sh শ **श**	Sh ষ **ष**	s স **स**	h হ **ह**		
.D ড় **ड़**	.Dh ঢ় **ढ़**	.n ং ं	H ঃ ः	.N ঁ ँ	

References

Begum, R., Husain, S., Dhwaj, A., Misra, D., Bai, L., & Sangal, R. (2008). Dependency annotation scheme for indian languages. In *Proceedings of the third international joint conference on natural language processing(IJCNLP)*. Hyderabad, India.

Bharati, A., Chaitanya, V., Sangal, R. (1999) Natural language processesing: A paninian perspective. New Delhi: Prentice-Hall of India.

Bharati, A., Sangal, R., Chaitanya, V., Kulkarni, A., Sharma, D. M., & Ramakrishnamacharyulu, K. V. (2002). Anncorra: building tree-banks in indian languages. In *Proceedings of the 3rd workshop on Asian language resources and international standardization* (Vol. 12, pp. 1–8), COLING '02.

Bhatt, R., Narasimhan, B., Palmer, M., Rambow, O., Sharma, D. M., & Xia, F. (2009). A multi-representational and multi-layered treebank for hindi/urdu. In *Proceedings of the third Linguistic annotation workshop*, ACL-IJCNLP '09, (pp. 186–189). Association for Computational Linguistics, Stroudsburg, PA, USA. URL http://dl.acm.org/citation.cfm?id=1698381.1698417

Black, E., Eubank, S., Kashioka, H., Magerman, D., Garside, R., & Leech, G. (1996). Beyond skeleton parsing: Producing a comprehensive large-scale general-English treebank with full grammatical analysis. In *Proceedings of the 17th international conference on computational linguistics* (COLING-96), (pp. 107–112).

Black, E. W., Garside, R., & Leech, G. N. (Eds.) (1993). *Statistically-driven computer grammars of English: The IBM/Lancaster approach. No. 8 in Language and Computers*. Amsterdam. http://books.google.de/books?id=Hkzr-LYVz2wC&lpg=PR5&ots=QJhw16OVS4&dq=Statistically-driven%20computer%20grammars%20of%20English&lr&pg=PP1#v=onepage&q&f=false

Chakravarty, B. (2010). *"uchchatara bangla vyakaran", a complete text book on higher bengali grammar*. Akshay Malancha.

Charniak, E., Blaheta, D., Ge, N., Hall, K., Hale, J., & Johnson, M. (2000). Bllip 1987–89 wsj corpus release 1. Linguistic Data Consortium.

Chatterji, S., Sarkar, T. M., Sarkar, S., & Chakraborty, J. (2009). Karak relations in bengali. In *Proceedings of 31st All-India conference of Linguists (AICL 2009)*, (pp. 33–36). Hyderabad, India.

Chatterji, S. K. (2003). *Bhasha-prakash bangala vyakaran [a grammar of the bangla language]*. Calcutta: Roopa and Company.

Chopde, A. (2000). Itrans "indian language transliteration package", a package for printing text in indian language scripts. http://www.aczone.com/itrans/.

Dandapat, S., Sarkar, S., & Basu, A. (2004). A hybrid model for part-of-speech tagging and its application to bengali. In *International conference on computational intelligence*, (pp. 169–172).

de Marneffe, M., & Manning, C. D. (2008). Stanford typed dependencies manual.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382.

Hajič, J., Böhmová, A., Hajičová, E., & Vidová-Hladká, B. (2000). The prague dependency Treebank: A three-level annotation scenario. In A. Abeillé (Ed.), *Treebanks: Building and using parsed corpora* (pp. 103–127). Amsterdam: Kluwer.

Hajič, J., Hajičová, E., & Rosen, A. (1996). Formal representation of language structures. *TELRI Newsletter*, *3*, 12–19.

Hajič, J., Vidová-Hladká, B., & Pajas, P. (2001). The prague dependency Treebank: Annotation structure and support. In *Proceedings of the IRCS Workshop on Linguistic Databases*, (pp. 105–114). Philadelphia, USA: University of Pennsylvania.

Karlsson, F., Voutilainen, A., Heikkilä, J., & Anttila, A. (Eds.) (1995). *Constraint Grammar: A language-independent system for parsing unrestricted text*. Berlin: Mouton de Gruyter.

Marcus, M.P., Marcinkiewicz, M.A., & Santorini, B. (1993). Building a large annotated corpus of english: the penn treebank. *Computational Linguistics 19*, 313–330. http://dl.acm.org/citation.cfm?id=972470.972475

McCord, M. C. (1990). Slot grammar: A system for simpler construction of practical natural language grammars. In R. Studer (Ed.), *Natural Language and Logic: Proceedings of the international scientific symposium, Hamburg, FRG*, (pp. 118–145). Berlin: Springer.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, *31*, 71–106. doi:10.1162/0891201053630264.

Santorini, B., & Marcinkiewicz, M.A. (1991). Bracketing guidelines for the penn treebank project. unpublished manuscript.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, *19*, 321–325.

Sharma, D.M., Sangal, R., Bai, L., Begam, R., Ramakrishnamacharyulu, K. (2007). Anncorra : Treebanks for Indian languages, annotation guidelines (manuscript).

Singh, A. K. (2011). Part-of-speech annotation with sanchay. In *Proceedings of the National Seminar On POS annotation for Indian Languages: Issues & Perspectives*. Mysore, India.

Xue, N., Xia, F., Chiou, F.D., Palmer, M. (2005). The penn chinese treebank: Phrase structure annotation of a large corpus. In *Natural Language Engineering*.