International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 01 | Jan - 2025 SJIF Rating: 8.448 ISSN: 2582-3930

# Modeling of Link Grammar Parser for Parsing of Kumauni Complex Sentences

Rakesh Pandey
HSB Govt. P.G. College, Someshwar (Almora)
Uttarakhand (India)
rakeshpandeyaits@gmail.com

**Abstract**

The Kumauni language, a regional dialect spoken in the Kumaun region of the Himalayas, remains relatively understudied. This research aims to develop a parsing tool to facilitate linguistic studies of Kumauni. The primary objective is to establish a method for analyzing the grammatical structures of complex sentences in the language. To achieve this, a set of existing complex Kumauni sentences was analyzed to derive grammatical rules, which were then incorporated into a Link Grammar Parser model. The resulting model offers a new tool for parsing Kumauni complex sentences, supporting further research and linguistic analysis.

**Keywords**

Complex sentence, Link Grammar Parser, Kumauni dialects, Clauses in sentences

## 1. Introduction

Parsing, also called clause analysis, is a traditional grammatical exercise that involves breaking a text into its components and explaining their forms, functions, and syntactic relationships. This process largely relies on understanding a language's conjugations and declensions, which can be particularly complex in highly inflected languages. For example, parsing the phrase "man bites dog" would involve identifying "man" as the singular noun serving as the subject, "bites" as the third-person singular present tense form of the verb "to bite," and "dog" as the singular noun functioning as the object. Sentence diagrams are sometimes used to visually represent these relationships.

In certain machine translation and natural language processing systems, computer programs parse written texts in human languages. However, parsing human sentences is challenging due to significant ambiguity in language structure. Human language is used to convey meaning across an infinite range of possibilities, only some of which are relevant in specific contexts. For instance, while "Man bites dog" and "Dog bites man" are distinct in English, another language might present both as "Man dog bites," relying on context to differentiate the meanings—assuming the distinction even matters. Crafting formal rules to capture informal linguistic behavior is difficult, despite the fact that some patterns clearly exist.

Building on the computational Panini grammar introduced by Bharati et al. (1995) and Patil et al. (2013), this study proposes Karaka links to define relationships between nominal words and verbs in sentences, as summarized in Table 1. Historically, parsing was a fundamental aspect of grammar education across English-speaking regions, essential for understanding written language, though such methods are no longer widely taught.

Unlike programming languages with structured grammars, natural language parsing algorithms cannot assume orderly grammatical properties. Some grammar formalisms are computationally challenging to parse, necessitating the use of context-free grammar

approximations for initial parsing attempts. Context-free grammar algorithms often employ variations of the CYK algorithm, enhanced by heuristics to eliminate unlikely analyses for efficiency.

Seungmi Lee (1997) introduced a re-estimation algorithm, an adaptation of the inside-outside algorithm for probabilistic dependency grammars (PDG), along with a best-first parsing (BFP) approach. Hoifung Poon et al. (2006) presented the first unsupervised method for learning semantic parsers using Markov logic. The dynamic programming-based Earley's algorithm (1970) applies to all CFGs, with a worst-case complexity of $O(N^3)$ and $O(N^2)$ for unambiguous grammars, widely used in computational linguistics.

Vaishali et al. (2014) proposed methods for linking complex sentence clauses and modeled them using the Link Grammar Framework for parsing. Pandey et al. (2010) adapted the Earley algorithm for parsing Kumauni sentences.

Different parsing algorithms generally place various restrictions on the grammar of the language to be parsed

- Top- down
- Bottom - up
- Recursive descent
- LL
- LR
- LALR
- SLR
- CYK
- GLR
- Simple precedence parser

- **Link Grammar:** The Link Grammar Parser is a syntactic tool for analyzing English sentences, built on the link grammar framework—an innovative theory of English syntax. It assigns a syntactic structure to a given sentence by creating labeled links that connect pairs of words. Additionally, the parser generates a "constituent" representation, highlighting elements like noun and verb phrases. Unlike traditional parsing methods that rely on part-of-speech tags and rules, link grammar classifies natural languages by establishing connections between word sequences. For a sentence to be recognized as part of a language under link grammar, it must satisfy three specific conditions.
  - Planarity: The links do not cross (when drawn above the words).
  - Connectivity: The links suffice to connect all the words of the sequence together.
  - Exclusion: Links must satisfy the linking requirements of each word in the sentence. Linking requirements are defined in the dictionary of the grammar.

The link grammar provides a more flexible approach to "tagging" compared to the Penn Treebank tags, with a focus on links or connectors between words. For words to form links, they must have compatible connectors, which are specified in a dictionary containing each word's linking requirements. For example, a simple dictionary might outline the linking needs for words like "a," "the," "cat," "snake," "Mary," "ran," and "chased," with visual diagrams representing their linking rules.

Link Grammar operates on a key natural language principle: if arcs are drawn between related words in a sentence, these arcs should not cross. Sleator et al. (1991) developed a parsing system capturing various aspects of English grammar through around 700 definitions specifying words and their linking requirements. J. Lafferty et al. (1995) introduced an algorithm for parsing sentences based on this grammar. A sentence is accepted by the system if three conditions are met: the linking requirements for all words are satisfied (connectivity property), no links cross (planarity property), and there is at most one link between any word pair (exclusion property).

Shailly Goyal et al. (2006) conducted an analytical study of Hindi complex sentence structures, proposing a parsing scheme that not only generates parse structures for complex Hindi sentences but also verifies compliance with planarity conditions for LG-based parsing. As part of the Indo-Aryan dialect continuum, Kumauni shares grammatical features with several other Indo-Aryan languages, including Nepali, Hindi, Rajasthani, Kashmiri, and Gujarati. Much of its grammar is also common to other Central Pahari languages. Unique grammatical traits in Kumauni and other Central Pahari languages are attributed to the influence of the now-extinct Khasas language, spoken by the region's earliest inhabitants. In Kumauni, the verb substantive derives from the root "ach." Classified as part of the Central subgroup of Pahari languages, Kumauni has a rich literary tradition, though its dialects do not yet qualify as a distinct language (Bhasa). Devdatta Sharma (1985), a prominent linguist, was the first to conduct a detailed linguistic study of Kumauni. Building on his work, this research focuses on processing Kumauni for grammar validation in input sentences. Parsing involves two components: a parser (procedural) and a grammar (declarative). While the parser remains constant, the grammar must be tailored to the specific language being parsed. By defining a grammar for Kumauni and using the Link Grammar Parser, this study parses complex Kumauni sentences based on a collection of pre-existing sentence structures.

2. **Kaaraks in Kumauni dialect:** Kaarak are the words which are commonly used before a noun, noun phrase, pronoun or verb. Also, connects with the noun, pronouns and phrases to other words in a sentence in order to make the sentence comprehensible for readers.

Example: Ram ne dande se ghode ko pita (राम ने डंडे से घोडे को पीटा।): “Ram bitted the horse with stick" - Here 'Ne' 'Se' and ‘Ko’ are kaarak.

Like as Hindi language, Kumauni has also eight Kaaraks;

1) Karta kaarak.(कर्ता कारक) - Nominative Case

2) Karma kaarak (कर्म कारक) - Instrument Case

3) Karan kaarak (करण कारक) - Ablative Case

4) Sampradan kaarak (संप्रदान कारक) - Possessive Case

5) Apadan Kaarak (अपादान कारक) - Objective Case

6) Sabandh kaarak (संबंध कारक) - Dative Case

7) Adhikaran kaarak (अधिकरण कारक) - Locative Case

8) Sambodhan kaarak (संबोधन कारक) - Vocative Case

It can be more understood from the table below:

| Kaarak | Sign | Meaning | Functionality |
|---|---|---|---|
| Karta | ले (le) | Who does work | Verb to Subject |
| Karma | को (ko) | The work done/ to be done | Verb to Object |
| Karan | बटी (bati) | By which karta does work | Verb to instrument of the activity |
| Sampradan | क, लिजी (k, ligi) | The work done for which | Verb to word which gives donation meaning |
| Apadan | से {अलग होना} (se) | The break | Verb to word which gives separation meaning |
| Sambandh | की, के, रा, री, रे (ki, ke, ra,ri, re) | Relation with other terms | indicates any relationship or connection that may exist between two individuals or objects |
| Adhikaran | में, बे (me, be) | Base of Karm (verb) | Verb to time and place of the activity |
| Sambodhan | हे !, अरे ! ऊजा ! (he! Are!, Uja!) | a sudden cry or remark expressing surprise | The Vocative case is used when directly addressing a person or group of people |

Table 1: Table of Kaaraks in Kumauni

**3. Complex sentence in Kumauni:** In Kumauni, as in other languages, complex sentences consist of an independent clause combined with one or more dependent (subordinate) clauses. These clauses are often connected by conjunctions, relative pronouns, or other linking words.

**Examples of Features in Kumauni Complex Sentences:**

I. **Conjunctions for Subordination:** Words like *ki* (that), *jaba* (when), and *agar* (if) are commonly used to link clauses.

II. **Relative Clauses:** Sentences often use relative pronouns such as *jo* (who, which) to describe or specify nouns.

III. **Conditional and Temporal Structures:** Words like *jab* (when) and *jab tak* (until) are used in time-bound complex sentences.

It usually precedes the correlative though other orders are also found. Each clause carries its own relative marker J and correlative marker T. Relative and correlative markers handled in our system are “Jail- Vail”, (जैल- वैल); “Jas-Tas” (जस - तस); “Jhar-Far” (झर - फर); “Gav-Gav” (गाव - गाव); Jaan-Waan” (जां-वां), “Jo-To” (जो-तो); “Jaik-Vaik (जैक-वैक); etc.

Being part of the Indo-Aryan dialect continuum Kumauni shares its grammar with other Indo-Aryan languages especially Nepali, Hindi, Rajasthani, Kashmiri and Gujarati. It shares much of its grammar with the other languages of the Central Pahari like Garhwali and Jaunsari. The peculiarities of grammar in Kumauni and other Central Pahari languages exist due to the influence of the now extinct language of the Khasas, the first inhabitants of the region. In Kumauni the verb substantive is formed from the root ach, as in both Rajasthani and Kashmiri. In Rajasthani its present tense, being derived from the Sanskrit present *rcchami*, I go, does not change for gender. But in Pahari and Kashmiri it must be derived from the rare Sanskrit particle **rcchitas*, gone, for in these languages it is a

participial tense and does change according to the gender of the subject. Thus, in the singular we have: - Here we have a relic of the old Khasa language, which, as has been said, seems to have been related to Kashmiri. Other relics of Khasa, again agreeing with north-western India, are the tendency to shorten long vowels, the practice of epenthesis, or the modification of a vowel by the one which follows in the next syllable, and the frequent occurrence of disaspiration. Thus, Khas siknu, Kumauni *sikhno*, but Hindi *sikhna*, to learn; Kumauni *yeso*, plural *yasa*, of this kind.

### 4. Modeling of Kumauni Complex Sentences for LG Parsing

Usually Kumauni complex sentence can be represented in various forms. Therefore we may have more than one parsing structure of the complex sentences. Our approach is to develop such linking scheme for Kumauni complex sentences which is consistent for all type of structures. In this respect the major confront dealing with complex sentences is crossing of the links. That is planarity rule. We can observ that, in general planarity cannot be maintained for Kumauni complex sentences. For example following complex sentence disobey the planarity rule if system constructs links in its customary style.

**Sentence:** Ji chhoral inaam jiti U Rameshak chyal chu (जी छोरल इनाम जीती उ रमेशक च्यल छू)/ The boy who won the prize is the son of Ramesh.

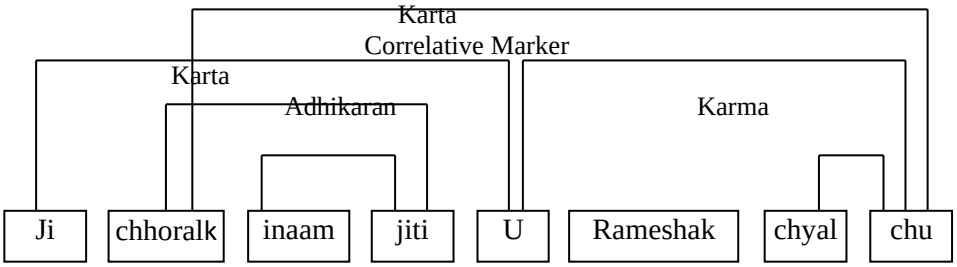


Fig. 1 Crossing of the Links

The crossing of the links occurs because of the correlative structure. In above example since chhorral (छोरल: boy) is subject of the verb phrase “is the son of Ramesh” (रमेशक च्यल छू: is the son of Ramesh), Karta karaka is also required in it and so crosses the correlative marker “U” (उ).

To overcome such conditions of crossing of links in complex sentences one can parse the sentence two levels:

1. In first level we can give the clausal links
2. In second level we can give the internal clausal links.

By the splitting the parse structure in two levels the upper level deals with relative-correlative marker and chunks of clauses and lower level deals with the words within the clause.

In this study we propose new links having valid and functional linkage between the words of complex sentences.

**Sentence:** Ji chhoral inaam jiti U Rameshak chyal chu (जी छोरल इनाम जीती उ रमेशक च्यल छू)/ The boy who won the prize is the son of Ramesh.

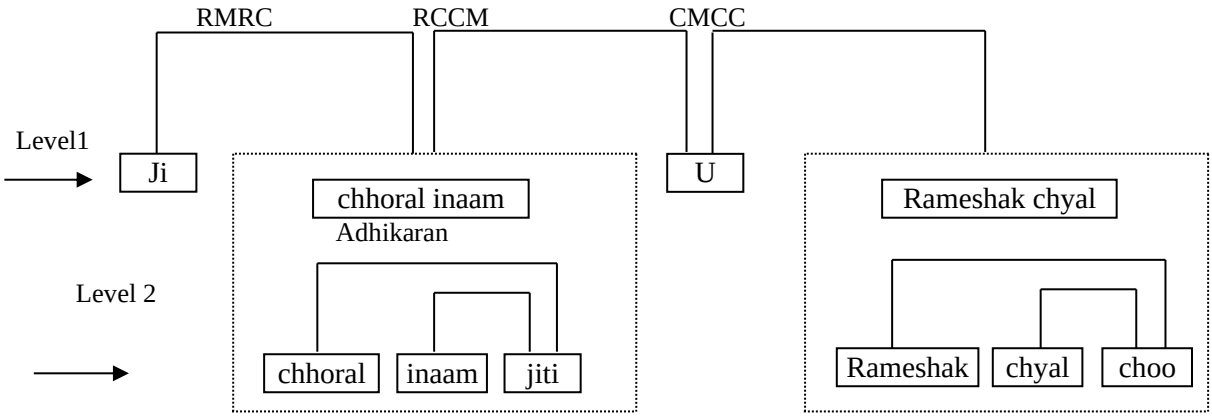


Figure 2 Two Level Linkage Parsing

Correlative Marker used in figure 2 is divided into two links RMRC and CMC. The links anticipated as shown in above figure 3 are RMRC which connects relative marker to relative clause, link CMC connects relative clause to correlative marker and link CMCC connects the correlative marker to correlative clause.

| ***Sr. No.*** | ***Function*** | | ***Link Name*** |
|---|---|---|---|
| | ***From*** | ***To*** | |
| 1 | Header | Main Clause | HM |
| 2 | Header | Complementizer | HCo |
| 3 | Main Clause | Complementizer | MCo |
| 4 | Complementizer | Complement Clause | CoCC |
| 5 | Complement Clause | Header | CCH |
| 6 | Subject | Header | SH |
| 7 | Complement Clause | Jaik Liji (“जैक लिजी”) | CCJL |
| 8 | Object | Complement Clause | OCC |
| 9 | Subject | Main Clause | SMC |
| 10 | Relative Marker | Relative Clause | RMRC |
| 11 | Relative Marker | Correlative Marker | RMCrM |
| 12 | Relative Clause | Correlative Marker | RCCrM |
| 13 | Correlative Marker | Relative Marker | CrMRM |
| 14 | Correlative Marker | Correlative Clause | CrMCrC |
| 15 | Correlative Clause | Relative Marker | CrCRM |
| 16 | Correlative Clause | Subject | CrCS |
| 17 | Relative Clause | Correlative Clause | RCCrC |
| 18 | Header | Subject | HS |

| | | | |
|---|---|---|---|
| 19 | Subject | Relative Clause | SRC |
| 20 | Adverbial Clause | Main Course | ACM |
| 21 | Main Clause | Conjunctive Particle | MCP |
| 22 | Conjunctive Particle | Adverbial Clause | CPAC |

Table 2: Proposed Links for Complex Sentence Structures

## 5. Modeled complex structure and proposed links in Kumauni LG parsing

The algorithm of proposed model can be demonstrated as:

Step1. Input sentence

Step2. Pre process

Step3. Apply parsing algorithm

If success

go to step 5.

else

Step4. Link dictionary with step 2

Go to step 2

Step5. Post processing

Step6. Parsed output

Stop

We propose several models according to the type of complex sentences existing for Kumauni dialects.

**Model (1):** In this model link proposed to connect complement type complex structure are HM which connects Header 'hi" to Main Clause, MCCo connect Main Clause to Complementizer "ki", CCoC connects Complementizer (Co) to Complement Clause (CC).

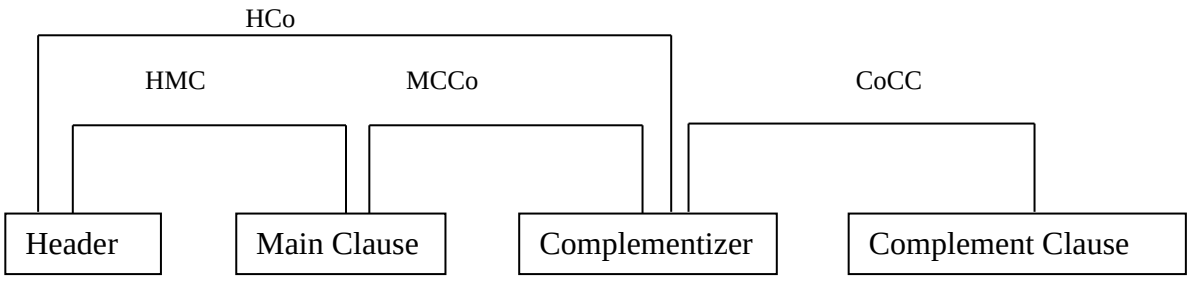


Figure 3: Model 1 for complex sentence structure

For example : Tau baat bari Khatarnaak chu ki baaghal Sureshak baakar kha haili (तौ बात बड़ी खतरनाक छू के बाघल सुरेशक बाकर खा हैली): It is very dangerous that tigers has eaten the goat of Suresh.

**Model (2):** This second model is developed for the condition if complementizer (Co) is absent. It is the variation of Complement Clause (CC). In such structures link CCH is used to join Complement Clause (CC) to Header (H).

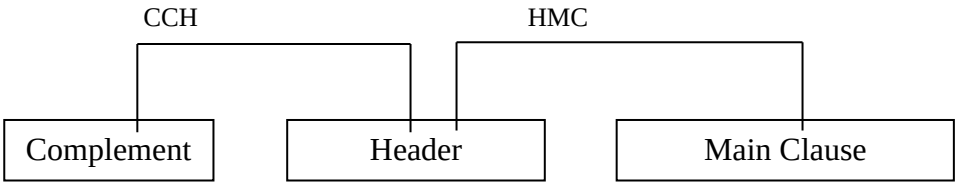


Figure 4: Model 2 for complex sentence structure

For example: Baaghal Sureshak baakar kha haili Tau baat bari Khatarnaak chu (बाघल सुरेशक बाकर खा हैली तौ बात बड़ी खतरनाक छू): Tiger has eaten the goat of Suresh it is very dangerous.

**Model (3):** This model is used for another variation of Complement Clause (CC), in this model header is absent and it is still grammatical. Link MCCo is used to join Main Clause (MC) to Complementizer (Co).

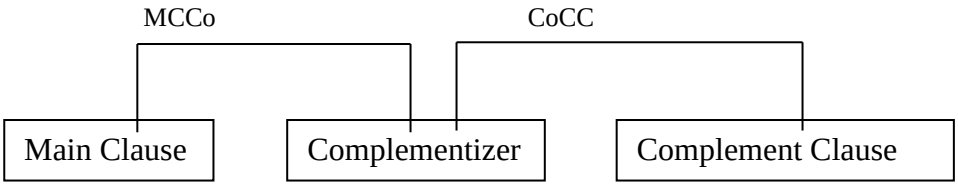


Figure 5: Model 3 for complex sentence structure

For example: Mee Jannu ki tum kis dagar pyaar karo (मी जानू की तुम कके दगढ़ प्यार करो।); I know that you love someone.

**Model (4):** If subject of Main Clause is separated from the Main Clause and position before complement clause without header. Link SH is proposed to connect subject with Header of Main Clause.

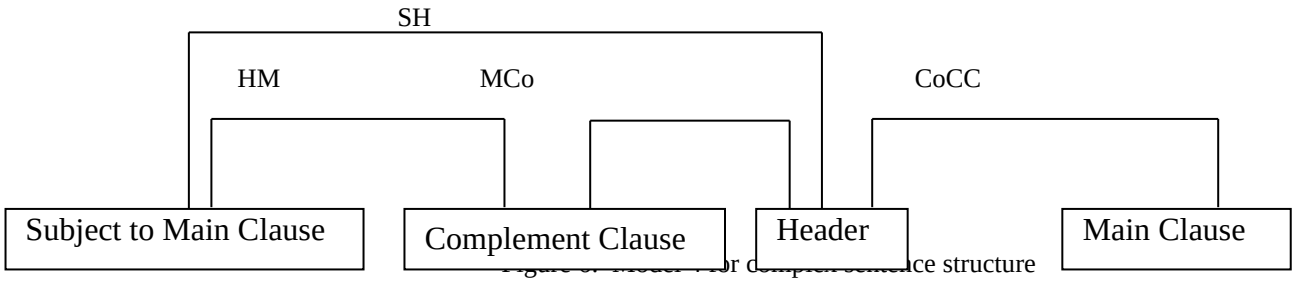


Figure 6: Model 4 for complex sentence structure

For Example: Meet um kake pyaar karo tas samajh chu (मी तुम कके प्यार करो तस समझ छू।); I think you love someone.

**Model (5):** In this model we have taken those sentences which contain other variations like deletion of relative marker that we can demonstrate in the following structure:

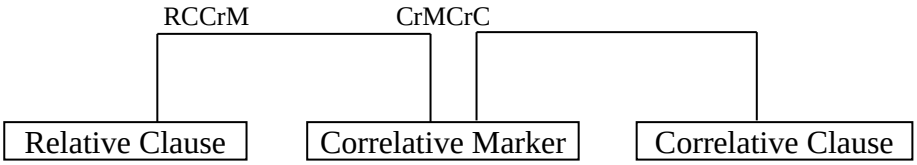


Figure 7: Model 5 for complex sentence structure

For Example: Jo dhok dino u chhor barh shath chhu (जो धोख दिनो उ छोर बढ़ शठ छू।); who cheats that boy is very cunning.

**Model (6):** This model has another variation of complex sentence

For example: Jo chhor barh shath chhu u chhor dhok dino (छोर जो बढ़ शठ छू उ छोर धोक दिनो),The boy who is cunning that boy cheats.

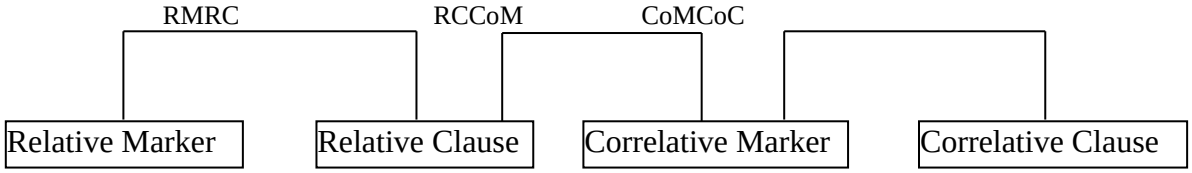


Figure 8: Model 6 for complex sentence structure

**6. Result** From the proposed models we can observe that in Correlative clause structures the following four types of patterns exist.

1. Free Relatives- These structures are headless relatives.
2. Gap Relatives- In these types of structures there is deletion of relative marker and common to both clauses.
3. Full Correlatives – In these sentences relative and correlative markers as well as clauses exists.
4. Multiple headed relatives- In these structures several Noun phrases are simultaneously relativized.

In this study we have modeled complex sentences in the form of possible valid linkage and proposed various links to connect the clauses in appropriate way. Our system identifies 16 such complex sentence structures.

## 7. Conclusion

This paper explained how link grammar parser can be used for parsing of Kumauni complex sentence. After intensive study of Kumauni complex sentences some links are proposed to develop connection between the clauses. Here we have proposed 22 new links for Complex Sentence Structures. To solve the issue of crossing of the links, two level links are also projected. This model can be used for other Indian languages to parse complex sentences.

## 8. Future Work

In this work, we have considered a limited number of Kumauni complex sentences for modeling of link grammar parser. We have also considered only twenty two links of complex sentence structures. In future work(s) related to the field of study covered in this paper,

an effort can be made to reflect on many more Kumauni complex sentences and more complex sentence structures, for developing a more effective model of Link grammar parser.

**References**

Bharati A., Chaitnya V. & Sangal R. (1995) *Natural Language Processing*: *A Paninian Perspective*, New Delhi: Prentice-Hall of India.

Devidatta, Sarma (1985) *The formation of Kumauni language*, (SILL: Series in Indian languages and linguistics), Bahri Publications.

Goyal Shailly and Chatterjee Niladri (2006) *A Scheme for Using Annotated English Complex Sentences to Parse Parallel Hindi Corpus*, In proceedings of the First National Symposium on Modeling and Shallow Parsing of Indian Languages-2006

Lafferty J., Grinberg D. & Sleator D. (1995) *A Robust Parsing Algorithm for Link Grammars*, Technical Report CMU-CS-95-125.

Lee S. and Choi K. (1997) *Reestimation and best-first parsing algorithm for probabilistic dependency grammar.* In Proceeding of WVLC-5, 1997.

Maamouri Mohamed, Bies Ann, Buckwalter Tim, Diab Mona, Habash Nizar, Rambow Owen, and Tabessi Dalila (2006) *Developing and using a pilot dialectal Arabic tree- bank*, In Proceedings of the Fifth International Confer- ence on Language Resources and Evaluation, LREC'06, Genoa, Italy.

Pandey Rakesh, Dhami Hoshiyar, (2010), *Parsing of Kumauni Language Sentences after Modifying Earley's Algorithm*, Dialectologia ©Universitat de Barcelona, Vol 7 (2011), PP- 75-93

Patil V. B., Pawar B. V. (2013) *Influence of Karaka Relation in the framework of Marathi Link Grammar Parser*, In National Conference on Advances in Computing (NCAC"13), ISBN 978-81-910591-7-5, 2013, PP. 255-258.

Patil V. B., Pawar B. V. (2014), *Developing Links of Compound Sentences for Parsing Through Marathi Link Grammar Parser*, International Journal on Natural Language Computing (IJNLC) Vol. 3, No.5/6, PP 1-9

Poon Hoifung and Doming Pedro (2009) *Unsupervised semantic parsing*, EMNLP '09 Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1 PP 1-10.

Sleator D. K. & Temperley D. (1991) *Parsing English with a Link Grammar*, Technical Report CMU-CS-91-196, Carnegic Mellon University, School of Computer Science.