Introducing Kashmiri Dependency Treebank

Linguistic Data Consortium for Indian Languages, CIIL Mysore

Shahid Mushtaq Bhat

Shahid.bhat3@gmail.com

ABSTRACT

Treebank is a basic language resource for training and testing syntactic parser which forms a key module in various NLP systems like machine translation system. This paper reports an ongoing research of building dependency treebank for Kashmiri (KashTreeBank) and discusses some main annotation issues. The paper is based on the pilot annotation of 500 sentences.

KEYWORDS: Corpus, Annotation, Dependency, V2 Phenomenon

1. Introduction

Treebank is a set of corpora annotated with skeletal syntactic information, such as POS tags for words level and syntactic tags for beyond word level (Kristin Jacque, 2006). It is essentially a machine readable repository of syntactic structures of a language. KashTreeBank[1] is a small scale Kashmiri dependency treebank, based on Paninian Computational Grammar-PCG (Bharati, et al. 1994). Sanchay (See Singh 2006) has been used for syntactic annotations. The present treebank is in the initial stage of development in which annotation of 2361 sentences has been taken up. So far, POS annotation of 888 sentences has been completed, out

1 KashTreeBank is a personal effort on part of the researcher which was initially conceived as a summer school project in IASNLP 2011.

Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012), pages 53–60,
COLING 2012, Mumbai, December 2012.

of which 500 sentences are further annotated at chunk & inter-chunk dependency level, manually by using the afore mentioned interface.

2. Kashmiri and Resource Poor Scenario

'Kashmiri' is one of the 22 scheduled languages as per the 8th schedule of the Indian constitution. It is mainly spoken in the greater region of "Kashmir". There are 6 million Kashmiri speakers (see Ethnologue, 2006). Kashmiri is a Dardic language (Grierson, 1915), classified with its sister language Shina, under a separate group within Indo-Aryan family. It is inflectionally rich language with extensive V2 phenomenon & pronominal clitics. It is mainly written in modified Persio-Arabic script with writing convention from right to left. It is the only language in Dardic group which has a written tradition. Since, it has been never used as an official language or the medium of instruction; the text is produced in limited domains, predominantly, in literature. Very little computational and language resources are available for Kashmiri. It is only 3-4 years back that some initial efforts started in this direction at LDC-IL and University of Kashmir (ldcil.org & kashmirizaban.com). These efforts resulted in some basic language resources.

3. Corpus

The treebanks are usually based on contemporary newspaper texts, which have the practical advantage of being relatively easily accessible. For example, the Wall Street Journal part of the Penn Treebank (See Marcus et al. 1993). This has been very influential as a model for treebanks across the world but newspaper corpus is not readily available for Kashmiri as it is for English. However, for the current work, the use of newspaper corpus was avoided in initially for the only reason that it comprises of very lengthy sentences which were very hard to annotate in the beginning. Therefore, 500 sentences corpus was created by manually inputting the text from

the short stories. The idea was that these stories consist of relatively shorter sentences (4 to 30 Words in length) which would be easier to annotate, hence, to prepare a basic guideline would be an easy job. Later on, 1861 sentences from a newspaper "Sangarmaal" (political domain) were added to the existing corpus.

4. Framework and The Grammar Formalism

For the current work, PCG has been used for syntactic annotation which is actually a variant of dependency grammar (Kiparsky & Staal, 1969). This model helps to capture the syntacto-semantic relations in a sentence. Sentence is considered as a series of modifier-modified relations with a primary modified, main verb, the root of the dependency tree. The elements which modify main verb are its arguments or adjuncts that participate in the action specified by the verb. The relations of these participants with the verb are called karaka. Since, Kashmiri is an inflectionally rich language; there are clear cut markers or postpositions (vibaktis) on the arguments and adjuncts that participate in an action/event. Such syntactic cues can be very instrumental in identifying the relation of an argument or adjunct with its root, the main verb. There is almost one-to-one relation between the karakas & the case markers/postpositions in many constructions. Such correspondences between karaka & vebhakti along with TAM features are very helpful in syntactic annotation and handling relatively free word-order (Bharati et al. 1993).

5. Some Important Annotation Issues

Initially, Hindi annotation guideline was used to experiment with Kashmiri corpus and to frame a rough Kashmiri annotation guideline. Though, it covered most of the dependency relations occurring in the corpus but there were many remaining issues related to V2 phenomenon, discontinuity in complex predicates,

distinction between coordinating and subordinating conjuncts, pronominal clitics, handling of discourse elements, etc which were resolved after initial pilot annotation of 500 sentences. These issues are briefly explained below with the help of some representative dependency trees:

I. **V2 Phenomenon and Discontinuous Verb Group**

1. asi **A:s** doshvun' bA:tsan tam'-sInz seyThaa

   we **had** both husband-wife it's immense

   nikhath **gA:mIts**

   hatred **went**

   We both husband wife had developed immense hatred of it.

In this example, shown in Fig.1, tense auxiliary "A:s" (آس/FRAGP) and main verb "gAmIts" (**گَمِڑ**/VGF) are discontinuous (non-contiguous) with three intervening NPs. The auxiliary verb occurs mostly at second position and the main verb at sentence final position of the sentence. It is called V2 phenomenon which is similar to German and Yiddish. Since the root of the sentence is finite verb group, the main issue was whether to posit AUX or VM as the root of the sentence; given the discontinuity in finite verb group (VGF). Initially, FRAGP tag (used to handle discontinuity in Hindi treebank) was used for AUX and it was treated as a root of the sentence. The relation between FRAGP & VGF was shown by arbitrary FRAGof arch (see Fig.1) Later on, the decision to posit FRAGP (AUX) as head vis-à-vis root of sentence was revised and VGF (VM) was taken as head. This decision was made in consonance with the theory behind PCG which holds that only content words can be heads.

II. **Complex Predicates and their Discontinuity**

2. zA:hir chu ki Akis **aasi** akh kitaab **pasand**
   obvious is that one **will-have** one book **like**

   tI beykis **aasi** byaakh kitaab **pasand.**

   and other **will-have** other book **like**

It is obvious that one likes one book and other likes other book.

Identification of complex predicates (CP) and their extraction is already a complex problem in syntactic parsing but it is more complicated in Kashmiri where both discontinuous and continuous CPs occurs. The noun/adjective/verb part of CP occurs apart from the light verb which takes second position due to V2 phenomenon. While annotation, the noun/adjective/verb part is being attached to the light verb (VGF) with relation label, Part-of (Pof), as shown in Fig.3, the nominal part of the CP "pasand" (پَسنْد/NP) is attached to the light verb "aasi" (آسۍ/VGF) by (Pof) arch.

III. **Coordinating and Subordinating Conjuncts**

In shown in Fig.3, no distinction is maintained between coordinating and subordinating or embedding phenomena as both relations are represented by the (ccof) archs in the tree. However, it is important to maintain the distinction by positing separate relations (ccof for coordination & ccsof for subordination). Coordination conjunct (CCP) is taken as head of the sentence as shown in Fig.3 as well as Fig.2, joining the roots (verb groups) of the two independent clauses by (ccof) relations.

IV. **Pronominal Clitics**

3. Yami-is yi behtar zon-**un** ti thovn-**as**

   For-whom whatever better know-**3C** that kept-**3C**

lekhith.

written

For whom whatever he deemed better he kept that in his/her destiny.

Pronominal clitics are the characteristic feature of Kashmiri. Certain pronominal arguments are dropped from the argument structure but are gets cliticised as verbal inflection shown in Fig.2, the light verb lekhith thovn-as (لیکھتھ تھونَس) is inflected for second person pronominal clitic (–as) and completes the argument structure even without the presence of free pronominal arguments. So it is not important to posit dummy argument, instead a feature of clitic can be added to the light verb head.

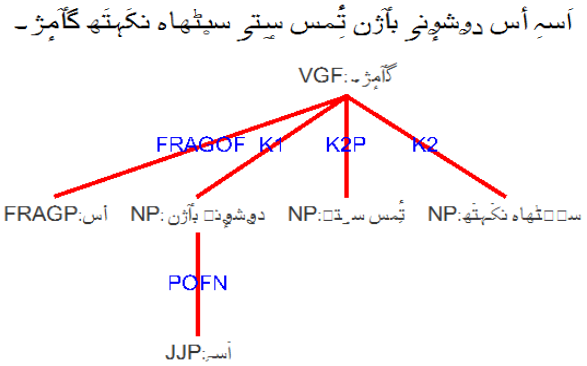

FIGURE.1 Simple Sentence showing V2 Phenomenon

یُمِس یہِ بہتَر زونُن تہِ تھونَس لیکھتھ ۔

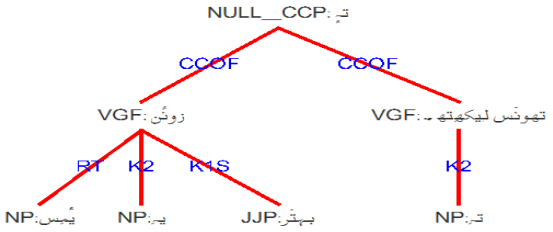

FIGURE.2 Sentence showing Pronominal Clitic

ظاہِر چُھ کہ أکس آسہٕ اَکھ کِتاب پَسنُد تہٕ بیکِس آسہِ بیاکھ کِتاب پَسند ۔

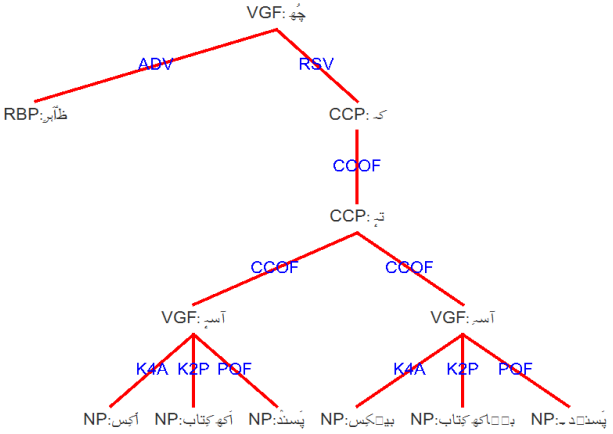

FIGURE.3 Sentence Showing Discontinuous Complex Predicate, Coordination and Sub-ordination

4. Conclusion

In this paper we explored some annotation issues raised during the pilot phase of annotation for *KashTreeBank*. We have used the already developed dependency annotation scheme of Hindi-Urdu to Kashmiri and found to a greater extent it could capture almost all the phenomenon of Kashmiri that occurred in 500 sentences except some phenomenon peculiar to Kashmiri as discussed above. The output of this work is a dependency annotation guideline for Kashmiri treebank. The guideline will be applied to the newspaper corpus, some part of which has been already annotated according with BIS POS annotation scheme. Since, POS annotation of the current 500 sentences has been carried out on the basis of ILMT scheme; the Kashmiri dependency guideline needs many changes before applying to the newspaper corpus in future.

References

Abeille, A./Clement, L./Toussenel, F. (2003). *Treebanks: Building and Using Parsed Corpora*. Dordrecht: Kluwer.

Bharati, Akshar, Chaitanya, Vineet and Sangal, Rajeev. (1995). *Natural Language Processing: A Paninian Perspective.* Prentice-Hall of India: New Delhi

Begum Rafiya, Husain, Samar, Dhwaj Arun, Sharma, Dipti Misra, Lakshmi Bai, Rajeev Sangal. (2008). Dependency Annotation Scheme for Indian languages. In *Proceedings of International Joint Conference on Natural Language Processing*.

Vempaty, Chaitanya, Naidu, Viswanatha, Husain, Samar, Kiran, Ravi, Bai, Lakshmi, Sharma, Dipti M., and Sangal, Rajeev. (2010). Issues in Analyzing Telugu Sentences towards Building a Telugu Treebank. In *Proceedings of CICLING* 2010.