Building a Hierarchical Annotated Corpus of Urdu: The URDU.KON-TB Treebank

Qaiser Abbas

University of Konstanz, Department of Linguistics,
Box D 185, 78457 Konstanz, Germany
qaiser.abbas@uni-konstanz.de
http://ling.uni-konstanz.de/pages/compling/

Abstract. This work aims at the development of a representative treebank for the South Asian language Urdu. Urdu is a comparatively under resourced language and the development of a reliable treebank for Urdu will have significant impact on the state-of-the-art for Urdu language processing. In URDU.KON-TB treebank described here, a POS tagset, a syntactic tagset and a functional tagset have been proposed. The construction of the treebank is based on an existing corpus of 19 million words for the Urdu language. Part of speech (POS) tagging and annotation of a selected set of sentences from different sub-domains of this corpus is in process manually and the work performed till to date is presented here. The hierarchical annotation scheme we adopted has a combination of a phrase structure (PS) and a hybrid dependency structure (HDS).

Keywords: Urdu, Treebank, POS, Phrase, Hybrid.

1 History and Introduction

The primary aim of this work is to build a treebank URDU.KON-TB. A treebank or parsed corpus is a text corpus of sentences, annotated with a syntactic structure (a tree structure), hence the name treebank. Similarly, corpus annotation is simply the process of the addition of interpretative linguistic information to a corpus, e.g. addition of tags/labels identifying the class of words in a text. This is so-called part of speech (POS) tagging [1]. A sample of a POS and syntactic annotation scheme is given in Figure 1 for the sentence as follows:

حامد نے شیر کو مارا.
Roman: Hamid ne sher ko mara.
English: Hamid killed the lion.

In this tree KP, PN, P, NN, VP and VB represents case phrase, proper noun, particle, noun, verb phrase and verb respectively. As Urdu is a case marking language, so a KP for case phrase is used and not CP. Similarly, particle P is used to tag case marker(CM) like *'ne'* and *'ko'* in above tree and not CM. These

A. Gelbukh (Ed.): CICLing 2012, Part I, LNCS 7181, pp. 66–79, 2012.

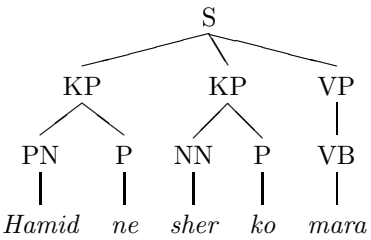


Fig. 1. Example tree for *"Hamid ne sher ko mara"*

issues of ambiguous tagging are discussed in [20]. Anyhow a phrase structure annotation scheme was adopted which has the advantages of simplicity, is easily convertable into bracketed sentences, 'light' on resources and the tree structure is relatively easy to read without specialized software tools. The leave nodes represent words and the nodes connected directly to leaves represent POS tags of the respective words. Other nodes up above the POS tags represent the syntactic annotation for a sentence S.

It is pertinent to note that the tree above is only for common understanding of human being. Computers can not process this tree annotation until the link between each node is converted into some computer readable form. So, bracketing of trees is needed here which will be described in design section. The tree does not include functional annotation information e.g. gramatical, semantic and thematic which is included in the current work of URDU.KON-TB treebank with meaningful POS and syntactic annotation.

POS tagging schemes are generally useful for differentiating words which have same spelling but different meanings, for example the word "present" may denote a noun *gift*, a verb to give someone a *present* or an adjective *not absent*. So, having a reliable and linguistically informed annotation scheme is very important. Moreover, there are many annotation schemes for corpora which include semantic (meanings of words), discourse (adding a information about anaphoric links), stylistic (adding information about speech and thought presentation), lexical (adding the identity of the lemma/base/stem of each word form in a text), etc [2]. Annotation can help automatic processing and analysis in many different ways. For example POS tagged corpora can generate frequency list or frequency dictionaries with grammatical classification e.g. the verb "leaves" and the noun "leaves" should be treated differently based on their frequency.

The newly developed 19 million word corpus by [3] available for the Urdu language is a huge corpus, but a balanced approach for selection of set of sentences from each domain of the corpus has been adopted. A tagset based on linguistics distinctions is developed. A portion of the Urdu corpus is first annotated with POS tags and then syntactic analysis is added on top of the POS annotation. The URDU.KON-TB treebank is enriched with sufficient semantic or other linguistic information. Our approach to treebank creation as per recommendation is completely manual at the moment and will be semi-automatically in future for speeding up the procedure on the whole 19 million

corpus. A detail of POS, syntactic, grammatical, semantic and thematic tagging is discussed in the design section, along with a sample bracketed sentence. A combination of a phrase structure (PS) and a hybrid dependency structure (HDS) has been adopted which is also useful for conversion into C-structure and F-structure. This outcome also allows linguistics community to use the resources of these languages in very effective way.

A simple but large difference between phrase and dependency annotation structure is that in phrase structure annotation, the nodes represents phrases/ constituents only e.g. KP, VP, etc as shown in Figure 1, while in dependency structure the nodes represents head words or head word plus its syntactic tag as shown in Figure 6(b). Some treebanks alongwith their annotation structure are as follows: the BulTreeBank[1] for the Bulgarian language follows HPSG (Head-driven Phrase Structure Grammar), the Penn Treebank[2] and ICE-GB[3] (International Corpus of the English-Great Britain) for English adopted the phrase structure scheme, the Prague Dependency Treebank[4] for the Czech language by Czech republic and the Quranic Arabic Dependency Treebank[5] for the Arabic language by University of Leeds, UK had adopted the dependency structure during treebank annotation. Almost 64 treebank exist for different languages in the world. At present, the languages for which more than three treebanks are available included Arabic, English, German, Italian, Latin and Japanese. A list of treebanks can be seen in [4]. The most popular treebank in the history was the Penn treebank for the English language in 1993. A short description of some treebanks is as follows.

The Penn Treebank for English: An annotated corpus containing 4.5 million words of the English language. In the first phase of this project, POS information [5] was encoded along with annotated syntactic structure (tree structure) in parallel with half of the corpus. A large number of work heavily relied on the Penn Treebank e.g. stochastic parsing [6,7,8], skeletal parsing [9,10], training POS taggers [11], disambiguating spoken sentences [12], linguistic theory and psychological modelling [13], grammar development etc. The Penn Treebank had limitations like no clear argument/adjunct relationship, trapping problem, inconsistencies in the annotation scheme, limited annotation and need of predicate-argument structure. Other Treebanks for English are the Susanne Corpus [14], the Lancaster Parsed Corpus [15], and International Corpus of English [16]. The German treebank TIGER [17] is based on the NEGRA treebank and the popular treebank for German is Tuba-D/Z. These treebanks (TIGER & Tuba-D/Z) contain 22000 and more than 50000 sentences respectively collected from German newspapers and also annotated with phrase and dependency structure

[1] http://www.bultreebank.org/ The project was funded by theVolkswagen Stiftung, Germany under the Programme "Cooperation with Natural and Engineering Scientists in Central and Eastern Europe".
[2] http://www.cis.upenn.edu/ treebank/
[3] http://www.ucl.ac.uk/english-usage/projects/ice-gb/index.htm
[4] http://ufal.mff.cuni.cz/pdt/
[5] http://corpus.quran.com/

حامد نے شیر کو افریقہ کے جنگل میں بندوق سے مارا ۔

Roman: Hamid ne sher ko Africa ke jungle
mein bandooq se mara.
English: Hamid killed the lion in the jungle
of Africa with the gun.

```
(S
    (KP (PN   حامد )   (P    نے ))
    (KP (NN   شیر )    (P    کو ))
    (PREP
        (NP
            (GP
                    (PN   افریقہ )
                    (P    کے ))
            (NN   جنگل ))
        (PRE  میں ))
    (SEP (NN بندوق )   (SE   سے ))
    (VP (VB   مارا ))
    (SM       . ))
```

Fig. 2. A sample bracketed sentence from NU-FAST Treebank

respectively. Each treebank used Stuttgart Tübingen POS Tagset along with 49 and 36 grammatical function labels respectively [18]. TIGER has a flat annotation scheme with no unary branching while the other one allows for this and contains a deeper hierarchical structure.

The selection of annotation scheme is totally dependent on the constituent ordering of the language. Phrase structure is good for fixed constituent order languages like English, Bulgarian, Chinese, etc., while dependency structure is good for free constituent order languages like Urdu, Hindi, German, etc. [19]. The current work builds on a previous study at constructing a NU-FAST treebank [20]. However, the design of that treebank proved to be too simple and flat. It neither contained detailed syntactic, morphological, semantic and thematic information, nor any information about displaced constituents/phrases, empty arguments or traces. In addition, it was based on a POS-Tagset that is currently revised by the author due to issues like the word *'ne'* in Figure 1, which should be tagged as case marker CM instead of particle P [21]. A bracketed sentence from the NU-FAST Treebank is given in Figure 2.

Another Hindi-Urdu treebanking effort is under way in a collaborative project[6] between five different universities at Colorado. However, the Urdu treebank being developed is comparatively small and is being done as part of a larger effort at establishing a treebank for Hindi. Although Urdu and Hindi share many structural features, there are some interesting differences and as the main effort of the Hyderabad-Colorado cooperation is focused on Hindi, many of the issues with

[6] http://verbs.colorado.edu/hindiurdu/

respect to Urdu are remaining unresolved. Our work take up these issues and proposing solutions for them with passage of time.

2 Design

The first objective achieved of this URDU.KON-TB treebank, is the analysis of the corpus. This 19 million word corpus is available at Centre for Language Engineering (CLE)[7] . This corpus is in a very balanced and normalized form already. The corpus has six different domains which are from C1 to C6. The corpus does not yet include ethical and religious domain C7(shown in italics), which is ready after sufficient inspection and investigation but not yet merged. The domains and sub-domains of the corpus with size distribution and number of distinct words are given in Table 1. For treebank work, Samples of 200 sentences from each domain are selected, which comes up with 1400 sentences. The work of manual annotation with PS & HDS has been completed for domains C1 to C3 and information extracted till to date is presented in this paper.

The second objective of this work achieved is to study and investigate encoding & annotation scheme which is a combination of PS & HDS annotation and find useful for the Urdu language. Since, Urdu is a free word order plus case marked language and head word can not be found in sequence like in English most of the time. Hence, HDS annotation solve the problem here alongwith the PS. As for as the encoding scheme related to POS tagging is concerned, an existing POS tagger [22] is being used to tag the words in a sentence computationally just to speed up the process and then these tags are manually corrected & updated during its annotation process. The existing tagger only has very basic functionality with a limited tagset. It was therefore decided by our integrated team (KN + UET)[8] that a new POS tagset should be developed. This linguistically motivated POS tagset is almost in existence after the completion of C1 to C3 domain's sentences. A sample of the newly developed POS tagset is given in Table 2 and dot '.' is used to add multiple subcategories in a main category e.g. V.LIGHT.PERF, which is a verb V as main category having light LIGHT and perfective PERF concepts as subcategories, concluding hierarchical structure.

The tagset in Table 2 represents a complete POS tagset extracted from the manually annotated sentences of domains C1 to C3 of the corpus. For example, adjectives are being dealt with in four ways. A general category of adjectives as ADJ, a manner category of adjectives as ADJ.MNR, a spatial category of adjectives as ADJ.SPT and a temporal category of adjectives as ADJ.TMP. The Relevant examples are provided in Figure 3.

The addition of this .SPT and .TMP after the syntactic tag ADJ for adjective represents spatial and temporal adjectives respectively, however this '.'

[7] CLE, University of Engineering and Technology (UET), Lahore, Pakistan (http://www.cle.org.pk).

[8] KN + UET is a team of University of Konstanz, DE and University of Engineering & Technology, PK.

Table 1. Existing Corpus at CLE

Domains	Sub Domains
C1. Sports/Games	C1.1.Sports (special events)
C2. News	C2.1. Local and international affairs C2.2. Editorials and opinions
C3. Finance	C3.1. Business, domestic and foreign market
C4. Culture/Entertainment	C4.1. Music, theatre, exhibitions, review articles on literature C4.2. Travel / tourism
C5. Consumer Information	C5.1. Health C5.2. Popular science C5.3. Consumer technology
C6. Personal communications	C6.1. Emails, online discussions, editorials,e-zines
C7. Ethics and Religious	C7.1. History, Online discussion, Preaching literature, e-magzines

Domains	Size	Distinct words
C1. Sports/Games	1,666,304	23,118
C2. News	8,957,259	67,365
C3. Finance	1,162,019	17,024
C4. Culture/Entertainment	3,845,117	59,214
C5. Consumer Information	1,980,723	34,151
C6. Personal communications	1,685,424	30,469
C7. Ethics and Religious	2,756,695	28,170
Total	22,053,541	132,511

dot notation can also be used for morphological purposes e.g. for the continuous/progressive verb form, a V.PROG tag is being used. Similarly, adverbs which are mostly used as a qualifier of verbs can also be used independently. Adverbs are categorized into six forms presented in Table 2. Some examples of adverbs are given in Figure 4. The examples quoted above and below are to give an idea how POS encoding/tagging has been assigned to given words in a sentence of a corpus.

In URDU.KON-TB treebank, an intermediate approach of PS and HDS has been adopted. In this annotation scheme, the PS approach is implemented on an outer level (physical) and the HDS approach is implemented at an inner level (logical). For example, in Figure 5, bold formated noun phrase NP is extracted from the hidden concept of noun lying among case phrase KP, noun N, coordination conjunction C.CORD and coordination phrase CP at the same level inside the NP. This is basically logical concept of HDS adopted, while the physical annotation without head words as in Figure 5 is PS. Due to this combination, it will also be very easy for linguists to obtain additionally the XLE[9]

[9] XLE is a rule based parser introduced by Xerox and Palo Alto Research Center (PARC) in 1993.

Table 2. The URDU.KON-TB Part of Speech Tagset

- ADJ (Adjective)
  - ADJ.MNR (Manner …)
  - ADJ.TMP (Temporal …)
  - ADJ.SPT (Spatial …)
- ADV (Adverb)
  - ADV.DEG (Degree …)
  - ADV.MNR (Manner …)
  - ADV.NEG (Negative …)
  - ADV.SPT (Spatial …)
  - ADV.TMP (Temporal …)
- C.CAUS (Causative Conjunction)
- C.CORD (Coordination …)
- C.SBORD (Subordinate …)
- CM (Case Marker)
- DATE.CAL (Calendar Date)
- DATE.Y (Year Date)
  - DATE.Y.CAL (Calendar …)
- IZF (Izafe)
- KER (Ker)
- N (Noun)
  - N.ADJ (Adjectival …)
  - N.CURR (Currency …)
  - N.PROP (Proper …)
    - N.PROP.DATE.M (Month…)
    - N.PROP.SPT (Spatial …)
  - N.SPT (Spatial …)
  - N.TMP (Temporal …)
- P (Pronoun)
  - P.DEM (Demonstrative …)
  - P.INDF (Indefinite …)
  - P.PERS (Personal …)
  - P.REF (Reflexive …)
    - P.REF.POSS (Possessive…)
  - P.REL (Relative …)
    - P.REL.DEM (…)
    - P.REL.PERS (Personal …)
- POSTP (Post Position)
  - POSTP.SPT (Spatial …)
  - POSTP.TMP (Temporal …)
- PREP (Pre Position)
- PT (Particle)
- Q (Quantifier)
  - Q.CARD (Cardinal …)
  - Q.FRAC (Factional …)
  - Q.ORD (Ordinal …)
- QW (Question Word)
- SM (Sentence Marker)
- U (Unit)
- V.IMPERF (Imperfective …)
- V.INF (Infinitive Verb)
  - V.INF.OBL (Oblique …)
- V.LIGHT (Light Verb)
  - V.LIGHT.IMPERF (…)
  - V.LIGHT.INF (Infinitive …)
  - V.LIGHT.PERF (Perfective …)
  - V.LIGHT.ROOT (Root …)
- V.LIGHTV.PERF (… Light-Verb …)
- V.MOD (Modal Verb)
- V.PASS.INF (Infinitive Passive …)
- V.PASS.LIGHTV.IMPERF (…)
- V.PERF (Perfective Verb)
- V.ROOT (Root Verb)
  - V.ROOT.LIGHT (Light …)
  - V.ROOT.LIGHTV (…)
  - V.ROOT.PERF (Perfective …)
- V.TB.PERF (Perfective To-Be Verb)
- VALA (Vala)
- VAUX.COP (Copula Verb-Auxiliary)
  - VAUX.COP.IMPERF (…)
  - VAUX.COP.PRES (Present …)
  - VAUX.COP.ROOT (Root …)
- VAUX.IMPERF (Imperfective …)
- VAUX.LIGHT (Light Verb-Auxiliary)
- VAUX.LIGHTV.IMPERF (…)
  - VAUX.LIGHTV.PERF (…)
- VAUX.MOD (Modal Verb-Auxiliary)
  - VAUX.MOD.IMPERF (…)
  - VAUX.MOD.PERF (…)
- VAUX.PASS.IMPERF (…)
- VAUX.PASS.PERF (… Passive …)
- VAUX.PASS.ROOT (Root …)
- VAUX.PERF (… Verb-Auxiliary)
- VAUX.PROG (Progressive …)
  - VAUX.PROG.PERF (…)
- VAUX.REP.HABT (Habitual …)
  - VAUX.REP.HABT.IMPERF (…)
  - VAUX.REP.HABT.LIGHT (…)
- VAUX.TB.IMPERF (… To-Be …)
- VAUX.TENS (Tense Verb-Auxiliary)
  - VAUX.TENS.COP (Copula …)
  - VAUX.TENS.LIGHT (…)

1. اچھا لڑکا (*Good boy*)
   Acha larka
   ADJ N
   Here the word *Acha* 'good' is used as an adjective
2. ملتانی کھسہ (*Multani shoe*)
   Multani khussah
   ADJ.SPT N
   Multani 'city name' is as a spatial adjective ADJ.SPT
3. گذشتہ سال (*Previous year*)
   Guzashta sal
   ADJ.TMP N
   Guzashta 'previous' is as a temporal adjective ADJ.TMP
4. جابرانہ حکومت (Cruel government)
   Jaberana hakoomat
   ADJ.MNR N
   Jaberana 'cruel' is manner adjective ADJ.MNR

Fig. 3. Examples of Adjective

1. تقریبا ساری دنیا میں (*Almost in the whole world*)
   Taqreeban sari dunia mein
   ADV Q N.SPT CM
   Taqreeban 'Almost' is used as an adverb ADV
2. عمارت مکمل نہ ہو سکی ۔ (*The building could not be completed.*)
   Emarat mukamal na ho saki .
   N ADJ ADV.NEG V.LIGHT.ROOT V.MOD SM
   na 'not' here is used as a negative adverb ADV.NEG
3. بہت اچھی لڑکی (*Very good girl*)
   Bohat achi larki
   ADV.DEG ADJ N
   bohat 'very' is used as a degree adverb ADV.DEG

Fig. 4. Examples of Adverbs

(Xerox Linguistic Environment) parser like C-structure/phrase structure and F-structure/dependency structure from this treebank. . A phrase/constituent structure annotation example is already depicted in Figure 1 and can also be seen in Figure 2. Bracketed notation in Figure 2 is equivalent to the C-structure of XLE parser. However, the dependency structure is distinct from phrase structure annotation as discussed earlier. Dependency structure is not dependent on a specific word order and hence well suited to free word order Urdu language. Mostly, dependency structure is presented in arrows pointing to head word from its dependent words or vice versa as shown in Figure 6(a). However, it can also be

اخبارات مسلمانوں کی لاشوں اور چیخ و پکار کی تصویروں
سے بھرے ہوتے ہیں ۔
Roman: Akh-barat musalmano ki lashon aur cheekh o pukar ki tasweeron se bhare hote hein.
English: The newspapers are used to full of pictures of Muslim's corpses and cries.

```
(S
   (NP.NOM-SUB
      (N اخبارات ))
   (KP.INST-OBL
     (NP
       (KP.POSS
         (NP
           (KP.POSS
             (N مسلمانوں) (CM کی))
           (N لاشوں )(C.CORD اور)
           (CP
             (N چیخ)(C.CORD و)(N پکار)))
         (CM کی))
       (N تصویروں))
     (CM سے))
   (VCMAIN
     (V.PERF بھرے )(VAUX.TB.IMPERF ہوتے)
     (VAUX.TENS ہیں ))
   ( SM . ))
```

Fig. 5. Physical and logical concept

presented in tree form but then the main verb should be used as root (predicate) of the sentence as in Figure 6(b).

A physical concept of phrase structure and logical concept of dependency structure is merged in the current development of URDU.KON-TB treebank. Moreover, normally dependency structure is limited to the word level e.g. concluding head word from given words, but we enhanced and implemented this idea further at constituent/phrase level which is named as hybrid dependency structure. It means concluding head constituent from given constituents i.e. concluding relationship among head constituent and dependent constituents (hybrid dependency). This whole scheme is named as combination of phrase structure and hybrid dependency structure. The need of such type of schemes is highly advocated in literature such as [19,23], etc. To build such schemes, it is very important that you must have some POS, syntactic and functional tagset. The syntactic tagset is shown in Table 3.

The bracketing form of Figure 6(c) can also be represented in the following bracketing form with some addition of functional tags which we have built in our

Table 3. The URDU.KON-TB Syntactic Tagset.

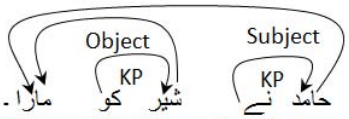
Tags	Description
ADJP	Adjective Phrase
ADVP	Adverb Phrase
CL.KER	Ker Clause
CP	Conjunction Phrase
DATEP	Date Phrase
KP	Case Phrase
KP.ACC	Accusative KP
KP.DAT	Dative KP
KP.ERG	Ergative KP
KP.INST	Instrumental KP
KP.POSS	Possessive KP
NP	Noun Phrase
NP.ACC	Accusative NP
NP.DAT	Dative NP
NP.ERG	Ergative NP
NP.NOM	Nominative NP
NP.OBL	Oblique NP
PP	Pre/Post Position Phrase
QP	Quantifier Phrase
S	Sentence
SBAR	Subordinate Clause
SBARQ	Question Subordinate Clause
SQ	Yes/No Question Sentence and Subconstituent of SBARQ
UP	Unit Phrase
VALAP	Vala Phrase
VALAP.NOM	Nominative VALAP
VCMAIN	Verb Complex Main
VCP	Verb Complex Predicate
VIP	Verb Infinitive Phrase

URDU.KON-TB treebank e.g. PRD, SUB, OBJ, etc.. This bracket form is very close to F-structure of XLE parser. Moreover, the functional tagset developed for the URDU.KON-TB treebank is given in Table 4.

```
[ PRED mara
          [ SUBJ Hamid
                   [ CM ne] ]
          [ OBJ sher
                   [ CM ko] ] ]
```

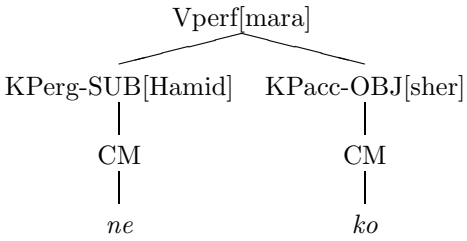
A sample of a bracketed sentence using POS, syntactical and functional tagging can be seen in Figure 7 and compare this tagging and encoded information to the existing NU-FAST treebank output given in Figure 2 earlier. This new output in Figure 7 is the achievement of our third objective of the project which

(a).



Roman: Hamid-ne sher-ko mara.
English: H amid killed the lion.

(b).



(c).

[V.PERF mara [KP.ERG-SUB Hamid [CM ne]] [KP.ACC-OBJ sher [CM ko]]]

Fig. 6. (a). Dependency arrow (b). Dependency tree (c). Dependency bracket form

Table 4. The URDU.KON-TB Functional Tagset

Grammatical	**Semantic/Thematic**
-PRD (Predlink)	-DEG (Degree)
-OBJ (Direct Object)	-SPT (Spatial)
-SUB (Subject)	-TMP (Temporal)
-OBJ2 (Indirect Object)	-MNR (Manner)
-OBL (Oblique)	-CMP (Comparative)
	-INST (Instrumental)
Miscelleneous	
* (for Empty Categories)	
-L (for linking displaced consititudents/categories)	
-1/-2/-3.... (for labelling of displaced constituents/categories)	

enables us to produce a linguistically enriched treebank. The symbol '//' (double slashes) is used to explain the terms in form of comments. In URUD.KON-TB treebank, the structures are identified and approach of understanding & labelling words is purely realistic and humanistic.

حامد نے شیر کو افریقہ کے جنگل میں بندوق سے مارا ۔
Hamid Ne Sher Ko Africa Ke Jangle Mein Bandooq Se Mara .
Hamid killed the lion in the jungle of Africa with the gun.

```
( S      //root S (SYNTACTICAL)
   ( KP.ERG-SUB //Ergative Case Phrase (SYNATICAL), and Subject (GRAMATICAL)
      ( N.PROP    حامد )  ( CM  نے ) )    //proper noun (POS)    //case marker(POS)
   ( KP.ACC-OBJ //Accusative Case  phrase (SYNATICAL), and Object  (GRAMATICAL)
      ( N     شیر )   ( CM  کو ) )              //noun (POS)        //case marker (POS)
   ( KP-SPT  //Case Phrase (SYNATICAL) and spatial (THEMATIC)
      ( NP   //Noun Phrase (SYNATICAL)
            ( KP.POSS  //possessive case phrase (SYNATICAL)
                  ( N.PROP      افریقہ ) ( CM    کے )) // proper noun (POS) //case marker (POS)
            ( N.SPT       جنگل ) ) //spatial noun (POS)
       ( CM  میں ) ) //case marker (POS)
   ( KP.INST            //Instrumental Case Phrase(SYNATICAL)
      ( N     بندوق ) ( CM  سے ) )       // noun (POS)    //case marker (POS)
   ( VCMAIN   //Verb Complex Main or Verb Phrase (SYNTACTICAL)
      ( V.PERF      مارا ) )      //perfective verb (MORPHOLOGICAL & SYNTACTICAL)
   ( SM       . ) )//sentence marker (SYNATICAL)
```

Fig. 7. A bracketing sentence from URDU.KON-TB tree bank with linguistically encoded information

3 Conclusion

A limited sized standard URDU.KON-TB treebank for the Urdu language is the main output yet after the first phase of this work. However, additional resources developed till to date contained a standard POS tagset, a syntactic tagset, a functional tagset and new tagged corpus. These resources will be enhanced further as the work progress. These resources can all be used for natural language processing (NLP) such as probabilistic parsing, training of POS taggers, disambiguation of spoken sentences, grammar development [24], language identification [25], sources for linguistic inquiry and psychological modelling, pattern matching, and in many applications of NLP and machine learning domains [26] & [27].

Acknowledgments. The author would like to express his gratitude to Prof. Dr. Miriam Butt, University of Konstanz for her encouragement, guidance and support. I am also indebted to the members of our integrated team in the various stages of this being continued research project.

References

1. Leech, G.: Adding linguistic annotation. In: Wynne, M. (ed.) Developing Linguistic Corpora: A Guide to Good Practice, ch. 3, pp. 17–29. Oxbow Books, Oxford (2005)

2. Garside, R., Leech, G.N., McEnery, T.: Corpus annotation: linguistic information from computer text corpora. Longman, London (1997)
3. Ijaz, M.: Urdu 5000 Most Frequently Used Words: Technical Report, Center for Research in Urdu Language Processing (CRULP), Lahore, Pakistan (2007)
4. Wallis, S.: Searching treebanks and other structured corpora. In: Lüdeling, A., Kytö, M. (eds.) Corpus Linguistics: An International Handbook. Handbücher zur Sprache und Kommunikationswissenschaft, ch. 34. Mouton de Gruyter, Berlin (2008)
5. Santorini, B.: Part-of-speech tagging guidelines for the Penn treebank project: Technical report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania (1990)
6. Brill, E.: Discovering the lexical features of a language. In: 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA (1991)
7. Brill, E., Magerman, D., Marcus, M.P., Santorini, B.: Deducing linguistic structure from the statistics of large corpora. In: DARPA Speech and Natural Language Workshop (1990)
8. Magerman, D., Marcus, M.P.: Parsing a natural language using mutual information statistics. In: AAAI (1990)
9. Pereira, F., Schabes, F.: Inside-outside re-estimation from partially bracketed corpora. In: 30th Annual Meeting of the Association for Computational Linguistics (1992)
10. Weischedel, R., Ayuso, D., Bobrow, R., Boisen, S., Ingria, R., Palmucci, J.: Partial parsing: a report of work in progress. In: 4th DARPA Speech and Natural Language Workshop (1991)
11. Meteer, M., Schwartz, R., Weischedel, R.: Studies in part of speech labelling. In: 4th DARPA Speech and Natural Language Workshop (1991)
12. Veilleux, M.N., Ostendorf, M.: Probabilistic parse scoring based on prosodic features. In: 5th DARPA Speech and Natural Language Workshop (1992)
13. Niv, M.: Syntactic disambiguation. The Penn Review of Linguistics 14, 120–126 (1991)
14. Sampson, G.: English for the computer: The SUSANNE corpus and analytic scheme. Clarendon Press, Oxford (1995)
15. Leech, G.: The Lancaster Parsed Corpus. ICAME Journal 16(124) (1992)
16. Greenbaum, S.: Comparing English worldwide: The International Corpus of English. Clarendon Press, Oxford (1996)
17. Dipper, S., Brants, T., Lezius, W., Plaehn, O., Smith, G.: The TIGER Treebank. In: Third Workshop on Linguistically Interpreted Corpora LINC 2001, Leuven, Belgium (2001)
18. Schiller, A., Teufel, S., Stoeckert, C.: Vorlaeufige Guidelines fuer das Tagging deutscher Textcorpora mit STTS(Deutsche): Technical Report, IMS-CL, University Stuttgart (1995)
19. Skut, W., Krenn, B., Brants, T., Uszkoreit, H.: An Annotation Scheme for Free Word Order Languages. In: Fifth Conference on Applied Natural Language Processing (ANLP), Washington, D.C (1997)
20. Abbas, Q., Karamat, N., Niazi, S.: Development of Tree-bank based probabilistic grammar for Urdu Language. International Journal of Electrical & Computer Science 09(09), 231–235 (2009) ISSN: 2077-1231
21. Butt, M., King, T.H.: The Status of Case. In: Dayal, V., Mahajan, A. (eds.) Clause Structure in South Asian Languages, pp. 153–198. Springer, Berlin (2005)

22. Sajjad, H., Schmid, H.: Tagging Urdu Text with Parts of Speech: A Tagger Comparison. In: 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2009 (2009)
23. Clark, A., Fox, C., Lappin, S.: The Handbook of Computational Linguistics and Natural Language Processing. Blackwell Handbooks in Linguistics, vol. 52, pp. 239–244. John Wiley and Sons (2010) ISBN: 1405155817, 9781405155816
24. Abbas, Q., Khan, A.H.: Lexical functional grammar for Urdu modal verbs. In: 5th IEEE (ICET) 2009 International Conference on Engineering and Technology, pp. 07–12 (2009)
25. Abbas, Q., Ahmed, M.S., Niazi, S.: Language Identifier for Languages of Pakistan Including Arabic and Persian. International Journal of Computational Linguistics (IJCL) 01(03), 27–35 (2010) ISSN: 2180-1266
26. Marcus, M.P., Santorini, B., Marcinkiewicz, M.A.: Building a large annotated corpus of English. Computational Linguistics (CL) 19(2), 313–330 (1993)
27. Bies, A., Ferguson, M., Katz, K., Macintyre, R.: Bracketing guidelines for Treebank II style penn treebank project: Technical Report, University of Pennsylvania (1995)