Urdu Treebank: (full corpus) v 1.0 (POS + Morph Analysis + Dependency Annotation)

Item Name:	Urdu Treebank: (full corpus) v 1.0 (POS + Morph Analysis + Dependency annotation)
Author(s):	Language Technologies Research Centre, IIIT-Hyderabad, India
Project No.	LTRC-TBIL-MCIT-66
Release Date:	August, 2016
Member Year(s):	2016
DCMI Type(s):	Text
Project(s):	Hindi-Urdu Treebank (HUTB) project
Application(s):	Information extraction, cross-lingual information retrieval, automatic content extraction, Natural Language Processing, Machine translation
Language(s):	Standard Urdu
License(s):	Creative Commons License Attribution-NonCommercial-ShareAlike 4.0 International.
Online Documentation:	ltrc.iiit.ac.in
Citation:	

1. Introduction

This file contains documentation on the Urdu Treebank: (full corpus) v 1.0 (POS + Morph Analysis + Dependency annotation) with Project No. LTRC-TBIL-MCIT-66.

The goal of the Urdu Treebank is to support the development of data-driven approaches and other natural language processing (NLP) applications, human language technologies, automatic content extraction (topic extraction and/or grammar extraction), cross-lingual information retrieval, information detection, and other forms of linguistic research on Modern Standard Urdu in general.

It is the child of the parent Hindi-Urdu Treebank (HUTB) project which is a collaborative effort of five universities in two countries:

University of Colorado Boulder
Columbia University
University of Massachusetts at Amherst (UMass)
University of Washington (UW)
International Institute of Information Technology (IIIT) in Hyderabad, India.

The overall objective of the Hindi-Urdu Treebank (HUTB) project is to build a multi-representational and multi-layered Treebank for Urdu and Hindi. In this release, we provide both syntactic (Treebank) annotation and annotation on part of speech (POS), chunking, Morph analysis and Dependency annotations for Urdu.

The Urdu Treebank, started in 2011 with the objective of annotating via human intervention is a large Urdu machine-readable text corpus of approx. 200,000 words. It was started with a view to build a multi-layered Treebank that will provide both syntactic and semantic annotations.

The development of the Treebank started in 2011 with raw sentences taken from news articles. The Urdu Treebank is developed following a generic pipeline.

The steps in the process of building the Urdu Treebank under this pipeline consists of

(i) Tokenization, (ii) Morph- Analysis, (iii) POS-tagging, (iv) Chunking, and (v) Dependency annotation. (Dependency Annotation is based on Paninian Grammar Framework.)

Annotation process commences with the tokenization of raw sentences. The tokens thus obtained are annotated with morphological and POS information. After morph-analysis and POS-tagging, words are grouped into chunks. All the above processing steps have been automated by high accuracy tools (rule-based or statistical) thus speeding up the manual process. The last process in this pipeline so far is the manual dependency annotation. The inter-chunk dependencies are marked leaving the dependencies between words in a chunk unspecified for the intra-chunk dependencies.

PropBanking is the next step in this generic pipeline which is aimed at establishing another layer of semantics on the Urdu Treebank. The Urdu Dependency Treebank is developed following this Treebanking pipeline for the newspaper articles using a team of expert linguistics annotators.

The tool used for the annotation is Sanchay (Singh and Ambati, 2010). All the annotations are represented in Shakti Standard Format (SSF). So far, ~7,000 sentences (around 200K words) have been annotated with dependency structure. Each sentence contains an average of 29 words and an average of 13.7 chunks of average length 2.0.

2. Tag-sets used:

* POS:

We have used the following POS tag-set to annotate POS information on the UTB.

Sl No.	***Category***	***Tag name***	***Example***
1.1	Noun اسم	NN	شہر،نام،کتاب،آب، گھر
1.2	NLoc	NST	آگے، پیچھے،نیچے،اندر، باہر
2.	Proper Noun	NNP	دہلی، حےدرآباد،لال قلعہ،محمد
3.1	Pronoun	PRP	مےں، وہ، تم،آپ
3.2	Demonstrative	DEM	یہ، وہ
4	Verb-finite	VM	جانا، کھانا، پینا
5	Verb Aux	VAUX	رہا، ہوئے،گا
6	Adjective	JJ	کمزور، کالا،ناتواں
7	Adverb	RB	*Only manner adverb یقیناً، فی الحال
8	Post position	PSP	میں،نے، سے، پر، تک، کو
9	Particles	RP	بھی،ہی، جی، صاحب، تو

10	Conjuncts	CC	اور، تو، چاہے لےکن
11	Question Words	WQ	کیا، کیوں، کےسے
12.1	Quantifiers	QF	بہت، کم
12.2	Cardinal	QC	1,2,3,67, 78,100,10000,111,
12.3	Ordinal	QO	پہا، دوسرا
12.4	Classifier	CL	عدد، نفر
13	Intensifier	INTF	ہی،بہت
14	Interjection	INJ	ارے،اوہ
15	Negation	NEG	نہیں،نہ
16	Quotative	UT	
17	Sym	SYM	- ، ') (} [
18	Compounds	*C	مغلِ اعظم، دردِ دل، دردِ جگر
19	Reduplicative	RDP	دوڑتے دوڑتے، کھاتے کھاتے
20	Echo	ECH	چائے،وائے،کھانا وانا
21	Unknown	UNK	انگرےزی لفظ، یا کسی اور زبان کا لفظ جو نا معلوم ہو

* Chunk Tag Set for Urdu:

Following Chunk tag-set is being used to annotate chunk/phrase information on the UTB.

Sl. No	*Chunk Type*	*Tag Name*	*Example*
1	Noun Chunk	NP	*((مےرانیاگھر))*_**NP** *“my new house”*
2.1	Finite Verb Chunk	VGF	مےں نے گھر پر ((کھایا_VM*))*_**VGF**
2.2	Non-finite Verb Chunk	VGNF	*mAin* میں *((**chAlte – chAlte**_**VM**))*_**VGNF** *gir gayA.* (گرگیا)
2.4	Verb Chunk (Gerund)	VGNN	*mujhe rAta meM ((**nAhanA**_**VM**))_**VGNN** acchA lagatA hai.* مجھے رات میں نہانا اچھا لگتا ہے
3	Adjectival CHunk	JJP	*nAdiyA ((khubsurAx_**JJ))_JJP** hE.* ندیا خوبصورت ہے
4	Adverb Chunk	RBP	*vaha* ((*dhIre-dhIre*_**RB**))_**RBP** *cala rahA thA.*وہ دھیرے دھیرے چل رہا تھا
5	Chunk for Negatives	NEGP	**((*binA*))_NEGP** ((*kucha*))_NP **((*bole*))_VG** ((*kAma*))_NP ((*nahIM calatA*))**_VG.** **بنا کچھ بولے کام نہیں چلتا**
6	Conjuncts	CCP	((*sAhid*))_NP ***((Ora))*_CCP** ((*hAmid*))_NP. **شاہد اور حامد**
7	Chunk Fragments	FRAGP	***sAhid*** (*jo merA baDZA bhAI hE*) ***ne*** *kahA*
8	Miscellaneous	BLK	

Sl. No	*Chunk Type*	*Tag Name*	*Example*

* Morph analysis:

Urdu Treebank contains Morph analysis at token level for the following:

1. Category

2. Lexical category

3. Gender

4. Number

5. Person

6. Case

7. Vibhakti/TAM

* Dependency labels used for Urdu :

S.No	*Labels*	*Description*	*Gloss*
1	k1	karta	doer/agent/subject
2	pk1, jk1, mk1		causer, causee, mediator-causer
3	k1s	vidheya karta - karta samanadhikarana	noun complement of karta
4	k2	karma	object/patient
5	k2p		Goal, Destination
6	k2g		secondary karma
7	k2s	karma samanadhikarana	object complement
8	K3	karana	instrument
9	k4	sampradana	recipient
10	k4a	anubhava karta	Experiencer
11	k5	apadana	source
12	K5prk	prakruti apadana	source material
13	k7t	kAlAdhikarana	location in time

14	k7p	deshadhikarana	location in space
15	k7	vishayadhikarana	location elsewhere
16	k7a		according to
17	k*u	sAdrishya	similarity/comparison
18	r6	shashthi	genitive/possessive
19	r6-k1, r6-k2		karta or karma of a conjunct verb (complex predicate)
20	r6v	kA	relation between a noun and a verb
21	adv	kriyAvisheSaNa	adverbs - ONLY 'manner adverbs' have to be taken here
22	Sent-adv		Sentential Adverbs
23	rd	relation prati	direction
24	rh	hetu	reason
25	rt	Tadarthya	purpose
26	ras-k*	upapada_ sahakArakatwa	associative
27	ras-neg		Negation in Associatives
28	rs	relation samanadhikaran	noun elaboration
29	rsp		relation for duratives
30	rad		address terms
31	nmod__relc, jjmod__relc, rbmod__relc		relative clauses, jo-vo constructions
32	Nmod		participles etc modifying nouns
33	vmod		verb modifier
34	jjmod		modifiers of the adjectives
35	pof		part of units such as conjunct verbs
36	ccof		co-ordination and sub-ordination
37	fragof		Fragment of
38	Enm		enumerator
39	rsym		a symbol

40	nmod__emph		nmod__emph
41	psp__cl		

3. Data

This release contains approx. 200,000 source tokens.

The corpus is released as SSF and CONLL format in UTF-8. It contains the Inter-chunk dependencies and Intra-chunk expanded data. For further information, kindly refer to README.

4. Samples

**** A sample raw Urdu sentence:***

ریاست میں انتخابی مہم کا بدتریج آغاز ہو رہا ہے

(In state) (of elections) (gradually)(commence)(is)

The gradual commencement of elections is happening in the state.

**** Sentence in SSF format:***

```
<Sentence id='1'>
1	((	NP	<fs name='NP' drel='k7p:VGF'>
1.1	ریاست	NN	<fs af='ریاست,n,f,sg,3,o,0,0' posn='10' name='ریاست'>
1.2	میں	PSP	<fs af='میں,psp,,,,,,' posn='20' name='میں'>
	))
2	((	NP	<fs name='NP2' drel='r6-k1:NP3'>
2.1	انتخابی	JJ	<fs af='انتخابی,adj,any,any,,o,,' posn='30' name='انتخابی'>
2.2	مہم	NN	<fs af='مہم,n,f,sg,3,d,0,0' posn='40' name='مہم'>
2.3	کا	PSP	<fs af='کا,psp,m,sg,,d,,' posn='50' name='کا'>
	))
3	((	RBP	<fs name='RBP' drel='adv:VGF'>
3.1	بتدریج	RB	<fs af='بتدریج,adv,any,any,,d,,' posn='60' name='بتدریج'>
	))
4	((	NP	<fs name='NP3' drel='pof:VGF'>
4.1     آغاز    NN      <fs af='آغاز,n,m,sg,3,d,0,0' posn='70' name='آغاز'>
        ))
5       ((      VGF     <fs name='VGF' stype='declarative' voicetype='active'>
5.1     ہو      VM      <fs af='ہو,v,any,any,any,,0,0' posn='80' name='ہو'>
5.2     رہا     VAUX    <fs af="رہ,v,m,sg,any,,یا,yA' posn='90' name='رہا'>
5.3     ہے      VAUX    <fs af='ہے,v,any,sg,3,,ہے,hE' posn='100' name='ہے'>
5.4     -       SYM     <fs af='-,s,,,,,,' posn='110' name='-'>
        ))
</Sentence>
```

Sentence in corresponding CONLL format:

```
1       ریاست   ریاست   n       NN      cat-n|gen-f|num-sg|pers-3|case-o|vib-0_میں|tam-0|chunkId-NP|stype-|voicetype-       5       k7p     _       _
2       مہم     مہم     n       NN      cat-n|gen-f|num-sg|pers-3|case-d|vib-0_کا|tam-0|chunkId-NP2|stype-|voicetype-       4       r6-k1   _       _
3       بتدریج  بتدریج  adj     RB      cat-adv|gen-any|num-any|pers-|case-d|vib-|tam-|chunkId-RBP|stype-|voicetype-       5       adv     _       _
4       آغاز    آغاز    n       NN      cat-n|gen-m|num-sg|pers-3|case-d|vib-0|tam-0|chunkId-NP3|stype-|voicetype- 5       pof     _       _
5       ہو      ہو      v       VM      cat-v|gen-m|num-sg|pers-3|case-|vib-0_رہ+یا_ہے|tam-0|chunkId-VGF|stype-declarative|voicetype-active      0       root    _       _
```

Related Publications

1. A Dependency Treebank of Urdu and its Evaluation. *Riyaz Ahmad Bhat and Dipti Misra Sharma***.** Proceedings of the 6th Linguistic Annotation Workshop, pages 157–165, Jeju, Republic of Korea, 12-13 July 2012.

2. A Proposition Bank of Urdu. *Maaz Anwar Nomani, Riyaz Ahmad Bhat, Ashwini Vaidya, Tafseer Ahmed, Martha Palmer and Dipti Misra Sharma.* Proceedings of the 10th edition of the Language Resources and Evaluation Conference, Portorož, Slovenia, 23-28 May 2016.

Acknowledgement

We acknowledge the efforts of Mrs. Anjum Parveen, Mr. Riyaz Ahmed Bhat, Mr. Maaz Anwar and Mrs. Sameena Yasmeen for annotating and validating the UTB. The work reported in this documentation is supported by the NSF grant (Award Number: CNS 0751202; CFDA Number: 47.070)