A Dependency Treebank of Urdu and its Evaluation

Riyaz Ahmad Bhat

Dipti Misra Sharma

IIIT, HYDERABAD

Outline

- Introduction
- Treebanking efforts and related work
- Urdu Dependency Treebank
- Issues
- Evaluation
- Conclusion

Outline

- Introduction
- Treebanking efforts and related work
- Urdu Dependency Treebank
- Issues
- Evaluation
- Conclusion

Aim of the paper

- To provide a description of Urdu dependency Treebank developed using Paninian Grammar Framework.

- To discuss:

  → the task of annotation, and
  → the validity/reliability of the manual annotation.

Hindi Vs. Urdu

Hindi and Urdu are two literary styles of a sub-dialect (Hindustani).

- Similar in Grammar and Core vocabulary at colloquial level;
- Different vocabulary at literary and formal levels (mutually unintelligible);
  - » Hindi Vocabulary-Sanskritised
  - » Urdu Vocabulary-Persianised
- Written in two different scripts:
  - Hindi is written in Devnagri Script,

    हिन्दी भाषा

    hindi baashaa

    - 'Hindi language'
  - Urdu is written in Persio-Arabic Script.

    اردو زبان

    zabaan urdu

    'Urdu language'

Computational Paninian Grammar [CPG] Model

- A Dependency Grammar framework
    - modifier-modified relations
    - main verb of the sentence - primary modified
    - modifiers' relations with verb called *karaka*
- Inspired by Paninian grammatical analysis of Sanskrit,
- Suitable for syntactic analysis of morphologically rich languages.

Computational Paninian Grammar [CPG] Model

"'*karaka*' is the name given to the relation subsisting between a noun and a verb in a sentence." (Shastri, 1990)

- Six "*karaka*" relations defined by Panini are central to the framework:
  - *karta 'agent'*
  - *karma 'patient'*
  - *karana 'instrument'*
  - *sampradaan 'recipient'*
  - *apaadaan 'source'*
  - *adhikarana 'location'*

- *The framework also provides relations other than "karaka"* relations, such as purpose, reason, possession etc.

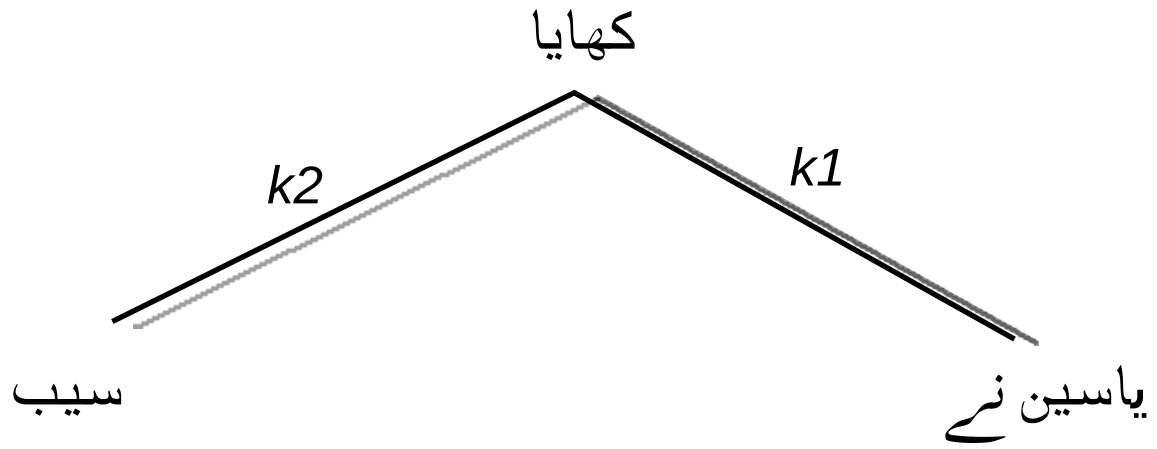
Example-1 shows *karaka* roles of verb “eat”:

یاسین نے سیب کھایا

Yasin-ne saeb khaya

Yasin-ERG apple-NOM eat-PaPERF

‘Yasin ate an apple.’



K1 - *karta*

K2 - *karma*

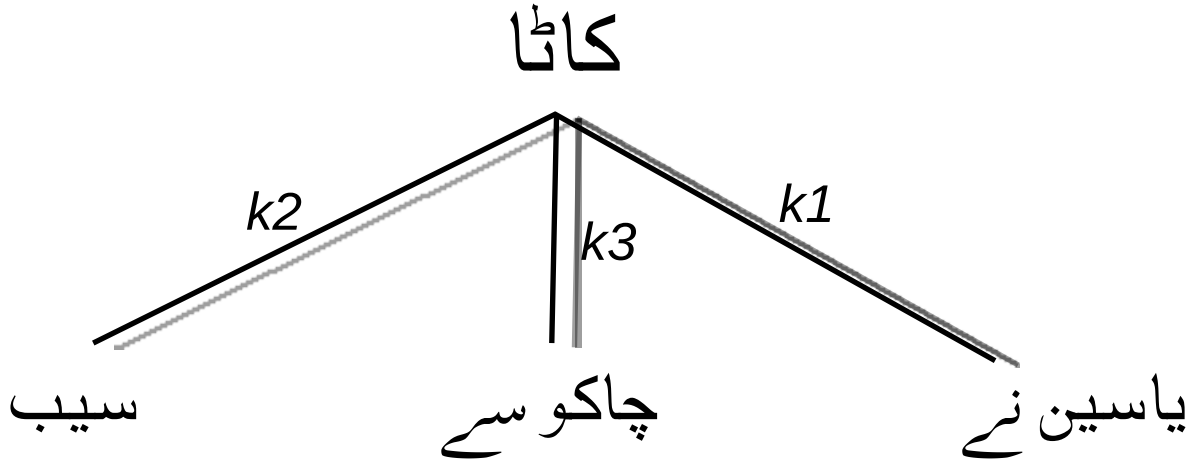
Example-2 shows *karaka* roles of verb “cut”:

یاسین نے چاکو سے سیب کاٹا

Yasin-ne chaku-se saeb kata

Yasin-ERG knife-INST apple-NOM cut-PaPERF

‘Yasin cut the apple with a knife.’



K1 - *karta*
K2 - *karma*
K3 - *karana*

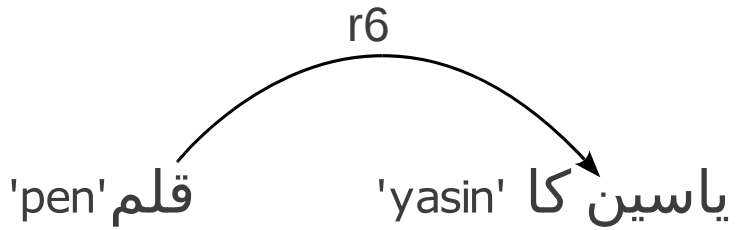
Example-3 shows a genitive construction showing possession (non-karaka relation):

یاسین کا قلم

Yasin-ka kalam

Yasin-GEN pen

'Yasin’s pen.'



Outline

- Introduction
- Treebanking efforts and related work
- Urdu Dependency Treebank
- Issues
- Evaluation
- Conclusion

Treebanks

- Treebanks play an increasingly important role in NLP tasks such as parsing.

- They can be an indispensable resource for linguistic investigations.

- Some of the Treebanks are:
  - Penn treebank (PTB)
    → Phrase structure analysis – English

  - Prague Dependency Treebank (PDT)
    → Dependency analysis - Czech

  - Hyderabad Dependency Treebank (HyDT)
    → Dependency analysis – Hindi

Outline

- Introduction
- Treebanking efforts and related work
- Urdu Dependency Treebank
- Issues
- Evaluation
- Conclusion

Urdu Dependency Treebank

- 0.1M words (around 3366 sentences) manually annotated with:

  → Morph features,
  → POS tags,
  → Chunk types, and
  → Inter-chunk Dependencies.

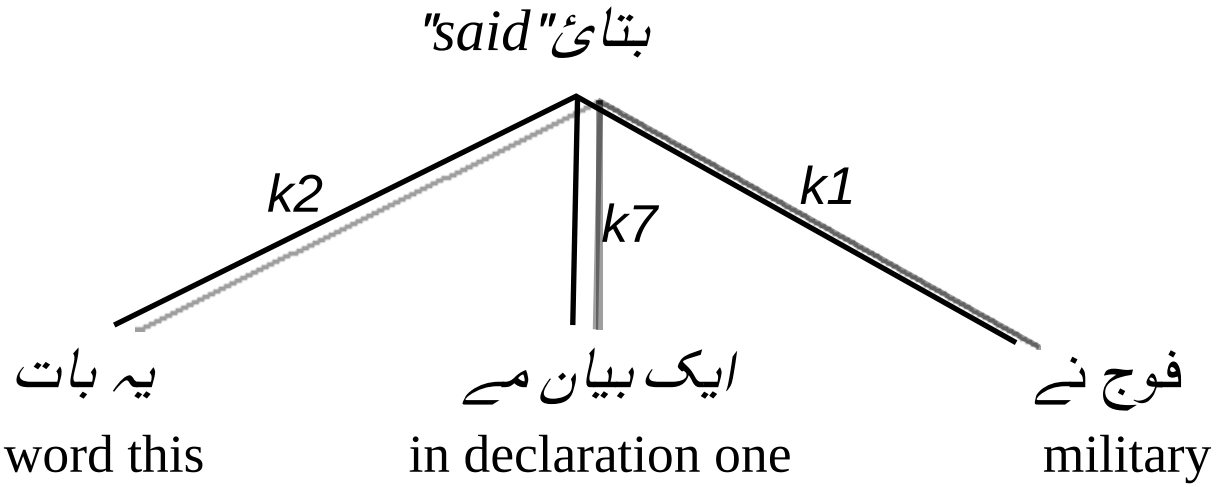
  - Treebank Statistics:

Corpus Type	Sentences	Words / sentence	Chunks / sentence
Newspaper articles	3366	29	13.7

Examples from the Treebank

فوج نے ایک بیان مے یہ بات بتائ۔

foj-ne ek bayan mem ye baat batayi
military-ERG one declaration in this word-NOM said-PaPERF
Military has revealed this matter in a declaration.

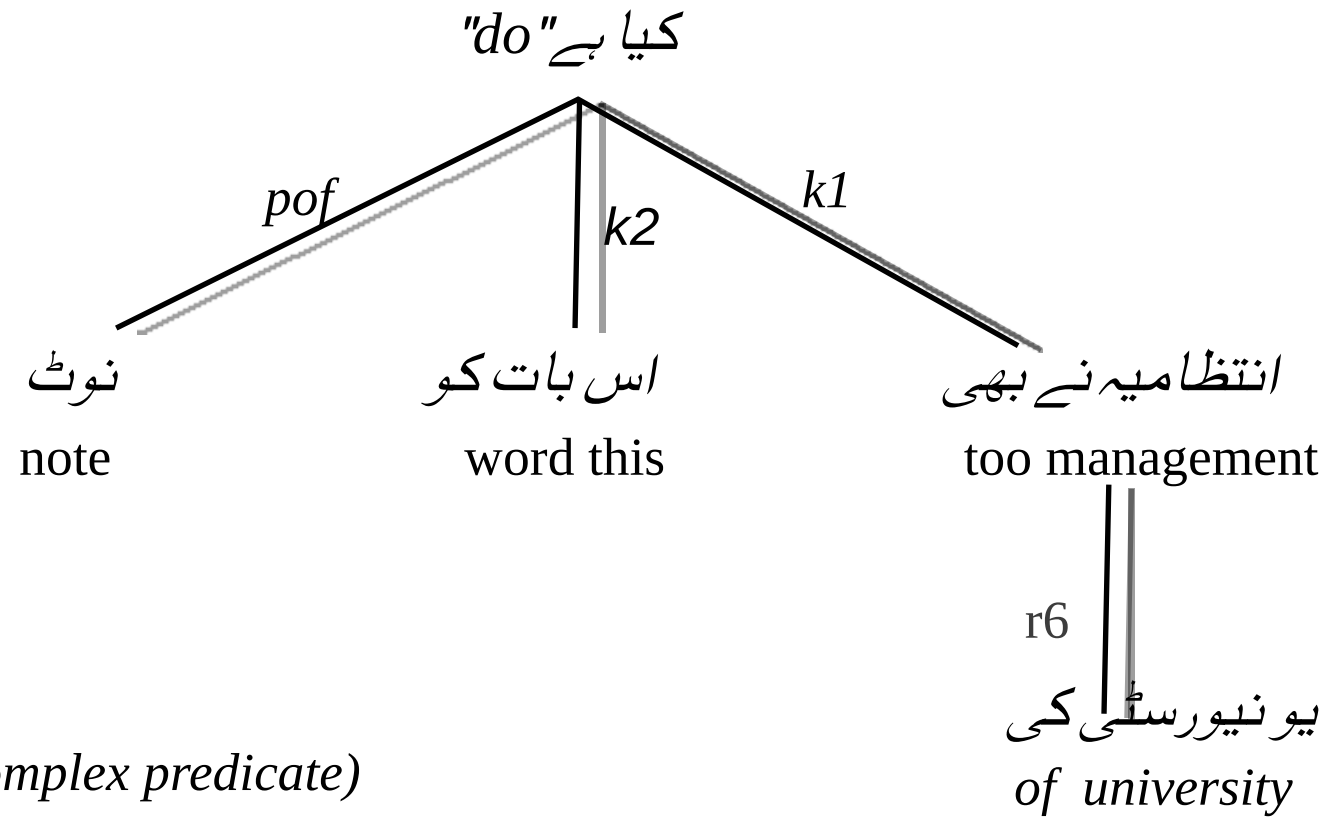


k1 - *karta*
k2 - *karma*
k7 - *adhikarana*

Examples from the Treebank

یونیورسٹی کی انتظامیہ نے بھی اس بات کو نوٹ کیا ہے

university-ki intizamiya-ne bhi is baat-ko note kiya hai.
University-GEN management-ERG too this word-ACC note do-PrPERF
The management of the University too has noted this point.



k1 – *karta*
k2 – *karma*
pof – *part of (complex predicate)*
r6 – genitive

IIIT, HYDERABAD

Outline

- Introduction
- Treebanking efforts and related work
- Urdu Dependency Treebank
- Issues
- Evaluation
- Conclusion

- **Differences with Hindi:**
  - → ***Ezafe:***
    - a loan construction from Persian,
    - contains an enclitic short vowel "*e*" joining two nouns, a noun and an adjective or an adposition and a noun in a possessive relation or a nominal modification.
    - head initial (Urdu is a head final language)

Examples:

ساحبِ تکھت

sahb-e takht

owner-Ez throne

'The owner of the throne.'

روزِ روشن

rooz-e rooshan

day-Ez bright

'Bright day.'

- **Annotating Ezafe:**

→ Modifier and head of an ezafe are chunked separately, both can take modifiers and project their own phrases.

- ***Examples-4***:

مالکِ دو جہاں

malik-e dho jahan

owner-Ez two worlds

‘The owner of two worlds.’

- ***Examples-5***:

تلاوتِ قلامِ پاک

tilawat-e qalam-e pak

recitation-Ez writing-Ez pure

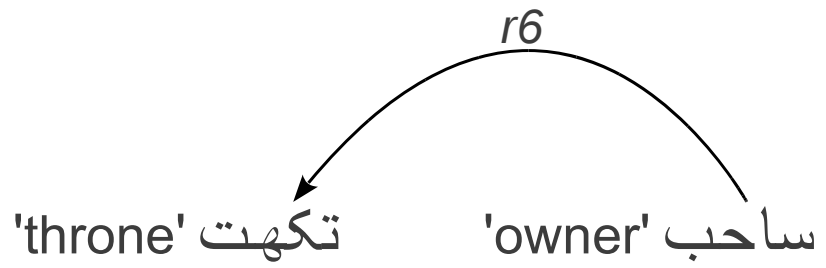
‘Recitation of the pure word.’

- In example-4 modifier noun جہاں 'world' is itself modified by دو 'two'.
- Example-5 shows a recursive ezafe construction where head noun تلاوتِ 'recitation' is modified by another ezafe قلامِ پاک

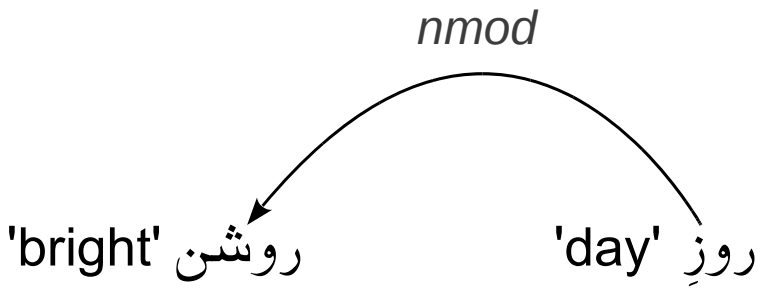
- **Annotating Ezafe:**

  Ezafe in urdu show possession and nominal modification:

  → Ezafe showing possession are annotated similar to genitives,



  → Ezafe showing nominal modification are annotated with an "*nmod*" relation.



- **Word Segmentation**

→ In Urdu writing *space* character is used for:

- generating correct shaping of words
  Example:
  ضرورت مند *"needy"* is a single word, a space is used after 'ت' in oder to prevent it from combining with the following character 'م', generating a *visually* wrong token ضرورتمند .

- separating words.
  Example:
  اردو مرکز "Urdu center" a space is used between اردو and مرکز to show them as separate words.

- **Word Segmentation**

→ In Urdu *space* character is thus an unreliable cue for word boundary.

- Words with spaces are broken into multiple tokens during tokenization,

  Example:

  Tokenizer divides ضرورتSPACE مند“*needy*” a single word into two tokens ضرورت and مند .

- Such erroneous tokens are corrected before further stages of treebanking,

  “_” 'underscore' is used to join the fragments of such words to ensure they are treated as one word with proper visual shape ضرورت_مند.

Outline

- Introduction
- Treebanking efforts and related work
- Urdu Dependency Treebank
- Issues
- Evaluation
- Conclusion

- **Inter-Annotator Agreement (IAA):**

  → to ensure validity of manual annotation,
  → to measure the annotators level of understanding of annotation guidlines,
  → greater the agreement more reliable and consistent the annotations are.

- **Measuring inter-annotator agreement:**

  → two annotators annotated same data set of 5600 words,
  → 2595 annotations (edges) marked with 39 labels,
  → agreement measured for every edge in a tree with respect to dependency label marked,
  → agreement scores calculated using Cohen's kappa.

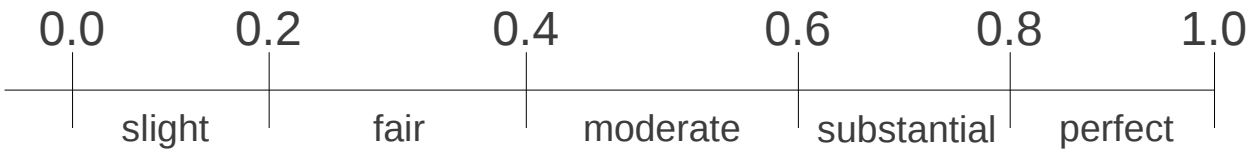
- **Cohen's Kappa (Cohen 1960):**

→ The kappa coefficient κ is calculated as:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

Notation: Pr(a) . . . observed (or "percentage") agreement
Pr(e) . . . expected agreement by chance

→ Scale for the interpretation of Kappa (Landis and Koch (1977))



- **Results and Discussion:**

→ Kappa Statistics:

No. of Annotations	Agreement	Pr(a)	Pr(e)	Kappa
2595	1921	0.74	0.097	0.71

0.71 kappa score shows a substantial agreement between the annotators.

	Relations	Ann.1	Ann.2	Agr.	Disagr.
1	*ras — k4*	0	1	0	1
2	*ras — k1*	4	6	3	4
3	*ras — k2*	1	3	0	4
4	*pof__idiom*	1	0	0	1
5	*r6 — k1*	10	8	4	10
6	*r6 — k2*	63	50	43	27
7	*rbmod*	2	0	0	2
8	*pof*	325	271	243	110
9	*rt*	43	48	38	15
10	*k3*	11	8	6	7
11	*rs*	1	8	1	7
12	*k2s*	21	30	17	17
13	*k2p*	4	3	2	3
14	*k1*	346	320	254	158
15	*rd*	13	3	2	12
16	*k2*	249	298	179	189
17	*nmod__relc*	27	30	13	31
18	*k7*	160	156	123	70
19	*jjmod*	23	8	8	15
20	*k5*	15	28	12	19
21	*k4*	46	50	34	28
22	*nmod__k2inv*	2	3	2	1
23	*rh*	21	15	7	22
24	*k4a*	10	12	7	8
25	*k7a*	5	6	4	3
26	*adv*	47	45	30	32
27	*nmod__k1inv*	0	1	0	1
28	*fragof*	6	7	5	3
29	*k7p*	46	44	29	32
30	*k7t*	67	71	53	32
31	*nmod__emph*	1	2	0	3
32	*k1s*	62	70	41	50
33	*r6*	297	335	258	116
34	*k1u*	0	1	0	1
35	*vmod*	102	98	63	74
36	*nmod*	91	96	48	91
37	*ccof*	436	486	389	144
38	*sent — adv*	1	0	0	1
39	*r6v*	5	5	3	4

Agreement and Disagreement between the Annotators.

- **Agreement Analysis:**

→ Disagreement on basic Karaka Roles:

- **Case syncretism** i.e. one to many mapping between case markers and case roles.

	نے 'ne'	کو 'ko'	کا 'ka'	سے 'se'	مے 'mem'	پر 'par'
k1	100	22	1	0	0	0
k2	0	46	1	15	0	0
k3	0	0	0	2	0	0
k4	0	17	0	19	0	0
k4a	0	2	0	0	0	0
k5	0	0	0	14	0	0
k7	0	0	1	1	60	70
k7t	0	5	2	11	6	0
k7p	0	0	0	0	19	10
r6	0	0	89	0	0	0
rh	0	0	0	5	0	0

Agreement among the Annotators on Karaka roles given a Case Marker.

- Agreement on 735/965 case marked nominals due to clear Karaka-case marker mapping;

- Disagreement on 230/965 due to case syncretism.

- **Agreement Analysis:**

  - Examples of Case Syncretism:

Example-6:

نادیا کو کحانی یاد آئ

nadya-ko kahani yaad aayi

nadiya-Dat story-NOM memory come-PST+PRF

'nadiya remembered the story.'

Example-7:

یاسین نے نادیا کو کتاب دی

yasin-ne nadiya-ko kitab di.

yasin-ERG nadiya-DAT book-NOM give-PST+PRF

'Yasin gave Nadiya a book.'

*Nadiya-**ko*** is an **exprencier subject** (k4a) in example-6 while it is **recipient** (k4) in example-7.

- **Agreement Analysis:**

  - **Indentification of Complex Predicates:** Disagreement due to similar syntactic distribution of a *part of complex predicate (**pof**)* and karaka role of a verb.

    → Out of 110 disagreements for label ‘pof’, annotators differ 81 (74%) times in marking a given dependency structure either with a *‘pof’* relation or with *‘karta-agent’*, *‘k1s-noun complement’ or ‘karma-theme’*.

Example-8:

یاسین نے نادیا کو کتاب دی

yasin-ne nadiya-ko kitab di.

yasin-ERG nadiya-DAT book-NOM give-PST+PRF

‘Yasin gave Nadiya a book.’

Example-9:

یاسین نے نادیا کو دھمکی دی

yasin-ne nadiya-ko dhamki di.

yasin-ERG nadiya-ACC threat give-PST+PRF

‘Yasin threatened Nadiya.’

- **Agreement Analysis:**

کتاب *"Book"* in example-8 and دھمکی *"threat"* in example-9 have similar syntactic context, in the former کتاب *"book"* is ***theme*** of the verb دی *"give"* and in later دھمکی *"threat"* forms a ***complex predicate*** with دی *"give"* (دھمکی دینا *"threaten"*).

Example-10:

یاسین نے نادیا سے چابی لی

yasin-ne nadiya-se chabi li.

Yasin-ERG Nadiya-ABL key-NOM take-PST+PRF

'Yasin took key from Nadiya.'

Example-11:

یاسین نے نادیا سے مدد لی

yasin-ne nadiya-se madad li.

yasin-ERG nadiya-ABL help take-PST+PRF

'Yasin took help from Nadiya.'

Similary چابی *"key"* is *theme* of the verb لی "take" in example-10 and مدد *"help"* is *part of* the complex predicate *"مدد لینا"* in example-11.

Outline

- Introduction
- Treebanking efforts and related work
- Urdu Dependency Treebank
- Issues
- Evaluation
- Conclusion

Conclusion

- Presented a CPG based dependency treebank of Urdu.
- Discussed:
  - Ezafe Construction,
  - Problem of word segmentation.
- Evaluation:
  - Calculated an IIA based evaluation of manual dependency annotations;
    - » Annotators show similar enough understanding of the annotation guidelines.
    - » Annotations in Urdu Treebank must be substantially consistent given the high kappa score.

References

- R. Begum, S. Husain, A. Dhwaj, D.M. Sharma, L. Bai,and R. Sangal. 2008. Dependency annotation scheme for Indian languages. In Proceedings of IJCNLP. Citeseer.

- R. Begum, K. Jindal, A. Jain, S. Husain, and D. Misra Sharma. 2011. Identification of conjunct verbs in hindi and its effect on parsing accuracy. Computational Linguistics and Intelligent Text Processing, pages 29–40.

- A. Bharati, V. Chaitanya, R. Sangal, and KV Ramakrishnamacharyulu. 1995. Natural Language Processing: A Paninian Perspective. Prentice-Hall of India.

- Bharati, M. Bhatia, V. Chaitanya, and R. Sangal. 1996. Paninian grammar framework applied to english. Technical report, Technical Report TRCS-96-238, CSE, IIT Kanpur.

- A. Bharati, R. Sangal, and D.M. Sharma. 2007. Ssf: Shakti standard format guide. Technical report, Technical report, IIIT Hyderabad.

- Bharati, D.M. Sharma, S. Husain, L. Bai, R. Begum, and R. Sangal. 2009. Anncorra: Treebanks for Indian languages guidelines for annotating Hindi treebank (version–2.0).

- D.N.S. Bhat. 1991. Grammatical relations: the evidence against their necessity and universality. Psychology Press.

- R. Bhatt, B. Narasimhan, M. Palmer, O. Rambow, D.M. Sharma, and F. Xia. 2009. A multi-representational and multi-layered treebank for hindi/urdu. In Proceedings of the Third Linguistic Annotation Workshop, pages 186–189. Association for Computational Linguistics.

- T. Bögel, M. Butt, and S. Sulger. 2008. Urdu ezafe and the morphology-syntax interface. Proceedings of LFG08.

- F. Bond, S. Fujita, and T. Tanaka. 2008. The Hinoki syntactic and semantic treebank of Japanese. Language Resources and Evaluation, 42(2):243–251.

- C. Bosco and V. Lombardo. 2004. Dependency and relational structure in treebank annotation. In Proceedings of Workshop on Recent Advances in Dependency Grammar at COLING'04.

- S. Brants, S. Dipper, S. Hansen, W. Lezius, and G. Smith. 2002. The tiger treebank. In Proceedings of the workshop on treebanks and linguistic theories, pages 24-41.

- J. Carletta, S. Isard, G. Doherty-Sneddon, A. Isard, J.C. Kowtko, and A.H. Anderson. 1997. The reliability of a dialogue structure coding scheme. Computational linguistics, 23(1):13–31.

- H. Chaudhry and D.M. Sharma. 2011. Annotation and issues in building an English dependency treebank.

- J. Cohen et al. 1960. A coefficient of agreement for nominal scales. Educational and psychological measurement, 20(1):37–46.

- A. Culotta and J. Sorensen. 2004. Dependency tree kernels for relation extraction. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, page 423. Association for Computational Linguistics.

- N. Durrani and S. Hussain. 2010. Urdu word segmentation. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 528–536. Association for Computational Linguistics.

- E. Hajicová . 1998. Prague dependency treebank: From analytic to tectogrammatical annotation. Proceedings of TSD98, pages 45–50.

- J. Hajič. 1998. Building a syntactically annotated corcpus: The prague dependency treebank. Issues of valency and meaning, pages 106–132.

- E. Hajičová , A. Abeillé , J. Hajič , J. Mírovský , and Z. Urešová . 2010. Treebank annotation. In Nitin Insadurkhya and Fred J. Damerau, editors, Handbook of Natural Language Processing, Second Edition. CRC Press, Taylor and Francis Group, Boca Raton, FL. ISBN 978-1420085921.

- J.R. Landis and G.G. Koch. 1977. The measurement of observer agreement for categorical data. Biometrics, pages 159–174.

- G.S. Lehal. 2010. A word segmentation system for handling space omission problem in urdu script. In 23rd International Conference on Computational Linguistics, page 43.

- M.P. Marcus, M.A. Marcinkiewicz, and B. Santorini. 1993. Building a large annotated corpus of english: The penn treebank. Computational linguistics, 19(2):313–330.

- C.P. Masica. 1993. The Indo-Aryan Languages. Cambridge Univ Pr, May.

- T.W. Mohanan. 1990. Arguments in Hindi. Ph.D. Thesis, Stanford University.

- K. Oflazer, B. Say, D.Z. Hakkani-Tür, and G. Tür. 2003. Building a turkish treebank. Abeillé (Abeillé, 2003), pages 261–277.

- M. Palmer, R. Bhatt, B. Narasimhan, O. Rambow, D.M. Sharma, and F. Xia. 2009. Hindi syntax: Annotating dependency, lexical predicate-argument structure, and phrase structure. In The 7th International Conference on Natural Language Processing, pages 14–17.

- O. Rambow, C. Creswell, R. Szekely, H. Taber, and M. Walker. 2002. A dependency treebank for english. In Proceedings of LREC, volume 2.

- F. Reichartz, H. Korte, and G. Paass. 2009. Dependency tree kernels for relation extraction from natural language text. Machine Learning and Knowledge Discovery in Databases, pages 270–285.

- S.M. Shieber. 1985. Evidence against the contextfreeness of natural language. Linguistics and Philosophy, 8(3):333–343.

- L. Uria, A. Estarrona, I. Aldezabal, M. Aranzabe, A. Díaz ı de Ilarraza, and M. Iruskieta. 2009. Evaluation of the syntactic annotation in epec, the reference corpus for the processing of basque. Computational Linguistics and Intelligent Text Processing, pages 72–85.

- A. Vaidya, S. Husain, P. Mannem, and D. Sharma. 2009. A karaka based annotation scheme for english. Computational Linguistics and Intelligent Text Processing, pages 41–52.

- L. Van der Beek, G. Bouma, R. Malouf, and G. Van Noord. 2002. The alpino dependency treebank. Language and Computers, 45(1):8–22.

- C. Yong and S.K. Foo. 1999. A case study on interannotator agreement for word sense disambiguation.

Thank You!