A Dependency Treebank of Urdu and its Evaluation

Riyaz Ahmad Bhat
Dipti Misra Sharma
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.
“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”:

\[
\text{یاسین نے سیب کھایا}
\]

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”:

\[
\text{یاسین نے چاکو سے سیب کاٹا}
\]

Yasin-ne chaku-se saeb kara

‘Yasin cut the apple with a knife.’

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

\[
\text{یاسین کا قلم}
\]

Yasin-ka kalam

‘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:

<table>
<thead>
<tr>
<th>Corpus Type</th>
<th>Sentences</th>
<th>Words / sentence</th>
<th>Chunks / sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newspaper articles</td>
<td>3366</td>
<td>29</td>
<td>13.7</td>
</tr>
</tbody>
</table>
Military has revealed this matter in a declaration.

Examples from the Treebank

"said"  بتائے

فوج نے ایک بیان مے یہ بات بتائے۔
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
The management of the University too has noted this point.

Examples from the Treebank

University-GEN management-ERG too this word-ACC note do-PrPERF

The 6th Linguistic Annotation Workshop
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 تکهت

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:

ماَلِک دو جِہاں
malik-e dho jahan
owner-Ez two worlds
‘The owner of two worlds.’

• Examples-5:

تِلَوَت قَلَام 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 تلاوت 'recitation' is modified by another ezafe قلام پاک.
**Annotating Ezafe:**

Ezafe in Urdu show possession and nominal modification:

→ Ezafe showing possession are annotated similar to genitives,

\[ r6 \]

'Ezafe' تکهت ساحب 'owner'

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

\[ nmod \]

'bright' روشن 'day' روز
• 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 order 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 ضرورت SPACE مند "needy" a single word into two tokens مند ضرورت.

- 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_منت.
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 guidelines,
→ 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 $\kappa$ 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))

<table>
<thead>
<tr>
<th>$\kappa$</th>
<th>slight</th>
<th>fair</th>
<th>moderate</th>
<th>substantial</th>
<th>perfect</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results and Discussion:

→ Kappa Statistics:

<table>
<thead>
<tr>
<th>No. of Annotations</th>
<th>Agreement</th>
<th>Pr(a)</th>
<th>Pr(e)</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>2595</td>
<td>1921</td>
<td>0.74</td>
<td>0.097</td>
<td>0.71</td>
</tr>
</tbody>
</table>

0.71 kappa score shows a substantial agreement between the annotators.
<table>
<thead>
<tr>
<th>Relations</th>
<th>Ann.1</th>
<th>Ann.2</th>
<th>Agr.</th>
<th>Disagr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 \ ras - k4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2 \ ras - k1</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3 \ ras - k2</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>4 \ pof_idiom</td>
<td>10</td>
<td>8</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>5 \ r6 - k1</td>
<td>63</td>
<td>50</td>
<td>43</td>
<td>27</td>
</tr>
<tr>
<td>6 \ r6 - k2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>7 \ rbmod</td>
<td>325</td>
<td>271</td>
<td>243</td>
<td>110</td>
</tr>
<tr>
<td>8 \ pof</td>
<td>43</td>
<td>48</td>
<td>38</td>
<td>15</td>
</tr>
<tr>
<td>9 \ k3</td>
<td>11</td>
<td>8</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>10 \ rs</td>
<td>1</td>
<td>8</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>11 \ k2s</td>
<td>21</td>
<td>30</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>12 \ k2p</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>13 \ k1</td>
<td>346</td>
<td>320</td>
<td>254</td>
<td>158</td>
</tr>
<tr>
<td>14 \ rd</td>
<td>13</td>
<td>3</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>15 \ k2</td>
<td>249</td>
<td>298</td>
<td>179</td>
<td>189</td>
</tr>
<tr>
<td>16 \ nmod_relc</td>
<td>27</td>
<td>30</td>
<td>13</td>
<td>31</td>
</tr>
<tr>
<td>17 \ k7</td>
<td>160</td>
<td>156</td>
<td>123</td>
<td>70</td>
</tr>
<tr>
<td>18 \ jjmod</td>
<td>23</td>
<td>8</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>19 \ k5</td>
<td>15</td>
<td>28</td>
<td>12</td>
<td>19</td>
</tr>
<tr>
<td>20 \ k4</td>
<td>46</td>
<td>50</td>
<td>34</td>
<td>28</td>
</tr>
<tr>
<td>21 \ nmod_k2inv</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>22 \ rh</td>
<td>21</td>
<td>15</td>
<td>7</td>
<td>22</td>
</tr>
<tr>
<td>23 \ k4a</td>
<td>10</td>
<td>12</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>24 \ k7a</td>
<td>5</td>
<td>6</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>25 \ adv</td>
<td>47</td>
<td>45</td>
<td>30</td>
<td>32</td>
</tr>
<tr>
<td>26 \ nmod_k1inv</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>27 \ fragof</td>
<td>6</td>
<td>7</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>28 \ k7p</td>
<td>46</td>
<td>44</td>
<td>29</td>
<td>32</td>
</tr>
<tr>
<td>29 \ k7t</td>
<td>67</td>
<td>71</td>
<td>53</td>
<td>32</td>
</tr>
<tr>
<td>30 \ nmod_emph</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>31 \ k1s</td>
<td>62</td>
<td>70</td>
<td>41</td>
<td>50</td>
</tr>
<tr>
<td>32 \ r6</td>
<td>297</td>
<td>335</td>
<td>258</td>
<td>116</td>
</tr>
<tr>
<td>33 \ k1u</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>34 \ vmod</td>
<td>102</td>
<td>98</td>
<td>63</td>
<td>74</td>
</tr>
<tr>
<td>35 \ nmod</td>
<td>91</td>
<td>96</td>
<td>48</td>
<td>91</td>
</tr>
<tr>
<td>36 \ ccof</td>
<td>436</td>
<td>486</td>
<td>389</td>
<td>144</td>
</tr>
<tr>
<td>37 \ sent_adj</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>38 \ r6v</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>'ne'</th>
<th>'ko'</th>
<th>'ka'</th>
<th>'se'</th>
<th>'mem'</th>
<th>'par'</th>
</tr>
</thead>
<tbody>
<tr>
<td>k1</td>
<td>100</td>
<td>22</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>k2</td>
<td>0</td>
<td>46</td>
<td>1</td>
<td>15</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>k3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>k4</td>
<td>0</td>
<td>17</td>
<td>0</td>
<td>19</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>k4a</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>k5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>k7</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>60</td>
<td>70</td>
</tr>
<tr>
<td>k7t</td>
<td>0</td>
<td>5</td>
<td>2</td>
<td>11</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>k7p</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>10</td>
</tr>
<tr>
<td>r6</td>
<td>0</td>
<td>89</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>rh</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

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 کہانی یاد آی
nadiya-Dat story-NOM memory come-PST+PRF
‘nadiya remembered the story.’

Example-7:
یاسین نے نادیا کو کتاب دی
yasin-ERG nadiya-DAT book-NOM give-PST+PRF
‘Yasin gave Nadiya a book.’

Nadiya-ko is an exprenier 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:**

\[
\begin{align*}
\text{یاسین نے نادیا کو کتاب دی} & \\
yasin-ne & \quad nadiya-ko & \quad kitab & \quad di. \\
yasin-ERG & \quad nadiya-DAT & \quad book-NOM & \quad give-PST+PRF
\end{align*}
\]

‘Yasin gave Nadiya a book.’

**Example-9:**

\[
\begin{align*}
\text{یاسین نے نادیا کو دھمکی دی} & \\
yasin-ne & \quad nadiya-ko & \quad dhamki & \quad di. \\
yasin-ERG & \quad nadiya-ACC & \quad threat & \quad give-PST+PRF
\end{align*}
\]

‘Yasin threatened Nadiya.’
Agreement Analysis:

“Book” in example-8 and “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 “take help” 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


> C. Yong and S.K. Foo. 1999. A case study on interannotator agreement for word sense disambiguation.
Thank You!