Adding an Annotation Layer to the Hindi/Urdu Treebank

Annette Hautli, Sebastian Sulger, Miriam Butt

TLT10, Heidelberg
January 6th, 2012
Aim of the talk

*Hindi/Urdu Treebank*
new treebank resource
Aim of the talk

- **Hindi/Urdu Treebank**
  - new treebank resource

- **Urdu ParGram grammar**
  - LFG grammar for Hindi/Urdu

augment with??
Aim of the talk

Hindi/Urdu Treebank
new treebank resource

augment with??

Urdu ParGram grammar
LFG grammar for Hindi/Urdu

Improved Hindi/Urdu Treebank
Can the HUTB benefit from the Urdu ParGram grammar? If so, how?
Outline of the talk:

Overview: Hindi/Urdu resources

The Hindi/Urdu Treebank

The Urdu ParGram grammar

An additional annotation layer for the HUTB

Conclusion
Outline of the talk

Overview: Hindi/Urdu resources

The Hindi/Urdu Treebank

The Urdu ParGram grammar

An additional annotation layer for the HUTB

Conclusion
Hindi/Urdu resources

- Hindi/Urdu: structurally very similar, differences in vocabulary, orthography
- different scripts: Arabic script for Urdu, Devanagari for Hindi
- often regarded as varieties of a single language (*Hindustani*)
- ∼240 million native speakers
- but: under-resourced language
- only a few lexical resources, no large corpora, no broad-coverage parsers
Outline of the talk

Overview: Hindi/Urdu resources

The Hindi/Urdu Treebank

The Urdu ParGram grammar

An additional annotation layer for the HUTB

Conclusion
The Hindi/Urdu treebank (HUTB)

- new treebank resource for Hindi/Urdu
- collaborative research effort of five universities (University of Colorado at Boulder, Columbia University, University of Massachusetts at Amherst, University of Washington, International Institute of Information Technology in India)
- project started in 2008 (work in progress)
- goal: building a multi-representational, multi-layered treebank
- described e.g. Palmer et al. (2007) and Bhatt et al. (2009)
- see other talks at this conference
HUTB: multi-layered, multi-representational

- three levels of annotation: two syntactic levels, one lexical-semantic level
HUTB: multi-layered, multi-representational

- three levels of annotation: two syntactic levels, one lexical-semantic level
- syntactic encoding 1: annotation of phrase structure
  - inspired by Chomskyan approach to syntax
  - binary trees, traces (movement)
HUTB: multi-layered, multi-representational

- three levels of annotation: two syntactic levels, one lexical-semantic level
- syntactic encoding 1: annotation of phrase structure
  - inspired by Chomskyan approach to syntax
  - binary trees, traces (movement)
- syntactic encoding 2: annotation of dependencies
  - Computational Pāṇinian Grammar (CPG) (Bharati et al. 1995)
  - system of kaṇaka relations (cf. thematic roles) to model head-argument relations (≈30 kaṇaka relations)
HUTB: multi-layered, multi-representational

- three levels of annotation: two syntactic levels, one lexical-semantic level
- syntactic encoding 1: annotation of phrase structure
  - inspired by Chomskyan approach to syntax
  - binary trees, traces (movement)
- syntactic encoding 2: annotation of dependencies
  - Computational Pāṇinian Grammar (CPG) (Bharati et al. 1995)
  - system of kaṟaka relations (cf. thematic roles) to model head-argument relations (~30 kaṟaka relations)
- lexical-semantic encoding: PropBank roles (Palmer et al. 2005)
  - semantic roles (Arg0, Arg1 etc.) assigned to verb arguments
  - PropBank roles are mapped onto kaṟaka roles
An example from the treebank

दूतावास ऐंध्कारियों ने उसे अच्छी से हत में पाया

dUtvAs adHikAryOn=nE usE accHl
embassy.Masc.Sg officer.Masc.Pl=Erg he.Acc good.Fem.Sg
sEhat=mEN pA-yA
health.Fem.Sg=Loc find-Perf.Masc.Sg

‘Embassy officers found him in a healthy condition.’
Syntactic annotation

((( NP  
   XC हृतावास
   NN एड्कारियों
   PSP मे
 ))
(( NP  
   PRP उसे
 ))
(( NP  
   JJ अच्छी
   NN सेहत
   PSP मे
 ))
(( VGF  
   VM पाया
   SYM .
 )))

<fs drel='k1:VGF' name='NP'>
<fs drel='k2:VGF' name='NP2'>
<fs drel='k7:VGF' name='NP3'>

Figure: Sample sentence from the Hindi/Urdu Treebank
Syntactic annotation

\[
(( \text{NP} \quad <\text{fs} \ drel='k1:VGF' \ name='NP'>
\begin{align*}
\text{XC} & \text{ दृतावास} \\
\text{NN} & \text{ एध्यारियों} \\
\text{PSP} & \text{ ने}
\end{align*}
\))
\]

Figure: Closer look at NP \textit{dUtAvAs adHikAriyOn=nE}
Syntactic annotation

((( NP
  XC दृतावास
  NN एथ्कारियों
  PSP ने
))

Figure: Closer look at NP *dUtAvAs adHikAriyOn=nE*

- preterminal nodes assign part of speech to lexical items (XC, NN, etc.)
Syntactic annotation

\[
(( \text{NP} <\textit{fs} \text{ drel='k1:VGF' name='NP'}> \\
  \text{XC द्रतावास} \\
  \text{NN एथ्कारियों} \\
  \text{PSP ने} \\
 ))
\]

Figure: Closer look at NP \textit{dU}t\textit{A}v\textit{A}s \textit{adH}i\textit{kA}ri\textit{yOn}=nE

- preterminal nodes assign part of speech to lexical items (XC, NN, etc.)
- nodes are grouped into constituents by bracketing
Syntactic annotation

(( NP <fs drel='k1:VGF' name='NP'>
    XC द्वारा
    NN एक्सेसरीयों
    PSP ने
  ))

**Figure:** Closer look at NP dUtAvAs adHikAriyOn=nE

- preterminal nodes assign part of speech to lexical items (XC, NN, etc.)
- nodes are grouped into constituents by bracketing
- dependencies are attached to constituents (drel)
Lexical-semantic annotation

<table>
<thead>
<tr>
<th>पा ‘to find’</th>
<th></th>
<th>पा ‘to find’</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Arg0</strong></td>
<td>agent</td>
<td>हृतावास ऐंध्यारियों ‘embassy officers’</td>
</tr>
<tr>
<td><strong>Arg1</strong></td>
<td>patient (theme)</td>
<td>उसे ‘him’</td>
</tr>
<tr>
<td><strong>ArgM-MNR</strong></td>
<td>modifier (manner)</td>
<td>अच्छी सेहत ‘good health’</td>
</tr>
</tbody>
</table>

**Table:** PropBank frame for पा ‘to find’
Lexical-semantic annotation

<table>
<thead>
<tr>
<th>पा ‘to find’</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Arg0</strong></td>
</tr>
<tr>
<td><strong>Arg1</strong></td>
</tr>
<tr>
<td><strong>ArgM-MNR</strong></td>
</tr>
</tbody>
</table>

Table: PropBank frame for पा ‘to find’

- PropBank roles are mapped onto कर्का roles
Lexical-semantic annotation

<table>
<thead>
<tr>
<th>पा ‘to find’</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arg0</td>
</tr>
<tr>
<td>Arg1</td>
</tr>
<tr>
<td>ArgM-MNR</td>
</tr>
</tbody>
</table>

Table: PropBank frame for पा ‘to find’

- PropBank roles are mapped onto kaṇaka roles
- e.g., agent Arg0 mapped onto kaṇaka role k1 (dUtvAvAs adHikAriyOn=nE)
Outline of the talk

Overview: Hindi/Urdu resources

The Hindi/Urdu Treebank

The Urdu ParGram grammar

An additional annotation layer for the HUTB

Conclusion
Introducing the Urdu ParGram grammar

- research project at University of Konstanz
- part of an international research program called ParGram (Parallel Grammars)
  - aimed at developing broad-coverage grammars
  - couched within LFG framework (c-structure for constituents, f-structure for dependencies/grammatical relations)
  - uses XLE grammar development platform to *manually* construct grammars (Crouch et al. 2011)
  - described e.g. in Butt and King (2007), Bögel et al. (2009)
A grammar example

दूतावास ऐत्थ्वारियों ने उसे अच्छी सेहत में पाया

dUtAvAs  adHikAriyOn=nE  usE  accHI
embassy.Masc.Sg officer.Masc.Pl=Erg he=Acc good.Fem.Sg
sEhat=mEN     pA-yA
health.Fem.Sg=Loc find-Perf.Masc.Sg

‘Embassy officers found him in a healthy condition.’
Output (f-structure)

"dUtAvAs adHikAriyOn nE usE accHI sEhat mEN pAyA"

```
[ PRED 'pA< [2:dUtAvAs adHikAriyOn], [25:vuh]> ]
[ SUBJ NTYPE [NSYN common] ]
  2 CASE erg, GEND masc, NUM pl, PERS 3
[ PRED 'vuh' ]
  [ PRED 'sEhat' ]
  [ ADJUNCT { [ PRED 'accH' ]
      [ ADJUNCT { [ NTYPE [NSYN common] ]
          [ NTYPE [NSYN pronoun] ]
          25 CASE acc, NUM sg, PERS 3, PRON-TYPE pers
          TNS-ASP [ASPECT perf, MOOD indicative]
          68 CLAUSE-TYPE decl, PASSIVE -, VTYPE main
```
What information can the Urdu ParGram grammar add to the HUTB?

- LFG f-structures explicitly encode dependency information
- f-structures: not only head-argument dependencies, but rather highly articulated feature/value combinations that can serve as input to semantic computations
What information can the Urdu ParGram grammar add to the HUTB?

- LFG f-structures explicitly encode dependency information
- f-structures: not only head-argument dependencies, but rather highly articulated feature/value combinations that can serve as input to semantic computations
  - f-structures contain more detailed information than the HUTB annotation scheme
What information can the Urdu ParGram grammar add to the HUTB?

- LFG f-structures explicitly encode dependency information
- f-structures: not only head-argument dependencies, but rather highly articulated feature/value combinations that can serve as input to semantic computations
- f-structures contain more detailed information than the HUTB annotation scheme

ASPECT perf, TENSE pres, DEGREE positive, MODALITY can, ADDRESS rude, DEIXIS proximal, MOOD imperative, PROPER-TYPE location, ...
Outline of the talk

Overview: Hindi/Urdu resources

The Hindi/Urdu Treebank

The Urdu ParGram grammar

An additional annotation layer for the HUTB

Conclusion
An additional annotation layer for the HUTB

- proposition: add an additional annotation layer to the HUTB
- additional layer: XLE *triples*
An additional annotation layer for the HUTB

- proposition: add an additional annotation layer to the HUTB
- additional layer: XLE triples
  - information harvested from Urdu ParGram grammar parses
An additional annotation layer for the HUTB

- proposition: add an additional annotation layer to the HUTB
- additional layer: XLE *triples*
  - information harvested from Urdu ParGram grammar parses
  - feature-value information extracted from f-structures in the form of *triples*
An additional annotation layer for the HUTB

- proposition: add an additional annotation layer to the HUTB
- additional layer: XLE *triples*
  - information harvested from Urdu ParGram grammar parses
  - feature-value information extracted from f-structures in the form of *triples*
  - flexible, internal XLE process allows for the translation of f-structure features into triples format (features may be deleted or flattened)
An additional annotation layer for the HUTB

- proposition: add an additional annotation layer to the HUTB
- additional layer: XLE *triples*
  - information harvested from Urdu ParGram grammar parses
  - feature-value information extracted from f-structures in the form of *triples*
  - flexible, internal XLE process allows for the translation of f-structure features into triples format (features may be deleted or flattened)
  - has been done for parts of the Wall Street Journal section of PennTreebank (PARC700, King et al. (2003))
The XLE triples format

\[
\text{TNS-ASP [ASPECT perf, MOOD indicative]}
\]

\[
\text{68 [CLAUSE-TYPE decl, PASSIVE -, VTYPE main]}
\]
The XLE triples format

\[
\begin{align*}
\text{TNS-ASP} & \quad [\text{ASPECT perf, MOOD indicative}] \\
68 & \quad [\text{CLAUSE-TYPE decl, PASSIVE -}, \text{ VTYPE main}]
\end{align*}
\]

\[\downarrow\]

flattening of TNS-ASP feature, deletion of CLAUSE-TYPE feature
The XLE triples format

\[
\begin{array}{c}
\text{TNS-ASP} \quad [\text{ASPECT perf, MOOD indicative}] \\
\text{68} \quad [\text{CLAUSE-TYPE decl, PASSIVE -}, \text{VTYPE main}]
\end{array}
\]

\[
\downarrow
\]

flattening of TNS-ASP feature, deletion of CLAUSE-TYPE feature

\[
\downarrow
\]

resulting triples selection:

\[
\begin{align*}
\text{aspect}(pA, \text{perf}) \\
\text{mood}(pA, \text{indicative}) \\
\text{passive}(pA, -) \\
\text{vtype}(pA, \text{main})
\end{align*}
\]
### Included Features: Overview

<table>
<thead>
<tr>
<th>Grammatical function labels</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>subj</td>
<td>subject</td>
<td>obj</td>
</tr>
<tr>
<td>obl</td>
<td>oblique</td>
<td>comp</td>
</tr>
<tr>
<td>xcomp</td>
<td>open complement</td>
<td>predlink</td>
</tr>
<tr>
<td>adjunct</td>
<td>adjunct phrase</td>
<td>conj</td>
</tr>
<tr>
<td>topic</td>
<td>topic phrase</td>
<td>focus</td>
</tr>
<tr>
<td>poss</td>
<td>possessive phrase</td>
<td>mod</td>
</tr>
<tr>
<td></td>
<td></td>
<td>object</td>
</tr>
<tr>
<td></td>
<td></td>
<td>compl. clause</td>
</tr>
<tr>
<td></td>
<td></td>
<td>copula constr.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>conjunction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>focus phrase</td>
</tr>
<tr>
<td></td>
<td></td>
<td>modifier clause</td>
</tr>
</tbody>
</table>

**Table:** GF labels for an additional dependency annotation
### Included Features: Overview

<table>
<thead>
<tr>
<th>feature</th>
<th>value</th>
<th>feature</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>address</td>
<td>e.g. rude, familiar</td>
<td>adv-type</td>
<td>e.g. loc, sadv</td>
</tr>
<tr>
<td>adjunct-type</td>
<td>e.g. loc</td>
<td>aspect</td>
<td>e.g. prog</td>
</tr>
<tr>
<td>case</td>
<td>e.g. erg, acc, dat</td>
<td>causative</td>
<td>direct, indirect</td>
</tr>
<tr>
<td>adjunct-type</td>
<td>e.g. and, because</td>
<td>deixis</td>
<td>e.g. proximal</td>
</tr>
<tr>
<td>case</td>
<td>masc, fem</td>
<td>mood</td>
<td>e.g. imperative</td>
</tr>
<tr>
<td>coord-form</td>
<td>e.g. ezafe</td>
<td>num</td>
<td>sg, pl</td>
</tr>
<tr>
<td>gend</td>
<td>e.g. must, can</td>
<td>tense</td>
<td>e.g. past</td>
</tr>
<tr>
<td>mod-type</td>
<td>card, ord</td>
<td>passive</td>
<td>+</td>
</tr>
<tr>
<td>modality</td>
<td>e.g. pers, rel</td>
<td>proper-type</td>
<td>e.g. location, name</td>
</tr>
<tr>
<td>number-type</td>
<td>e.g. main, copular</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table:** Other feature labels for an additional dependency annotation
Modality in Hindi/Urdu

- Hindi/Urdu features only two *dedicated* modals: *sak* ‘can’, *cAhlyE* ‘need’ (defective paradigm: only *cAhlyE*)
Modality in Hindi/Urdu

- Hindi/Urdu features only two dedicated modals: *sak* ‘can’, *cAhlyE* ‘need’ (defective paradigm: only *cAhlyE*)
- all other modality: expressed *constructionally* by combining main verb with one of the following: *pA* ‘find’, *paR* ‘fall’, *hO* ‘be’ (Bhatt et al. 2011)
Modality in Hindi/Urdu

- Hindi/Urdu features only two dedicated modals: sak ‘can’, cAhlyE ‘need’ (defective paradigm: only cAhlyE)
- all other modality: expressed *constructionally* by combining main verb with one of the following: pA ‘find’, paR ‘fall’, hO ‘be’ (Bhatt et al. 2011)
- example:
  yAsIn vuh kar sak-A
  Yassin.Masc.Sg that.Sg.Nom do can-Perf.Masc.Sg
  ‘Yassin could do that.’
Modality in the HUTB

- modal constructions annotated on PropBank annotation layer
- PropBank label ARG0–Mod marks modals
Modality in the HUTB

- modal constructions annotated on PropBank annotation layer
- PropBank label ARG\text{-}Mod marks modals
- but: modal meaning not further identified:

> Modal constructions in Hindi convey notions such as ability, desire, obligation, permission, etc. In Prop[on]bank, we will annotate the following cases using the ARG\text{-}Mod label.

(PB\textunderscore guidelinesNov2010.pdf, p. 29)
Modality in the Urdu ParGram grammar

- Hindi/Urdu modals recently implemented in the Urdu ParGram grammar
- f-structure features not only encode modality as present in some construction
- detailed attribute-value pairs account for the nature of the expressed modality
- modality subsumed under f-structure feature [MODALITY]
Modality in the additional layer for the HUTB

"yAsIn vuh kar sakA"

Figure: F-structure for *yAsIn vuh kar sakA* ‘Yassin could do that.’
Modality in the additional layer for the HUTB

```
pred(root,sak)
subj(sak,yAsIn)
xcomp(sak,kar)
subj(kar,yAsIn)
obj(kar,vuh)
modality(sak,can)
```

**Figure:** Proposed set of triples for modals
Modality in the additional layer for the HUTB

\[
pred(root, sak) \\
subj(sak, yAsIn) \\
xcomp(sak, kar) \\
subj(kar, yAsIn) \\
obj(kar, vuh) \\
modality(sak, can)
\]

**Figure:** Proposed set of triples for modals

\[\rightarrow \text{attribute-value pair } \text{[MODALITY can]} \text{ retained to form a part of the set of triples for this sentence}\]
Tense/aspect in Hindi/Urdu

- elaborate system of expressing different temporal and aspectual notions in Hindi/Urdu
  - e.g. progression, continuation, habituality, iteration, perfective, imperfective
- no differentiation between these concepts in the HUTB
- temporal auxiliaries get VAUX, main verbs get VM
Tense/aspect in Hindi/Urdu

- elaborate system of expressing different temporal and aspectual notions in Hindi/Urdu
  - e.g. progression, continuation, habituality, iteration, perfective, imperfective
- no differentiation between these concepts in the HUTB
- temporal auxiliaries get VAUX, main verbs get VM
- we propose: triples information on tense, aspect and mood
Tense/aspect in the HUTB

Lalu Yadav was working.

Lalu Yadav  work.Masc.Sg  do  Prog.Masc.Sg  be.Impf.Masc.Sg
‘Lalu Yadav was working.’
Tense/aspect in the HUTB

lAlu yAdav kAm kar rahA tHA
Lalu Yadav work.Masc.Sg do Prog.Masc.Sg be.Impf.Masc.Sg
‘Lalu Yadav was working.’

[...]
(( VGF
   VM kar
   VAUX rahA
   VAUX tHA
))

Figure: HUTB tense/aspect annotation
Tense/aspect in the Urdu ParGram grammar

"lAlU yAdav kAm kar rahA tHA"

```
[PRED 'kar<[2:lAlU yAdav], [22:kAm]'>
    [PRED 'lAlU yAdav'
        [SUBJ NTYPE NSEM [PROPER [PROPER-TYPE name]]
            [NSYN proper
                2[CASE nom, GEND masc, NUM sg, PERS 3]
            ]
        ]
    ]

[PRED 'kAm'
    [OBJ NTYPE NSEM [COMMON count]
        [NSYN common
            22[CASE nom, GEND masc, NUM sg, PERS 3]
        ]
    ]

[TNS-ASP ASPECT prog, MOOD indicative, TENSE past]
54[CLAUSE-TYPE decl, PASSIVE -, VTYPE main]
```

**Figure:** F-structure for tense/aspect representation
Tense/aspect in the additional layer for the HUTB

subj(kar,lAlU yAdav)
obj(kar,kAm)
agentive(kar,+)
aspect(kar,prog)
tense(kar,past)
mood(kar,indicative)

Figure: Proposed set of triples for tense/aspect
Tense/aspect in the additional layer for the HUTB

subj(kar, lAlU yAdav)
obj(kar, kAm)
agentive(kar, +)
aspect(kar, prog)
tense(kar, past)
mood(kar, indicative)

**Figure:** Proposed set of triples for tense/aspect

- flattened TNS–ASP information from the f-structure
Tense/aspect in the additional layer for the HUTB

subj(kar, lAlU yAdav)
obj(kar, kAm)
agentive(kar, +)
aspect(kar, prog)
tense(kar, past)
mood(kar, indicative)

**Figure:** Proposed set of triples for tense/aspect

- flattened TNS–ASP information from the f-structure
- aspect, tense and mood triples kept in the dependency triples
Outline of the talk

Overview: Hindi/Urdu resources

The Hindi/Urdu Treebank

The Urdu ParGram grammar

An additional annotation layer for the HUTB

Conclusion
Concluding remarks I

- HUTB: resource for Hindi/Urdu aimed at multi-layered treebanking
Concluding remarks I

- HUTB: resource for Hindi/Urdu aimed at multi-layered treebanking
- HUTB can be extended using the Urdu ParGram grammar
Concluding remarks 1

- HUTB: resource for Hindi/Urdu aimed at multi-layered treebanking
- HUTB can be extended using the Urdu ParGram grammar
- Urdu ParGram grammar can provide linguistically informed analyses for several phenomena:
  - modality, tense/aspect/mood, adjunct type, proper noun type, multiword entities ...
Concluding remarks I

- HUTB: resource for Hindi/Urdu aimed at multi-layered treebanking
- HUTB can be extended using the Urdu ParGram grammar
- Urdu ParGram grammar can provide linguistically informed analyses for several phenomena:
  - modality, tense/aspect/mood, adjunct type, proper noun type, multiword entities ... 
- f-structure triples provide detailed functional information not present in any of the HUTB layers
Concluding remarks I

- HUTB: resource for Hindi/Urdu aimed at multi-layered treebanking
- HUTB can be extended using the Urdu ParGram grammar
- Urdu ParGram grammar can provide linguistically informed analyses for several phenomena:
  - modality, tense/aspect/mood, adjunct type, proper noun type, multiword entities ...
- f-structure triples provide detailed functional information not present in any of the HUTB layers
- triples on a separate layer of HUTB enhance its usability and information content
Concluding remarks II

- short-term goals (current work):
  - randomly select HUTB sentences
  - automatically add triples annotation layer
  - manually check result
Concluding remarks II

- short-term goals (current work):
  - randomly select HUTB sentences
  - automatically add triples annotation layer
  - manually check result

- ultimate target: provide dependency annotation layer similar to PARC700
  - additional bonus: parsers for Hindi and parsers for Urdu may be evaluated against a single gold standard
  - HUTB provides gold standard for both languages
  - long-term goal: generally improve training of linguistically informed parsers for Hindi and Urdu
This research was funded by the Deutsche Forschungsgemeinschaft (DFG).

Thank you for your attention!
References


