URDU TREEBANKS

Dr. Tafseer Ahmed
DHA Suffa University, Karachi

Presentation Plans

- Representation/Modeling schemes

- **Konstanz Urdu Treebank**
- Hindi-Urdu Treebank

- Dependency Structures

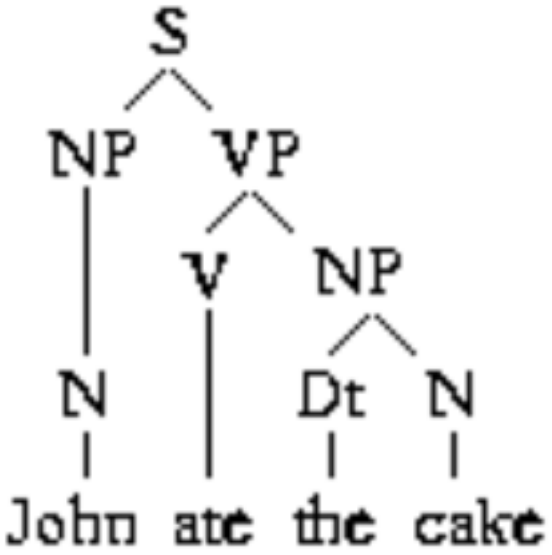
Syntactic Representation

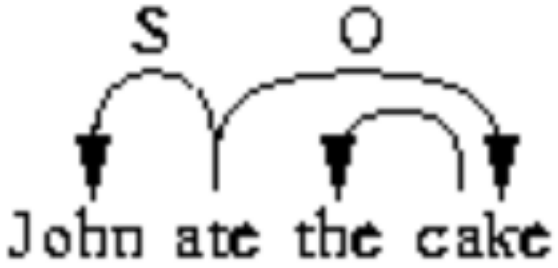
Phrase Structure

vs

Dependency Structure

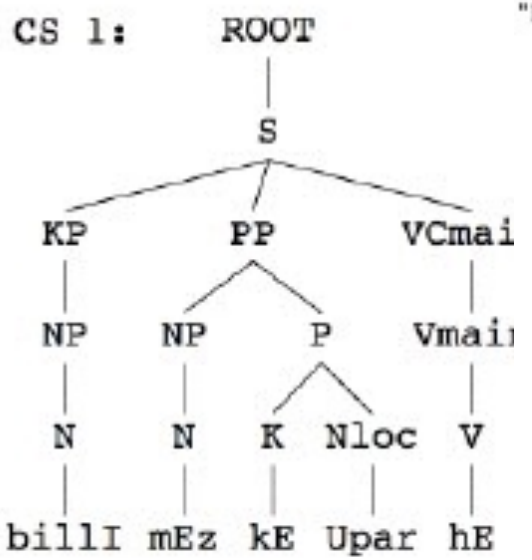
Phrase Structure vs Dependency





Constituent and Functional Structures

- Lexical Functional Grammar (LFG)



"billI mEz kE Upar hE"

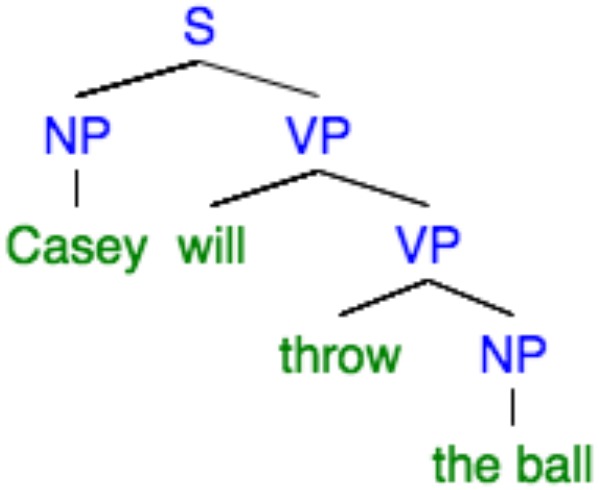
```
PRED       'hO<[1:bill], [35:Upar]>'
SUBJ     1 [PRED  'bill'
            NTYPE [NSEM [COMMON count]
                   NSYN common]
            CASE nom, GEND fem, NUM sg, PERS 3]
PREDLINK   [PRED  'Upar<[17:mEz]>'
            OBJ  [PRED  'mEz'
                  CHECK [_NMORPH obl]
                  NTYPE [NSEM [COMMON count]
                         NSYN common]
           17     GEND fem, NUM sg, PERS 3]
            CHECK [_NMORPH obl]
            SPEC  [POSS [CASE gen]]
           35 GEND masc, NUM sg, PSEM loc, PTYPE sem]
CHECK      [_VMORPH    [_MTYPE infl]]
           [_RESTRICTED-]
TNS-ASP    ASPECT impf, MOOD indicative, TENSE pres
39 CLAUSE-TYPE decl, PASSIVE -, VTYPE copular
```

Important Phrase Structure models

- Penn Treebank
- Universal Multilingual Phrase Labels (Han et. al. 2014)
- Hindi Urdu Treebank (having chunks)

A parse tree example

- (S
  (NP Casey)
  (VP will
  (VP throw
  (NP the ball)
  )))



Konstanz Urdu Treebank

- German Academic Exchange Service (DAAD) funding for collaboration for 2 years (an extension for 1 year is applied.)

- Collaborators
  - Dr. Miriam Butt, University of Konstanz, Germany
  - Dr. Sarmad Hussain, Centre of Language Engineering, KICS, UET, Lahore
  - Dr. Tafseer Ahmed, DHA Suffa University, Karachi

Layers of Annotation

(S (PP-SUBJ:Agent us nE)
(PP-OBJ:Theme aik kitAb)
(VC kharIdI thI))

Syntactic_Phrase_Label-Grammatical_Function:Semantic_Role

Layer 1: Syntactic Phrase Labels

S, SBAR,

VC,

NP,

AdjP, QP, DMP, ValaP

AdvP,

PP, PrP

X

Inspired by Universal tagset with minor changes

Noun Phrase

Examples:

- (NP kitAb)
- (NP acHcHI kitAb)
- (NP pAncH acHcHI kitAbEN)

Postpositional and Prepositional Phrase (PP and PrP)

- The phrases having **case markers** and **other postpositions** are marked as **PP**.
- The phrases having **prepositions** e.g. sivAE will be marked as **PrP**.
- Examples

(PP tum nE)

(PP gHar kA)

(PP gHar tak)

(PrP sivAE is kE)

Verb Complex (VC) - examples

- (S (NP vuh) (NP kitAb) **(VC parH rahI hE)**)
- (S ……. **(VC gir gayA tHA)**)
- (S (NP vuh) (NP sabaq) (NP yAd) **(VC kar rahA tHA)**)

  noun part of noun+verb complex predicate is not a part of verb complex

- (S ……. **(VC gir saktA hE)**)

Discontinuous Phrases

- (S (NP-SUBJ ye)
- **(VC#1 rO)**
- (ADVP kiyoN)
- **(VC#1 rahA hay))**

- (S (NP-SUBJ ye)
- (ADVP kiyoN)
- **(VC rO rahA hay))**

Layer 2&3: Function Tags

Penn Treebank uses function tags.

(S (NP**-SBJ** He)
(VP left
(NP**-TMP** yesterday)))

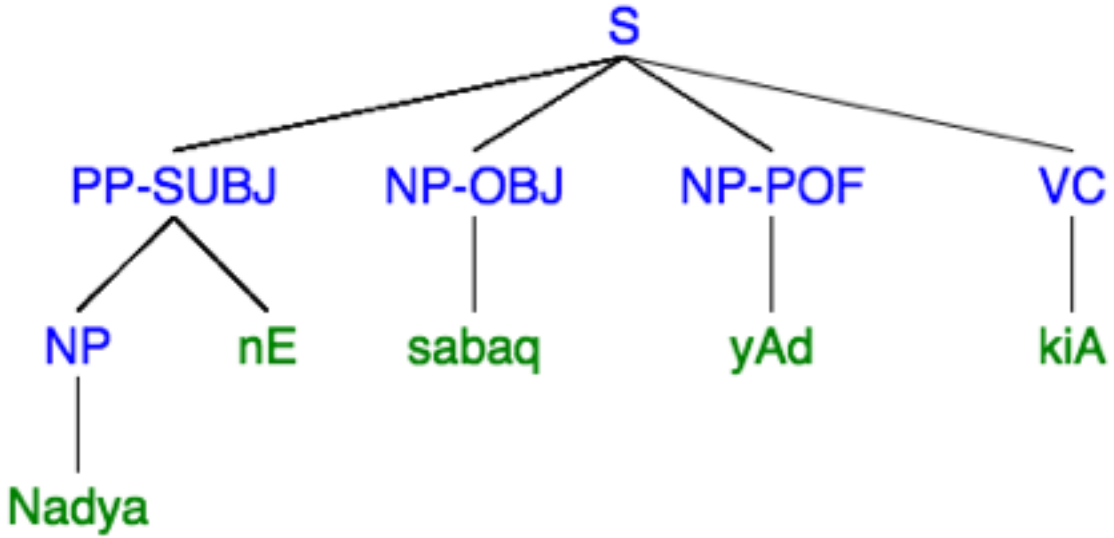
Hindi Urdu Treebank uses semantic/karaka roles as labels.

Grammatical Functions

- SUBJect
- OBJect
- OBLique
- ADJunct
- PreDicate Link
- Part Of Function
- InterJectioN

Part Of Function

- (S
- (PP-SUBJ (NP Nadya) nE)
- (NP sabaq)
- **(NP-POF yAd)**
- (VC kiA)



PreDicate Link

- Copular Constructions

- (S (NP-SUBJ laRkI) **(ADJP-PDL aqalmand)** (VC hE))

- (S (NP-SUBJ laRkA) **(PP-PDL (NP daftar) mEN)**
- (VC hE))

- (S (NP-SUBJ vuh) **(NP-PDL sadar)** (vC ban gayA))

Layer 3: Semantic Roles

- Semantic Roles as attributes of functions
- Propbank roles (Kingsbury & Palmer 2002) used as starting point

Propbank

- **Arg0** Prototypical agent, actor, experiencer
- **Arg1** Prototypical patient, theme
- **Arg2** Beneficiary, receiver
- **Arg3** Instrument
- **Argm** modifiers

انہوں نے جلدی سے کنوئیں سے پانی نکالا

Arg0: انہوں نے

Argm-mnr: جلدی سے

Arg2-sou: کنوئیں سے

Arg1: پانی

Rel: نکالا

Some Examples

(S (NP-SUBJ:A0 vuh)
(NP-OBL:A1 gHar)
(VC ponhcHI))

(S (PP-SUBJ:A0 us nE)
(PP-OBL:A1 mujH par)
(NP-POF bHarosA)
(VC kiA))

Function Tag Examples

- **Adjuncts (-ADJ)**
  :TMP, :LOC, :DIR, :EXT, :MNR, :BEN

Example:

(S (PP-SUBJ:A0 laRkI nE)
(NP-ADJ:TMP kal)
(PP-ADJ:MNR tEzI sE)
(NP-OBJ:A1 gARI)
(VC cHalAI))

Hindi Urdu Treebank (HUTB)

- a multi-representational and multi-layered treebank for Hindi and Urdu.

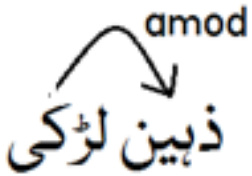
- University of Colorado Boulder
- Columbia University
- University of Massachusetts at Amherst (UMass)
- University of Washington (UW)
- International Institute of Information Technology (IIIT) in Hyderabad, India

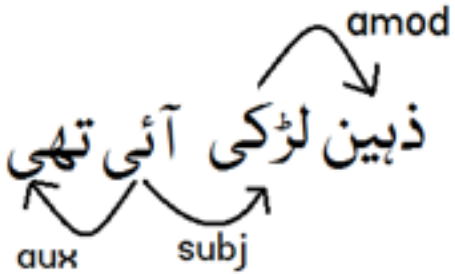
HUTB Representations

- **Dependency structure:** Paninian grammar (Panini 400 BC)
- **Phrase structure:** variant of Minimalism (Chomsky 1995)
- **Propbank:** semantic roles (Kingsbury and Palmer, 2003)

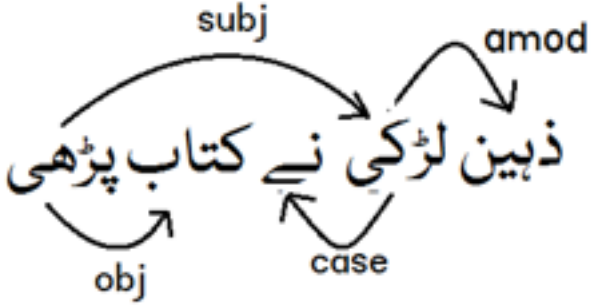
Dependency Structures

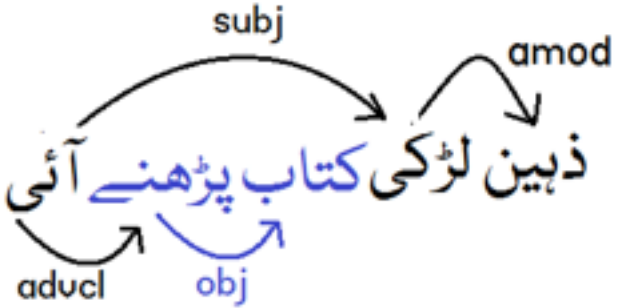
Dependency Structures





DS - More Examples





“Universal” Dependencies

Core dependents of clausal predicates

Nominal dep	*Predicate dep*	
nsubj	csubj	
nsubjpass	csubjpass	
dobj	ccomp	xcomp
iobj		

Non-core dependents of clausal predicates

Nominal dep	*Predicate dep*	*Modifier word*
nmod	advcl	advmod
		neg

Special clausal dependents

Nominal dep	*Auxiliary*	*Other*
vocative	aux	mark
discourse	auxpass	punct
expl	cop	

Noun dependents

Nominal dep	*Predicate dep*	*Modifier word*
nummod	acl	amod
appos		det
nmod		neg

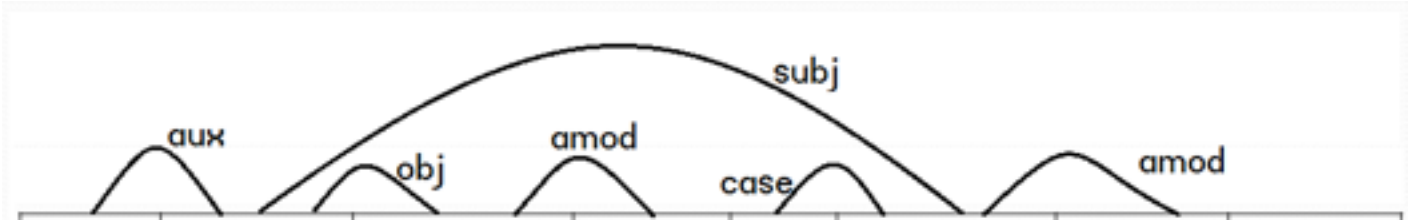
Compounding and unanalyzed

compound	mwe	goeswith
name	foreign	

Coordination

conj	cc	punct

An Example



تھیں	پڑھیں	کتابیں	اچھی	نے	لڑکیوں	ذہین	
ہے	پڑھ	کتاب	اچھا	نے	لڑکی	ذہین	**Lemma**
Aux	Verb	NN	Adj	AdP	Noun	Adj	**POS**
	Form= Perf Gend= Fem Pers=3 Num=Pl	Gend=Fem Num=Pl Form=Nom	Gend= Fem Num= Pl Form= Nom		Gend= Fem Num=Pl Form=Obl	Gend= Fem Num=Pl Form= Obl	**Features**

CoNLL Format

- CoNLL (Conference on Natural Language Learning) format
- Representing graph (and other tags) in text file

Id
Word
Lemma
Coarse Grained POS
Fine Grained POS
Features
Host
Dependency Type

CoNLL Format