

# Tagset and Tagging Urdu Corpus

By Atif Gulzar

National University of Computer and Emerging Sciences, Pakistan

# Outline

---

- Urdu Parts of Speech
- Designing Tagset for Urdu
- Syntactic POS Tagset for Urdu
- Challenges in Tagging Urdu Text
- Tagging Quality and Assurance

## Urdu Parts of Speech (cont.)

- Annotation can be done primarily at five levels: phonology, morphology, syntax, semantics, and pragmatics
- The level of annotation describes the type of information and number of tags required for it
- Researchers has proposed different number of tagsets for Urdu ranging from 10 (Schmidt, 1999) to 350 (Hardie, 2003)

# Designing Tagset for Urdu

- It should be comprehensive
  - Fineness vs. Coarseness
- What should be the starting point?
  - Penn Treebank tagset is used as a baseline tagset
- Syntactic tagset for Urdu

# Basic Assumptions

- It is decided that tags will not be subcategorized for finer distinctions
- However, some basic tags are subcategorized if the following assumptions are inclusively true
  - sub-tags does not complicated the manual tagging process and in some cases these adds the manual tagging
  - these can carry additional linguistic information
  - these can easily be merged to single tag if needed

## Syntactic POS Tagset for Urdu (cont.)

- It consists of 46 tags divided into 15 main groups

<b>Nouns</b>		
	Noun	جہاز، زمین، درخت، لڑکا
	Prepositional noun	باہر، اندر، نیچے، اوپر
	Proper noun	لاہور، احمد
	Proper noun continue	ساجد شوگر مل
	Combined Noun	موسم <NNC> گرما <NNCR>
	Combined Noun continue	
<b>Pronouns</b>		
	Pronoun	میں، ہم، تم، آپ، یہ
	Reflexive pronoun	خود، آپ
	Relative pronoun	جو، جن، جنہوں
	Possessive Pronoun	میرا، تمہارا، ہمارا
	Possessive reflexive Pronoun	اپنا

## Syntactic POS Tagset for Urdu (cont.)

<b>Demonstratives</b>		
	Demonstrative	ہم، تم، آپ، یہ
	Relative demonstrative	جو، جن، جن
<b>Verbs</b>		
	Verb	لکھنا، کھانا، جانا
	Light Verb	صاف کیا، بسر کیا
	Infinitive Verb	تیرنا
	Infinitive light verb	صاف کرنا، بسر کرنا
	Verb to be	ہے، ہوں، ، ہوا، تھا
	Aspectual auxiliary	رہا، کرنا، چکا
	Tense auxiliary	ہے، ہیں، ہوں، تھا، تھے، تھیں
<b>Adverb</b>		
		تقریباً، آہستہ، نہیں، مت
<b>Adjective</b>		
		ظالم، خوبصورت، کمزور

## Syntactic POS Tagset for Urdu (cont.)

<b>Measurements</b>		
	Quantifier	کچھ ، چند،تمام، اتنے ، سب
	Cardinal	ایک، دو، تین، چار بیالیس
	Ordinal	پہلا، دوسرا، تیسرا، چوتھا
	Fractional	چوتھائی، ڈھائی، اڑھائی
	Multiplicative	دگنا، دہرا، تہرا
	Measuring unit	پون، پائو، کلو
	Intensifier	بہت، نہایت، بڑا
<b>Conjunctions</b>		
	Coordinating Conjunction	اور، یا
	Subordinating Conjunction	کہ، کیونکہ
<b>Semantic marker</b>		کا ، کو ، کی، نے
<b>Particles</b>		
	Intensifier Particle	ہی، بھی، تو، نا
	Adjectival particle	سا، سے، سی
	Adverbial Particle	سے

# Syntactic POS Tagset for Urdu

Special		
	KER	کے، کر
	Pre-Mohmil	آمنے سامنے، اردگرد
	Post-Mohmil	پانی وانی، ٹھیک ٹھاک
	WALA	والا، والی، والے
	DATE	2007, 1999
Interjection		واہ، سبحان اللہ، اچھا
Question word		کیا، کیوں، آیا
Punctuation markers		
	Sentence marker	.
	Phrase marker	, ;
	Symbol	%, “, #
Unknown		

## Challenges in Tagging Urdu Text (cont.)

- Proper Nouns

- Syntactically the function of proper noun is same as of noun. However, tagging proper nouns do not require any extra effort while tagging manually and annotated corpus tagged with proper noun can be used for certain application e.g. name entity recognition.

- Compound proper nouns

پاکستان سٹیل کارپوریشن کراچی میں واقع ہے۔  
NNPC NNPC NNPC is  
situated in Karachi corporation Steel Pakistan

“Pakistan Steel Corporation is situated in Karachi”

## Challenges in Tagging Urdu Text (cont.)

- Proper nouns vs. adjectives
  - In Urdu, most of the proper nouns are derived from adjectives.

نواز	شریف	لاہور	میں	رہتا	ہے۔
NNP	NNPC	NNP	CM	VB	AUXT
Nawaz	Shareef	Lahore	in	live	-es

“Nawaz **Shareef** lives in Lahore”

اسلم	شریف	آدمی	ہے۔
NNP	JJ	NN	VBT
Aslam	pious	man	is

“Aslam is a **pious** man”

## Challenges in Tagging Urdu Text (cont.)

- Noun vs. proper noun
  - In Urdu, unlike English there is no clear orthographic distinction of proper nouns
    - Only those words are tagged Proper Noun where there is no confusion
- Noun vs. adjective
  - In Urdu inflected forms of adjective behave as noun

ہمیں بڑوں کا ادب کرنا چاہیے۔

AUXA VBL NN CM NN PR

Should respect elders we

“We should respect **elders**”

## Challenges in Tagging Urdu Text (cont.)

- Noun vs. adjective

ہے تاجر ہے  
VBT NN PR  
is businessman he  
“He is businessman”

ہے تاجر شخص ہے  
VBT JJ NN PD  
is businessman man this  
“This man is businessman”

## Challenges in Tagging Urdu Text (cont.)

- Prepositional noun

- In Urdu, some time temporal and special nouns also behave as preposition

وہ کرسی کے اوپر بیٹھا ہے۔  
AUXT VB **NNCM** CM NN PR  
is sitting **on** chair he  
“His is sitting **on** chair”

وہ اوپر کو گیا ہے۔  
AUXT VB CM **NN** PR  
has gone **upstairs** he  
“He has gone **upstairs**”

## Challenges in Tagging Urdu Text (cont.)

- Light Verb
  - These verbs make compound verbs by combining a noun or adjective and a verb

اس نے کمرہ صاف کیا۔

VBL JJ NN CM PR

clean room He

“He cleaned the room”

## Challenges in Tagging Urdu Text (cont.)

- KER tag
  - This is a special word that joins two verb phrases and shows the completion of 1<sup>st</sup> verb

وہ کھانا کھا کر چلا گیا۔

AUXA VB **KER** VB NN PR

-ed leave after eat meal he

“He left after eating meal”

## Challenges in Tagging Urdu Text (cont.)

- WALA tag
  - WALA (والا) has varied used in Urdu. It modifies and combines with the phrase to give an extended meaning as a noun phrase or adjective phrase

دودھ والی قلفی میٹھی ہے۔  
BT JJ NN **WALA** NN  
is sweet ice cream 'made of' milk  
“The ice cream made of milk is sweet”

دودھ والا آیا ہے۔  
AUXT VB **WALA** NN  
has come milkman  
“Milkman has come”

## Challenges in Tagging Urdu Text

- Mohmils
  - Mohmils are such words that do not carry their own meaning
  - These rhymes with the preceding or following word
  - Semantically they give a sense of etcetera

وہ کھانا شانا کھا رہا ہے۔  
PR NN **MOPO** VB AUXA AUXT  
he meal **etc** eat -ing is  
“He is eating meal (and etc.).”

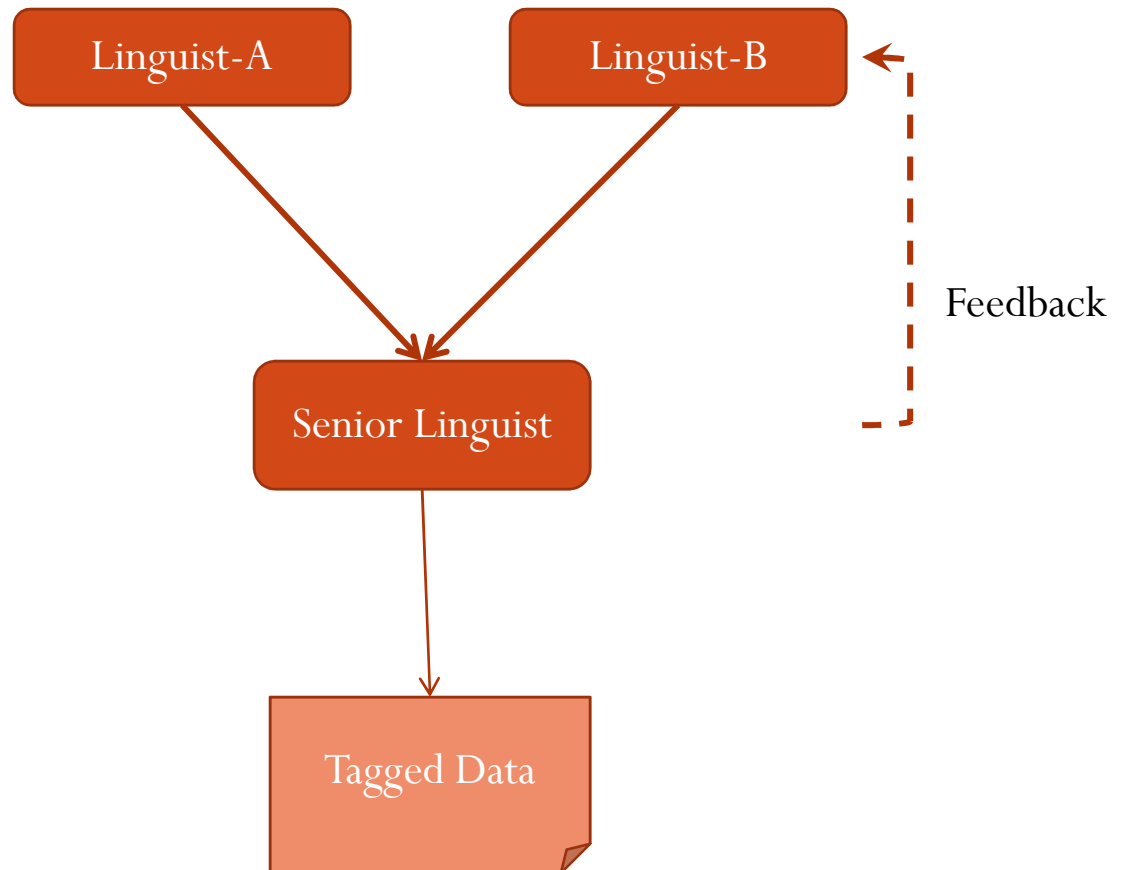
## Tagset Testing and Quality Assurance (cont.)

- Three Urdu linguists are hired for tagging
- A sample corpus of same 5,000 words is given to each linguist
- The results and differences are then discussed and tagging guidelines are designed to ensure the consistency in tagged data and as reference for linguists

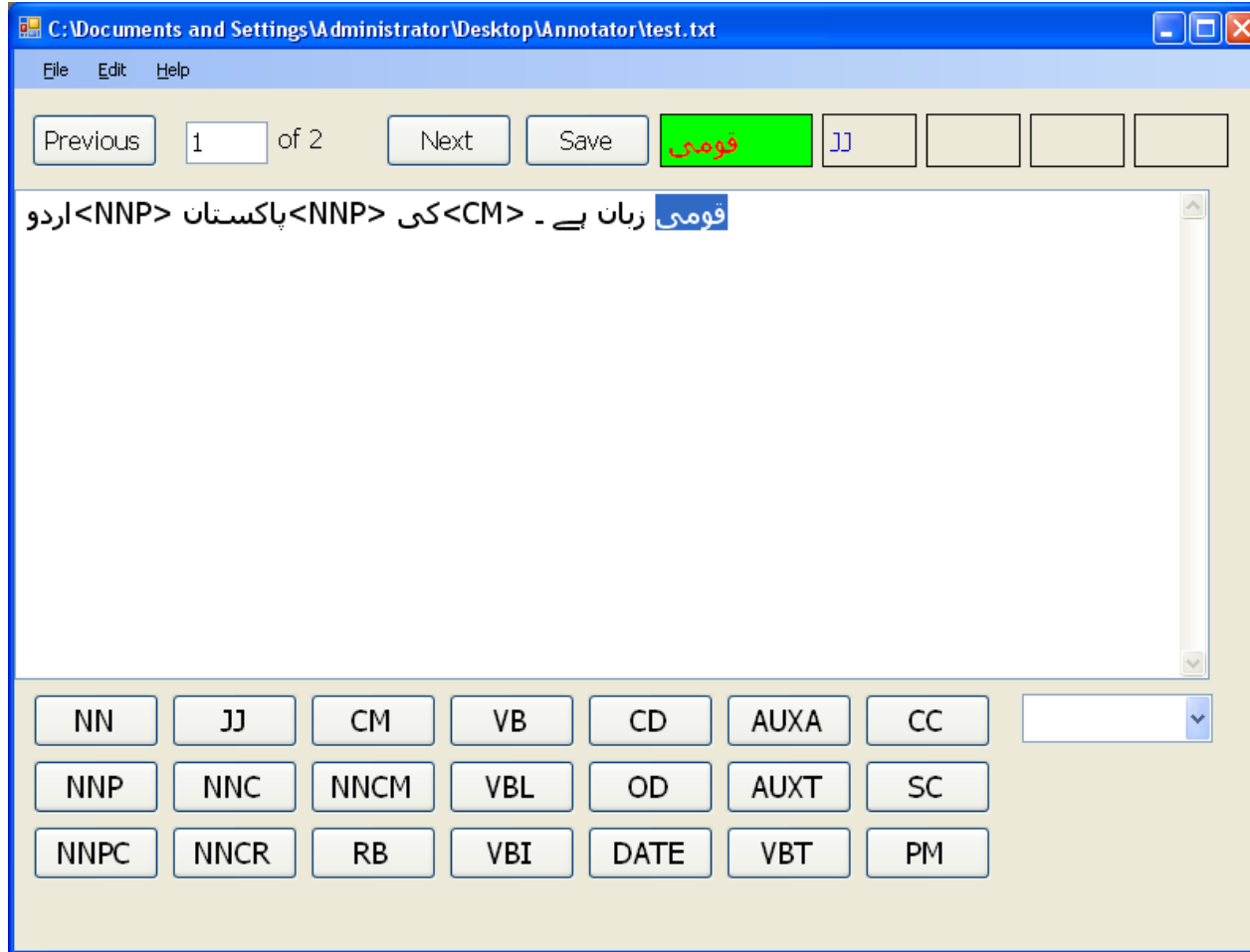
# Tagset Testing and Quality Assurance

Daily task of 1000 words is assigned to each linguist

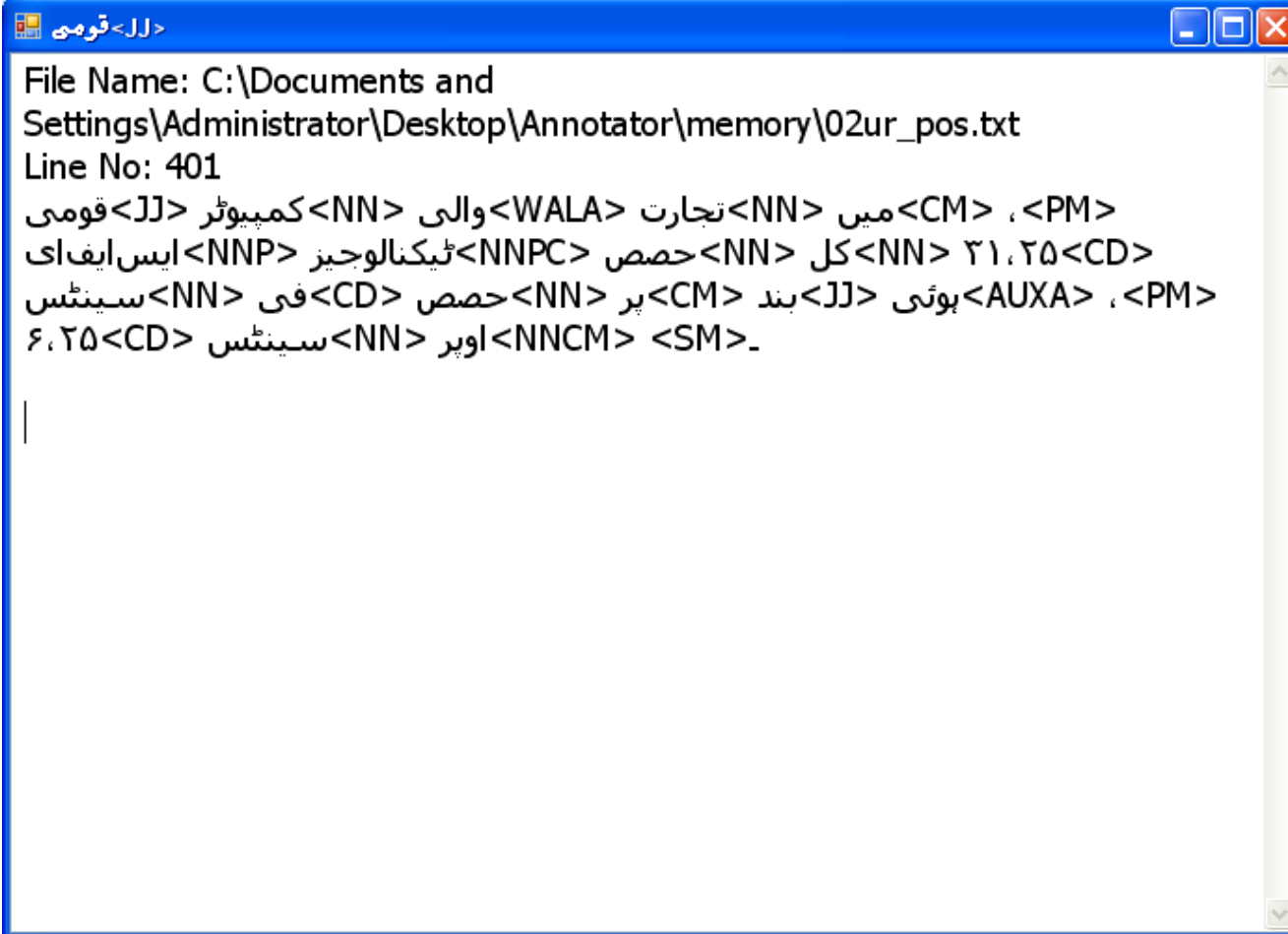
Translation is reviewed by senior Linguist



# Tagging Tool



# Tagging Tool



File Name: C:\Documents and Settings\Administrator\Desktop\Annotator\memory\02ur\_pos.txt  
Line No: 401  
<PM> ، <CM> میں <NN> تجارت <WALA> والی <NN> کمپیوٹر <JJ> قومی  
<CD> ۲۱، ۲۵ <NN> کل <NN> حصص <NNPC> ٹیکنالوجیز <NNP> ایس ایف ای  
<PM> ، <AUXA> ہوئی <JJ> بند <CM> پر <NN> حصص <CD> فی <NN> سینٹس  
<SM>۔ <NNCM> اوپر <NN> سینٹس <CD> ۶، ۲۵

# Thanks

---

The corpus is available at  
[http://crulp.org/software/ling\\_resources/UrduNepaliEnglishParallelCorpus.htm](http://crulp.org/software/ling_resources/UrduNepaliEnglishParallelCorpus.htm)