CLPL5: Part of Speech Tagset

Tafseer Ahmed

Part of speech

- a **part of speech** (also a word class, a lexical class, or a lexical category) is a linguistic category of words (or more precisely lexical items), which is generally defined by the syntactic or morphological behaviour of the lexical item in question.

John	saw	the	saw
Noun	Verb	Article	
Noun			

“Traditional” POS Tag set

- Noun
- Pronoun
- Adjective
- Verb
- Adverb
- Preposition
- Conjunction
- Interjection

Tag set sizes

- 3 Tags: اسم ، فعل، حرف : Arabic (“Tradition”)
- 8 Tags: English (“Tradition”)
- 48 Tags :Penn Treebank Tagset
- 282 Tags: Hardie’s Urdu Tagset
- 35 Tags: CLE Urdu Tagset

Granularity Problem

- **Adjective:**
  - good, bad, اچھا ، برا — **adjective**
  - some, many, چند ، کئی — **quantifier**
  - first, second, پہلا ، دوسرا — **ordinal**

- **Verb:**
  - go, read, جا، پڑھ — **main verb**
  - is, was, ہے ، تھا — **helping verb / auxiliary**
  - can, may, سک، چاہیے — **modal verb**

POS tagset for Computation

Table 2
The Penn Treebank POS tagset.

1. CC	Coordinating conjunction	25. TO	*to*
2. CD	Cardinal number	26. UH	Interjection
3. DT	Determiner	27. VB	Verb, base form
4. EX	Existential *there*	28. VBD	Verb, past tense
5. FW	Foreign word	29. VBG	Verb, gerund/present participle
6. IN	Preposition/subordinating conjunction	30. VBN	Verb, past participle
7. JJ	Adjective	31. VBP	Verb, non-3rd ps. sing. present
8. JJR	Adjective, comparative	32. VBZ	Verb, 3rd ps. sing. present
9. JJS	Adjective, superlative	33. WDT	*wh*-determiner
10. LS	List item marker	34. WP	*wh*-pronoun
11. MD	Modal	35. WP$	Possessive *wh*-pronoun
12. NN	Noun, singular or mass	36. WRB	*wh*-adverb
13. NNS	Noun, plural	37. #	Pound sign
14. NNP	Proper noun, singular	38. $	Dollar sign
15. NNPS	Proper noun, plural	39. .	Sentence-final punctuation
16. PDT	Predeterminer	40. ,	Comma
17. POS	Possessive ending	41. :	Colon, semi-colon
18. PRP	Personal pronoun	42. (	Left bracket character
19. PP$	Possessive pronoun	43.)	Right bracket character
20. RB	Adverb	44. "	Straight double quote
21. RBR	Adverb, comparative	45. ‘	Left open single quote
22. RBS	Adverb, superlative	46. “	Left open double quote
23. RP	Particle	47. ’	Right close single quote
24. SYM	Symbol (mathematical or scientific)	48. ”	Right close double quote

Sample Text - English

one/CD **charge**/NN **of**/IN **filing**/NN **a**/DT **false**/JJ **return**/N **and**/CC **was**/VBD **fined**/VBN **$**/$ **5,000**/CD **and**/CC **sentenced**/VBN **to**/TO **18**/CD **months**/NNS **in**/IN **prison**/NN **./.**

Urdu Tagset - Issues

- Coarse grained tags
  - better for machine learning
  - easy / fast annotation
- Fine grained tags
  - more (morphological) information

Urdu Tagset - Issues

- Granularity
  - Hardie (282 tags)

NNUM1N	بھائی	Common unmarked masculine singular nominative noun
NNUM1O	بھائی	Common unmarked masculine singular oblique noun
NNUM1V	بھائی	Common unmarked masculine singular vocative noun
NNUM2N	بھائی	Common unmarked masculine plural nominative noun
NNUM2O	بھائیوں	Common unmarked masculine plural oblique noun

Urdu Tagset - Issues

- Syntactic versus functional behavior
  - Noun (NN)

میں نے پانی پیا۔

vs.

میں نے سبق یاد کیا۔

vs.

میں گھر کے اندر گیا۔

Urdu Tagsets

- http://www.cle.org.pk/Downloads/langproc/UrduPOStagger/Urdu%20POS%20Tagset%200.3.pdf
- http://verbs.colorado.edu/hindiurdu/guidelines_docs/Chunk-POS-Annotaion-Guidelines.doc
- https://universaldependencies.org/u/pos/index.html

Sample Text - Urdu

CLE POS Tagged Data

دنیا/NN کا/PSP ہر/JJ فرد/NN کامیابی/NN کا/PSP آرزومند/NN ہے/VBF
ناکامی/NN سے/PSP سب/JJ گھبراتے/VBF ہیں/PU. AUXT/ عزت/PU/، NN دولت/PU/، NN
راحت/NN اور/CC عافیت/NN کی/PSP زندگی/NN کے/PSP سبھی/PRP شیدائی/NN ہیں/VBF
/PU. لیکن/SC اصل/JJ کامیابی/NN کیا/RB چیز/NN ہے/VBF ؟/PU اور/CC حقیقی/JJ
عزت/NN و/CC راحت/NN کس/PDM طرح/NN نصیب/NN ہوتی/VBF ہے/AUXT ؟/PU اس/PDM
بھید/NN سے/PSP بہت/Q کم/Q لوگ/NN واقف/NN ہیں/VBF ./PU اگر/SCP آپ/PRP
حقیقی/JJ کامیابی/NN کے/PSP گُر/NN جاننا/VBI چاہتے/VBF ہیں/AUXT تو/SC
ڈاکٹر/NN زاہد/NNP منیر/NNP عامر/NNP کی/PSP تازہ/JJ تصنیف/PU/' NN آئینہ/NN
کردار/PU/' NN پڑھیے/VBF ./CD ۱۱۲/PU صفحوں/NN کی/PSP اس/PDM کتاب/NN کا/PSP
ایک/CD ایک/CD حرف/NN بصیرت/NN کے/PSP دریچے/NN کھولنے/VBI پر/PSP مامور/NN
/PU. VBF/ہے

“Universal” Tagset

- ADJ: adjective
- ADP: adposition
- ADV: adverb
- AUX: auxiliary verb
- CONJ: coordinating conjunction
- DET: determiner
- INTJ: interjection
- NOUN: noun
- NUM: numeral
- PART: particle
- PRON: pronoun
- PROPN: proper noun
- PUNCT: punctuation
- SCONJ: subordinating conjunction
- SYM: symbol
- VERB: verb
- X: other

Universal Tagset

- NOUN (nouns) کتاب، شہر
- PROPN (proper noun) کراچی، کامران
- VERB (verbs) پڑھتا, لکھتی
- AUX (auxiliary) ہے ، رہا
- ADJ (adjectives) اچھا ، چند، پہلا
- ADV (adverbs) روزانہ، بہت، تقریباً

“Universal” Tagset

- PRON (pronouns) میں، تم، وہ
- DET (determiners and articles) وہ
- ADP (prepositions and postpositions) نے ، اندر
- NUM (numerals) 2, دو
- CONJ (conjunctions) اور، یا، لیکن
- ‘.’ (punctuation marks) -، ؟
- X (a catch-all for other categories)

Morphological Information

Not Secure | lindat.mff.cuni.cz/services/udpipe/

Id	Form	Lemma	UPosTag	XPosTag	Feats
# newdoc					
# newpar					
# sent_id = 1					
# text = لڑکیاں لائبریری میں کتابیں پڑھ رہی تھیں۔					
1	لڑکیاں	لڑکی	NOUN	NN	Case=Nom|Gender=Fem|Number=Plur|Person=3
2	لائبریری	لایبریری	NOUN	NN	Case=Acc|Gender=Fem|Number=Sing|Person=3
3	میں	میں	ADP	PSP	AdpType=Post
4	کتابیں	کتاب	NOUN	NN	Case=Nom|Gender=Masc|Number=Plur|Person=3
5	پڑھ	پڑھ	VERB	VM	Voice=Act
6	رہی	رہ	AUX	VAUX	Aspect=Perf|Gender=Fem|Number=Sing|VerbForm=Part
7	تھیں	تھا	AUX	VAUX	Gender=Fem|Mood=Ind|Number=Plur|Person=3|Tense=Past|VerbForm=Fin