Challenges of Processing South Asian Languages (CPSAL)

Kengatharaiyer **Sarves**waran
University of Konstanz, Germany
University of Jaffna, Sri Lanka.

Tafseer Ahmed
Senior NLP Scientist
Alexa Translations

Course outline

- **Topics (Tentative):**
  - Day 01: Languages, Scripts, and Encoding of South Asian Languages.
  - Day 02: Phonology, Transliteration and Morphology of South Asian Languages.
  - **Day 03: More on Morphology, Part of Speech and Multi-word tokenisation**
  - Day 04: Syntax, Morphosyntax, and Semantics of South Asian Languages.
  - Day 05: Deep Learning for South Asian Languages and winding up the course.

Pronominal suffixes

Topics

- POS tagsets, tagging, and Challenges
- Multi-word tokenisation
- The Universal Dependencies

Part of Speech tagging

- **Part-of-speech tagging is the process of assigning a part-of-speech to each word in a text.**
- **The (ART) quick (ADJ) brown (ADJ) fox (N) jumps (V) over (PREP) the (ART) lazy (ADJ) dog (N).**
- **Tagging involves a lot of disambiguation**
  - words are ambiguous - need the context to disambiguate.
- **Tagsets**
  - How much information can or should one capture using POS?
  - Flat vs Hierarchical tagsets

Tagsets

Part of Speech tagsets

- Language dependent, mostly.
- The Penn POS tagset.

Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coordinating conjunction	*and, but, or*	PDT	predeterminer	*all, both*	VBP	verb non-3sg present	*eat*
CD	cardinal number	*one, two*	POS	possessive ending	*'s*	VBZ	verb 3sg pres	*eats*
DT	determiner	*a, the*	PRP	personal pronoun	*I, you, he*	WDT	wh-determ.	*which, that*
EX	existential 'there'	*there*	PRP$	possess. pronoun	*your, one's*	WP	wh-pronoun	*what, who*
FW	foreign word	*mea culpa*	RB	adverb	*quickly*	WP$	wh-possess.	*whose*
IN	preposition/ subordin-conj	*of, in, by*	RBR	comparative adverb	*faster*	WRB	wh-adverb	*how, where*
JJ	adjective	*yellow*	RBS	superlatv. adverb	*fastest*	$	dollar sign	$
JJR	comparative adj	*bigger*	RP	particle	*up, off*	#	pound sign	#
JJS	superlative adj	*wildest*	SYM	symbol	+,%, &	“	left quote	‘ or “
LS	list item marker	*1, 2, One*	TO	“to”	*to*	”	right quote	’ or ”
MD	modal	*can, should*	UH	interjection	*ah, oops*	(	left paren	[, (, {, <
NN	sing or mass noun	*llama*	VB	verb base form	*eat*	)	right paren	],), }, >
NNS	noun, plural	*llamas*	VBD	verb past tense	*ate*	,	comma	,
NNP	proper noun, sing.	*IBM*	VBG	verb gerund	*eating*	.	sent-end punc	. ! ?
NNPS	proper noun, plu.	*Carolinas*	VBN	verb past part.	*eaten*	:	sent-mid punc	: ; ... – -

POS tagsets - CLE Urdu tagset

Categories	Types	POS Tag
1. Noun	1.1 Common	NN
	1.2 Proper	NNP
2. Verb	2.1 Main Verb Infinitive	VBI
	2.2 Main Verb Finite	VBF
3. Auxiliary	3.1 Aspectual	AUXA
	3.2 Progressive	AUXP
	3.3 Tense	AUXT
	3.4 Modals	AUXM
4. Pronoun	4.1 Personal	PRP
	4.2 Demonstrative	PDM
	4.3 Possessive	PRS
	4.4 Relative Demonstrative	PRD
	4.5 Relative Personal	PRR
	4.6 Reflexive	PRF
	4.7 Reflexive Apna	APNA
5. Nominal Modifier	5.1 Adjective	JJ
	5.2 Quantifier	Q
	5.3 Cardinal	CD
	5.4 Ordinal	OD
	5.5 Fraction	FR
	5.6 Multiplicative	QM
6. Adverb	6.1 Common	RB
	6.2 Negation	NEG
7. Adposition	7.1: Preposition	PRE
	7.2:Postposition	PSP
8. Conjunction	8.1 Coordinate Conjunction	CC
	8.2 Subordinate Conjunction	SC
	8.3 SCKar	SCK
	8.4 Pre-sentential	SCP
9. Interjection	9.1 Interjection	INJ
10. Particle	10.1 Common	PRT
	10.2 Vala	VALA
11. Symbol	11.1 Common	SYM
	11.2 Punctuation	PU
12. Residual	12.1: Foreign Fragment	FF

Tagsets: Unified POS Standard for Indian Languages

Sl. No	Category			Label	Annotation Convention**	Remarks
	Top level	Subtype (level 1)	Subtype (level 2)			
1	Noun			N	N	
1.1		Common		NN	N_NN	
1.2		Proper		NNP	N_NNP	
1.3		Verbal		NNV	N_NNV	The verbal noun sub type is only for languages such as Tamil and Malayalam)
1.4		Nloc		NST	N_NST	
2	Pronoun			PR	PR	
2.1		Personal		PRP	PR_PRP	
2.2		Reflexive		PRF	PR_PRF	
2.3		Relative		PRL	PR_PRL	
2.4		Reciprocal		PRC	PR_PRC	
2.5		Wh-word		PRQ	PR_PRQ	
2.6		INDEFINITE		PRI	PR_PRI	
3	Demonstrative			DM	DM	
3.1		Deictic		DMD	DM_DMD	
3.2		Relative		DMR	DM_DMR	
3.3		Wh-word		DMQ	DM_DMQ	
3.4		Indefinite		DMI	DM_DMI	
4	Verb			V	V	
4.1		Main		VM	V_VM	
4.1.1			Finite	VF	V_VM_VF	
4.1.2			Non-finite	VNF	V_VM_VNF	
4.1.3			Infinitive	VINF	V_VM_VINF	
4.1.4			Gerund	VNG	V_VM_VNG	
4.2		Verbal		VN	V__VN	
4.2.1			Finite	VAUX	V_VAUX_VF	
4.2.2			Non-finite	VNF	V_VAUX_VNF	
4.2.3			Infinitive	VINF	V_VAUX_VINF	
4.2.4			Gerund	VNG	V_VAUX_VNG	
4.2.5			PARTICIPLE NOUN	VNP	V_VAUX_VNP	

Tagsets: Unified POS Standard for Indian Languages

5	Adjective			JJ		
6	Adverb			RB		Only manner adverbs
7	Postposition			PSP		
8	Conjunction			CC	CC	
8.1		Co-ordinator		CCD	CC_CCD	
8.2		Subordinator		CCS	CC_CCS	
8.2.1			Quotative	UT	CC_CCS_UT	
9	Particles			RP	RP	
9.1		Default		RPD	RP_RPD	
9.2		Classifier		CL	RP_CL	
9.3		Interjection		INJ	RP_INJ	
9.4		Intensifier		INTF	RP_INTF	
9.5		Negation		NEG	RP_NEG	
10	Quantifiers			QT	QT	
10.1		General		QTF	QT_QTF	
10.2		Cardinals		QTC	QT_QTC	
10.3		Ordinals		QTO	QT_QTO	

11	Residuals			RD	RD	
11.1		Foreign word		RDF	RD_RDF	A word written in script other than the script of the original text
11.2		Symbol		SYM	RD_SYM	For symbols such
11.3		Punctuation		PUNC	RD_PUNC	Only for punctuations
11.4		Unknown		UNK	RD_UNK	
11.5		Echowords		ECH	RD_ECH	

POS tagsets - Amrita tagset (Tamil)

- **Echo words and Reduplication words**

S.N	TAG	DESCRIPTION		S.N	TAG	DESCRIPTION
1	<NN>	NOUN		17	<VINT>	VERB INFINITE
2	<NNC>	COMPOUND NOUN		18	<CNJ>	CONJUNCTION
3	<NNP>	PROPER NOUN		19	<CVB>	CONDITIONAL VERB
4	<NNPC>	COMPOUND PROPER NOUN		20	<QW>	QUESTION WORDS
5	<ORD>	ORDINALS		21	<COM>	COMPLEMENTIZER
6	<CRD>	CARDINALS		22	<NNQ>	QUANTITY NOUN
7	<PRP>	PERSONAL PRONOUN		23	<QTF>	QUANTIFIERS
8	<PRIN>	PRONOUN INTROGATIVE		24	<PPO>	POSTPOSITIONS
9	<PRID>	PRONOUN INDEFINITE		25	<DET>	DETERMINERS
10	<ADJ>	ADJECTIVE		26	<INT>	INTENSIFIER
11	<ADV>	ADVERB		27	<ECH>	ECHO WORDS
12	<VNAJ>	VERB NON FINITE ADJECTIVE		28	<EMP>	EMPHASIS
13	<VNAV>	VERB NON FINITE ADVERB		29	<COMM>	COMMA
14	<VBG>	VERBAL GERUND		30	<DOT>	DOT
15	<VF>	VERB FINITE		31	<QM>	QUESTION MARK
16	<VAX>	VERB AUXILARY		32	<RDW>	REDUPLICATION WORDS

POS tagsets: Universal POS tagset

Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

Reduplication in Tamil

1) Morphological reduplication:
   ovvoṉṟu one by one < *oṉṟu* one[3]
   ivviraṇṭu two by two < *iraṇṭu* two
   vevvēṟu various < *vēṟu* different[4]
2) Semantic reduplication:
   parantu viri- < (to be spread out + to expand, spread out)
   utavi ottācai < (aid, help, assistance + aid, help, assistance)
   alaintu tiri- < (to go to and fro, to roam + to walk about, wander, go here and there)
3) Lexical reduplication:
   teruvukkut teru from street to street < *teru* street
   avacara avacaramākap pōṉār He went in great haste. < *avacaram* haste
   tēmpit tēmpi aḻutāḷ She wept while sobbing again and again. < *tēmpu-* to sob
   marattōṭu maramāka niṉṟāṉ He stood very close to the tree. (Lit. He stood as if he became a tree with a tree).
   muṭiyavē muṭiyātu[5] just not possible < *muṭi-* to be possible
4) Phrasal reduplication:
   talaitūkkit talaitūkkip pārttārkaḷ They were looking lifting up their heads. < *talai tūkku-* to lift up one's head
   āḷ māṟṟi āḷ māṟṟi kāvaliruppārkaḷ They took turns in guarding. < *āḷ māṟṟu-* man change (Lit. People changed people changed guarding).
   nalla paiyaṉ killa paiyaṉ 'good boy and so forth' < (*nalla* good; *paiyaṉ* boy)
5) Phonetic reduplication:
   pēccu mūccu signs of life, consciousness < (ability to talk, speech + breath)
   aṭakkam oṭukkamāka being humble or modest < (humility, submission, subordination + self-control)
   kōṇal māṇal unevennes < (*kōṇal* being crooked, bent, askew + *māṇal* has no meaning by itself)

Reduplication in Urdu

bacce ko	*ek tâfî*	*do*
child-S DAT	one toffee	give

'give a toffee to the child' (definite occurrence)

baccoN ko	*ek-ek tâfî*	*do*
child-P DAT	one-one toffee	give

'give a toffee to each child, one toffee per child'

https://shs.hal.science/halshs-00449691/document

Echo words

- Common among South Asian Languages

Tamil:

kōppi	**kīppi**	**kuṭikkiṟīṅkaḷā?**
He	kīppi	kuṭi-kkiṟ-īṅkaḷ-ā?
coffee	?	drink-PRES-3Pl-QPART
NOUN	?	VERB

“Do you need coffee or something?”

How will a POS tagger deal with it?

Rich morphology

- Morphologically rich languages will have more words!
- POS tagger training require more data / more rules.
- Malayalam is a Dravidian language.

Language	Corpus size	Unique words
English	10 million	97,734
Turkish	10 million	4,17,775
Malayalam	10 million	14,27,392

https://thottingal.in/blog/2019/09/10/bis-pos-tagset-review/

Challenges

Ambiguous lexical categories - A common problem

earnings growth took a **back/JJ** seat
a small building in the **back/NN**
a clear majority of senators **back/VBP** the bill
Dave began to **back/VB** toward the door
enable the country to buy **back/RP** about debt
I was twenty-one **back/RB** then

Ambiguous lexical categories - A common problem

āṇṭāḷ	kaṇṇaṉai	aṭaintāḷ.
āṇṭāḷ	kaṇṇaṉ-ai	reached.
āṇṭāḷ.NOM	kaṇṇaṉ-ACC	reached.
PROPNOUN	PROPNOUN	VERB

“āṇṭāḷ reached Kannan”

araci	nāṭṭai	āṇṭāḷ.
araci	nāṭṭ-ai	āṇ-ṭ-āḷ.
Queen.NOM	country-ACC	
rule-PAST-3SgF.		
NOUN	NOUN	VERB

“The queen ruled the country.”

Challenges - Noun/Verbs in Tamil

paṭittavaḷ	ceyvāḷ
paṭi-tt-avaḷ	cey-v-āḷ
study-PAST-3SgF	do-FUT-3SgF
NOUN?	VERB

“Female who studied (or an educated female) will do”

kaṭai	naṭattupavarkaḷ	paṭittavaṉait	terivuceytaṉar
kaṭai	naṭattu-p-avarkaḷ	paṭi-tt-avaṉ-ait	terivu-cey-t-aṉar
shop	operate-FUT-3Pl	study-PAST-3SgM-ACC	select-do-PAST-3Pl
NOUN	NOUN?	NOUN?	VERB

“Shop operators selected the man who studied (or an educated man)”

Challenges - Demonstratives in Tamil

appai	periyatu
a-ppai	periyatu
DEM-bag	big
NOUN?	NOUN

“That bag is big”

Challenges - Complimentisers in Tamil

- Tamil does not have complementizers of the that-type as in English.
- *enṛu* - a past participle form of *en* “say” is used.
- eṉpatu - a present form of en “say” is used.

[avaṉ	piḻai	ceytāṉ	eṉpatai]	rām	nirūpittāṉ
avaṉ	piḻai	cey-t-āṉ	eṉp-a-atu-ai	rām	nirūpittāṉ
He	mistake	do-PAST-3SgM	COMP-REL-3SgN-ACC	Ram	proved
NOUN	NOUN	VERB	COMP?	NOUN	VERB

‘Ram proved (the fact) that he made mistakes.’

(Butt, 2020)

Confusing categories - Sindhi

Pronominal suffixes in Sindhi

xat-u	likH-iyO-maan-si
Letter.M-Nom.Sg	write-PastPart.Sg-1P.Sg-3P.Sg
I wrote him/her (a) letter.	

Multi-word tokenisation

Multi-word tokenisation

- **How this can be implemented?**
  - Identifying multi-word tokens.
  - Preserving the original word forms.

paṭittavaḷ	**ceyvāḷ**
paṭi-tt-avaḷ	**cey-v-āḷ**
paṭitta + avaḷ	
VERB + NOUN	**VERB**

“Female who studied (or an educated female) will do”

Multi-word tokenisation - extract more important information

- Useful to exact syntactic information

avaṉ vantukoṇṭirukkiṟāṉ
avaṉ vantu-koṇṭ-iru-kkiṟ-āṉ
He come-hold-be-PRES-3SgM
PRON VERB
“He has been coming”

vantukoṇṭirukkiṟāṉ -> vantu VERB
koṇṭu AUX
iru AUX

Multi-word tokenisation - extract more important information

Polar questions:

avaṉa	vārukiṟāṉ
avaṉ-a	vāru-kiṟ-āṉ
he-INTG	come-PRES-3SgM
NOUN?	VERB

“Is he coming?”

avaṉ-a ->	avaṉ	VERB
	a	PART

Pronominal on case markers - Sindhi

هن مون کي ڪتاب ڏنو

Huni	mūN=kHē	kitābu	dinū
he.Obl	1P.Sg.Obl=Dat	book.Sg.M	give.PastPart.M.Sg

He gave me a book.

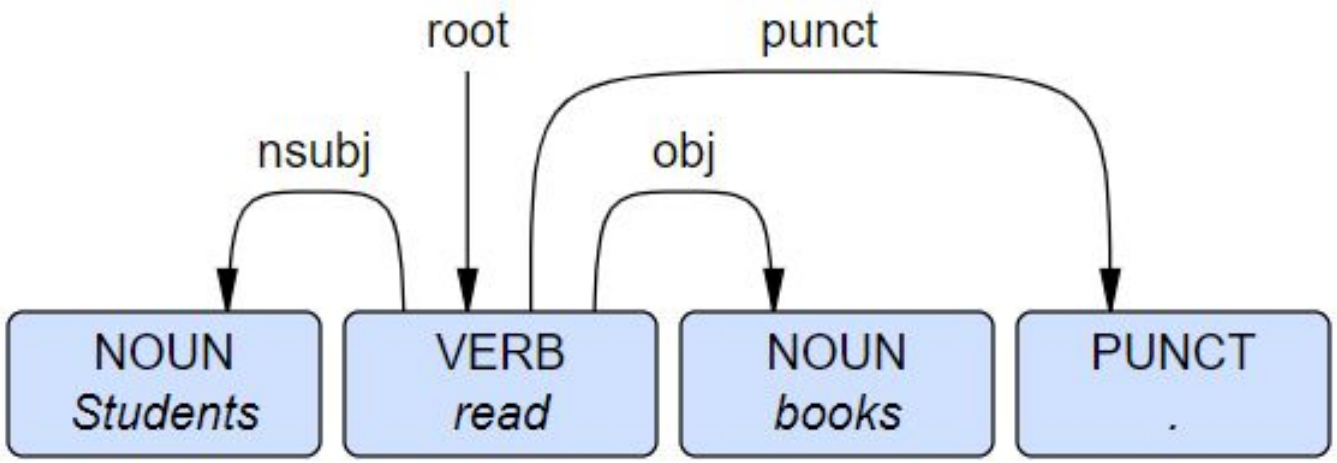
هن کيم ڪتاب ڏنو

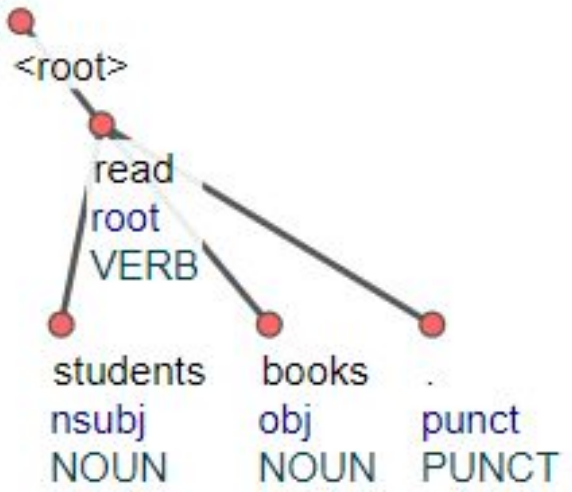
Huni	kHē-mi	kitAbu	dinU
3P.M.Obl	Dat-1P.Sg	book.Sg.M	give.PastPart.M.Sg

He gave me a book.

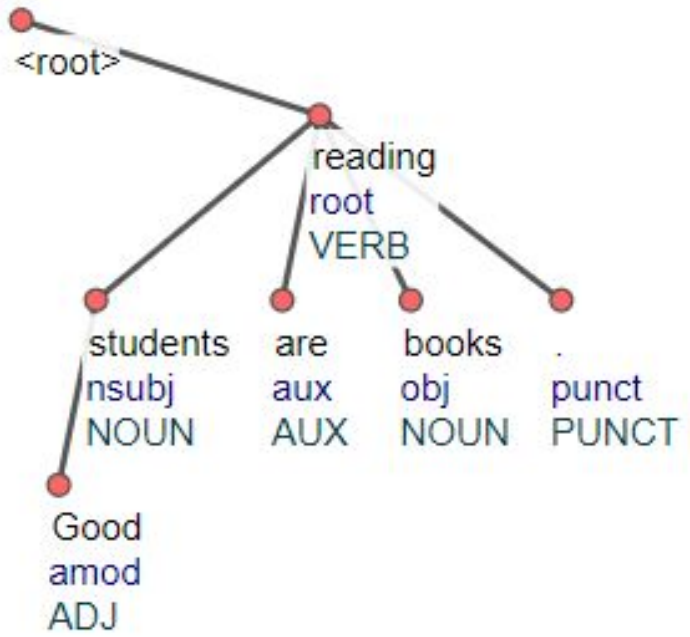
Universal Dependencies
a lightning fast introduction

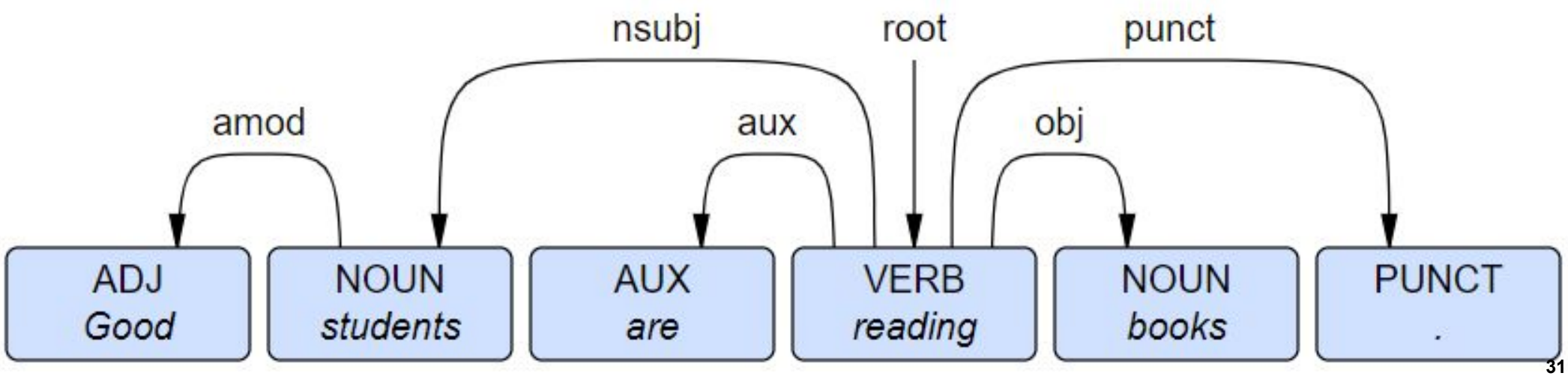
Students read books.



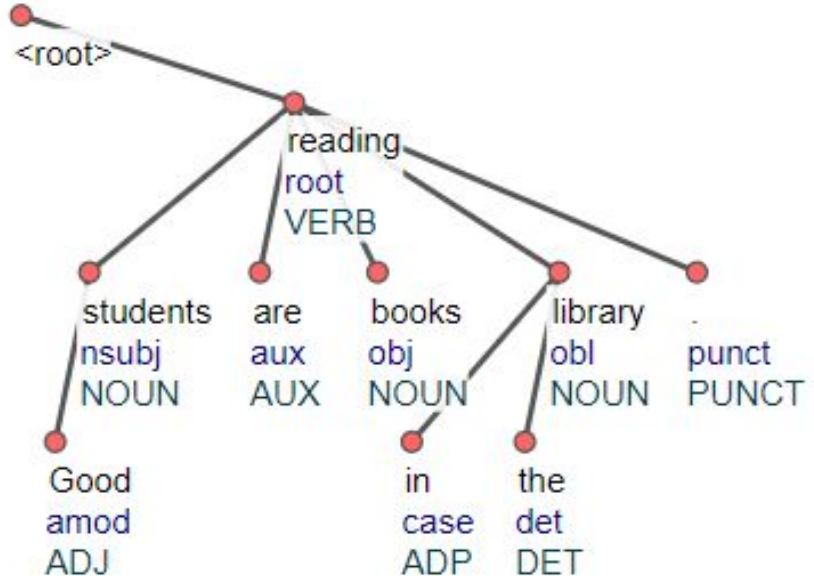


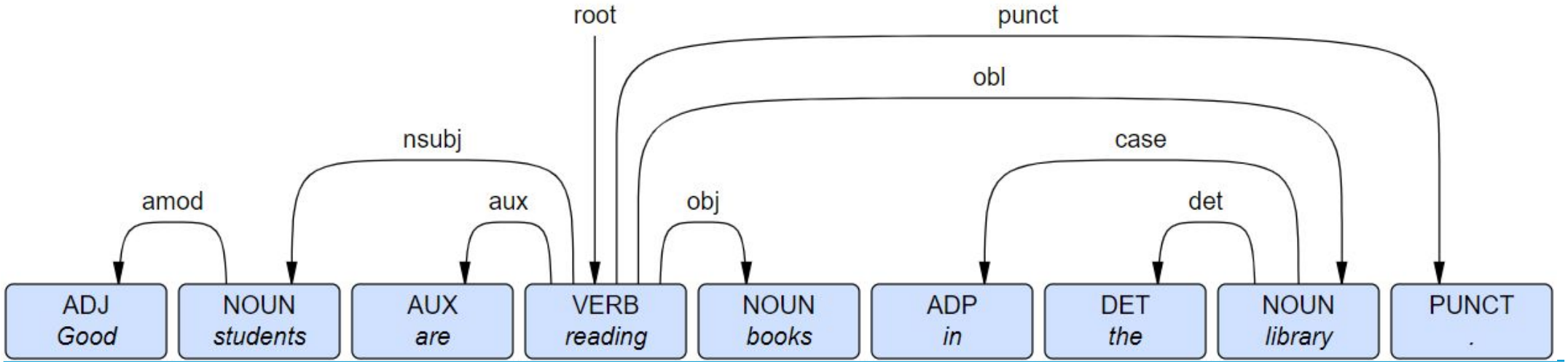
Good students are reading books.





Good students are reading books in the library.





CoNLL format

Id	Form	Lemma	UPosTag	XPosTag	Feats	Head	DepRel
1	Good	good	ADJ	JJ	Degree=Pos	2	amod
2	students	student	NOUN	NNS	Number=Plur	4	nsubj
3	are	be	AUX	VBP	Mood=Ind|Number=Plur|Pers	4	aux
4	reading	read	VERB	VBG	Tense=Pres|VerbForm=Part	0	root
5	books	book	NOUN	NNS	Number=Plur	4	obj
6	in	in	ADP	IN	_	8	case
7	the	the	DET	DT	Definite=Def|PronType=Art	8	det
8	library	library	NOUN	NN	Number=Sing	4	obl
9	.	.	PUNCT	.	_	4	punct

Dependency Relations

	Nominals	Clauses	Modifier words	Function Words
Core arguments	nsubj obj iobj	csubj ccomp xcomp		
Non-core dependents	obl vocative expl dislocated	advcl	advmod* discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case
Coordination	**Headless**	**Loose**	**Special**	**Other**
conj cc	fixed flat	list parataxis	compound orphan goeswith reparandum	punct root dep

Multiword tokenisation in the Universal Dependencies

#Text = It's just fun

Token id	Surface form	Lemma	POS	XPOS	Morph	Dep. id	Dep. name		
1-2	It's	_	_	_	_	_	_	_	_
1	It	it	PRON	PRP	Case=Nom|Gender=Neut|Number=Sing|Person=3|PronType=Prs	4	nsubj	_	_
2	's	be	AUX	VBZ	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin	4	cop	_	_
3	just	just	ADV	RB	_	4	advmod	_	_
4	fun	fun	ADJ	JJ	_	0	root	_	_

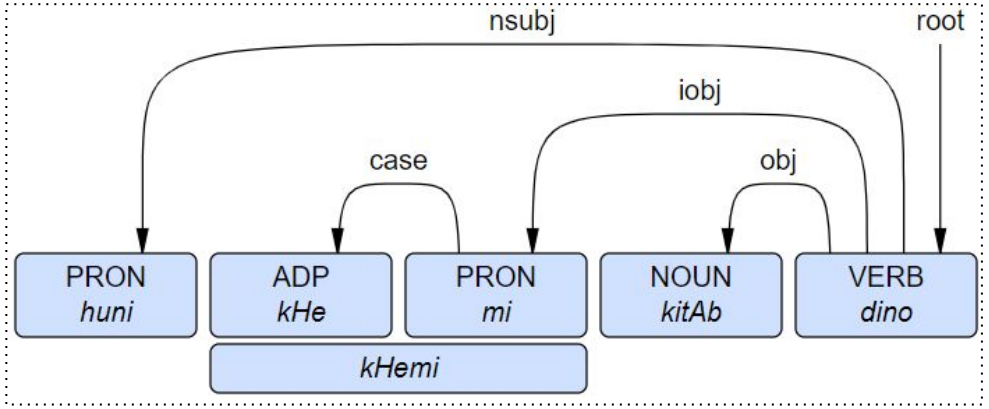
Pronominal Suffix on case markers - Sindhi

هن کیم کتاب ڏنو

Huni	kHē-mi	kitAbu	dinU
3P.M.Obl	Dat-1P.Sg	book.Sg.M	give.PastPart.M.Sg

He gave me a book.

```
# text = هن کیم کتاب ڏنو
1       huni    -       PRON    PRP     Case=Acc|Number=Sing|Person=3|PronType=Prs      5       nsubj   _
2-3     kHemi   -       -       -       -       -       -       -
2       kHe     -       ADP     PSP     Case=Acc,Dat|Number=Sing|Person=1|PronType=Prs  3       case    _
3       mi      -       PRON    PRP     Case=Acc,Dat|Number=Sing|Person=1|PronType=Prs  5       iobj    _
4       kitAb   -       NOUN    NN      Case=Nom|Gender=Masc|Number=Sing|Person=3        5       obj     _
5       dinU    .       VERB    VM      Gender=Masc|Number=Sing|Person=3|VerbForm=PastPart       0       root
```



UD - important websites

- **Paper**
  De Marneffe, M. C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. Computational linguistics, 47(2), 255-308.
- **Main website and corpora**
  https://universaldependencies.org/
- **Visualization**
  https://urd2.let.rug.nl/~kleiweg/conllu/
- **UDPipe Parser**
  https://lindat.mff.cuni.cz/services/udpipe/
- **Models**
  https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-5150

  https://github.com/ufal/udpipe/tree/udpipe-2