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

Urdu/Hindi Morphology

Urdu(/Hindi) Noun

Form	Number	Gender		
Nominative	Singular	Masculine	لڑکا آیا	laRk-ā āyā boy-M.Sg.Nom came
Nominative	Plural	Masculine	دو لڑکے آئے	do laRk-ē āē two boy-M-Pl.Nom came
Oblique	Singular	Masculine	لڑکے نے کہا	laRk-ē=nē kahā boy-M.Sg.Obl=ērg said
Oblique	Plural	Masculine	لڑکوں نے کہا	laRk-oN=nē kahā boy-M.Pl.Obl=Erg said

Urdu/Hindi Noun Morphology

	Transl.	Sg.Nom	Sg.Obl	Pl.Nom	Pl.Obl
Masc. ending with *ā* sound	*boy*	laRk-ā لڑکا	laRk-ē لڑکے	laRk-ē لڑکے	laRk-ON لڑکوں
Masc. Other endings	*house*	gHar گھر	gHar گھر	gHar گھر	gHar-ON گھروں
Fem. ending with ī sound	*girl*	laRk-ī لڑکی	laRk-ī لڑکی	laRk-īāN لڑکیاں	laRk-īON لڑکیوں
Fem. Other endings	*book*	kitāb کتاب	kitāb کتاب	kitāb-ēN کتابیں	kitāb-ON کتابوں

Borrowed Morphology (Urdu)

Singular Word	Language of Origin	Native Plural	Borrowed Plural
darj-ah درجہ	Arabic	darj-ē درجے	darj-āt درجات
kampiuTar کمپیوٹر	English	kampiūTar-ON کمپیوٹروں	kampiūTar-z کمپیوٹرز
kitāb کتاب	Arabic	kitāb-ēN کتابیں	kutub کتب

Case Markers (Urdu/Hindi)

لڑکی نے لڑکے کو دوربین سے دیکھا

laRki=nē laRkē=ko dūrbīN=sē dēkHA

girl.F.Sg.Obl=Erg boy.M-Sg.Obl=Acc telescope.F.Sg.Obl=inst *saw*

'The girl saw the boy with the telescope.'

Case	**Case Marker**
Ergative	nē
Accusative, Dative	ko
Ablative, īnstrument	sē
Locative	mēN ('in'), par ('on'), …
Genitive	k- (kā, kī, kē)

Morphological Case (Sindhi)

http://www.languagesgulper.com/eng/Sindhi.html

Nominative	**Oblique**	**Ablative**
gHar-u house.M-Sg.Nom	gHar-a house.M-Pl.Obl	gHar-**āN** house.M-Sg.Abl
gHar-āN house.M-Pl.Nom	gHar-ani house.M-Pl.Obl	gHar-ani-**āN** house.M-Pl.Obl-Abl

یاسین گھر کان آیو

Yasīn gHar=kHāN āyo

Yaseen house=Abl *came*

'Yaseen came from the house.'

یاسین گھران آیو

Yasin gHar-āN āyo

Yaseen house-Abl *came*

'Yaseen came from the house.'

Morphological Case (Punjabi)

- gHar-ON
  home.M-Abl

- gHar-iC
  home.M-Loc_in

Urdu/Hindi Verb Morphology

Word	Transliteration	Features
لکھ	likH	write
لکھتا	likH-tā	write-impf.M.Sg
لکھتی	likH-tī	write-īmpf.F.Sg
لکھی	liKH-ī	write-Perf.M.Sg
لکھوں	likH-ON	write-Subj.1P.Sg
لکھیے	likH-iē	write-Perc.2P

Urdu/Hindi Verb Morphology

Imperfective	**likH-tā** M.Sg	**likH-tē** M.Pl	**likH-tī** F.Sg	**likH-tīN** F.Pl
Perfective	**likH-ā** M.Sg	**likH-ē** M.Pl	**likH-ī** F.Sg	**likH-īN** F.Pl
Subjunctive	**likH-ūN** 1P.Sg	**likH-ē** 1P.Sg/2P.Sg	**likH-ēN** Pl	
Future	**likH-ūN-gā** 1P.M.Sg **likH-ūN-gī** 1P.F.Sg	**likH-ē-gA** 2/3P.M.Sg **likH-ē-gī** 2/3P.F.Sg	**likH-ēN-gē** M.Pl **likH-ēN-gī** F.Pl	
Percative	**likH-iē** (2P.Pl)			

Urdu/Hindi Verb Morphology

laRk-ī	ā-ī
girl-F.Sg.Nom	come-Perf.F.Sg
‘A/The girl came.’	

laRk-īāN	ā-īN
girl-F.Pl.Nom	come-Perf.F.Pl
‘Girls came.’	

ammi	ā-īN
mother.**Hon**-F.Sg.Nom	come-Perf.F**.Hon**
‘Mother came.’	(Honorific)

Pronominal Suffixes (Sindhi)

Root: puT-u (son-M.Sg)

Suffix Person	Suffix Number	Word		Alternate phrase
1	Sg	پٹم	puTu-mi (my son)	منهنجو پٽ munHanjO puTu
1	Pl	پٹون	puTu-ūN (our son)	اسان جو پٽ Asān=jo puTu
2	Sg	پٹين	puTu-ēN (your.sg son)	توهان جو پٽ tuHān=jo puTu
2	Pl	پٹوا	puTu-va (your.pl son)	توهان جو پٽ tuHān=jo puTu
3	Sg	پٹس	puTu-si (his son)	هن جو پٽ hun=jo puTu
3	Pl	پٹن	puTu-ni (their son)	سندن پٽ sundanu puTu

Pronominal Suffixes (Sindhi)

on Case Markers

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 ڪيم 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.

Pronominal Suffixes (Sindhi)

On Verbs

خط لکیومانس

xat-u likH-iyO-maan-si

Letter.M-Nom.Sg write-PastPart.Sg-1P.Sg-3P.Sg

I wrote him/her (a) letter.

Pronominal Suffixes

https://prr.hec.gov.pk/jspui/bitstream/123456789/10182/1/PhD%20Thesis%20%28Mutee%20u%20Rahman%29%20.pdf

http://www.sindhiadabiboard.org/Catalogue/Lasaniyat/Book5/Book_page7.html

Tokenization

Urdu script

an-ginat	ان گنت	uncountable
an-kahī	ان کہی	untold
aasaan-tareen	آسان ترین	easiest

Tokenization

Word With Space	Transliteration and gloss	Meaning
ان گنت	an-ginat not-counted	uncountable
ان کہی	an-kahl not-told	untold
انجان	an-jAn not-known	unknown
اہم ترین	aham-tarIn important-most	most important
شادی شدہ	SHAdI-SHudah marriage-became	married.Adj
گم شدہ / گمشدہ	gum-SHudah lost-became	lost.Adj

Tokenization

Stuttgart

Transliteration	Google Search Results
شٹٹگارٹ	15
شٹٹ گارٹ	39

The rule(s) for space insertion/deletion (including the exceptions) is/are yet to be explored.

Morphology Induction

- **Linguistica**
  - https://linguistica-uchicago.github.io/lxa5/
  - Unsupervised Learning from Corpus
  - Learning
  - Affixes, Signatures, and associated words
  - An Example is in the Python notebook

Morphology Induction

- **Morfessor**
  - https://morfessor.readthedocs.io/
  - Pre-trained segmentation models
  - Models can be trained

Morphological Features

UniMorph

Number	Greater paucal	GPAUC
Number	Greater plural	GRPL
Number	Inverse	INVN
Number	Paucal	PAUC
Number	Plural	PL
Number	Singular	SG
Gender	Bantu Noun Classes	BANTU1-23
Gender	Feminine	FEM
Gender	Masculine	MASC
Gender	Nakh-Daghestanian Noun Classes	NAKH1-8
Gender	Neuter	NEUT

https://unimorph.github.io/

Tamil morphological analyser

https://nlp-tools.uom.lk/thamizhi-morph/parse-sentence.php

Thank you

Tafseer Ahmed
Senior NLP Scientist, Alexa Translations
tafseer@gmail.com

Kengatharaiyer Sarveswaran
University of Jaffna, Sri Lanka.
University of Konstanz, Germany
sarves@univ.jfn.ac.lk
sarves.github.io

Subwords for Deep Learning

What are the challenges of rich morphology to:

- **Morphological Analysis/Generation**
- **Part of Speech tagging**
- **Word Embedding**

Morphological Features

ParGram(Parallel Grammar)

Nominal Features

ANIM: ±

CASE: ACC, DAT, GEN, NOM

GEND: FEM, MASC, NEUT

NTYPE: COUNT, MASS, PROPER

NUM: PL, SG

PERS: 1, 2, 3

REFL: ±

Morphological Features

Universal Dependency

Lexical features	Inflectional features	
	*Nominal**	*Verbal**
PronType	Gender	VerbForm
NumType	Animacy	Mood
Poss	NounClass	Tense
Reflex	Number	Aspect
Foreign	Case	Voice
Abbr	Definite	Evident
	Degree	Polarity
		Person
		Polite
		Clusivity

https://universaldependencies.org/u/fe