Computational Linguistics

for Pakistan Languages

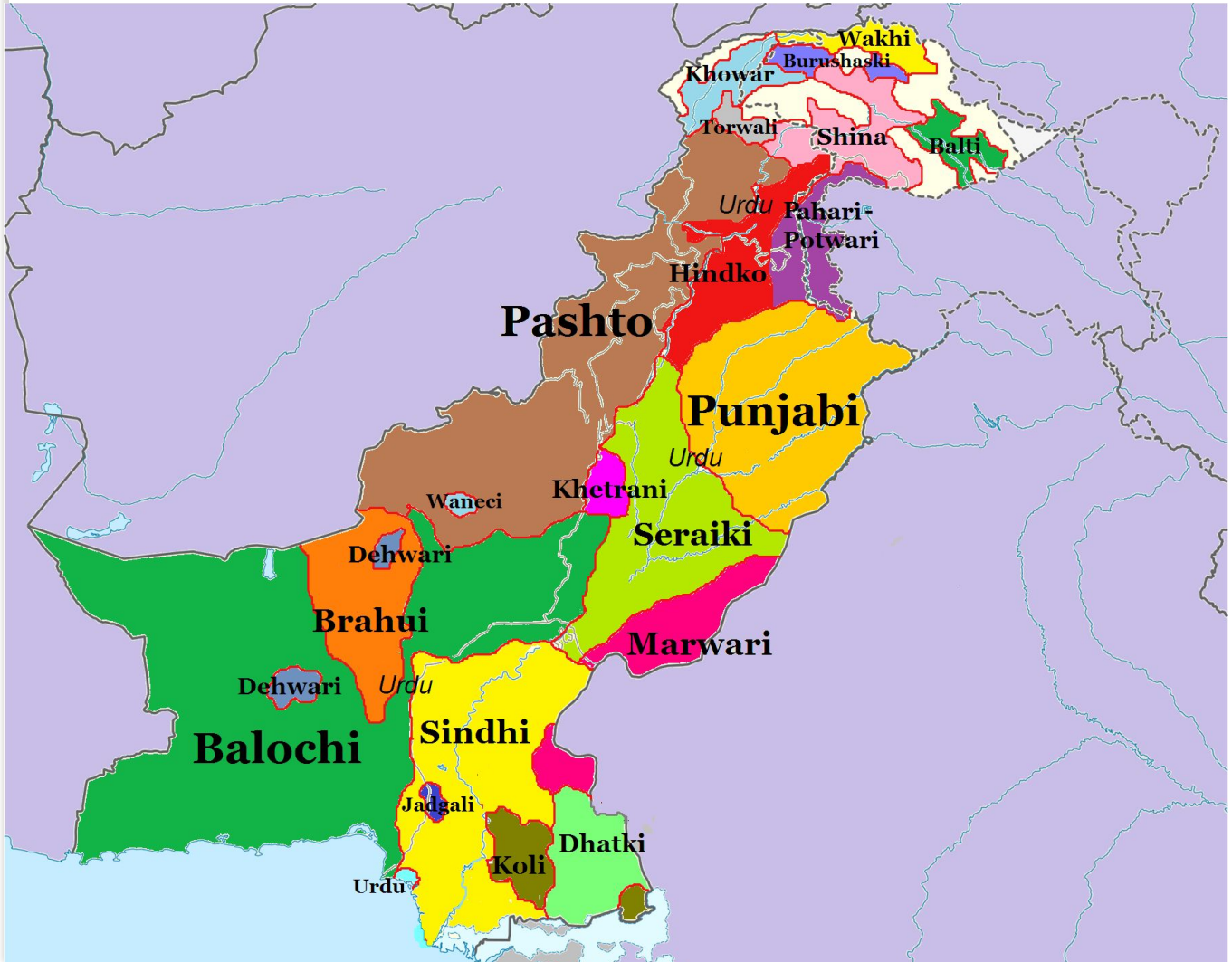
Tafseer Ahmed

Languages in Pakistan

- **Languages: 73**, Indigenous: 65
- Institutional: 7, Developing: 17, Vigorous: 39, Trouble: 8, Dying: 2
- Families: Indo-Aryan, Iranian, Dravidian, Tibetan, Isolate

- https://www.ethnologue.com/country/PK

Overview


Wakhi
Burushaski
Khowar
Torwali
Shina
Balti
Urdu
Pahari-
Potwari
Hindko
Pashto
Punjabi
Urdu
Khetrani
Waneci
Seraiki
Dehwari
Brahui
Marwari
Dehwari
Urdu
Sindhi
Balochi
Jadgali
Dhatki
Koli
Urdu

Linguistic Analysis

- **Traditional Grammars**
  - paragraphs explaining the grammar

- **Formal Grammars**
  - Formal grammar rules
  - Panini (4th century BC) , Chomsky (1928-)

Computer Science

- **Computational Linguistics (CL)**
  - modeling of natural (human) languages

- **Natural Language Processing (NLP)**
  - processing of natural languages

Different Techniques of NLP

- Linguistic Analysis
- Non-Linguistic Methods
  - Bag of Words
  - Sequence to Sequence

Linguistic Analysis

- Deep Method *(not related to deep learning)*
- Inspired by Linguistics

- Token → Part of Speech → Morphology → Syntax → Semantics → Discourse → ...

Linguistic Features

- Linguistic Features are used as features in many machine learning systems.

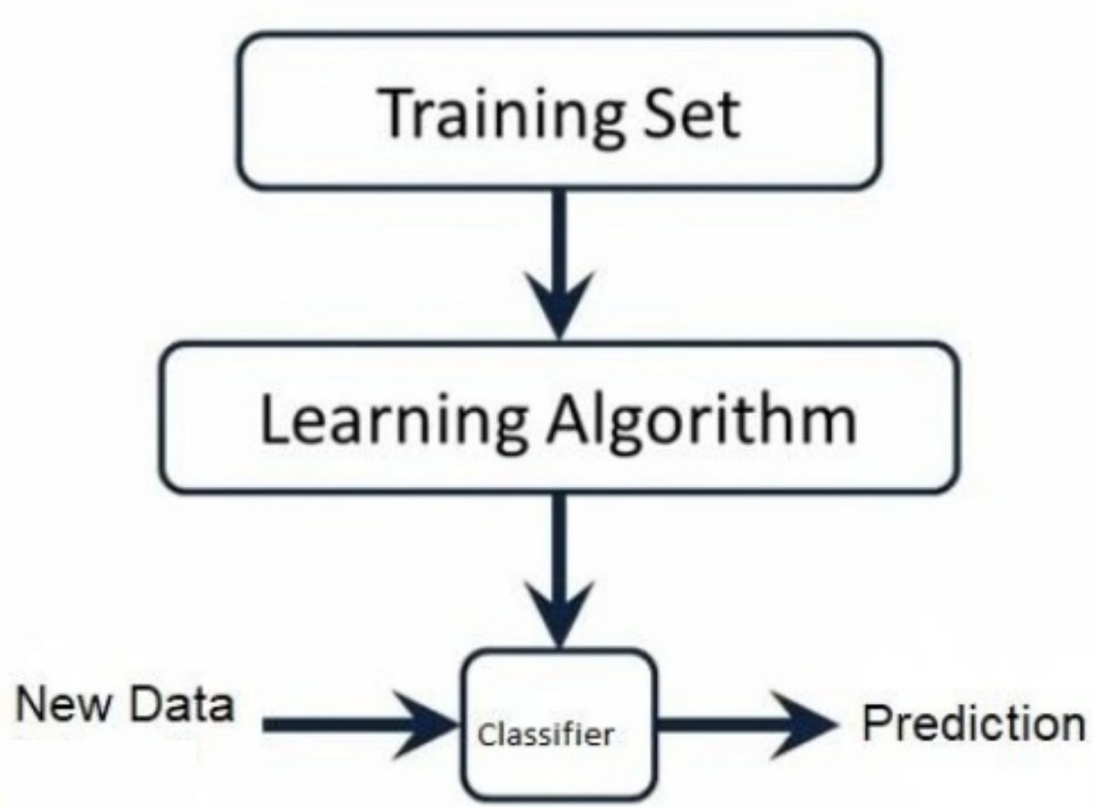
- Many current and old systems uses linguistic analysis, however modern trend is End-to-End systems.

Non-Linguistic Methods

- **Bag of Words - Order does not matter**
  - Text Classification, Text Clustering, Topic Modeling, ...

- **Sequence Modeling - Order does matter**
  - Language Modeling, Machine Translation, ...

Supervised Machine Learning



A sample feature set in SVMTool

word features	$w_{-3}, w_{-2}, w_{-1}, w_0, w_{+1}, w_{+2}, w_{+3}$
PoS features	$p_{-3}, p_{-2}, p_{-1}, p_0, p_{+1}, p_{+2}, p_{+3}$
ambiguity classes	a_0, a_1, a_2, a_3
may_be's	m_0, m_1, m_2, m_3
word bigrams	$(w_{-2}, w_{-1}), (w_{-1}, w_{+1}), (w_{-1}, w_0)$ $(w_0, w_{+1}), (w_{+1}, w_{+2})$
PoS bigrams	$(p_{-2}, p_{-1}), (p_{-1}, a_{+1}), (a_{+1}, a_{+2})$
word trigrams	$(w_{-2}, w_{-1}, w_0), (w_{-2}, w_{-1}, w_{+1}),$ $(w_{-1}, w_0, w_{+1}), (w_{-1}, w_{+1}, w_{+2}),$ (w_0, w_{+1}, w_{+2})
PoS trigrams	$(p_{-2}, p_{-1}, a_{+0}), (p_{-2}, p_{-1}, a_{+1}),$ $(p_{-1}, a_0, a_{+1}), (p_{-1}, a_{+1}, a_{+2})$
sentence_info	punctuation ('.', '?', '!')
prefixes	s_1, $s_1 s_2$, $s_1 s_2 s_3$, $s_1 s_2 s_3 s_4$
suffixes	s_n, $s_{n\text{-}1} s_n$, $s_{n\text{-}2} s_{n\text{-}1} s_n$, $s_{n\text{-}3} s_{n\text{-}2} s_{n\text{-}1} s_n$
binary word features	initial Upper Case, all Upper Case, no initial Capital Letter(s), all Lower Case, contains a (period / number / hyphen ...)
word length	integer

Linguistic Analysis - Goals (of this series)

- Learning and creating **Linguistic Representations** that
  - are widely used in computer applications
  - deals with peculiar features of Pakistani languages

- Training the **annotators** for creating linguistically correct training dataset

Linguistic Analysis - Goals (of this series)

- Creating linguistic analyzers by using **small training datasets**
  - Unsupervised Learning
  - Word Embeddings
  - Transfer Learning
  - .....
- Using the **libraries and tools** of computational linguistics

Not the Goals/Focus of this series

- Text Mining
- Corpus Linguistics
- NLP algorithms - in depth
- Machine (including Deep) Learning - in depth
  - However, the topics related to linguistic analysis will be discussed

Linguistic Analysis

Part of Speech (PoS)

ذہین	لڑکی	نے	روزانہ	ایک	اچھی	کتاب	پڑھی
ADJ	NOUN	ADP	ADV	NUM	ADJ	NOUN	VERB

PoS - Universal 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	

Building PoS Tagger using small dataset

- The available PoS Tagger(s) for Urdu are not very good.
- Annotated Data is required for other Pakistani languages

Morphological Features

تھیں	پڑھی	کتابیں	اچھی	نے	لڑکیوں	ذہین	
ہے	پڑھ	کتاب	اچھا	نے	لڑکی	ذہین	**Lemma**
AUX	VERB	NOUN	ADJ	ADP	NOUN	ADJ	**POS**
	Form= Perf Gend= Fem Pers=3	Gend=Fem Num=Pl Form=Nom	Gend= Fem Num= Pl Form=		Gend= Fem Num=Pl Form=Obl	Gend= Fem Num=Pl Form= Obl	**MORPH. Features**

Morphology - Pakistani Languages

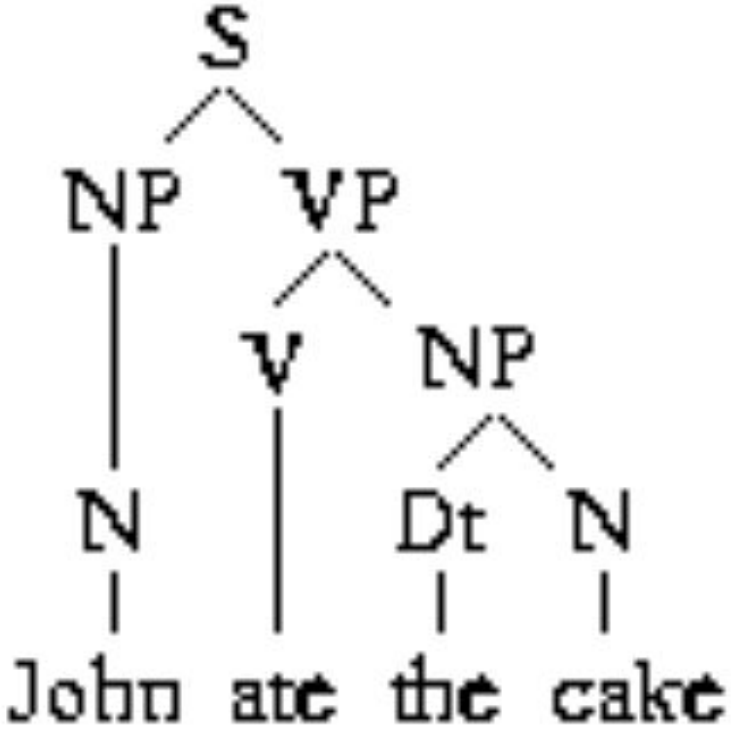
- **Rich Morphology**
- Nominative and Oblique Forms
- Inflectional Case
- Causatives
- Reduplication
- Pronominal Suffixes

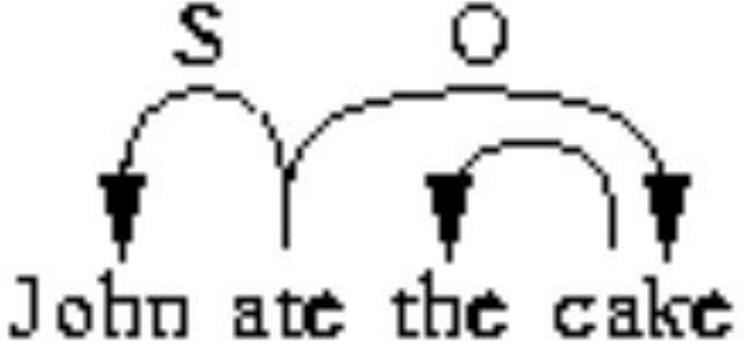
Derivational Morphology

Morphological Analysis

- FInite State Transducer
- Morph. Feature Prediction
- Unsupervised Learning and Prediction

Syntax: Phrase Structure vs Dependencies





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

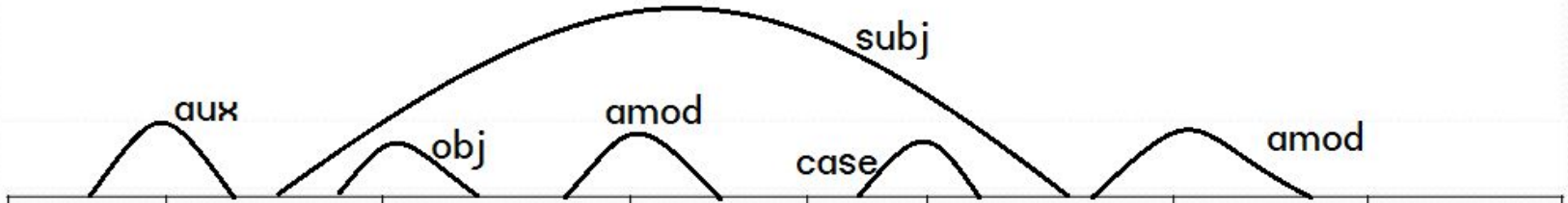
Nominal dep	*Predicate dep*	*Modifier word*
nummod	acl	amod
appos		det
nmod		neg

Compounding and unanalyzed

compound	mwe	goeswith
name	foreign	

Coordination

conj	cc	punct



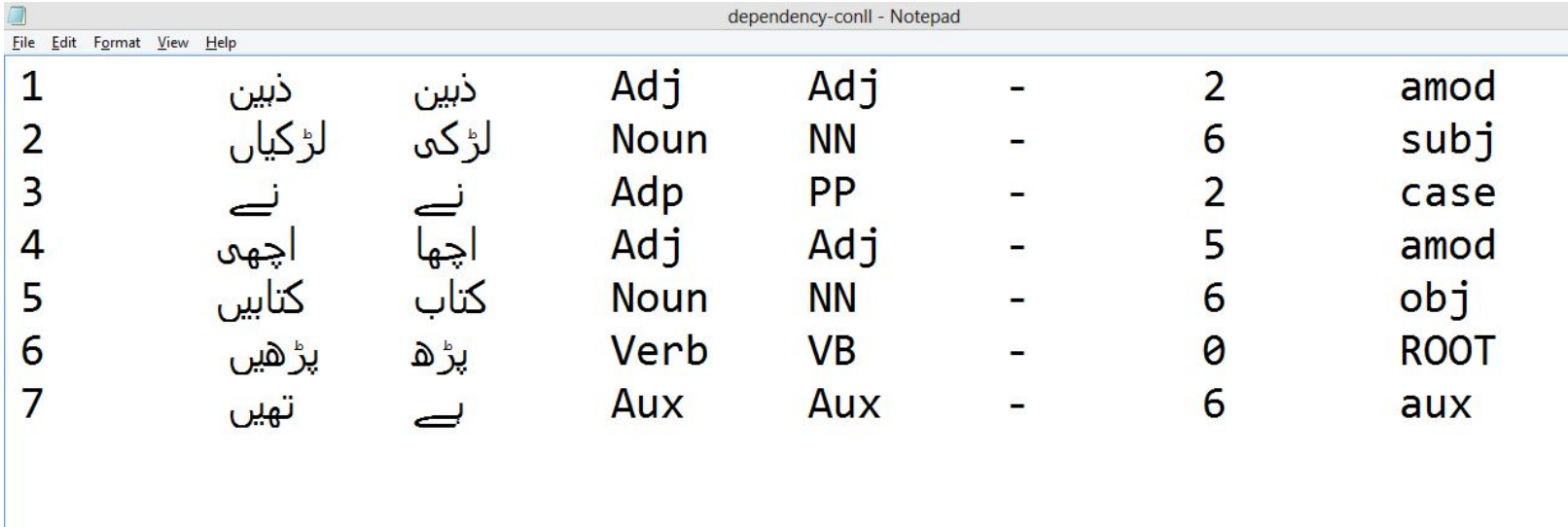
تھیں	پڑھیں	کتابیں	اچھی	نے	لڑکیوں	ذہین	
ہے	پڑھ	کتاب	اچھا	نے	لڑکی	ذہین	**Lemma**
Aux	Verb	NN	Adj	AdP	Noun	Adj	**POS**
	Form=Perf Gend=Fem Pers=3 Num=Pl	Gend=Fem Num=Pl Form=Nom	Gend=Fem Num=Pl Form=Nom		Gend=Fem Num=Pl Form=Obl	Gend=Fem Num=Pl Form=Obl	**Features**

Universal Dependencies - To Do

- Revisiting annotation scheme for Urdu (and other South Asian languages)
- Creating UD_banks for other Pakistani languages

Annotation

- **GATE** (General Architecture for Text Engineering), Sheffield
- **Brat**, Manchester, Tokoyo,.....
- **Webanno**, Darmstad
- .....



Semantic Role

- The glass[patient] broke.
- I[agent] break the glass[patient].
- The glass[patient] was broken by me[agent].

Urdu PropBank

Predicate: *DagmagA*

DagmagA: created by Tafseer - Hindi dagmagA

Roleset id: DagmagA.01 , *To stagger/tremble/shake*

Roles:

Arg1: *The entity that staggers/trembles/shakes*

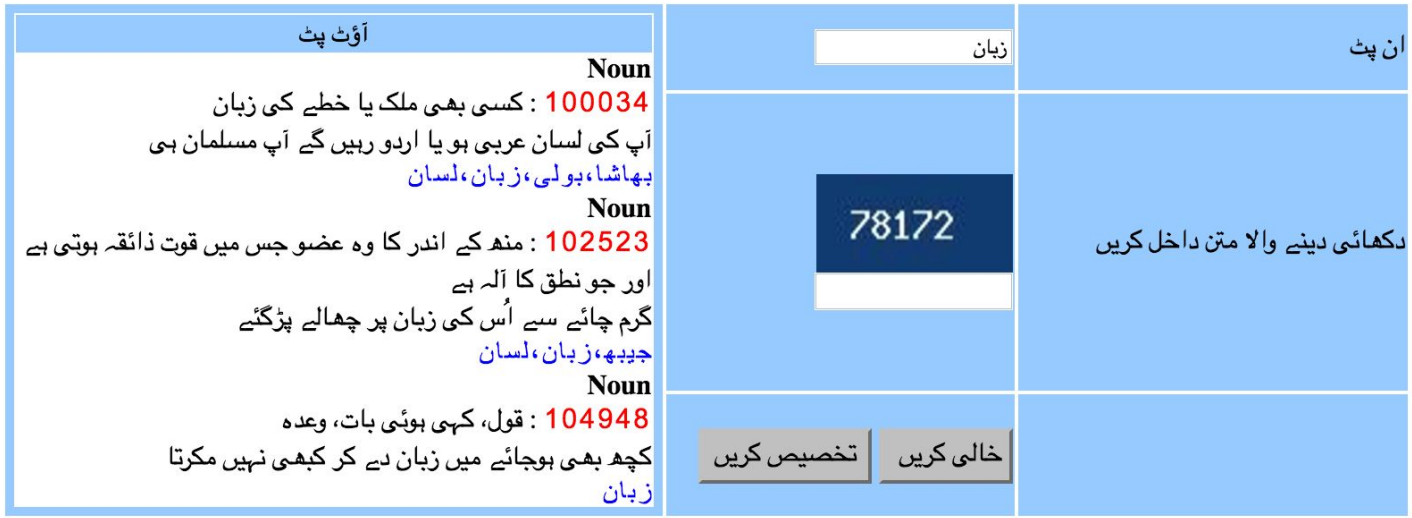
Example:

http://www.express.pk/story/44282/
ہیلی کاپٹر بُری طرح ڈگمگایا

Arg1: ہیلی کاپٹر
Argm-mnr: بُری طرح
Rel: ڈگمگایا

Lexical Resources

- Urdu Wordnet
- http://www.cle.org.pk/clestore/urduwordnet.htm



Distributional Semantics

- “You shall know a word by the company it keeps.” John Rupert Firth

- Word Embeddings
- Topic Models
- ……


```
: model.most_similar( 'سورج', topn
```

```
: [('0.8165232539176941 ,'چاند),
   ('0.8143321275711106 ,'سُورج),
   ('0.7290103435516357 ,'آفتاب),
   ('0.7133749723434448 ,'ماہتاب),
   ('0.707730770111084 ,'تارے),
   ('0.702995777130127 ,'ستاروں),
   ('0.69676274061203 ,'ستارے),
   ('0.6915234327316284 ,'کرنوں),
   ('0.6877450942993164 ,'مریخ),
   ('0.6747698783874512 ,'غروب),
   ('0.6725152730941772 ,'بادلوں),
   ('0.6688274741117279 ,'تاروں),
   ('0.6525486707687378 ,'گہن),
   ('0.6493672132492065 ,'شفق),
   ('0.6463797092437744 ,'شعاعوں)]
```

Discourse Analysis

- ..... He wrote that last year. ...

  Who is HE?

  What is THAT?

  When is the LAST YEAR?

Filling the gaps – Pro Drop

- وہ کمرے میں آیا، کرسی ڈھونڈی اور بیٹھ گیا

- وہ کمرے میں آیا، (اس نے) کرسی ڈھونڈی اور (وہ) (اس پر) بیٹھ گیا

Thank You

شکریہ

Contact: tafseer@gmail.com