

# Urdu Corpus Processing Fundamentals

Sarmad Hussain

## Important Note

Much of this material is from:

*Speech and Language Processing: An Introduction to Natural Language  
Processing, Computational Linguistics, and Speech Recognition*

By Daniel Jurafsky, James H. Martin  
Published by Pearson Prentice Hall, 2008  
ISBN 0131873210, 9780131873216

*These slides are for Computational Linguistics courses at NUCES;  
**For course use only and not for further circulation or reuse due to  
possible copyright violations; Please purchase original book***

## Challenges

- Non-Unicode Encoding
  - Inpage
- Unicode Encoding
  - Selection of Urdu Characters
  - Normalization
  - Redundancy Filtering
- Word Segmentation

## Word Segmentation

- Normally done through
  - Space
  - Punctuation Marks
- Not all languages use space to separate words
  - Chinese 人人生而自由，在尊嚴和權利上一律平等
  - Thai เราทุกคนเกิดมาอย่างอิสระ
  - Lao ມະນຸດທຸກຄົນເກີດມາບົ່ງຄູ່ດ້ວຍສິດສິດທິ
  - Khmer មនុស្សទាំងអស់កើតមកមានសេរីភាពនិងភាពស្មើ
  - Burmese လူတိုင်းသည် တူညီလိင်တူလိပ်သော
  - Dzongkha འགྲོ་བ་མི་རིགས་ལ་ར་དབང་ཆ་འདྲ་མཉམ་

## Word Segmentation in Urdu

- Urdu does not have a concept of space
  - Uses character shaping to indicate boundaries
- Users use space in this context
  - Not only to create word boundaries
  - Also to create correct shaping
- Space cannot be consistently interpreted as word boundary marker

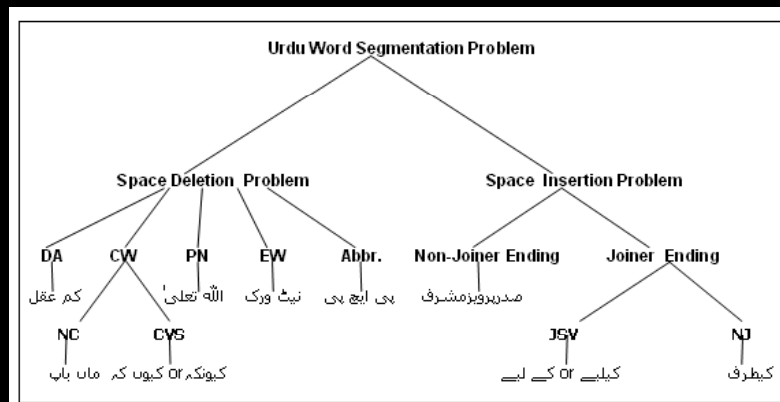
## Word Segmentation in Urdu

- Word boundary necessary for language processing for ALL applications
  - A necessary pre-requisite for language processing applications

## Word Segmentation in Urdu Anatomy of the Problem

- A single word has spaces inside it
  - Due to writing conventions
    - یونیورسٹی، خوب صورت، اسلام آباد
- Multiple words do not have a space between them
  - Due to words ending in non-joiners
    - نوجوان آرہے
      - “young man is on this side”
      - “nine young men are coming!”
  - Due to ambiguous definition of a word
    - اچکا، کیلے، اسکا

## Word Segmentation in Urdu Anatomy of the Problem



Key: DA=Derivational Affixes, CW=Compound Words, PN=Proper Nouns,  
EW=English Words, Abbr. =Abbreviation, NC=Normal Compounds  
CVS=Compounds with Spelling Variation, JSV= Joiners with Spelling Variation,  
NJ=Normal Joiners

## Word Segmentation Solution

- Step 1: Generate all legitimate segmentations
- Step 2: Rank segmentations
- Details later...
- At this time, careful with corpus processing and using space as a delimiter for tokenization

## Summary for Corpus Cleaning

1. Take input
2. Convert to standard encoding (Unicode)
3. Map to selected subset of character to reduce encoding redundancy
4. Normalize text
5. Define what is a “word” and segment text into words for further processing