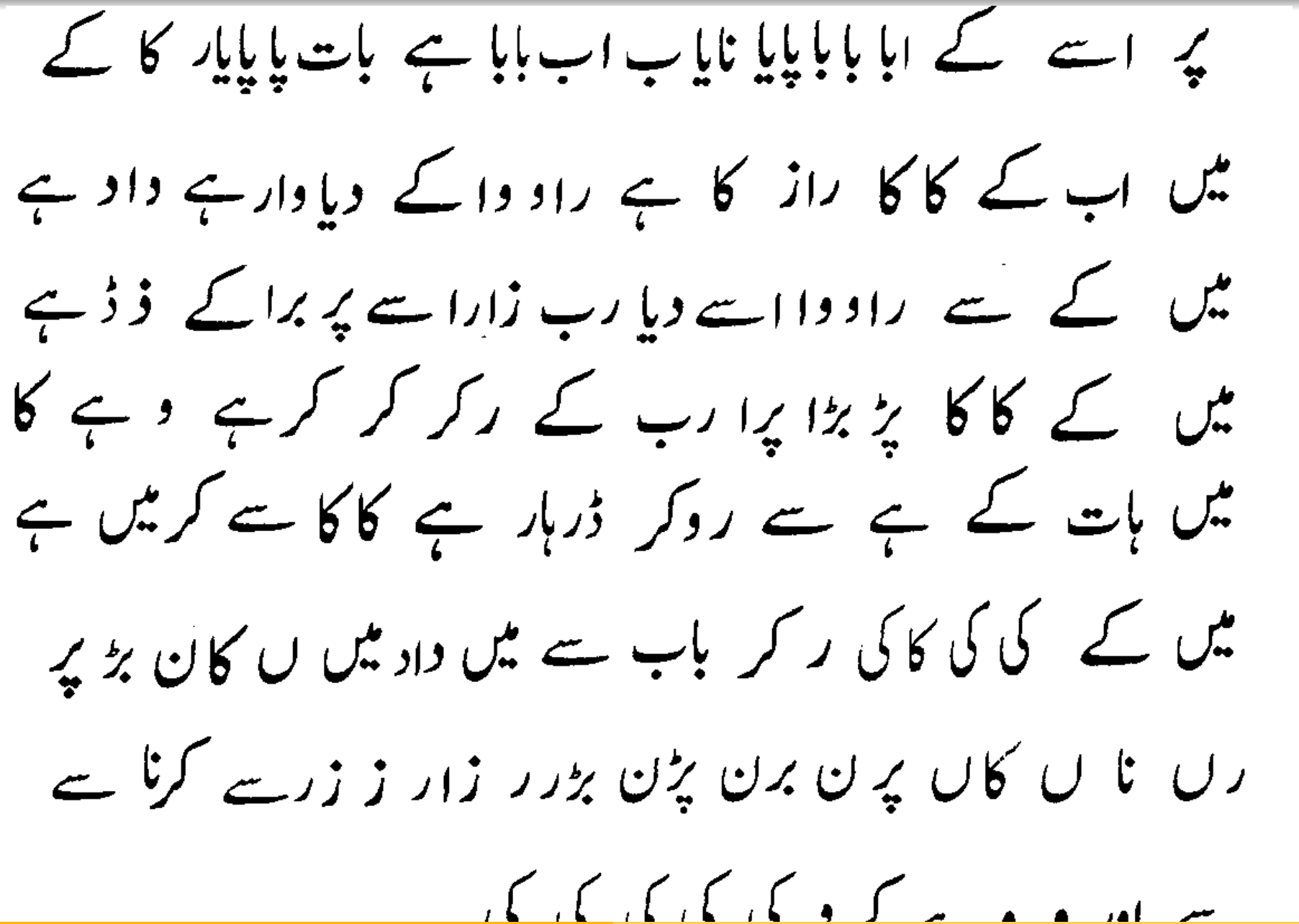
URDU Optical Character Recognizer

Benazir Mumtaz, Hina Khalid, Muhammad Kamran

ISSALE 2014

September 26, 2014

- Pre-processing
  - Line segmentation
  - Ligature segmentation
  - Connected Component identification
- Recognition
- Post-processing
  - Ligature formation



+اسے کے ا + + + + + + ب اب + + ہے + ت + + + ر کا کے

میں اب کے کا کا را + کا ہے راووا کے د + وار ہے واد ہے

میں کے سے راووا اسے د + رب + اوا سے + برا کے + + ہے

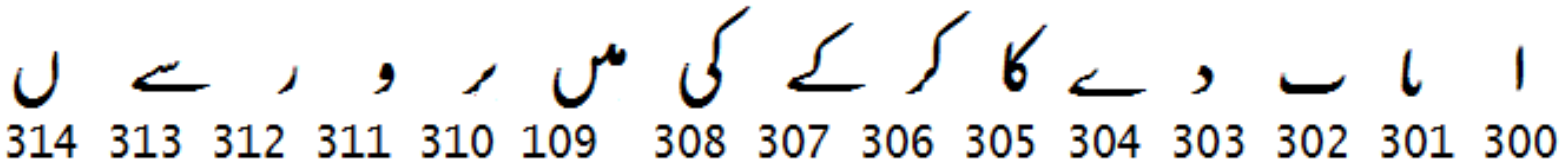
میں کے کا کا + + ا + ارب کے + کر + کے ہے وہے کا

+ + ت کے ہے سے + وکے + ر + رہے کا کا سے + میں ہے

+ کے کی کا کا + + کے + ب سے میں داو میں + کا + + +

ر + + کا + + + + بر + + + + رر + ار + + رسے + + +

- ligature Shapes : 15 main strokes



- diacritics : 5 diacritics
- lines: 8 lines

Main Body Type	Total Tokens in Document Image	Total Unique Syllable in Document Image
300	18	1
301	18	5
302	8	2
303	9	3
304	11	1
305	12	1
306	8	1
307	10	1
308	8	1
309	8	1
310	12	4
311	11	1
312	23	2
313	9	1

- Line Segmentation: [8:8] = 100%
- Ligature Segmentation: [179:180]=100%
- Connected Component Disambiguation:

MB in Document Image	MB Detected	Diacritics in Document Image	Diacritic Detected
180	180	88	88

Training Accuracy : 97.5 %

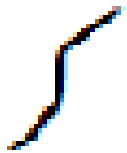
```
BB_Area <= 1020
|   Width <= 17
|   |   Width <= 9: three_hundred (35.0)
|   |   Width > 9
|   |   |   Height <= 31: three_hundred_eleven (35.0)
|   |   |   Height > 31: three_hundred_one (35.0)
|   Width > 17
|   |   Width <= 21
|   |   |   Height <= 25: three_hundred_twleve (35.0)
|   |   |   Height > 25: three_hundred_three (35.0)
|   |   Width > 21
|   |   |   Height <= 22: three_hundred_two (35.0)
|   |   |   Height > 22: three_hundred_ten (35.0)
BB_Area > 1020
|   Height <= 45
|   |   Width <= 58: three_hundred_four (56.0)
|   |   Width > 58
|   |   |   Height <= 27: three_hundred_thirteen (28.0)
|   |   |   Height > 27
|   |   |   |   BPC  <= 530: three_hundred_four (11.0)
|   |   |   |   BPC  > 530
|   |   |   |   |   Height <= 29: three_hundred_thirteen
(5.0)
```

Training Accuracy : 91.8 %

```
Height <= 14
|   Width <= 14: three_hundred_twenty (36.0/1.0)
|   Width > 14: three_hundred_twentyone (35.0)
Height > 14
|   Width <= 14
|   |   Height <= 16: three_hundred_twentyfive (19.0)
|   |   Height > 16
|   |   |   Height <= 17
|   |   |   |   BB_Area <= 209
|   |   |   |   |   BPC <= 77: three_hundred_twentyfour
(10.0)
|   |   |   |   |   BPC > 77: three_hundred_twentyfive
(5.0/2.0)
|   |   |   |   BB_Area > 209: three_hundred_twentyfive
(18.0/5.0)
|   |   |   Height > 17: three_hundred_twentyfour (18.0)
|   Width > 14: three_hundred_twentythree (19.0)
```

Class Type	Total Samples	Correctly Recognized	% Accuracy
301	15	3	20%
302	15	2	13%
303	15	3	20%
305	15	15	100%
306	15	15	100%
307	15	15	100%
308	15	15	100%
309	15	15	100%
310	15	15	100%
311	15	0	0
312	15	3	20%
314	15	15	100%

Class Type	Total Samples	Correctly Recognized	% Accuracy
300	18	18	100
301	18	0	0
302	8	8	100
303	9	8	88
304	11	8	72
305	11	11	100
306	8	6	75
307	10	10	100
308	8	8	100
309	8	8	100
310	12	2	16
311	11	6	55
312	23	18	78
313	9	9	100
314	8	4	50

310	320		
303	320		
303	325		
325	310	323	
325	310	320	
306			
301	324		
308			

Tesseract vs Decision Trees

64.45% — 63.75 %