A Comprehensive Image Dataset of Urdu Nastalique Document Images

QURAT UL AIN AKRAM, ANEETA NIAZI, FARAH ADEEBA,
SABA UROOJ, SARMAD HUSSAIN, SANA SHAMS

National ICT R&D Fund

CLE Nastalique OCR

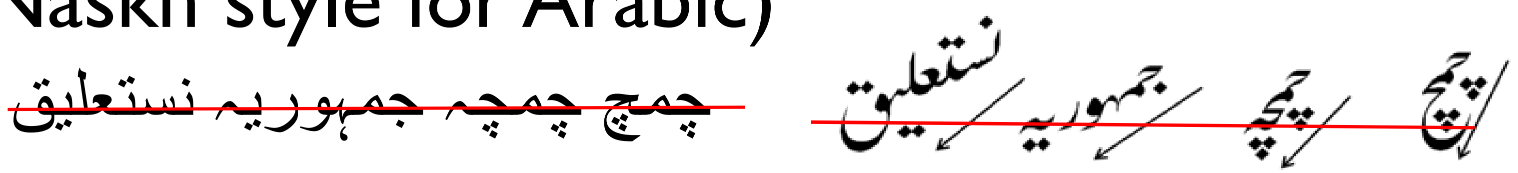
Al-Khawarizmi Institute of Computer Science, UET, Lahore

Road Map

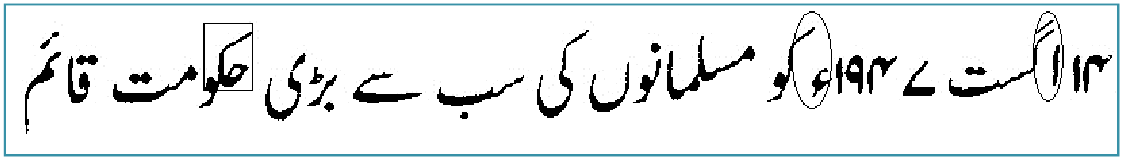
- Nastalique complexity
- Corpus development process
- Corpus statistics
- Conclusion

Challenges for Nastalique OCR

Diagonality and cursiveniess of Nastalique (vs. Naskh style for Arabic)



Character & ligature overlapping



Thick-thin stroke variation

Challenges of Nastalique OCR

- RASM joining

بونے نے کہا ''نہیں ملکہ صاحبہ ان میں سے بھی میرا

- Diacritics and RASM joining

بڑی تصویر میں ڈاکٹر صاحب ٹارچ سے

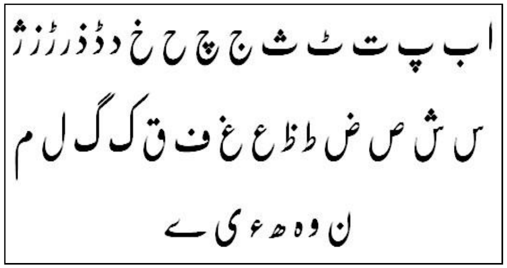
- Broken

Corpus development process

- Corpus collection
- Corpus development from books
- Corpus Organization

Corpus design

- **Design criteria**
  - **Character set and symbols**
    - Urdu alphabet given in Figure
    - Latin digits (0, 1, 2, 3, 4, 5, 6, 7, 8, 9)
    - English characters (A-Z, a-z)
    - Urdu digits (۰, ۱, ۲, ۳, ۴, ۵, ۶, ۷, ۸, ۹)
    - Urdu aerab (ٔ ٗ ِ ُ َ ً)
    - Other symbols of Urdu, as follows:
  - (ؔ ؐ ؏ ؑ ؒ ؓ ؟ () ؛ - " ' ، :)
  - **Font style and font size**
    - Noori Nastalique font style
    - 14-40 font size



Corpus design

- **Categories based on font size**
  - **<u>Normal text books</u>** (having normal text printed at most frequent 14 and 16 font sizes)
  - **<u>Children books</u>** (images having normal text in children and poetry books cover18 to 22 font sizes)
  - **<u>Headings</u>** (images having headings text covering 24-40 font sizes)
- **Multiple Domains**

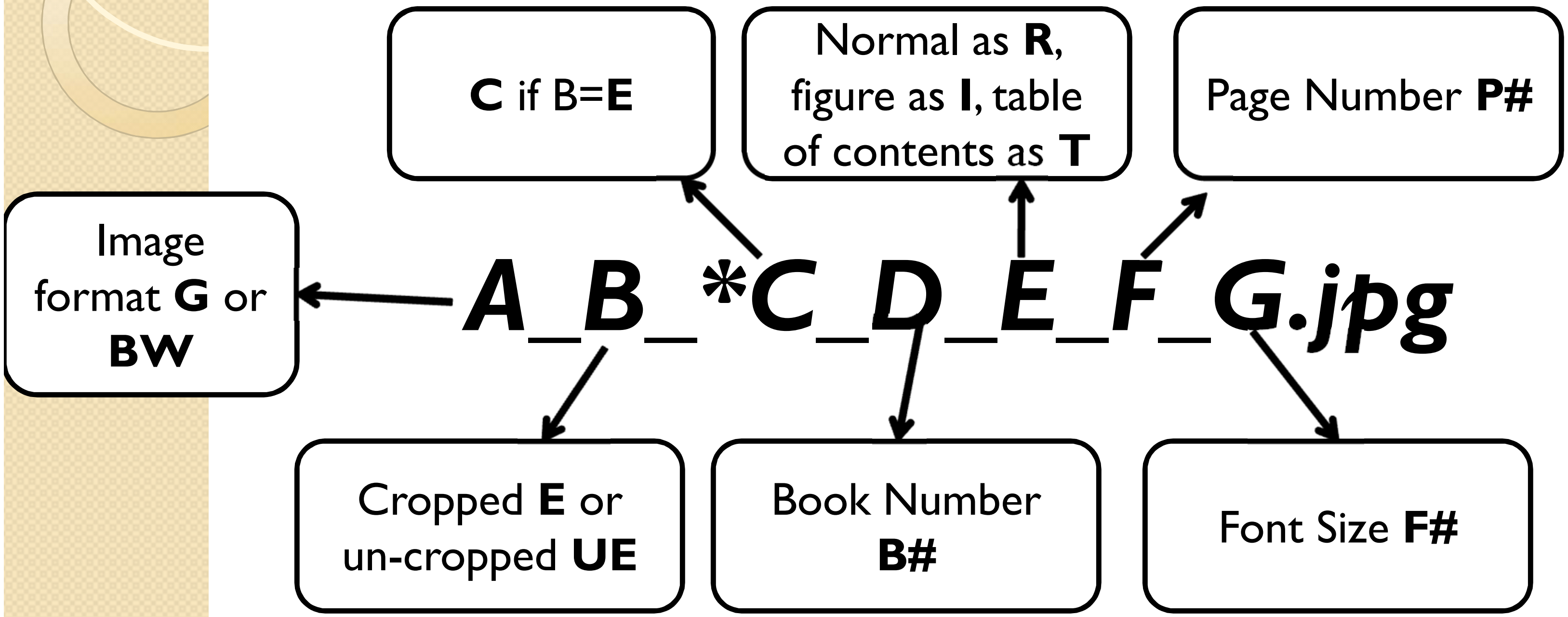
  Multiple domains for each font size category to address the coverage of a balanced corpus
- **Publishers and Publication Date Variety**
  - Multiple publishers of different cities
  - Within a city, variety of publishers
  - Publication date
- **Page/Printing Quality**

Corpus development from books

- Normal text books
  - 100 books for each font size (14 and 16)
  - Five pages, table of content (TOC) and image/figure
- Children books
  - 30 books for each font size (18, 20, 22)
  - Five pages from each book
- Headings
  - 20 books for each font size(24, 28, 32, 36 and 40)
  - 10 headings
- Scanning process
  - 300 DPI
  - Black and White (BW) (cropped and un-cropped)
  - Gray Scale (G) (cropped and un-cropped)
  - Heading textual area is cropped

Corpus Organization

Naming convention of images

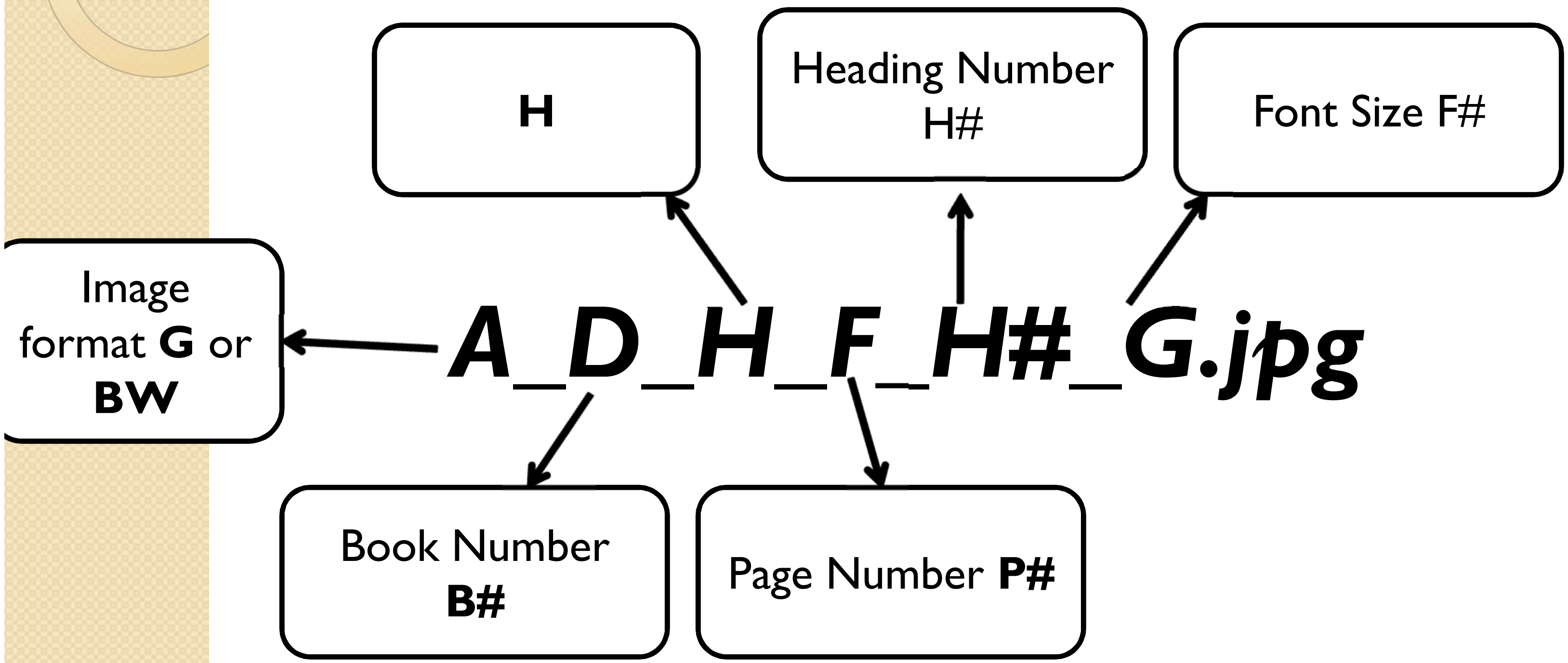


e.g.

BW_UE_B13_R_P26_F14.jpg

Corpus Organization

Naming convention for Headings



e.g.

G_B149_H_P34_H1_F32.jpg

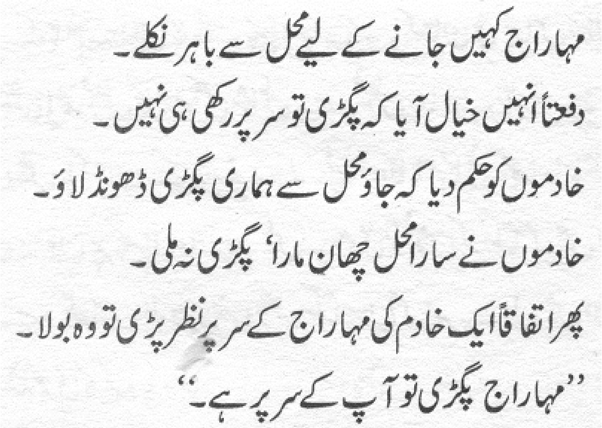
Metadata

Book ID	Total number of pages	Print quality (Poor, Average, Good)	Transparency (1-transparent, 3-not transparent)	Paper Quality(1-rough, 3-smooth)	Domain	Book Name	Author Name
B01	150	1	2	2	Poetry	نغمہء آب	ڈاکٹر این گوپی
B01	150	1	2	2	Poetry	نغمہء آب	ڈاکٹر این گوپی
B01	150	1	2	2	Poetry	نغمہء آب	ڈاکٹر این گوپی
B01	150	1	2	2	Poetry	نغمہء آب	ڈاکٹر این گوپی
B01	150	1	2	2	Poetry	نغمہء آب	ڈاکٹر این گوپی

Book ID	Publisher	Year of Publication	Country, City	Page number	File Name
B01	نوشین پبلیکشنز	2005	India, Hyderabad	Image	G_UE_B01_I_P_F14_F22
B01	نوشین پبلیکشنز	2005	India, Hyderabad	102	G_UE_B01_R_P102_F14
B01	نوشین پبلیکشنز	2005	India, Hyderabad	103	G_UE_B01_R_P103_F14
B01	نوشین پبلیکشنز	2005	India, Hyderabad	104	G_UE_B01_R_P104_F14
B01	نوشین پبلیکشنز	2005	India, Hyderabad	105	G_UE_B01_R_P105_F14
B01	نوشین پبلیکشنز	2005	India, Hyderabad	107	G_UE_B01_R_P107_F14

Text corpus of images (ground truth)

- Typed by two typist
- Typo Mistakes in image remains as is.
- 2,843 images are typed

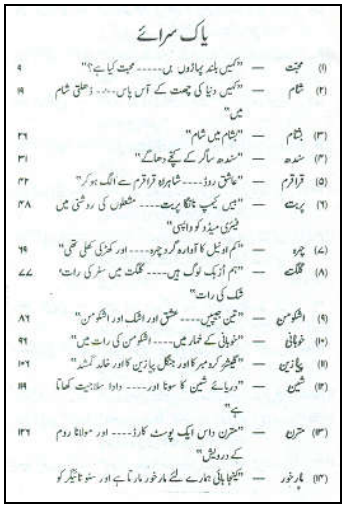

مہاراج کہیں جانے کے لیے محل سے باہر نکلے۔
دفعتاً انہیں خیال آیا کہ پگڑی تو سر پر رکھی ہی نہیں۔
خادموں کو حکم دیا کہ جاؤ محل سے ہماری پگڑی ڈھونڈ لاؤ۔
خادموں نے سارا محل چھان مارا' پگڑی نہ ملی۔
پھر اتفاقاً ایک خادم کی مہاراج کے سر پر نظر پڑی تو وہ بولا۔
''مہاراج پگڑی تو آپ کے سر پر ہے۔''

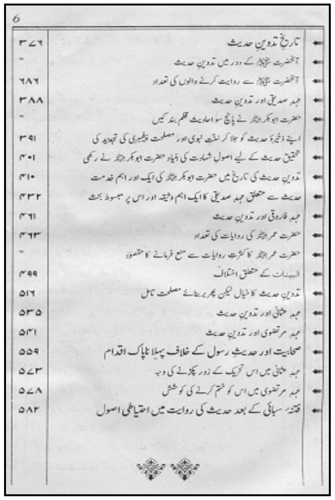
مہاراچ کہیں جانے کے لیے محل سے باہر نکلے-
دفعتاً انہیں خیال آیا کہ پگڑی تو سر پر رکھی ہی نہیں-
خادموں کو حکم دیا کہ جاؤ محل سے پگڑی ڈھونڈ لاؤ-
خادموں نے سارا محل چھان مارا پگڑی نہ ملی-
پھر اتفاقاً ایک خادم کی نظر مہاراج کے پر پڑی تو وہ بولا-
مہاراج پگڑی تو آپ کے سر پر ہے "-

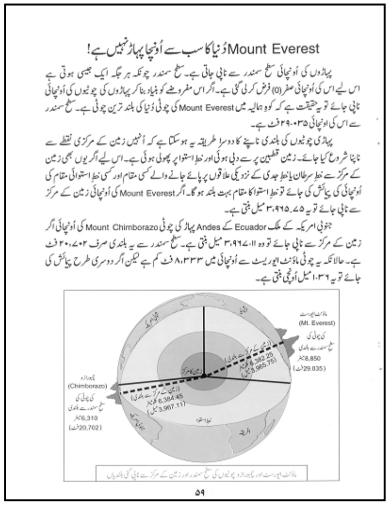
↔

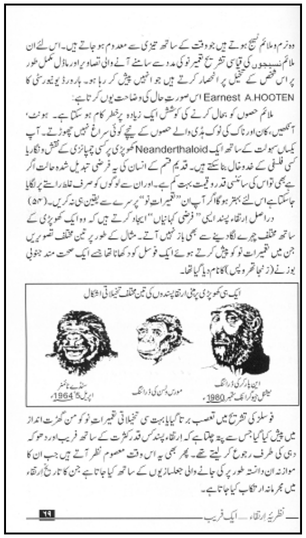
مہاراج کہیں جانے کے لیے محل سے باہر نکلے
دفعتاً انہیں خیال آیا کہ پگڑی تو سر پر رکھی ہی نہیں-
خادموں کو حکم دیا کہ جاؤ محلسے پگڑی ڈھونڈ لاؤ-
خادموں نے سارا محل چھان مارا پگڑی نہ ملی-
پھر اتفاقاً ایک خادم کی نظر مہاراج کے پر پڑی تو وہ بولا-
مہاراج پگڑی تو آپ کے سر پر ہے "-

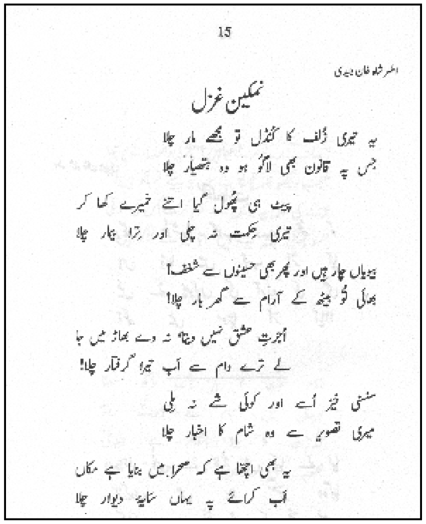
Image samples

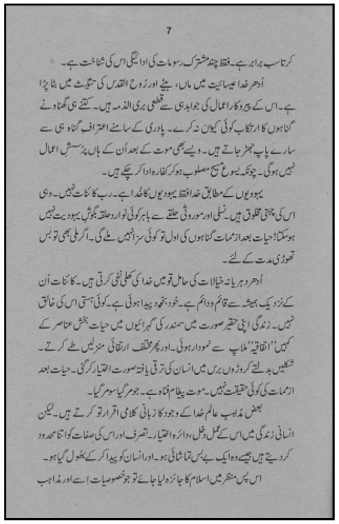

پاک سرائے




Mount Everest دنیا کا سب سے اُونچا پہاڑ نہیں ہے!




نمکین غزل



قومی اسمبلی سے خطاب

Corpus statistics

Font Size	Book/ Magazine count	Number of document images	Domains	Authors
14	101	593	18	76
16	116	595	19	100
18	30	150	10	23
20	45	149	2	24
22	56	151	2	21
24	21	461	18	24
28	21	202	6	21
32	23	186	9	21
36	31	226	7	22
40	26	199	7	22

Font wise letters, digits, aerab and symbols statistics

Font size	Total Char	Unique Urdu Char	Unique Urdu Digits	Unique English Char	Unique Latin Digits	Unique Urdu Aerab	Unique Symbols
14	726,385	45	10	52	10	12	32
16	579,730	45	10	51	10	13	31
18	111,178	44	10	34	10	10	22
20	101,559	44	10	20	10	9	15
22	81,718	43	8	3	10	10	18
24	7,807	43	3	10	6	7	18
28	2,730	42	3	16	4	6	12
32	5,519	40	0	4	10	9	11
36	3,462	42	4	11	3	7	13
40	2,949	42	0	0	0	9	12

Font wise lines and ligature statistics of corpus

Font Size	Total Document images	Lines	Total Lig.	Unique Lig.	Avg. Lines per image	Avg. Lig. per Line
14	591	13,712	386,648	6,452	23	28
16	528	11,080	306,080	5,938	20	27
18	150	2,622	60,056	2,872	18	23
20	149	2,318	54,657	2,204	16	24
22	151	1,857	43,121	1,865	12	23
24	461	463	3,961	883	1	9
28	202	203	1,424	502	1	7
32	186	274	2,874	616	2	11
36	226	260	1,776	537	1	7
40	199	222	1,510	498	1	7

Conclusion

- A total of 2, 912 images from 413 books
- Image corpus and parallel text corpus are publically available at www.cle.org.pk/clestore

THANKS