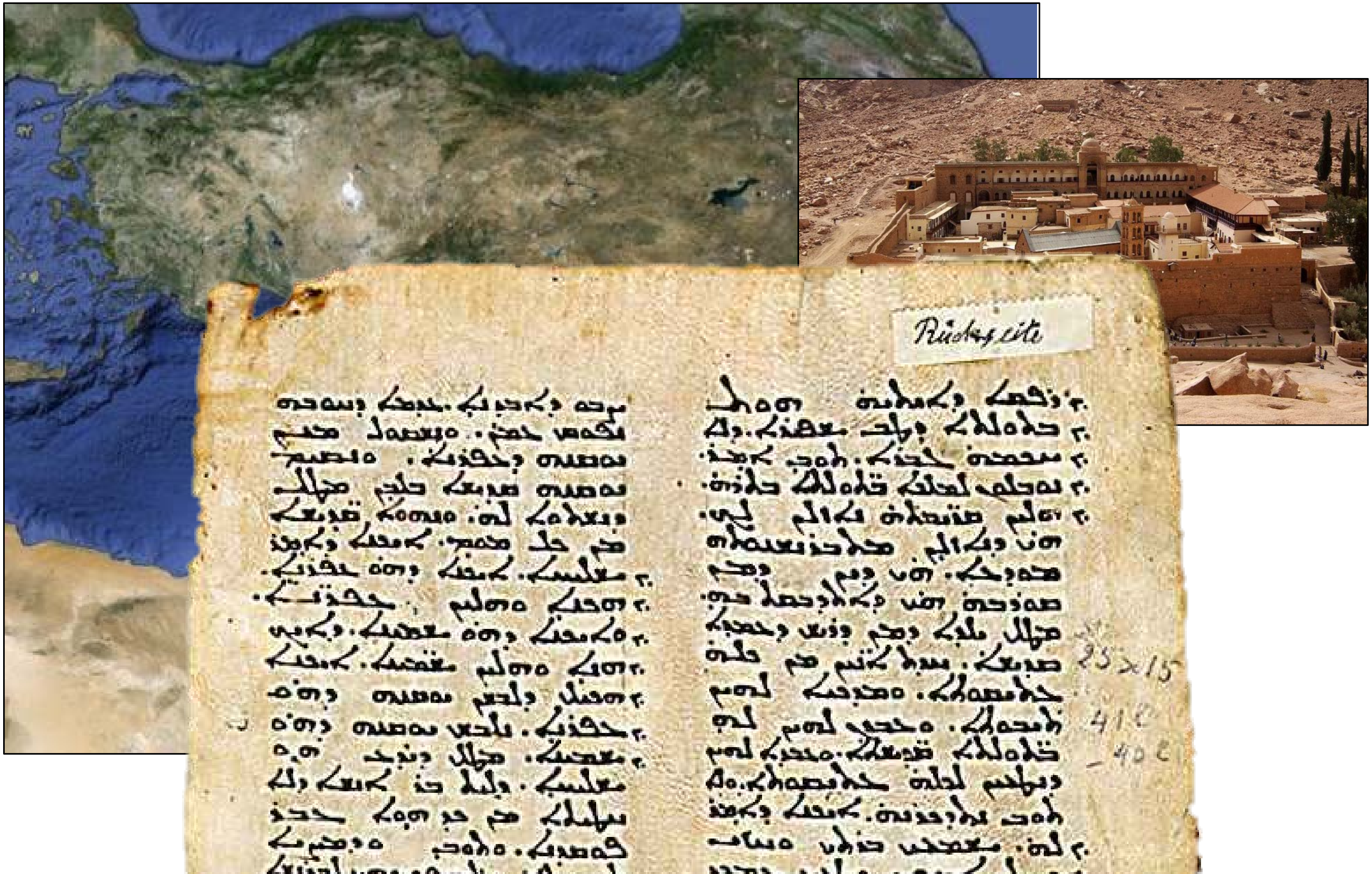
A Character Style Library for Syriac Manuscripts

Nicholas Howe
Smith College
nhowe@smith.edu

Alice Yang
Smith College
ayang@smith.edu

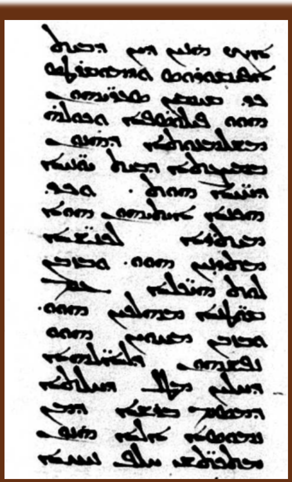
Michael Penn
Mt. Holyoke College
mpenn@mtholyoke.edu

Syriac & Early Christianity

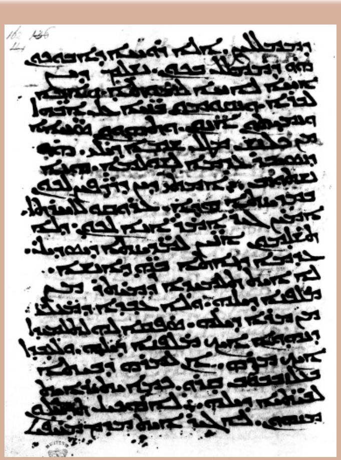


Paleography

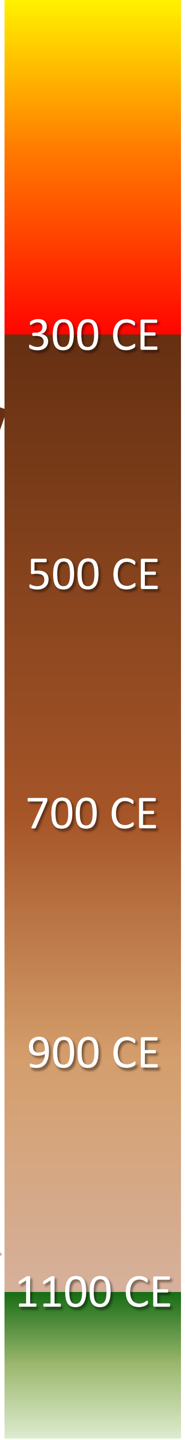
- Most documents not securely dated
- Writing style changes over time
- Some documents have known dates
- Use these to calibrate dates of others



BL Add. 12150: 411 CE

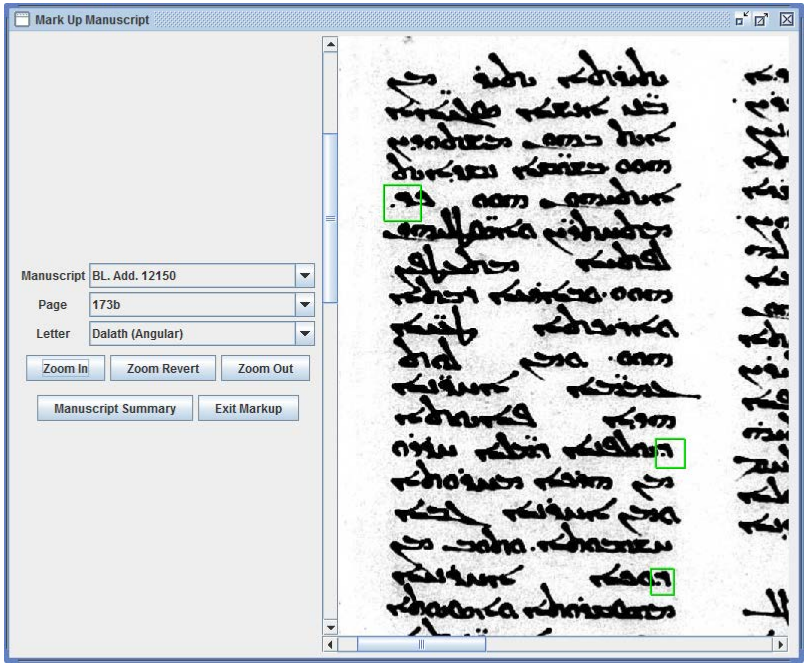


BL Add. 14490: 1089 CE



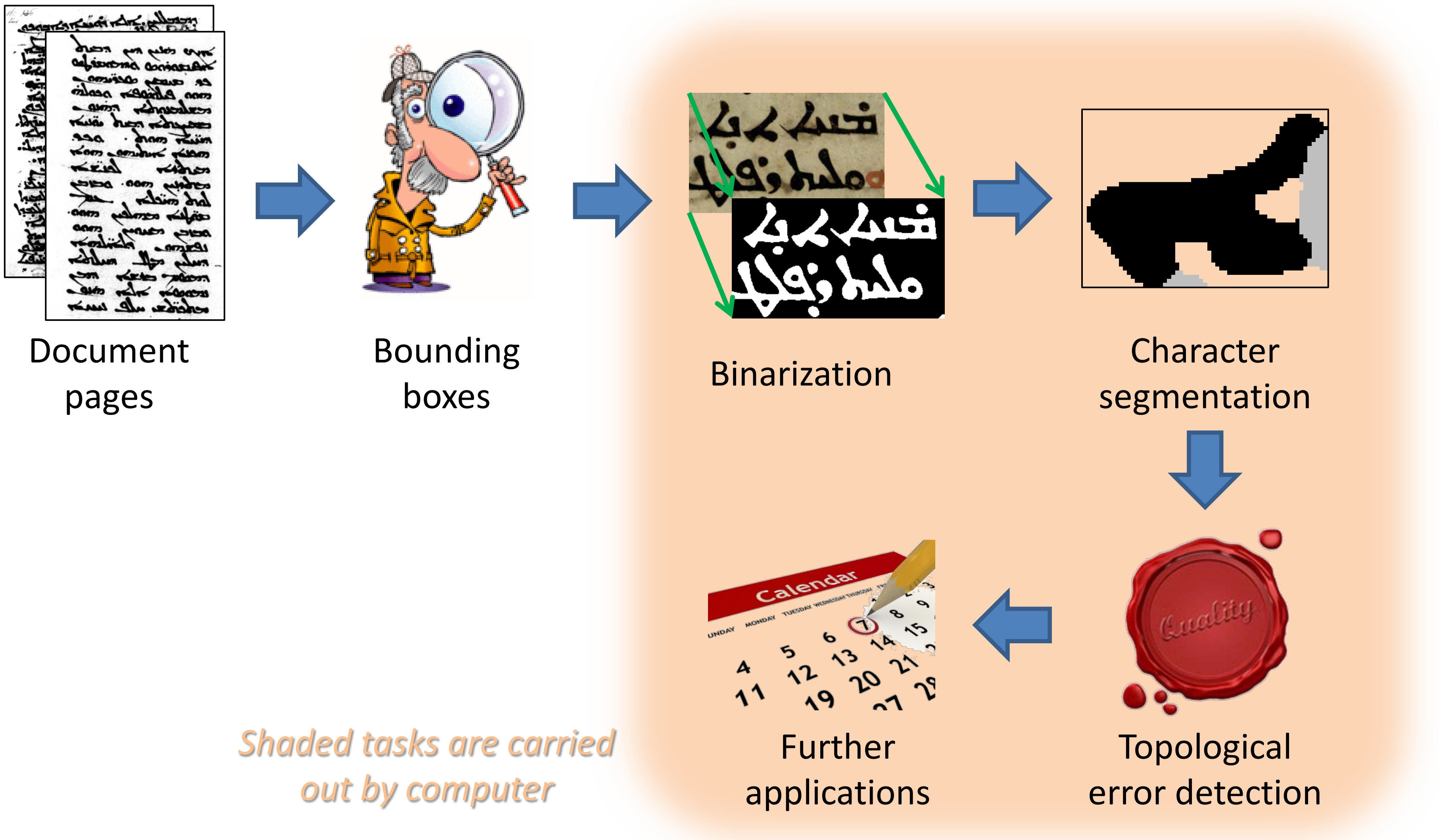
Human Annotation

- Humans identify character samples
- 5 per document per character
- Bounding box only
  - – More detail too time-consuming
  - – Need automatic character segmentation



Manuscript: BL. Add. 12150
Page: 173b
Letter: Dalath (Angular)

Workflow Sketch



Document pages

Bounding boxes

Binarization

Character segmentation

Topological error detection

Further applications

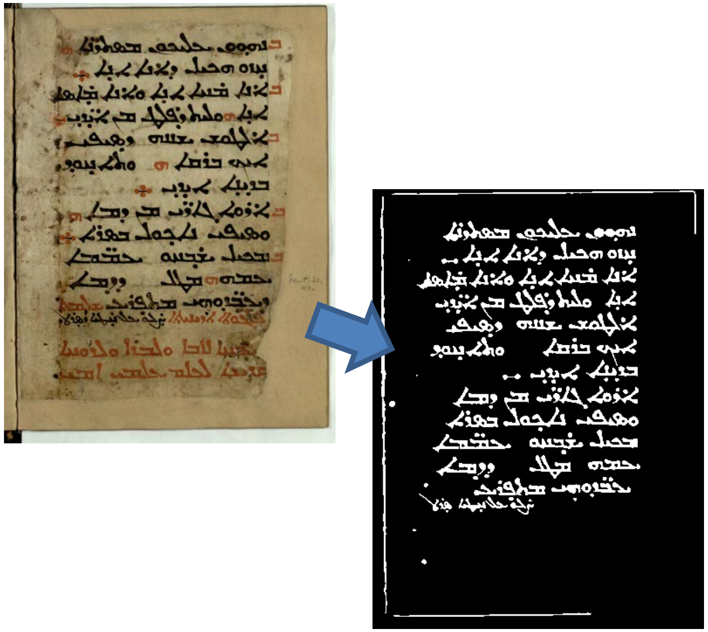
Shaded tasks are carried out by computer

Computational Steps

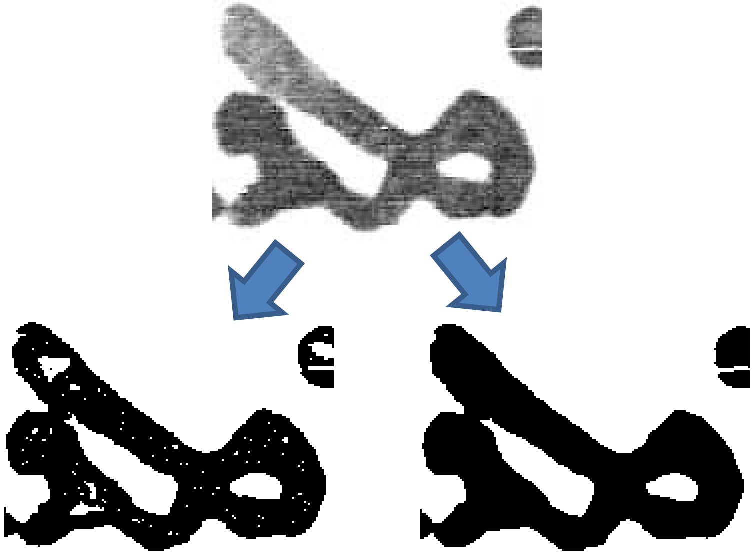
1. Binarization
   - Take heterogeneous sources to known format
   - Uses Howe’s binarization (Laplacian energy min.)
2. Character segmentation
   - Connected letters make problem tricky
   - Evaluated two part-structured models
3. Postprocessing/quality control
   - Possible to detect errors in prior stages

Binarization

- Most documents binarize well – automatic
- Two problem areas: red text & high resolution



Red text lost

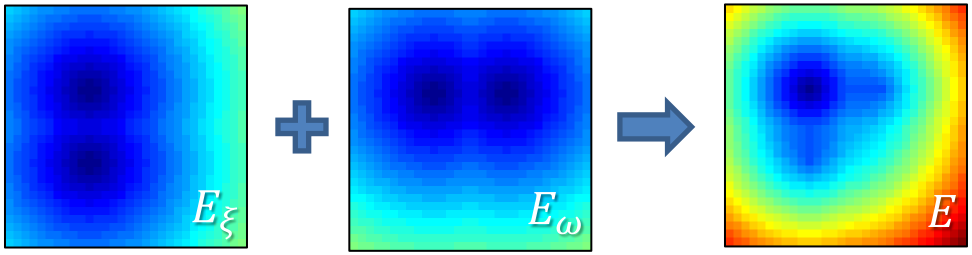


Without smoothing

With smoothing

Part-Structured Models

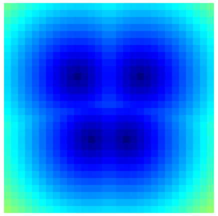
- Complex model is made of simple parts in a spatial relationship
- Proposed layout of parts is a ***configuration***
- Likelihood of configuration has two factors:
  - Do observations support layout of parts? E_ω
  - Does layout of parts match expected offsets? E_ξ



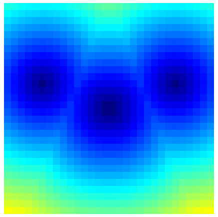
$$E = E_\xi + \lambda E_\omega$$

Part-Structured Localization

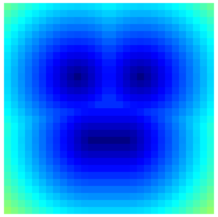
- Part detectors do some localization



Eyes

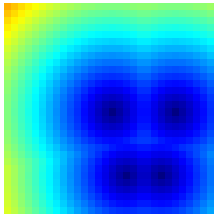


Nose

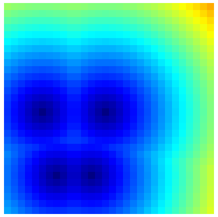


Mouth

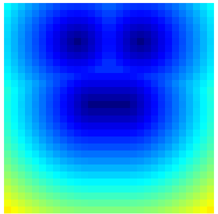
- Offset detections and combine



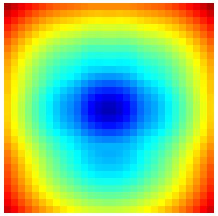
Left eye to nose



Right eye to nose



Mouth to nose

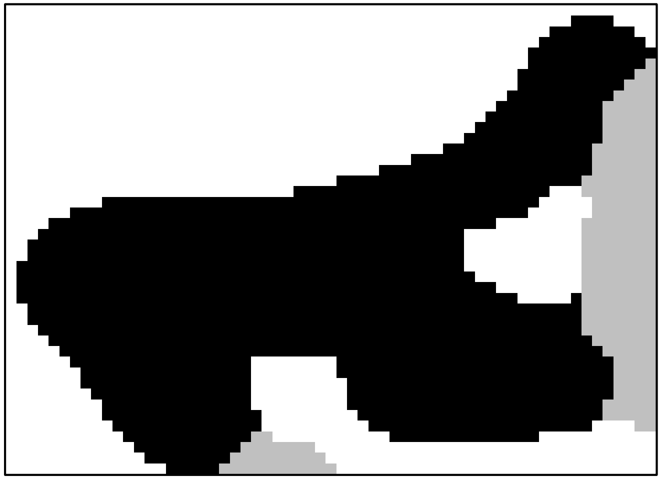


Combined nose likelihood

Accounting for subordinate parts clarifies nose position

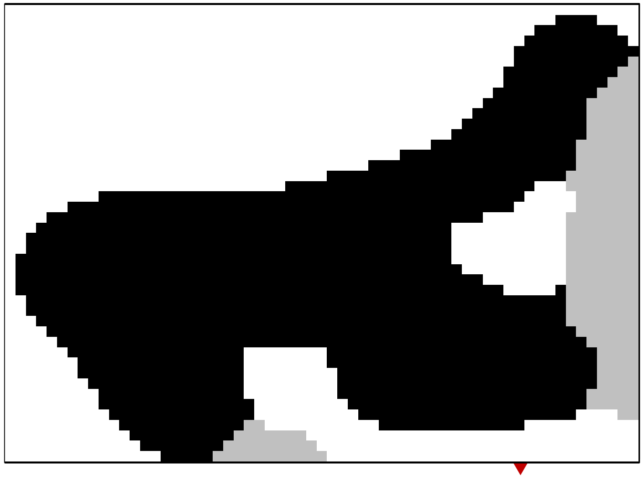
Given nose position, can place subordinate parts

Model #1: Inkballs



- Parts are disks of ink placed on medial axis
- Model built from sample character
- Matching & segmentation:
  - Find minimal energy configuration
  - Render model to classify medial points
  - Attribute pixels based on nearest medial point

Model #2: Boundary Trace



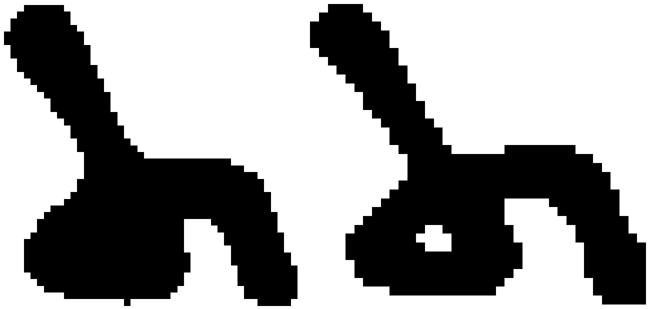
- Parts are oriented edge segments
- Arranged in double ring around letter
- Matching & segmentation:
  - – Find minimal energy
  - – Identify closed loop
  - – Attribute points

Similar to active contours/snakes except:

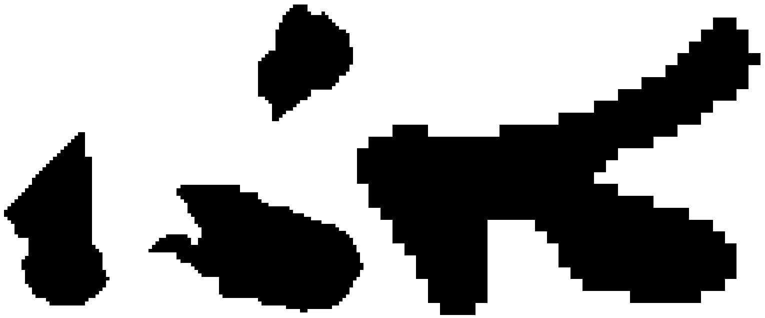
- *Prior on shape from model character*
- *Direct optimization*

Automatic Quality Control

- Topological considerations catch some errors
  - *Unimplemented: feedback to binarization step*



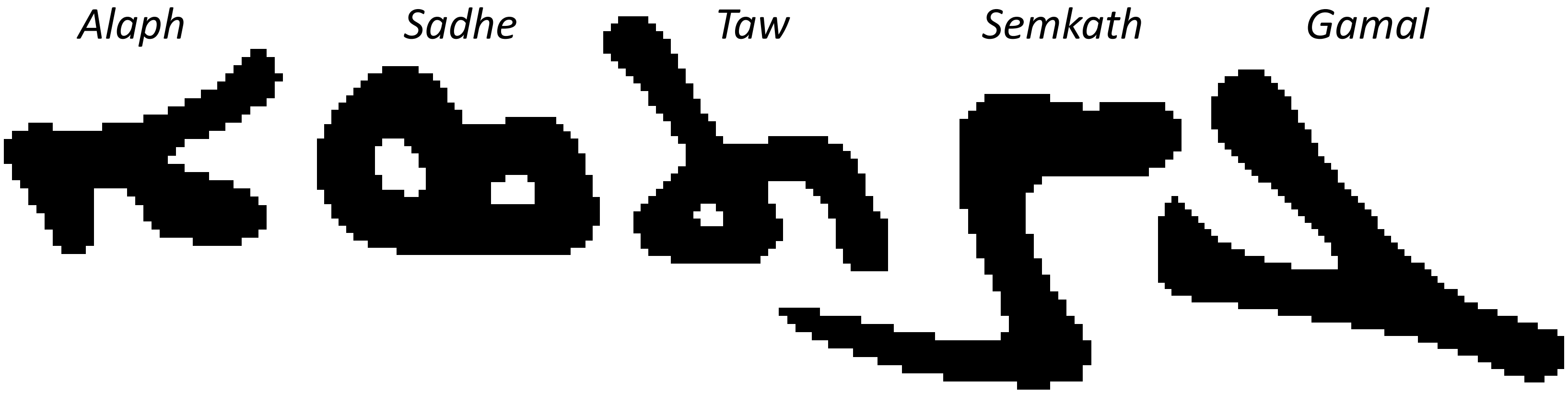
Filled holes



Broken characters

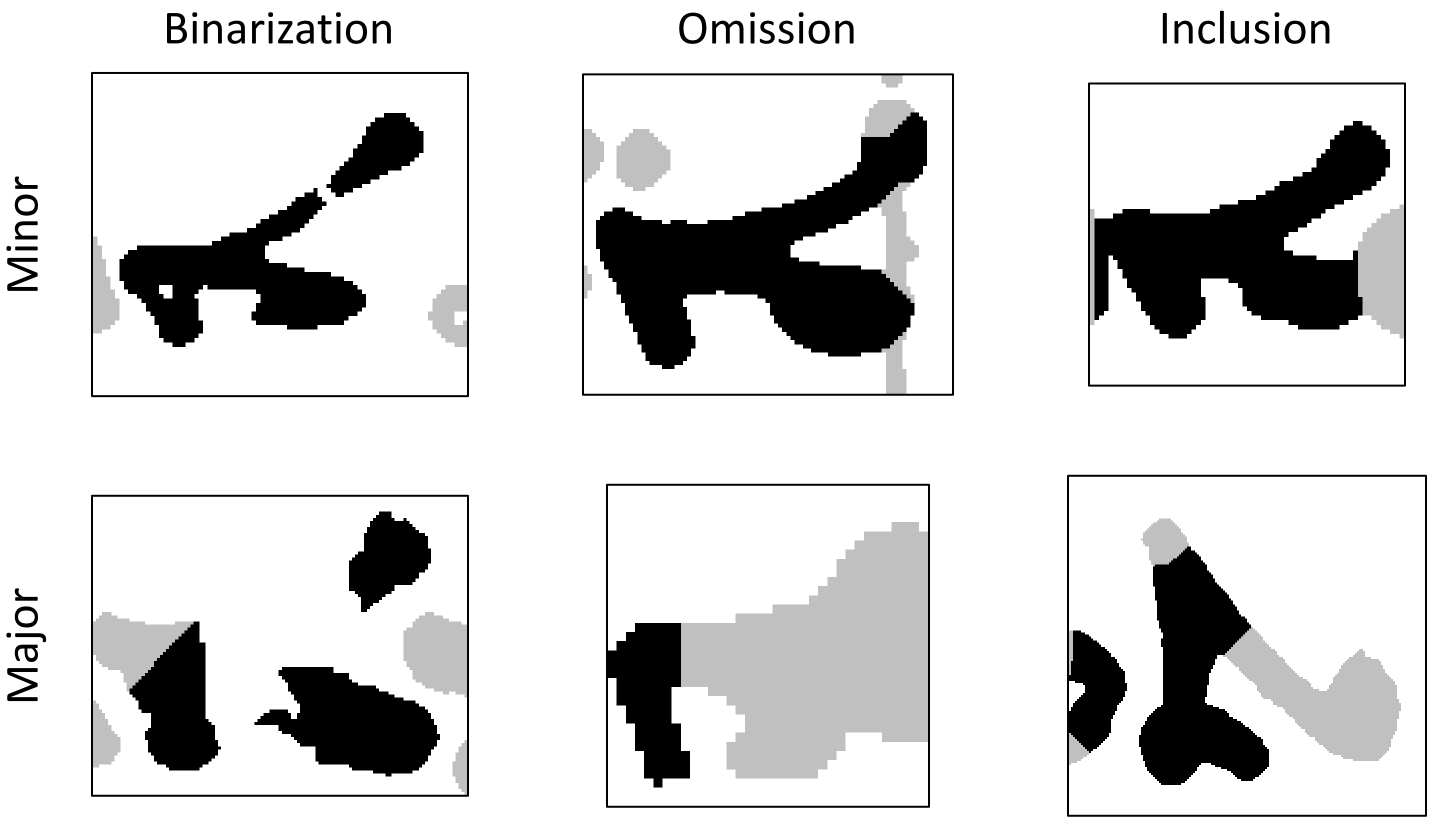
Manual Evaluation

- Evaluating 60,000 results is impractical
- Selected 1 sample per document x 5 letters



- Human expert rated quality of each result

Error Types



Binarization

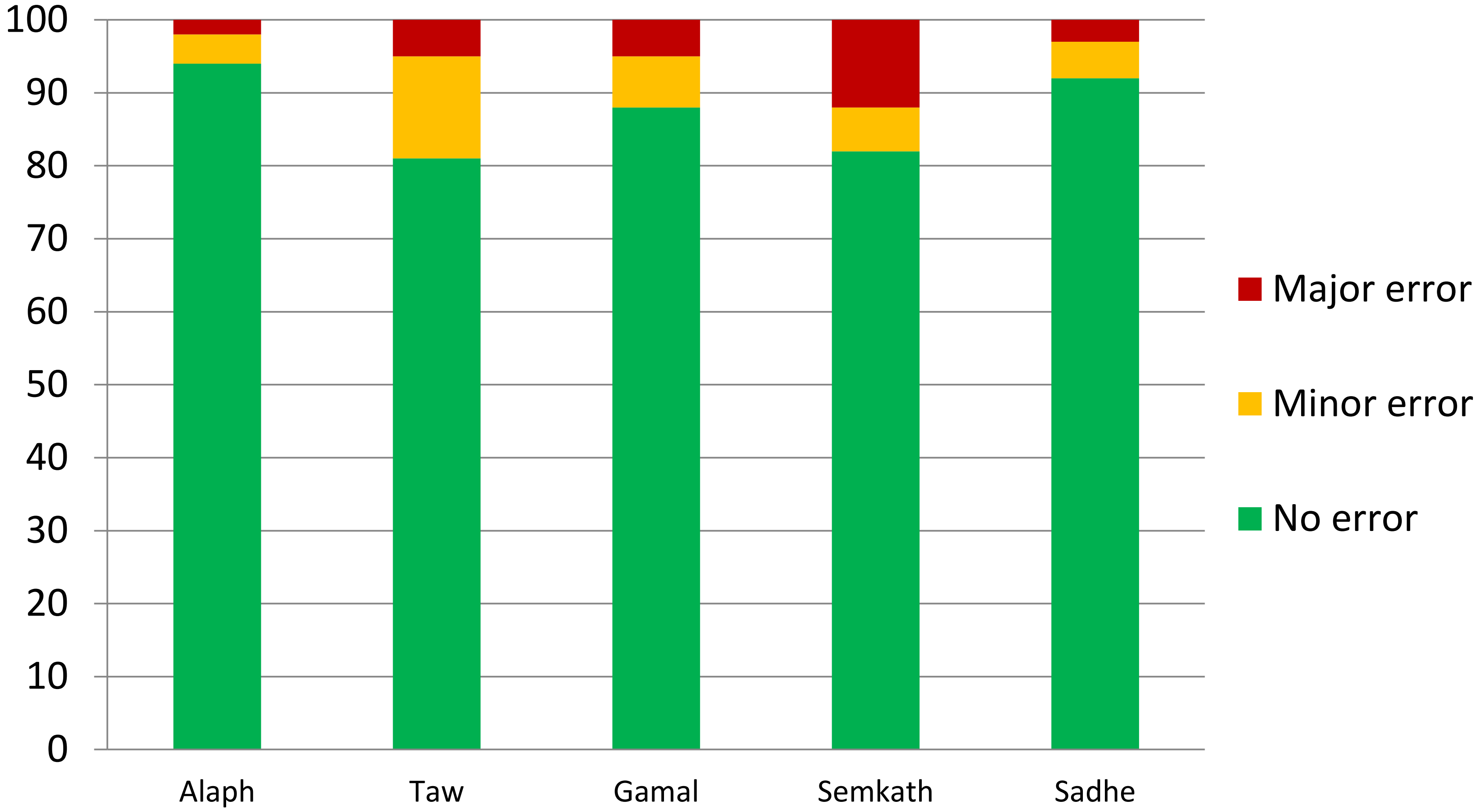
Omission

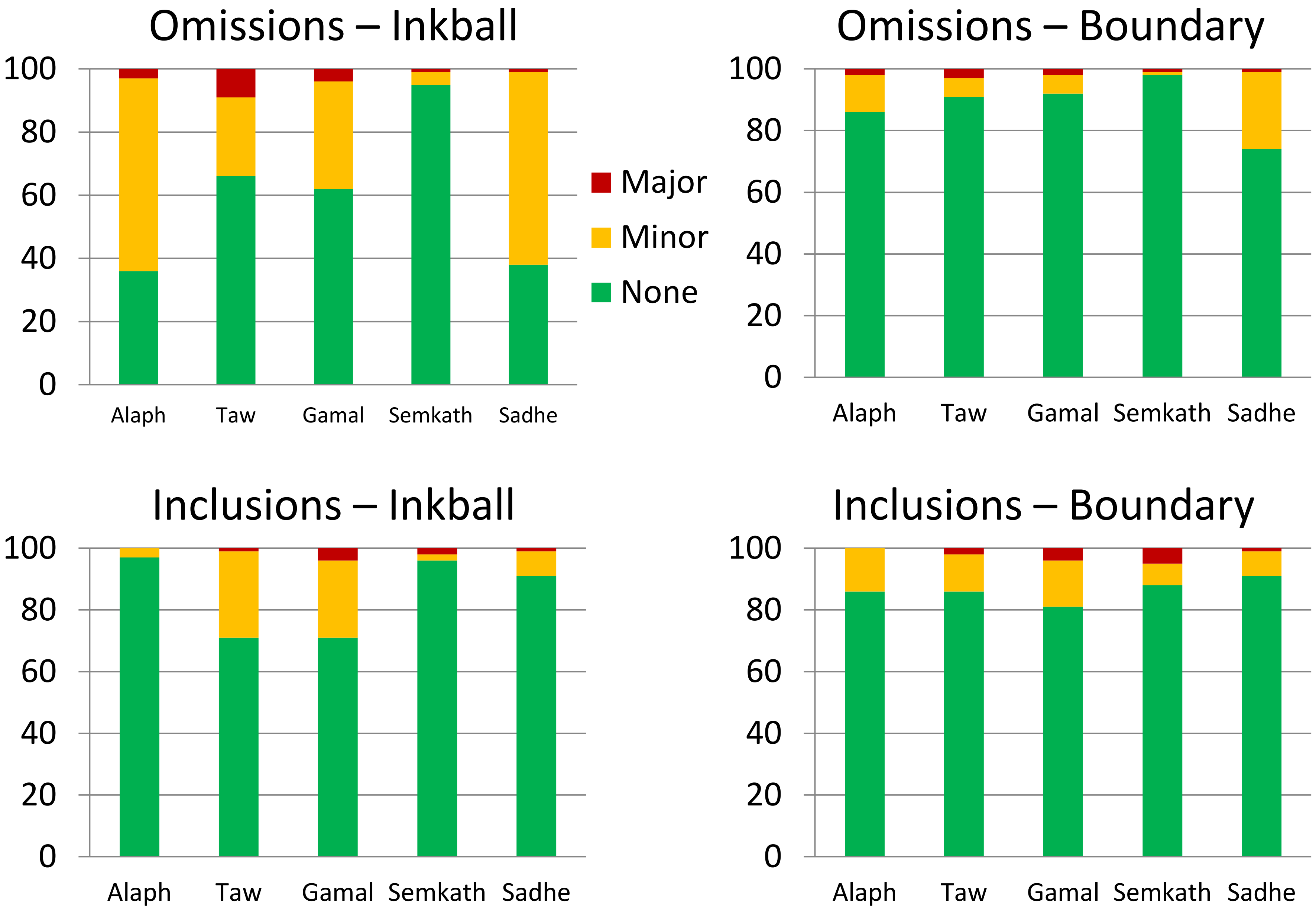
Inclusion

Minor

Major

Results – Binarization Quality





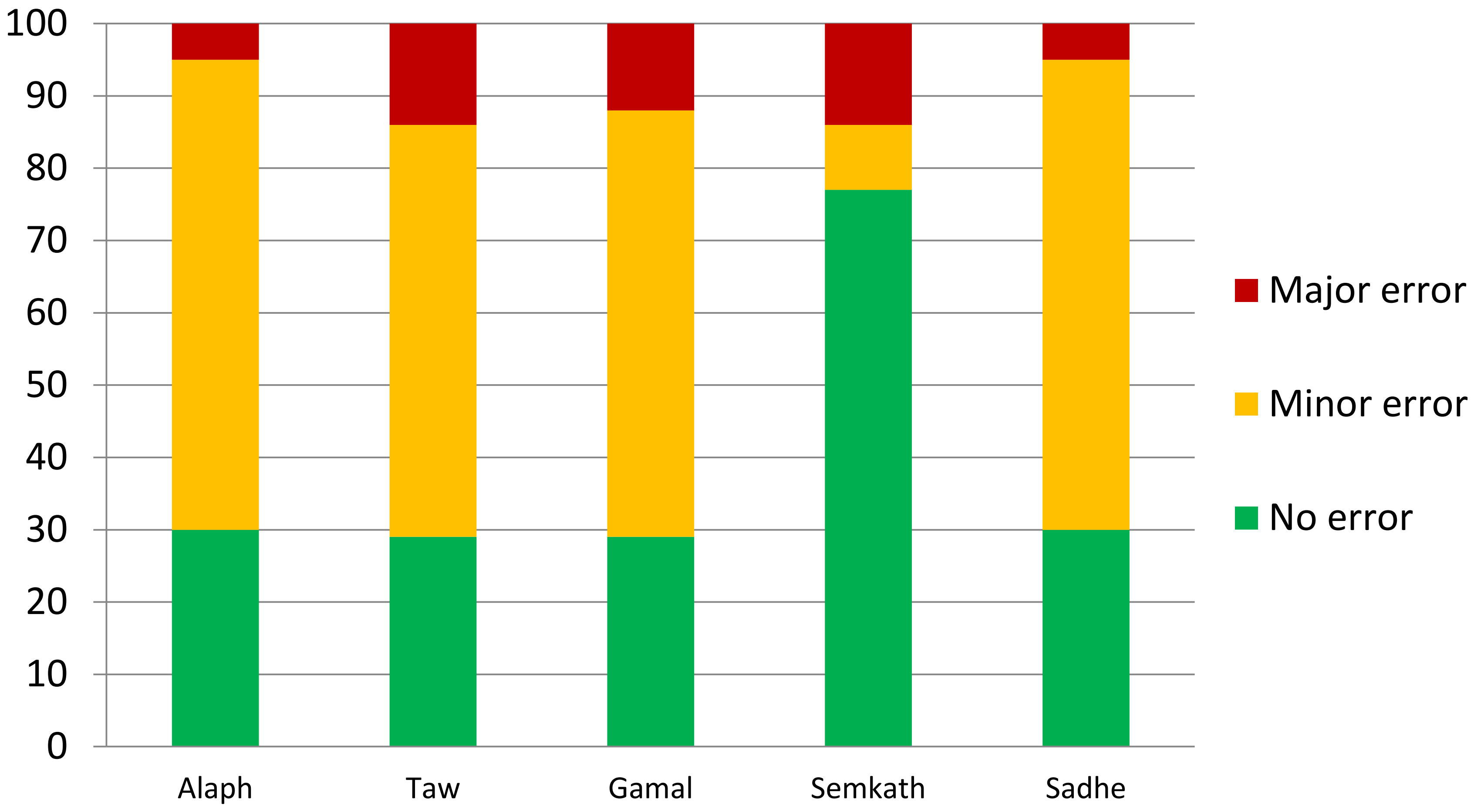
Omissions – Inkball

Omissions – Boundary

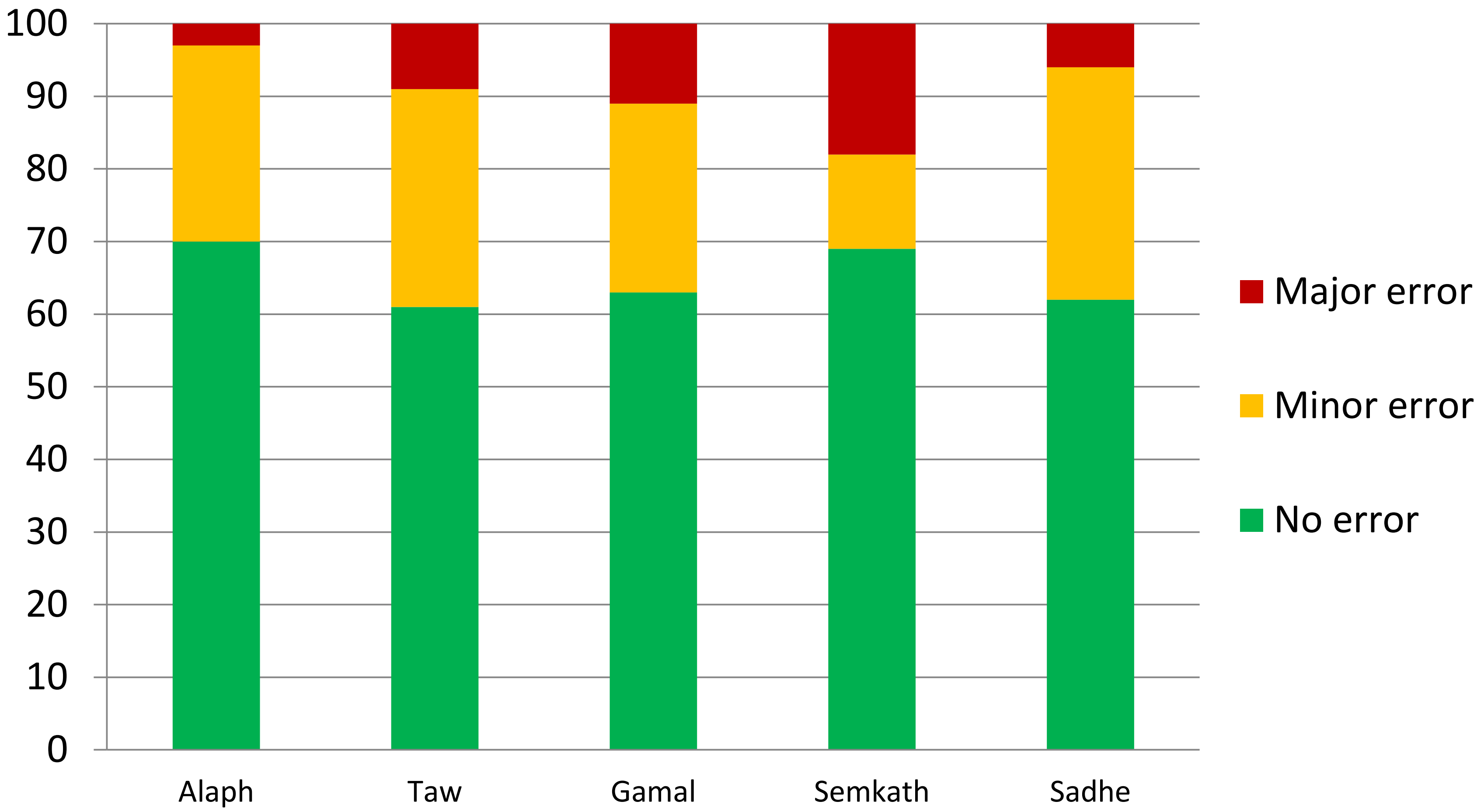
Inclusions – Inkball

Inclusions – Boundary

Overall Results – Inkball



Overall Results – Boundary



Conclusion

- Mixture of human and machine effort

+

- Boundary models give best results:
  - At least 60% of samples are error-free
  - Fewer than 20% show major errors
  - 5 samples/character/document → likely one is good
- Next step: **paleographic dating**