Workshop on Computational Methods in the Humanities 2018 (COMHUM 2018), Lausanne.

Clustering Writing Components from Medieval Manuscripts

UPPSALA UNIVERSITET

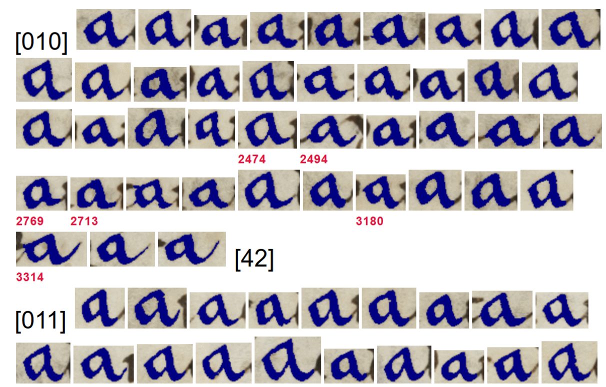
Mats Dahllöf

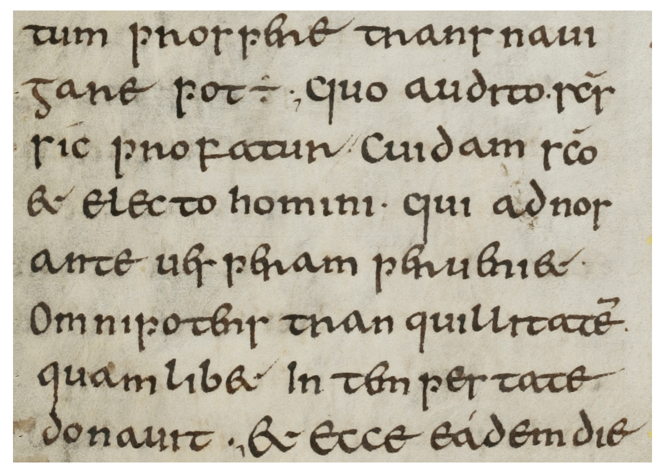
mats.dahllof@lingfil.uu.se

Department of Linguistics and Philology

Purpose

- **Extraction** and **clustering** of **writing components.**
- Partial transcription in combination with human annotation of the clusters.
- Qualitative palaeographic analysis.





[Schaffhausen SB, Gen. 1, e-codices]

“Pipeline”

- **Extraction of components**
  - Binarization (ink-background separation)
  - Connected component labelling
  - Segmentation
- **Clustering**
  - Core clustering
  - Removal of small clusters
  - Extension by classification

In experiments, we took 20 000 **components.**

Overview of clusters, Gen. 1

1 *1715* 4.7	**2** *1171* 5.9	**3** *827* 4.6	**4** *755* 6.0	**5** *675* 4.0	**6** *667* 4.6	**7** *460* 6.8
8 *406* 4.1	**9** *306* 4.2	**10** *289* 4.1	**11** *273* 6.7	**12** *271* 3.6	**13** *252* 5.8	**14** *251* 5.9
15 *204* 4.3	**16** *112* 5.4	**17** *101* 4.8	**18** *90* 6.8	**19** *65* 5.5	**20** *54* 6.7	

In cells: “central” instance, number, size, and width.
From experiment with our “baseline” settings.

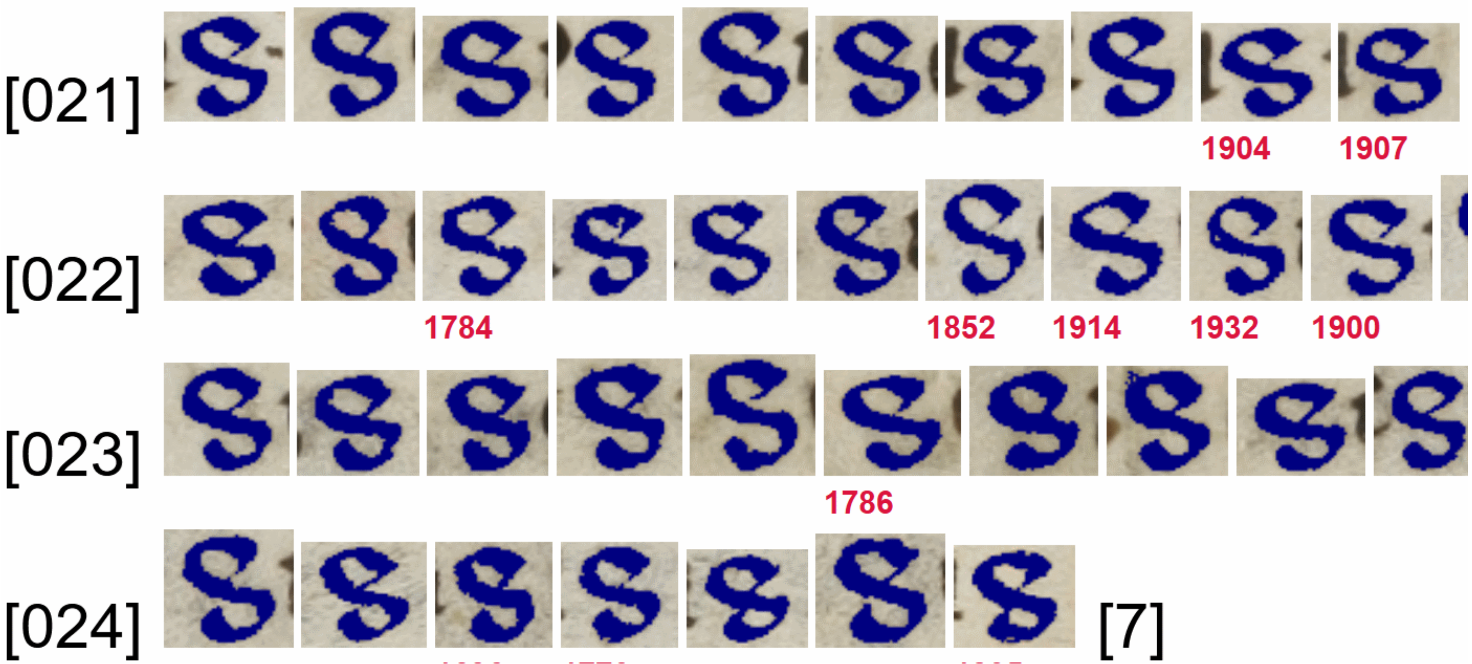
Gen. 1, ⟨a⟩ whole cluster

Size=1171. (27 p.) High precision for ⟨a⟩.

Gen. 1, ⟨a⟩ cluster, subset

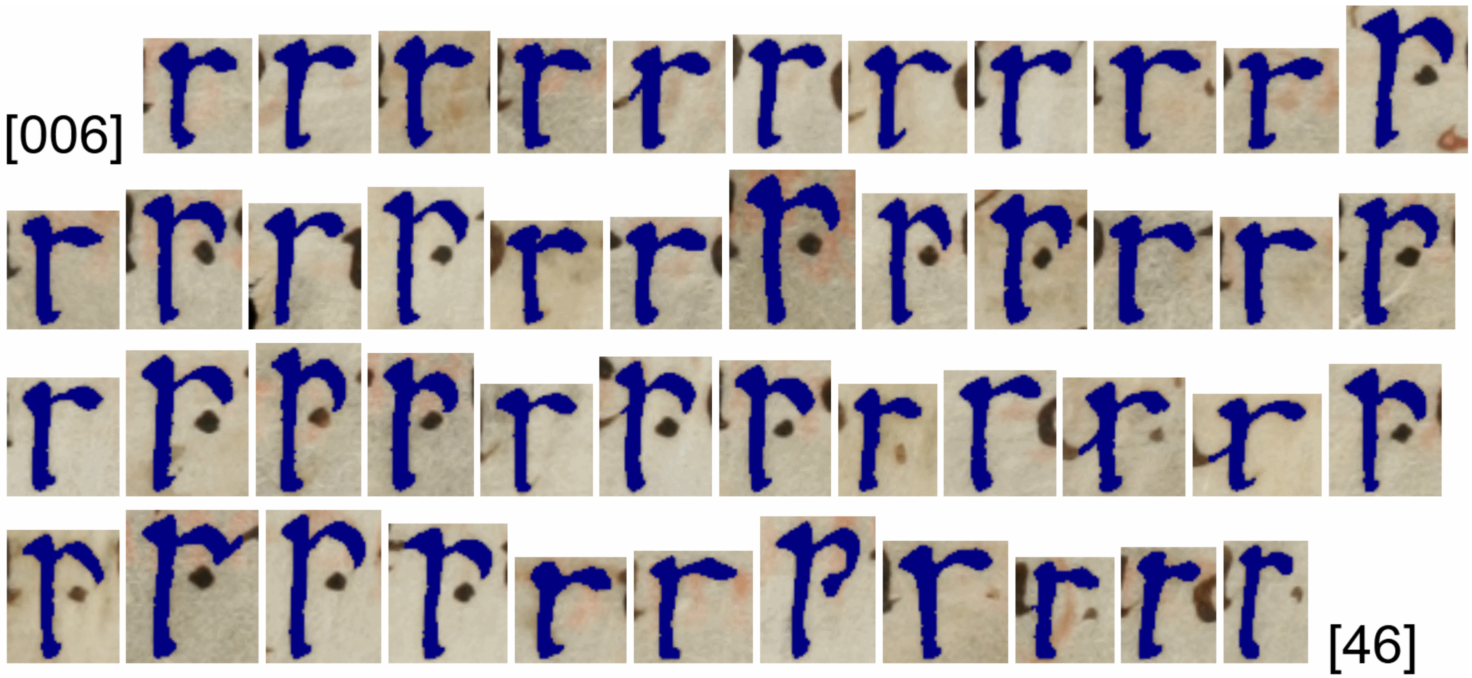
Size=1171. (27 p.) High precision for ⟨a⟩.

Gen. 1, ⟨s⟩ (one allograph) cluster



Size=101. 100% precision for this ⟨s⟩ allograph.

Gen. 1, largest cluster



Mixture of ⟨p⟩ and ⟨s⟩ (another allograph).

Another example: C 61 (b), clusters

1 *2889* 5.5	**2** *1594* 3.7	**3** *1180* 4.2	**4** *905* 6.3	**5** *511* 3.3	**6** *508* 3.8	**7** *461* 5.8
8 *456* 3.9	**9** *436* 3.2	**10** *261* 5.1	**11** *237* 4.1	**12** *205* 6.6	**13** *158* 6.8	**14** *152* 3.8
15 *148* 3.6	**16** *134* 6.9	**17** *130* 5.0	**18** *117* 4.3	**19** *114* 6.1	**20** *111* 7.1	**21** *109* 3.3
22 *94* 7.5	**23** *88* 5.3	**24** *63* 6.0	**25** *60* 6.6	**26** *52* 5.7	**27** *41* 7.5	

In cells: “central” instance, number, size, and width.

C 61 (b), cluster for ⟨ä⟩

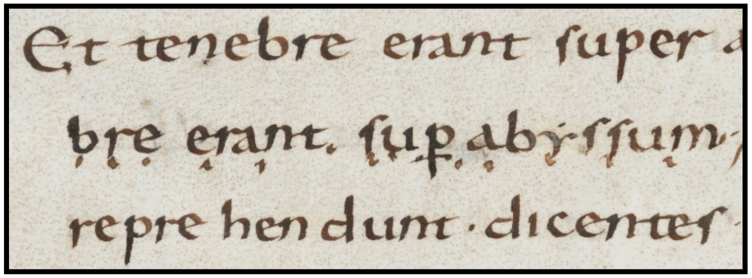
How to evaluate clustering results?

- **For which categories** are clusters established?
- **Precision** of a cluster: fraction of elements belonging to the right category.
- **Recall**: fraction of real instances of the right category assigned to the cluster.

Component Extraction

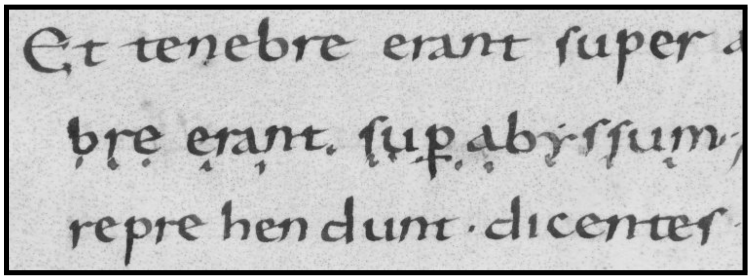
- **Binarization** (ink-background separation).
- **Connected component labelling** finds regions of ink.
- **Segmentation**, guided by estimated stroke width, w_s. (w_s is estimated as the most common width of sequences of continuous horizontal ink pixels separated by at least two pixels of background.)

Binarization



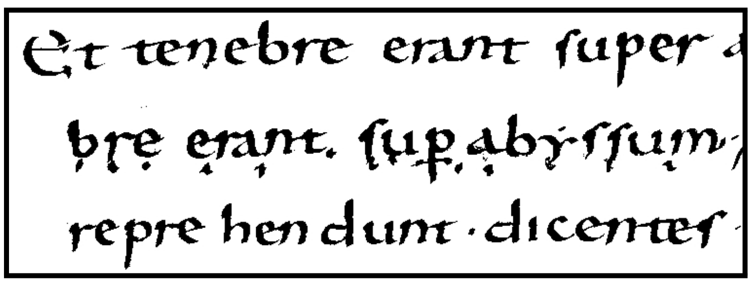
colour

(red, green, blue)



greyscale

(dark–light)



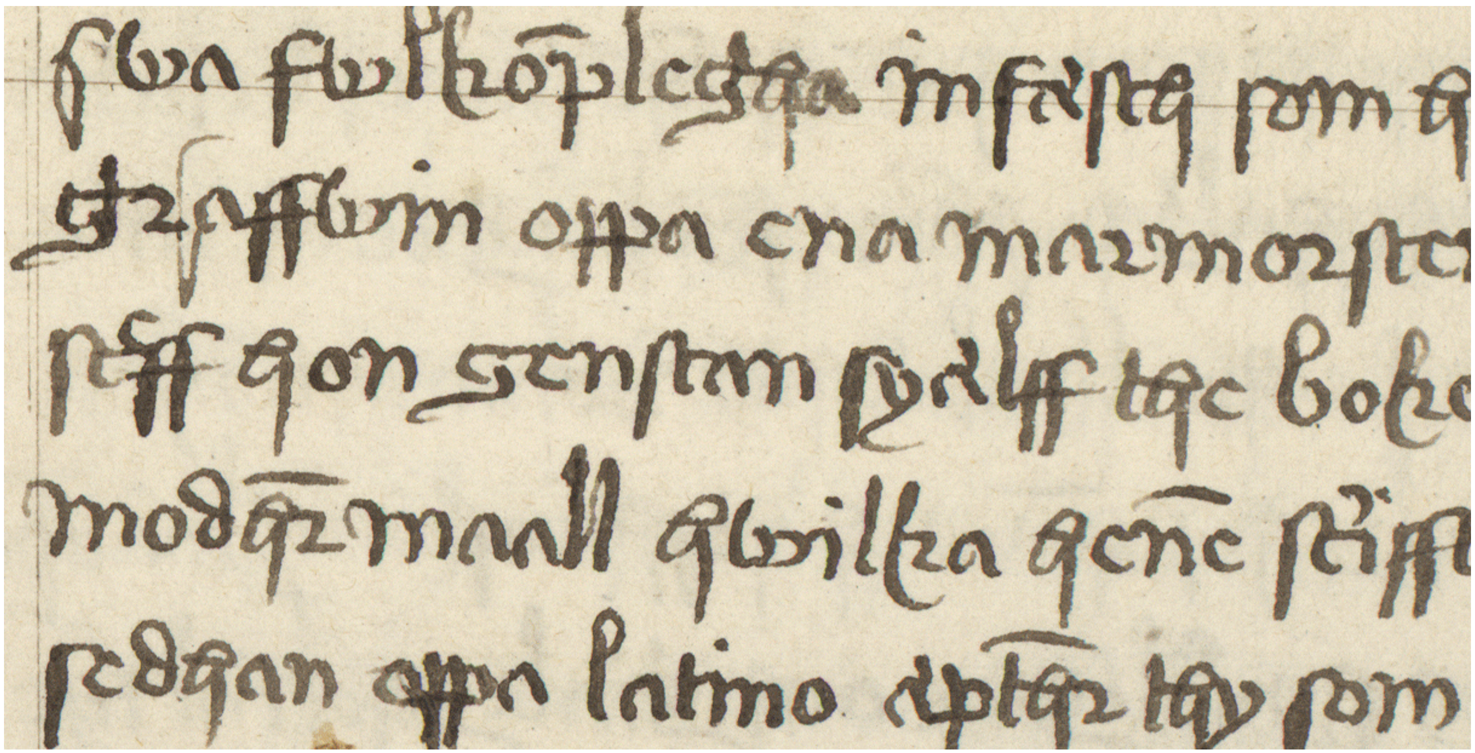
binarized

(ink, background)

Segmentation, five parameters

- width $\in [w_{\mathrm{mn}}, w_{\mathrm{mx}}]$
- t_{i} is the thickest amount of ink that allows a cut to be made.
- height $\in [h_{\mathrm{mn}}, h_{\mathrm{mx}}]$
- In our experiments, $(t_{\mathrm{i}}, w_{\mathrm{mn}}, w_{\mathrm{mx}}, h_{\mathrm{mn}}, h_{\mathrm{mx}}) = (1.0w_{\mathrm{s}}, 3.0w_{\mathrm{s}}, 8.0w_{\mathrm{s}}, 3.0w_{\mathrm{s}}, 15.0w_{\mathrm{s}})$.

An example. C 61 (b)



An example. C 61 (b)

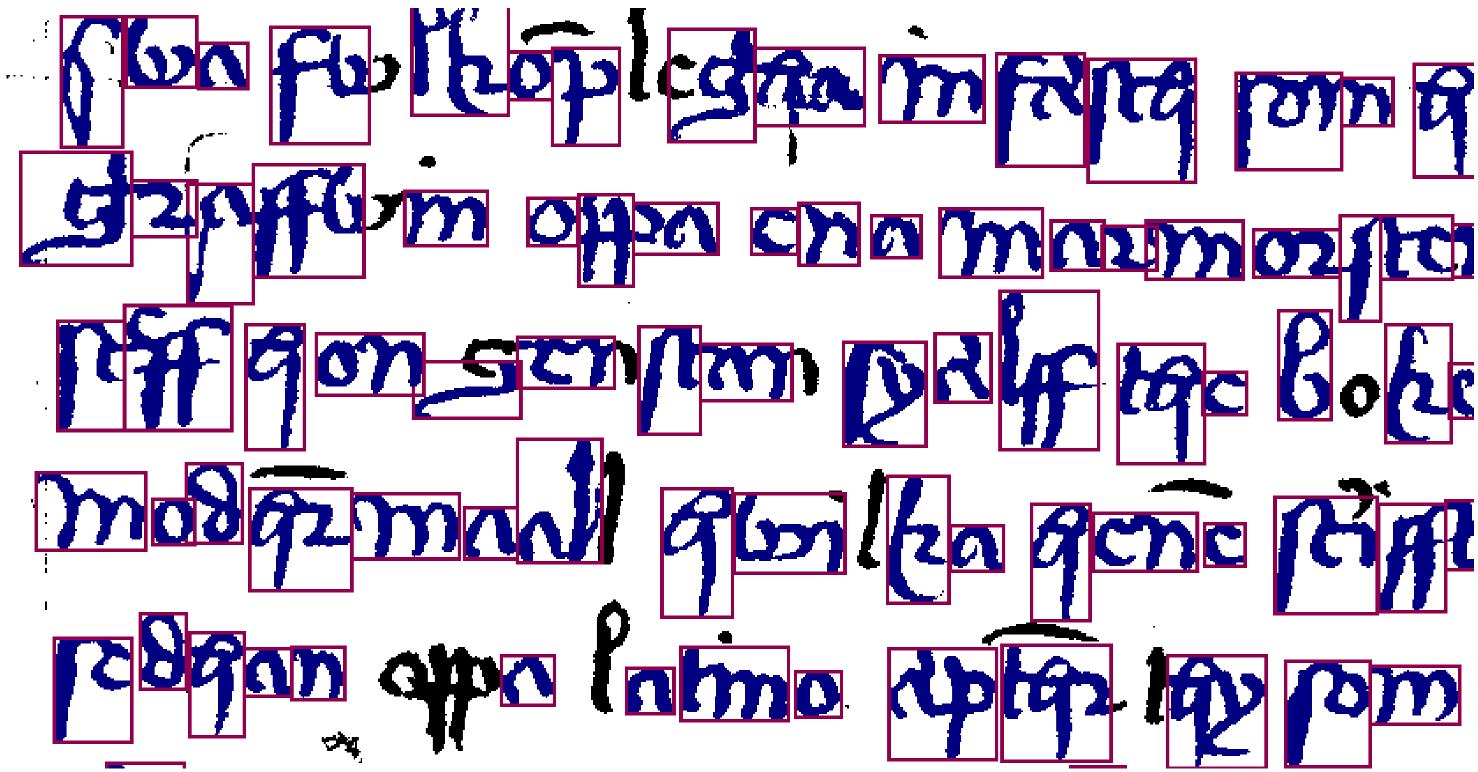
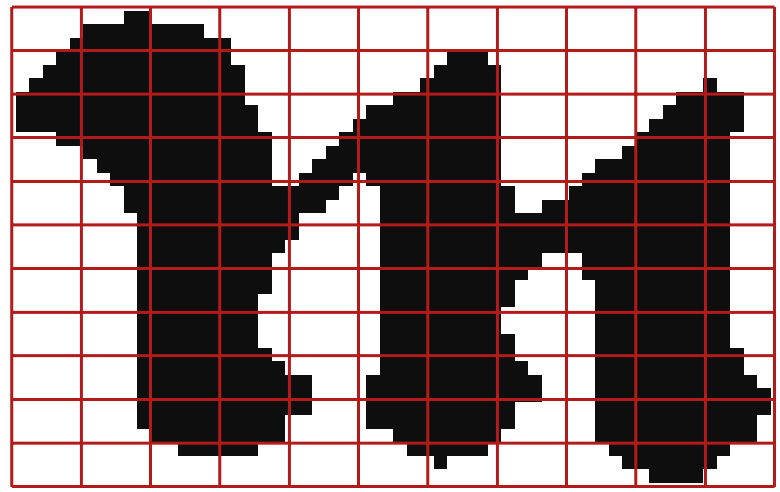


Image features



Distribution of ink as captured by a grid of 11×11 equal subrectangles over the bounding box.

Each value is the ratio of the number of ink pixels to the size of the subrectangle region.

Similarity of components

- So, we associate each component with a 121-dimensional (11×11) vector.
- We apply Euclidean distance on these:

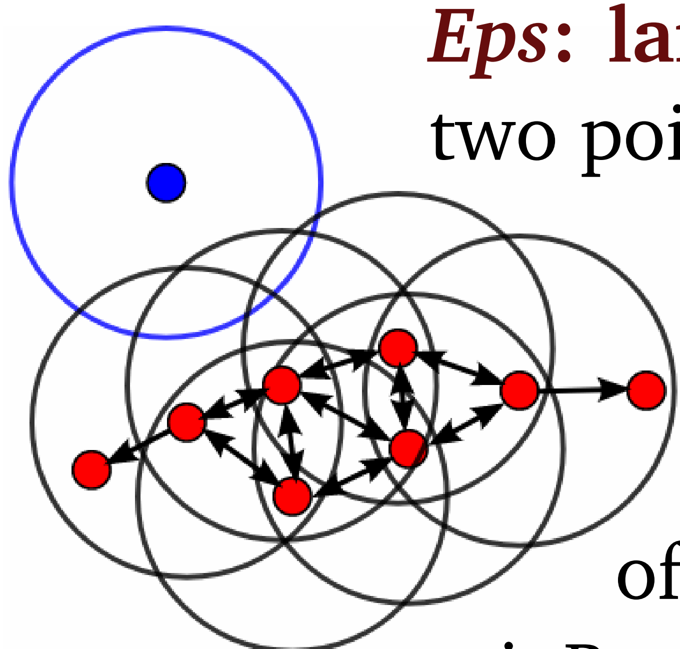
$$distance(I,J)=\sqrt{\sum_{i=1}^{n}(I_i-J_i)^2}$$

- **The feature model and distance metric define a distance** (dissimilarity) between any two image components.

Clustering

- **“Core” clustering** by density-based algorithm (DBSCAN).
- **Removal** of small clusters (size < 40 here).
- **Classification** assigns some not clustered components to the remaining core clusters (“nearest neighbour” to centroids).

DBSCAN guided by two parameters



Eps: **largest distance** between two points counted as neighbours.

minPts: **minimal number** of neighbouring points required for the formation of a same-cluster dense region.

$minPts = 11$ in our experiments.

(Illustration: $minPts = 4$.)

Eps estimation

- Image distance is difficult to “use” in an intuitive way.
- *Eps* estimated from the probability (p_{Eps}) that two randomly selected components are at least that close to each other.
- Baseline setting: $p_{Eps} = 0.0007$.
- This makes *Eps* sensitive to the data set.

Extension of clusters by classification

- DBSCAN typically leaves some data unclustered.

  Small clusters are removed.
- In the last step some of the unclustered components are assigned to existing clusters. This is based on a “nearest neighbour” (to cluster centroids) procedure.

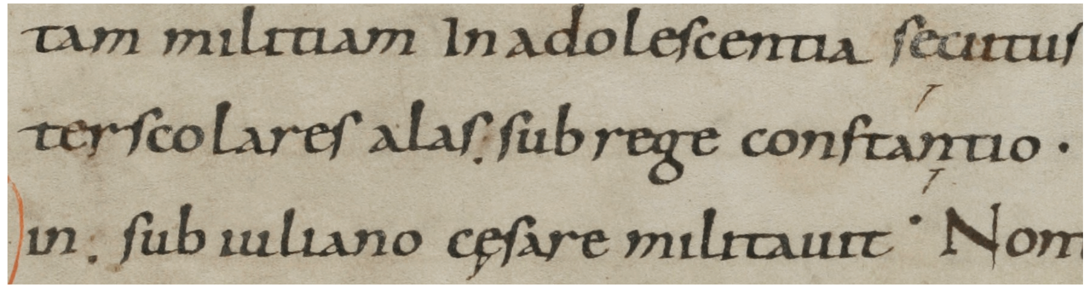
Data in experiments

- **Irish** (7th/8th C.): Gen. 1. and CS 60.
- **Carolingian minuscule** (St. Gallen) CS 557 (9th C.) and CS 564 (12th C.).
- **Textualis** (14th C., Swedish): B 59 and B 10.
- **Cursiva recentior** (late 15th C., Swedish): C 61(a) and C 61(b).

High-resolution images (JPEG or TIFF), published open access.

Two settings

CS 557 (44 p.).



Settings	**letters**	**ligatures, bigrams**	**mixtures, etc.**
Baseline $p_{Eps}=0.0007$ 17 clusters 5.0k elements	**b**:241+93, $\mathbf{d}_1$:273, **g**:64, **h**:84, **l**:85, **m**:585, **n**:144, **o**:676, **p**:597, **q**:195, **r**:409+86, **s**:1141, **v**:191	**is**:104, **ss**:56	–
More generous $p_{Eps}=0.0014$ 27 clusters 9.5k elements	**a**:594, **b**:404, **c**:111, $\mathbf{d}_1$:309, $\mathbf{d}_2$:167, **e**:423, **E**:54, **g**:112, **h**:108, **l**:109*, **m**:960, **n**:685, **o**:841, **p**:628, **q**:245, **r**:609+138†, **s**:1149, **t**:64, **u**:969, **v**:229	**co**:104, **er**:184, **&**:104, **is**:113, **ri**:64, **ss**:66	–

Precision levels: default $>$ 99.5%. $\star$: $>$ 98%. $\dagger$: $>$ 80%. ($minPts = 11$.)

Recall and precision (%) – a few cases

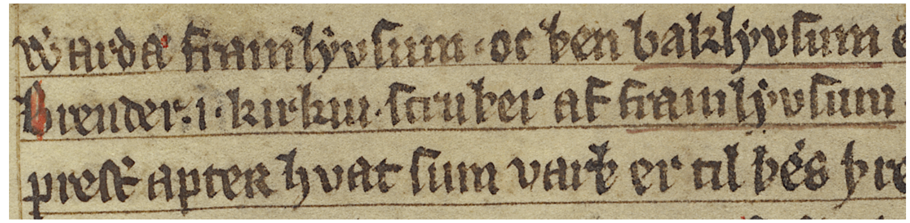
	"Core" $p_{Eps}=0.0007$			Baseline w. classif.			More generous $p_{Eps}=0.0014$		
Manus.	**e**	**m**	**o**	**e**	**m**	**o**	**e**	**m**	**o**
Gen. 1 *recall*	10	46	54	14	54	58	46	68	83
prec.	100	100	100	100	100	100	100	100	100
CS 557 *recall*	0	13	46	0	44	61	8	71	76
prec.	–	100	100	–	100	100	100	100	100
CS 564 *recall*	0	46	1	4	58	17	7	61	36
prec.	–	97	100	100	65	100	100	57	100
C 61(b) *recall*	28	58	25	28	63	28	31	65	36
prec.	100	100	100	100	97	100	100	96	29

Overclustering (B 59, largest cluster)

Baseline settings. Size: 3846.

Three settings

B 59 (44 p.)

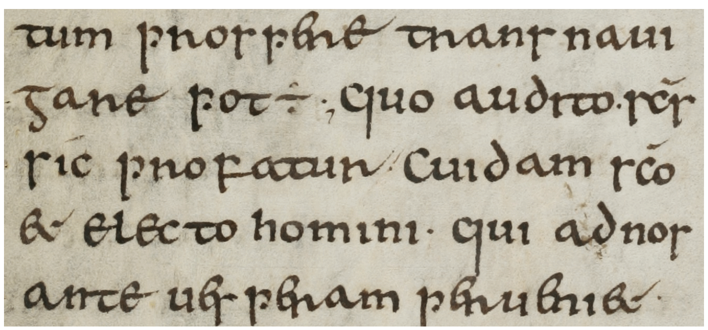


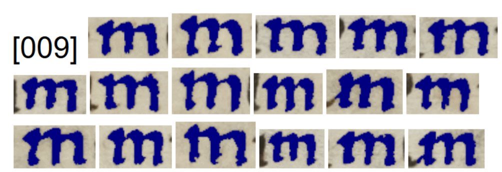
Manuscript	letters	ligatures, bigrams	mixtures, etc.
Baseline $p_{Eps}=0.0007$ 23 clusters 14.8k elements	**a**:2938†, **m**:361†, **n**:1562‡, **o**:488, **s**:1220‡	**al**:57, **bo**:47, **fa**:105⋆, **fi**:157, **gh**:63, **gi**:94†, **ll**:71†, **sk**:256†, **sti**:53†	**a**⋯:367, ⋯**a**:290†, **e**⋯:110, **sk**∪**st**:52†. *Useless:* 3846+2172+223+190+109
More reluctant $p_{Eps}=0.00035$ 18 clusters 8.9k elements	**a**:2634⋆, **d**:200, **h**:71, **k**:456⋆, **m**:289†, **n**:1348‡, **o**:359, **s**:1144‡, **þ**:355†	**ar**:122⋆, **sk**:236†, **ta**:98†	**v**∪**b**:693‡. *Useless:* 507+157+75+65+64
More reluctant $minPts=22$ 16 clusters 6.6k elements	**a**:2852⋆, **h**:93, **k**:477⋆, **m**:335†, **n**:1433‡, **o**:455, **s**:1170‡, **þ**:386†,	**ar**:229‡, **fi**:63, **sk**:223†, **ta**:143†,	**v**∪**b**:943‡. *Useless:* 599+155+66

Precision levels: default > 99.5%. ⋆: > 98%. †: > 80%. ‡: > 60%.

Methodological overview

- “Unsupervised” learning.
- Theoretical and experimental tuning of parameters.
- Human interpretation of output.
- Useful precision, but possibly low recall.





Conclusions

- Clustering allows us to find classes of letters and other writing elements in handwriting.
- Almost all components in the pipeline can be modified (binarization, segmentation, feature model, similarity metric, clustering algorithm and settings).

Thank you!

Work funded by the Swedish Research Council (Vetenskapsrådet, Dnr 2012-5743) and Riksbankens Jubileumsfond (Dnr NHS14-2068:1).