Importance of Textlines in Historical Document Classification

Martin Kišš, Jan Kohút, Karel Beneš, Michal Hradiš

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Page-level label(s) without any further details

Font & script classification

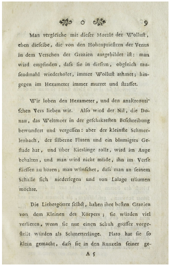
Up to *N* possible labels per page

Localization

One label per page

Dating

(not-before; not-after) intervals



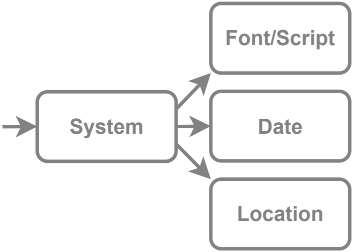
o 9

Man vergleiche mit dieſer Morale der Wolluſt, eben dieſelbe, die von den Hohenprieſtern der Venus in dem Verschen der Grazien ausgebildet iſt: man wird empfinden, daſs ſie in dieſem, obgleich tauſendmahl wiederholet, immer Wolluſt athmet; hingegen im Hexameter immer murret und ſtraffet.

Wir loben den Hexameter, und den anakreontiſchen Vers lieben wir. Alſo wird der Nil, die Donau, das Weltmeer in der geſchikteſten Beſchreibung bewundert und vergeſſen: aber der kleinſte Schmerlenbach, der ſilberne Fluten und ein blumigtes Geſtade hat, und über Kieslinge rollt, wird im Auge behalten, und man wird nicht müde, ihn im Verſe flieſſen zu hören; man wünſchet, daſs man an ſeinem Schalle ſich niederlegen und von Lalage träumen möchte.

Die Liebesgötter ſelbſt, haben ihre beſten Grazien von dem Kleinen des Körpers; ſie würden viel verlieren, wenn ſie nur einen Schuh gröſſer vorgeſtellt würden als Schmetterlinge. Plato hat ſie ſo klein gemacht, daſs ſie in den Runzeln ſeiner ge-

A 5



System

Font/Script

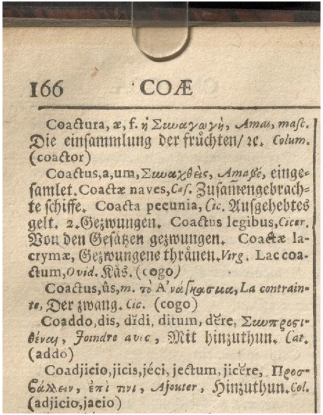
Date

Location

Data examples

Font/script

Fraktur, greek, italic, antiqua



166 COÆ

Coactura, æ, f. ἡ Συναγωγή, Amas, maſc. Die einſammlung der früchten/ ꝛc. Colum. (coactor)

Coactus,a,um, Συναχθεὶς, Amaſſé, eingeſamlet. Coactæ naves, Caſ. Zuſam̃engebrachte ſchiffe. Coacta pecunia, Cic. Auſgehebtes gelt. 2. Gezwungen. Coactus legibus, Cicer. Von den Geſätzen gezwungen. Coactæ lacrymæ, Gezwungene thränen. Virg. Lac coactum, Ovid. Käß. (cogo)

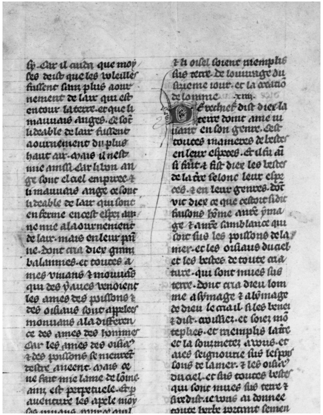
Coactus, ûs, m. τὸ Ἀνάγκασμα, La contrainte, Der zwang. Cic. (cogo)

Coaddo, dis, dĭdi, ditum, dĕre, Συμπροσθεῖναι, Joindre avec, Mit hinzuthun. Cat. (addo)

Coadjicio, jicis, jéci, jectum, jicĕre, Προσβάλλειν, ἐπί τινι, Ajouter, Hinzuthun. Col. (adjicio, jacio)

Localization

Paris

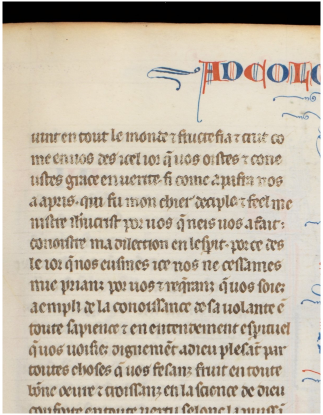


ſꝑ. Car il cuida que moy
ſes deiſt que les voleilles
fuſſent ſanz plus aour
nement de lair qui eſt
entour la terre. et que li
mauvais anges. ce ſōt
li deable de lair fuſſent
aournement du plus
haut air. Mais il neſt
mie ainſi. Car li bon an
ge ſont el ciel empiree &
li mauvais ange ce ſont
li deable de lair qui ſont
enferme en ceſt eſpez air
ne mie a laournement
de lair. mais en leur pu
ne. Dont cria diex grans
balaines. et toutes a
mes vivans & mouvās
qui des yaues venoient
les ames des poiſſons &
des oiſiaus ſont apelees
mouvans a la differen
ce des ames des hommes
Car les ames des oiſia⁹
& des poiſſons ſe muerēt
de leſtre avecent. Mais ce
ne fait mie lame de lōme.
ainz eſt perpetuele. Et p
aventure les apele moy

& li oiſel ſoient remplis
ſus terre. De louange du
ſizieme iour. et la creatiō
de lomme. xiiij.
Errechief diſt diex la
terre donne ame vi
vant en ſon genre. Ceſt
toutes manieres de beſtes
en leur eſpeces. Et il fu ai
ſi fait & fiſt diex les beſtes
de la tre ſelonc leur eſpe
ces. & en leur genres. Dōt
vit diex ce que ceſtoit ſi dit
faiſons hōme a nře yma
ge & a nře ſamblance qui
ſoit ſus les poiſſons de la
mer. et les oiſiaus du ciel
et les beſtes de toute crea
ture. qui ſont meues ſus
terre. Dont cria dieu lom
me a ſymage & a lymage
de dieu le cria il. ſi les benei
& diſt. croiſſiez. et ſoiez mō
teplies. et raemplis la tre
et la ſoumetez a vous. et
aiez ſeignourie ſus les poiſ
ſons de la mer. & les oiſia⁹
du ciel. et ſus toutes beſtes
qui ſont meues ſus terre &
ſi redist. Je vous ai donnee
toute herbe portant ſemen

Dating

1260-1269



ADCOL

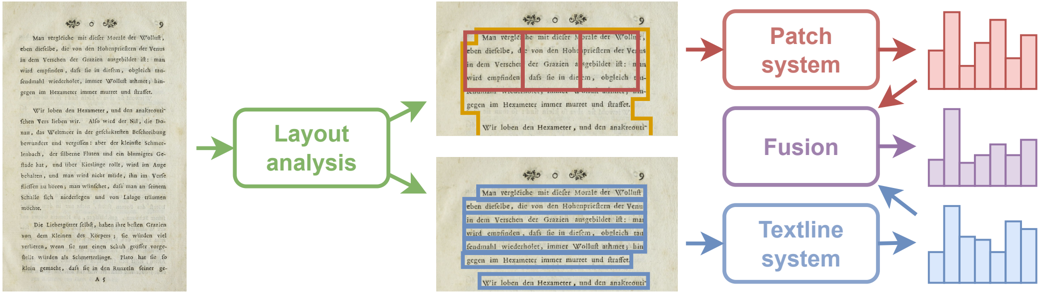
viue en tout le monde & fructefia & croiſt co
me en vos des icel ior q̄ vos oiſtes & conne
uſtes grace en verite. ſi come epafra nos
a apris. qui fu mon chier deciple & feel me
niſtre iheſucriſt por nos q̄ neis nos a fait
conoiſtre ma dilection en leſpit. por ce des
le ior q̄ nos euſmes ice nos ne ceſſames
mie prianz por vos & requeranz q̄ vos ſoiez
aempli de la conoiſſance de ſa volante ē
toute ſapience & en entendement eſpirituel
q̄ vos voiſiez dignemēt a dieu pleſāt par
toutes choſes q̄ vos feſanz fruit en toute
bōne oeuure & croiſſanz en la ſcience de dieu

Dataset

- custom training-validation dataset splits
- the aim to create class-balanced validation set
- publicly available at `pero.fit.vutbr.cz/hdc_dataset`

Task	Training	Validation	Testing
Font	35 382	239	5 506
Script	7 594	419	1 256
Location	5 397	65	325
Date	10 294	1 000	2 516

Our approach



9

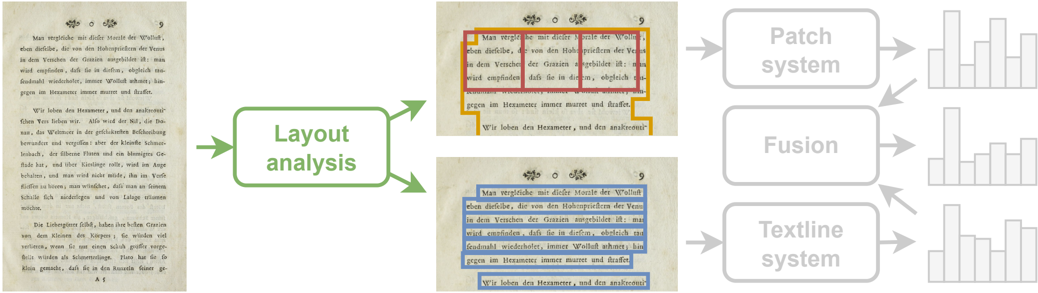
Man vergleiche mit dieſer Morale der Wolluſt, eben dieſelbe, die von den Hohenprieſtern der Venus in dem Verschen der Grazien ausgebildet iſt: man wird empfinden, daſs ſie in dieſem, obgleich tauſendmahl wiederholet, immer Wolluſt athmet; hingegen im Hexameter immer murret und ſtraffet.

Wir loben den Hexameter, und den anakreontiſchen Vers lieben wir. Alſo wird der Nil, die Donau, das Weltmeer in der geſchikteſten Beſchreibung bewundert und vergeſſen: aber der kleinſte Schmerlenbach, der ſilberne Fluten und ein blumigtes Geſtade hat, und über Kieslinge rollt, wird im Auge behalten, und man wird nicht müde, ihn im Verſe flieſſen zu hören; man wünſchet, daſs man an ſeinem Schalle ſich niederlegen und von Lalage träumen möchte.

Die Liebesgötter ſelbſt, haben ihre beſten Grazien von dem Kleinen des Körpers; ſie würden viel verlieren, wenn ſie nur einen Schuh gröſſer vorgeſtellt würden als Schmetterlinge. Plato hat ſie ſo klein gemacht, daſs ſie in den Runzeln ſeiner ge-

A 5

Our approach - Layout analysis



9

Man vergleiche mit dieſer Morale der Wolluſt, eben dieſelbe, die von den Hohenprieſtern der Venus in dem Verschen der Grazien ausgebildet iſt: man wird empfinden, daſs ſie in dieſem, obgleich tauſendmahl wiederholet, immer Wolluſt athmet; hingegen im Hexameter immer murret und ſtraffet.

Wir loben den Hexameter, und den anakreontiſchen Vers lieben wir. Alſo wird der Nil, die Donau, das Weltmeer in der geſchikteſten Beſchreibung bewundert und vergeſſen: aber der kleinſte Schmerlenbach, der ſilberne Fluten und ein blumigtes Geſtade hat, und über Kieslinge rollt, wird im Auge behalten, und man wird nicht müde, ihn im Verſe flieſſen zu hören; man wünſchet, daſs man an ſeinem Schalle ſich niederlegen und von Lalage träumen möchte.

Die Liebesgötter ſelbſt, haben ihre beſten Grazien von dem Kleinen des Körpers; ſie würden viel verlieren, wenn ſie nur einen Schuh gröſſer vorgeſtellt würden als Schmetterlinge. Plato hat ſie ſo klein gemacht, daſs ſie in den Runzeln ſeiner ge-

A 5

Layout analysis

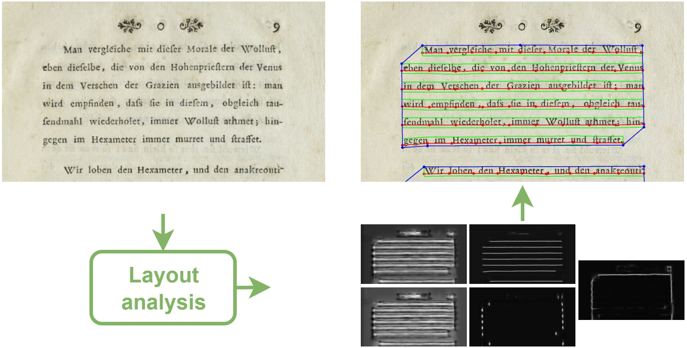
Patch system

Fusion

Textline system

Our approach - Layout analysis

- ParseNet (U-net-based) neural network[1]
- Text lines and regions detection
- Trained on PERO Layout dataset



9

Man vergleiche mit dieſer Morale der Wolluſt, eben dieſelbe, die von den Hohenprieſtern der Venus in dem Verschen der Grazien ausgebildet iſt: man wird empfinden, daſs ſie in dieſem, obgleich tauſendmahl wiederholet, immer Wolluſt athmet; hingegen im Hexameter immer murret und ſtraffet.

Wir loben den Hexameter, und den anakreonti-

Layout analysis

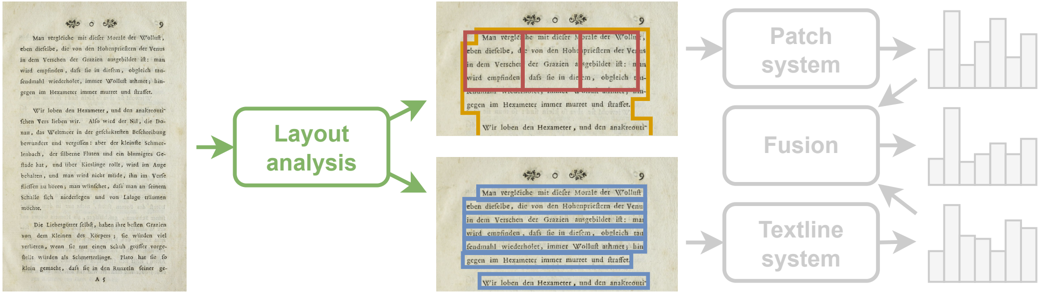
9

Man vergleiche mit dieſer Morale der Wolluſt, eben dieſelbe, die von den Hohenprieſtern der Venus in dem Verschen der Grazien ausgebildet iſt: man wird empfinden, daſs ſie in dieſem, obgleich tauſendmahl wiederholet, immer Wolluſt athmet; hingegen im Hexameter immer murret und ſtraffet.

Wir loben den Hexameter, und den anakreonti-

[1] Kodym, Oldřich, and Michal Hradiš. "Page layout analysis system for unconstrained historic documents."

Our approach - Layout analysis



9

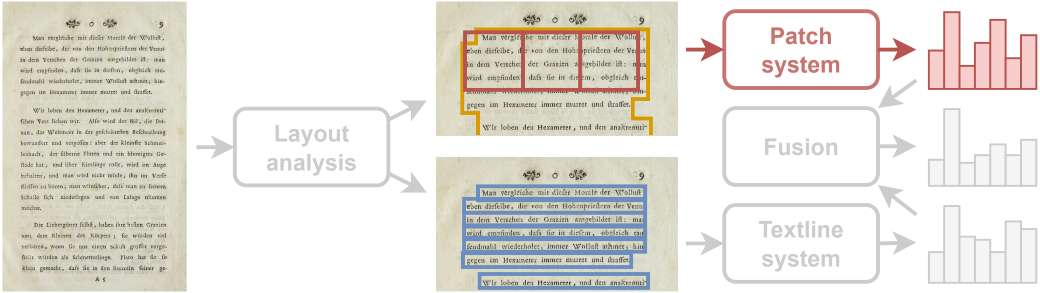
Man vergleiche mit dieſer Morale der Wolluſt, eben dieſelbe, die von den Hohenprieſtern der Venus in dem Verschen der Grazien ausgebildet iſt: man wird empfinden, daſs ſie in dieſem, obgleich tauſendmahl wiederholet, immer Wolluſt athmet; hingegen im Hexameter immer murret und ſtraffet.

Wir loben den Hexameter, und den anakreontiſchen Vers lieben wir. Alſo wird der Nil, die Donau, das Weltmeer in der geſchikteſten Beſchreibung bewundert und vergeſſen: aber der kleinſte Schmerlenbach, der ſilberne Fluten und ein blumigtes Geſtade hat, und über Kieslinge rollt, wird im Auge behalten, und man wird nicht müde, ihn im Verſe flieſſen zu hören; man wünſchet, daſs man an ſeinem Schalle ſich niederlegen und von Lalage träumen möchte.

Die Liebesgötter ſelbſt, haben ihre beſten Grazien von dem Kleinen des Körpers; ſie würden viel verlieren, wenn ſie nur einen Schuh gröſſer vorgeſtellt würden als Schmetterlinge. Plato hat ſie ſo klein gemacht, daſs ſie in den Runzeln ſeiner ge-

A 5

Our approach - Patch system



9

Man vergleiche mit dieſer Morale der Wolluſt, eben dieſelbe, die von den Hohenprieſtern der Venus in dem Verschen der Grazien ausgebildet iſt: man wird empfinden, daſs ſie in dieſem, obgleich tauſendmahl wiederholet, immer Wolluſt athmet; hingegen im Hexameter immer murret und ſtraffet.

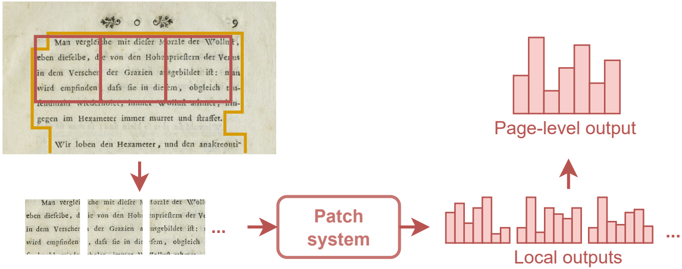
Wir loben den Hexameter, und den anakreontiſchen Vers lieben wir. Alſo wird der Nil, die Donau, das Weltmeer in der geſchikteſten Beſchreibung bewundert und vergeſſen: aber der kleinſte Schmerlenbach, der ſilberne Fluten und ein blumigtes Geſtade hat, und über Kieslinge rollt, wird im Auge behalten, und man wird nicht müde, ihn im Verſe flieſſen zu hören; man wünſchet, daſs man an ſeinem Schalle ſich niederlegen und von Lalage träumen möchte.

Die Liebesgötter ſelbſt, haben ihre beſten Grazien von dem Kleinen des Körpers; ſie würden viel verlieren, wenn ſie nur einen Schuh gröſſer vorgeſtellt würden als Schmetterlinge. Plato hat ſie ſo klein gemacht, daſs ſie in den Runzeln ſeiner ge-

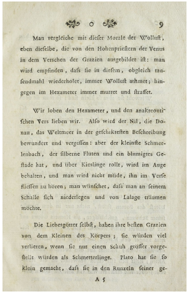
A 5

Our approach - Patch system

- ResNeXt-50-based neural network
- Training on four scales of non-overlapping square patches with three training strategies
- The page-level output is calculated as the mean of the local outputs



Our approach - Patch system



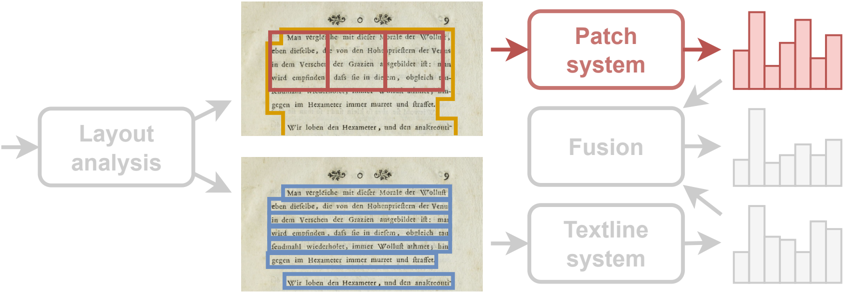
9

Man vergleiche mit dieſer Morale der Wolluſt, eben dieſelbe, die von den Hohenprieſtern der Venus in dem Verschen der Grazien ausgebildet iſt: man wird empfinden, daſs ſie in dieſem, obgleich tauſendmahl wiederholet, immer Wolluſt athmet; hingegen im Hexameter immer murret und ſtraffet.

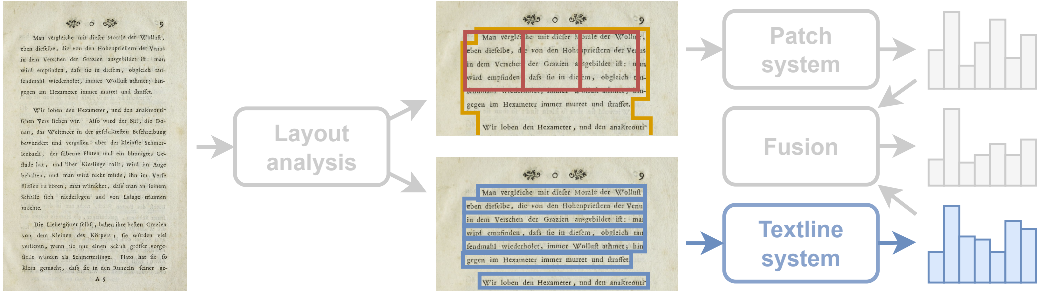
Wir loben den Hexameter, und den anakreontiſchen Vers lieben wir. Alſo wird der Nil, die Donau, das Weltmeer in der geſchikteſten Beſchreibung bewundert und vergeſſen: aber der kleinſte Schmerlenbach, der ſilberne Fluten und ein blumigtes Geſtade hat, und über Kieslinge rollt, wird im Auge behalten, und man wird nicht müde, ihn im Verſe flieſſen zu hören; man wünſchet, daſs man an ſeinem Schalle ſich niederlegen und von Lalage träumen möchte.

Die Liebesgötter ſelbſt, haben ihre beſten Grazien von dem Kleinen des Körpers; ſie würden viel verlieren, wenn ſie nur einen Schuh gröſſer vorgeſtellt würden als Schmetterlinge. Plato hat ſie ſo Klein gemacht, daſs ſie in den Runzeln ſeiner ge-

A 5



Our approach - Textline system



9

Man vergleiche mit dieſer Morale der Wolluſt, eben dieſelbe, die von den Hohenprieſtern der Venus in dem Verschen der Grazien ausgebildet iſt: man wird empfinden, daſs ſie in dieſem, obgleich tauſendmahl wiederholet, immer Wolluſt athmet; hingegen im Hexameter immer murret und ſtraffet.

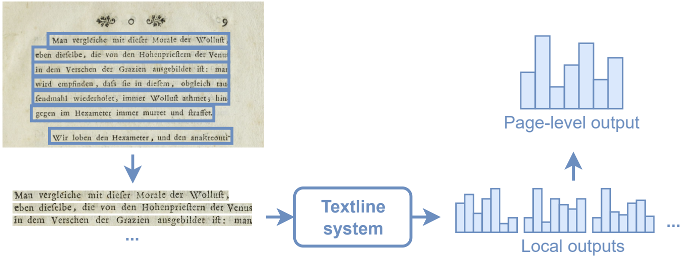
Wir loben den Hexameter, und den anakreontiſchen Vers lieben wir. Alſo wird der Nil, die Donau, das Weltmeer in der geſchikteſten Beſchreibung bewundert und vergeſſen: aber der kleinſte Schmerlenbach, der ſilberne Fluten und ein blumigtes Geſtade hat, und über Kieslinge rollt, wird im Auge behalten, und man wird nicht müde, ihn im Verſe flieſſen zu hören; man wünſchet, daſs man an ſeinem Schalle ſich niederlegen und von Lalage träumen möchte.

Die Liebesgötter ſelbſt, haben ihre beſten Grazien von dem Kleinen des Körpers; ſie würden viel verlieren, wenn ſie nur einen Schuh gröſſer vorgeſtellt würden als Schmetterlinge. Plato hat ſie ſo klein gemacht, daſs ſie in den Runzeln ſeiner ge-

A 5

Our approach - Textline system

- VGG-based neural network + global average pooling
- Arbitrarily long height-normalized textline image as an input
- The page-level output is calculated as the mean (classification tasks) or median (dating task) of the local outputs

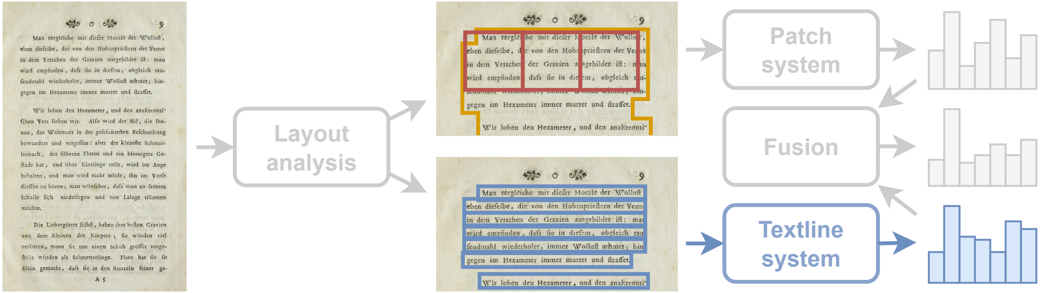


9

Man vergleiche mit dieſer Morale der Wolluſt,
eben dieſelbe, die von den Hohenprieſtern der Venus
in dem Verschen der Grazien ausgebildet iſt: man
wird empfinden, daſs ſie in dieſem, obgleich tau-
ſendmahl wiederholet, immer Wolluſt athmet; hin-
gegen im Hexameter immer murret und ſtraffet.
Wir loben den Hexameter, und den anakreonti-

Man vergleiche mit dieſer Morale der Wolluſt,
eben dieſelbe, die von den Hohenprieſtern der Venus
in dem Verschen der Grazien ausgebildet iſt: man
...

Our approach - Textline system



9

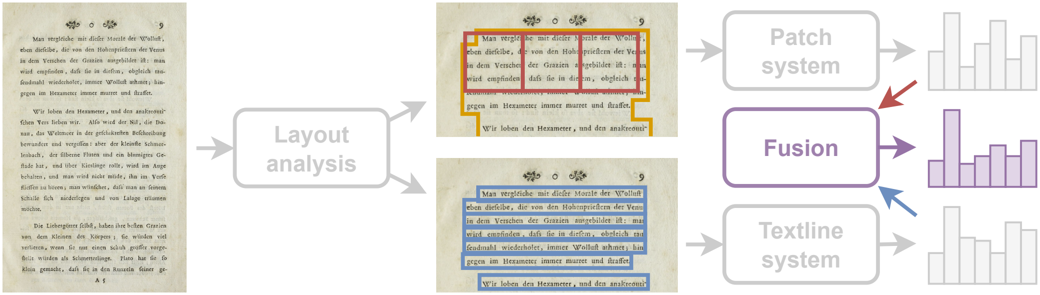
Man vergleiche mit dieſer Morale der Wolluſt, eben dieſelbe, die von den Hohenprieſtern der Venus in dem Verschen der Grazien ausgebildet iſt: man wird empfinden, daſs ſie in dieſem, obgleich tauſendmahl wiederholet, immer Wolluſt athmet; hingegen im Hexameter immer murret und ſtraffet.

Wir loben den Hexameter, und den anakreontiſchen Vers lieben wir. Alſo wird der Nil, die Donau, das Weltmeer in der geſchikteſten Beſchreibung bewundert und vergeſſen: aber der kleinſte Schmerlenbach, der ſilberne Fluten und ein blumigtes Geſtade hat, und über Kieslinge rollt, wird im Auge behalten, und man wird nicht müde, ihn im Verſe flieſſen zu hören; man wünſchet, daſs man an ſeinem Schalle ſich niederlegen und von Lalage träumen möchte.

Die Liebesgötter ſelbſt, haben ihre beſten Grazien von dem Kleinen des Körpers; ſie würden viel verlieren, wenn ſie nur einen Schuh gröſſer vorgeſtellt würden als Schmetterlinge. Plato hat ſie ſo klein gemacht, daſs ſie in den Runzeln ſeiner ge-

A 5

Our approach - Fusion



9

Man vergleiche mit dieſer Morale der Wolluſt, eben dieſelbe, die von den Hohenprieſtern der Venus in dem Verschen der Grazien ausgebildet iſt: man wird empfinden, daſs ſie in dieſem, obgleich tauſendmahl wiederholet, immer Wolluſt athmet; hingegen im Hexameter immer murret und ſtraffet.

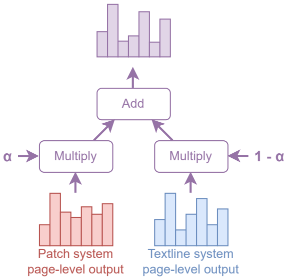
Wir loben den Hexameter, und den anakreontiſchen Vers lieben wir. Alſo wird der Nil, die Donau, das Weltmeer in der geſchikteſten Beſchreibung bewundert und vergeſſen: aber der kleinſte Schmerlenbach, der ſilberne Fluten und ein blumigtes Geſtade hat, und über Kieslinge rollt, wird im Auge behalten, und man wird nicht müde, ihn im Verſe flieſſen zu hören; man wünſchet, daſs man an ſeinem Schalle ſich niederlegen und von Lalage träumen möchte.

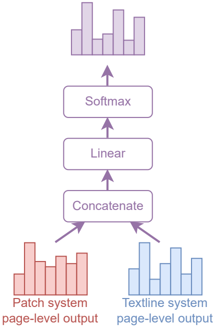
Die Liebesgötter ſelbſt, haben ihre beſten Grazien von dem Kleinen des Körpers; ſie würden viel verlieren, wenn ſie nur einen Schuh gröſſer vorgeſtellt würden als Schmetterlinge. Plato hat ſie ſo klein gemacht, daſs ſie in den Runzeln ſeiner ge-

A 5

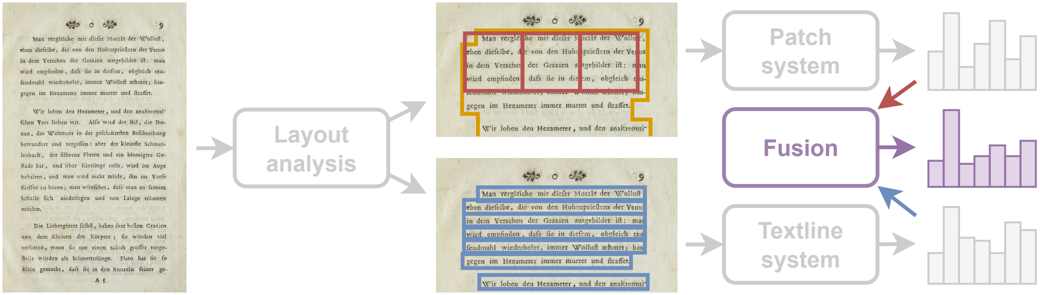
Our approach - Fusion

- Linear (left image) and log-linear (right image) fusions
- Page-level outputs from the patch and the textline systems as the input
- Optimized on the validation sets using 10-fold cross validation





Our approach - Fusion



9

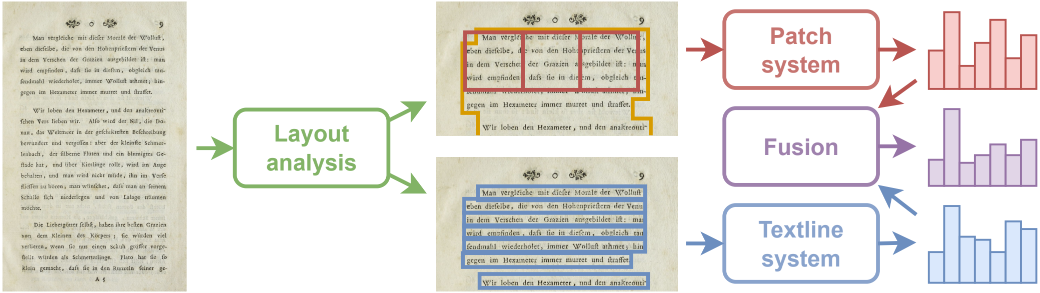
Man vergleiche mit dieſer Morale der Wolluſt, eben dieſelbe, die von den Hohenprieſtern der Venus in dem Verschen der Grazien ausgebildet iſt: man wird empfinden, daſs ſie in dieſem, obgleich tauſendmahl wiederholet, immer Wolluſt athmet; hingegen im Hexameter immer murret und ſtraffet.

Wir loben den Hexameter, und den anakreontiſchen Vers lieben wir. Alſo wird der Nil, die Donau, das Weltmeer in der geſchikteſten Beſchreibung bewundert und vergeſſen: aber der kleinſte Schmerlenbach, der ſilberne Fluten und ein blumigtes Geſtade hat, und über Kieslinge rollt, wird im Auge behalten, und man wird nicht müde, ihn im Verſe flieſſen zu hören; man wünſchet, daſs man an ſeinem Schalle ſich niederlegen und von Lalage träumen möchte.

Die Liebesgötter ſelbſt, haben ihre beſten Grazien von dem Kleinen des Körpers; ſie würden viel verlieren, wenn ſie nur einen Schuh gröſſer vorgeſtellt würden als Schmetterlinge. Plato hat ſie ſo klein gemacht, daſs ſie in den Runzeln ſeiner ge-

A 5

Our approach



9

Man vergleiche mit dieſer Morale der Wolluſt, eben dieſelbe, die von den Hohenprieſtern der Venus in dem Verschen der Grazien ausgebildet iſt: man wird empfinden, daſs ſie in dieſem, obgleich tauſendmahl wiederholet, immer Wolluſt athmet; hingegen im Hexameter immer murret und ſtraffet.

Wir loben den Hexameter, und den anakreontiſchen Vers lieben wir. Alſo wird der Nil, die Donau, das Weltmeer in der geſchikteſten Beſchreibung bewundert und vergeſſen: aber der kleinſte Schmerlenbach, der ſilberne Fluten und ein blumigtes Geſtade hat, und über Kieslinge rollt, wird im Auge behalten, und man wird nicht müde, ihn im Verſe flieſſen zu hören; man wünſchet, daſs man an ſeinem Schalle ſich niederlegen und von Lalage träumen möchte.

Die Liebesgötter ſelbſt, haben ihre beſten Grazien von dem Kleinen des Körpers; ſie würden viel verlieren, wenn ſie nur einen Schuh gröſſer vorgeſtellt würden als Schmetterlinge. Plato hat ſie ſo klein gemacht, daſs ſie in den Runzeln ſeiner ge-

A 5

Loss functions - Font & script classification

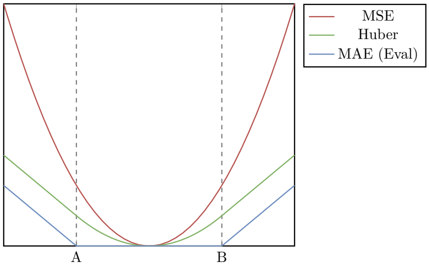
- Up to N possible fonts/scripts per page
    - e.g. $\mathcal{T} \in \{fraktur, greek, italic, antiqua\}$
- Page-level annotations without location information
- L_{hard} selects the most probable page-level label based on network output
- L_{soft} considers all page-level labels weighted by the network output probability

$$L_{hard} = \min_{i \in \mathcal{T}} [-\log(f(x)_i)] \quad (1)$$

$$L_{soft} = \sum_{i \in \mathcal{T}} -\log(f(x)_i) \cdot f(x)_i. \quad (2)$$

Loss functions - Dating

- *(not-before; not-after)* interval instead of single point annotations
  - *e.g. (1260; 1269)*
- Modified Huber loss for intervals
- Pulls the output of the network more towards the middle of the interval



Results

	Font Acc. ↑	Script Acc. ↑	Location Acc. ↑	Date MAE ↓
Textline system	98.42 %	88.54 %	69.85 %	**21.91** years
Patch system	95.68 %	80.26 %	75.08 %	32.45 years
Linear fusion	98.27 %	**88.84** %	70.77 %	21.99 years
Log-linear fusion	**98.48** %	88.60 %	**79.69** %	—
Baseline	—	55.22 %	62.46 %	—
The North LTU	82.80 %	74.12 %	43.69 %	79.43 years
CLUZH	95.66 %	35.25 %	—	—
NAVER Papago	97.17 %	—	—	—