

OCR4all and LAREX

Insights and Prospects

Christian Reul

Centre for Philology and Digitality „Kallimachos“ (ZPD)
University of Würzburg



16.12.2021



OCR4all – Motivation and Basics

- Open-source OCR tools powerful but can be overwhelming for non-technical users
 - Complicated setup (missing dependencies, ...)
 - No comfortable GUI but unfamiliar command line usage
 - ...
- Idea behind OCR4all
 - Comprehensible for and applicable by any given user
 - Entire workflow encapsulated into a single Docker/VBox image
→ Platform independence, easy installation
 - Fully controllable via comfortable web GUI
 - Incorporates and combines several open-source tools
 - Fully [open-source](#)

Currently available Submodules

- OCRopus
 - <https://github.com/ocropus/ocropy>
 - Preprocessing (Binarization, Deskewing)
 - Automatic (Line)Segmentation
- LAREX
 - <https://github.com/OCR4all/LAREX>
 - Initially designed for high-quality interactive region segmentation including fine-grained semantic classification
 - Meanwhile comprehensive editing tool
 - Region- and line coordinates
 - Semantic typification and reading order
 - Text (ground truth production for training)
- Calamari
 - <https://github.com/Calamari-OCR>
 - Text recognition and model training
 - Evaluation
- Kraken
 - <https://github.com/mittagessen/kraken>
 - Automatic segmentation
 - Potential further option for text recognition and model training

EXECUTE > FINALIZE CURRENT PROCESS AND EXIT ✕ CANCEL ✕

☰ Process selection

Preprocessing	<input checked="" type="checkbox"/>
Noise Removal	<input type="checkbox"/>
Segmentation (Dummy)	<input checked="" type="checkbox"/>
Line Segmentation	<input checked="" type="checkbox"/>
Recognition	<input checked="" type="checkbox"/>

⚙ Settings

ℹ Status

EXECUTE > FINALIZE CURRENT PROCESS AND EXIT ✕ CANCEL ✕

👁 Pages

✓ Select all ▼

✓ Page 0001



✓ Page 0002



✓ Page 0003





10 et

7 **1** **2** **3** **4** **5** **6**

fy font desquelles ie me tais. **1** Pour des folz belistriés qui de robes ceulz qui n'ont pas du pain a manger & a l'adventure qu'ilz n'osent demander de honte. Les vieilles gens / pourres des vies / labres / auengles / helas pense p / car certes vous en tenez bres compte devant celluy qui nous crea

2 Des conditiōs courroux et **3** grandes mauuaities des fēmes

4 Plusieurs asnes cheuaucheroient **5** Quilz ont fait & grief infame **6** Au pouure asne & grans tormēs **7** Car luy ont toz tous ses bons mēs

8 Si colligo q vincta nemo magis gaudet q se ra

9 Ntendes folz e' stourdis & vous congnoistres les mauuaities des fēmes. Aussi fēmes approchez vo' & vo' ozzes bone matiere. Mes versetz ditez & escriptz Voult broiet des fēmes

10 Ntendes folz e' stourdis & vous congnoistres les mauuaities des fēmes. Aussi fēmes approchez vo' & vo' ozzes bone matiere. Mes versetz ditez & escriptz Voult broiet des fēmes

11 fut sage & p'fite tāt q sa grāt vint & sageffe avec ses bēnes mēres reddēt par tout les bēnes ie foveray mais ie blainceay les mauuaities ta

12 **13** **14**

Settings

Regions

Reading Order

- 0 r10-page_number
- 1 r5-paragraph
- 2 r13-heading
- 3 r2-header
- 4 r3-header
- 5 r14-drop_capital
- 6 r15-paragraph
- 7 r9-marginalia
- 8 r12-marginalia
- 9 r11-marginalia
- 10 r8-marginalia
- 11 r16-marginalia

Parameters

SEGMENT

SAVE RESULT

LOAD RESULT

Fueillet

fy font desquelles ie me tais. Pour des folz belistriés qui de robes ceulz qui n'ont pas du pain a manger & a l'adventure qu'ilz n'osent demander de honte. Les vieilles gens / pourres des vies / labres / auengles / helas pense p / car certes vous en tenez bres compte devant celluy qui nous crea

Des conditiōs courroux et grandes mauuaities des fēmes

Plusieurs asnes cheuaucheroient
Le nest que monte y eust femme
Au moyē de quoy ne voudroient
Monte dessus po euse diffame

Quilz ont fait & grief infame
Au pouure asne & grans tormēs
Car luy ont toz tous ses bons mēs
Car luy ont toz tou fcs bons mēs

Si colligo q vincta nemo magis gaudet q se ra



Ntendes folz e' stourdis & vous congnoistres les mauuaities des fēmes. Aussi fēmes approchez vo' & vo' ozzes bone matiere. Mes versetz ditez & escriptz Voult broiet des fēmes

puer. ecci. Eccl. xxv. puerbi. xxi.

Quilz ont fait & grief infame

Quilz ont fait 7 grief infame

Au pouure asne & grans tormēs

Au pouure asne 7 grans tormēs

Car luy ont toz tous ses bons mēs

Car luy ont toz tou fcs bons mēs

Ⓔ
Ⓒ

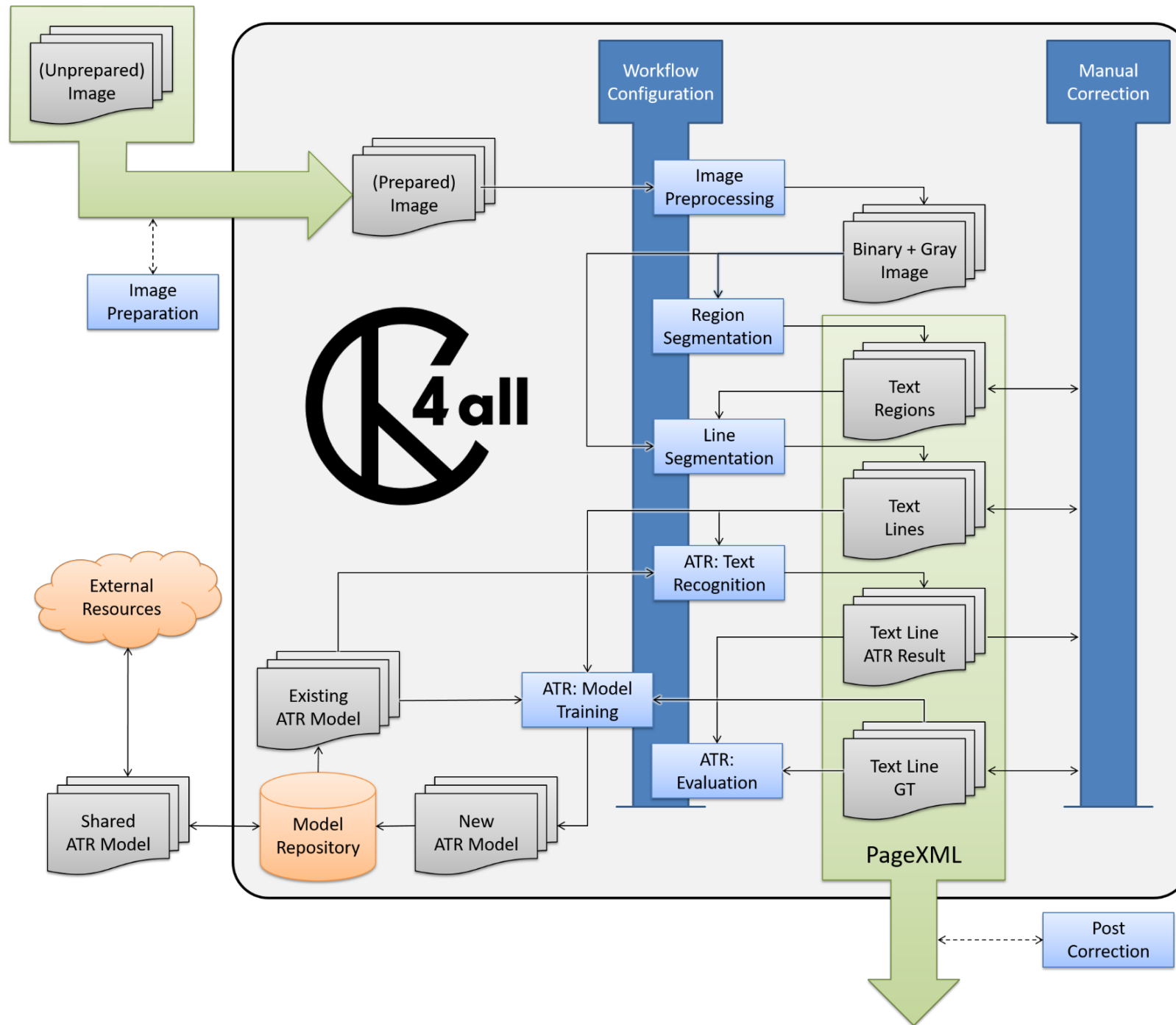
Ntendes folz e'

Ntendes folz e'

LOAD SAVE

ā	æ	þ	ç	ç	d	ð
ē	ē	ff	ff	ff	g	g
ī	ī	ll	m	n	ō	œ
þ	þ	p	p	p	p	p
q	q	q	q	q	q	q
q̄	q̄	q̄	q̄	q̄	q̄	q̄
r	r	z	s	f	h	ff
t	ç	ū	'	o	ð	
v	v	w	x	y	z	&
o	o	z	z	'	l	ç

+ [] []





www.ocr4all.de


Short Summary of Selected Evaluations

- Printings before 1600
 - Material: complex layout, highly variant typography
 - Approach: highly interactive segmentation, thorough book-specific GT production and training
 - Results:
 - No difference between un-/experienced user in terms of quality (average CER < 0.5%)
 - Experienced users more efficient (ca. factor 2 both for segmentation and GT production)
- 19th century Fraktur
 - Material: standard novels with moderate layout
 - Approach: fully automatic segmentation and application of highly performant mixed model
 - Results: average CER < 1%, mostly (considerably) < 0.5%
- For details please cf. the corresponding [paper](#)

Ongoing and Future Work

- DFG-funded project *OCR4all-libraries*
- Application scenarios and material
 - Collaborative work
 - Handwritten Text Recognition
- Individual tools and workflow steps
 - LAREX
 - Calamari
 - Fully automatic segmentation
 - (Interactive) Postcorrection
- Usability optimization

OCR4all-libraries – Full-Text Recognition of Historical Collections

- Applied for and approved within the third [OCR-D](#) phase
- Cooperation between the GEI Braunschweig and Uni Würzburg (HCI, ZPD)
- Two year funding since 07/2021
- Goals: OCR-D  OCR4all
 1. Comfortable application of OCR-D solutions via OCR4all
 - Full control without using the CLI
 - Maximizing usability and user experience
 - Applicable by non-technical users
 2. Support optimizing the OCR result within OCR4all
 - Comparison, optimization, and exchange of workflow configurations
 - Definition of sets, management of GT, as well as comfortable training
 - Also helpful for more technical users

Application Scenario: Collaborative Work

- OCR4all initially developed for local use by a single user
- ... but collaborative use of a centrally administered server instance
 - already possible
 - more and more in demand
 - meanwhile standard at the University of Würzburg
- Additional requirements compared to single usage
 - Project, user and task management
 - Resource management
 - Backup and versioning
- Viable interim solutions available (complete rewrite of the OCR4all backend)
- Fully comprehensive, flexible and future-proof all-round care package requires further external funding (DFG application imminent)

Scheduler service

Start: 2021-12-15 14:06:28
 State: **paused** (since 2021-12-15 14:08:53)

 Run


Scheduled

ID	Created	Target	Description	Action
5	2021-12-15 14:13:32	project_04	Folios import (v1.0)	<input type="button" value="Down"/> <input type="button" value="End"/> <input type="button" value="Cancel"/>
6	2021-12-15 14:14:11	project 02 (workflow 1/p2)	Workflow folios (v1.0)	<input type="button" value="Begin"/> <input type="button" value="Up"/> <input type="button" value="Down"/> <input type="button" value="End"/> <input type="button" value="Cancel"/>
4	2021-12-15 14:12:43	project_03 (workflow 2/p3)	Workflow folios (v1.0)	<input type="button" value="Begin"/> <input type="button" value="Up"/> <input type="button" value="Cancel"/>

Running

ID	Created	Start	Steps	Progress	Target	Description	Action
----	---------	-------	-------	----------	--------	-------------	--------

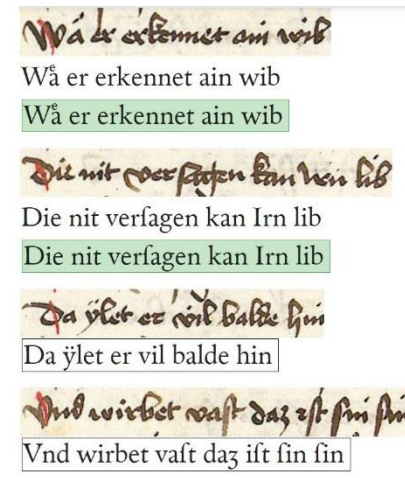
Done

 Expunge

ID	Created	Start	End	State	Steps	Progress	Target	Description
3	2021-12-15 14:09:27		2021-12-15 14:09:33	canceled	1	0%	Project 01 (ws_01)	Workflow folios (v1.0)
2	2021-12-15 14:07:11	2021-12-15 14:07:11	2021-12-15 14:07:11	interrupted	1	0%	project 02 (workflow 2/p2)	Workflow folios (v1.0)
1	2021-12-15 14:06:28	2021-12-15 14:06:28	2021-12-15 14:06:28	completed	1	100%	project 02	Folios import (v1.0)

Material: Handwritten Text Recognition

- Recognition of printed and handwritten documents conceptually very similar
- As of now, adapting OCR4all towards HTR rather straight-forward
- Apparently already widely used
- TODO: strong mixed models (work in progress, cf. talk in January)
- Several project applications in preparation
 - University of Würzburg, Berlin, Bremen, ...
 - Current focus: German manuscripts from the 14/15th and 19th century

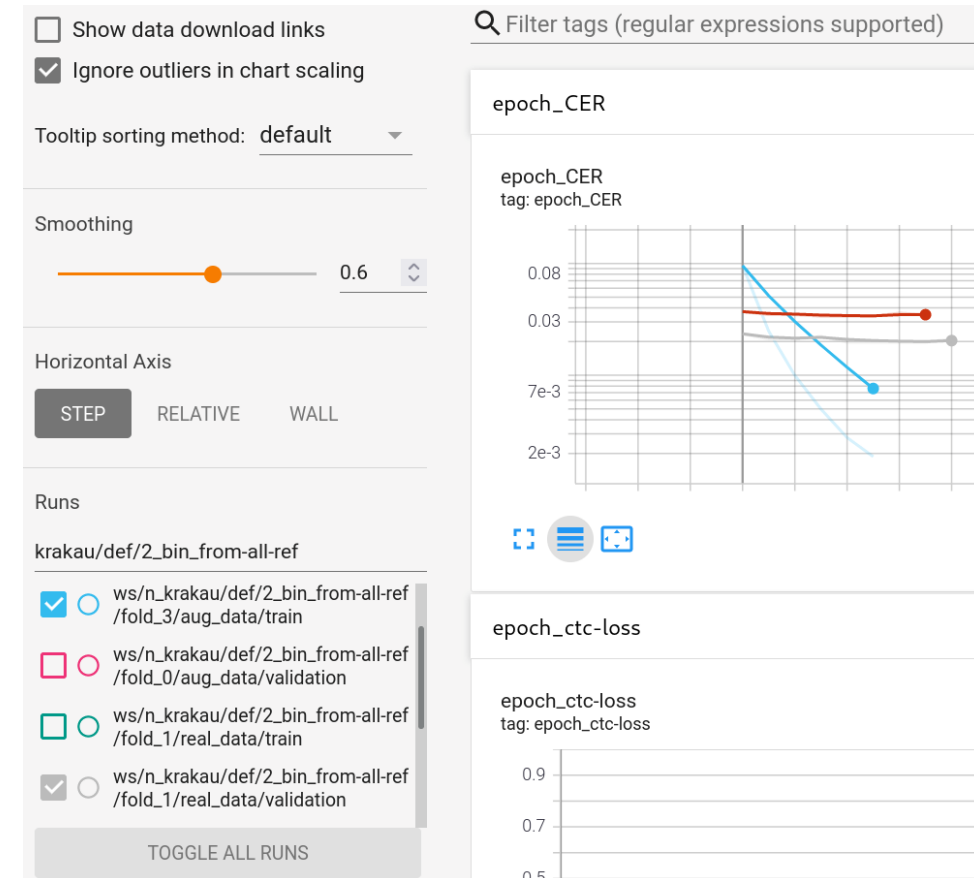


LAREX

- Outgrew itself (region segmentation only → universal correction tool)
- Many (even smaller) adaptations/extensions not as straight forward as they should be
- Complete rewrite in an component-based way (using Vue.js)
- Base application:
 - Basic (manual) correction functionality only
 - Easily deployable as Electron App
- Further functionality via extensions (also from external sources)
- Wishlist (tiny selection):
 - Calling OCR(-D) processors directly from LAREX
 - Page/region/line level
 - Generic solution for arbitrary processors
 - Soften single page view (e.g. show 25 most uncertain lines from all pages/works/...)
 - Enable editing of words and glyphs

Calamari

- Upgraded to Calamari 2
 - More accurate
 - Word/glyph + confidence output
 - Tensorboard support
 - ...
- Trained several highly performant mixed [models](#)
 - Printings: [paper](#) at HIP21
 - HTR: journal paper in preparation (cf. talk in January)
- Wishlist (for OCR4all):
 - Flexible model download within OCR4all
 - from various sources
 - (also for various OCR engines)
 - Full and seamless integration of Tensorboard



Automatic Segmentation

- Several promising, fully open-source, baseline-based, trainable (!) solutions:
 - [Kraken](#) ([Kiessling: A Modular Region and Text Line Layout Analysis System](#))
 - Developments at the Chair for Artificial Intelligence at Uni Würzburg (Fischer, Hartelt, Gehrke, Puppe)
 - ...
- Offer various solutions → let users choose the best one for their concrete use case and material
- Currently working on improving existing models for application on medieval manuscripts

Und neme da so bilde wol
Welchs wib and welch' man
An rechten dingten mit rechten sin
Der nimmet so uel und so gut
Köfz bilde wann er mit
Der ist zu dem besten ie bereit
Gentliche wib sint so gemait
Wann sie mögent hören nicht
Daz amc andin wib gesticht
Daz zuehet zu unechten dingten
Die sprechent ons mag unsechtigen
Krit in uns die hat es getan
Und ward ic halber syt ic man
Ein bideckes wib sol
Daz getar ich gezeiten wol
Doch das feiwen ob an wib
Hät nach rechte von lib
Wann ich sage ic es für war
Der und die reuget sic gar
Die mit amc andin constete
Wenet wachen ic missetete
Da so an bideckewid sol
Loug sin and tut mit wol
Ani ande wib der missetete
Böllet sin gut wib maechen stete
Die sol da so sin wal bewaet
Daz sie mit kome an ic wart
Die feiwe pflent neme sin
Affin der feiwen ungewin
Die da helena woz genant
Hü bewachen über alle lant
Was sie an gewaltig küniginne
Die het sol schone und listel sinne
Die schone bracht große conge sticht
Schone ist an sin an wicht
An feiwe sol habin die sinne
Wer mit ic rede so der minne

Die sol ich habin den mit
Was man rede uel al gut
Daz sie antworte alle feist
Jach nach der man ist
Alld darnach er hab kaget
So ist die feiwe und er gewest
An feiwe hat ande sinne gnuet
Daz sie hoffnet sy und gefuget
Hab ich geleide die sint gut
Mit schoner rede künfte mit
Wol sie dann sinne mere
So hat die zucht von die leze
Besage mit was sie sinne hat
An bedarf ic mit zu potestat
An man sol haben künfte wil
Wol feiwen zucht die wib
Daz an feiwe sy an arzen wän list
Die edel und bideck ist
An unelich stat den feiwen wol
Jedoch an ieglich feiwe sol
Dabn die leze und die sinne
Daz sie sich künfte vor summine
Wan hauffet minne die das
Daz unminne hauffet das
Ich stult mit minne ist die gut
Wer ic mit pnschte hat
Schöne feiwe geburt diehtu minne
Gut condeicht die an sinne
Schöne ist erwicht danne sy
Sin und unguete dy
Welch' man mit sinne hat
Der gut sinen feinden lösen wart
Wer sine sin ist wolgeborn
Des adel ist vil gut wolden
An ieglich richum ist erwicht
Wirt er mit sinne getalt nicht
Minne wirt die zu summine

Postcorrection

- Fully automatic
 - No own approach planned
 - Monitoring applicability and transferability of OCR-D solutions
 - Cooperation with yet to be filled AI professorships (NLP, Computational Humanities, ..) at the University of Würzburg
- Interactive / semi-automatic
 - Confidence-based interactive error spotting
 - Incorporating dictionaries
 - Combination of both?

Using Confidence Information

0002.bin 100% 135%

SEGMENTS LINES TEXT

dan war vmb er fol nit vil wort re

dan war vmb er fol nit vil wort re

en od och nit mit nremā anders δ

en od och nit mit nremā anders δ

ch reden. alle zit ym die gefuntheit

ch reden. alle zit ym die gefuntheit

ver bergen noch vſchigē vnd nüm

ver bergen noch vſchigē vnd nüm

nit endür. Dar vmb iſt des Cyru

nit endür. Dar vmb iſt des Cyru

Da wider zü famen zebzingen oder

da wider zü famen zebzingen oder

gātz zemachen als es vor iſt gewef

gātz zemachen als es vor iſt gewef

Settings

- Show Diff
- Show Prediction
- Show Confidence

Threshold 0.90

- Word Confidence
- Glyph Confidence

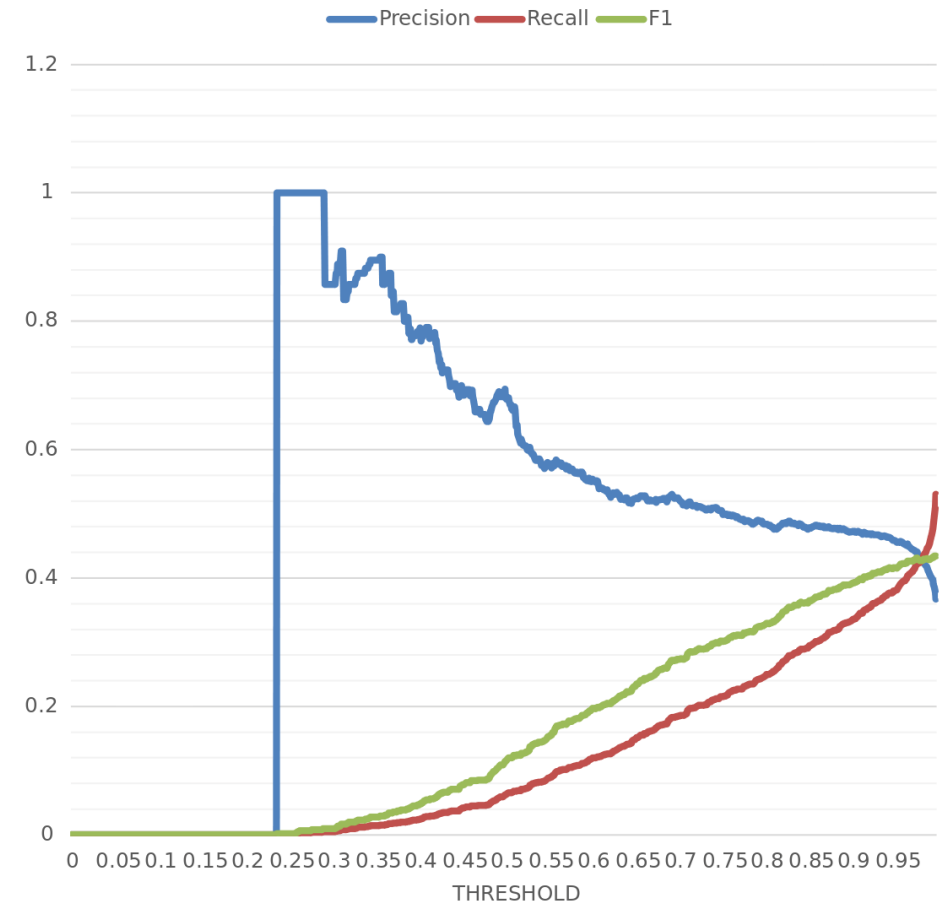
- Only show lines below threshold

- Show Glyph alternatives above threshold

Threshold Glyph 0.30

- Only show lines with Alternatives

... of the OCR engines to identify and highlight suspicious spots



Which confidence threshold to choose?
Trade-off! (Precision/Recall problem)

Incorporating Dictionaries

Developments in DFG project
[Camerarius digital](#) (Norbert Fischer)

The screenshot displays the Camerarius digital software interface. The top menu bar includes SEGMENTS, BASELINES, LINES, and TEXT. The main window shows a Latin manuscript page with OCR text. The text is: *Defensor. Tunetisq; caput nunc obsidet urbem Maurorum imperij. Quo cum se inferre pararet Exp:ans Siculo uenientes & quore naues, Tunc insigne crucis posteaquam Christidos una Extulit ille manu sacræ uenerabile signum Militiæ, positis genibus sic esse locutus Armorum sese firur cingente corona. O soci, neq; enius parua aut contempta manus est Copia in his uestre CAROLO duce & auspice castris. Hæc nos su, cepi præcedent signa ducti, Hæc ego capta mea dextra, si quaeritis, ipse, Complectar uosq; ante geram comprehensa fideli. Quare si qua ducis reuerentia, si qua potentis Numiuis, & si qua est communis cura salutis, Pro se quisq; uiri tota huc incumbite mente, Per uos ut summi cælorum gloria Regis, Perq; ut uos uestre grandescat gloria Genis, Siue itala de stirpe sati seu sætidos oræ Cultores nostri regni, seu bellica misit Ad graue discrimen sancti Germania Martis, Cernitis hoc toto nil usquam ex agmine deficit, Sunt naues quibus inductis cælabitur æquor, Quas neq; Perfarum quondam feruilia regis Agmina, nõ etiam mare prouertæ hoc quoq; in ipsum Athides, aut opere aut superent felicibus armis. Nec tamen unquam aliæ memorantur clasibus istis Instruæ melius falsas fluxisse per undas, Solis est annis opus & formidine casso Peccore, in imbellem tales proficiscimur hostem.*

The right side of the interface shows the OCR text with a search window for the word "esiræ". The search results show "esiræ not in Dictionary!" and a button "+ ESIRAE". The bottom right corner has buttons for "AUTOMATIC" and "RERUN OCR".

Usability Optimization

- Good usability and user experience indispensable for user-centred application
- Increase in complexity of OCR4all unavoidable
- Nonetheless, the use of the system must be as low-threshold as possible
- Cooperation with the Chair for Human Computer Interaction at the University of Würzburg (Prof. Dr. Marc Latoschik, Kristof Korwisi)
 - Ongoing user studies closely linked to the development process
 - Targeted GUI optimizations
 - Expansion of user manuals to incorporate new functionality



Usability Optimization

■ Methods:

- Expert review
- Use case-driven user studies
 - One-to-one sessions
 - Group session

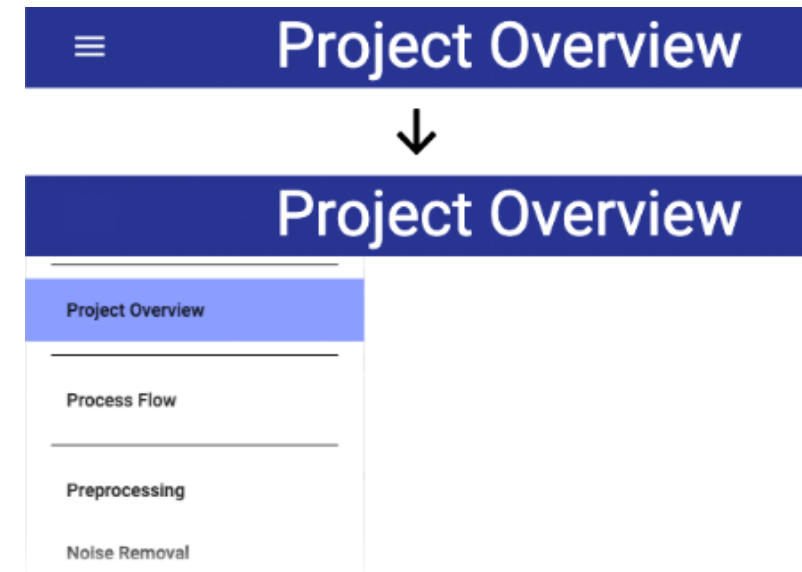
■ Outcomes:

- Identified several critical (where is the main menu???) and non-critical usability problems
- Some issues addressed right away (user manuals PDF → MD, ...)
- Input for OCR4all frontend rewrite

■ Upcoming:

- Synchronization measures OCR4all ↔ LAREX („Look and Feel“)
- Dedicated LAREX studies
- OCR4all: next iteration

1. Laden Sie das Projekt „GNM“ mit der Nummer Ihres Breakoutrooms (Breakoutroom 1 = 001-GNM, ...).
2. Führen Sie den Schritt Preprocessing mit den vorgegebenen Werten aus.
3. Führen Sie den Schritt Segmentation aus, verwenden Sie dazu bitte LAREX.
4. Segmentieren Sie die Seite mit den vorgegebenen Einstellungen und speichern sie das Ergebnis ab.



Upcoming “Milestones”

- Next release (before Christmas?)
 - Calamari 2
 - Kraken segmentation
 - LAREX confidence view
 - ...
- Major update (version 1.0?)
 - Full rewrite of the backend
 - Rewrite of the OCR4all frontend (also using Vue.js)
 - Elimination of numerous small-ish but annoying bugs
- Full support of OCR-D processors
- Full rewrite of the LAREX frontend