Transkribus

A research platform for the digitisation, transcription, recognition and searching of historical documents

Günter Mühlberger

University Innsbruck,

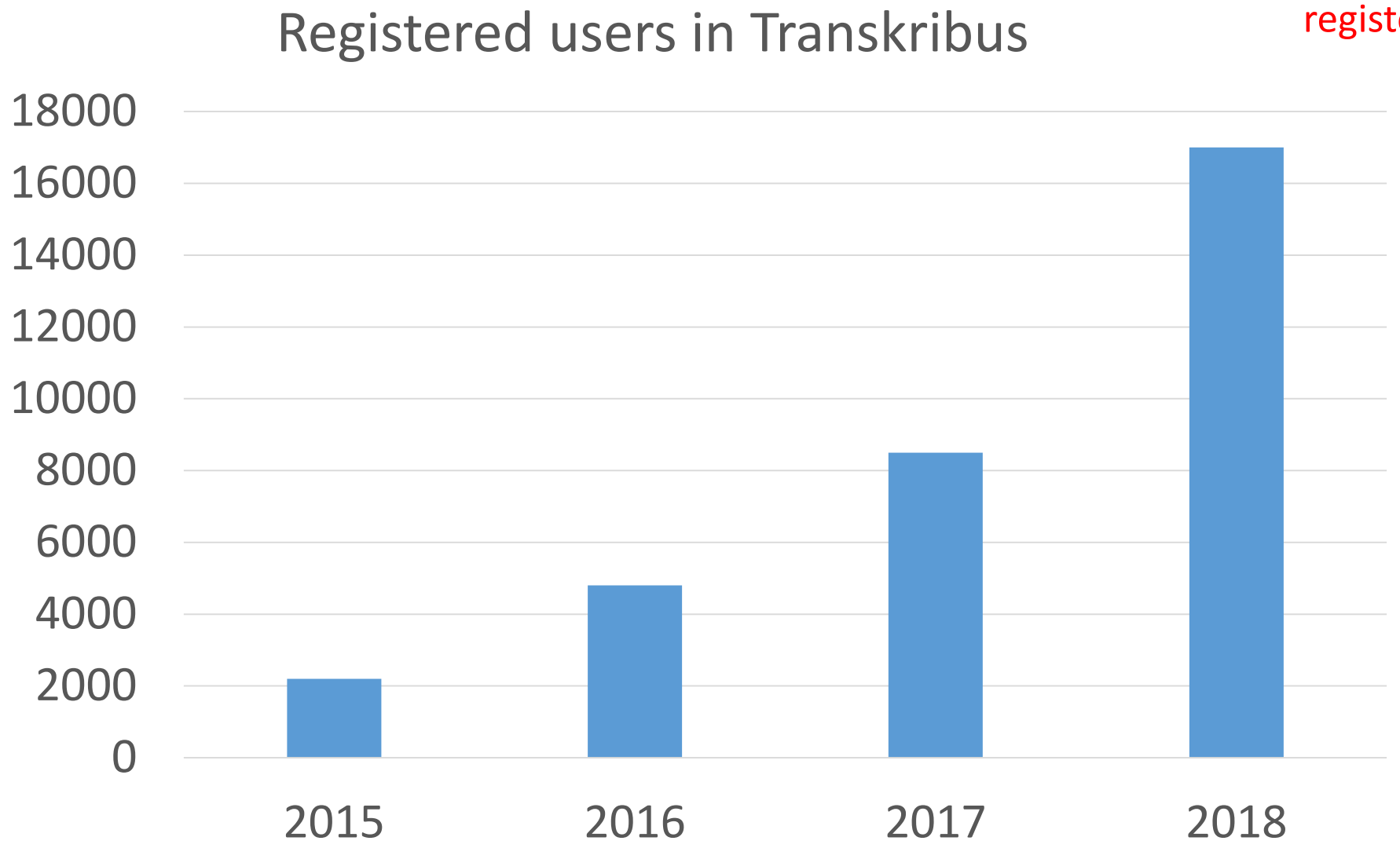
Digitisation and Digital Preservation Group

READ

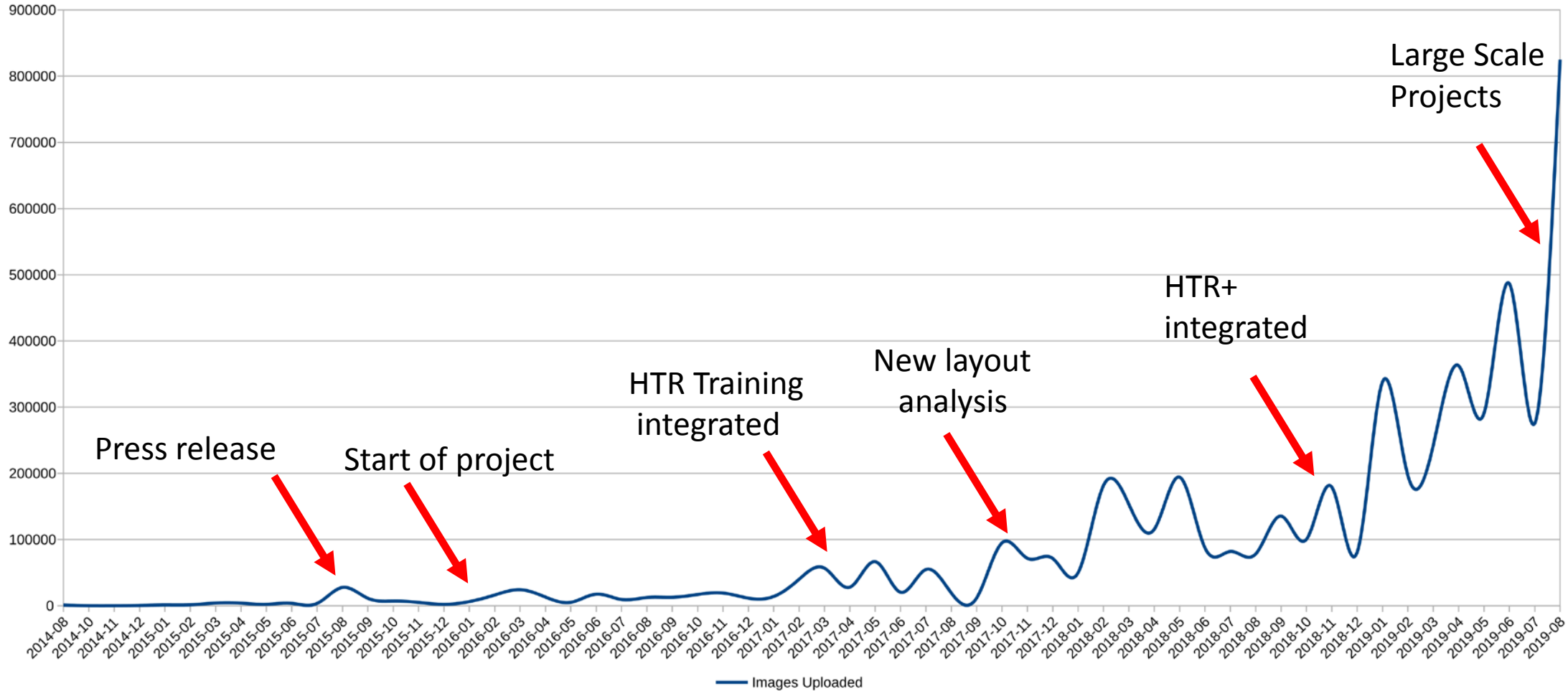
Agenda

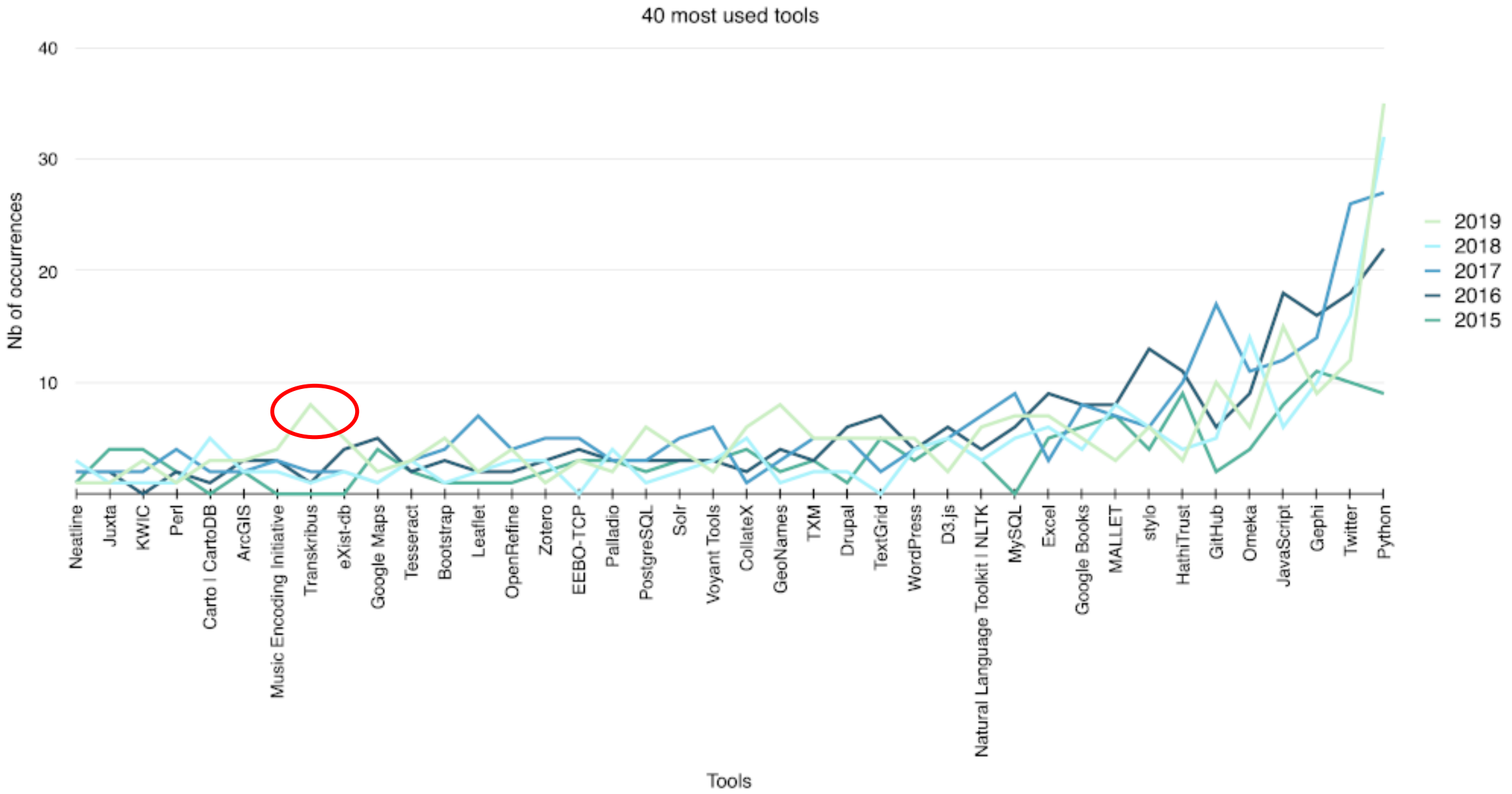
- Introduction
- New features
- READ-COOP SCE
- Q&A

Yesterday: 30,000+ registered users!



Images uploaded per month





Source: WHICH DH TOOLS ARE ACTUALLY USED IN RESEARCH?
BY LAURE BARBOT, FRANK FISCHER, YOANN MORANVILLE AND IVAN POZDNIAKOV — 06 DEC 2019
https://weltliteratur.net/dh-tools-used-in-research/

October 2019

- New Users : 1351
- Active Users / Unique Logins : 2647
- Created Documents: 8904
- Exported Documents: 1406
- Layout Analysis Jobs: 9594 – 249.249 images
- HTR Jobs : 7597 – 295.130 images
- Training runtime: 1799h
- Images Uploaded by users: 1.065.995

Transkribus User Conferences – 2017 + 2018

Join us for 2020 in Innsbruck – 6/7 February 2020!

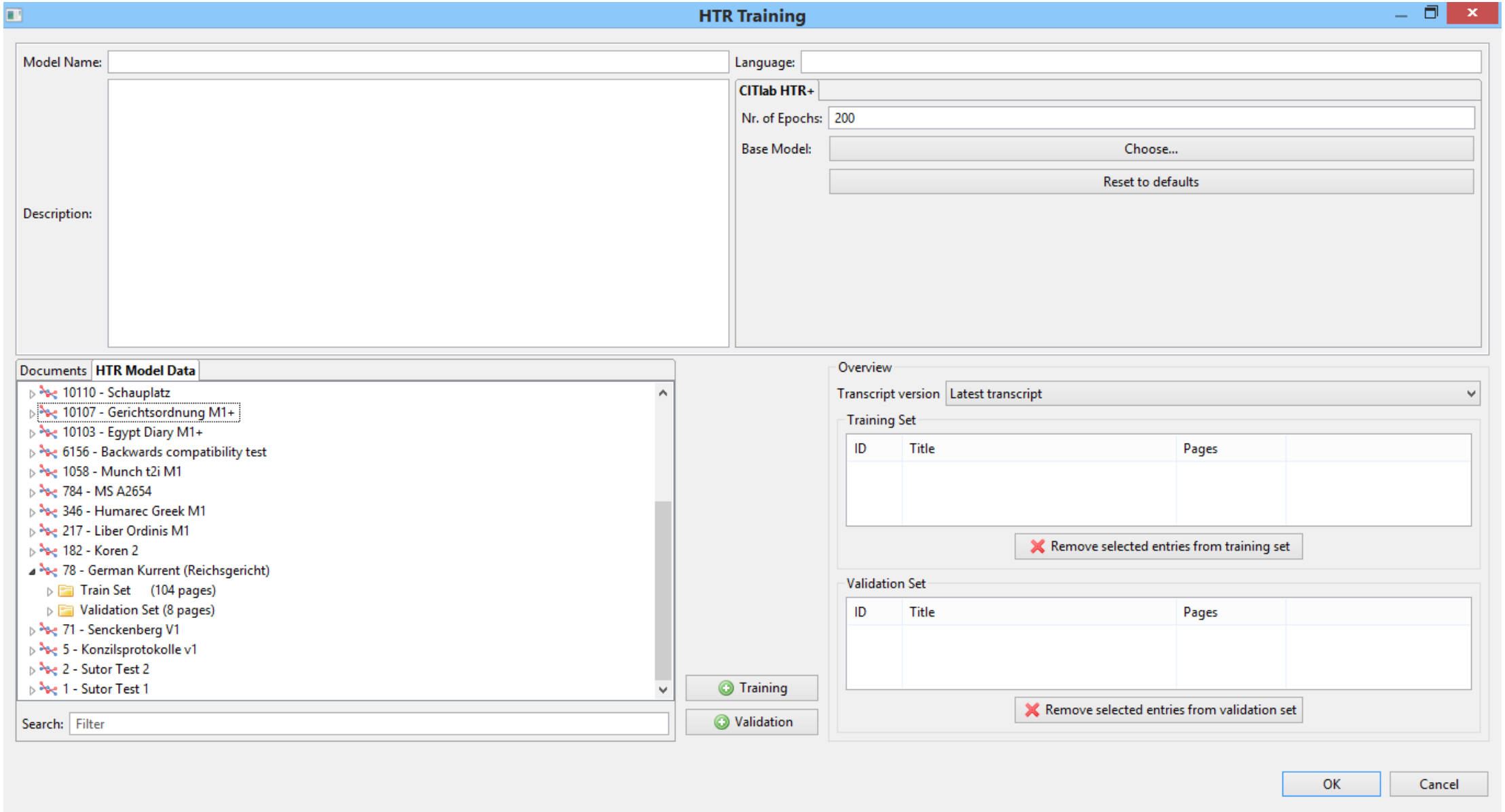
Running projects

- National Archives Finland
- University Archive Greifswald
  - German Science Funds
- Vienna Library
- National Archives Netherlands
- Picturae
  - Amsterdam: Crowd leert computer lezen
  - Local chronicls
- Georgian papers
  - Mellon foundation
- Cadastre documents
  - Tyrolian government
- NewsEye project
  - H2020 Project

New features

New features

- Direct HTR (with default layout analysis)
- Training with existing models
- Training with base-models
- Implementation of P2Pala
- Text2Image matching
- Sample compare


HTR Training
Model Name:
Language:
Description:
CITlab HTR+
Nr. of Epochs:
200
Base Model:
Choose...
Reset to defaults
Documents
HTR Model Data
10110 - Schauplatz
10107 - Gerichtsordnung M1+
10103 - Egypt Diary M1+
6156 - Backwards compatibility test
1058 - Munch t2i M1
784 - MS A2654
346 - Humarec Greek M1
217 - Liber Ordinis M1
182 - Koren 2
78 - German Kurrent (Reichsgericht)
Train Set (104 pages)
Validation Set (8 pages)
71 - Senckenberg V1
5 - Konzilsprotokolle v1
2 - Sutor Test 2
1 - Sutor Test 1
Search:
Filter
Training
Validation
Overview
Transcript version
Latest transcript
Training Set
ID
Title
Pages
Remove selected entries from training set
Validation Set
ID
Title
Pages
Remove selected entries from validation set
OK
Cancel

Choose a model

Search: Filter Technology: Show all

Name	Language	Curator	Technology	Created	ID
French_18thC_Print	French	info@caromein.nl	CITlab HTR+	05.12.19	19166
ONB_Newseye_GT_M1+ CITlab	german	Unknown	CITlab HTR+	03.12.19	19105
Frakturschrift Mein Kochbuch	German	guenter	PyLaia	02.12.19	19071
PyLaia NAF Court Records M10	Swedish	guenter	PyLaia	02.12.19	19070
Dutch_Gothic_Print	Dutch (16th, 1...	info@caromein.nl	CITlab HTR+	28.11.19	18944
Noscemus GM v1	Latin (partly G...	stefan.zathammer...	CITlab HTR+	22.11.19	18743
NAF test with basemodel	Swedish	guenter	CITlab HTR+	24.08.19	16243
NAN/NHA_GT_M3+	Dutch	vincent.noppe@n...	CITlab HTR+	23.08.19	16203
Dutch Notarial Model 18th Century	Dutch	jirsireinders1989@...	CITlab HTR+	17.07.19	15708
Tuomiakirjat flat subset 4 v1	Swedish	guenter	CITlab HTR+	14.05.19	13550
NZZ Gold Standard M1+	German	guenter	CITlab HTR+	23.04.19	12664
Combined_Full_VKS_2	Church Slavonic	achim.rabus@slav...	CITlab HTR+	14.04.19	12425
HIMANIS Chancery M1+	Latin, French	guenter.hackl@tra...	CITlab HTR+	14.04.19	12423
Egypt Diary M2+	English	guenter	CITlab HTR+	18.03.19	11685
ONB_Newseye_GT_M1+	german	guenter.hackl@tra...	CITlab HTR+	15.02.19	10810
German Kurrent M1+	German	guenter	CITlab HTR+	31.01.19	10384
VMC_Test_4+	Russian Churc...	achim.rabus@slav...	CITlab HTR+	25.01.19	10124
Schauplatz	German	guenter	CITlab HTR+	24.01.19	10110
Gerichtsordnung M1+	German	guenter	CITlab HTR+	24.01.19	10107
Egypt Diary M1+	English	guenter	CITlab HTR+	24.01.19	10103
Passau 4 HTR+	German	Unknown	CITlab HTR+	22.10.18	7146
Backwards compatibility test	English	philip	CITlab HTR	21.09.18	6156
Passau_HTR_plus	Unknown	Unknown	CITlab HTR	07.06.18	3979
Munch t2i M1	Norwegian	guenter	CITlab HTR	07.11.17	1058
MS A2654	Arabic	guenter	CITlab HTR	20.09.17	784
Beckett T2I French	French	Unknown	CITlab HTR	22.08.17	663
Beckett T2I English	English	Unknown	CITlab HTR	22.08.17	662
Humarec Greek M1	Greek	guenter	CITlab HTR	09.05.17	346

Details

Name: French_18thC_Print

Language: French

Description:

This model is based on printed texts in French (Romantype Font) that was used in Flanders (Low Countries), during the 18th century. The type of sources used for this model, are books of ordinances, which contained the norms ('laws') at the time.

This model has been the result of one of the KB National Library of the Netherlands Researcher-in-Residence position 2019. The project was called

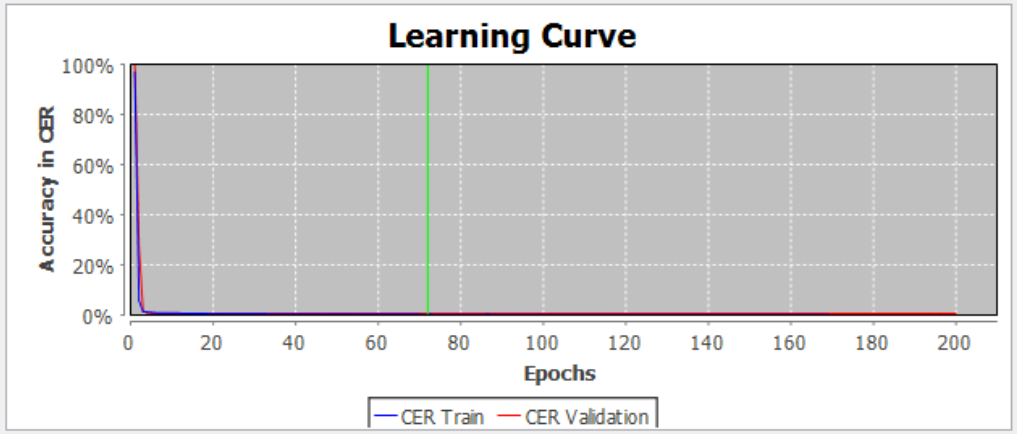
Parameters:

Nr. of Epochs	200
HTR Base Model ID	19148
HTR Base Model Name	French_18thC_Print

Nr. of Words: 38487

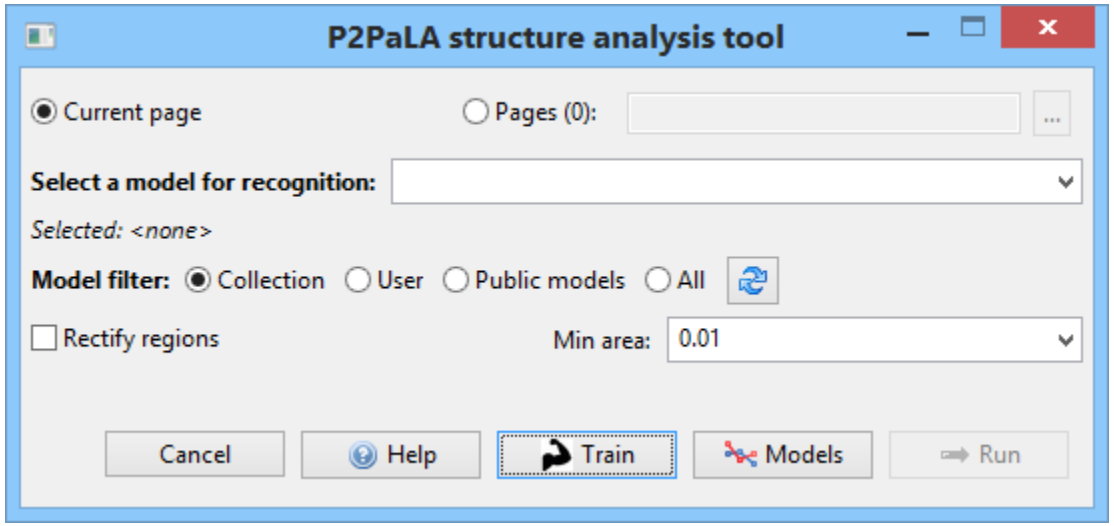
Nr. of Lines: 3883

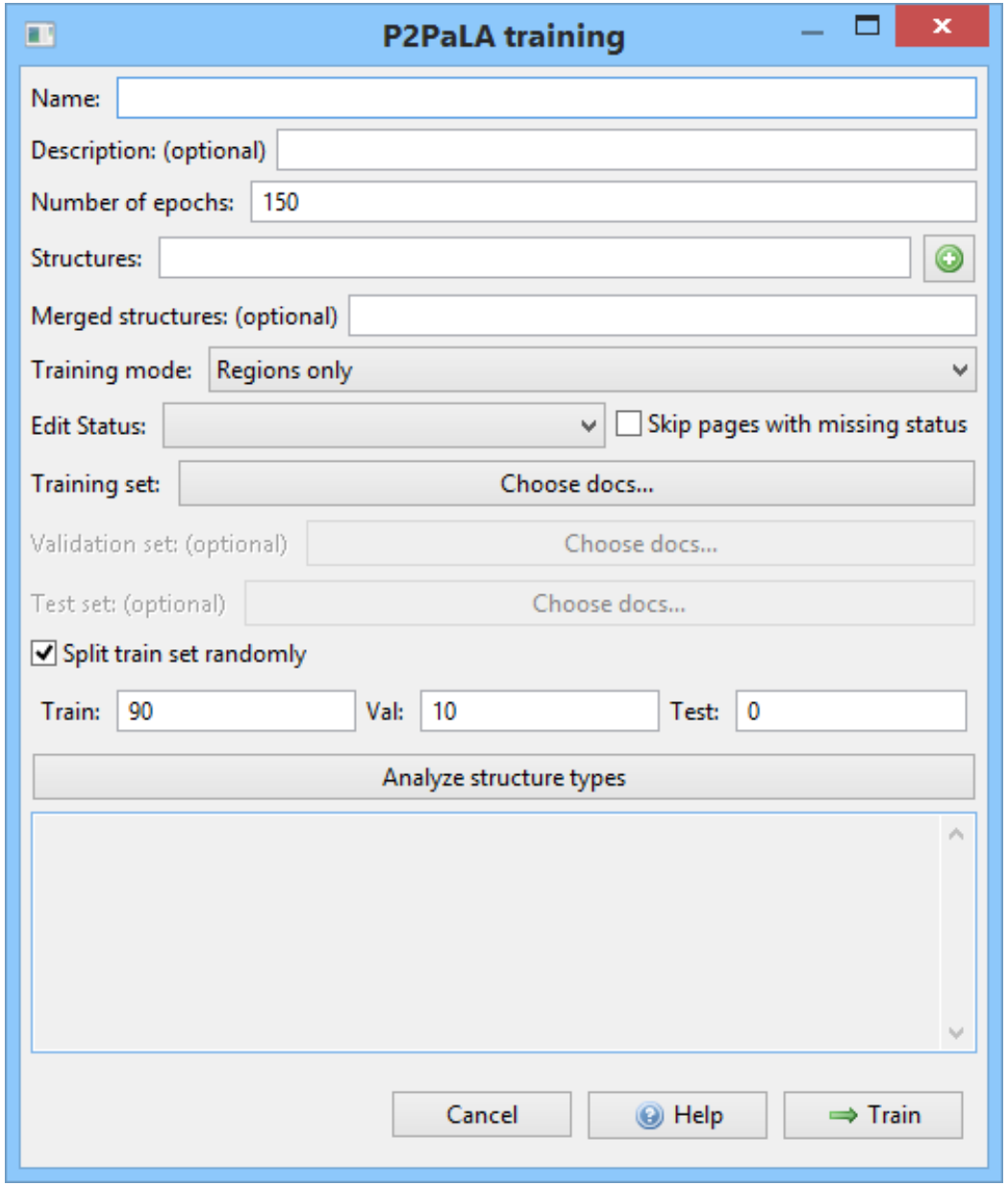
Save | Show Train Set | Show Validation Set | Show Characters

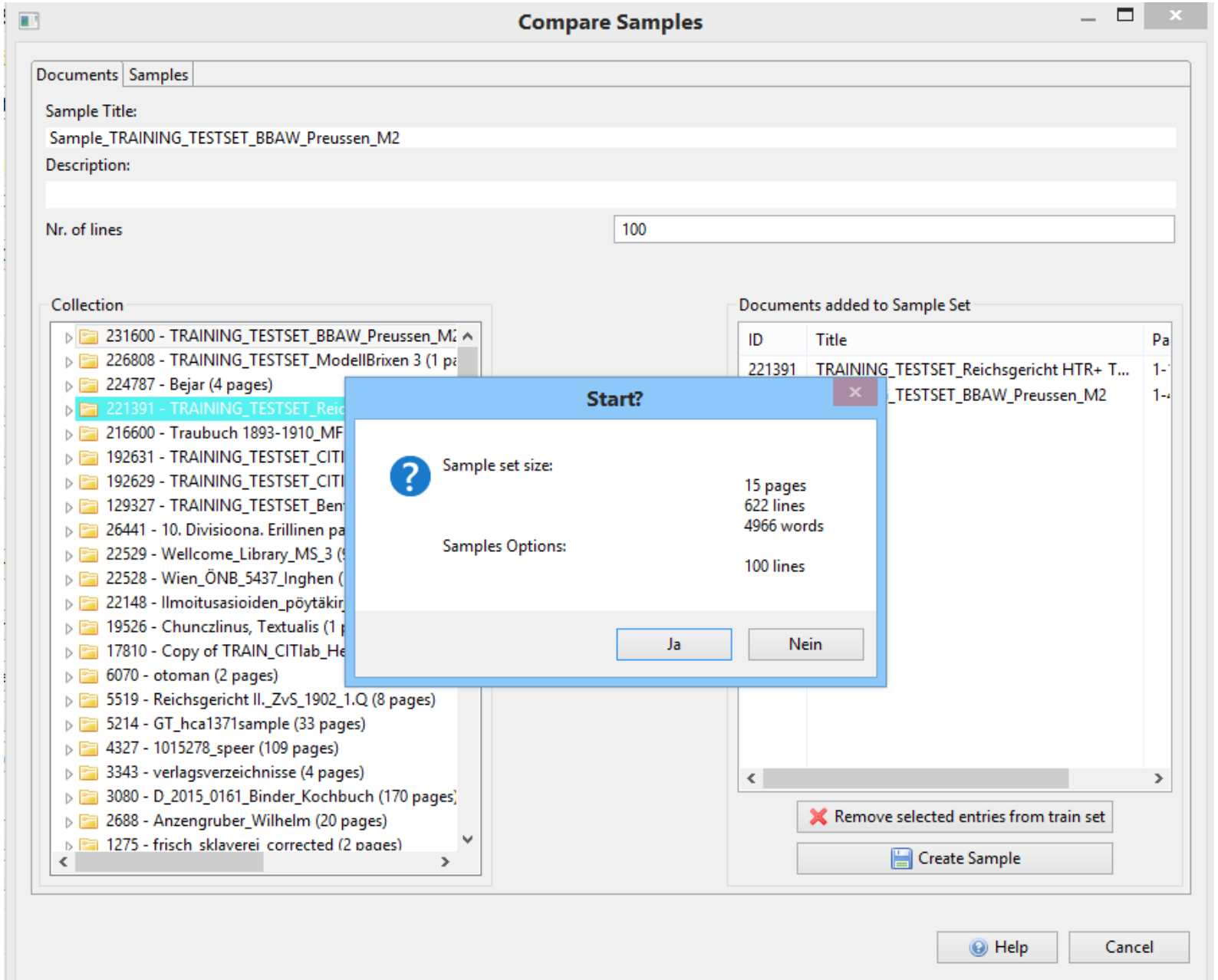


CER on Train Set: 0.33% CER on Validation Set: 0.65%

OK | Cancel

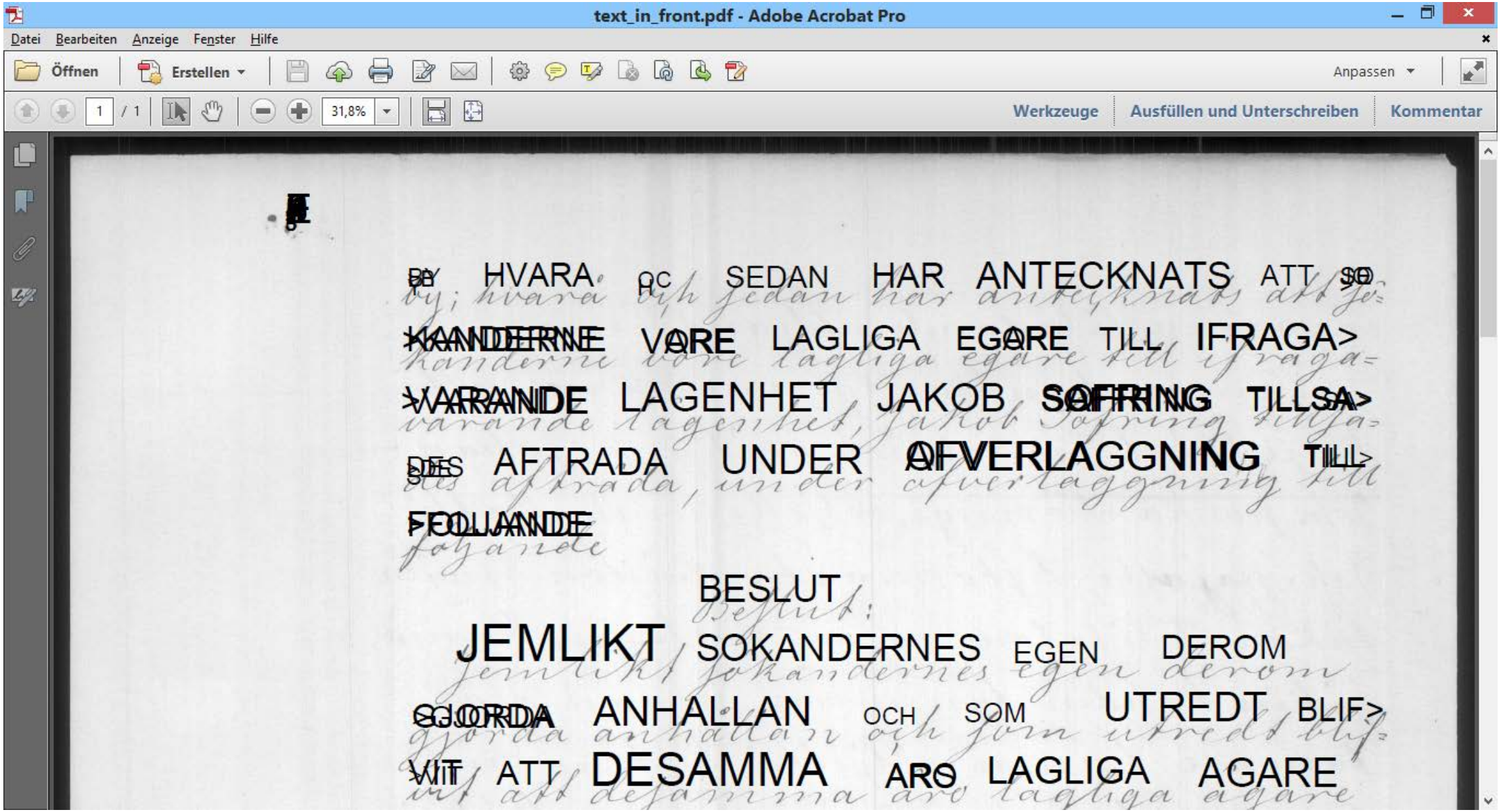

P2PaLA structure analysis tool
Current page
Pages (0):
Select a model for recognition:
Selected: <none>
Model filter:
Collection
User
Public models
All
Rectify regions
Min area:
0.01
Cancel
Help
Train
Models
Run


P2PaLA training
Name:
Description: (optional)
Number of epochs:
150
Structures:
Merged structures: (optional)
Training mode:
Regions only
Edit Status:
Skip pages with missing status
Training set:
Choose docs...
Validation set: (optional)
Choose docs...
Test set: (optional)
Choose docs...
Split train set randomly
Train:
90
Val:
10
Test:
0
Analyze structure types
Cancel
Help
Train


Compare Samples
Documents
Samples
Sample Title:
Sample_TRAINING_TESTSET_BBAW_Preussen_M2
Description:
Nr. of lines
100
Collection
231600 - TRAINING_TESTSET_BBAW_Preussen_M2
226808 - TRAINING_TESTSET_ModellBrixen 3 (1 pa
224787 - Bejar (4 pages)
221391 - TRAINING_TESTSET_Reic
216600 - Traubuch 1893-1910_MF
192631 - TRAINING_TESTSET_CITI
192629 - TRAINING_TESTSET_CITI
129327 - TRAINING_TESTSET_Ben
26441 - 10. Divisioona. Erillinen pa
22529 - Wellcome_Library_MS_3 (
22528 - Wien_ÖNB_5437_Inghen (
22148 - Ilmoitusasioiden_pöytäkir
19526 - Chunczlinus, Textualis (1 p
17810 - Copy of TRAIN_CITlab_He
6070 - otoman (2 pages)
5519 - Reichsgericht II._ZvS_1902_1.Q (8 pages)
5214 - GT_hca1371sample (33 pages)
4327 - 1015278_speer (109 pages)
3343 - verlagsverzeichnisse (4 pages)
3080 - D_2015_0161_Binder_Kochbuch (170 pages)
2688 - Anzengruber_Wilhelm (20 pages)
1275 - frisch_sklaverei_corrected (2 pages)
Documents added to Sample Set
ID
Title
Pa
221391
TRAINING_TESTSET_Reichsgericht HTR+ T...
1-
_TESTSET_BBAW_Preussen_M2
1-
Remove selected entries from train set
Create Sample
Start?
Sample set size:
15 pages
622 lines
4966 words
Samples Options:
100 lines
Ja
Nein
Help
Cancel

Keyword spotting – probabilistic index


Google Kalend
40 most used
Transkribus - K
Connection suc
VIS:online
https://transkr
Gäste WLAN
Gäste WLAN
transkribus.eu/r/kansallisarkisto/en
Apps
Kalender
Home - EU Researc...
READ Call
READ WEB
READ WUI
KWS
Kalender
PROJEKTE - Transkr...
Sandbox
KWS SEARCH
www.cardcomplete...
THE NATIONAL ARCHIVES OF FINLAND
Search
Browse
Info
Help
ENG
Search Finnish Court Records
Search and browse district court notification records from 1810 to 1870.
Enter search term ...
Search


text_in_front.pdf - Adobe Acrobat Pro
Datei Bearbeiten Anzeige Fenster Hilfe
Öffnen
Erstellen
Anpassen
1 / 1
31,8%
Werkzeuge
Ausfüllen und Unterschreiben
Kommentar
HVARA
SEDAN HAR ANTECKNATS ATT
VARANDE LAGENHET JAKOB
AFTRADA UNDER
BESLUT
JEMLIKT SOKANDERNES EGEN DEROM
ANHALLAN
DESAMMA
LAGLIGA AGARE


Google Kalend
40 most used
Transkribus - K
Connection suc
VIS:online
https://transkr
Gäste WLAN
Gäste WLAN
transkribus.eu/r/kansallisarkisto/en
Apps
Kalender
Home - EU Researc...
READ Call
READ WEB
READ WUI
KWS
Kalender
PROJEKTE - Transkr...
Sandbox
KWS SEARCH
www.cardcomplete...
THE NATIONAL ARCHIVES OF FINLAND
Search
Browse
Info
Help
ENG
Search Finnish Court Records
Search and browse district court notification records from 1810 to 1870.
Enter search term ...
Search

Further plans

- Sharing of models
  - With and without sharing the training/validation data
  - Tagging of models
- Quick evaluation on best models
  - Upload image, run several models at once in order to get the best results from public models
- 2020?
  - No master plan – finish running projects with satisfied customers and keep the platform going and as open as possible – even with a pricing model in the background
  - Motivate other countries such as Sweden, Denmark, Austria to follow the Dutch example
  - Support usage of Transkribus for students

READ-COOP SCE

- Successfully implemented
  - 9th November officially registered
  - This week: tax number!
- Very positive feedback from many institutions and private persons
  - More than 30 members
- Coop as a powerful tool
  - A tool which gives us the chance to collaborate on a new level – earn money, invest money – however with a general purpose defined by the members
  - Flexible, international, collaborative, democratic,…
- Pricing model
  - Main idea is to keep the platform services free and to apply no hard limits, however if institutions are working professionally with Transkribus services are charged
  - Currently per image file – easy to calculate – in the longer run better measurements (computing time, number of words, etc…) might be more appropriate
  - 12-20 Cent per image file as a rule of thumb, lower prices for printed material

Q&A

- Publishing public models outside of Transkribus
  - ZENODO: Training, validation and testset (contains actual result)
- Better baselines
  - You can train baselines specifically with P2PaLa: if successful we could combine it with the general training of an HTR model
- Structure types
  - Several structure types can be trained (and recognized)
  - HTR can be done specifically for structure types
- Table recognition
  - Naverlabs is working on this – table recognition competition ICDAR 2019
- Crowd-sourcing
  - Plans with Picturae to make it easier with VeleHanden

Thanks a lot for your attention!

More information

https://read.transkribus.eu/

https://transkribus.eu/

https://read.transkribus.eu/coop/

https://scantent.eu/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 674943.