

Pre pokročilých používateľov

15th February 2024



Vaši sprievodcovia pre dnešnú cestu



Mirjam
User Success
Team



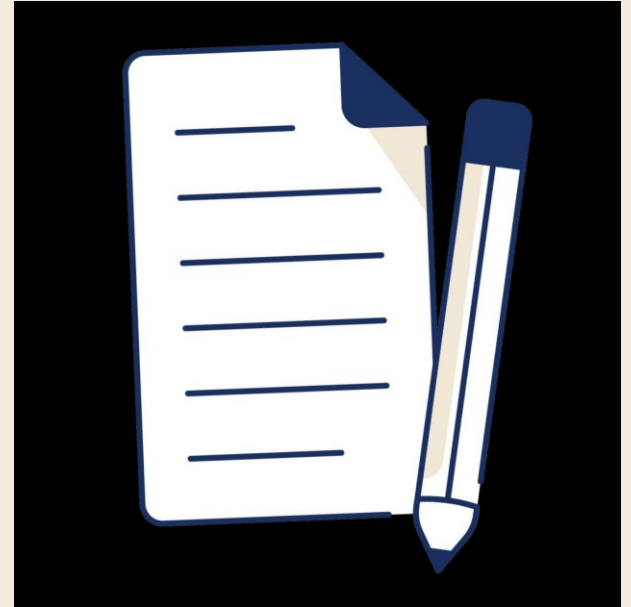
Sara
User Success
Team

s.mansutti@readcoop.eu

m.elattal@readcoop.eu

Obsah workshopu

- 1. Úvod
- 2. Trénovanie & Značkovanie/Tagovanie
- 3. Analýza rozloženia & Základné čiary
- 4. Polia modelov (Beta) & Modely tabuliek
- 5. Zverejňovanie – Stránky Transkribus
-

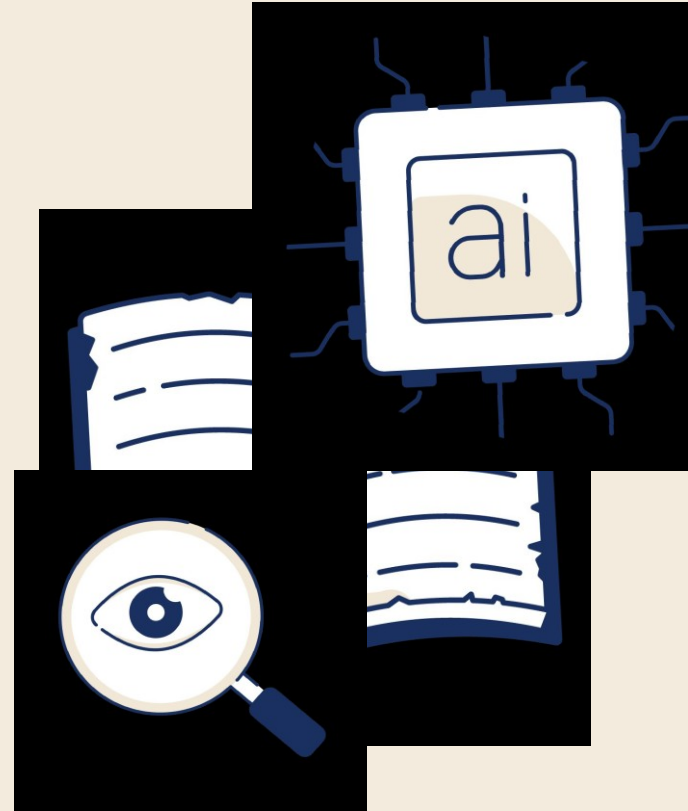




Úvod

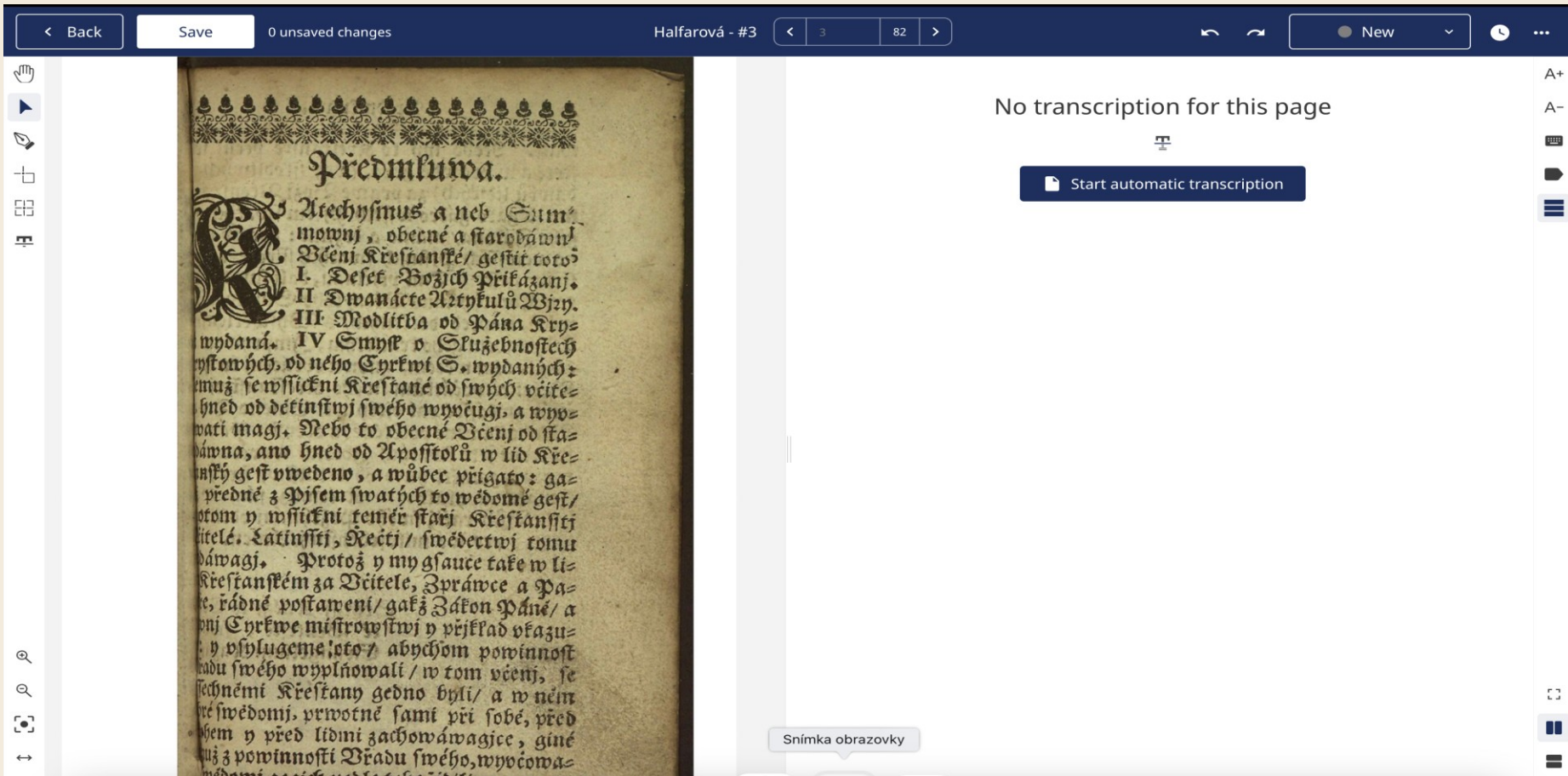
Čo je Transkribus?

Transkribus je váš partner, ktorý pomocou umelej inteligencie (AI) zjednodušuje časovo náročnú a namáhavú prácu s historickými dokumentmi.



Halfarová; Katechizmus, print 1661

Back Save 0 unsaved changes Halfarová - #3 3 82 New



Predmluva.

Prechyslus a neb Sum-
mowj, obecne a starodawni
Vceni Krestanske/ gestit toto
I. Deset Bozich Prikazani.
II Dwanacte Uctyhulú Biry.
III Modlitba od Pána Krys
wydana. IV Smysl o Sluzebnostech
yftowhch. od neho Cyrkwí S. wydaných:
emuz se wssickni Krestane od swych wci-
hned od definstwj swého wyocugi, a wyo-
wati magi. Debo to obecne Vceni od sta-
dawná, ano hned od Apofftolú w lid Kře-
stah gest wvedeno, a wúbec přigato: ga-
předně z Písem swatých to wedomě gest/
otom y wssickni teměr staj Krestanstj
titele. Latinstj, Krectj / swedectwj tomu
dawnagi. Protož y my gsauce take w li-
Krestanském za Vcitele, Zprávce a Pa-
e, rádne postaveni/ gakž Zákon Páne/ a
onj Cyrkwe místrowstwj y přiklad wkažu-
y wshlugeme; otož abychom powinnost
radu swého vyplňowali / w tom vceni, se
schněmi Krestany gedno byli/ a w něm
w swědomj. prvotně sami při sobě, před
hem y před lidmi zachowáwagjce, gine
ž z powinnosti Prádu swého, wyocowa-
wědami z nich w dle se b

No transcription for this page

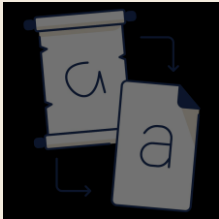
Start automatic transcription

Snímka obrazovky

Začínáme: Prvé kroky v Transkribuse

- 1. Registrácia a prehľad používateľského rozhrania
- 2. Vytvorenie zbierky
- 3. Nahrávanie súborov
- 4. Použitie kreditu

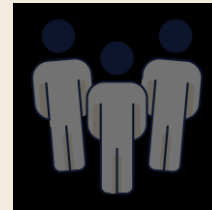
Čo Transkribus umožňuje?



**Manuálny a automatický prepis
ručne písaných a tlačených
dokumentov**



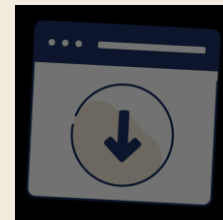
**Trénovanie modelov
umelej inteligencie**



Spolupráca



**Tagovanie štruktúry
a obsahu
dokumentov**



**Export dokumentov v
rôznych formátoch**

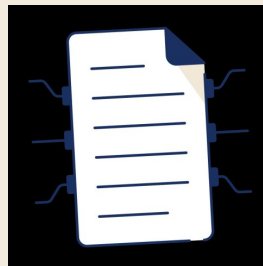
Trénovanie modelov umelej inteligencie

Strojové učenie:



Umožňuje strojom učiť sa z (označených alebo neoznačených) údajov, identifikovať vzorce a robiť predpovede s minimálnym zásahom človeka.

Trénovanie modelov AI



Modely umelej inteligencie:

algoritmy vytvorené počas tréningového procesu systému strojového učenia

predstavujú výstup tréningu/školenia získané vedomosti.

<https://help.transkribus.org/text-recognition>

Trénovanie modelov AI

- **Ground Truth** (Training Data, Základná pravda):
Označené údaje pre tréning, ktoré umožňujú modelu identifikovať vzory a robiť predpovede pre tieto označenia na základe nových údajov.
= všetky strany, ktoré boli prepísané ručne

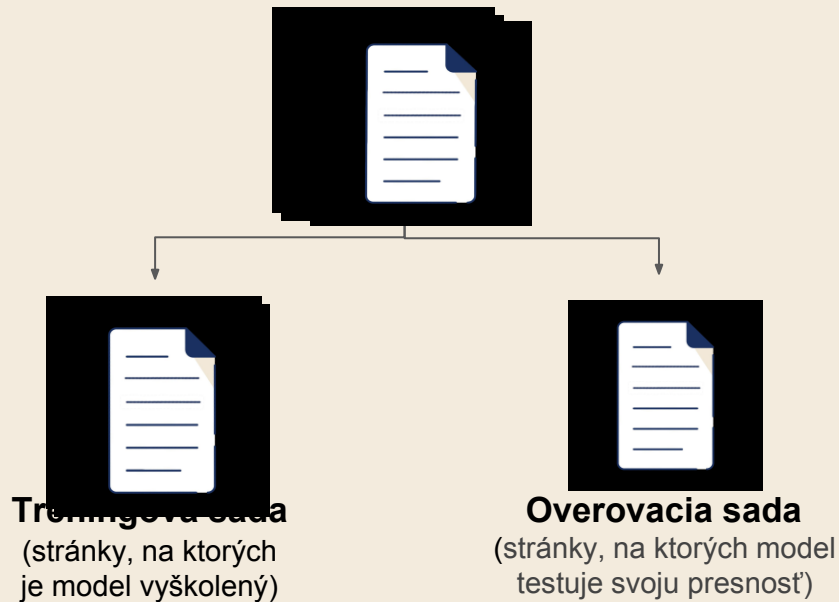
Tréningová sada (Training set)

Súbor príkladov, ktoré sa používajú na úpravu parametrov modelu
= dáta, na ktorých sú postavené poznatky v neurónovej sieti

- **Overovacia sada (Validation Set)**

Súbor príkladov, ktoré sa používajú na objektívne posúdenie výkonnosti modelu
= údaje použité na doladenie parametrov modelu počas jeho tréningu

Ground Truth (Základná pravda)



Dobrá overovacia sada: to je 10% tréningovej sady + obsahuje všetky príklady (znaky, glyfy)

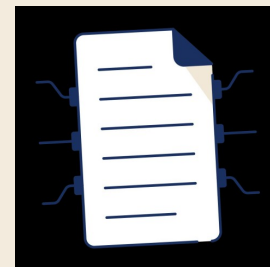


Tréning modelov

<https://help.transkribus.org/text-recognition>



Trénovanie modelov AI



Modely trénovateľné s Transkribusom:

Text

Riadky

Bloky textu

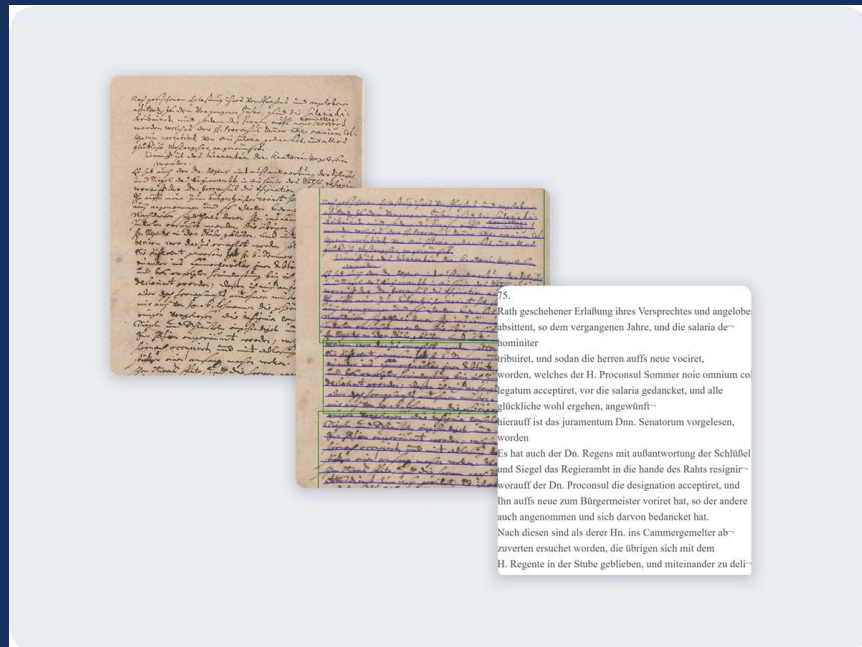
A screenshot of the Transkribus model training interface. It features a list of model types on the left and a '+ Train New Model' button on the right. Arrows point from the labels 'Text', 'Riadky', and 'Bloky textu' to the corresponding model types in the list.

- Text Recognition Model
- Baselines Model
- Field Model
- Table Model

+ Train New Model

Analýza rozloženia (segmentácia)

- 1. Automatická analýza rozloženia
- 2. Rozšírené nastavenia konfigurácie rozloženia
- 3. Manuálna úprava rozloženia
- 4. Základné modely
- 5. Modely polí
- 6. Tabuľky Modely
- 7. Noviny



Trénovanie textových modelov

Modely textu

Pred tréningom modelu:

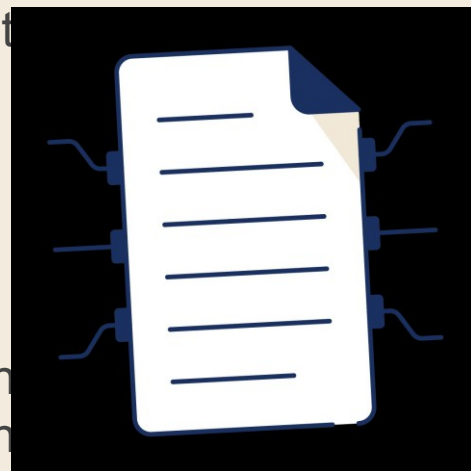
potrebujete **25 až 75 strán (5000-15000 slov)** prepísaného materiálu (**GT_Základná pravda**), v závislosti od typu dokumentu (tlačený alebo písaný rukou)

2 možnosti:

1. Ručný prepis stránky

<https://help.transkribus.org/transcribing-manually>

2. Použitie hotového modelu, ktorý bol trénovaný na podobnom skripte (ak je k dispozícii) a manuálna oprava prepisu



Textové modely

1. možnosť: manuálny prepis dokumentov

1. Vyberte stránky, ktoré chcete zahrnúť do **GT_Základnej pravdy**
2. Spustíte rozpoznávanie rozloženia textu – segmentácia (Layout Recognition)
3. Prepísať od začiatku:

Označte slová, ktoré nemôžete prečítať ako nejasné alebo "medzera"

Riadky, ktoré zostali prázdne: sa v tréningu neberú do úvahy

Skratky: udržiavané/riešené/označené: záleží na tom, čo očakávate ako konečný výstup

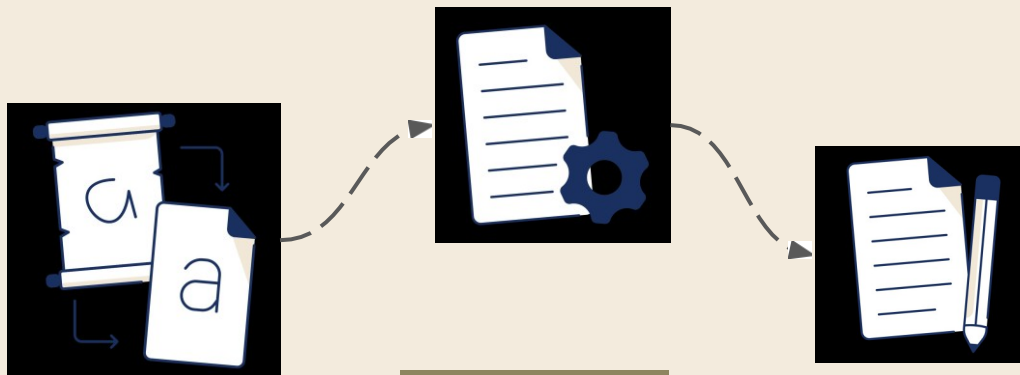
Uložte stránku ako GT "Základnú pravdu"!



Textové modely

2. možnosť: použitie modelu/supermodelu a následná oprava automatických prepisov

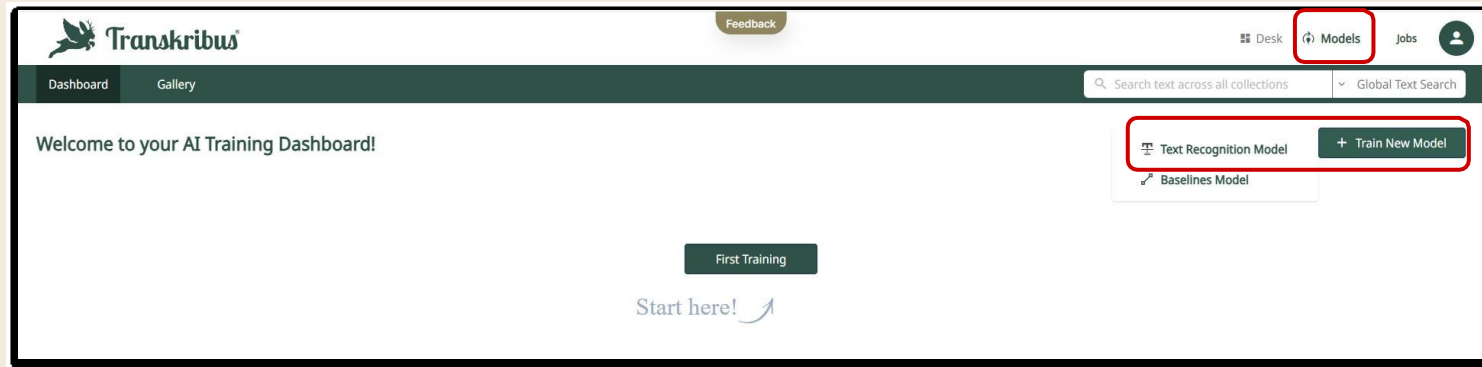
1. Vyberte stránky, ktoré chcete zahrnúť do základnej pravdy
2. Spustenie rozpoznávania textu
3. Oprava automatických prepisov
4. Uložte stránku ako "Základnú pravdu"



Textové modely

Po vytvorení prepisov (Základná pravda):

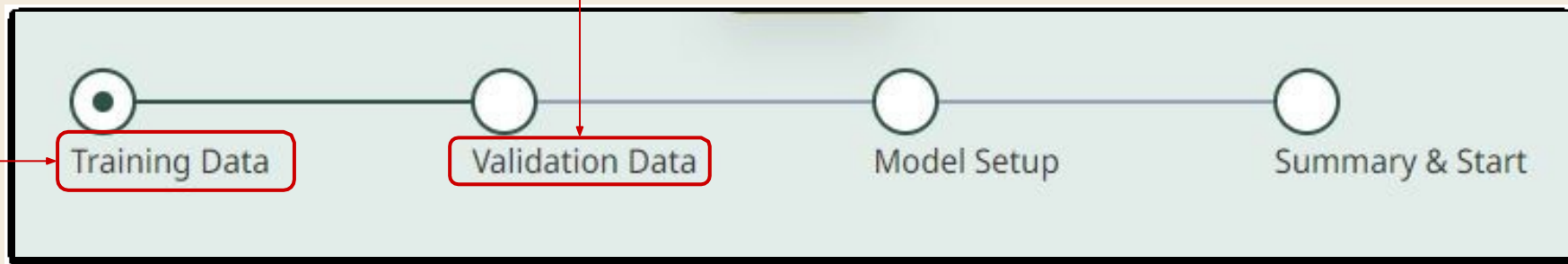
- prejdite do sekcie "Modely"
- kliknite na "Train New Model - Text Recognition Model"
- vyberte zbierku s prepismi Ground Truth



Textové modely

○ Vyberte stránky na:

1. Tréning/školenie (stránky, na ktorých je model školený)
2. Validáciu (strany, na ktorých model testuje svoju presnosť). Dobrá validačná sada: 10% tréningovej sady + obsahuje všetky príklady




Textové modely

Rozšířené možnosti:

Base Model Recommended

Select a pre-existing model to use as the base for your own model.

 Select Model

Advanced Settings (optional) ^

Training Cycles optional

Training Cycles

Enter the number of times you want the model to go through the entire training dataset.

Early stopping optional

Early stopping

Enter when you want to use early stopping to prevent overfitting.

Reverse Text (RTL) Optional

Select if you want the text to be written in a right-to-left direction.

Textové modely

Rozšírené možnosti:

- **Základný model (Base model):** pomocou základného modelu (Base model) tréning nezačína od nuly, ale od toho, čo sa už naučilo v tréningovom procese tohto modelu



Textové modely

Rozšírené možnosti:

- **Tréningové cykly (Training cycles (epochs)):** Maximálny počet prechodov modelu cez celú množinu tréningových údajov. Pri prvom tréningu ponechajte predvolený počet 100 tréningových cyklov
- **Predčasné zastavenie (Early stopping):** Minimálny počet cyklov tréningu. Predvolená hodnota je 20: ak po 20 epochách CER validačnej sady neklesne, tréning sa zastaví

Textové modely

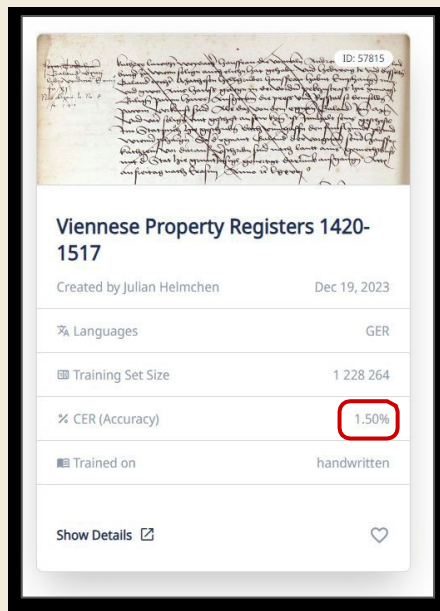
Rozšírené možnosti:

- **Obrátený text (Reverse text (RTL)):** Ak bol text na obrázku napísaný sprava doľava, ale v textovom editore bol prepísaný zľava doprava
- **Použitie existujúcich polygónov (Use existing line polygons):** Pozn.: používať iba v prípade, že ste upravili mnohoúhelníky v *Transkribus Expert*
- **Tréning s rozpisom skratiek (Train Abbrevs with expansion):** Trénuje model tak, aby automaticky označoval skratky a pridal ich rozpis
- **Vynechať riadky s tagmi nejasné/medzera (Omit lines by tag unclear/gap):** Táto možnosť vynecháva riadky obsahujúce slová označené ako gap/unclear.

Textové modely

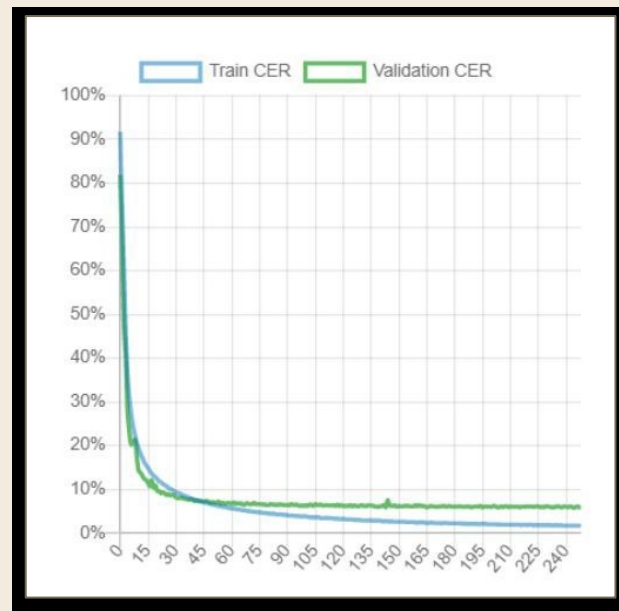
Po dokončení tréningu sa môžete pozrieť na podrobnosti modelu:

1. CER (Chybovosť znakov = Character Error Rate)
2. Krivka učenia



The screenshot shows a model card for 'Viennese Property Registers 1420-1517'. The card includes a thumbnail of a handwritten document, the model name, creator (Julian Helmchen), creation date (Dec 19, 2023), languages (GER), training set size (1 228 264), and CER (Accuracy) of 1.50%. The CER value is highlighted with a red circle. The model is trained on handwritten data.

| Property | Value |
|-------------------|-----------------|
| Created by | Julian Helmchen |
| Created | Dec 19, 2023 |
| Languages | GER |
| Training Set Size | 1 228 264 |
| CER (Accuracy) | 1.50% |
| Trained on | handwritten |

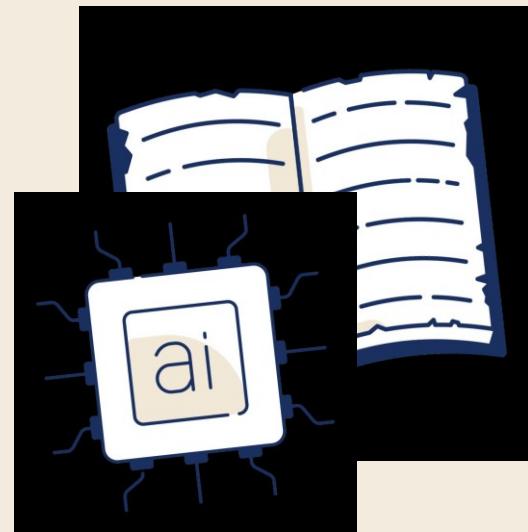


Textové modely

| | CER (chybovosť znakov) | Tréningová sada |
|---|------------------------|-------------------------------------|
| Tlačný text | 0,5-2% | ~ 5.000 words / 25 pages |
| Jedna ruka - jednoduché písanie | 2-4% | 10.000+ words / 50+ pages |
| Niekoľko rúk - zistené | 4-6% | 10.000+ words per hand / 150+ pages |
| Veľa rúk - z toho istého obdobia a regiónu – nie všetky zistené počas tréningu | 6-8% | 100.000+ words / 500+ pages |

Textové modely

- Ruky, ktoré nie sú nijako zistené, alebo načmárané poznámky oveľa horšie výsledky, tak potom:
- Zdvojnásobte počet tréningových dát 20-25% zníženie chybovosti
- **Existujúce modely** sa môžu použiť ako východiskový krok (Base model - základný model) na zníženie požadovaného množstva nových údajov



Textové modely

Verejný holandský rukopisný vzor: [Dutch Margaretha Turnor 17th Century](#)

Trained by The Utrecht Archives; Training set: 178 pages, Validation set: 20 pages

Dutch Margaretha Turnor 17th



by The Utrecht Archives

Nov 28, 2022

🌐 Languages

DUT

📄 Training Set Size

36 289

📊 % CER (Accuracy)

3.10%

📅 Centuries

17

📖 Trained on

handwritten

Model ID

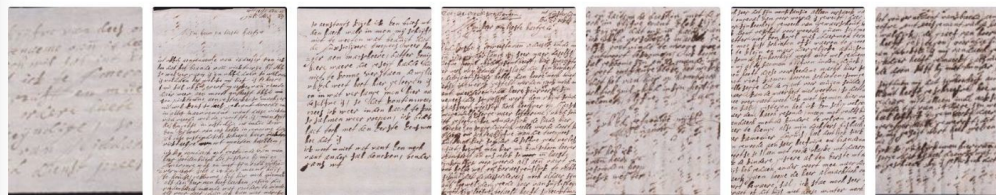
48329

Model description

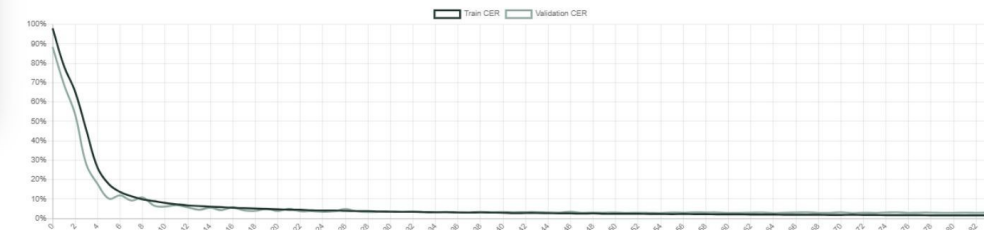
This is the first model created by the Utrecht Archives. It is based on a thousand letters Margaretha Turnor wrote to her husband during the late 17th century. She managed the castle of Amerongen, while her husband worked abroad as a diplomat for the Dutch Republic. Her letters provide an insight into family life in the Dutch Republic as well as the political situation in the country.

Training data

[View all >](#)



Training stats



Textové modely

Verejný model írskej gaelčiny: [Irish, Gaelic and Roman type \(Seanchló agus Cló Rómhánach\)](#)

Trained by Gerard Farrell; Training set: 243 pages, Validation set: 3 pages

Public Model

Irish, Gaelic and Roman type (Seanchló agus Cló Rómhánach) v.3

by farrelgn@tcd.ie Nov 4, 2023

🌐 Languages IRI

📄 Training Set Size 70 965

📊 CER (Accuracy) 1.20%


📖 Trained on print

[Edit](#) [Show Description](#)

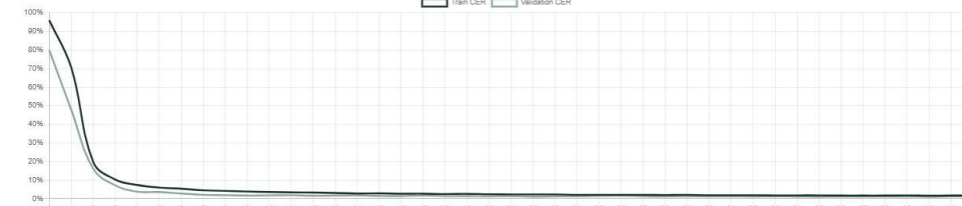
Model description

Model for reading Irish Gaelic (Gaelige) type or seanchló (common pre-mid-20th century). Can also read Irish in the standard Roman typeface used today. This model was trained on over 70,000 words of material in various typefaces from the 17th century to the early 20th, leaning more heavily towards books published from the mid-19th century in Cló Newman. The model can, however, handle text printed in earlier fonts, such as Cló Petrie, which was used in O'Donovan's edition of the Annals of the Four Masters, and the earlier Cló Moxon used in Bedell's Irish version of the Old Testament (1685). Dotted consonants are transcribed as the consonant followed by a 'h', following modern Irish convention, and the Tironian 'y' is transcribed as 'agus'. Around 30% of the training material also consisted of modern printed Irish texts.

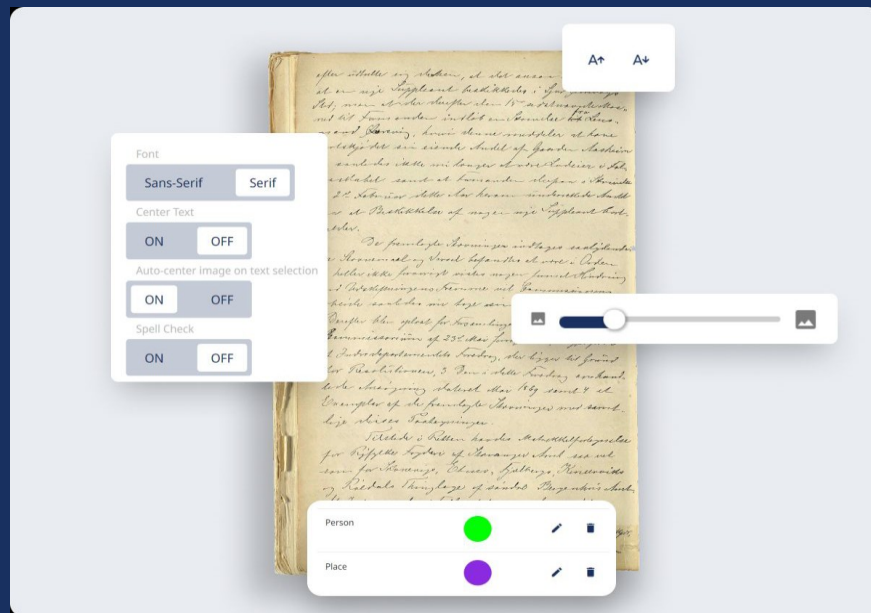
Training data



Training stats



| Iteration | Train CER | Validation CER |
|-----------|-----------|----------------|
| 0 | 95% | 85% |
| 1 | 30% | 25% |
| 2 | 15% | 12% |
| 3 | 12% | 10% |
| 4 | 11% | 10% |
| 5 | 10% | 10% |
| 6 | 10% | 10% |
| 7 | 10% | 10% |
| 8 | 10% | 10% |
| 9 | 10% | 10% |
| 10 | 10% | 10% |
| 11 | 10% | 10% |
| 12 | 10% | 10% |
| 13 | 10% | 10% |
| 14 | 10% | 10% |
| 15 | 10% | 10% |
| 16 | 10% | 10% |
| 17 | 10% | 10% |
| 18 | 10% | 10% |
| 19 | 10% | 10% |
| 20 | 10% | 10% |
| 21 | 10% | 10% |
| 22 | 10% | 10% |
| 23 | 10% | 10% |
| 24 | 10% | 10% |
| 25 | 10% | 10% |
| 26 | 10% | 10% |
| 27 | 10% | 10% |
| 28 | 10% | 10% |
| 29 | 10% | 10% |
| 30 | 10% | 10% |
| 31 | 10% | 10% |
| 32 | 10% | 10% |
| 33 | 10% | 10% |
| 34 | 10% | 10% |
| 35 | 10% | 10% |
| 36 | 10% | 10% |
| 37 | 10% | 10% |
| 38 | 10% | 10% |
| 39 | 10% | 10% |
| 40 | 10% | 10% |
| 41 | 10% | 10% |
| 42 | 10% | 10% |



Tagovanie/Značkovanie

<https://help.transkribus.org/tagging>

Tagging

a. Štrukturálne tagy (Structural Tags):

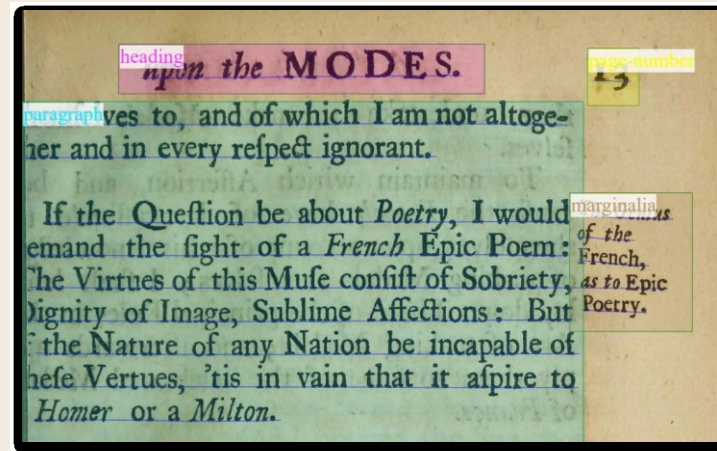
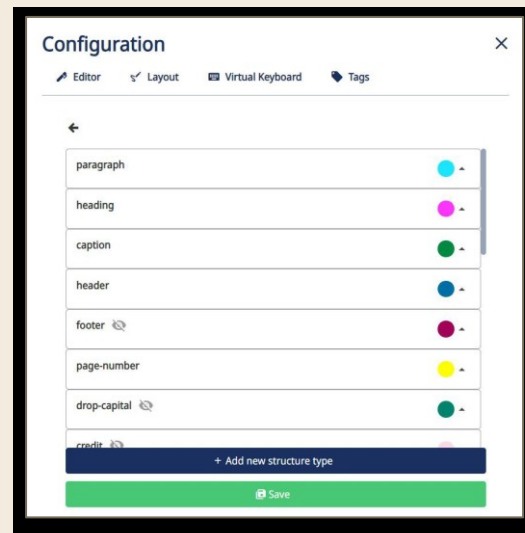
Slúžia na označenie prvkov štruktúry dokumentu

Editor dokumentov: prejdite na **Konfigurácia**
Rozloženie (Layout)

Riadenie typov štruktúry (Manage Structure Types)

Povoľte viditeľnosť značiek, ktoré chcete použiť/pridajte ďalšie značky

Vyberte tvar , kliknite pravým tlačidlom myši a pridajte štrukturálnu značku



Tagovanie/značkovanie

b. Textové tagy/značky: slúžia na označenie prepisu a pridanie atribútov vo vnútri textov

Textový editor: v editore vyberte kurzorom slovo, kliknite na príslušnú značku a pridajte vlastnosť

Správa textových značiek:

Konfigurácia Upravujte značky v nastaveniach kolekcie: pridávajte / odstraňujte značky a upravujte atribúty

[Example](#)

A screenshot of a text editor interface. The main text area shows a document snippet starting with "Augh 1st 1914" and "War declanded between Austria + Servia in morning papers". The word "Austria" is highlighted in purple. A toolbar above the text contains icons for Bold (B), Italic (I), Strikethrough (ABC), Underline (U), Subscript (x₂), and Superscript (x²). Below the toolbar, a dropdown menu is open, showing the word "place" selected and underlined. Below the dropdown, there are three input fields: "Wikidata ID", "country", and "placeName". To the right of the "placeName" field is a red square icon with a white document symbol.

Skratky

According to your needs, you can decide to train the model to:

1. **Ponechajte skrátenú formu v prepise: jednoducho prepíšte skratky ako sú v dokumente**

Nerozpisujeme

→ output: Skratka v texte

2. **Rozpisovanie skratiek:** Neurónové siete sú často schopné naučiť sa rozpoznávať a používať rozšírenia, najmä ak sa objavujú často napíšte rozšírenie skratky do prepisu, venujte dôslednú pozornosť

Rozpisujeme skratky (pozorne, rovnako)

→ output: Skratky. + rozšírenia v texte

3. **Tagujeme a trénujeme skratky vrátane rozpisu :** označte skratku a pridajte zodpovedajúci rozpis do vlastnosti "Rozšírenie" Pri trénovaní modelu vyberte možnosť trénovať skratky

Tagy vrátane rozšírení

→ output: možnosť získať iba skratky, skratky. po ktorých nasledujú ich rozpis alebo náhrada

Skratky

V konfigurácii tréningu
začiarknite políčko

**Train Abbrevs with
expansion
(Trénovať model s
rozpisom skratiek)**

Text Recognition Model

Training Data ✓ Validation Data ✓ Model Setup ○ Start ○

Remove Title

X Diary of John Henry Fisher - Copy

< Back

English ⓘ

Search

Centuries

Base Model Recommended

Select a pre-existing model to use as the base for your own model.

Select Model

Advanced Settings (optional)

Training Cycles optional

100

Enter the number of times you want the model to go through the entire training dataset.

Early stopping optional

20

Enter when you want to use early stopping to prevent overfitting.

Reverse Text (RTL) optional

Select if you want the text to be written in a right-to-left direction.

Use existing line polygons for training optional

Train Abbrevs with expansion optional

Omit lines by tag optional


unclear

gap

Skratky

- Verejný model [UCL–University of Toronto #7](#) trénovaný na riešenie skratiek v stredovekých rukopisoch
 - Training set: 330 pages, Validation set: 30

UCL–University of Toronto #7




by Bentham Project (University College London), DEEDS-project (University of Toronto) Dec 13, 2022

| | |
|---------------------|-------------|
| 🌐 Languages | LAT |
| 📄 Training Set Size | 140 158 |
| 📊 % CER (Accuracy) | 1.70% |
| 📅 Centuries | 13-15 |
| 📖 Trained on | handwritten |
| # Model ID | 48734 |

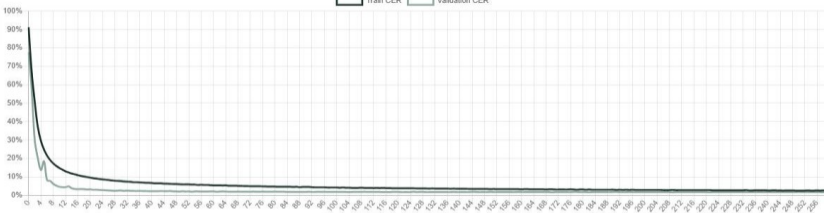
Model description

Seventh iteration of the collaborative UCL–University of Toronto model for processing medieval Latin manuscripts, particularly those containing a large quantity of abbreviated words. E-mail: criley@ucl.ac.uk.

Training data View all >



Training stats



| Epoch | Train CER | Validation CER |
|-------|-----------|----------------|
| 0 | 100% | 100% |
| 10 | ~10% | ~10% |
| 20 | ~5% | ~5% |
| 30 | ~3% | ~3% |
| 40 | ~2% | ~2% |
| 50 | ~1.7% | ~1.7% |
| 60 | ~1.7% | ~1.7% |
| 70 | ~1.7% | ~1.7% |
| 80 | ~1.7% | ~1.7% |
| 90 | ~1.7% | ~1.7% |
| 100 | ~1.7% | ~1.7% |

[Example](#)

Skratky

- Model trénovaný na stredovekých latinských dokumentoch (1520) na rozpoznávanie značky "skratka" vrátane vlastníctva "rozpisu skratiek" Training set: 177 pages, Validation set: 30 pages

| Hertziana_1520_abbrevs | |
|---|-------------|
|  | |
| 🌐 Languages | VAR |
| 📄 Training Set Size | 59 225 |
| 📊 CER (Accuracy) | 19.80% |
| 📖 Trained on | handwritten |
| # Model ID | 38873 |

[Example](#)

مکتب جمعی

دینی ، اجتماعی ، تربیوی ، ادبی ، علمی و فنی در .

سنه : ۱ — ۱۵ اگستوس ۱۳۳۶ — نومرو : ۱

مقصد (۱)

خالفك افكارىنى تئور مقصدىله چقاردىمىز بو مجموعه قارشىمنده دهرىن برهمنونيت دوئيوز و بولى
قىمىلى برونظيفه نلقى ايدىيورز . اكر . بووظيفه ايلهده ، مملكتك بك محتاج اولدىهي معارى دويمولرىنى
فعاليت دولقربلى . . اوباندىرمقده خدمت ايدىبيليرىنك بختبارز .
« مکتب جمعیسى » ، ساغاد ، مطبعه وسائره اجر تدرىنك بك بهالى اولدىهي بوزمانده عرفان توليد
اينمك ، هر كس ايجون فاندلى اولمق، اوزره ساخا انتشاره آتلمق .

Tréningové modely pre RTL písmo

RTL skripty

5 verejných modelov ([public models](#)) pre rôzne RTL skripty v Transkribus

2 verzie osmansko-tureckého tlačového modelu

Vaybertaytsh typ písma (jidiš)

Rukopis jidiš (model Dybbuk)

Zmes historických hebrejských písiem a jazykov (DiJeSt 2.0)



The screenshot shows the 'Text Recognition' section of the Transkribus interface. On the left, there are filters for 'Favorite Models' (0), 'Public Models' (5), and 'Private Models' (20871). Below these is a search bar and a 'Languages' filter. The main area displays a table of models with columns for Name, Words, Language, and CER.

| Name | Words | Language | CER |
|------------------------------------|---------|--------------------|-------|
| OttomanTurkish_Print_v2 | 248 083 | TUR | 7.60% |
| Vaybertaytsh.YidTakNL | 66 497 | YID, HEB | 0.90% |
| OttomanTurkish_Print_1 | 180 854 | TUR | 7.20% |
| The Dybbuk for Yiddish Handwriting | 144 985 | YID | 4.40% |
| DiJeSt 2.0 | 773 726 | HEB, YID, LAD, JUD | 2.00% |

RTL skripty

Ako v súčasnosti prepisovať a trénovať údaje RTL v Transkribuse:

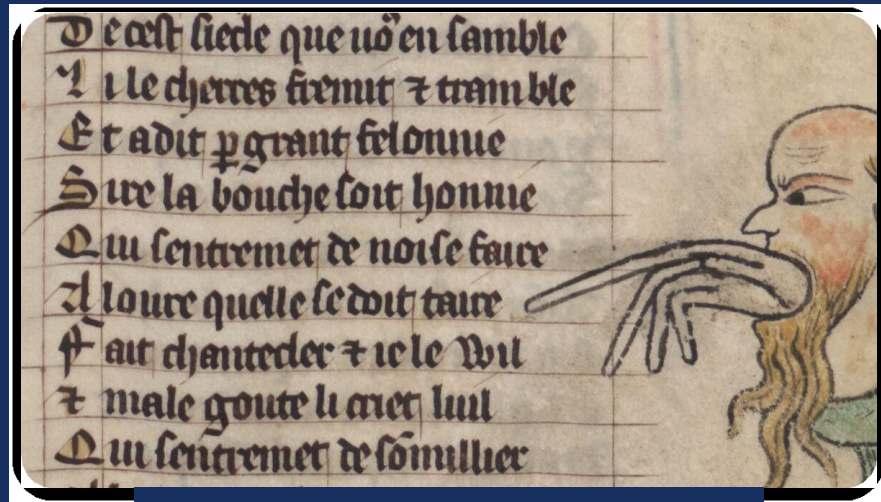
Manuálne spustenie segmentácie (rozpoznávania rozloženia) alebo označovanie rozloženia (oblasti textu + základné čiary) manuálne

- Prepis textu **zleft-to-right** v textovom editore
- V konfigurácii tréningu Rozšírené nastavenia vyberte **Reverse Text (RTL)** tak, aby bol výstupný text napísaný v smere sprava doľava

[Example](#) DiJeSt 2.0 model

Vízia:

- Podpora RTL pre webovú aplikáciu
- Prispôsobovanie konfigurácie tréningu

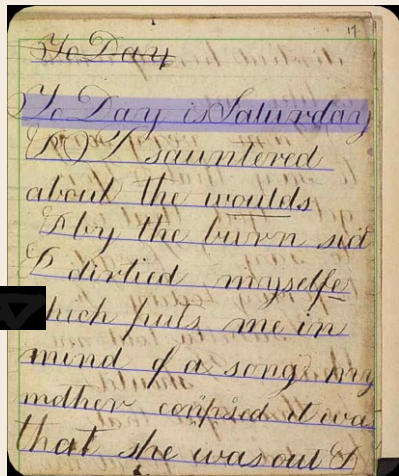


Paris, BnF, Fr. MS 12584 (13th century)

Rozpoznávanie rozloženia (Segmentácia)

Čo sa stane, keď sa stránka rozpozná?

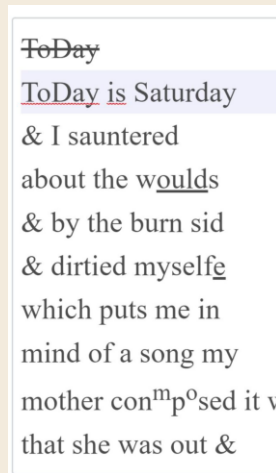
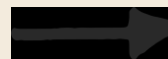
 Recognize



1. krok

Rozpoznávanie rozloženia

(Základné čiary (Baselines) & Bloky textu (Text regions))



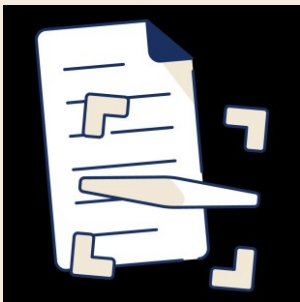
2. krok

Rozpoznávanie textu

Rozpoznávanie rozloženia (segmentácia)

1. krok

Rozpoznávanie rozloženia

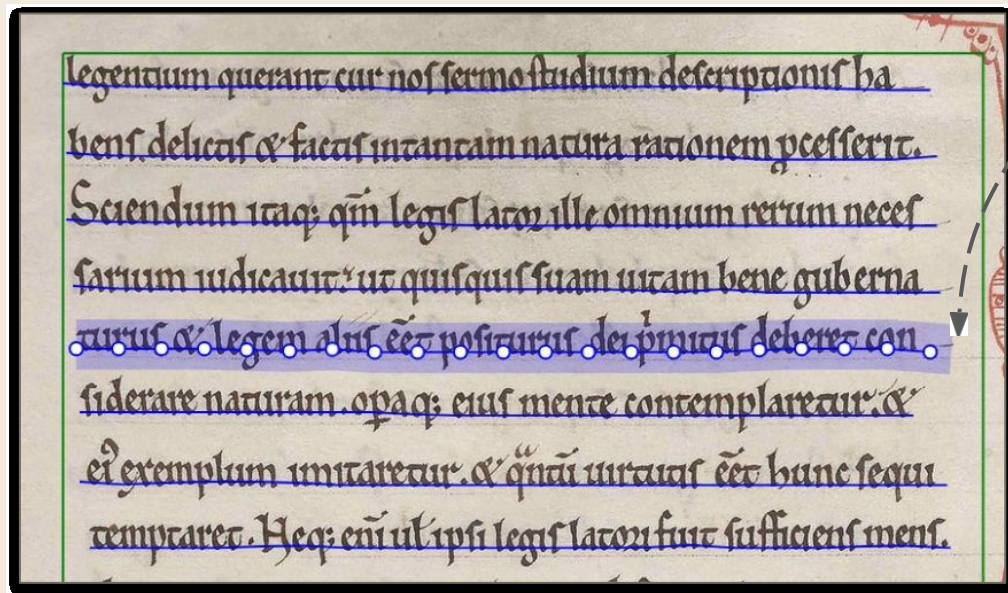


- Analýza rozloženia obrazu dokumentu
- Obrázok je potrebné rozdeliť na textové oblasti a základné čiary
- Základ pre rozpoznávanie a pre transkripciu (prepis)

Tri piliere rozloženia (segmentácie)

1) **Základná čiara** (Baseline):

Členená čiara prebiehajúca pozdĺž spodnej časti riadka rukou písaného textu



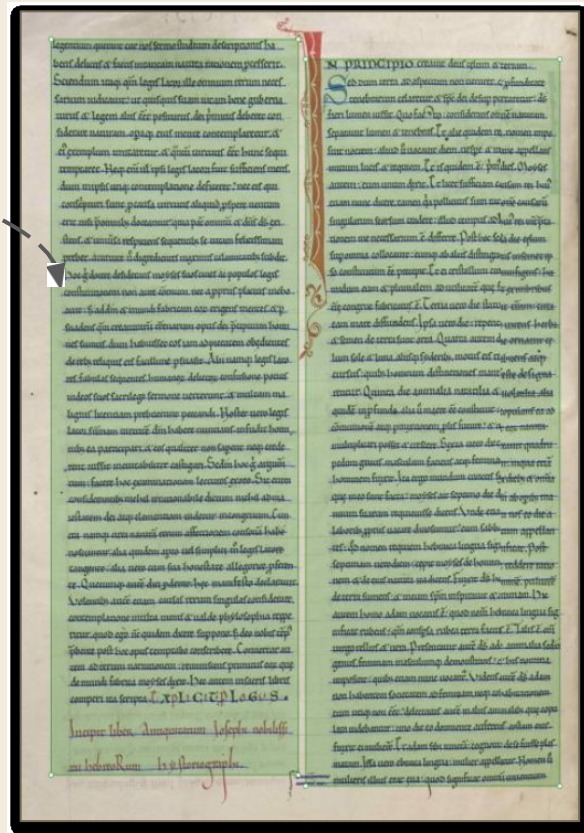
Tri piliere rozloženia (segmentácie)

1) Základné čiary (Baselines)

2) **Bloky textu (Text region):**
obdĺžnikový tvar obklopujúci text

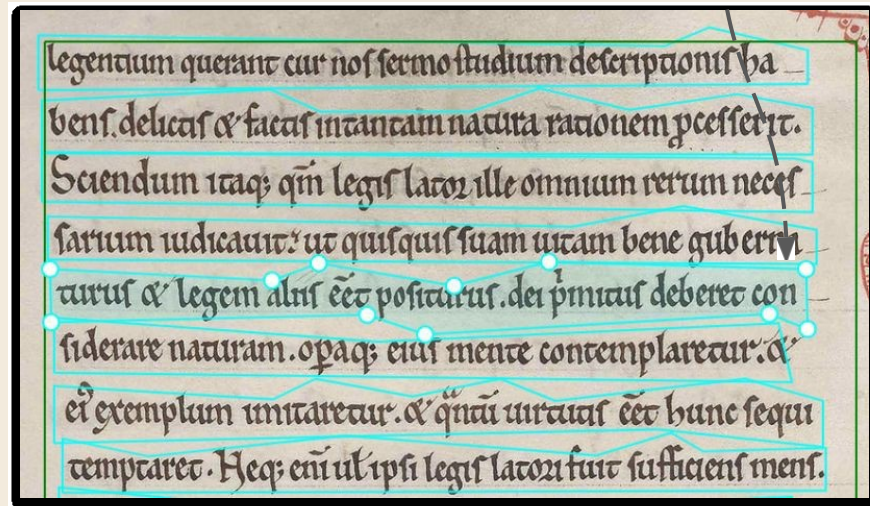
Pri predvolenej analýze rozloženia sú základné čiary zoskupené do blokov textu (textových oblastí na základe ich súradníc (prístup zdola nahor))

Bloky textu (Text region)



Tri piliere rozloženia (segmentácie)

- 1) Baseline
- 2) Text region
- 3) Polygóny riadku (Line Polygons:** mnohoúhelníky, obklopujúce všetok rukou písaný text v riadku



Pri spustení tréningu textu alebo rozpoznávania textu sa mnohoúhelníky čiar vypočítajú algoritmom, počnúc **základnými čiarami**

Tréning a rozpoznávanie textu prebiehajú na úrovni **základných čiar**

Rozpoznávanie rozloženia (segmentácia)

Kvalitu konečného rozpoznania (segmentácie) môže ovplyvniť:

1) Nepresné základné čiary (baselines):

- Zistí sa príliš málo základných čiar (východiskových hodnôt) alebo príliš veľa základných čiar (východiskových hodnôt)

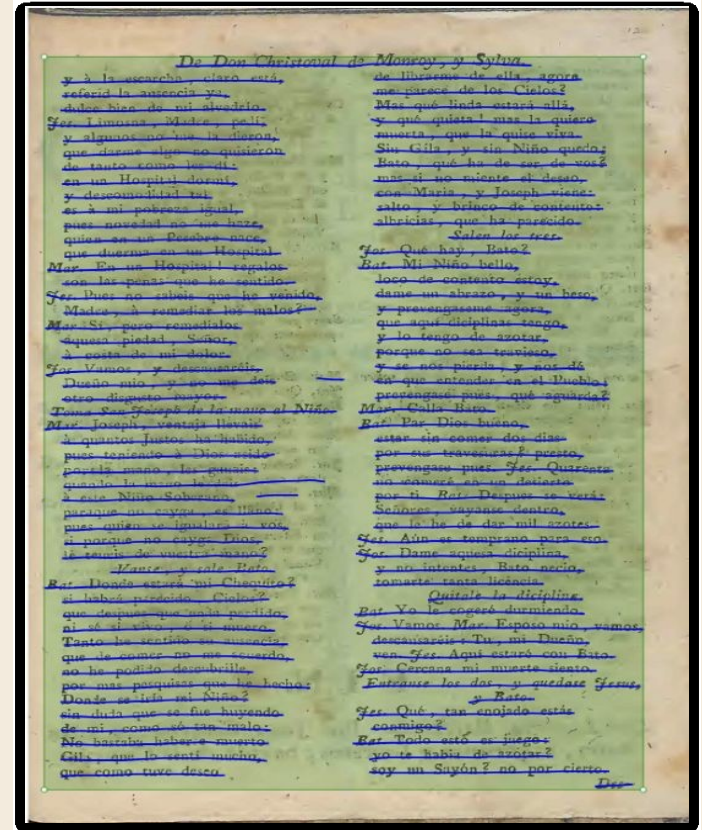


Rozpoznávanie rozloženia (segmentácia)

Kvalitu konečného rozpoznania (segmentácie) môže ovplyvniť:

2) Nepresné bloky textu:

- Nesprávne poradie čítania riadkov;
- Príliš málo blokov textu/príliš veľa blokov textu (text regións)



Rozpoznávanie rozloženia (segmentácia)

Kvalitu konečného rozpoznania (segmentácie) môže ovplyvniť:

3) Nepresné polygóny (Inaccurate polygons):

- Aj keď sú základné čiary správne, modely nedokážu správne prepísať text.
- Riadkové mnohoúhelníky nepokrývajú väčšinu tela písmen/
Polygóny čiar zahŕňajú aj ďalšie (neželané) prvky na strane



Sept 1 / the 1836
~~of the~~
I have been the driver
of the team, through the
woods, not to the
one in with the
thought I did not
go & I am
I have given my
near as much
than 20 times.
I have to have
to the way, it
is with you &
believe me,
your most
after. - A truly
Delan Weston.

Nepresné základné čiary

Nepresné základné čiary



Example

Nepresné základné čiary

Riešenia:

- 1) **Použitie iného verejného modelu základných čiar (Baseline model)**
- 2) **Zmeňte pokročilé nastavenia (advanced settings)**
- 3) **Vytrénujte model základnej čiary (Train a baseline model)**

Nepresné základné čiary

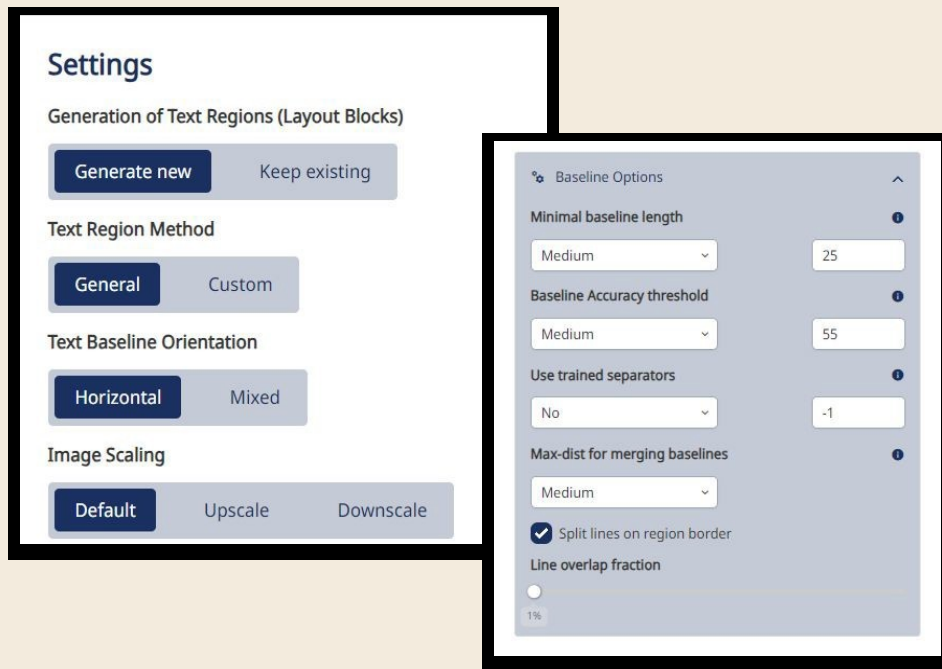
1) Použitie iného verejného modelu základných čiar (Baseline model)

:

- Zmiešaná orientácia riadkov (Mixed Line Orientation)
- Horizontálna orientácia riadkov (Horizontal Line Orientation)
- Univerzálne riadky (Universal Lines)

Nepresné základné čiary

2) Zmeňte pokročilé nastavenia (advanced settings)



The image shows two overlapping panels from a software settings interface. The left panel, titled "Settings", contains the following sections:

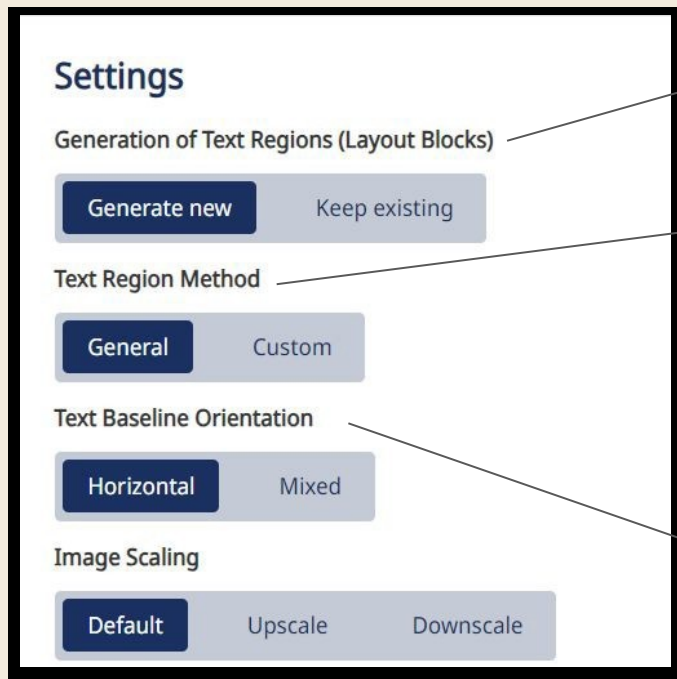
- Generation of Text Regions (Layout Blocks)**: Two buttons, "Generate new" (active) and "Keep existing".
- Text Region Method**: Two buttons, "General" (active) and "Custom".
- Text Baseline Orientation**: Two buttons, "Horizontal" (active) and "Mixed".
- Image Scaling**: Three buttons, "Default" (active), "Upscale", and "Downscale".

The right panel, titled "Baseline Options", contains the following settings:

- Minimal baseline length**: A dropdown menu set to "Medium" and a numeric input field set to "25".
- Baseline Accuracy threshold**: A dropdown menu set to "Medium" and a numeric input field set to "55".
- Use trained separators**: A dropdown menu set to "No" and a numeric input field set to "-1".
- Max-dist for merging baselines**: A dropdown menu set to "Medium".
- Split lines on region border**: A checked checkbox.
- Line overlap fraction**: A slider control set to "1%".

Nepresné základné čiary

2) Zmeňte pokročilé nastavenia (advanced settings)



Generate new: Generovať ďalšie textové oblasti /

Keep existing: Zachovať existujúce oblasti textu (použite to s poľami a tabuľkami)

Po zistení sú riadky zoskupené do textových oblastí. K dispozícii sú dve metódy zoskupovania:

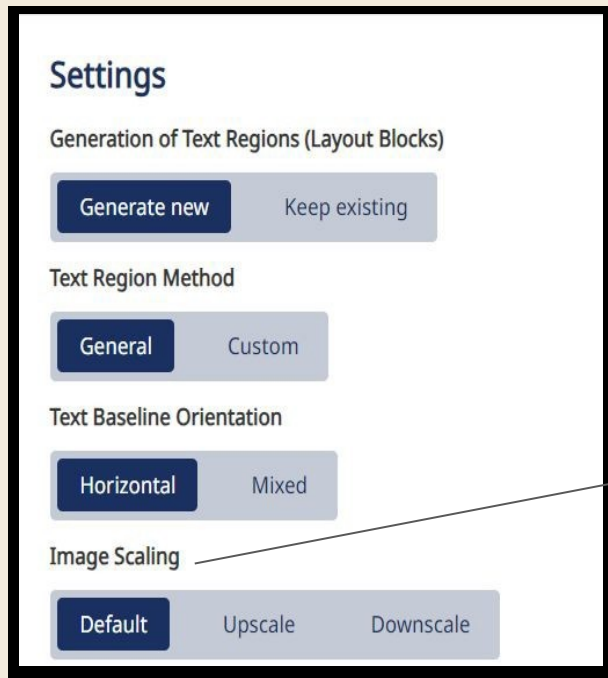
General (Všeobecné): zoskupí čiary zľava doprava

Custom (Vlastné): aglomeračné zoskupovanie založené na bode úplne vľavo každej čiary

Voľba **General**: Výber orientácie riadka textu na zlepšenie klastrovania (zoskupovania)

Nepresné základné čiary

2) Zmeňte pokročilé nastavenia (advanced settings)

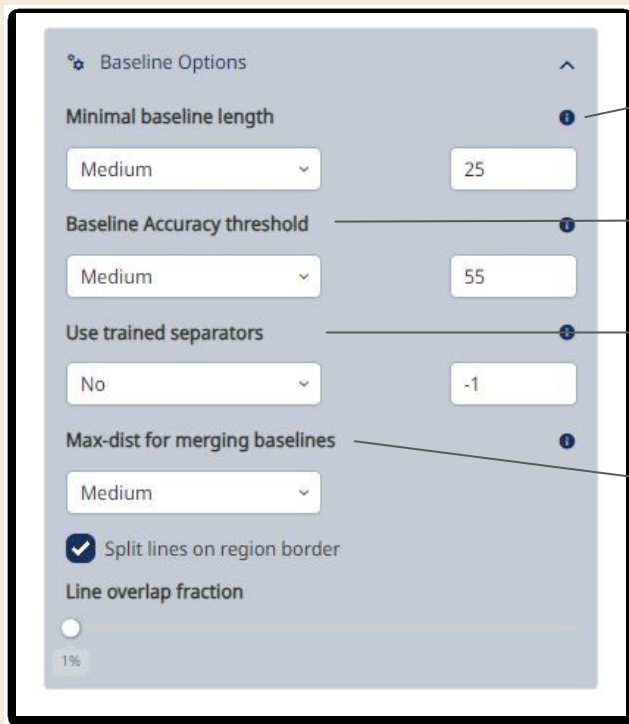


Škálovanie obrázka:

Upscale obrázky s nízkym rozlíšením alebo
Downscale obrázky s vysokým rozlíšením
(túto funkciu použite len v prípade, že
rozpoznávanie rozloženia nezistí žiadne alebo
len niekoľko riadkov)

Nepresné základné čiary

2) Zmeňte pokročilé nastavenia (advanced settings)



Minimálna dĺžka základnej čiary

(Minimal baseline length): Minimálna dĺžka riadkov **v pixeloch** (pre tabuľky je lepšie nastaviť ho na hodnotu Nízka)

Prah presnosti základnej čiary (Baseline Accuracy threshold):

Stredné a nízke poskytujú lepšie výsledky

Použitie trénovaných separátorov (Use trained separators) Ak zvýšite túto hodnotu, okolité čiary sa zvyčajne zlučujú

Max vzdialenosť pre spojenie základných čiar (Distance for merging baselines):

Low: Zlúčia sa iba najbližšie čiary

Medium

High: vzdialené základné čiary sa zlúčia

Nepresné základné čiary

2) Zmeňte pokročilé nastavenia (advanced settings)

Baseline Options

Minimal baseline length

Medium 25

Baseline Accuracy threshold

Medium 55

Use trained separators

No -1

Max-dist for merging baselines

Medium

Split lines on region border

Line overlap fraction

1%

Zlúčiť čiary v rámci bloku (Split lines on region border)

Iba ak zachováte existujúce bloky textu:

Delené čiary na hranici regiónu:

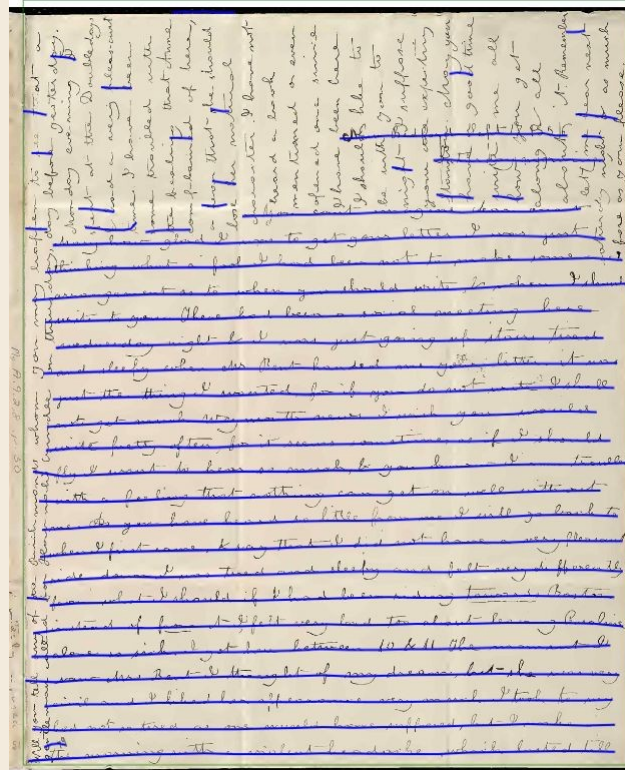
Aby čiary striktno dodržiavali hranicu regiónu.

Dôležité pre tabuľky!

Nepresné základné čiary

Example 1

Example 2



Nepresné základné čiary

3) Ak vám verejné modely a rozšírené nastavenia neposkytnú dobrý výsledok, tak:

Trénujte Model pre základné čiary (Baselines model) vášho špecifického dokumentu

Všetky stránky musia mať podobné rozloženie!

| MURRAY, MARGARET D. | | |
|---|--|---------|
| 10/20/08 | Marks Received on Examination. | Jacket |
| 10/20/08 | Recommendation of Exam. Board. | " |
| 10/28/08 | Authority of Sec. to appoint. | B 14 |
| 10/28/08 | Appointed. Reported 11/8/08 | Jacket. |
| 11/18/09 | Req. Trans. to Mare Island, Cal. | E 10 |
| 2/3/10 | Req. trans. to Wash. & Req. for Mare Island, Cal. withdrawn. | Jacket |
| 3/21/13 | Telegrams re- resignation. Miss Taylors | jacket |
| Bureau M. & S., Navy Department, Incc. 1 Jan. '11 | | |

| (2) MURRAY, MARGARET D. | | |
|--|--|---------|
| 3/20/13 | Tenders resignation. | Jacket. |
| 5/10/13 | Authority of Dept. to accept. | " |
| 5/18/13 | Resigned. (M.I.) | " |
| 4/14/15 | Miss Delano req. infor. (ans. 4/17/15. K 9 | |
| 3/24/14 | 3/R to Ruff | |
| 2826 Calvert St., Baltimore, Md. | | |
| 4/15/34 - 2101 Sh. Paul St. Balti. Md. | | |

Tréning modelu základných čiar (Baseline Model)

| MURRAY, MARGARET D. | | |
|--|---|---------|
| 10/20/08 | Marks Received on Examination. | Jacket |
| 10/20/08 | Recommendation of Exam. Board. | " |
| 10/28/08 | Authority of Sec. to appoint. | B 14 |
| 10/28/08 | Appointed. Reported 11/2/08 | Jacket. |
| 11/18/09 | Req. Trans. to Mare Island, Cal. | E 10 |
| 2/3/10 | Req. trans. to Wash. & Req. for Mare Island, Cal. withdrawn. | Jacket |
| 3/21/13 | Telegrams re- resignation. Miss Taylors | jacket |
| Bureau U. & S. Navy Department, 16,000. 1 Jan. '11 | | |

| (2) MURRAY, MARGARET D. | | |
|---|--|---------|
| 3/20/13 | Tenders resignation. | Jacket. |
| 5/10/13 | Authority of Dept. to accept. | " |
| 5/16/13 | Resigned. (M.I.) | " |
| 4/14/15 | Miss Delano req. infor. (ans. 4/17/15. K 9 | |
| 3/24/19 | <i>3/A to R-FF</i> | |
| 2826 Calvert St., Baltimore, Md. | | |
| <i>1/15/34 - 2101 St. Paul St. Balti. Md.</i> | | |

| MURRAY, MARGARET D. | | |
|--|---|---------|
| 10/20/08 | Marks Received on Examination. | Jacket |
| 10/20/08 | Recommendation of Exam. Board. | " |
| 10/28/08 | Authority of Sec. to appoint. | B 14 |
| 10/28/08 | Appointed. Reported 11/2/08 | Jacket. |
| 11/18/09 | Req. Trans. to Mare Island, Cal. | E 10 |
| 2/3/10 | Req. trans. to Wash. & Req. for Mare Island, Cal. withdrawn. | Jacket |
| 3/21/13 | Telegrams re- resignation. Miss Taylors | jacket |
| Bureau U. & S. Navy Department, 16,000. 1 Jan. '11 | | |

| (2) MURRAY, MARGARET D. | | |
|---|--|---------|
| 3/20/13 | Tenders resignation. | Jacket. |
| 5/10/13 | Authority of Dept. to accept. | " |
| 5/16/13 | Resigned. (M.I.) | " |
| 4/14/15 | Miss Delano req. infor. (ans. 4/17/15. K 9 | |
| 3/24/19 | <i>3/A to R-FF</i> | |
| 2826 Calvert St., Baltimore, Md. | | |
| <i>1/15/34 - 2101 St. Paul St. Balti. Md.</i> | | |

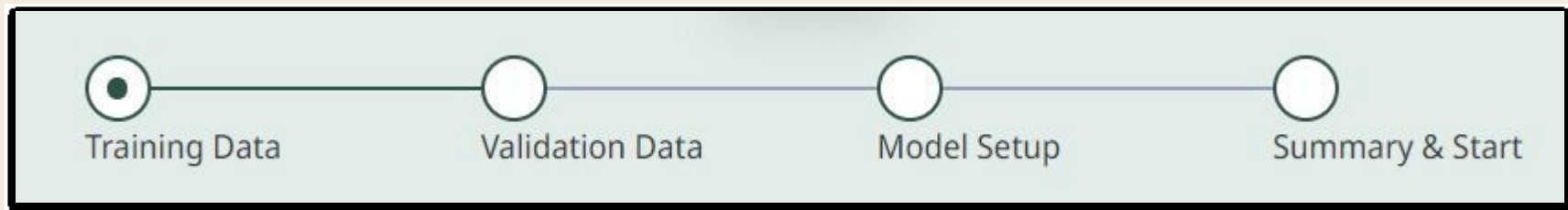
Tréning modelu základných čiar (Baseline Model)

Pripravte si aspoň 50 strán GT so správnymi základnými čiarami:

- Nakreslite všetky základné čiary manuálne alebo opravte automatické rozpoznávanie rozloženia
- Nakreslite základné čiary iba pre časti, ktoré chcete prepísať

Tréning modelu základných čiar (Baseline Model)

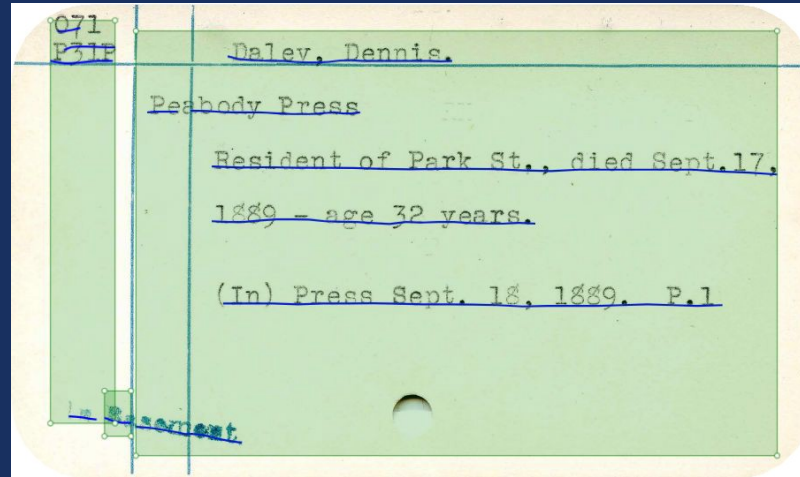
- Vyberte tréningové údaje (Training Data)
- Vyberte overovacie údaje (Validation Data)
- Nastavenie modelu (Model setup)
- Rozšírené nastavenia



Modely pre základné čiary (Baselines Models)

Po zaškolení môžete použiť svoj prispôsobený *Model pre základné čiary* (Baselines model) pre váš dokument! Zobrazí sa v zozname vašich súkromných Modelov rozloženia (Layout Models)

The screenshot displays the Text Recognition interface. At the top, there is a 'Text Recognition' header with a 'Layout' tab selected. Below this, a search bar contains 'NL-RISA_199_226'. To the right, there is a 'Start Recognition' button and credit information: 'Credits needed: 0.00' and 'Available: 0.00'. The main area is divided into two sections. On the left, there is a sidebar with 'Favorite Models' (0), 'Public Models' (7), and 'Private Models' (3), with 'Private Models' highlighted. Below the sidebar is a 'Filter' section with a search bar and 'Private Type' options: 'Own' (selected) and 'Shared'. The main list shows three models: 'v3', 'v2', and 'v1'. The 'v2' model is highlighted. Below the list, there are two entries: 'Medieval manuscript with glosses' with 2,366 words and 'Notes and miscellaneous materiel' with 8,468 words. On the right, a detailed view of the 'Private Model' 'v2' is shown, including the user 's.mansutti@readcoop.eu', the date '19/12/2022', and the 'CER (Accuracy)' of '5.19%', which is circled in red. Other details include 'Languages', 'Training Set Size', and 'Trained on' (handwritten).

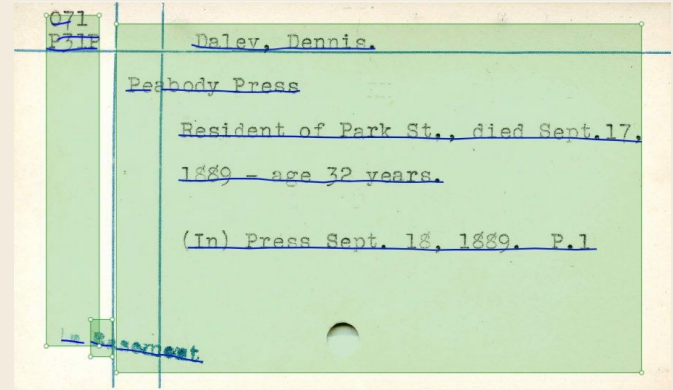
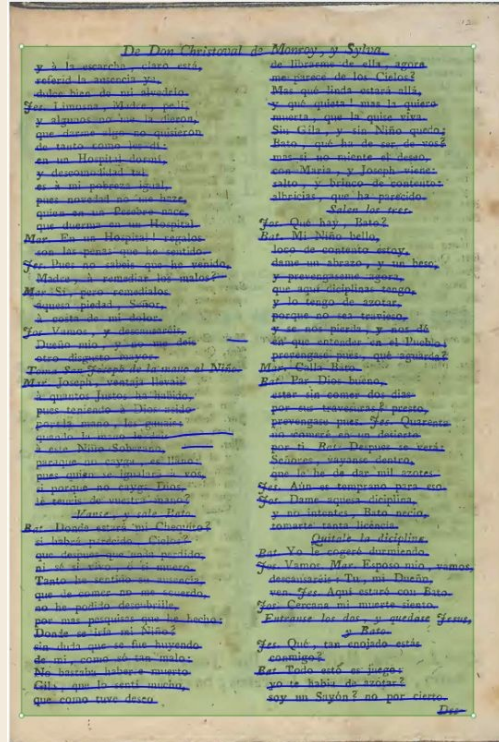


Nepresné bloky textu

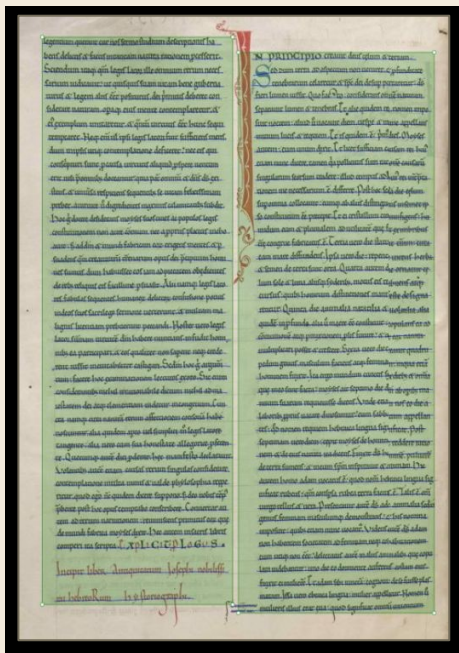
Nepresné bloky textu

Example 1

Example 2



Analýzy rozloženie/segmentácia (rozpoznanie textu)



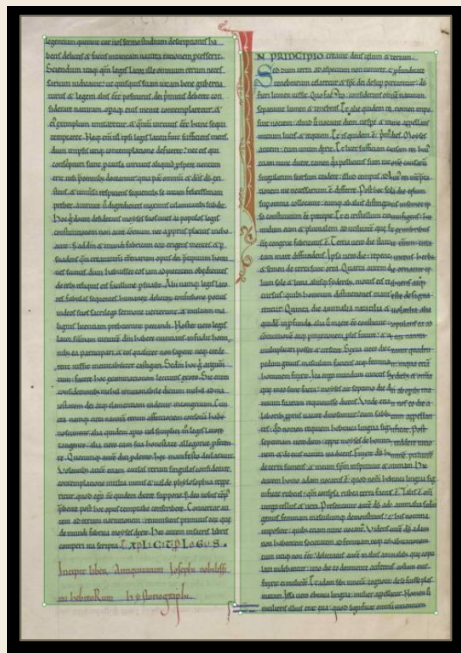
Bloky textu:

Prístup zdola nahor

(s predvoleným rozpoznávaním textu a rozloženia):

1. Rozpoznanie základných čiar
2. Agregácia východiskových hodnôt v textových oblastiach na základe ich súradníc
3. Základné čiary a polygóny sa tvoria v momente rozpoznávania textu (Text Recognition)

Analýzy rozloženie/segmentácia (rozpoznanie textu)



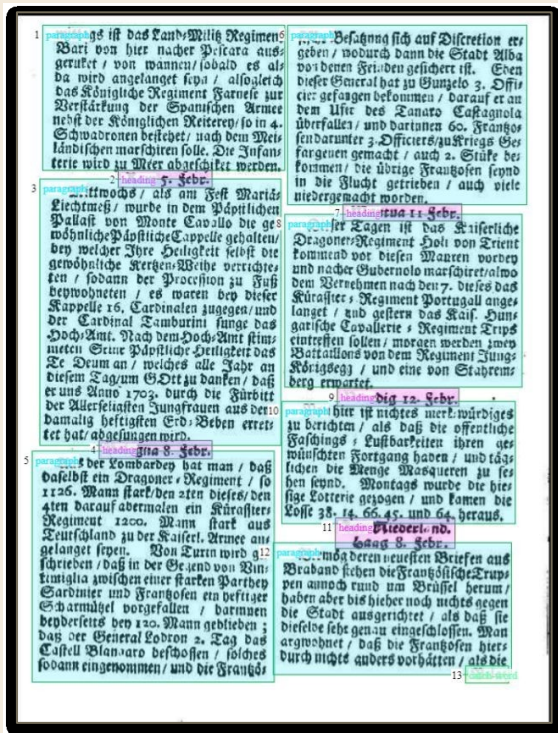
Bloky textu:

Prístup zdola nahor

V tomto prístupe môžete upraviť iba nastavenia:

1. Metóda oblasti textu
2. Orientácia základnej čiary textu

Analýzy rozloženie/segmentácia (rozpoznanie textu)



Bloky textu:

Prístup zhora nadol

1. Rozpoznávanie blokov textu pomocou **Modelu poľa (Field Model)**: *polia sú v blokoch stránky*
2. Rozpoznanie základných čiar (Layout Recognition)
 1. *Základné čiary a polygóny sa tvoria v momente rozpoznávania textu (Text Recognition)*

Vormerkblatt

Name: **Name:** H u r t h, Oberstlttn.

Geburtsjahr u. Ort: **Year:** 1884 **Place:** au

Heimatzuständigkeitsort **Place:** au, Sudetenland

$\left. \begin{array}{l} \text{vor} \\ \text{nach} \end{array} \right\}$ dem Umsturz 1918:

Assentjahr: 1903

Modely pol'a

Modely pol'a (Beta)

| | | | | | |
|--|---------|-------------------------------|---------------------------------------|-----------------|------------------|
| Haupt-Grundbuchheft (Offentjahrgang) | | 1904... | Blatt-Nr. | 625 | |
| Vor- und Zuname | | Johann Klüpfel <i>Klupfau</i> | | | |
| Geburts- | Ort | <i>Innsbrück</i> | Heimatsberechtigt in | Orts-gemeinde | <i>Innsbrück</i> |
| | Bezirk | <i>Innsbrück</i> | | Bezirk | <i>Innsbrück</i> |
| | Comitat | <i>⁄</i> | | Comitat | <i>⁄</i> |
| | Land | <i>Tirol</i> | | Land | <i>Tirol</i> |
| | | | Geburts-jahr | 18.83 | |
| | | | Religion | <i>kathol.</i> | |
| | | | Kunst, Gewerbe, sonstiger Lebensberuf | <i>Leinwand</i> | |
| seit 1904 nach der Losreihe auf drei Jahre in der Reserve und zwei Jahre in der Landwehr, zum 3. Aug. d. Tirol. Kav. Regt. | | | | | |

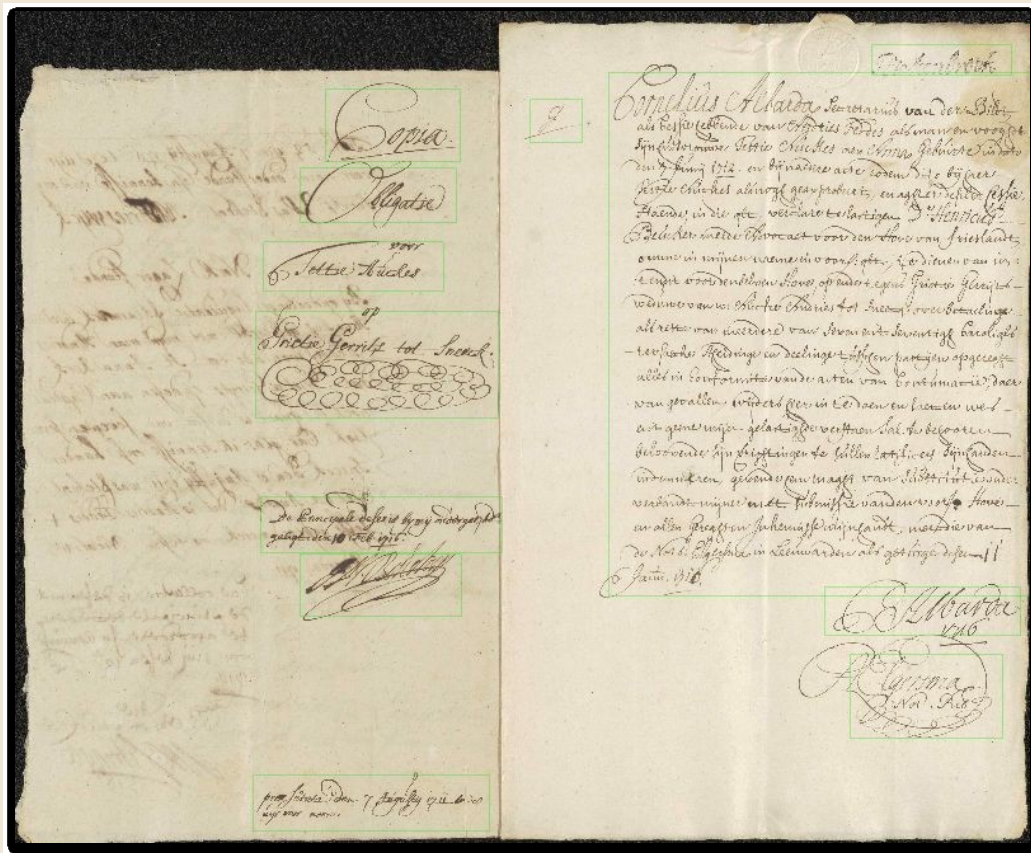
Modely poľa (Beta)

Modely poľa je možné trénovať na:
**automatické rozpoznávanie a
označovanie určitých prvkov (dát)
rozloženia dokumentu.**

- Bloky textu - Textové oblasti (polia)
- Priradenie značiek štruktúry pre tieto oblasti

| | | | | | | | | |
|--|---------|--------------------------------|----------------------|---------------|-----------|---------------------------------------|----------|-----------|
| Haupt-Grundbuchheft (Offentjahrgang) | | 1904... | Blatt-Nr. | | 625 | | | |
| Vor- und Zuname | | Name: Johann Hüpfauer Hüpfauer | | | | | | |
| Geburts- | Ort | Innsbrück | Heimatsberechtigt in | Orts-gemeinde | Innsbrück | Geburts-jahr | Jahrgang | 1883 |
| | Bezirk | Innsbrück | | Bezirk | Innsbrück | | Religion | kathol. |
| | Comitat | ✓ | | Comitat | ✓ | Kunst, Gewerbe, sonstiger Lebensberuf | | Lieferant |
| | Land | Tirol | | Land | Tirol | | | |
| Juli 1904 nach der Losreihe auf drei Jahre in der en Jahre in der Reserve und zwei Jahre in der Landwehr, zum 3. Aug. d. Tirol. Milit. 3. Quart. | | | | | | | | |

Blogy textu (Text regions)



Noviny: Segmentácia rozloženia



Segmentácia formulára

all

Vater: vater separiert 100% 1951

Mutter: mutter separiert 100% 3 D

Staatsangehörigkeit: Staatsangehörigkeit 99%

Personalakt.:

Familienname: name 100%

Vornamen: vorname 99%

Geburts-tag: datum 100% Geburtsort: ort 99%

Glaubensbek.: religion 100% Kreis: rov.

Beruf: 1. Beruf 100% 3.
4. 5. 6.

Mitglied und Seiltug
i. d. NSDAP
oder einer ihrer
Gliederungen

| Familienangehörige | Geburts- tag | mo- nat | jahr | Geburtsort (Kreis, Provinz) Standesamt | Glaubens- bek. | Aus- zugs- verm. | Mitglied und Seiltug i. d. NSDAP oder einer ihrer Gliederungen | Vermerke |
|--------------------|-----------------|------------|------|--|-------------------|------------------------|--|----------|
| Kinder: | | | | | | | | |
| | | | | | | | | |

Verheiratet seit verheiratetvu0020seit 99% Standesamt Standesamt 97%

Standesamtsnummer 99% dem verheiratet mit 95%

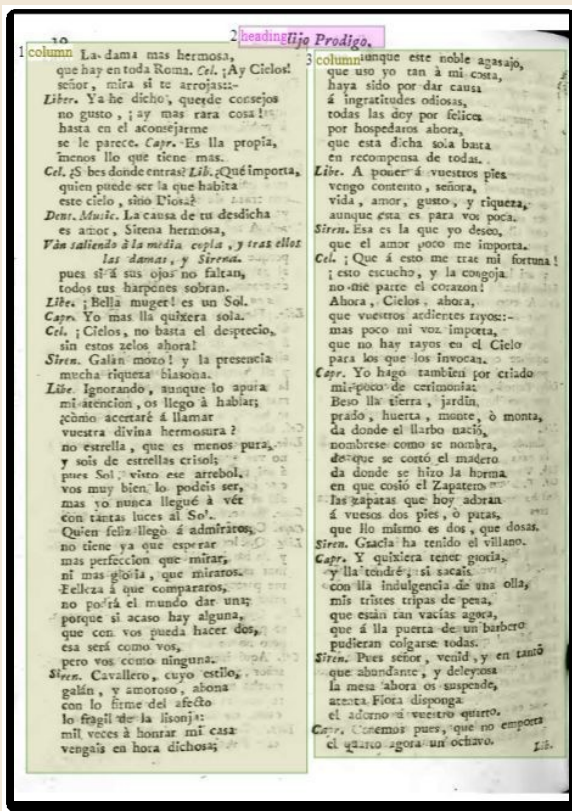
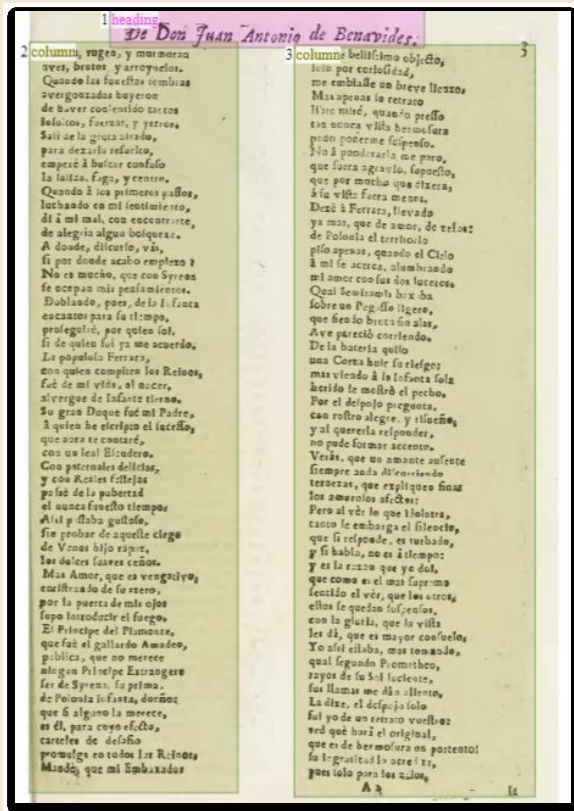
geb. verheiratet mit geboren am 97%

in verheiratet mit geboren in 100%

Wohnung: Wohnung 98%

Verdruck Nr. 304b (Wahlbild unvollst.) 9.48 380 000 In 150 Nr. 945 Staatsdruckerei Berlin 2706

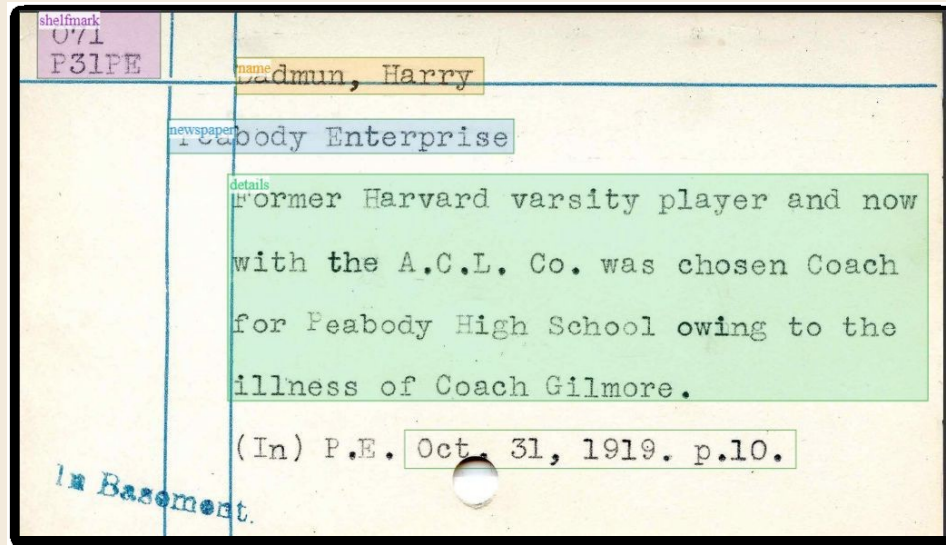
Viac stĺpcové rozloženie textu



Modely poľa (Beta)

Pripravte si cca 50 strán tréningových dát:

- Nakreslite textovú oblasť okolo relevantných informácií, ktoré chcete extrahovať
- Prirad'ujte štrukturálne značky (voliteľné)



Modely poľa (Beta)



 Desk

 Models

 Sites

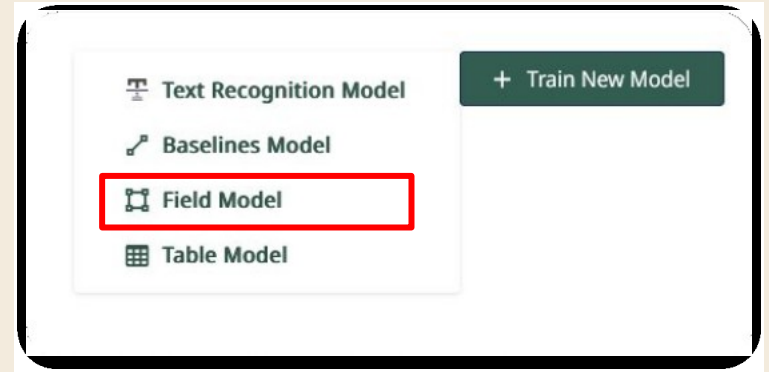
Jobs



Modely (Models) Transkribus je miesto, kde môžete trénovať a spravovať svoje modely.

Modely poľa (Beta)

- Tréningové údaje (Training Data)
- Výber značky (tagov) (Tag Selection)
- Overovacie údaje (Validation Data)
- Nastavenie modelu (Model Setup)
- Rozšírené nastavenia (Cykly tréningu a miera učenia)



Spracovanie dokumentov s poľami

- 1) **Vytvorenie Ground Truth pre rozpoznávanie polí:**
 - minimálne 50 strán
 - Viac strán so zložitým rozložením

Spracovanie dokumentov s poľami

- 1) Vytvorenie Ground Truth pre rozpoznávanie polí
- 2) **Trénovanie modelu rozpoznávania polí**

Spracovanie dokumentov s poľami

- 1) Vytvorenie Ground Truth pre rozpoznávanie polí
- 2) Trénovanie modelu rozpoznávania polí
- 3) **Použitie modelu rozpoznávania polí na zostávajúce strany**

Spracovanie dokumentov s poľami

- 1) Vytvorenie Ground Truth pre rozpoznávanie polí
- 2) Trénovanie modelu rozpoznávania polí
- 3) Použitie modelu rozpoznávania polí na zostávajúce strany
- 4) **Spustenie rozpoznávania rozloženia na detekciu čiar:**

Nastavenia:

- **Model základnej čiary (Baseline model):** Horizontal/Mixed Text Line Orientation/Model trained by you
- **Zachovanie existujúcich blokov - oblastí textu** (môže pomôcť) Minimálna dĺžka základnej čiary: (low) nízka
- **Rozdelené čiary na hranici regiónu**

Spracovanie dokumentov s poľami

- 1) Vytvorenie Ground Truth pre rozpoznávanie polí
- 2) Trénovanie modelu rozpoznávania polí
- 3) Použitie modelu rozpoznávania polí na zostávajúce strany
- 4) Spustenie rozpoznávania rozloženia na detekciu čiar
- 5) Rozpoznávanie textu**
- 6) Verejný model / Privátny model, ktorý ste vyškolili, → možnosť aplikovať rôzne modely v rôznych oblastiach

Spracovanie dokumentov s poľami

- 1) Vytvorenie Ground Truth pre rozpoznávanie polí
- 2) Trénovanie modelu rozpoznávania polí
- 3) Použitie modelu rozpoznávania polí na zostávajúce strany
- 4) Spustenie rozpoznávania rozloženia na detekciu čiar
- 5) Rozpoznávanie textu
Verejný model / Vami vytrénovaný model
- 6) Korekcie (optional)
- 7) **Export**

Spracovanie dokumentov s poľami

| | A | B | C | D | E | F |
|---|---------------------|-------------|-------------------|--------------------|--|--------------------|
| 1 | TranskribusFilename | shelfmark | name | newspaper | details | reference |
| 2 | 00729.jpg | 071 D218 | Dynan, Mary E. | Salem Evening News | Received diploma in October, 1910 from N.E. Institute of Anatomy, Sanitary Science and Embalming. Was first Peabody girl to graduate as an embalmer, also the youngest in the state. | Oct. 10, 1910. P.5 |
| 3 | 00730.jpg | 071 | Dynan, Mary E. | Salem Evening News | Of 17 Franklin St. was granted an Undertaker's license from the Board of Health. She passed a successful examination in embalming before the State Board and was the first woman in town to be granted such a license. | June 10, 1911. P.5 |
| 4 | 00731.jpg | | Dynan, Timothy I. | Salem Evening News | Who died at his home, 30 Chestnut St. was a baker by trade and an active member of organized labor. He was employed by Jackson and Tortat until he met with an accident 5 years ago. | July 22, 1920. P.7 |
| 5 | 00732.jpg | | Dynamite | Salem Evening News | Left unguarded, boys wander into magazine of the Essex Trap rock where enough is stored to blow the city to pieces. They take two sticks with them. | June 21, 1918. P.2 |

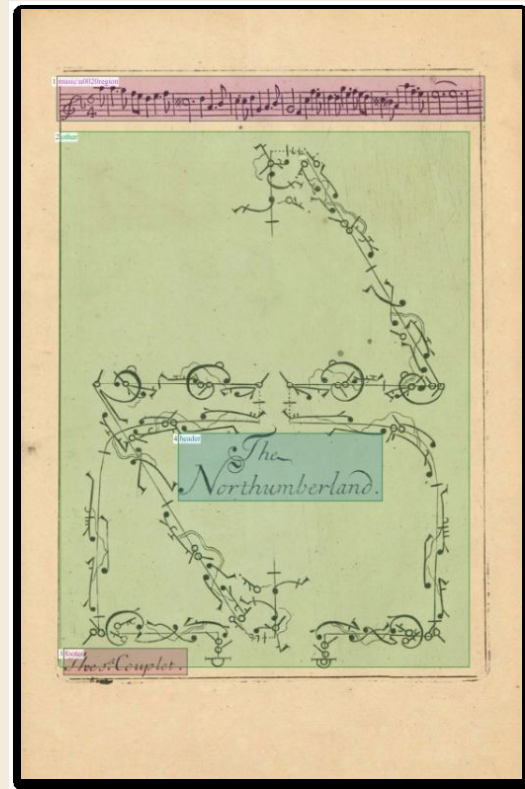
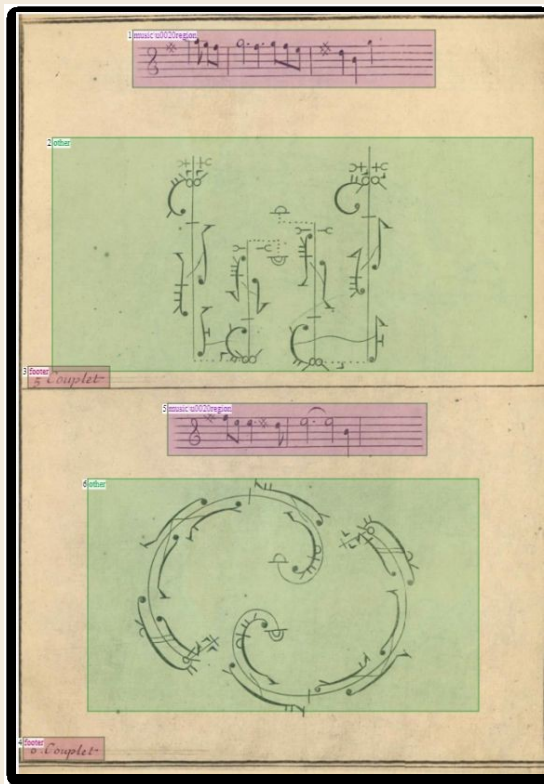
Príklady Modelov polí

Ground Truth:

30 strán

5 tagov

[Example](#)



Príklady Modelov polí

Example

1 **Header**
2 **Anno 1746.** (Num. 17.) 26. Februarii.
3 **Wienersches DIARIUM.**
4 **Imprimt** Ihrer K. k. Kaiserl. auch zu Singarn, und Wáheim K. k. Maj. Streybeten
5 **In dem neuen Michaeler-Haus/ bey Joh. Peter v. Spelen.**
6 **beobachtet Italien.**
7 **Venna 29. Jan.**
8 **Ergangenen Samstag** kamen abermalen zwey Catalonische Schiffe mit 3600. Spanischen Soldaten von Villa franca/und Sonnens tags Abends eine unferige Fehlfte von Calvi alhier an / von wannen sie den 19. mit Staats- Briefen ab / und nach der Lwoorno gegangen / und da sie auch von diesen letzteren gleich wieder abgegangen / hat selbe den 21. frúhe in diesem Gegebenen ein Engländisches Krieges-Schiff mit vier ektigen Flaagen sich anken gesehen. Briefe von Barcello unter dem 15. dieses geben/ daß dieselb nach und nach immerhin Drucken und Fahr- Züge antommen / weßche jetzerzeit gar bald nach hiesigen Gegenden befördert werden : es sollen auch wúrklich einige Cavallerie- Regimente zur Verstárkung der Spanischen Armee alßchon nach unsern Gránken im wúrklichen Anmarck seyn. Seit einigen Tagen hat man angefangen die von Savonia nebst vielen anderen Kriegs-geráthe hieher gekommenen Artillerie zu der Spanischen Armee nach der Kommanden abzuschicken. So seynd auch Dienstags 4. Sambdichn / Schiffe/ und 5. andere Neapolitanische Fahr-zeuge in 11. Tagen von Gaetra mit Artillerie und anderen Kriegs-Vorrat hiet an-

9 **zukommen** ; Auch seynd auf einer Franzósischer Tartane 150. Mann von dem Franzósischen Regiment Lothringen von Villa franca hier angelangt / von welchem Ort selbe zugleich mit 3. andern mit dertey Truppen beladenen hiet her bestimmten Schiffen abgefeglet. Eine Kroensische Pinke / so in 12. Tagen von Masilien gekommen / hat die vornehmsten Officiers des Regimentes Modena alhier an das Land gesehet. Zwischen Dimerikog und gestern kam men 3. Catalonische Schiffe mit 4000. Recruten für verschiedene Spanische Regimenter nebst ihren Officieren und mit 1500. Stúck-Kugeln alhier an.

10 **Woch Gelegenheit** da dieser Tag beyde Kónigliche Majestáten zu Vortice zu Mittag geseisset / wurde unter diesen dem Russischen Reichs / Russes Konsten Grafen von Woronzow nebst seiner Frauen Gemáhlin der gantz hiesige Kónigl. Pallast / samt allen Kleinodien / wie auch denen Kóniglichen Vermáhlungs-Kutschen / und alle úbrige kostbare Einrichtung gezeiget. Vergangene Woche ist eine hiesige wol- ausgerústete Volleotte mit 14000. Ducaten zu Bezahlung unserer Truppen in der Lombardey von hier nachter Venna abgehóret worden. Gestern Vort-

1 **besetzt** ist das Land-Milít Regiment Bari von hier nachter Picara ausgezúcket / von wannen / sobald es als da wird angelangt seyn / alßgleich das Kónigliche Regiment Farnese zur Verstárkung der Spanischen Armee nebst der Kóniglichen Reiteren / so in 4. Schwadronen beisset / nach dem Weislandischen marschiren solle. Die Infanterie wird zu Viter abarschicket werden.

2 **besetzt** / als am Fest Mariæ Liechtmess / wurde in dem Päpstlichen Pallast von Monte Cavallo die gewöhnliche Päpstliche Capelle gehalten / bey welcher Ihre Heiligkeit selbst die gewöhnliche Kirggen-Weide verrichteten / sodann der Proceßion zu Fuß beynaheten / es waren bey dieser Kapelle 16. Cardinale zugegen / und der Cardinal Tamburini sung das Hoch-Miss. Nach dem Hoch-Miss stimmten Seine Päpstliche Heiligkeit das Te Drum an / welches alle Tage an diesem Tagum G. Det zu danken / daß er uns Anno 1702. durch die Fürbitte der Allerheiligsten Jungfrauen aus dem damalig heftigsten Erd-Wehen errettet hat / abgefungen wird.

3 **besetzt** der Kommander hat man / daß daselbst ein Dragoner- Regiment / so 126. Mann stark/ den 2ten dieses/ den 1ten darauf abermalen ein Kürassiers Regiment 1200. Mann stark aus Teutschland zu der Kaiserl. Armee angelangt segen. Von Turin wird geschrieben / daß in der Stadt und Umgránda zwischen einer starken Partbey Sardiner und Franzosen ein heftiger Scharmútel vorzúfallen / darinnen beyderseits der 120. Mann geblieben ; das der General Lodron 2. Tag das Castell Pianaro beschoßen / solches sodann eingenommen / und die Franzó-

4 **besetzt** sich auf Discretion ergeben / wodurch dann die Stadt Alba von denen Franzen gesichert ist. Eben dieser General hat zu Gungelo 3. Officier gefangen bekommen / darauf er an dem Ufer des Tanato Casagnola úberfallen / und darinnen 60. Franzosen darunter 3. Officiers / zu Kriegs Gefangen gemacht / auch 2. Stúpe der kommen / die úbrige Franzosen seynd in die Flucht getrieben / auch viele niedergemacht worden.

5 **besetzt** / als am Fest Mariæ Liechtmess / wurde in dem Päpstlichen Pallast von Monte Cavallo die gewöhnliche Päpstliche Capelle gehalten / bey welcher Ihre Heiligkeit selbst die gewöhnliche Kirggen-Weide verrichteten / sodann der Proceßion zu Fuß beynaheten / es waren bey dieser Kapelle 16. Cardinale zugegen / und der Cardinal Tamburini sung das Hoch-Miss. Nach dem Hoch-Miss stimmten Seine Päpstliche Heiligkeit das Te Drum an / welches alle Tage an diesem Tagum G. Det zu danken / daß er uns Anno 1702. durch die Fürbitte der Allerheiligsten Jungfrauen aus dem damalig heftigsten Erd-Wehen errettet hat / abgefungen wird.

6 **besetzt** hiet ist nichts merk-würdiges zu berichten / als daß die öffentliche Fassung / Luftbarkeiten ihren gewúnschten Fortgang haoren / und täglich die Menge Wassuren zu sehen seynd. Montags wurde die hiesige Lotterie gezogen / und kamen die Kasse 38. 14. 66. 45. und 64. heraus.

7 **besetzt** / als am Fest Mariæ Liechtmess / wurde in dem Päpstlichen Pallast von Monte Cavallo die gewöhnliche Päpstliche Capelle gehalten / bey welcher Ihre Heiligkeit selbst die gewöhnliche Kirggen-Weide verrichteten / sodann der Proceßion zu Fuß beynaheten / es waren bey dieser Kapelle 16. Cardinale zugegen / und der Cardinal Tamburini sung das Hoch-Miss. Nach dem Hoch-Miss stimmten Seine Päpstliche Heiligkeit das Te Drum an / welches alle Tage an diesem Tagum G. Det zu danken / daß er uns Anno 1702. durch die Fürbitte der Allerheiligsten Jungfrauen aus dem damalig heftigsten Erd-Wehen errettet hat / abgefungen wird.

8 **besetzt** / als am Fest Mariæ Liechtmess / wurde in dem Päpstlichen Pallast von Monte Cavallo die gewöhnliche Päpstliche Capelle gehalten / bey welcher Ihre Heiligkeit selbst die gewöhnliche Kirggen-Weide verrichteten / sodann der Proceßion zu Fuß beynaheten / es waren bey dieser Kapelle 16. Cardinale zugegen / und der Cardinal Tamburini sung das Hoch-Miss. Nach dem Hoch-Miss stimmten Seine Päpstliche Heiligkeit das Te Drum an / welches alle Tage an diesem Tagum G. Det zu danken / daß er uns Anno 1702. durch die Fürbitte der Allerheiligsten Jungfrauen aus dem damalig heftigsten Erd-Wehen errettet hat / abgefungen wird.

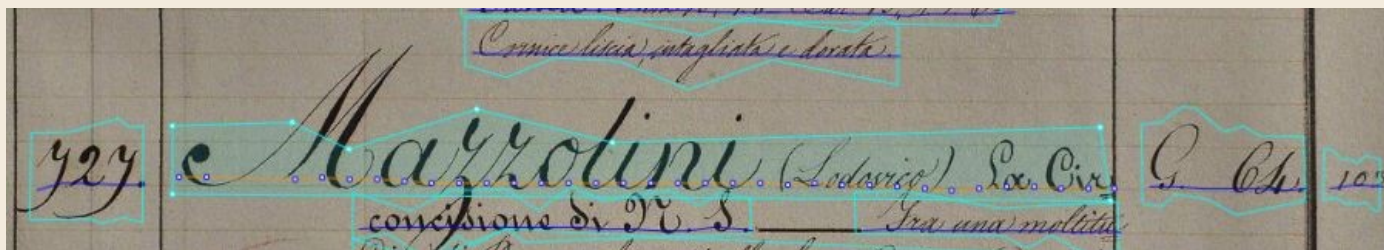


Nepresné polygóny (mnohouholníky)

Inaccurate Polygons

[Example 1](#)

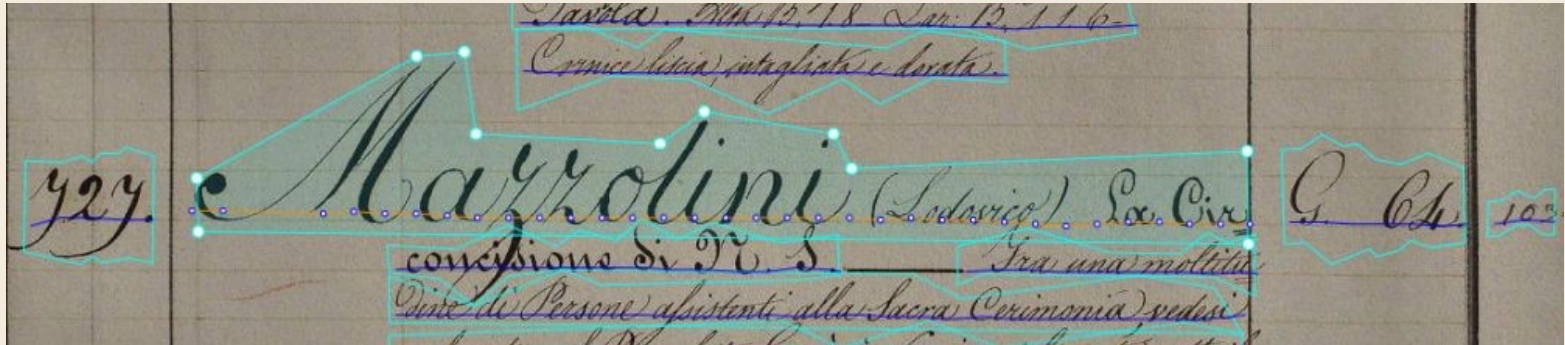
[Example 2](#)



Field Model trained on Line Polygons

Prepare about 50 pages of training data:

- Adjust the line polygons manually



Field Model trained on Line Polygons

- Training data
- Tag selection: TRAIN ON LINE POLYGONS
- Validation data
- Model setting
- Advanced settings (Training Cycles and Learning Rate)

Field Recognition Model

Training Data Tag Selection Validation Data Model Setup Start

| Remove | Title | Example polygons |
|--------|-------|------------------|
| X | | Example polygons |

< Back

Next >

1 documents selected

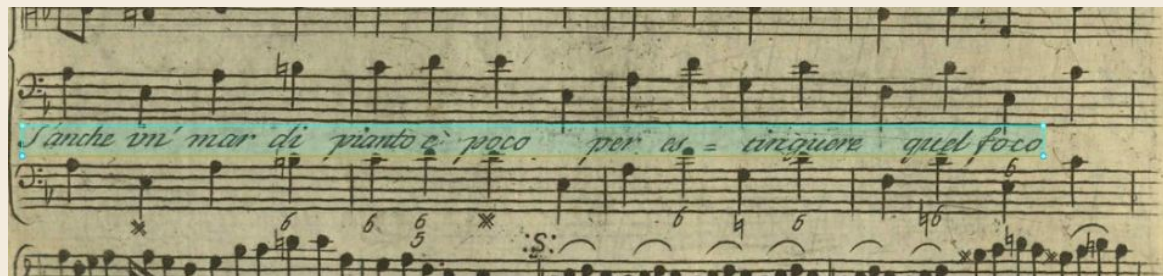
Recognise untagged regions
Select if you want include untagged regions in your training.

Train on line polygons
Instead of training on tags, your model will be trained on line polygons.

Field Model trained on Line Polygons



| | Region 1 |
|----------|----------|
| 1 | 44 |
| 2 | - |
| 3 | 1 |
| 4 | - |
| 5 | error |
| 6 | - |
| 7 | 2 |
| 8 | --- |
| Region 2 | |
| 1 | 27 |



| | Region 1 |
|----------|--|
| 1 | 44 |
| Region 2 | |
| 1 | 21 |
| Region 3 | |
| 1 | non sperare di smorzare col tuo pianto l'ira mi- |
| Region 4 | |
| 1 | -a |
| Region 5 | |
| 1 | s'anche in' mar di pianto poco per es-tinguere quel foco |
| Region 6 | |
| 1 | ch'arde gel di gelo-si-a per estinguere quel fo-co |
| Region 7 | |
| 1 | ch'arde al gel di gelo-si-a |
| Region 8 | |
| 1 | Da Capo |

APPLICATIONS FOR EDUCATIONAL INSTITUTIONS 1

| NAME OF INSTITUTION | TOWN | AMOUNT | OBJECT | DATE | DISPOSITION |
|------------------------------------|----------------------|--------|------------------------------|-----------|--------------------------------------|
| Acadia University (1) | F Wolfville, N.S. | | General | 1/9/11 | Has done share, 4/23/12 (amo. 2) |
| American Institute | Americus, Ga. | 10000 | Building | 1/17/11 | D + Low |
| Albemarle Normal & Ind. Inst. | Albemarle, N. C. | | General | 2/4/11 | Low |
| Asbury College | Wilmore, Ky. | | Buildings & Industrial Plant | 2/13/11 | D |
| Adrian College | P Adrian, Mich. | | Endowment | 3/13/11 | Denominational, 3/6/11 |
| Alabama State Normal School | Florence, Ala. | 25000 | Building | 3/10/11 | State institution, 2/15/11 |
| Antioch College | Yellow Springs, O. | 100000 | Endowment | 3/23/11 | Not sufficiently developed, 2/17/11 |
| American Church, Inst. for Negroes | New York, N. Y. | | General | 4/10/11 | D |
| Amherst College | P Amherst, Mass. | 50000 | Increase Salaries | F 5/10/11 | Has done share, 5/19/11 |
| Alma College | P Alma, Mich. | | Library Building | 5/18/11 | Denominational, 3/24/11 |
| American International College | Springfield, Mass. | | General | 1/10/11 | Not sufficiently developed, 2/10/11 |
| Alberta Ladies' College (1) | Red Deer, Alta. | | General | 12/15/11 | Not sufficiently developed, 1/14/11 |
| Allen University | Columbia, S.C. | | Library Building | 4/26/12 | Low |
| Acadia University (2) | F Wolfville, N. S. | 25000 | Library Building | 5/10/12 | Has done share, 4/23/12; D |
| Abingdon Presbytery | Orange, Va. | 500 | Building | 5/14/12 | D + Low |
| Amity College | College Springs, Ia. | | Endowment | 5/13/12 | Not sufficiently developed, 10/15/11 |
| Albert Lea College | P Albert Lea, Minn. | | Endowment | 6/18/12 | Denominational, 1/12/11 |
| Anderson College (1) | Anderson, S. C. | | General | 8/10/12 | D + Low |
| Alabama University of | University, Ala. | | Memorial Building | 10/10/12 | State institution, 1/15/11 |

Table Models

Modely pre tabuľky (Beta)

97


| APPLICATIONS FOR EDUCATIONAL INSTITUTIONS | | | | | |
|---|--------------------------|--------|--------------------------------------|--------|---|
| NAME OF INSTITUTION | TOWN | AMOUNT | OBJECT | DATE | DISPOSITION |
| Lutheran Ladies' Seminary | Reading, Minn. | | Convent | 1/10/0 | Not sufficiently developed, 2/10/07 |
| Lombard College | P. Yaleburg, Ill. | 50 000 | Science Building | 2/1/0 | D |
| Loyan Female College | Russellville, Ky. | 15 000 | Building and Equipment | 2/1/06 | Low |
| Livingston College (1) | P. Salisbury, N.C. | | Dormitory | 2/1/0 | D |
| Linden Hall Seminary | Leitch, Pa. | 50 000 | Library and Science Building | 2/1/0 | Seminary, 12/12/10 |
| Livingston College (2) | P. Salisbury, N.C. | | Land | 2/1/0 | D |
| Laurinburg Norm. & Ind. Inst. | Laurinburg, N.C. | | General | 2/1/0 | Normal |
| Lenoir College (1) | Hickory, N.C. | 70 000 | Science Bldg., Gymnasium, Auditorium | 2/1/0 | Not sufficiently developed, 2/10/06 |
| La Grange College (1) | La Grange, Mo. | | Endowment and Buildings | 2/1/0 | Not sufficiently developed, 2/10/06 |
| Lexington College | Lexington, Mo. | | Endowment | 2/1/06 | Not sufficiently developed, 2/10/06 |
| Lafayette College (1) | P. Easton, Pa. | | Engineering Bldg. & Endowment | 2/1/0 | Has done share, 2/1/0 |
| Lehigh College | P. Hopkinton, Iowa | 12 500 | Dept. of Agriculture | 1/1/0 | Has done share, 1/10/07, D |
| Lincoln Memorial University (1) | P. Cumberland Gap, Tenn. | | Building | 1/10/0 | Has done share, 1/10/0 |
| Lincoln Institute | Jefferson, Ct. Mo. | | Library Building | 2/1/06 | Low |
| Lutheran College (projected) | Seguin, Tex. | | Building | 2/1/0 | D |
| Leeds Industrial School | Lewisburg, Pa. | | Building | 2/1/06 | Low |
| La Grange College (2) | La Grange, Mo. | | Library Building | 2/1/06 | Not sufficiently developed, 2/10/06 (see 1) |
| Lindenwood College for Women | P. St. Charles, Mo. | 10 000 | Building | 2/1/06 | D |

97

| APPLICATIONS FOR EDUCATIONAL INSTITUTIONS | | | | | |
|---|--------------------------|--------|--------------------------------------|--------|---|
| NAME OF INSTITUTION | TOWN | AMOUNT | OBJECT | DATE | DISPOSITION |
| Lutheran Ladies' Seminary | Reading, Minn. | | Convent | 1/10/0 | Not sufficiently developed, 2/10/07 |
| Lombard College | P. Yaleburg, Ill. | 50 000 | Science Building | 2/1/0 | D |
| Loyan Female College | Russellville, Ky. | 15 000 | Building and Equipment | 2/1/06 | Low |
| Livingston College (1) | P. Salisbury, N.C. | | Dormitory | 2/1/0 | D |
| Linden Hall Seminary | Leitch, Pa. | 50 000 | Library and Science Building | 2/1/0 | Seminary, 12/12/10 |
| Livingston College (2) | P. Salisbury, N.C. | | Land | 2/1/0 | D |
| Laurinburg Norm. & Ind. Inst. | Laurinburg, N.C. | | General | 2/1/0 | Normal |
| Lenoir College (1) | Hickory, N.C. | 70 000 | Science Bldg., Gymnasium, Auditorium | 2/1/0 | Not sufficiently developed, 2/10/06 |
| La Grange College (1) | La Grange, Mo. | | Endowment and Buildings | 2/1/0 | Not sufficiently developed, 2/10/06 |
| Lexington College | Lexington, Mo. | | Endowment | 2/1/06 | Not sufficiently developed, 2/10/06 |
| Lafayette College (1) | P. Easton, Pa. | | Engineering Bldg. & Endowment | 2/1/0 | Has done share, 2/1/0 |
| Lehigh College | P. Hopkinton, Iowa | 12 500 | Dept. of Agriculture | 1/1/0 | Has done share, 1/10/07, D |
| Lincoln Memorial University (1) | P. Cumberland Gap, Tenn. | | Building | 1/10/0 | Has done share, 1/10/0 |
| Lincoln Institute | Jefferson, Ct. Mo. | | Library Building | 2/1/06 | Low |
| Lutheran College (projected) | Seguin, Tex. | | Building | 2/1/0 | D |
| Leeds Industrial School | Lewisburg, Pa. | | Building | 2/1/06 | Low |
| La Grange College (2) | La Grange, Mo. | | Library Building | 2/1/06 | Not sufficiently developed, 2/10/06 (see 1) |
| Lindenwood College for Women | P. St. Charles, Mo. | 10 000 | Building | 2/1/06 | D |

Modely tabuliek automaticky rozpoznávajú riadky a stĺpce a tým zlepšujú extrakciu a analýzu tabuľkových údajov.

Modely pre tabuľky(Beta)

- Modely sa učia rozpoznávať riadky, stĺpce alebo obe
 - Zatiaľ žiadne všeobecné modely, ale školenia pre konkrétne zbierky/dokumenty
 - Nie sú potrebné oddeľovače (separátory)
 - S dostatkom tréningových údajov dokáže model spracovať viacero typov tabuliek
- 

Modely pre tabuľky

Ground Truth tvorba v editore:

Tabuľka

- Stĺpce
- Riadky

| APPLICATIONS FOR EDUCATIONAL INSTITUTIONS | | | | | | 1 |
|---|-----------------------|--------|------------------------------|----------|--------------------------------------|---|
| NAME OF INSTITUTION | TOWN | AMOUNT | OBJECT | DATE | DISPOSITION | |
| Acadia University (1) | Woolfville, N.S. | | General | 1/9/11 | Has done share, 11/23/12 (ans 2) | |
| Americus Institute | Americus, Ga. | 10000 | Building | 11/7/11 | D + Low | |
| Albemarle Normal & Ind. Inst. | Albemarle, N.C. | | General | 2/1/11 | Low | |
| Asbury College | Wilmore, Ky. | | Buildings & Industrial Plant | 2/13/11 | D | |
| Adrian College | Adrian, Mich. | | Endowment | 3/13/11 | Denominational, 3/1/11 | |
| Alabama State Normal School | Montgomery, Ala. | 25000 | Building | 3/10/11 | State institution, 2/15/11 | |
| Antioch College | Hillsboro Springs, O. | 100000 | Endowment | 3/2/11 | Not sufficiently developed, 2/15/11 | |
| American Church Inst. for Negroes | New York, N.Y. | | General | 4/1/11 | D | |
| Amherst College | Amherst, Mass. | 50000 | Increase Salaries | F 5/1/11 | Has done share, 5/14/11 | |
| Alma College | Alma, Mich. | | Library Building | 5/18/11 | Denominational, 3/24/11 | |
| American International College | Springfield, Mass. | | General | 1/10/11 | Not sufficiently developed, 3/2/109 | |
| Alberta Ladies' College (1) | Red Deer, Alberta | | General | 12/15/11 | Not sufficiently developed, 1/14/11 | |
| Allen University | Columbia, S.C. | | Library Building | 4/26/12 | Low | |
| Acadia University (2) | Woolfville, N.S. | 25000 | Library Building | 5/15/12 | Has done share, 4/23/12; D | |
| Kingdon Presbytery | Stafford, Va. | 500 | Building | 5/10/12 | D + Low | |
| Amity College | College Springs, Ia. | | Endowment | 5/13/12 | Not sufficiently developed, 10/15/11 | |
| Albert Lea College | Pell City, Ala. | | Endowment | 6/18/12 | Denominational, 11/2/11 | |
| Anderson College (1) | Anderson, S.C. | | General | 8/10/12 | D + Low | |
| Alabama University of | University, Ala. | | Memorial Building | 10/20/12 | State institution, 1/15/109 | |

Modely pre tabuľky

Stránky GT:

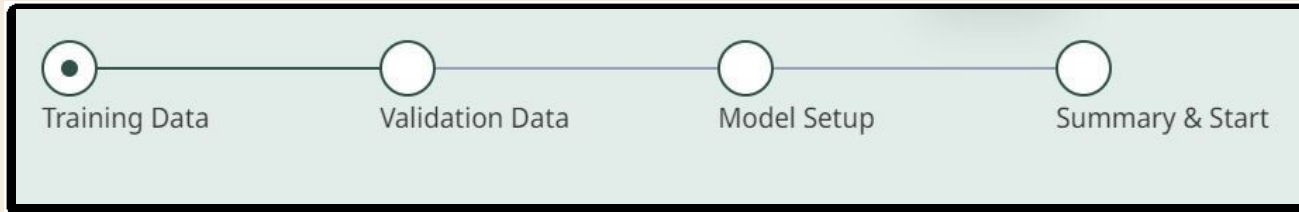
- **Jednoduché dabuľky:** 20 strán GT
- **Ťažké tabuľky:** 50 strán GT
- **mix rôznych tabuliek:** 50 až 100 strán GT v závislosti od počtu tabuliek

| APPLICATIONS FOR EDUCATIONAL INSTITUTIONS | | | | | | 1 |
|---|----------------------|--------|------------------------------|----------|--------------------------------------|---|
| NAME OF INSTITUTION | TOWN | AMOUNT | OBJECT | DATE | DISPOSITION | |
| Acadia University (1) | Woolfville, N.S. | | General | 1/9/11 | Has done share, 11/23/12 (ans. 2) | |
| Americus Institute | Americus, Ga. | 10000 | Building | 11/7/11 | D + Low | |
| Albemarle Normal & Ind. Inst. | Albemarle, N.C. | | General | 2/1/11 | Low | |
| Asbury College | Wilmore, Ky. | | Buildings & Industrial Plant | 2/13/11 | D | |
| Adrian College | Adrian, Mich. | | Endowment | 3/13/11 | Denominational, 3/1/11 | |
| Alabama State Normal School | Montgomery, Ala. | 25000 | Building | 3/10/11 | State institution, 2/15/11 | |
| Antioch College | Yellow Springs, O. | 100000 | Endowment | 3/2/11 | Not sufficiently developed, 2/12/11 | |
| American Church, Inst. for Negroes | New York, N.Y. | | General | 4/1/11 | D | |
| Amherst College | Amherst, Mass. | 50000 | Increase Salaries | F 5/1/11 | Has done share, 5/1/11 | |
| Alma College | Alma, Mich. | | Library Building | 5/18/11 | Denominational, 3/24/11 | |
| American International College | Springfield, Mass. | | General | 11/1/11 | Not sufficiently developed, 3/2/11 | |
| Alberta Ladies' College (1) | Red Deer, Alberta | | General | 12/15/11 | Not sufficiently developed, 11/15/11 | |
| Allen University | Columbia, S.C. | | Library Building | 4/26/12 | Low | |
| Acadia University (2) | Woolfville, N.S. | 25000 | Library Building | 5/1/12 | Has done share, 4/23/12; D | |
| Kingdon Presbytery | Stafford, Va. | 500 | Building | 5/14/12 | D + Low | |
| Amity College | College Springs, Ia. | | Endowment | 5/13/12 | Not sufficiently developed, 10/15/12 | |
| Albert Lea College | Pellott, La., Minn. | | Endowment | 6/18/12 | Denominational, 11/2/12 | |
| Anderson College (1) | Anderson, S.C. | | General | 8/15/12 | D + Low | |
| Alabama University of | University, Ala. | | Memorial Building | 10/2/12 | State institution, 1/15/12 | |

Modely pre tabuľky

Tréning (beta.transkribus.eu):

- Training data
- Validation data
- Model setting
- Advanced settings: Training Cycles and Learning Rate



Modely pre tabuľky

Ground Truth: 20 strán

| 2 APPLICATIONS FOR EDUCATIONAL INSTITUTIONS | | | | | |
|---|------------------|---------|--------------------|----------|---|
| NAME OF INSTITUTION | TOWN | AMOUNT | OBJECT | DATE | DISPOSITION |
| Allegheny College (1) | P Meadville, Pa. | | Chemistry Building | 12/12/12 | Has done share, 12/17/12 |
| Allegheny College | Allegheny, Ore. | | Endowment | 12/16/12 | Discontinuation, 12/17/12 |
| Assiut College | Assiut, Egypt | | Building | 12/17/12 | Outside field of work (Egypt), 12/17/12 |
| Allegheny College (2) | P Meadville, Pa. | | Library Building | 12/21/12 | Has done share, 12/17/12 (ans. 1) |
| Atlanta Normal & Ind. Inst. (1) C | Atlanta, Ga. | | General and Land | 12/21/12 | Low; Gen 1914, 4/15/14 |
| Alabama School of Trade & Ind. | Rayland, Ala. | 25 000 | Land | 3/10/13 | Planning stage, 3/18/13 |
| American University (projected) | Washington, D.C. | 140 000 | Building | 3/12/13 | D |
| Adelphi College | Brooklyn, N. Y. | | Endowment | 5/2/13 | Not disposed, 5/5/13 |
| Urbington Lit. & Ind. School (1) C | Anacostia, D.C. | 7 000 | Building | 4/1/13 | Low; Gen 1914, 2/18/14 |
| Allegheny College (1) | Meadville, Pa. | | Building | 12/10/13 | Discontinuation & not developed, 4/1/14 |
| Austin College (1) | Sherman, Tex. | | Library Endowment | 1/10/14 | Discontinuation, 1/15/14 |
| Atlanta University | C P Atlanta, Ga. | | Endowment | 1/15/14 | Gen 1914, 2/20/14 |
| Allegheny County Academy (1) | Cumberland, Md. | | Endowment | 1/17/14 | Academy, 1/5/14 |
| Austin College (2) | Sherman, Tex. | | Organ | 1/17/14 | No organs for institutions, 1/20/14 |

Modely pre tabuľky

Rozpoznávanie s tabuľkovými modelmi

Processed pages

| 46 | | | | | |
|--|--------------------------------|---------------------|--------------------------------|---------------------|---|
| APPLICATIONS FOR EDUCATIONAL INSTITUTIONS | | | | | |
| NAME OF INSTITUTION | TOWN | AMOUNT | OBJECT | DATE | DISPOSITION |
| Emporia, College of (2) Elgin Academy (2) | P Emporia, Kan. Elgin, Ill. | | Organ Endowment | 7/24/14 10/16/14 | No organs for institutions, 10/6/14 Gen 1914, 10/30/14 |
| "Ewing School" | Ewing, Va. | | Building | 6/3/15 | |
| Emory and Henry College | P Emory, Va. | 25 000 | Endowment | 12/1/15 | |
| Elk Creek Training School | Elk Creek, Va. | | Dormitory | 12/10/15 | |
| Elisee High School | Hemp, N.C. | | Piano | 1/3/16 | |
| Emporia, College of (2) | P Emporia, Kan. | | Rebuilding of Carnegie Library | | "Gen 1914" 2/2/16 |
| Elmira College (1) 1917 ↓ | P Elmira, N.Y. | 20 000 or 25 000 | Library Building | 1/4/16 | "Gen 1914" 2/7/16 |
| Ellsworth College (2) | P Iowa Falls, Iowa | | Buildings and endowment | 3/3/17 | "Gen 1914," 4/6/17 |
| Elmira College (2) | P Elmira, N.Y. | 50 000 | Buildings and endowment | 3/10/17 | "Gen 1914," 4/6/17 |
| Edenton Ind. & Norm. College | Edenton, N.C. | | General | 10/8/17 | |
| Emory and Henry College | Emory, Va. | | Enlargement, and equipment | 7/6/18 | |
| Elizabethtown College | Elizabethtown Pa. | 50 000 | Library | 1/28/19 | Gen. |
| Esqfield Preparatory School | Esqfield, Pa. | | Building | 3/20/19 | Gen. 3/26/19 |
| Ellsworth College | Iowa Falls, Va. | | Buildings and endowment | 5/1/19 | Gen 6/2/19 |

Spracovanie dokumentov s tabuľkami

- 1) **Vytvorenie GT „základnej pravdy“ pre rozpoznávanie tabuliek:**

Spracovanie dokumentov s tabuľkami

- 1) **Vytvorenie GT „základnej pravdy“ pre rozpoznávanie tabuliek**
- 2) **Trénovanie modelu rozpoznávania tabuliek**

Spracovanie dokumentov s tabuľkami

- 1) **Vytvorenie GT „základnej pravdy“ pre rozpoznávanie tabuliek**
- 2) **Trénovanie modelu rozpoznávania tabuliek**
- 3) **Použitie modelu rozpoznávania tabuľky na zostávajúce strany**

Processing documents with tables

- 1) Vytvorenie GT „základnej pravdy“ pre rozpoznávanie tabuliek
- 2) Trénovanie modelu rozpoznávania tabuliek
- 3) Použitie modelu rozpoznávania tabuľky na zostávajúce strany
- 4) Spustenie rozpoznávania rozloženia na detekciu riadkov:
- 5) **Nastavenia:**
- 6) **Model Základnej čiary (Baseline model):** Horizontal/Mixed Text Line Orientation/Model trained by you
 - Zachovanie existujúcich oblastí textu
 - Zmena mierky obrázka
 - Minimálna dĺžka základnej čiary:Low
 - Rozdelené čiary na hranici regiónu

Processing documents with tables

- 1) **Vytvorenie GT „základnej pravdy“ pre rozpoznávanie tabuliek**
- 2) **Trénovanie modelu rozpoznávania tabuliek**
- 3) **Použitie modelu rozpoznávania tabuľky na zostávajúce strany**
- 4) **Spustenie rozpoznávania rozloženia na detekciu riadkov:**
 - **Rozpoznávanie textu (Text Recognition)**
Verejný model / Súkromný model, ktorý ste trénovali

Processing documents with tables

1. **Vytvorenie GT „základnej pravdy“ pre rozpoznávanie tabuliek**
2. **Trénovanie modelu rozpoznávania tabuliek**
3. **Použitie modelu rozpoznávania tabuľky na zostávajúce strany**
4. **Spustenie rozpoznávania rozloženia na detekciu riadkov**
5. **Rozpoznávanie textu (Text Recognition)**
6. **Korekcie (Correction (voliteľné))**
7. **Export (Excel)**

Modely polí a tabuliek: Súhrn



začnite s približne 40-60 stranami GT

50 strán pre Modely polí

- jednoduché tabuľky: 10/20 strán
- Zložité tabuľky: 30-50 strán
- Mix rôznych tabuliek: minimálne 50 strán

Príprava tréningových údajov pomocou editora rozloženia

- Oblasti kreslenia a tagovania pre modely polí (= priradiť tagy štruktúry)
- Kreslenie tabuliek pre tabuľkové modely



Pracovný postup pre prácu s tabuľkami a poľami:

1. rozpoznať oblasti alebo tabuľky
2. potom základné čiary
3. potom text



Výpočty presnosti transkripcie

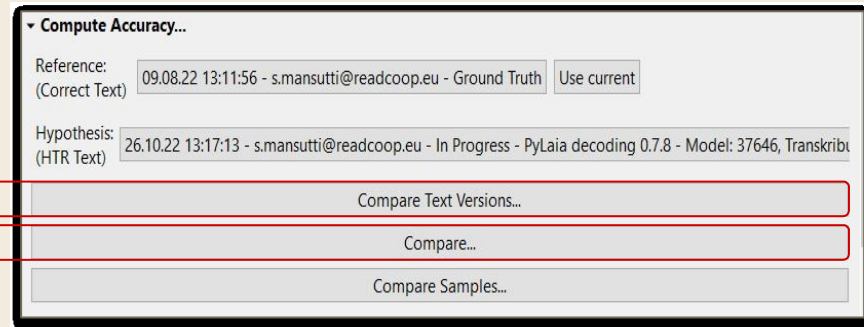


Výpočty presnosti transkripcie

Dve verzie tej istej stránky:

1. **Reference** (Ground Truth)
2. **Hypothesis** (HTR Automatic Transcription)

- **Porovnajte textové verzie (pozrite si rozdiely medzi dvoma vybratými verziami)**
- **Porovnať...(Compare)**
- **(porovnáva tieto dva prepisy a vypočítava chybovosť slov a chybovosť znakov)**



Výpočty presnosti transkripcie

Porovnať textové verzie

Ground Truth - model "Transkribus
English handwriting M3b" bez jazykového
modelu:

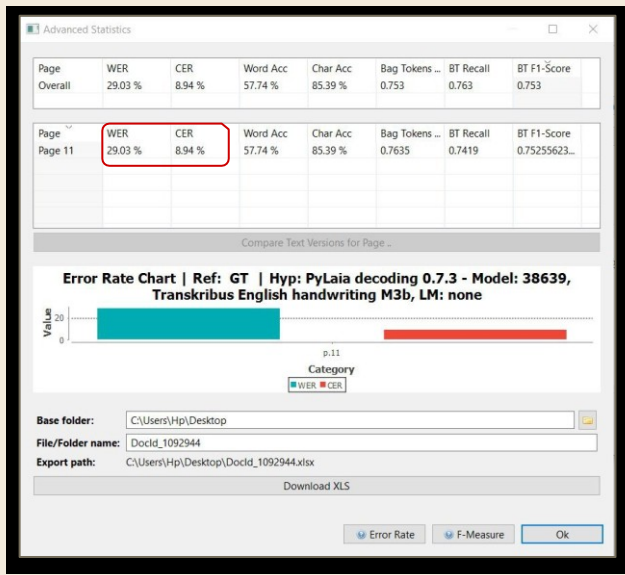


Ground Truth - "Transkribus anglický
rukopis M3b" model s jazykovým
modelom:

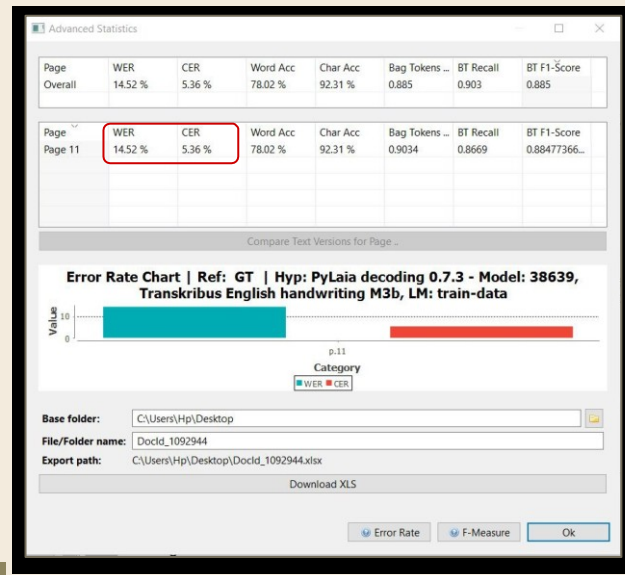


Výpočty presnosti transkripce

Porovnat'...Ground Truth - Model
"Transkribus English handwriting M3b"
bez jazykového modelu:



Ground Truth - "Transkribus anglický rukopis M3b" model s jazykovým modelom:



Výpočty presnosti transkripcie

Compare → Advanced Compare → Baselines

Compare: Advanced Compare

Type: Baselines

Pages (2): 1

Options: default (case sensitive)

Reference: GT Select hypothesis by toolname: TrHtr recognition 2.3.0 - Model: 51170, The Text Titan I

Compare

Previous Advanced Compare Results

| Created | Status | Queries | Duration | Scope | Type | Results |
|-------------------|-----------|--|-----------|--------------|------------|---|
| 26.09.23 10:35:06 | Completed | Page(s) : 1 Ref: GT Hyp : TrHtr recognition 2.3.0 - Model: 51... | 0.52 sec. | Document ... | Baselin... | P/R/F1: 0.74/0.98/0.84 (p1: 0.74/0.98/0.84) |
| 26.09.23 10:34:46 | Completed | Page(s) : 1 Ref: GT Hyp : Transkribus LA 0.0.5, Model: 49272... | 0.60 sec. | Document ... | Baselin... | P/R/F1: 0.87/0.99/0.93 (p1: 0.87/0.99/0.93) |
| 26.09.23 10:34:36 | Completed | Page(s) : 1 Ref: GT Hyp : Transkribus LA 0.0.5, Model: 51962... | 0.59 sec. | Document ... | Baselin... | P/R/F1: 0.93/0.97/0.95 (p1: 0.93/0.97/0.95) |

Options Cancel

Predvolené rozloženie s rozpoznávaním textu

Základný model orientácie zmiešanej čiary

Základný model univerzálnych línií

Kontrola kvality

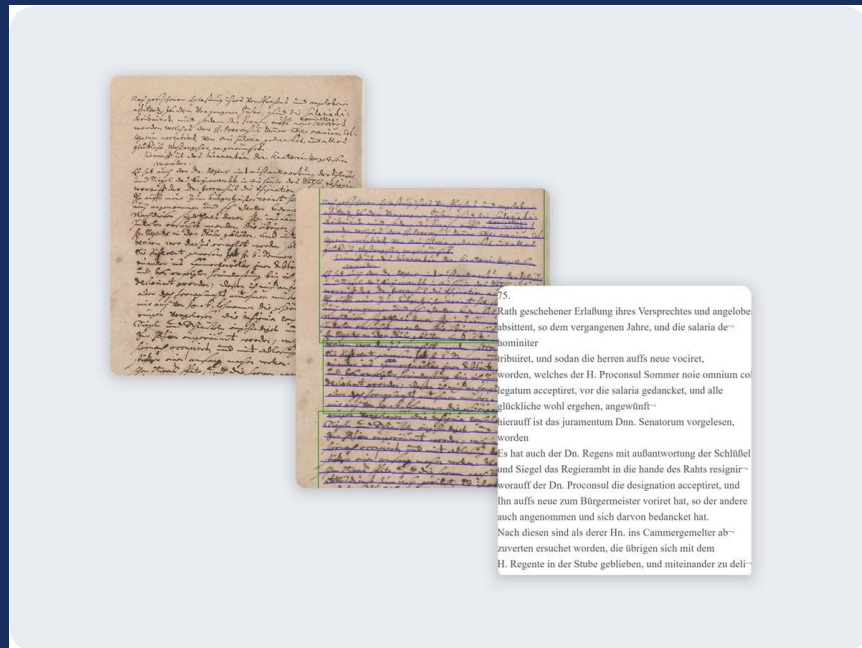
The screenshot displays the Transkribus Quality Control interface. At the top, the Transkribus logo is on the left, and navigation links for Desk, Models, Sites, Connect, and Jobs are on the right. Below the navigation bar, the breadcrumb path is "Quality Control > Sample 01 > Task 01".

Five evaluation metrics are shown in a row:

- Size: 100 Pages
- Layout Evaluation: 100%
- Transcript Evaluation: 98%
- Tags Evaluation: 97%
- Attributes Evaluation: 98%

Below the metrics is a table with the following data:

| PAGE | STATUS | ERRORS | |
|----------|--------|------------------|--------------------------|
| Page #33 | Error | Transcript | See Page |
| Page #78 | Error | Transcript, Tags | See Page |
| Page #84 | Error | Tags | See Page |
| Page #98 | Error | Tags | See Page |



Publikačné modely v Transkribus

Publikačné modely



Používatelia sa rozhodnú publikovať svoje vlastné modely, pretože

Sú hrdí na svoju prácu, a preto ju chcú sprístupniť aj ostatným používateľom, ktorí pracujú s podobnými skriptami a jazykmi

Musia publikovať čo najviac

Majú záujem o spoluprácu s inými vedcami na súvisiacich projektoch

Môžu vedieť o iných kolegoch alebo výskumných projektoch, ktoré by chceli použiť model, ale nemôžu zdieľať tréningové údaje

[Zenodo](#) Komunita pre publikovanie súborov údajov GT
plánuje zahrnúť priame rozšírenie od spoločnosti Transkribus

Publikačné modely

Ako publikovať model:

Kontaktujte nás prostredníctvom info@readcoop.eu alebo prostredníctvom [contact form/help center](#) aby ste nás informovali, že chcete zverejniť svoj model v rámci spoločnosti Transkribus

- Požiadavky: veľkosť tréningovej sady ~ 50 000 slov, CER 7%-5% alebo nižšia . Ak ide o model vyškolený na skript alebo jazyk, ktorý zatiaľ nemôžeme ponúknuť, tieto kritériá neplatia
- Poskytnúť stručný opis modelu, ktorý pomôže ostatným používateľom pochopiť použitý obsah školenia; Užitočné je aj pridanie reprezentatívneho obrázka alebo úryvku
- Povedzte nám, kto by mal byť uvedený ako tvorca modelu - môže to byť jedna alebo viac osôb alebo celý výskumný projekt
- Viditeľnosť tréningových údajov: môžu byť zachované v súkromí (z dôvodov ochrany údajov) alebo zdieľané, aby boli aj údaje o školeniach verejné



 Desk

 Models

 Sites

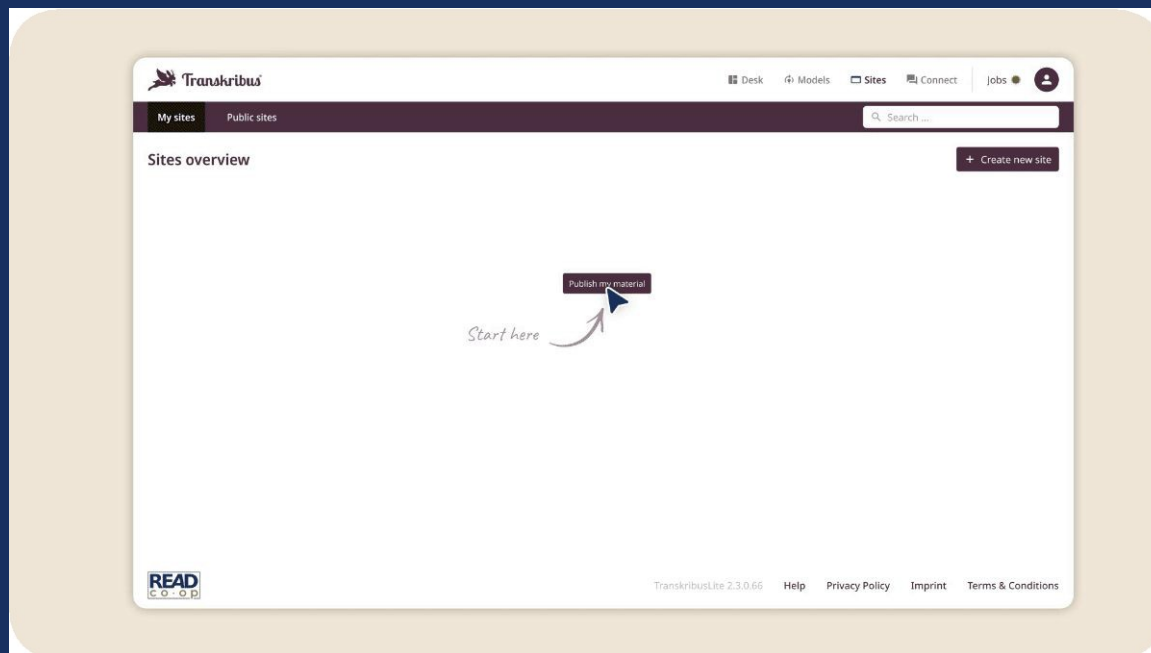
 **Connect**

Jobs 



Plánované na rok 2024

Transkribus **Connect** je miesto, kde sa **exchange** stane.



Transkribus stránky



 Desk

 Models

 Sites

Jobs



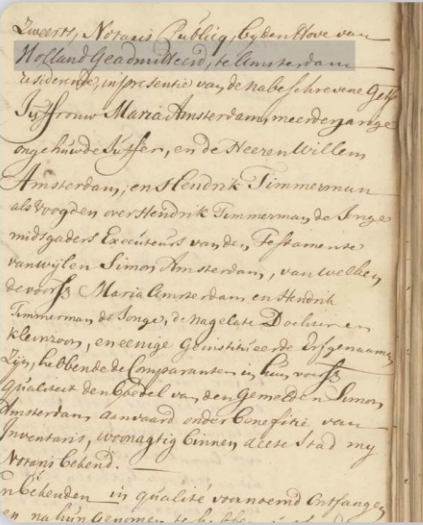
Transkribus **Connect** je miesto, kde sa **exchange** stane.

Plány predplatného

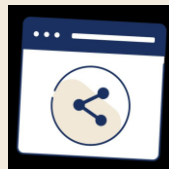
| | | |
|---|--|---|
| <h2>Individual</h2> <hr/> <p>0 €</p> <p>Ideal for Genealogists & Students /month incl. 20% VAT*</p> <hr/> <ul style="list-style-type: none">✓ AI Text Recognition✓ Custom AI Training✓ DOCX & PDF Export <p>Choose plan</p> | <h2>Scholar</h2> <hr/> <p>14.9 €</p> <p>Tailored for Individual Researchers /month incl. 20% VAT*</p> <hr/> <ul style="list-style-type: none">✓ Collaboration Tools✓ Advanced AI Tools✓ Transkribus Sites <p>Choose plan</p> | <h2>Organisation</h2> <hr/> <p>—</p> <p>For Research & Cultural Institutions</p> <hr/> <ul style="list-style-type: none">✓ User Management✓ Dedicated Success Manager✓ API Access <p>Get in Touch</p> |
|---|--|---|

100 Free Credits / Month

Transkribus stránky - vlastnosti



Op Heden den 27sten Maart Ao. 1790
Compareerden voor my Philip
Zweerts, Notaris Publicq, by den Hove van
Holland geadmitteerd, te Amsterdam
residerende, in presentie van de nabeschrevene Get.
Juffrouw Maria **Amsterdam** meerderjange
ongehuwde Suffer, en de Heeren Willem
Amsterdam, en Hendrik Timmerman
als Voogden over Hendrik Timmerman de Jonge
midsgaders Executeurs van den Testamente
van wijlen Simon **Amsterdam**, van welken
devoorsz Maria **Amsterdam** en Hendrik
Timmerman de Jonge, de nagelate Dochter en
kleinzoon, en eenige geinstitueerde erfgenaamen
Zijn, hebbende de Comparanten en hun voorsz
qualiteit den boedel van, den gemelden Simon
Amsterdam, aanvaard onder Conscriptie van
Inventaris, woonagtig binnen dese Stad my
Notaris bekend.
En bekenden in qualite voornoemd Ontfangen
en na hungenomen te hebben, uithanden
van Juff. Anna westerveen weduwe van
simon **Amsterdam** voornoemd alle de meubilen



Jednoduché zdieľanie materiálu

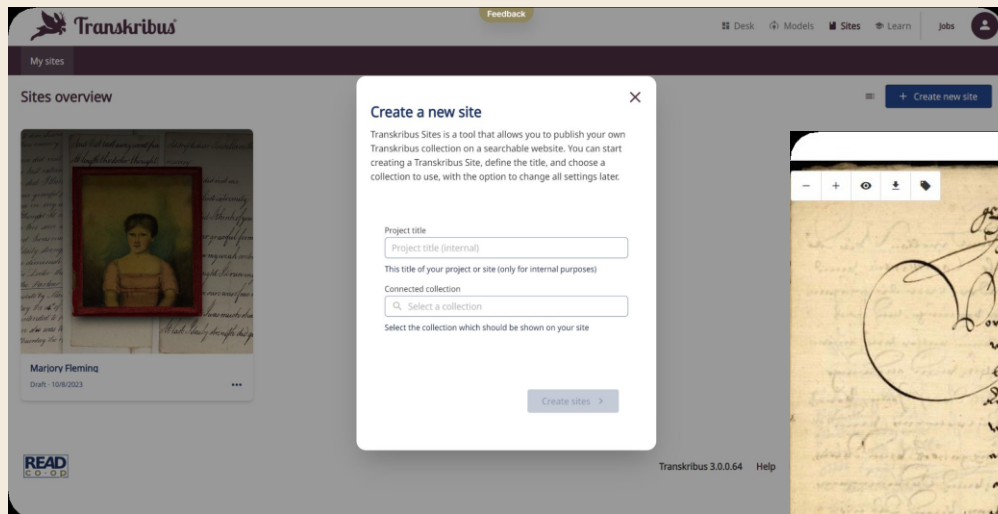


Pohľad strana vedľa strany
(obrázok-prepis)



Vylepšené možnosti vyhľadávania

Transkribus stránky

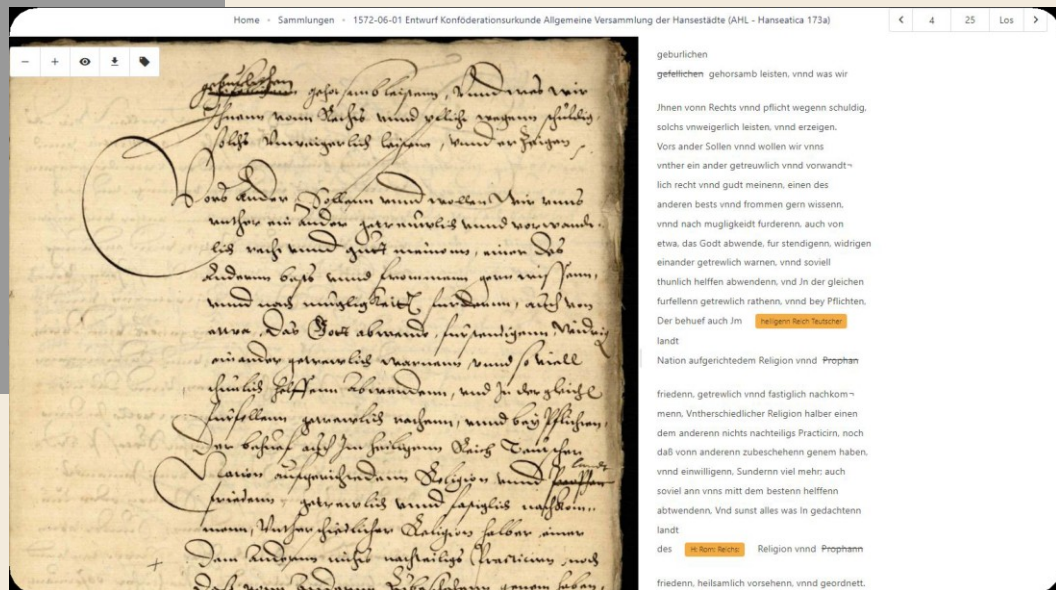


The screenshot shows the Transkribus interface with a 'Create a new site' modal dialog open. The dialog has a title bar with a close button (X) and a '+ Create new site' button. The main text reads: 'Transkribus Sites is a tool that allows you to publish your own Transkribus collection on a searchable website. You can start creating a Transkribus Site, define the title, and choose a collection to use, with the option to change all settings later.'

Below the text are three input fields:

- Project title**: A text input field with the placeholder 'Project title (internal)'.
- This title of your project or site (only for internal purposes)**: A text input field.
- Connected collection**: A dropdown menu with the placeholder 'Select a collection'.

Below the fields is a label: 'Select the collection which should be shown on your site'. At the bottom right of the dialog is a 'Create sites >' button. The background shows a 'Sites overview' page with a card for 'Marjory Fleming' (Draft, 10/8/2023) and a 'READ' logo.



The screenshot shows a document viewer interface for a handwritten manuscript. The top navigation bar includes 'Home', 'Sammlungen', and the document title '1572-06-01 Entwurf Konföderationsurkunde Allgemeine Versammlung der Hansestädte (AHL - Hanseatica 173a)'. On the right, there are navigation icons for back, page 4, 25, and forward.

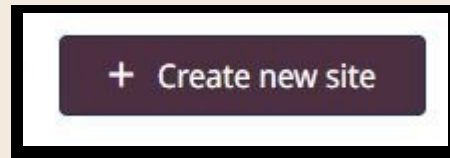
The main area displays a page of handwritten text in a cursive script. The text is partially obscured by a large, stylized initial 'D'. The viewer includes standard image controls: a minus sign, a plus sign, a refresh icon, a download icon, and a zoom icon.

On the right side, there is a transcription of the text. The first line is 'geburlichen'. The second line is 'geburlichen gehorsamb leisten, vnnnd was wir'. The third line is 'Ihnen vnnnd Rechts vnnnd pflicht wegnen schuldig, solchs vnnneiglich leisten, vnnnd erzeigen. Vns ander Sollen vnnnd wollen wir vnnns vnther ein ander getweulich vnnnd vorwandtlich recht vnnnd gndt meinnen, einen des anderen bests vnnnd frommen gern wissen, vnnnd nach muglichkeit fuderern, auch von etwa, das Godt abwende, fur stendigen, widrigen einander getweulich warnen, vnnnd soviel thunlich helfen abwendenn, vnnnd In der gleichen furtellenn getweulich rathenn, vnnnd bey Pflichten. Der behuef auch Im **Religien Recht** landt Nation aufgerichtetem Religion vnnnd Prophan friedenn, getweulich vnnnd fastiglich nachkommenn, Vnterschiedlicher Religion halber einen dem anderen nichts nachteiligs Practicirn, noch daß vnnn anderen zubeschehenn genem haben, vnnnd einwilligen, Sundern viel mehr: auch soviel ann vnnns mitt dem bestenn helfenn abwendenn, Vnnnd stund alles was In gedachtem landt des **Religien** Religion vnnnd Prophan friedenn, heilsamlich vorsehenn, vnnnd geordnet.

Vaša prvá stránka Transkribus

Vytvorenie novej stránky

- Názov projektu
- Vlastná webová adresa(app.transkribus.eu/sites/yourchosenname)
- Prepojené zbierky



Vaša prvá stránka Transkribus

3 editovateľné stránky:

- **Domov**
- **O**
- **Preskúmať**

upravovať stránky a zobrazovať aktualizácie súčasne, vedľa seba

Vaša prvá stránka Transkribus

Domov: (Home - Domovská stránka)

- Titul
- Stručný opis obsahu/stránky
- Obrázok pozadia domovskej stránky

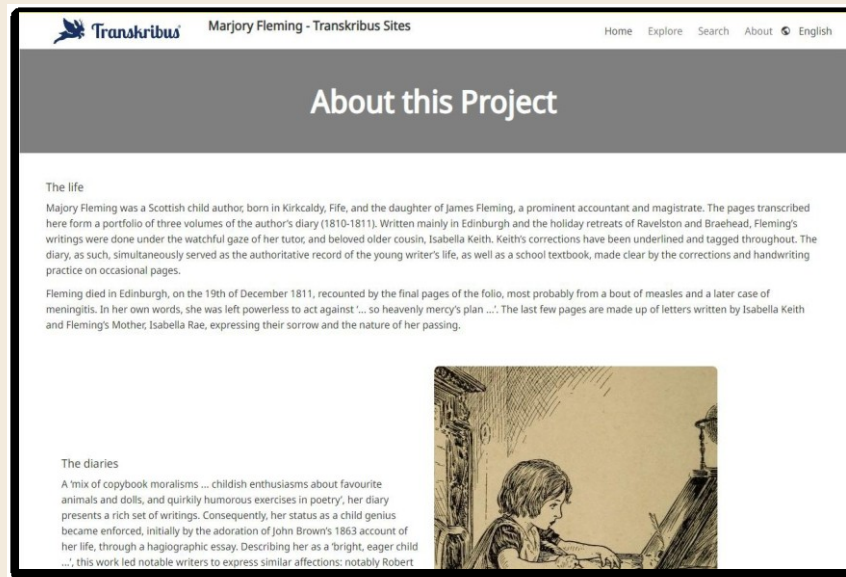


Vaša prvá stránka Transkribus

O (About)

(Vysvetlenie projektu,
obsah, tím...):

- Toľko sekcií, koľko chcete
- Každá časť: nadpis - text - obrázok (voliteľné)



The screenshot displays the 'About this Project' page for Marjory Fleming on the Transkribus website. The page features a dark header with the Transkribus logo and navigation links (Home, Explore, Search, About, English). The main content area is titled 'About this Project' and contains two sections: 'The life' and 'The diaries'. The 'The life' section provides a biographical overview of Marjory Fleming, mentioning her birth in Kirkcaldy, her family, and her education. The 'The diaries' section describes the content of her diaries, including moralisms, childlike enthusiasms, and humorous exercises. An illustration of a young girl writing at a desk is positioned to the right of the 'The diaries' text.

Transkribus Marjory Fleming - Transkribus Sites Home Explore Search About English

About this Project


The life

Majory Fleming was a Scottish child author, born in Kirkcaldy, Fife, and the daughter of James Fleming, a prominent accountant and magistrate. The pages transcribed here form a portfolio of three volumes of the author's diary (1810-1811). Written mainly in Edinburgh and the holiday retreats of Ravelston and Braehead, Fleming's writings were done under the watchful gaze of her tutor, and beloved older cousin, Isabella Keith. Keith's corrections have been underlined and tagged throughout. The diary, as such, simultaneously served as the authoritative record of the young writer's life, as well as a school textbook, made clear by the corrections and handwriting practice on occasional pages.

Fleming died in Edinburgh, on the 19th of December 1811, recounted by the final pages of the folio, most probably from a bout of measles and a later case of meningitis. In her own words, she was left powerless to act against '... so heavenly mercy's plan ...'. The last few pages are made up of letters written by Isabella Keith and Fleming's Mother, Isabella Rae, expressing their sorrow and the nature of her passing.

The diaries

A 'mix of copybook moralisms ... childish enthusiasms about favourite animals and dolls, and quirkily humorous exercises in poetry', her diary presents a rich set of writings. Consequently, her status as a child genius became enforced, initially by the adoration of John Brown's 1863 account of her life, through a hagiographic essay. Describing her as a 'bright, eager child ...', this work led notable writers to express similar affections: notably Robert

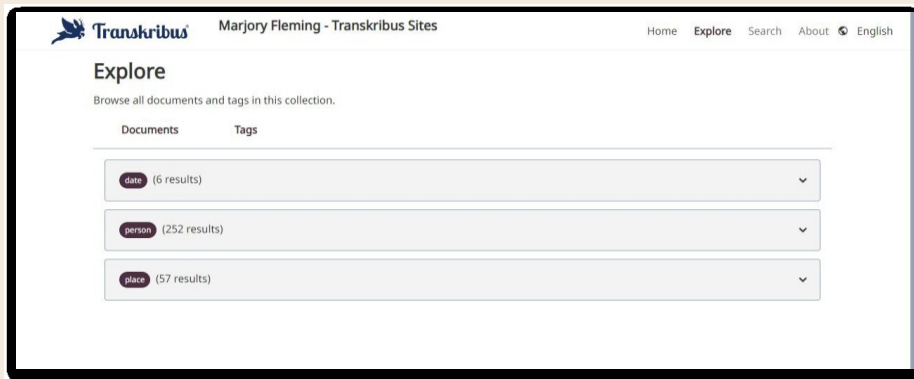


Vaša prvá stránka Transkribus

Preskúmať (Explore)

(Ako chcete nakonfigurovať stránku vyhľadávania):

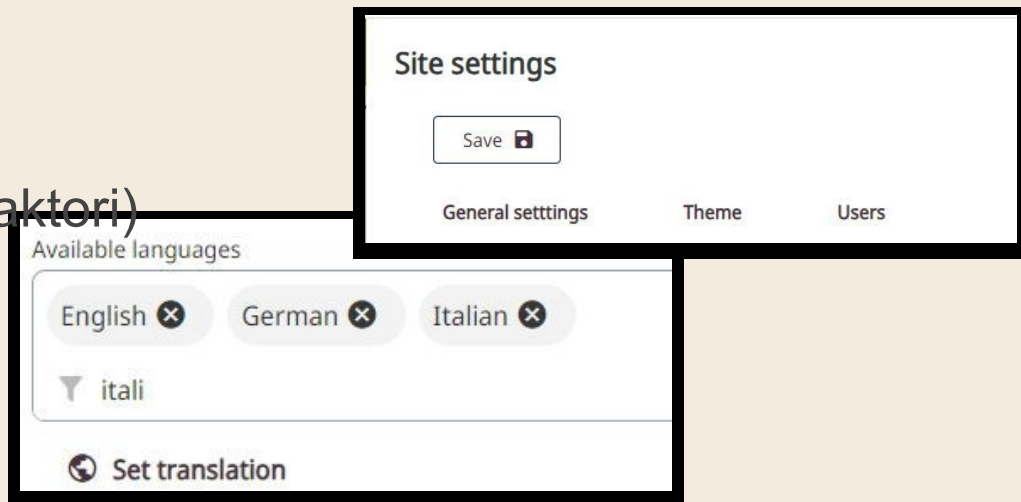
- Povolenie značiek prehľadávania
- Povolené značky (ak ste použili značky vo vašich dokumentoch Transkribus)
- Povoľiť filtre a filter rokov (na základe metadát dokumentov Transkribus)



Vaša prvá stránka Transkribus

Ďalšie nastavenia (Other settings):

- Jazyky + možnosť úpravy prekladov
- Súkromie
- Motív (logo a farba)
- Používatelia (vlastník, redaktori)





Čas na otázky





Hands-on
session
Praktické
sedenie



Help Center

<https://help.transkribus.org/>



Thank you!

Website: <https://transkribus.org/>

Email addresses:

s.mansutti@readcoop.eu

m.elattal@readcoop.eu

info@readcoop.eu





Unlocking the past, together

