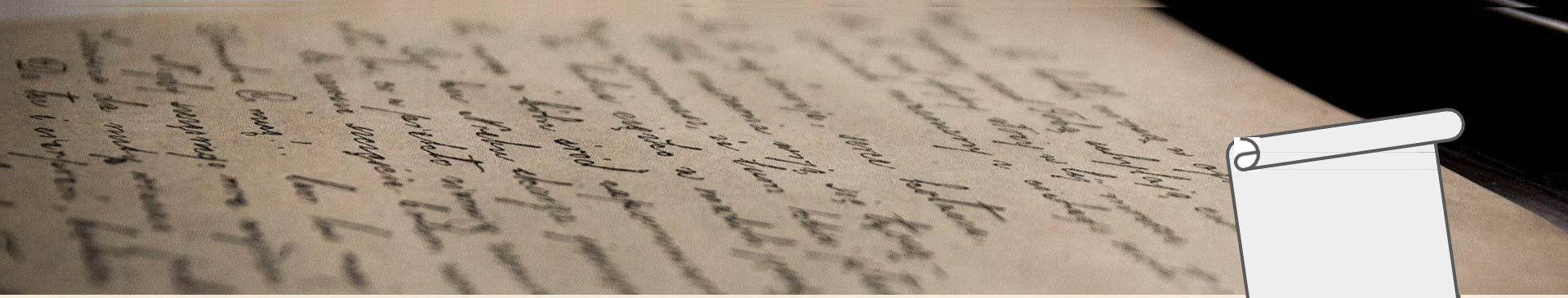


Pre pokročilých používateľov

15th February 2024

Vaši sprievodcovia pre dnešnú cestu

Mirjam
User Success
Team

Sara
User Success
Team

(Dušan Katuščák, doplnky a preklad)

s.mansutti@readcoop.eu

m.elattal@readcoop.eu

Digitálna knižnica - Texty

Sprístupnenie z digitálnych repozitárov:

1. **Digitálna knižnica obrázková (len nasnímané obrázky)**
2. **Digitálna knižnica plnotextová (full texts) (obrázky + OCR/HTR)**
3. **Digitálne knižnica hybridná (čiastočne obrázky + čiastočne OCR/HTR)**

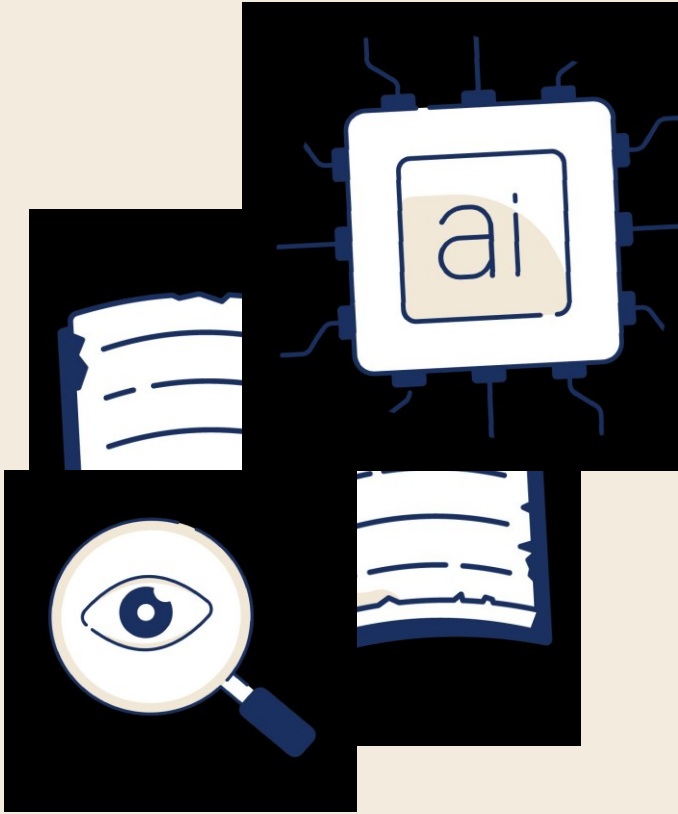
Obsah

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

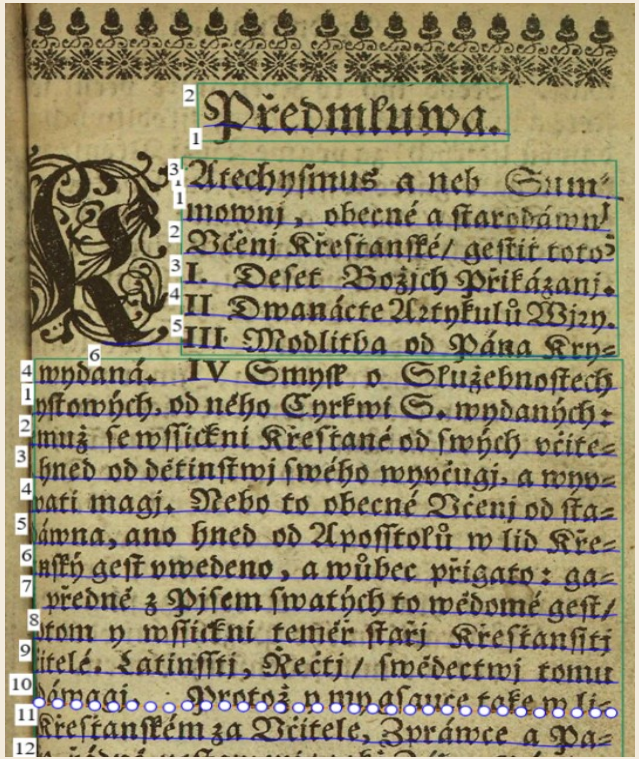
Úvod

Čo je Transkribus?

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



České dokumenty – práca študentov na Opavskej univerzite (Halfarová)



Region 2

1 Předmluwa.

Region 3

1 Atechysmus a neb Sum
2 mowni, obecné a starodawn | |
3 Vcenj Krestanské / gestit toto
4 I. Deset Bozich Prikázanj.
5 II Dwanácte Artykulu Wjry.
6 III Modlitba od Pána Kry=

Region 4

1 wydaná. IV. Smysk o. Skuzebnostech
2 ystowých. od něho Cyrkwi S. wydaných:
3 muzi se wssickni Krestane od swych: vcite= | |
4 hned od dětinstwí swého wyvčují, a wyo-
5 wati magj. Nebo to obecné Vcenj od sta= |
6 dawna, ano hned od Aposstolü w lid Kre= |
7 nský gest vwedeno, a wübec prigato: ga= | |
8 předné z Písem swatých to wědomě gest /
9 ptom y wssickni teměř staří Kresťansítí
10 litele: Latinssti, Recti / swědectwi tomu
11 dáwají. Protož y myjsauče take wli-

České dokumenty – práca študentov na Opavskej univerzite (Taufrová)

Čert a Prawda.
To geſt:
welmi pěkně ſmyſſlené, utěſſené
Hystorye a Rozpráwky
w několika ſtech,
kteréž ſe pro wyraženj mysli a pro
zasmánj při dlauhé chwili, y w každé
weselé ſpolečnoſti, časem také y drobet
pro wybrausenj rozumu dobře vžj-
wati mohau.
Wydané podlé rukopiſu
Hylarya, Jokoſa, Aſtucya.
Gakožto II. Díl
k Zrcadlu Poſſetiloſtj.
Kraméryuſowým nákladem.
W Praze, 1796.
k dostánj w Czeſké Expedycý w Dominy-
kánſké ulicy, u Hrabů w Nie. 373.

Čert a Prawda.
To geſt:
welmi pěkně smyſſlené, utěſſené
Hystorye a Rozpráwky
w několika ſtech,
kteréž se pro wyražení mysli a pro
zasmání při dlauhé chwíli, y w každé
weselé společnosti, časem také y drobet
pro wybrausení rozumu dobře uží-
wati mohau.
Wydané podlé rukopisu
Hylarya, Jokosa, Astucya.
Gakozto II. Dil
k Zrcadlu Possetilostj.

Region 2

Region 3

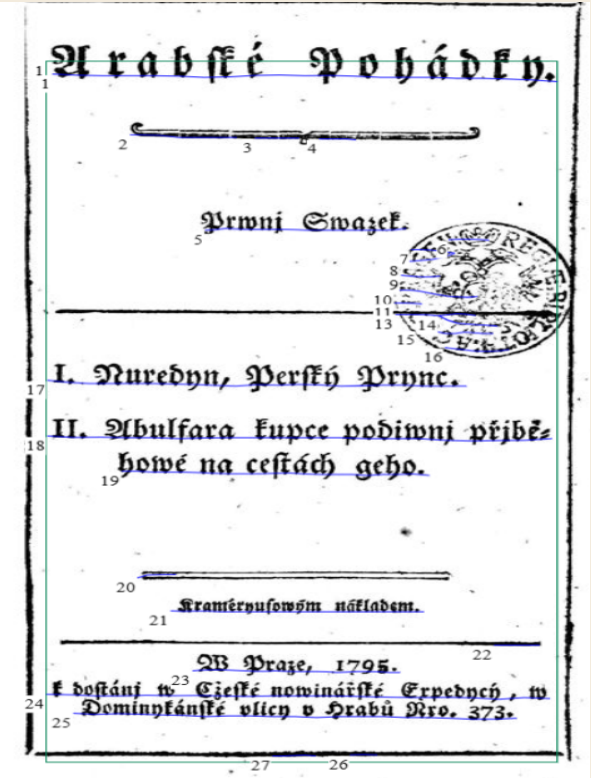
Kraméryusowým nákladem.

Region 4

Region 5

W Praze, 1796.
kdostání w České Expedycý w Dominy-
kánské ulicy, u Hrabů w Nie. 373.

České dokumenty – práca študentov na Opavskej univerzite (Kocianová, Olomouc, Žerotín, aféra s čarodejnicami...)



Arabské Pohádky.

Prwnj Swazek.

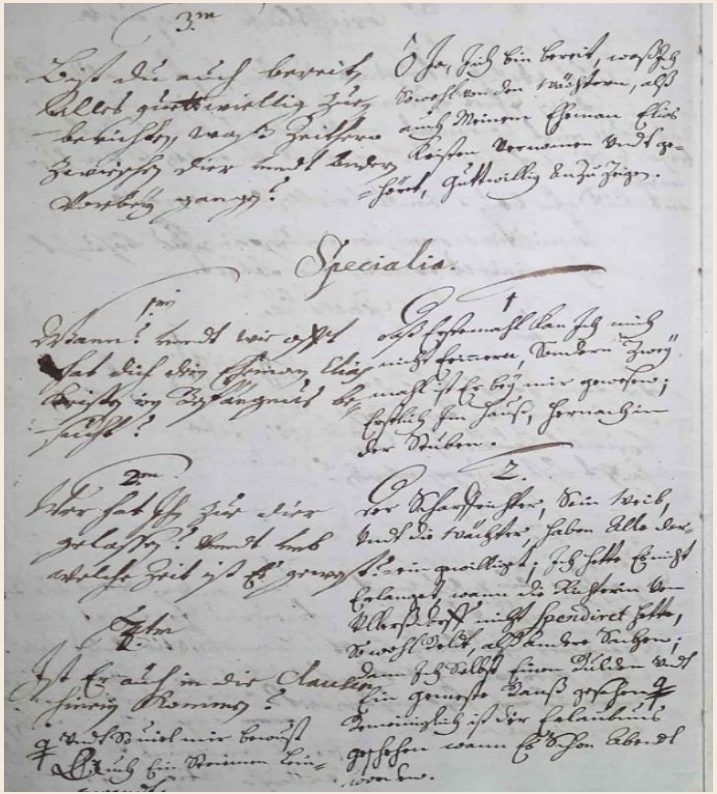
I. Nuredyn, Perský Prync.

II. Abulfara kupce podiwnj příbě-
howé na cestách geho.

Kramériusowým nákladem.

W Praze, 1795.

K dostánj w Czeské nowinářské Expedycy, w
Dominykánské ulicy v Hrabů Nro. 373.



Specialia.

České dokumenty – práca študentov na Opavskej univerzite (Gajdošová)

Náboženné Pjsně,

které

w německém gazyku složil

C. F. Gellert,

někdy

w Akadémii Lipské Včitel,

pro gegich pak užitečnost w slowenčinu

uwédl, a k nim některé nowé

přípogil

Samuel Čžerňanský

Cyrk. Ew. Bátowské

Sl. B. Kazatel.

w Preſſporku,

u Franciſſka Auguſtýna Pazko.

1787.

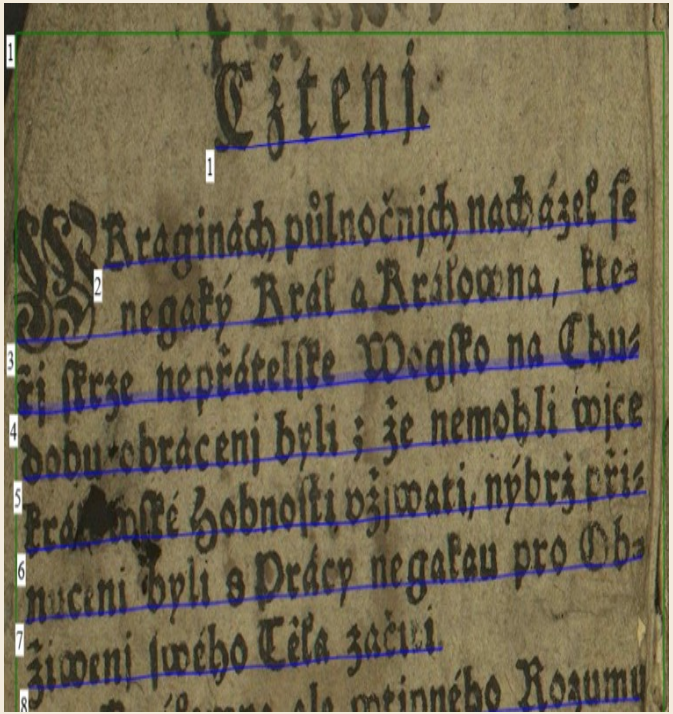
Region 1

Nábožné Písně

které
w německém gazyku složil
Gellert
C. J.
někdy
w Akadémii Lipské Učitel,
pro gegich pak užitečnost w slowenčinu
uwédl, a k nim některé nové
připogil
Samuel Čžerňanský
Cyrk. Ew. Bátowské
Sl. B. Kazatel.

W Preſſporku,
U Franciſſka Augustýna Pazko.
1787

České dokumenty – práca študentov na Opavskej univerzite (Málková)

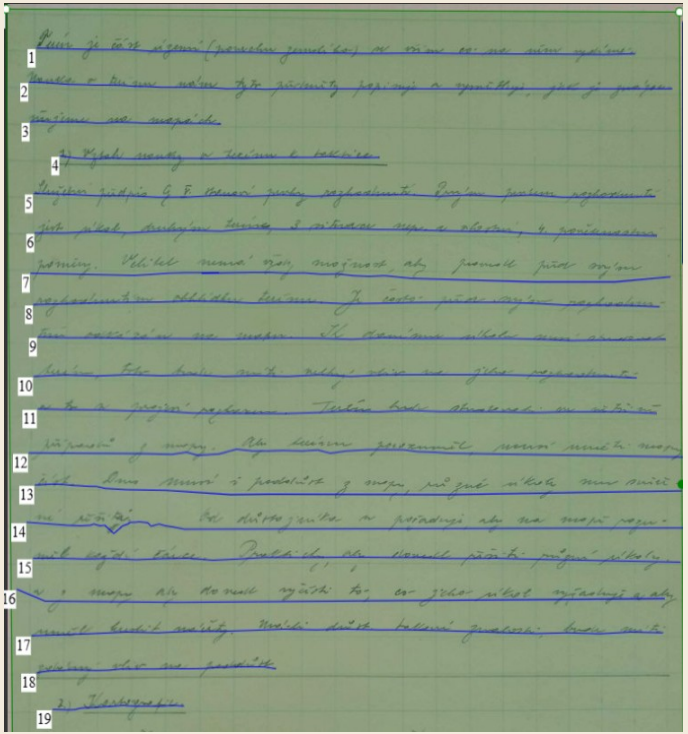


Region 1

Čzteni.

W kraginách půlnočnjich nacházel se negaký Král a Královna, kteřj skrze nepřátelske Wogsko na Chudobu obráceni byli ; že nemohli wjce králowské hodnosti užjwati, nýbrž přinuceni byli s Prácy negakau pro Obživeni swého Těla začiti.

České dokumenty – práca študentov na Opavskej univerzite (Němec)



Region 1

Terén je část území (povrchu zemského) se vším co na něm vydíme
Nauka o terénu nám tyto předměty popisuje a vysvětluje, jak je znázor-
ňujeme na mapách
1) Vztah nauky o terénu k taktice.
Služební předpis G V. stanoví prvky rozhodnutí. Prvým prvkem rozhodnutí
jest úkol, druhým terén, 3 situace nep. a vlastní, 4. povětrnostní
poměry. Velitel nemá vždy možnost, aby provedl před svým
rozhodnutím obhlídku terénu. Je často před svým rozhodnu-
tím odkázán na mapu. K danému úkolu musí stanovit
terén, toto bude míti velký vliv na jeho rozhodnutí.
a to se projeví rozkazem. Terén bude sledovati ve většině
případů z mapy. Aby terénu porozuměl, musí uměti mapy
číst. Dnes musí i provést z mapy různé úkoly mu svěře-
né řešiti. Od důstojníka se požaduje, aby na mapě rozu-
měl každé čárce. Prakticky, aby dovedl řešiti různé úkoly.
a z mapy aby dovedl vyčísti to, co jeho úkol vyžaduje a aby
uměl kreslit náčrty. Má-li důstojník takové znalosti, bude míti
zdárný vliv na poddůst.
2, Kartografie

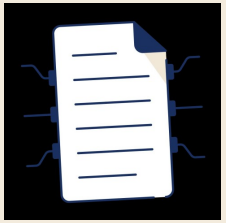
Začíname: Prvé kroky v Transkribuse

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

Vyhľadávanie

Čo Transkribus umožňuje?

Manuálny a automatický prepis ručne písaných a tlačených dokumentov



Trénovanie modelov umelej inteligencie

Spolupráca

Tagovanie štruktúry a obsahu dokumentov

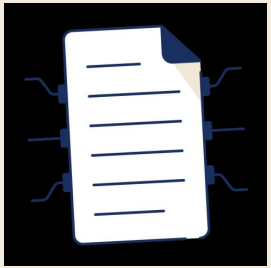
Export dokumentov v rôznych formátoch

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.

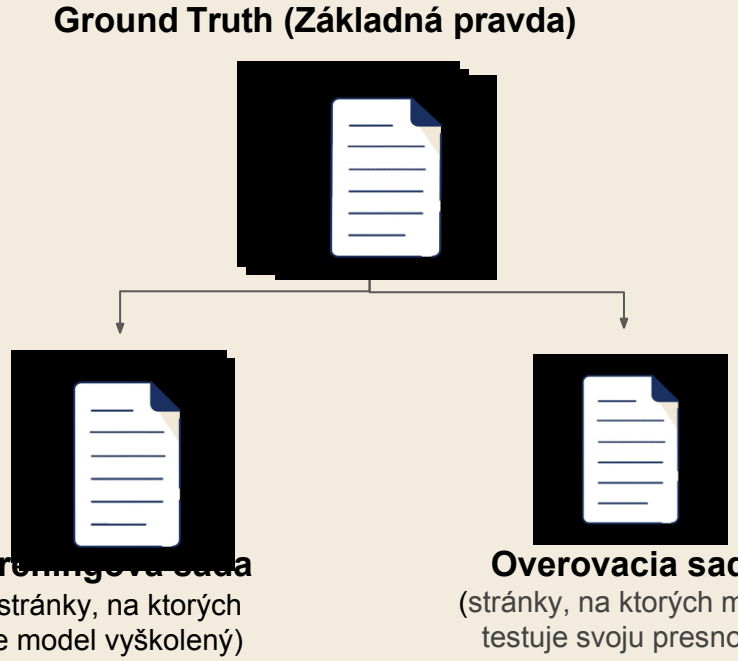
https://help.transkribus.org/text-recognition

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 trénovania

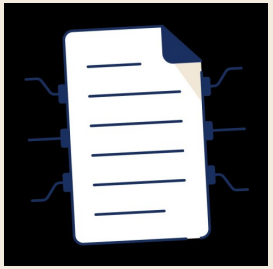


Dobrá overovacia sada: to je 10% tréningovej sady + obsahuje všetky príklady (znaky, glyfy)

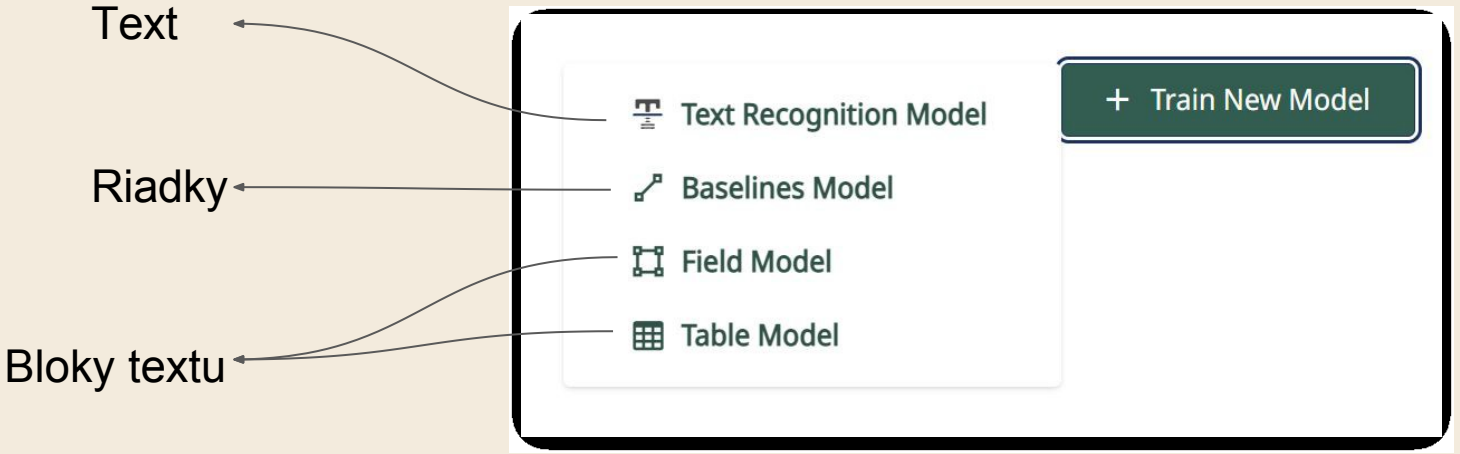
Tréning modelov

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

Trénovanie modelov AI

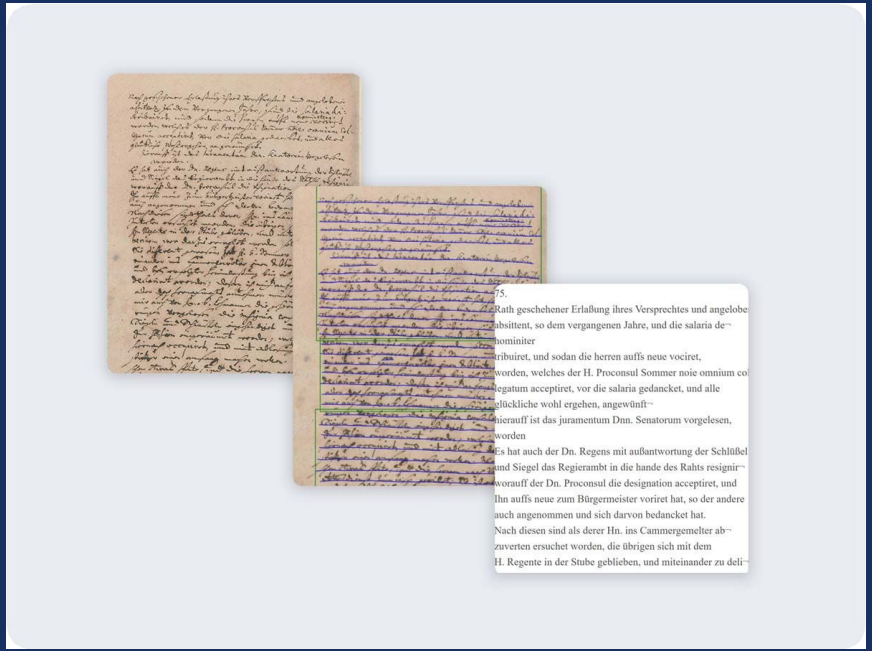


Modely trénovateľné s Transkribusom:



Analýza rozloženia (segmentácia)

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



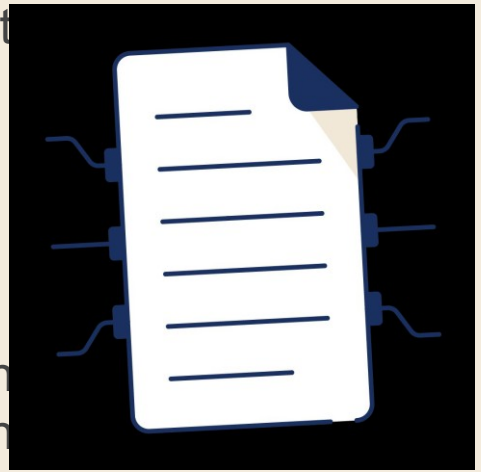
Trénovanie textových modelov

Transkribus

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 n podobnom skripte (ak je k dispozícii) a manuáln oprava prepisu

Textové modely

1. **možnosť: manuálny prepis dokumentov**

    1. Vyberte stránky, ktoré chcete zahrnúť do **GT_Základnej pravdy**
    2. Spustite 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

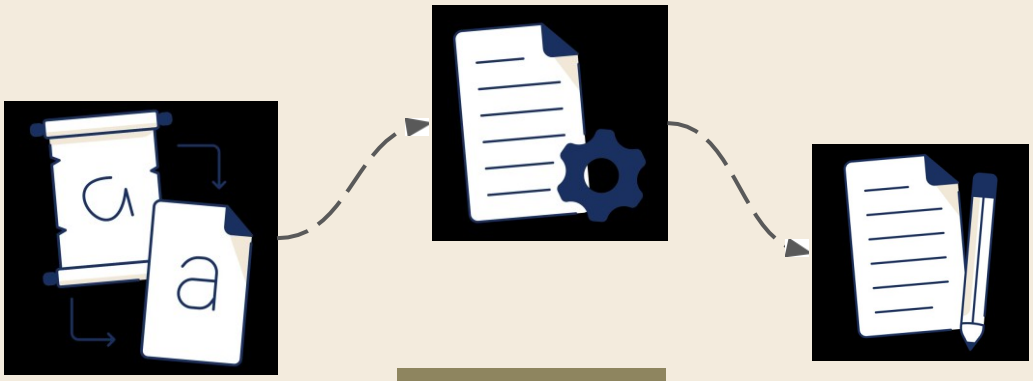
    Skratky: udržiavané/riešené/označené: záleží na tom, čo očakávate ako konečný výstup

    Uložte stránku ako GT "Základnú pravdu“!

Textové modely

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

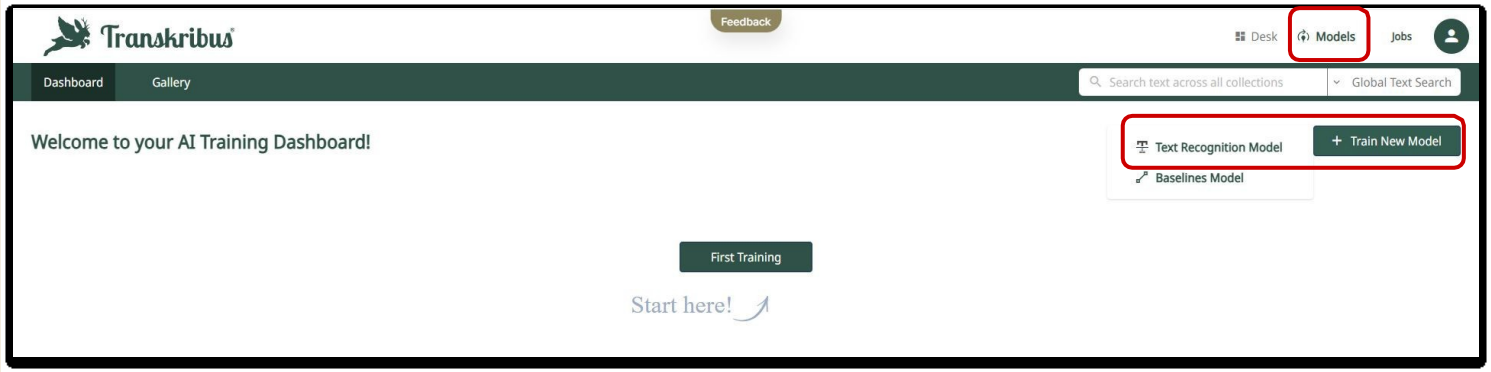
1. Vyberte stránky, ktoré chcete zahrnúť do Základnej pravdy (GT)
2. Spustenie rozpoznávania textu
3. Oprava automatických prepisov
4. Uložte stránku ako "Základnú pravdu„ (GT)



Textové modely

Po vytvorení prepisov (Základná pravda):

- prejdite do sekcie "Modely“
- kliknite na "Train New Model - Text Recognition Model"
- vyberte zbierku s prepismi (Základná pravda) Ground Truth



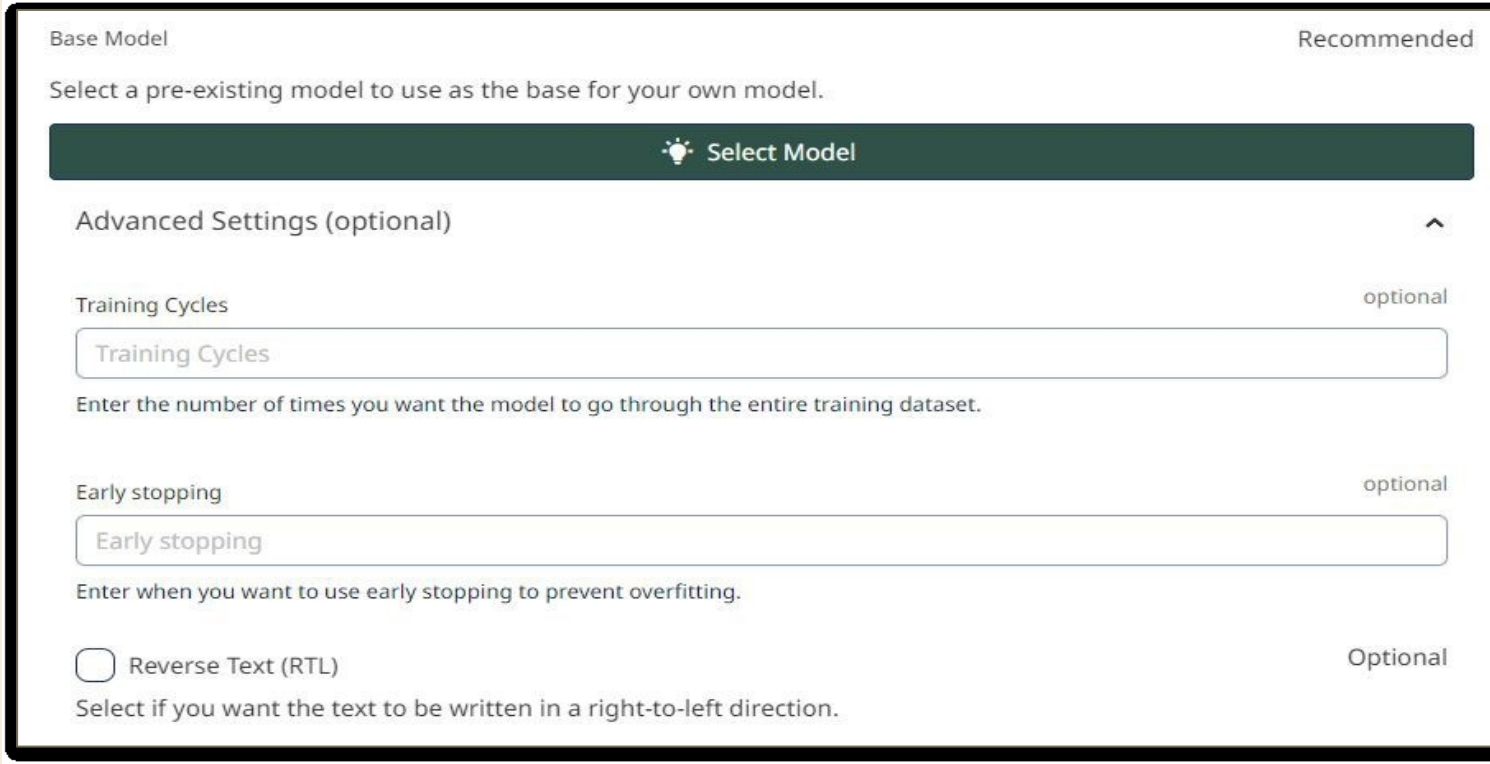
Textové modely

- 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

Rozšírené možnosti:



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

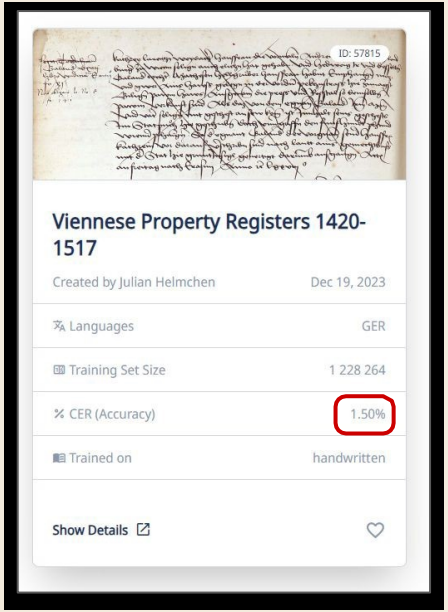
Rozšírené možnosti:

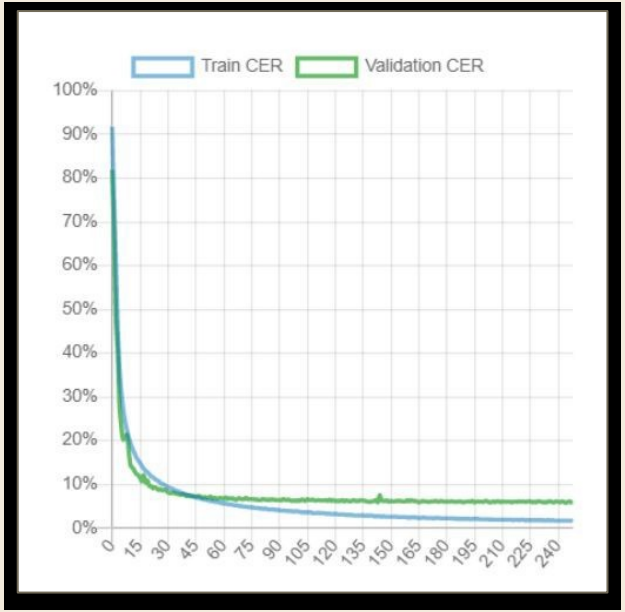
- **Obrátený text (Reverse text (RTL))**: Ak bol text na obrázku napísaný sprava doľava, ale v textovom editore bol prepísaný zľava doprava
- **Použitie existujúcich polygónov (Use existing line polygons):** Pozn.: používať iba v prípade, že ste upravili mnohouholníky v *Transkribus Expert*
- **Tréning 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/uclear.

Textové modely

Po dokončení tréningu sa môžete pozrieť na podrobnosti modelu:

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





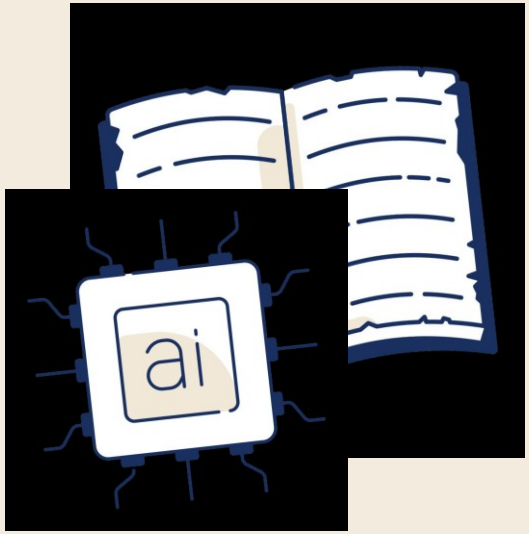
Textové modely

	CER (chybovosť znakov)	Tréningová sada
Tlačený 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

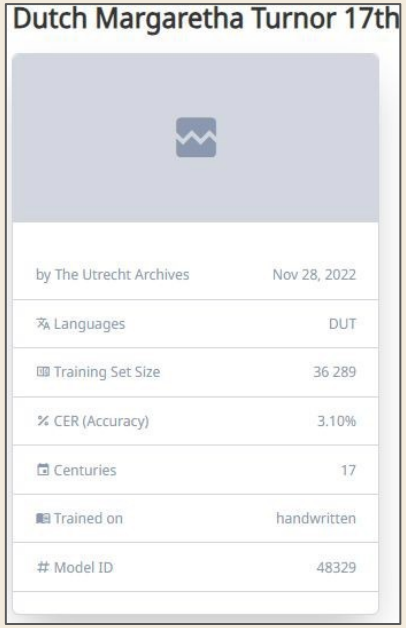
- Ruky, ktoré nie sú nijako zistené, alebo načmárané poznámky oveľa horšie výsledky, tak potom:
- Zdvojnásobte počet tréningových dát 20-25% zníženie chybovosti

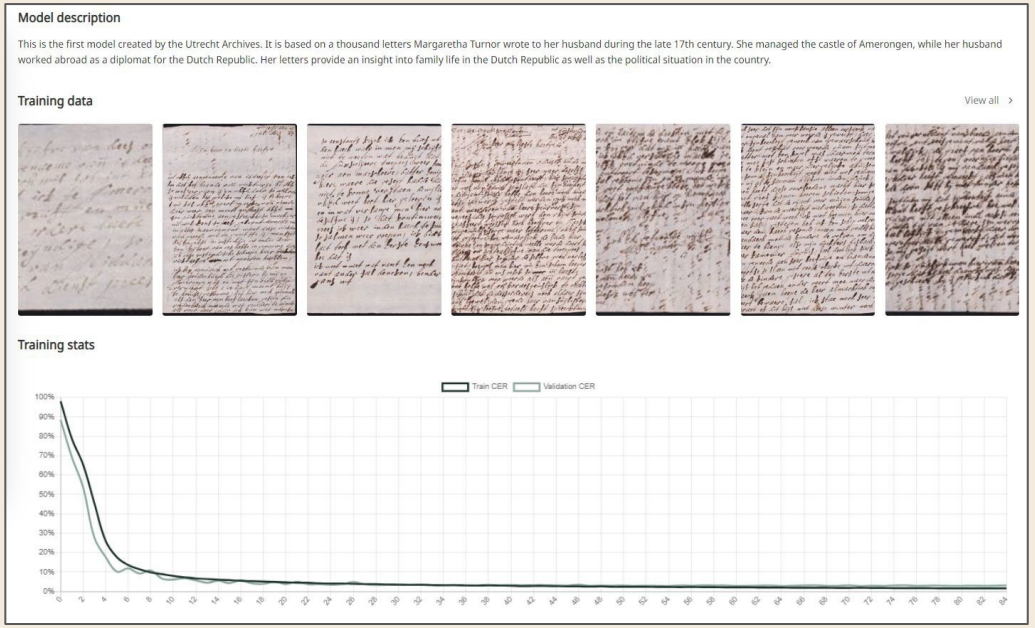
- **Existujúce modely** sa môžu použiť ako východiskový krok (Base model - základný model) na zníženie požadovaného množstva nových údajov



Textové modely

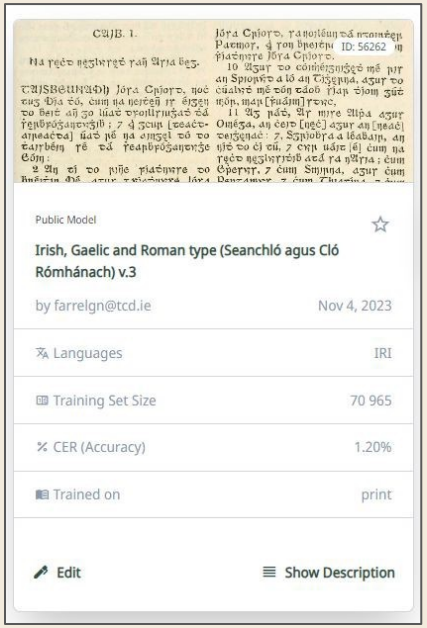
Verejný holandský rukopisný vzor:[Dutch Margaretha Turnor 17th Century]
Trained by The Utrecht Archives; Training set: 178 pages, Validation set: 20 pages

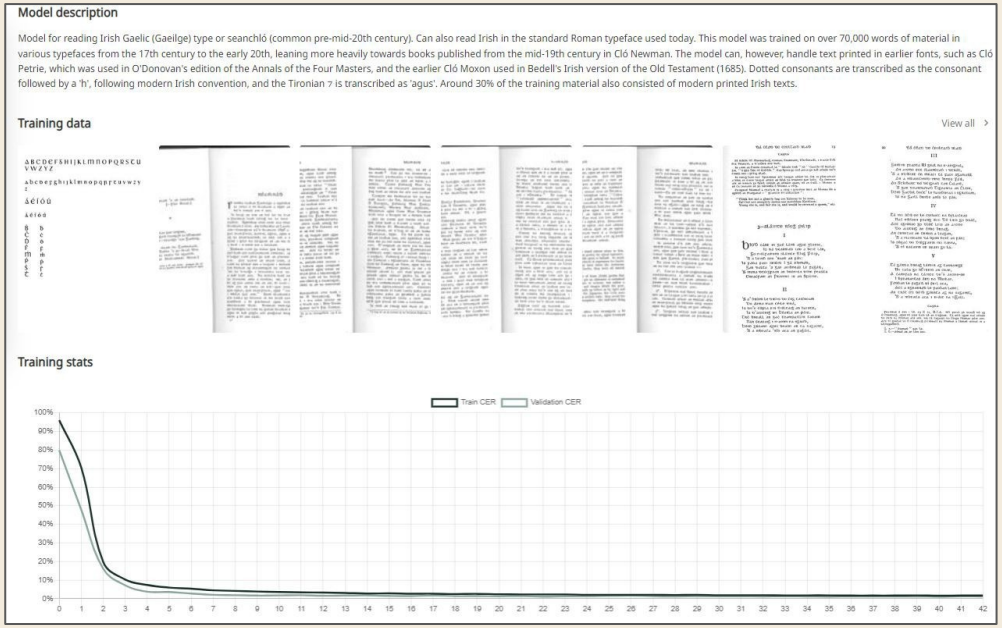


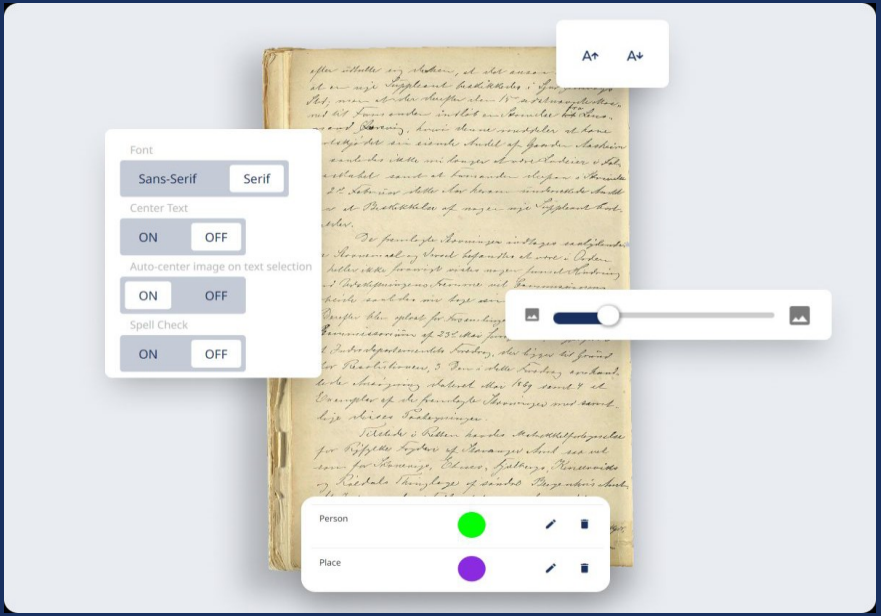


Textové modely

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







Tagovanie/Značkovanie

https://help.transkribus.org/tagging

Transkribus

Tagging

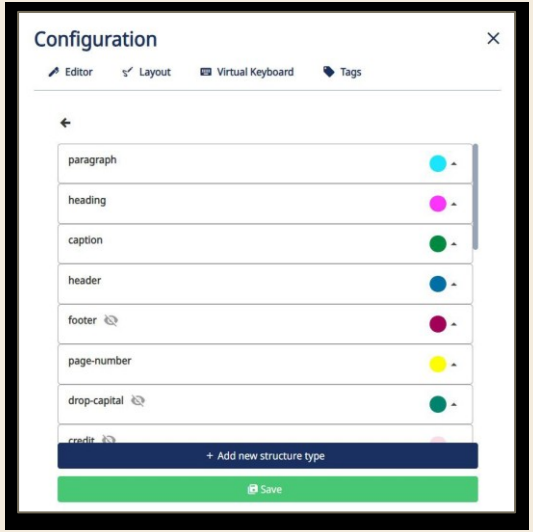
a. Štrukturálne tagy (Structural Tags):
Slúžia na označenie prvkov štruktúry dokumentu

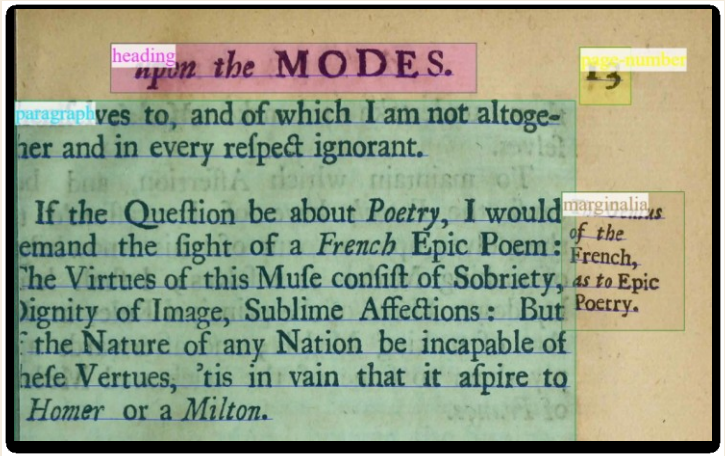
Editor dokumentov: prejdite na **Konfigurácia**
Rozloženie (Layout)

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



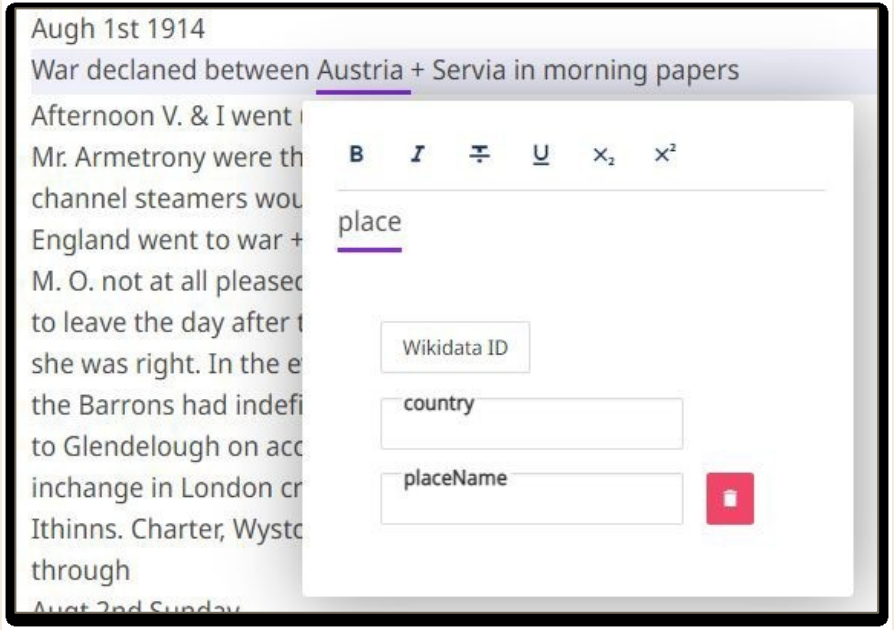


Tagovanie/značkovanie

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



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**

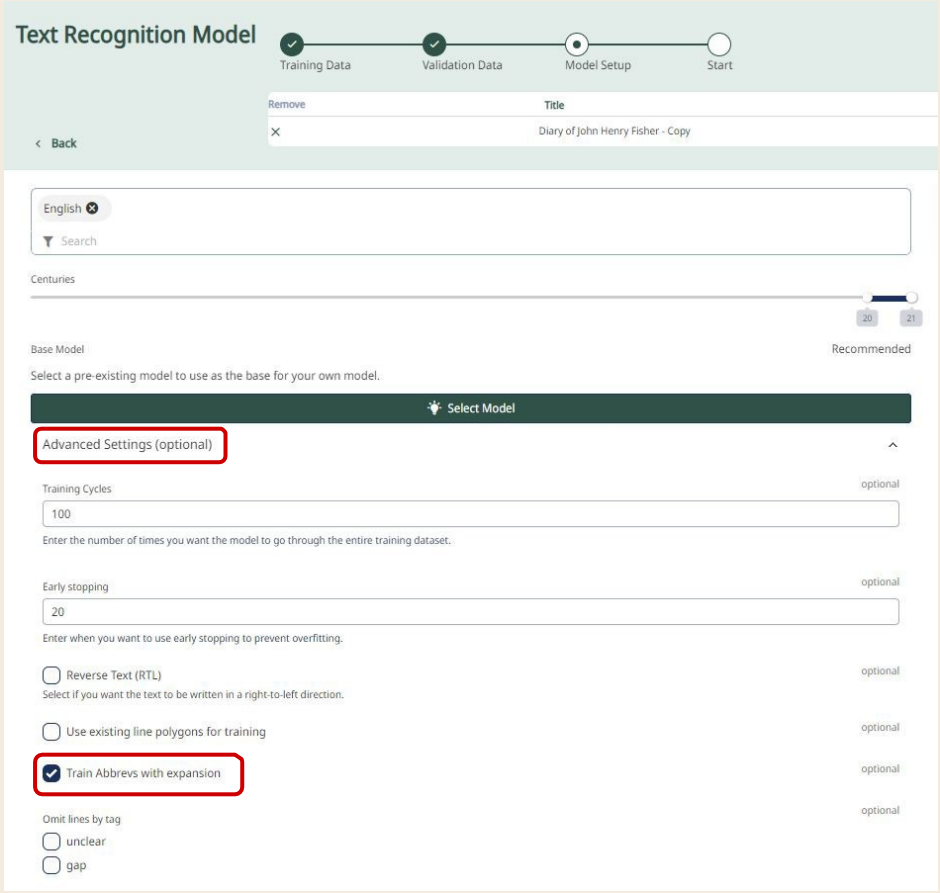
2. **Rozpisovanie skratiek**: Neurónové siete sú často schopné naučiť sa rozpoznávať a používať rozšírenia, najmä ak sa objavujú často napíšte rozšírenie skratky do prepisu, venujte dôslednú pozornosť
   Rozpisujeme skratky (pozorne, rovnako) → **output: Skratky. + rozšírenia v texte**

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

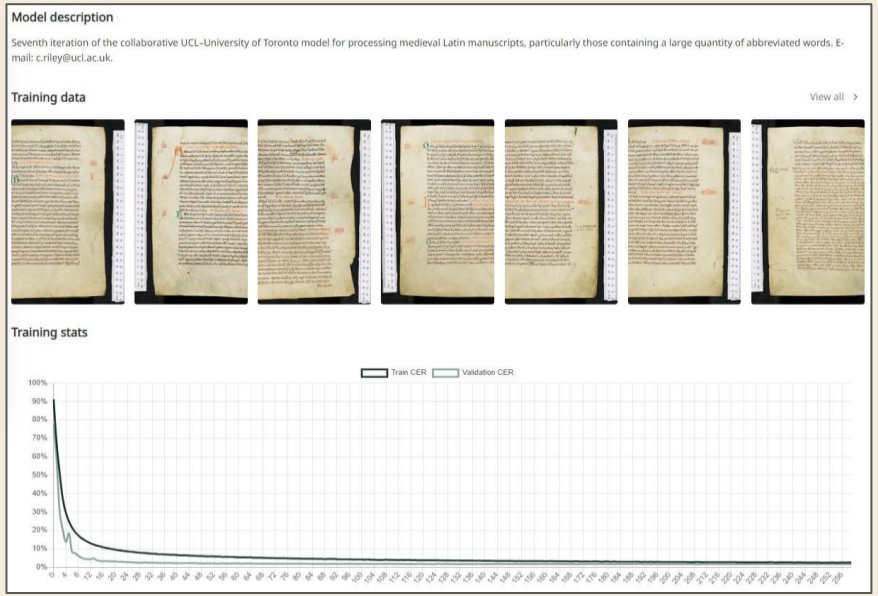
V konfigurácii tréningu začiarknite políčko

Train Abbrevs with expansion
(Trénovať model s rozpisom skratiek)



Skratky

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

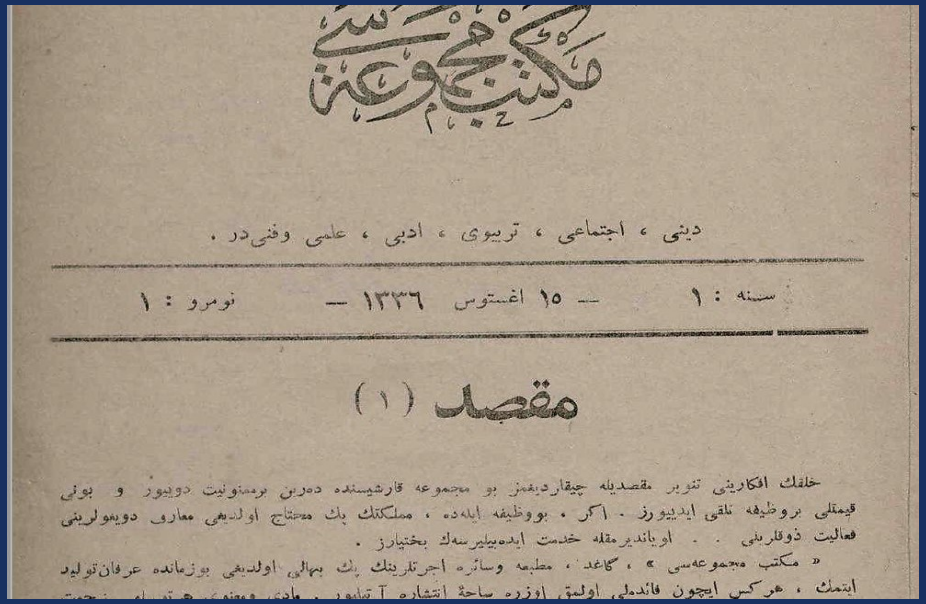


[Example](#)

Skratky

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

Example



مکتب مجموعه‌سی

دینی ، اجتماعی ، تربیوی ، ادبی ، علمی وفنی‌در .

سنه : ۱ — ۱۵ اغستوس ۱۳۳۶ — نومرو : ۱

مقصد (۱)

خلقك افكارینی تنویر مقصدیله چیقاردیغمز بو مجموعه قارشیسنده دهرین برمسئولیت دویوز و بونی قیمتلی بروظیفه تلقی ایدییورز . اگر ، بووظیفه ایده‌ه ، مملکتك بك محتاج اولدیغی معارف دویمولرینی فعالیت ذوقلرینی . . اوباندیرمقله خدمت ایده‌بیلیرسه‌ك بختیارز .

« مکتب مجموعه‌سی » ، گاهده ، مطبعه وسائره اجرتلرینك بك بهالی اولدیغی بوزمانده عرفان‌تولید ایتمك ، عرکس ایچون فائدهلی اولمق اوزره ساحهٔ انتشاره آتیلیور . مادی معنوی هر تورلو

Tréningové modely pre RTL písmo

Transkribus

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)

RTL skripty

Ako v súčasnosti prepisovať a trénovať údaje RTL v Transkribuse:

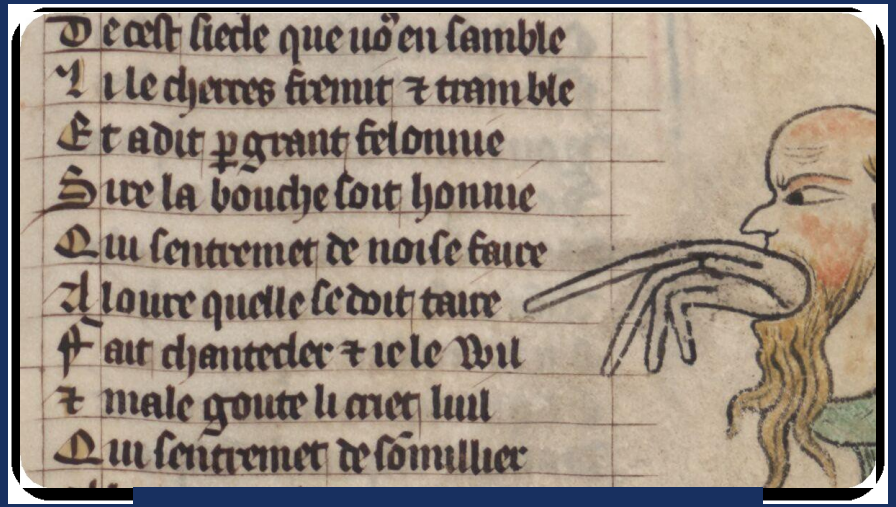
Manuálne spustenie segmentácie (rozpoznávania rozloženia) alebo označovanie rozloženia (oblasti textu + základné čiary) manuálne

- Prepis textu z **left-to-right** v textovom editore (zľava – doprava)
- 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

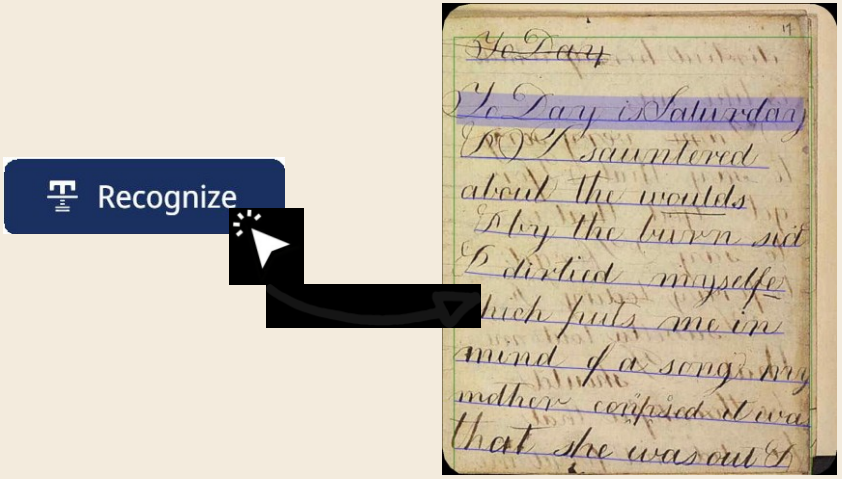


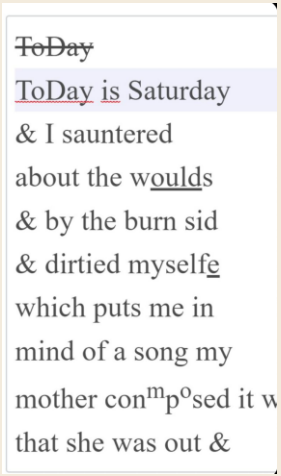
Paris, BnF, Fr. MS 12584 (13th century)

Rozpoznávanie rozloženia (Segmentácia)

Transkribus

Čo sa stane, keď sa stránka rozpozná?





1. krok

Rozpoznávanie rozloženia

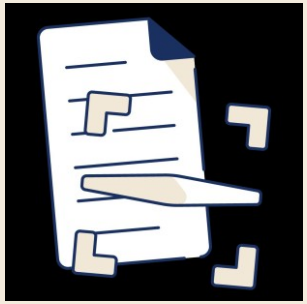
(Základné čiary (Baselines)& Bloky textu (Text regions)

2. krok

Rozpoznávanie textu

Rozpoznávanie rozloženia (segmentácia)

1. krok
Rozpoznávanie rozloženia

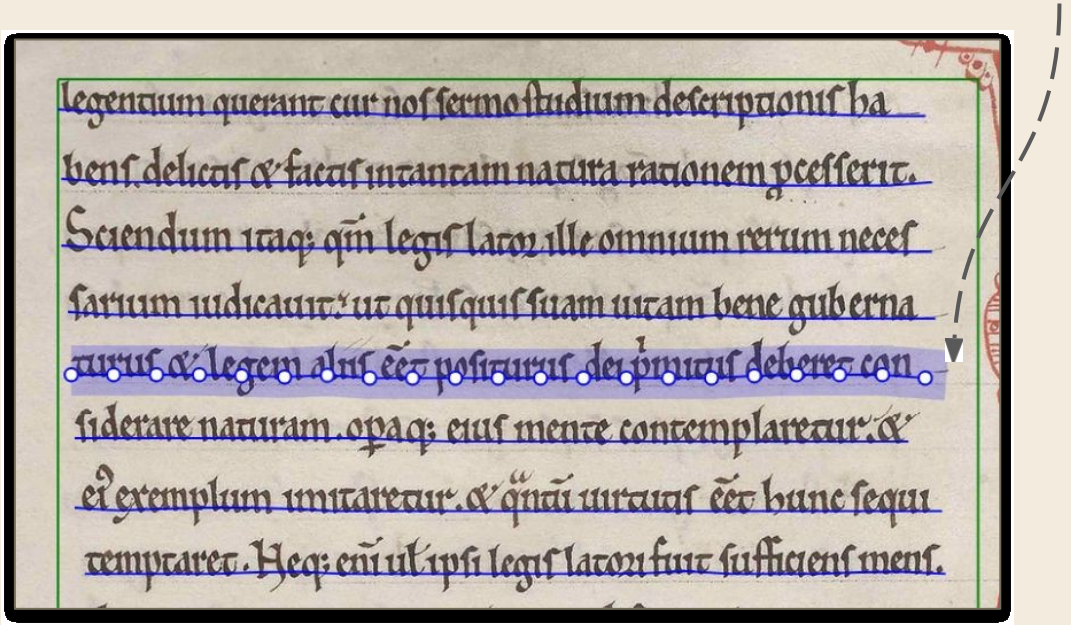


- Analýza rozloženia obrazu dokumentu
- Obrázok je potrebné rozdeliť na textové oblasti a základné čiary
- Základ pre rozpoznávanie a pre transkripciu (prepis)

Tri piliere rozloženia (segmentácie)

1) **Základná čiara (Baseline):**

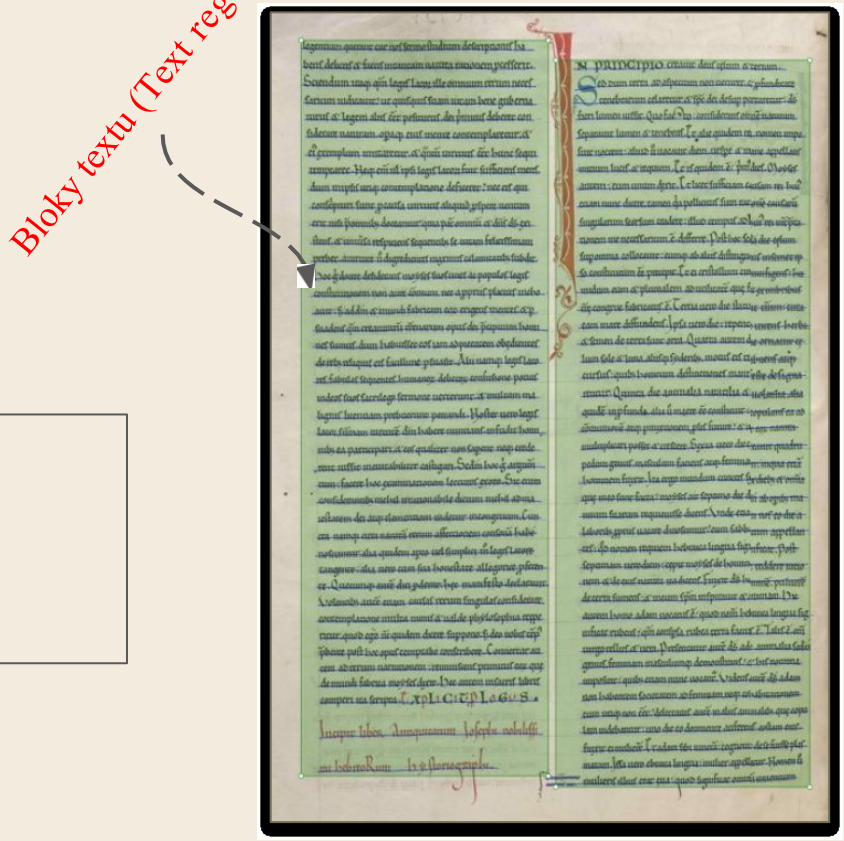
   Členená čiara prebiehajúca pozdĺž spodnej časti riadka rukou písaného textu



Tri piliere rozloženia (segmentácie)

1) **Základné čiary (Baselines)**

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



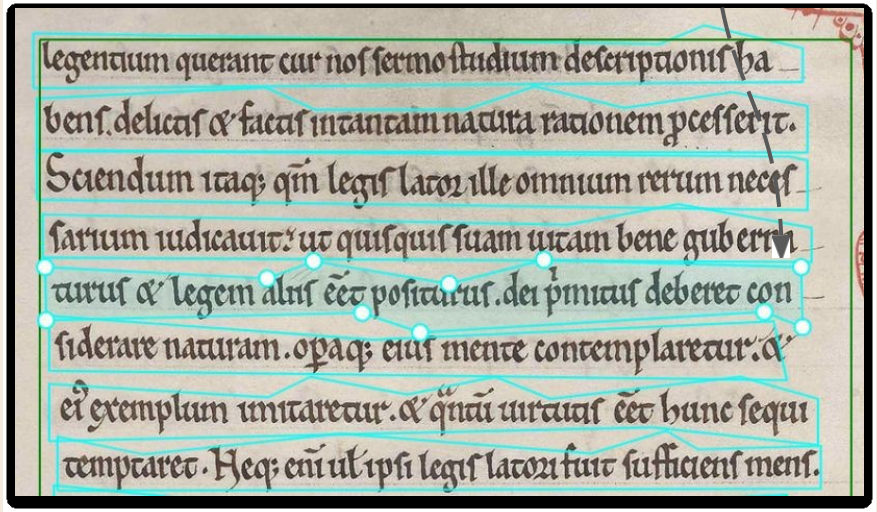
Pri predvolenej analýze rozloženia sú základné čiary zoskupené do blokov textu (textových oblastí na základe ich súradníc (prístup zdola nahor)

Tri piliere rozloženia (segmentácie)

1) Baseline

2) Text region

3) **Polygóny riadku (Line Polygons:** mnohouholníky, obklopujúce všetok rukou písaný text v riadku



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

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

Rozpoznávanie rozloženia (segmentácia)

Kvalitu konečného rozpoznania (segmentácie) môže ovplyvniť:

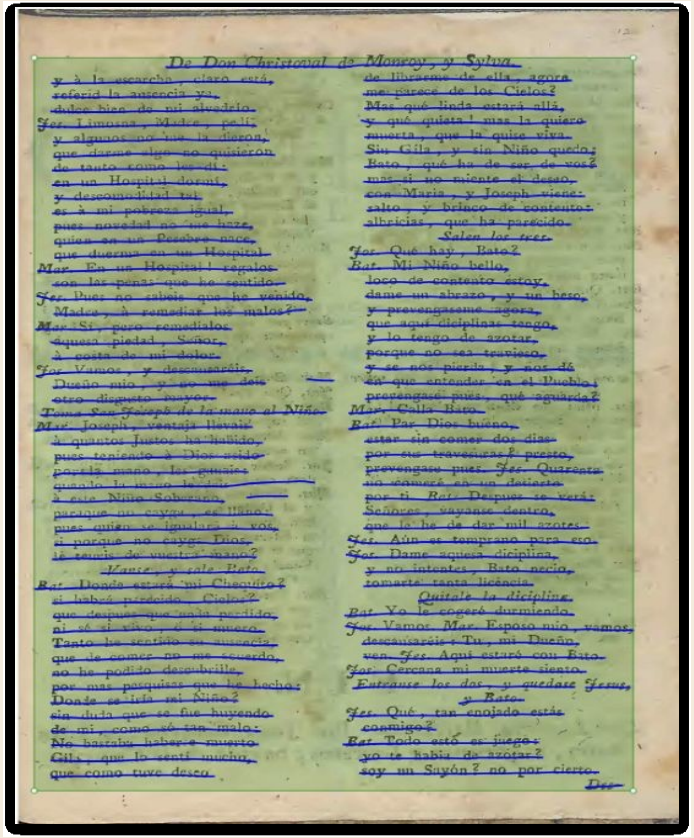
1) **Nepresné základné čiary (baselines):**
   - Zistí sa príliš málo základných čiar (východiskových hodnôt) alebo príliš veľa základných čiar (východiskových hodnôt)

Rozpoznávanie rozloženia (segmentácia)

Kvalitu konečného rozpoznania (segmentácie) môže ovplyvniť:

2) Nepresné bloky textu:

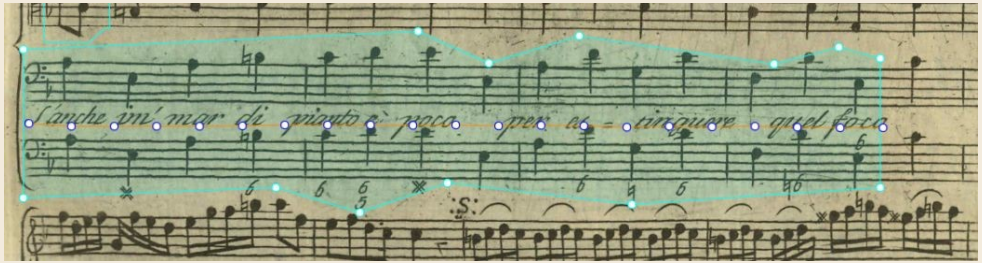
- Nesprávne poradie čítania riadkov;
- Príliš málo blokov textu/príliš veľa blokov textu (text regións)

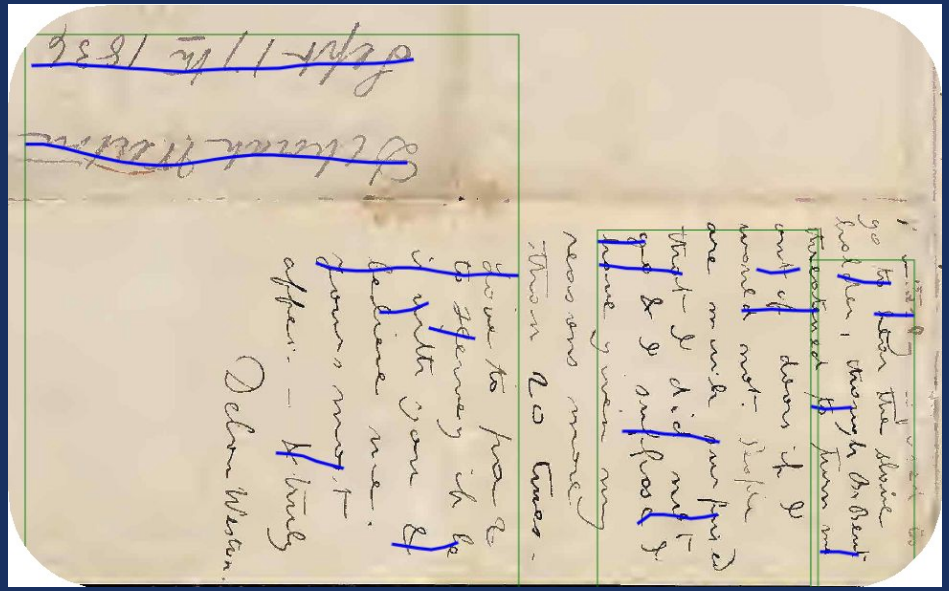


Rozpoznávanie rozloženia (segmentácia)

Kvalitu konečného rozpoznania (segmentácie) môže ovplyvniť:

3) **Nepresné polygóny (Inaccurate polygons):**
   - Aj keď sú základné čiary správne, modely nedokážu správne prepísať text.
   - Riadkové mnohouholníky nepokrývajú väčšinu tela písmen/ Polygóny čiar zahŕňajú aj ďalšie (neželané) prvky na strane





Nepresné základné čiary

Transkribus

Nepresné základné čiary

Morgen-Ausgabe

Berliner Tageblatt

und Handels-Zeitung

Chamberlain über China.

China auf dem Wege zur Freiheit.

Example

Nepresné základné čiary – čo robiť?

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

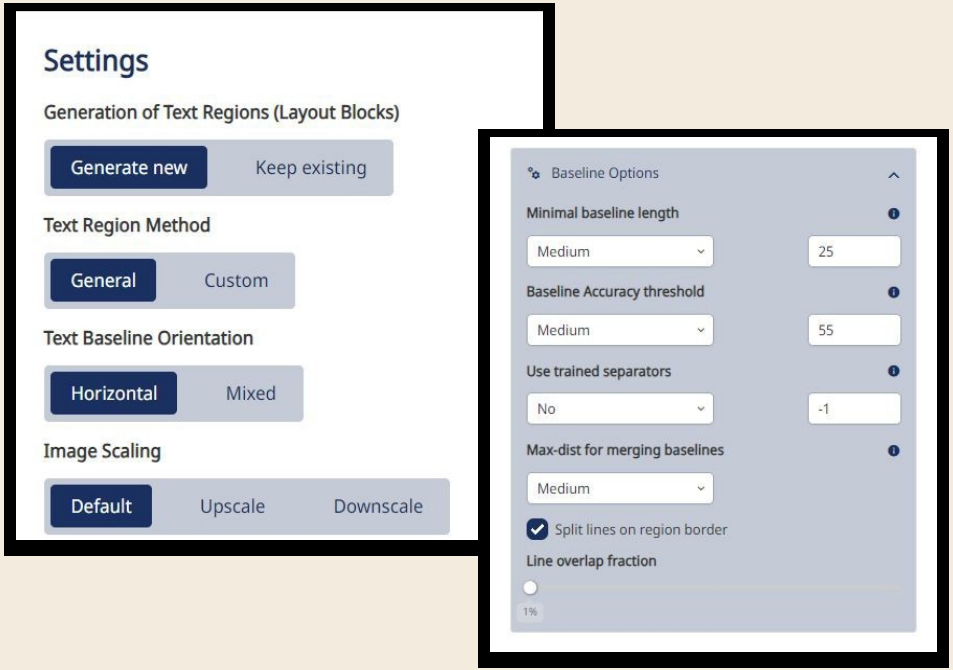
1) **Použitie iného verejného modelu základných čiar (Baseline model)**

:

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

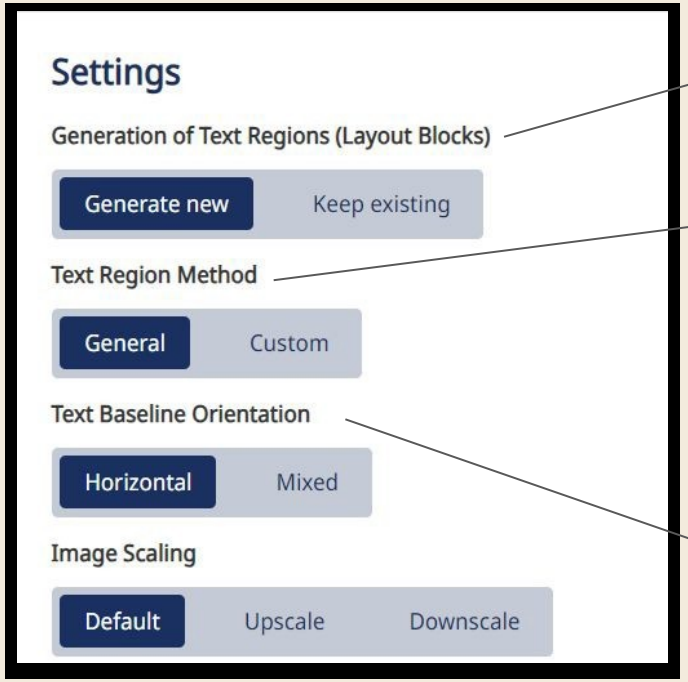
Nepresné základné čiary

2) **Zmeňte pokročilé nastavenia (advanced settings)**



Nepresné základné čiary

2) **Zmeňte pokročilé nastavenia (advanced settings)**



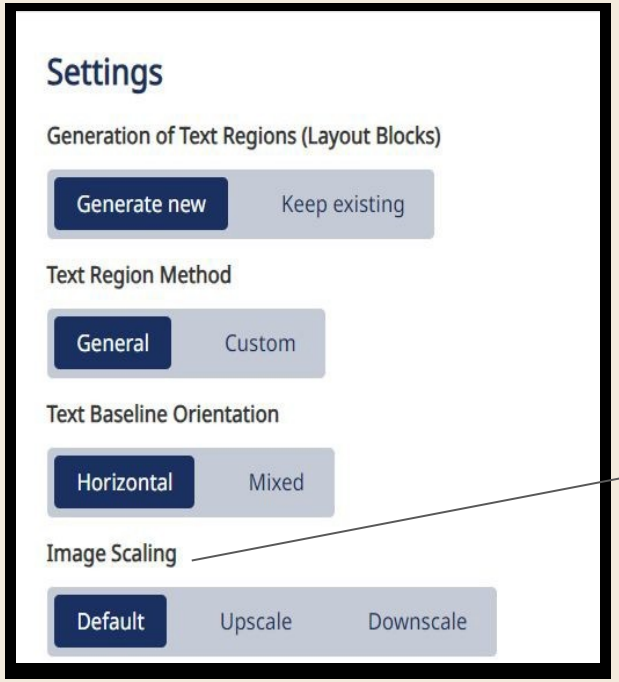
Generate new: Generovať ďalšie textové oblasti /
Keep existing: Zachovať existujúce oblasti textu (použite to s poľami a tabuľkami)

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

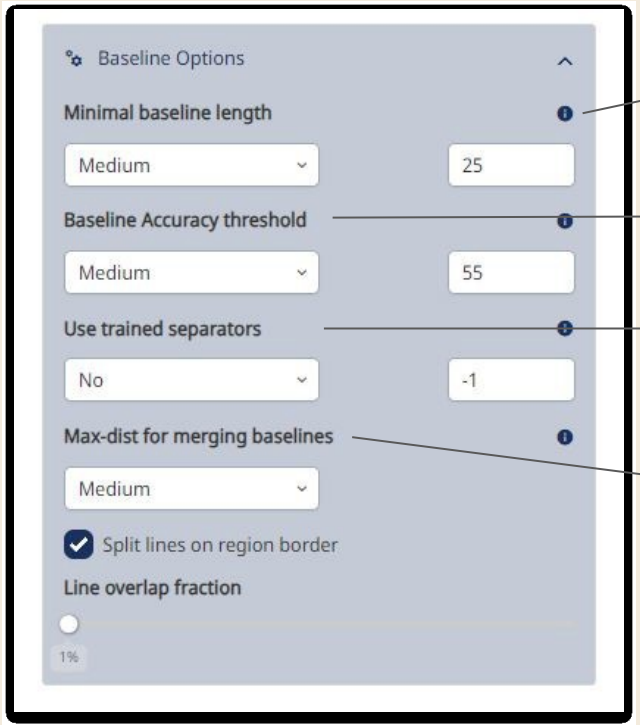
2) **Zmeňte pokročilé nastavenia (advanced settings)**



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 treined separators) Ak zvýšite túto hodnotu, okolité čiary sa zvyčajne zlučujú

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

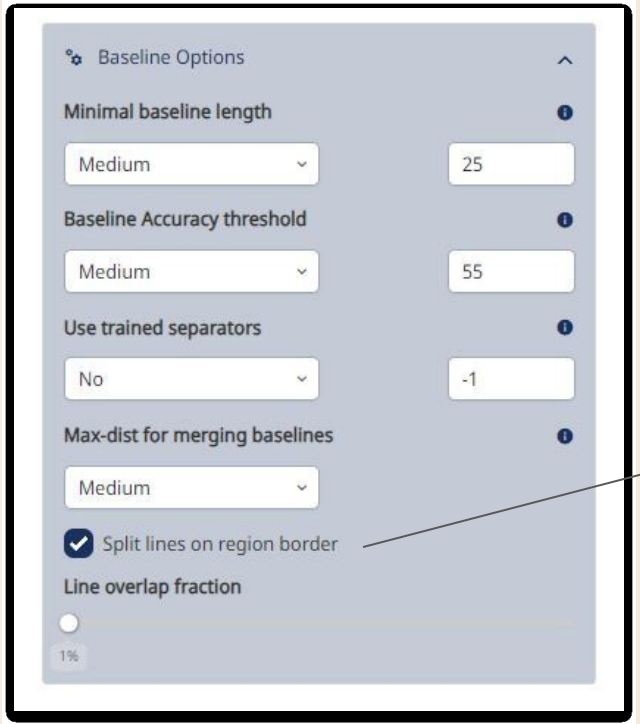
Low: Zlúčia sa iba najbližšie čiary

Medium

High: vzdialené základné čiary sa zlúčia

Nepresné základné čiary

2) **Zmeňte pokročilé nastavenia (advanced settings**

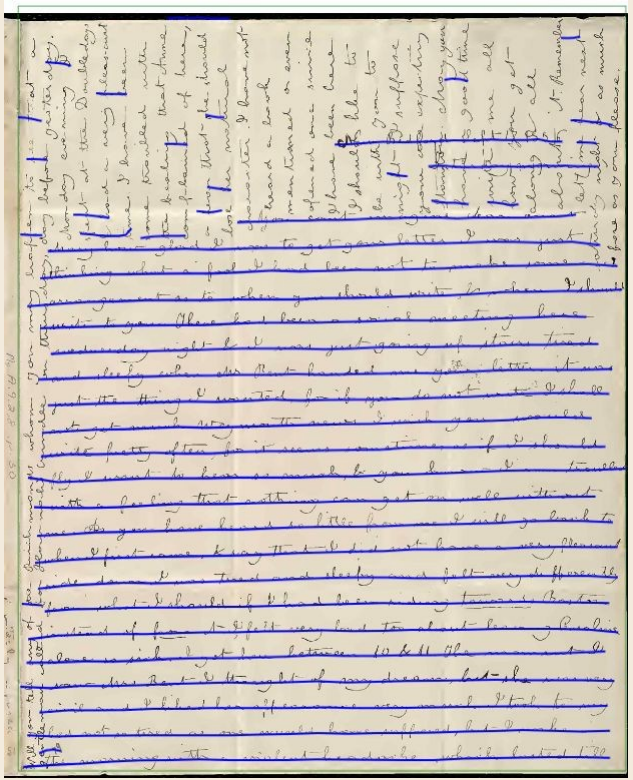


Rozdeliť čiary v rámci bloku (Split lines on region border)
Iba ak zachováte existujúce bloky textu:
Delené čiary na hranici regiónu:
Aby čiary striktne dodržiavali hranicu regiónu.
Dôležité pre tabuľky!

Nepresné základné čiary

Example 1

Example 2

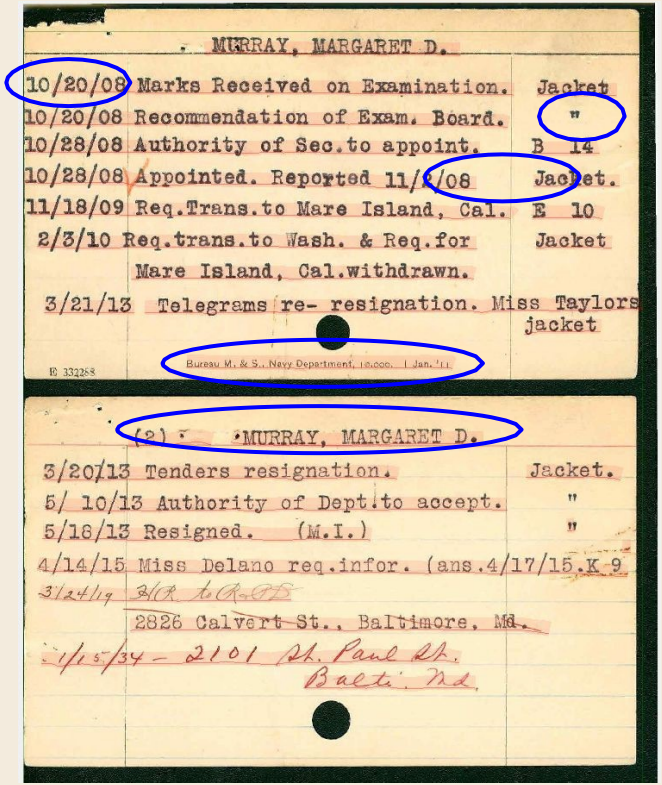


Nepresné základné čiary

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.

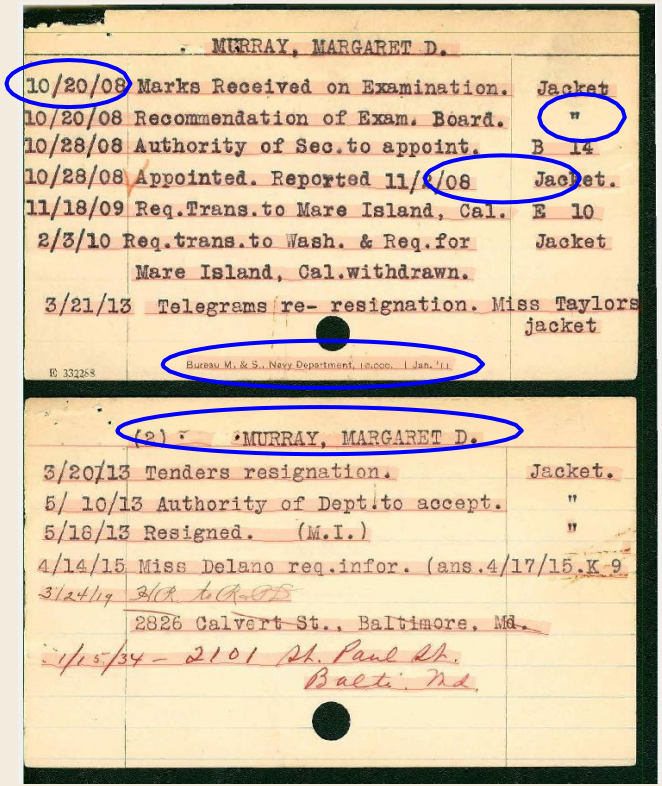
10/20/08 Marks Received on Examination. Jacket
10/20/08 Recommendation of Exam. Board. "
10/28/08 Authority of Sec.to appoint. B 14
10/28/08 Appointed. Reported 11/[illegible]/08 Jacket.
11/18/09 Req.Trans.to Mare Island, Cal. E 10
2/3/10 Req.trans.to Wash. & Req.for Jacket
Mare Island, Cal.withdrawn.
3/21/13 Telegrams re- resignation. Miss Taylors jacket

E 332268 Bureau M. & S., Navy Department, 10,000, 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 3/R to C-O-T
2826 Calvert St., Baltimore, Md.
1/15/34 - 2101 St. Paul St.
Balti. Md

Tréning modelu základných čiar (Baseline Model)



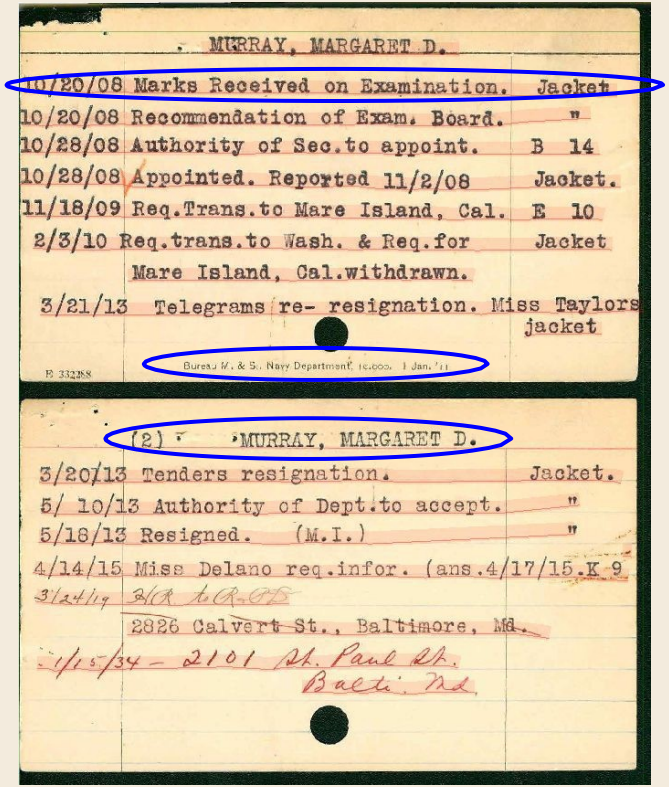
MURRAY, MARGARET D.

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

E 332288 Bureau M. & 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.	"
5/18/13	Resigned. (M.I.)	"
4/14/15	Miss Delano req.infor. (ans.4/17/15.K 9	
3/24/19	B/R to R.-C.P.S.	
	2826 Calvert St., Baltimore, Md.	
1/15/34	- 2101 St. Paul St. Balti. Md	



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

E 332288 Bureau M. & 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.	"
5/18/13	Resigned. (M.I.)	"
4/14/15	Miss Delano req.infor. (ans.4/17/15.K 9	
3/24/19	B/R to R.-C.P.S.	
	2826 Calvert St., Baltimore, Md.	
1/15/34	- 2101 St. Paul St. Balti. Md	

Tréning modelu základných čiar (Baseline Model)

Pripravte si aspoň 50 strán GT so správnymi základnými čiarami:

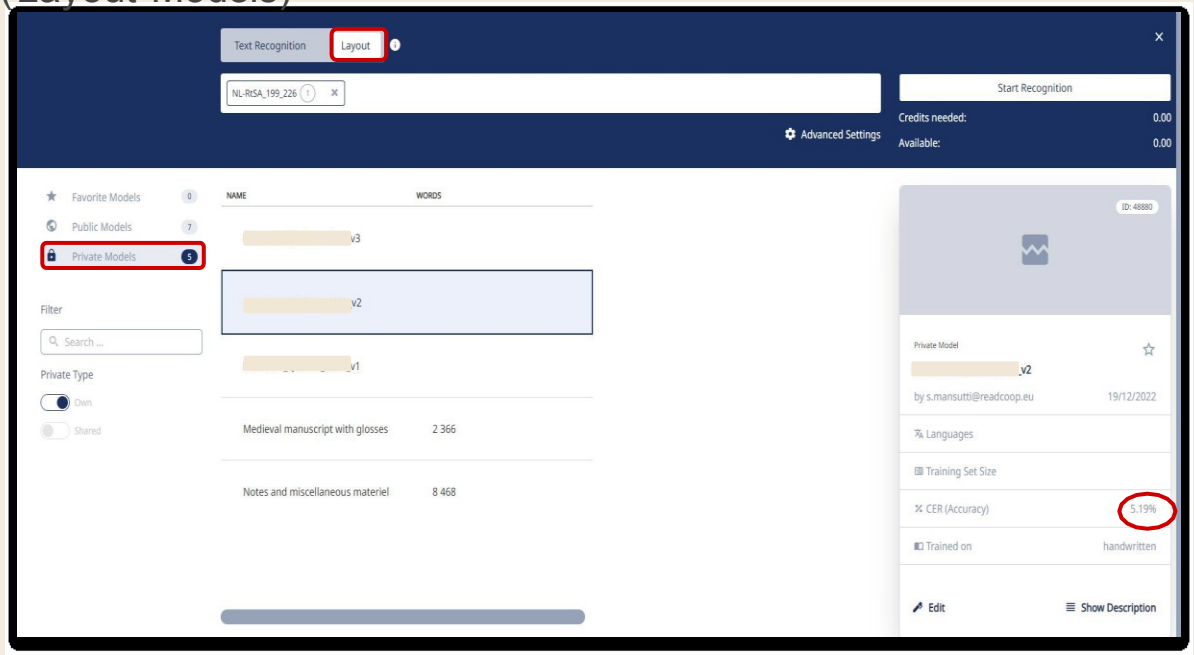
- Nakreslite všetky základné čiary manuálne alebo opravte automatické rozpoznávanie rozloženia
- Nakreslite základné čiary iba pre časti, ktoré chcete prepísať

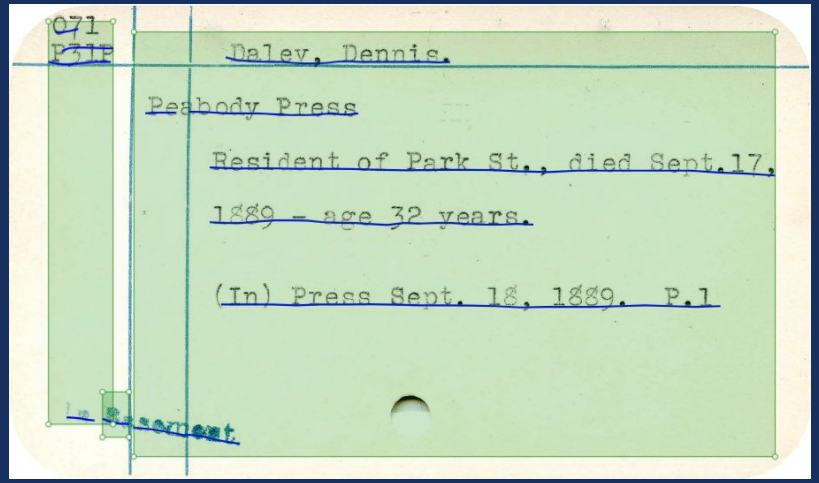
Tréning modelu základných čiar (Baseline Model)

- 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)





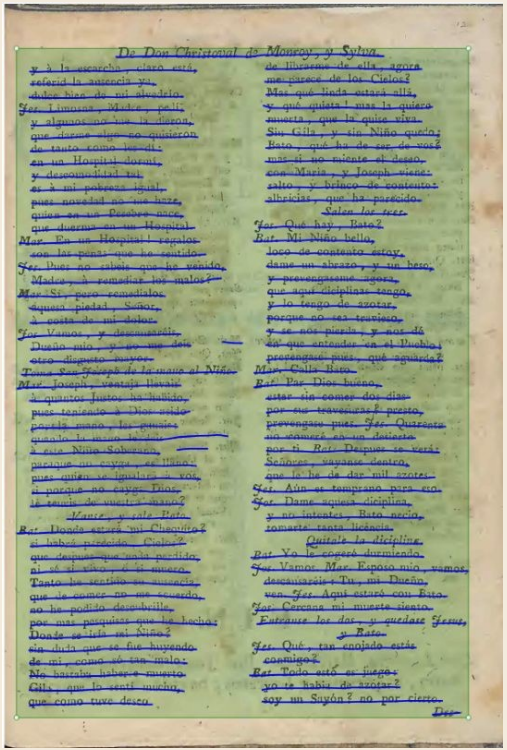
Nepresné bloky textu

Transkribus

Nepresné bloky textu

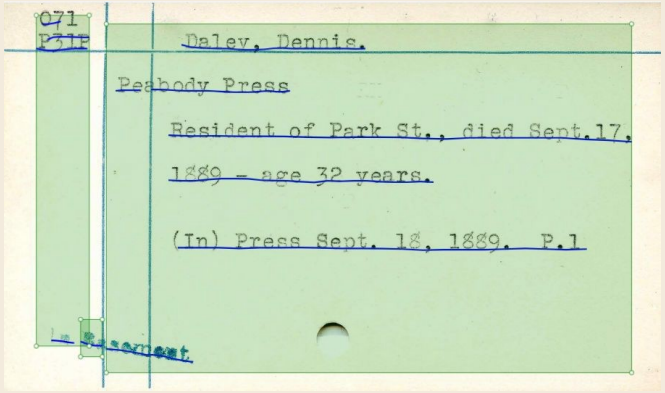
[Example 1]

[Example 2]



De Don Christoval de Monroy y Silva

y à la verdad, claro está,
referid la verdad ya,
dadme hoy de mi olvidado.
Yer. Limosna, Madre, pedí,
y algunos no me la dieron,
que duerme aquí no quisieron
de tanto como les dí:
en un Hospital dormí,
y descomunidad tal
es à mi pobreza igual,
pues novedad no me hizo,
quise en un Pesebre arroz,
que duerma en un Hospital.
Mar. En un Hospital! regalos
son los penas que he sentido.
Yer. Pues no sabeis que he venido,
Madre, à remediar los malos?
Mar. Sí, pero remedialos
alguna piedad, Señor,
à costa de mi dolor.
Yer. Vamos, y descansareis.
Dueño mio, y esposo mio,
este disgusto tuyo es
Para San Joseph de la mano el Niño.
Mar. Joseph, ventais llorais
à cuantos Justos ha habido
pues teniendo à Dios asido
perdia mano, los pesares
quando la mano le dais
à este Niño Soberano,
porque un cuyo se llama?
pues quien se igualara à vos,
ni porque no elijo Dios,
le tendrá de suerte el llanto?
Mares, si esta Pato
Bat. Donde estará mi Chequito?
si habrá perdido Cielos?
que despues que nada perdido,
ni el se ha ido, ni le quiero,
Tanto he sentido en ausencia,
que se comer no me acuerdo,
no he podido descubrille,
por mas pesquisas que he hechas:
Donde se habrá ido Niño?
sin duda que se fue huyendo
de mi, como só una malo:
No hacaba haber muerto
Gila, que lo sentí mucho,
que como tuve deseo
de librarme de ella, agora
me parece de los Cielos?
Mas qué linda estará allá,
y qué quieta! mas la quiero
muerta, que la quise viva.
Sin Gila, y sin Niño quedo;
Bato, qué ha de ser de vos?
más si no siente el deseo,
con Maria, y Joseph viene:
salto, y brinco de contentos
albricias, que ha parecido.
Salen los tres.
Yer. Que hay, Bato?
Bat. Mi Niño bello,
loco de contento estoy,
dame un abrazo, y un beso,
y prevenganse vuestros,
que aquí Discipulos tengo,
y lo tengo de azotar,
porque no sea travieso,
y se nos pierda, y por dó
sin que entiendan en el Pueblo,
prevenganse pues, qué aguardais?
Mar. Calla Bato.
Bat. Por Dios bueno,
estar dia comer dos dias
por sus travesuras? presto,
prevenganse pues. Yer. Quereis
no comeis en un dia?
por ti. Bat. Despues se verá:
Señores, vayanse dentro,
que lo he de dar mil azotes.
Yer. Aún es temprano para eso.
Yer. Dame aquesa disciplina,
y no intentes, Bato necio,
tomarte tanta licencia.
Quitale la disciplina.
Bat. Yo le cogeré durmiendo.
Yer. Vamos, Mar. Esposo mio, vamos,
descansareis: Tu, mi Dueño,
ven. Yer. Aquí estaré con Bato.
Yer. Cercana mi muerte siento.
Entranse los dos, y quedase Yesus, y Bato.
Yer. Qué, tan enojado estás
conmigo?
Bat. Todo esto es juego,
yo te habia de azotar?
soy un Sayón? no por cierto
Yer.



071
P31P

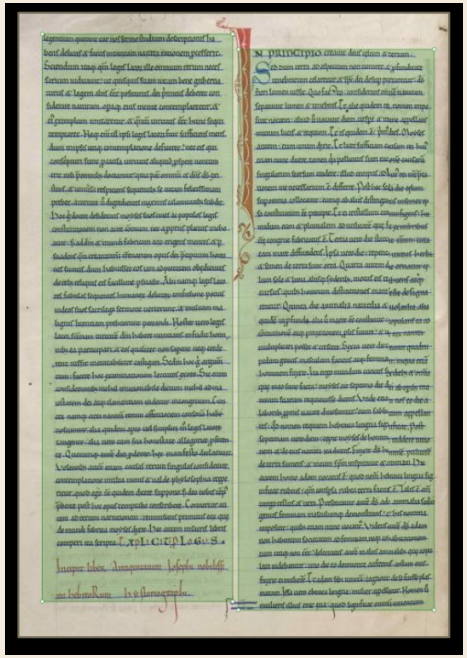
Daley, Dennis.

Peabody Press

Resident of Park St., died Sept.17, 1889 - age 32 years.

(In) Press Sept. 18, 1889. P.1

Analýzy rozloženie/segmentácia (rozpoznanie textu)

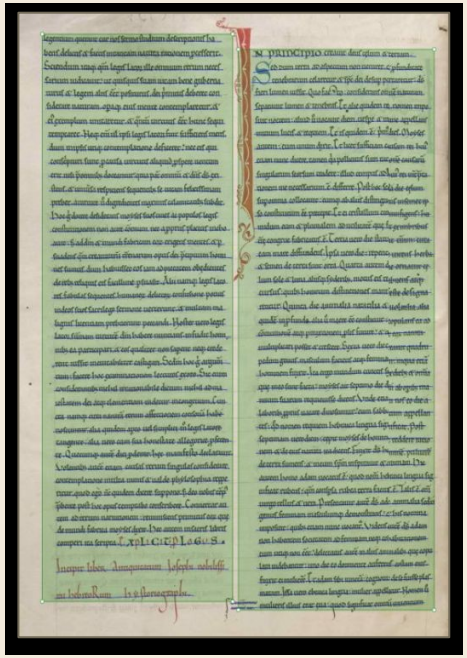


Bloky textu:

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 v momente *rozpoznávania textu* (Text Recognition)

Analýzy rozloženie/segmentácia (rozpoznanie textu)

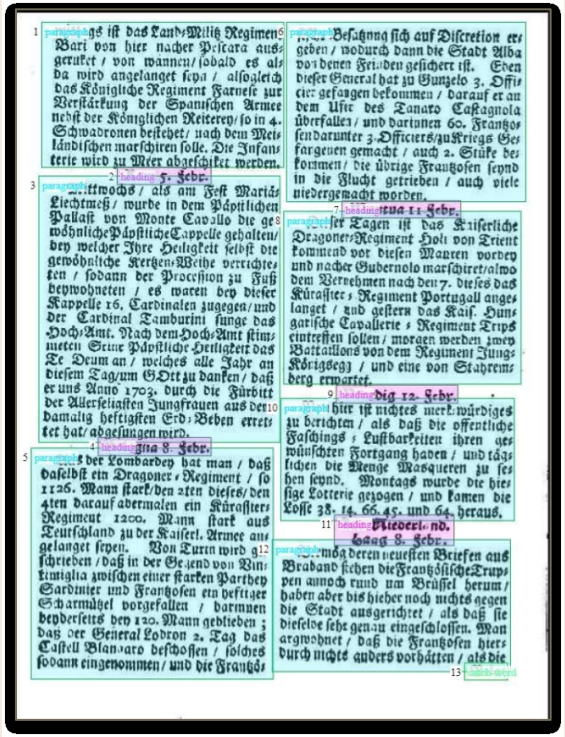


Bloky textu:

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

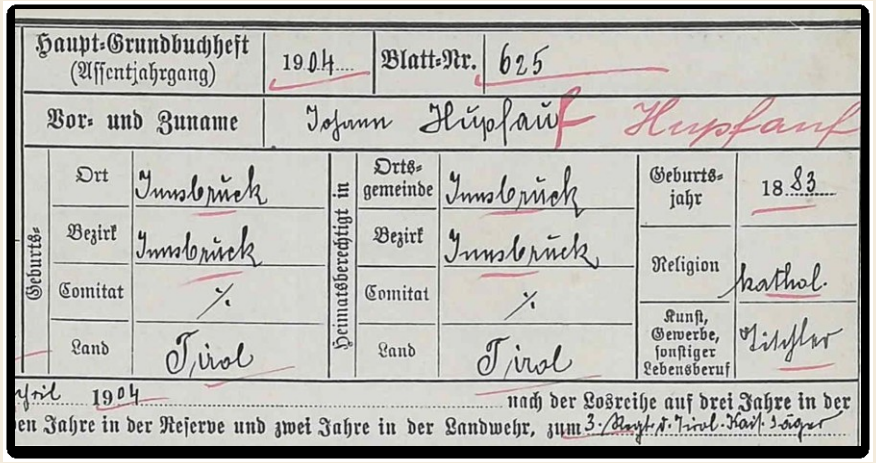
1. Rozpoznávanie blokov textu pomocou **Modelu poľa** (**Field Model**): *polia sú v blokoch stránky*
2. Rozpoznanie základných čiar (Layout Recognition)

1. *Základné čiary a polygóny* sa tvoria v v momente *rozpoznávania textu* (Text Recognition)

Modely poľa

Transkribus

Modely pol'a (Beta)



Haupt-Grundbuchheft (Assentjahrgang)	1904	Blatt-Nr.	625
Vor- und Zuname	Johann Hüpfauf Hupfauf		

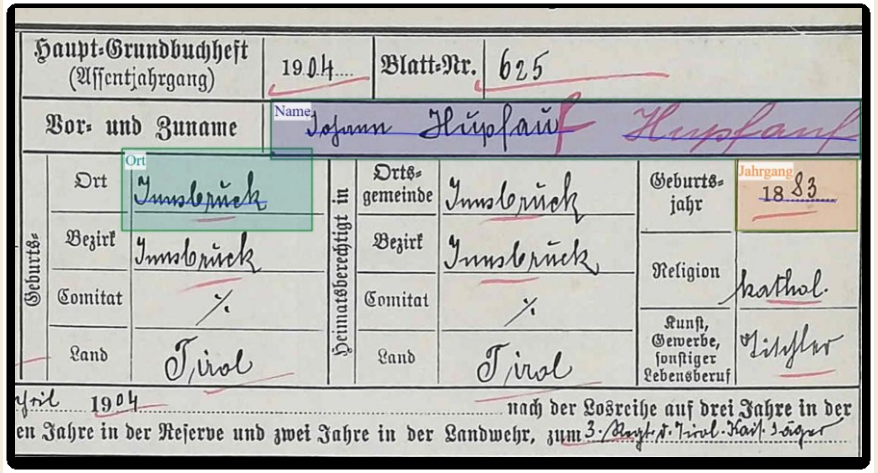
Geburts-		Heimatsberechtigt in			
Ort	Innsbruck	Ortsgemeinde	Innsbruck	Geburtsjahr	1883
Bezirk	Innsbruck	Bezirk	Innsbruck	Religion	kathol.
Comitat	/	Comitat	/	Kunst, Gewerbe, sonstiger Lebensberuf	Tischler
Land	Tirol	Land	Tirol		

Juli 1904 nach der Losreihe auf drei Jahre in der
en Jahre in der Reserve und zwei Jahre in der Landwehr, zum 3./Bayr. J. Tirol. Kais. Jäger

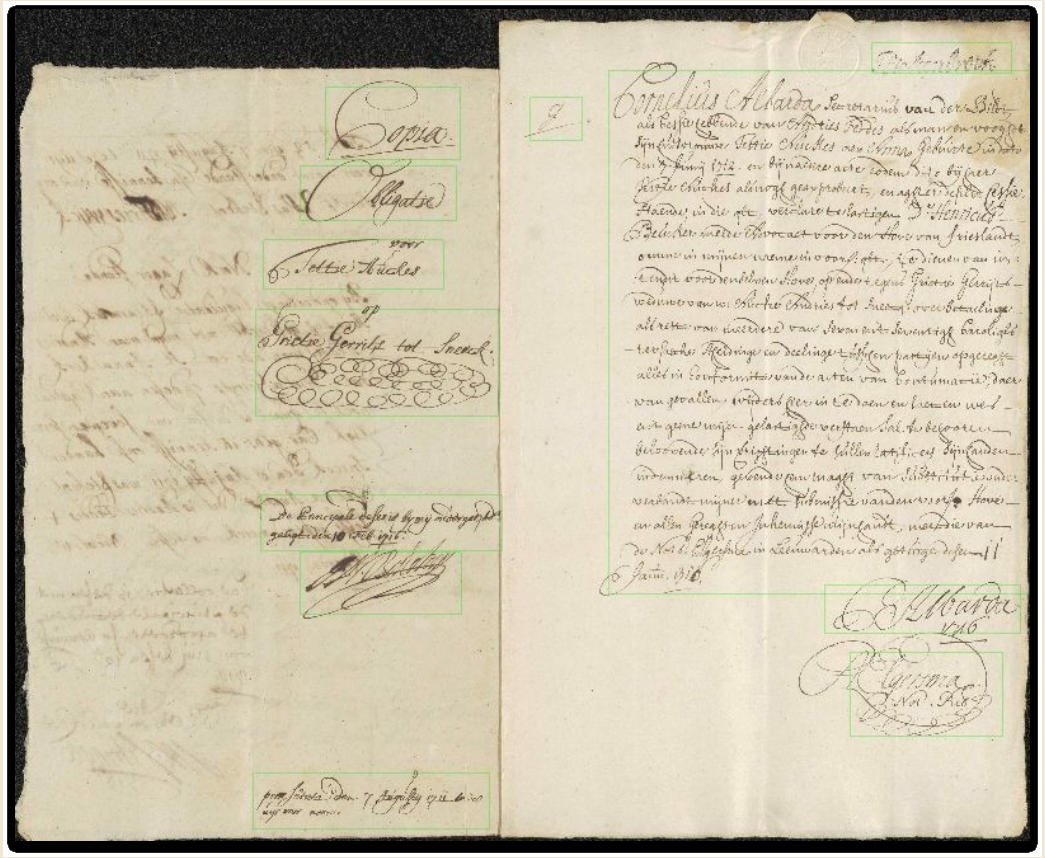
Modely poľa (Beta)

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

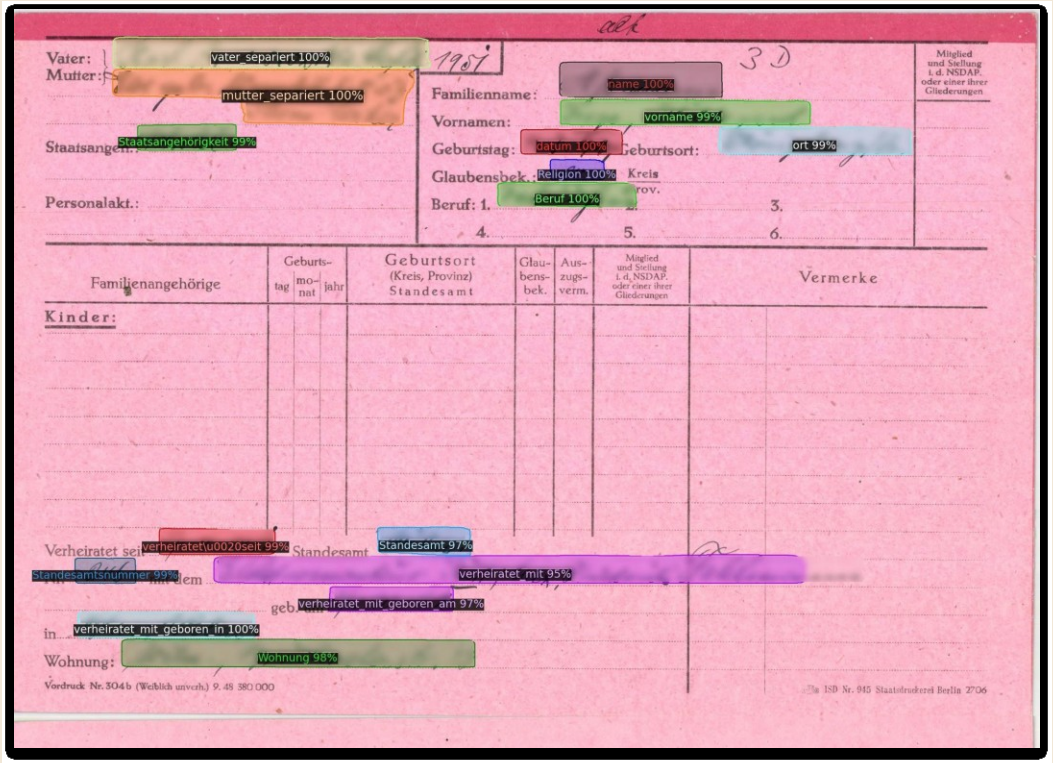


Blogy textu (Text regions)

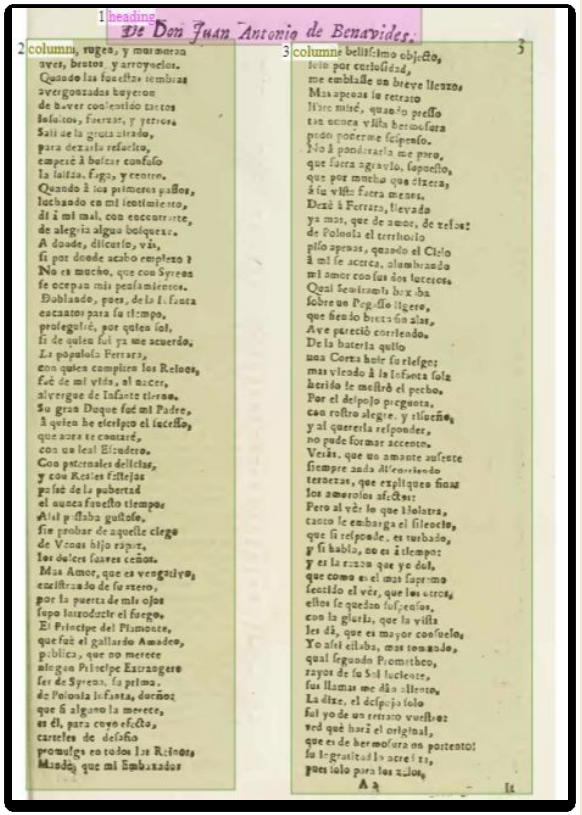


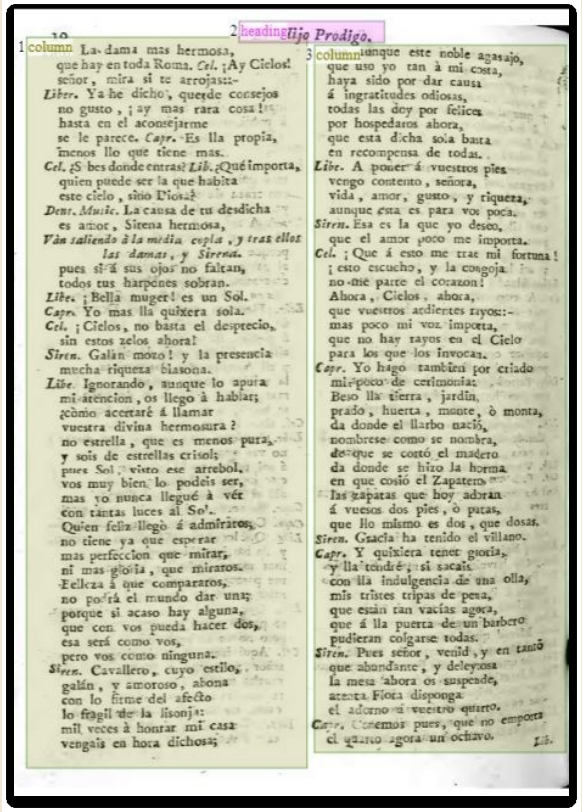
Noviny: Segmentácia rozloženia

Segmentácia formulára



Viac stĺpcové rozloženie textu

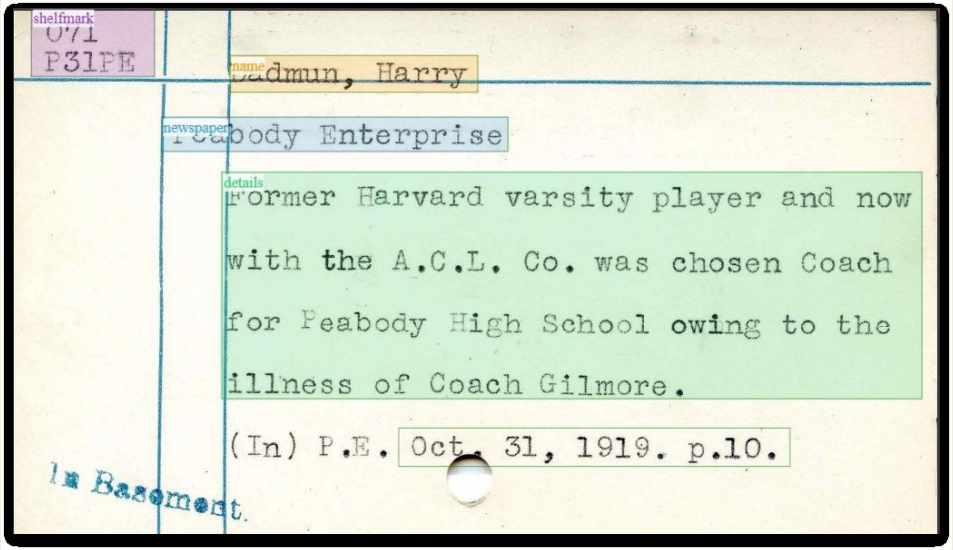




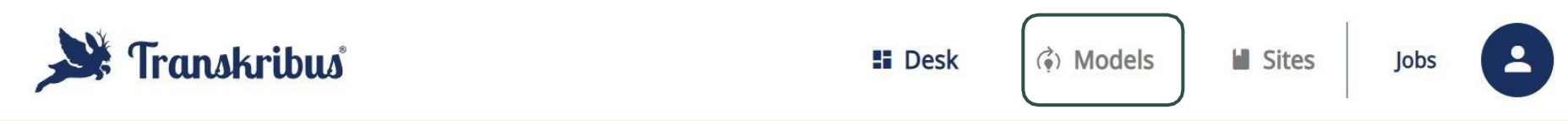
Modely poľa (Beta)

Pripravte si cca 50 strán tréningových dát:

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



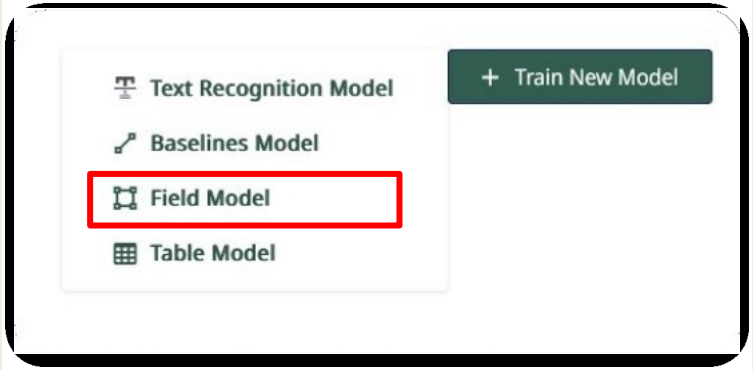
Modely poľa (Beta)



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**

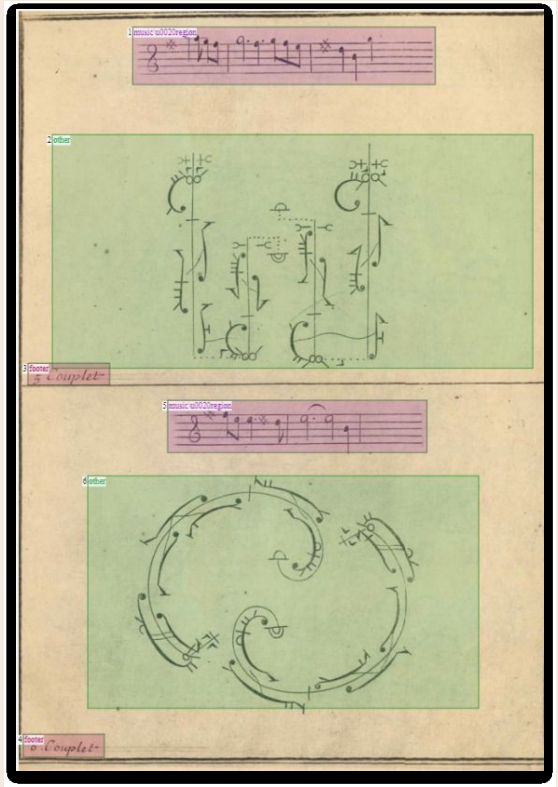
Spracovanie dokumentov s poľami

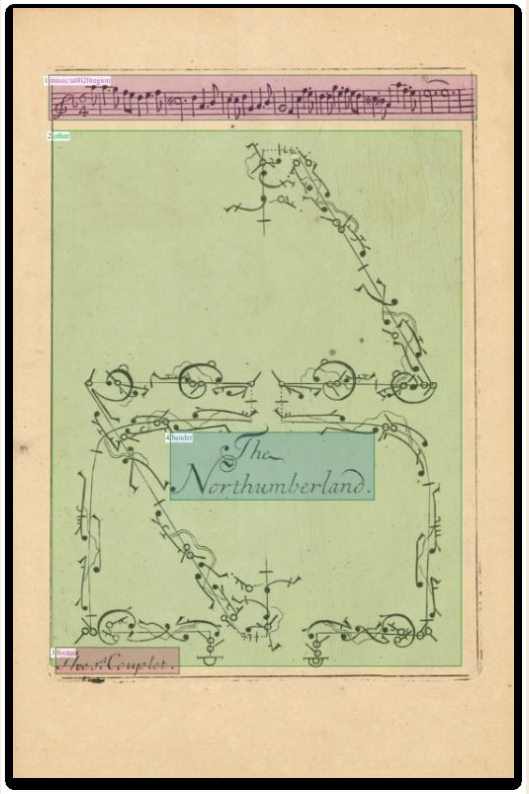
	A	B	C	D	E	F
1	TranskribusFilename	shelfmark	name	newspaper	details	reference
2	00729.jpg	071 D218	Dynan, Mary E.	Salem Evening News	Received diploma in October, 1910 from N.E. Institute of Anatomy, San- itary Science and Embalming. Was first Peabody girl to graduate as an embalmer, also the youngest 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.	Sallem Evening News	Who died at his home, 30 Chestnut St. was a baker b y trade and an active member of organized labor. He was em¬ ployed 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

Anno 1746. (Num. 17.) 26. Februarii.

Wienerisches

DIARIUM.

Ihrer Römisch-Kaiserl. auch zu Hungarn und Böheim Königl. Maj. Freyheit.

In dem neuen Michaeler-Haus / bey Joh. Peter v. Ghelen.

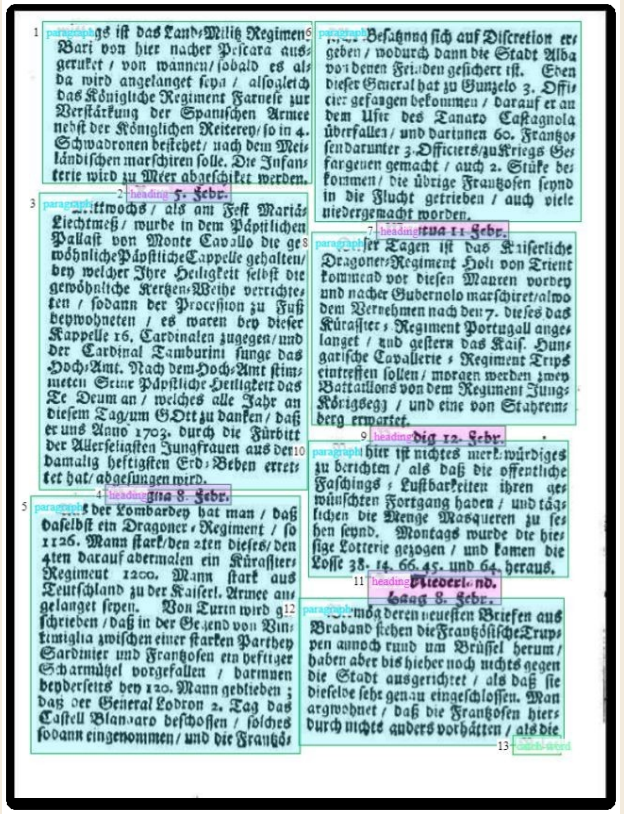
Aus Italien.

Genua 29. Jan.

Ergangenen Samstag kamen abermalen zwey Catalonische Schiffe mit 360. Spanischen Soldaten von Villa franca / und Sonntags Abends eine unserige Feluke von Calvi alhier an / von wannen sie den 19. mit Staats-Briefen ab / und nacher Livorno gegangen / und da sie auch von diesen letzteren gleich wieder abgegangen / hat selbe den 21. fruhe in dasigen Gegenden ein Englandisches Kriegs-Schif mit vier ekigten Flaggen sich ankern gesehen: Briefe von Barcellona unter dem 15. dieses geben / daß daselbst nach und nach immerhin Truppen und Fahr-Zeuge ankommen / welche jederzeit gar bald nach hiesigen Gegenden befördert werden; es sollen auch würklich einige Cavallerie-Regimenter zur Verstärkung der Spanischen Armee allschon nach unseren Gräntzen im würklichen Anmarsch seyn. Seit einigen Tagen hat man angefangen die von Savona nebst vielen anderen Kriegs Geräte hieher gekommene Artillerie zu der Spanischen Armee nach der Lombardey abzuschicken. So seynd auch Dienstags 4. Sambechin-Schiffe / und 5. andere Neapolitanische Fahr-Zeuge in 11. Tagen von Gaetta mit Artillerie und andererley Kriegs-Vorrat hier angekommen; Auch seynd auf einer Französischer Tartane 150. Mann von dem Frantzösischen Regiment Lothringen von Villa franca hier angelangt / von welchem Ort selbe zugleich mit 3. anderen mit derley Truppen beladenen hiesigen bestimmten Schiffen abgesegelet. Eine Livornische Pinke / so in 12. Tagen von Marsilien gekommen / hat die vornehmsten Officiers des Regiments Modena alhier an das Land gesetzet. Zwischen Dimerstag und gestern kamen 3. Catalonische Schiffe mit 400. Recruten für zerschiedene Spanische Regimenter nebst ihren Officieren und mit 1500. Stuk Kugeln alhier an.

Neapel 1. Febr.

Bey Gelegenheit da dieser Tagen beyde Königliche Majestäten zu Portici zu Mittag gespeiset / wurde unter dessen dem Russischen Reichs-Vice-Cantzler Grafen von Woronzow nebst seiner Frauen Gemahlin der gantze hiesige Königl. Pallast / samt allen Kleinodien / wie auch denen Königlichen Vermählungs-Kutschen / und alle übrige kostbare Einrichtung gezeiget. Vergangene Woche ist eine hiesige wohlausgerüstete Galeotte mit 15000. Ducaten zu Bezahlung unserer Truppen in der Lombardey von hier nacher Genua abgeschicket worden. Gestern Vor-



...ags ist das Land-Miliz Regimen... Bari von hier nacher Pescara ausgeruket / von wannen / sobald es alda wird angelanget seyn / alsogleich das Königliche Regiment Farnese zur Verstärkung der Spanischen Armee nebst der Königlichen Reiterey / so in 4. Schwadronen bestehet / nach dem Meiländischen marschiren solle. Die Infanterie wird zu Meer abgeschiket werden.

...ittwochs / als am Fest Mariä Liechtmeß / wurde in dem Päpstlichen Pallast von Monte Cavallo die gewöhnliche Päpstliche Capelle gehalten / bey welcher Ihre Heiligkeit selbst die gewöhnliche Kertzen-Weihe verrichteten / sodann der Procession zu Fuß beywohneten / es waren bey dieser Kappelle 16. Cardinalen zugegen / und der Cardinal Tamburini sunge das Hoch-Amt. Nach dem Hoch-Amt stimmeten Seine Päpstliche Heiligkeit das Te Deum an / welches alle Jahr an diesem Tag um GOtt zu danken / daß er uns Anno 1703. durch die Fürbitt der Allerseligsten Jungfrauen aus den damalig heftigsten Erd-Beben errettet hat / abgesungen wird.

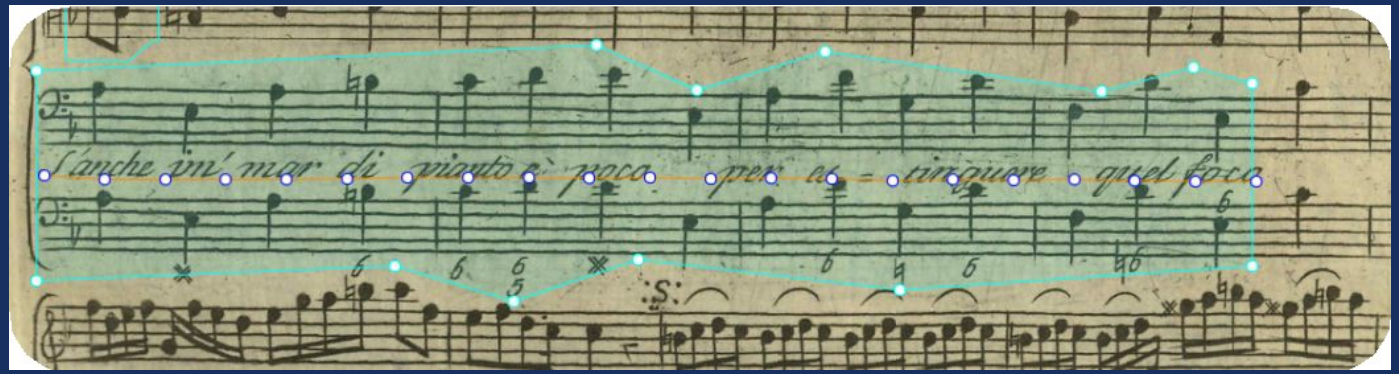
...s der Lombardey hat man / daß daselbst ein Dragoner-Regiment / so 1126. Mann stark / den 2ten dieses / den 4ten darauf abermalen ein Küraßier-Regiment 1200. Mann stark aus Teutschland zu der Kaiserl. Armee angelanget seyen. Von Turin wird geschrieben / daß in der Gegend von Vintimiglia zwischen einer starken Parthey Sardinier und Frantzosen ein heffitger Scharmützel vorgefallen / darinnen beyderseits bey 120. Mann geblieben; daß der General Lodron 2. Tag das Castell Blanzaro beschossen / solches sodann eingenommen / und die Frantzö-

...Besatzung sich auf Discretion ergeben / wodurch dann die Stadt Alba vor denen Feinden gesichert ist. Eben dieser General hat zu Gunzelo 3. Officier gefangen bekommen / darauf er an dem Ufer des Tanaro Castagnola überfallen / und darunter 60. Frantzosen darunter 3. Officiers zu Kriegs Gefangenen gemacht / auch 2. Stüke bekommen / die übrige Frantzosen seynd in die Flucht getrieben / auch viele niedergemacht worden.

...ser Tagen ist das Kaiserliche Dragoner-Regiment Holz von Trient kommend vor diesen Mauren vorbey und nacher Gubernolo marschiret / almo dem Vernehmen nach den 7. dieses das Küraßier-Regiment Portugall angelanget / und gestern das Kais. Hungarische Cavallerie-Regiment Trips eintreffen sollen / morgen werden zwey Battaillons von dem Regiment Jung-Königsegg / und eine von Stahremberg erwartet.

...hier ist nichts merkwürdiges zu berichten / als daß die offentliche Faschings-Lustbarkeiten ihren gewünschten Fortgang haben / und täglichen die Menge Masqueren zu sehen seynd. Montags wurde die hiesige Lotterie gezogen / und kamen die Losse 38. 14. 66. 45. und 64. heraus.

...mög deren neuesten Briefen aus Brabant stehen die Frantzösische Truppen annoch rund um Brüssel herum / haben aber bis hieher noch nichts gegen die Stadt ausgerichtet / als daß sie dieselbe sehr genau eingeschlossen. Man argwohnet / daß die Frantzosen hiers durch nichts anders vorhätten / als die



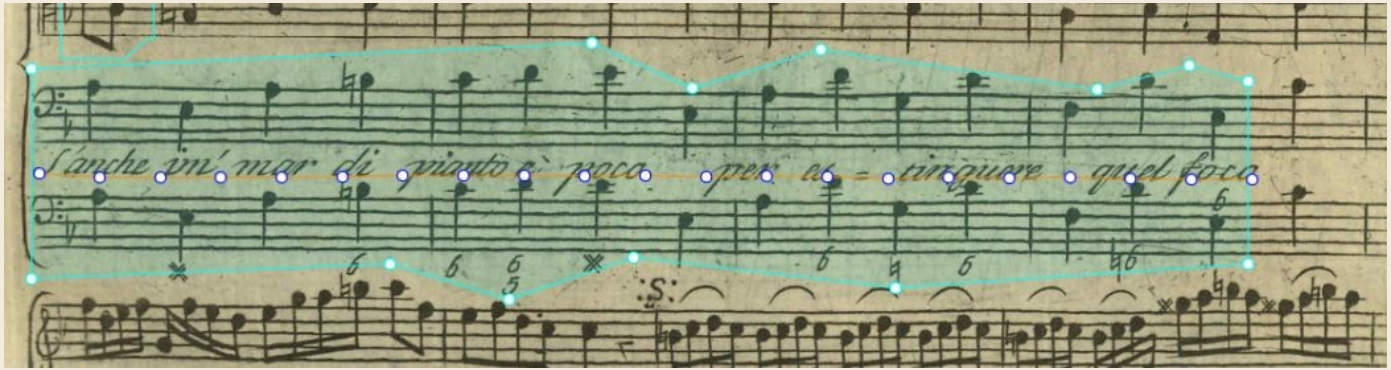
Nepresné polygóny (mnohouholníky)

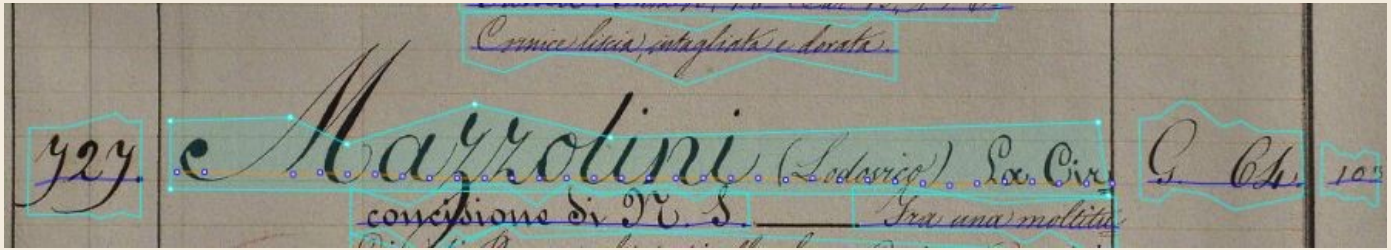
Transkribus

Inaccurate Polygons

Example 1

Example 2

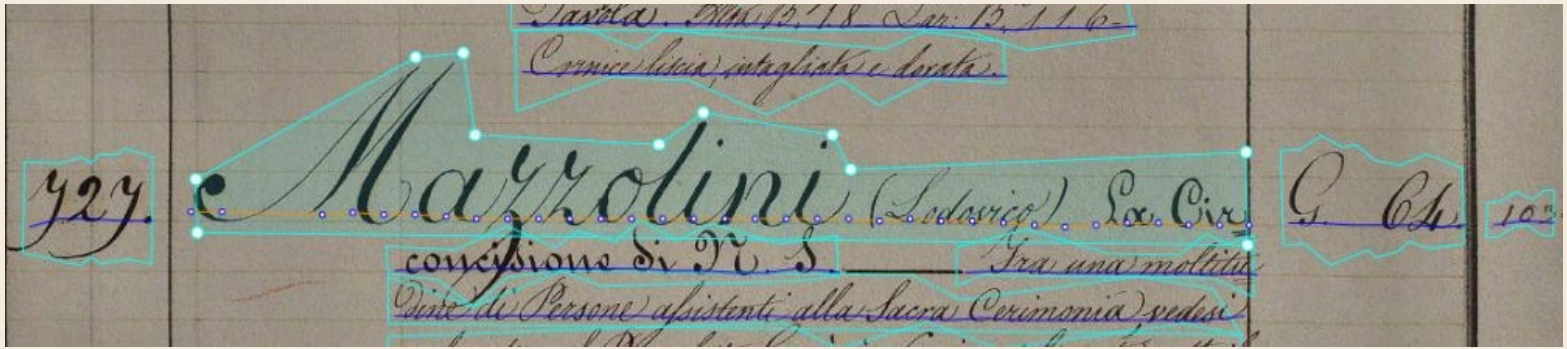




Field Model trained on Line Polygons

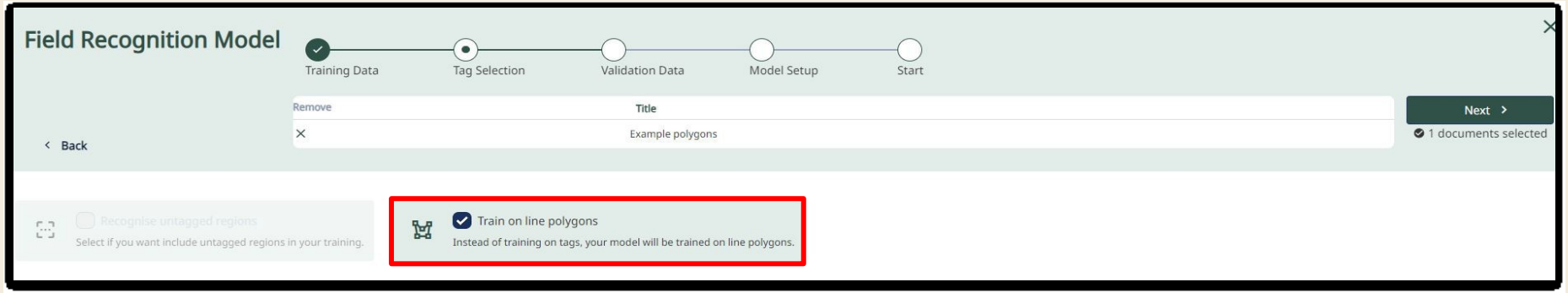
Prepare about 50 pages of training data:

- Adjust the line polygons manually

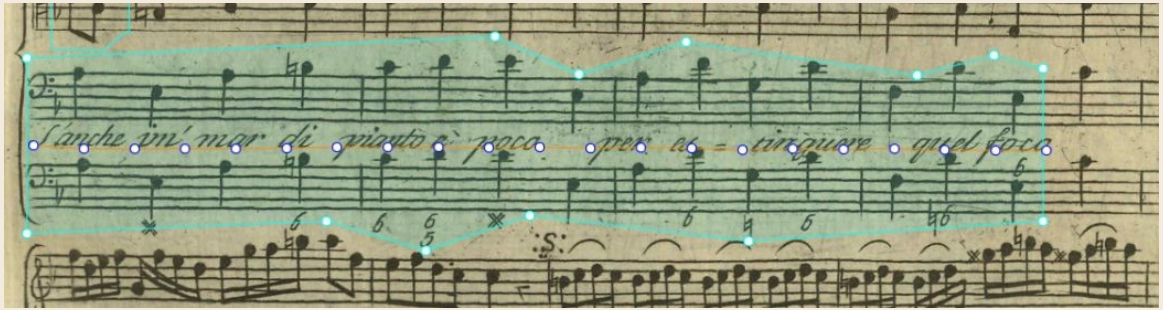


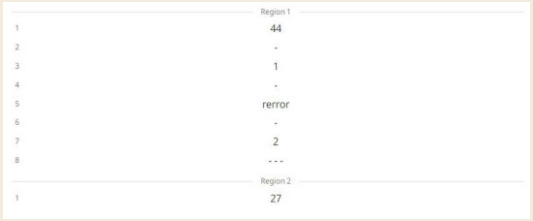
Field Model trained on Line Polygons

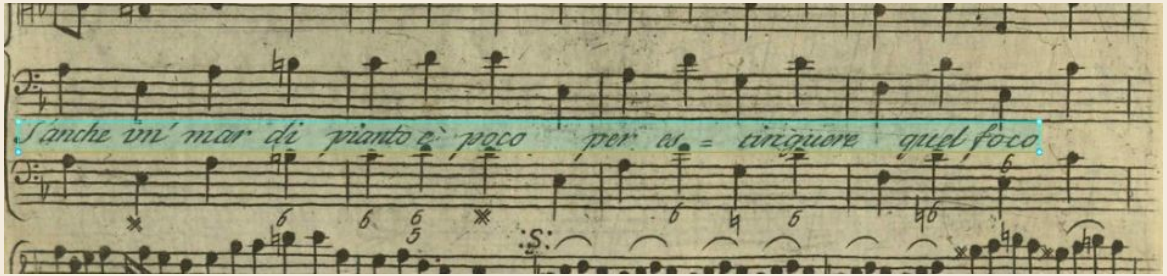
- Training data
- Tag selection: TRAIN ON LINE POLYGONS
- Validation data
- Model setting
- Advanced settings (Training Cycles and Learning Rate)

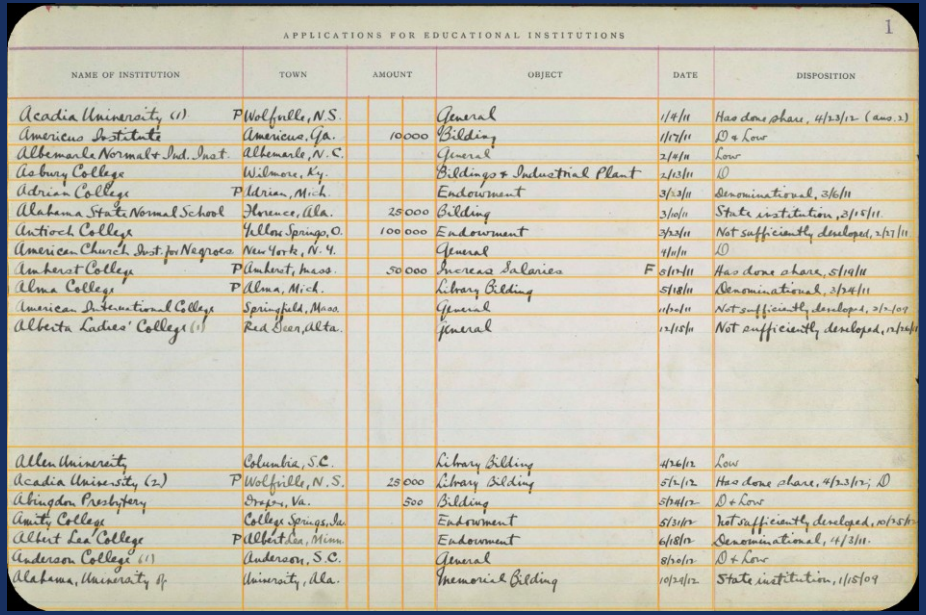


Field Model trained on Line Polygons









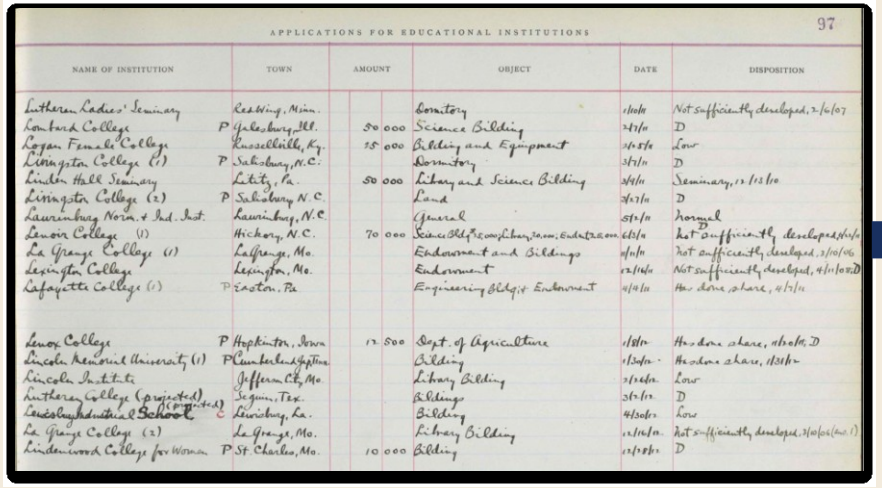
APPLICATIONS FOR EDUCATIONAL INSTITUTIONS 1

NAME OF INSTITUTION		TOWN	AMOUNT	OBJECT		DATE	DISPOSITION
Acadia University (1)	P	Wolfville, N.S.		General		1/9/11	Has done share, 4/23/12 (ans. 2)
Americus Institute		Americus, Ga.	10000	Bilding		11/17/11	D + Loss
Albemarle Normal + Ind. Inst.		Albemarle, N.C.		General		2/6/11	Loss
Asbury College		Wilmore, Ky.		Bildings + Industrial Plant		2/13/11	D
Adrian College	P	Adrian, Mich.		Endowment		3/2/11	Denominational, 3/6/11
Alabama State Normal School		Florence, Ala.	25000	Bilding		3/10/11	State institution, 3/15/11.
Antioch College		Yellow Springs, O.	100000	Endowment		3/22/11	Not sufficiently developed, 2/27/11
American Church Inst. for Negroes		New York, N.Y.		General		4/6/11	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 Bilding		5/18/11	Denominational, 5/24/11
American International College		Springfield, Mass.		General		11/6/11	Not sufficiently developed, 2/2/09
Alberta Ladies' College (1)		Red Deer, Alta.		general		12/15/11	Not sufficiently developed, 11/24/11
Allen University		Columbia, S.C.		Library Bilding		4/26/12	Loss
Acadia University (2)	P	Wolfville, N.S.	25000	Library Bilding		5/2/12	Has done share, 4/23/12; D
Abingdon Presbytery		Oreper, Va.	500	Bilding		5/14/12	D + Loss
Amity College		College Springs, Ia.		Endowment		5/21/12	Not sufficiently developed, 10/15/12
Albert Lea College	P	Albert Lea, Minn.		Endowment		6/18/12	Denominational, 11/3/11.
Anderson College (1)		Anderson, S.C.		General		8/20/12	D + Loss
Alabama, University of		University, Ala.		Memorial Bilding		10/24/12	State institution, 11/18/04

Table Models

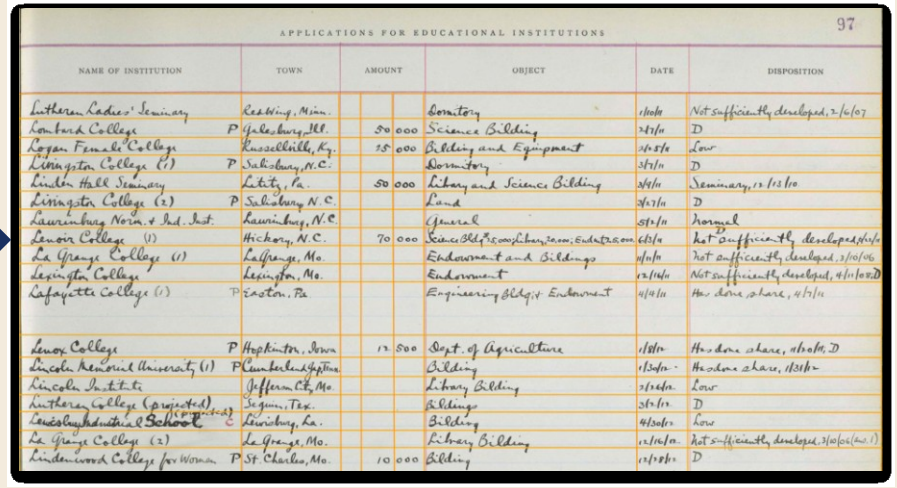
Transkribus

Modely pre tabuľky(Beta)



APPLICATIONS FOR EDUCATIONAL INSTITUTIONS 97

NAME OF INSTITUTION		TOWN	AMOUNT	OBJECT	DATE	DISPOSITION
Lutheran Ladies' Seminary		Red Wing, Minn.		Dormitory	1/10/10	Not sufficiently developed, 2/6/07
Lombard College	P	Galesburg, Ill.	50 000	Science Building	2/1/11	D
Logan Female College		Russellville, Ky.	15 000	Building and Equipment	2/15/11	Low
Livingston College (1)	P	Salisbury, N.C.		Dormitory	3/1/11	D
Linden Hall Seminary		Lititz, Pa.	50 000	Library and Science Building	3/4/11	Seminary, 12/13/10
Livingston College (2)	P	Salisbury N.C.		Land	3/17/11	D
Lawrenburg Norm. + Ind. Inst.		Lawrenburg, N.C.		General	5/1/11	Normal
Lenoir College (1)		Hickory, N.C.	70 000	Science Bldg + 2 com Library Dorm. Endowment	6/1/11	Not sufficiently developed, 4/1/10
La Grange College (1)		LaGrange, Mo.		Endowment and Buildings	11/10/11	Not sufficiently developed, 3/10/06
Lexington College		Lexington, Mo.		Endowment	12/14/11	Not sufficiently developed, 4/1/10
Lafayette College (1)	P	Easton, Pa.		Engineering Bldg + Endowment	11/4/11	Has done share, 4/1/11
Lenox College	P	Hopkinton, Iowa	12 500	Dept. of Agriculture	1/18/11	Has done share, 1/20/11, D
Lincoln Memorial University (1)	P	Cumberland Gap, Tenn.		Building	1/30/11	Has done share, 11/31/11
Lincoln Institute		Jefferson City, Mo.		Library Building	2/14/11	Low
Lutheran College (projected)		Seguin, Tex.		Buildings	3/1/11	D
Louisiana Industrial School	C	Lewisburg, La.		Building	4/30/11	Low
La Grange College (2)		La Grange, Mo.		Library Building	12/14/10	Not sufficiently developed, 3/10/06 (Am. 1)
Lindenwood College for Women	P	St. Charles, Mo.	10 000	Building	12/10/11	D



APPLICATIONS FOR EDUCATIONAL INSTITUTIONS 97

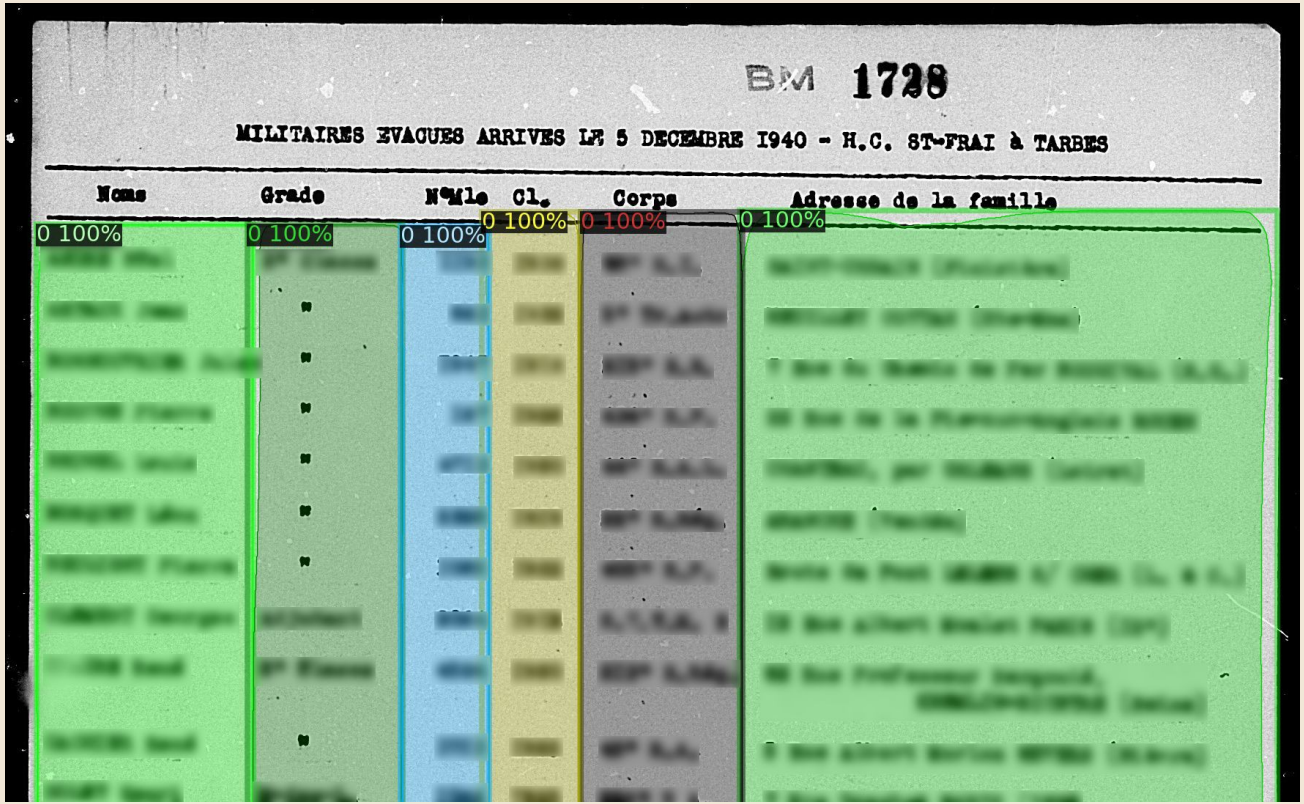
NAME OF INSTITUTION		TOWN	AMOUNT	OBJECT	DATE	DISPOSITION
Lutheran Ladies' Seminary		Red Wing, Minn.		Dormitory	1/10/10	Not sufficiently developed, 2/6/07
Lombard College	P	Galesburg, Ill.	50 000	Science Building	2/1/11	D
Logan Female College		Russellville, Ky.	15 000	Building and Equipment	2/15/11	Low
Livingston College (1)	P	Salisbury, N.C.		Dormitory	3/1/11	D
Linden Hall Seminary		Lititz, Pa.	50 000	Library and Science Building	3/4/11	Seminary, 12/13/10
Livingston College (2)	P	Salisbury N.C.		Land	3/17/11	D
Lawrenburg Norm. + Ind. Inst.		Lawrenburg, N.C.		General	5/1/11	Normal
Lenoir College (1)		Hickory, N.C.	70 000	Science Bldg + 2 com Library Dorm. Endowment	6/1/11	Not sufficiently developed, 4/1/10
La Grange College (1)		LaGrange, Mo.		Endowment and Buildings	11/10/11	Not sufficiently developed, 3/10/06
Lexington College		Lexington, Mo.		Endowment	12/14/11	Not sufficiently developed, 4/1/10
Lafayette College (1)	P	Easton, Pa.		Engineering Bldg + Endowment	11/4/11	Has done share, 4/1/11
Lenox College	P	Hopkinton, Iowa	12 500	Dept. of Agriculture	1/18/11	Has done share, 1/20/11, D
Lincoln Memorial University (1)	P	Cumberland Gap, Tenn.		Building	1/30/11	Has done share, 11/31/11
Lincoln Institute		Jefferson City, Mo.		Library Building	2/14/11	Low
Lutheran College (projected)		Seguin, Tex.		Buildings	3/1/11	D
Louisiana Industrial School	C	Lewisburg, La.		Building	4/30/11	Low
La Grange College (2)		La Grange, Mo.		Library Building	12/14/10	Not sufficiently developed, 3/10/06 (Am. 1)
Lindenwood College for Women	P	St. Charles, Mo.	10 000	Building	12/10/11	D

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

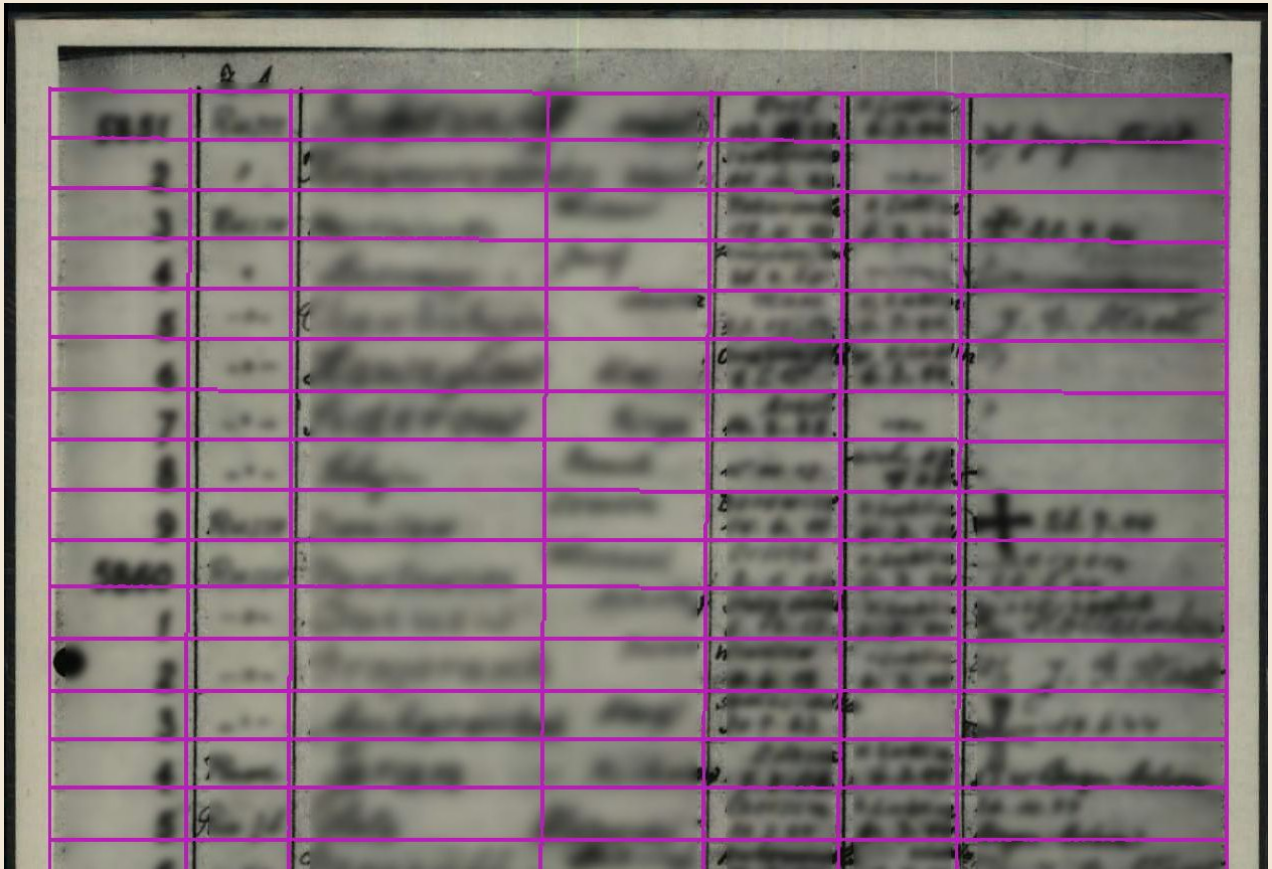
Modely pre tabuľky(Beta)

- 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

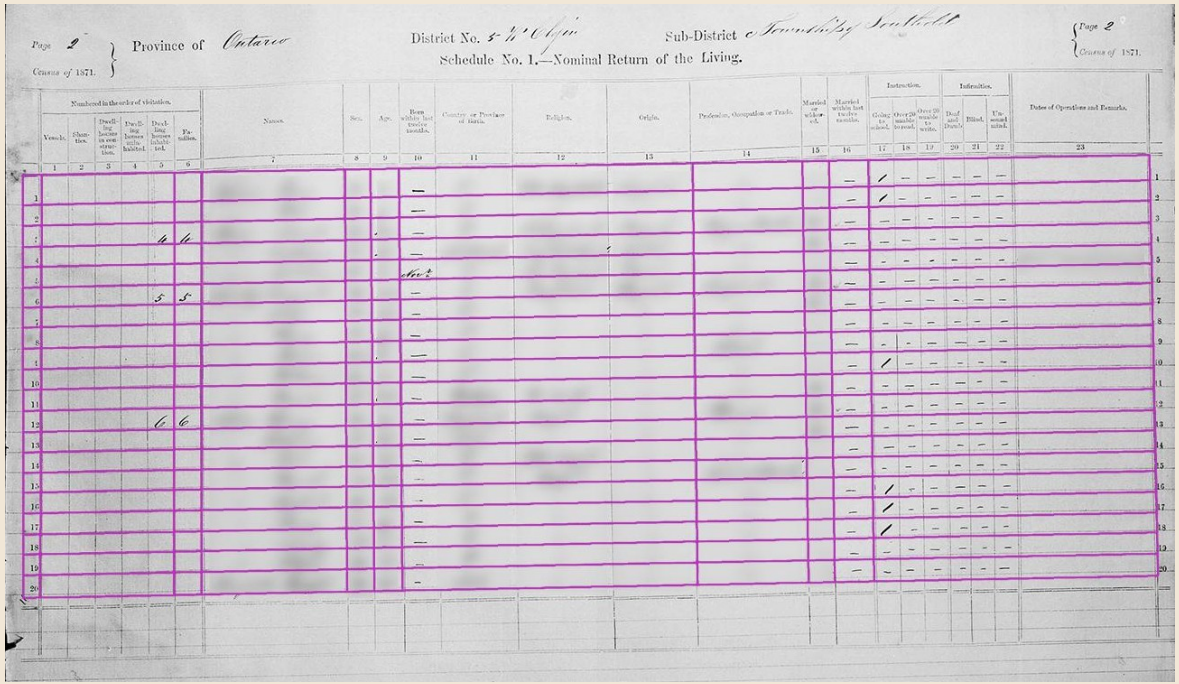
Riadky a stĺpce



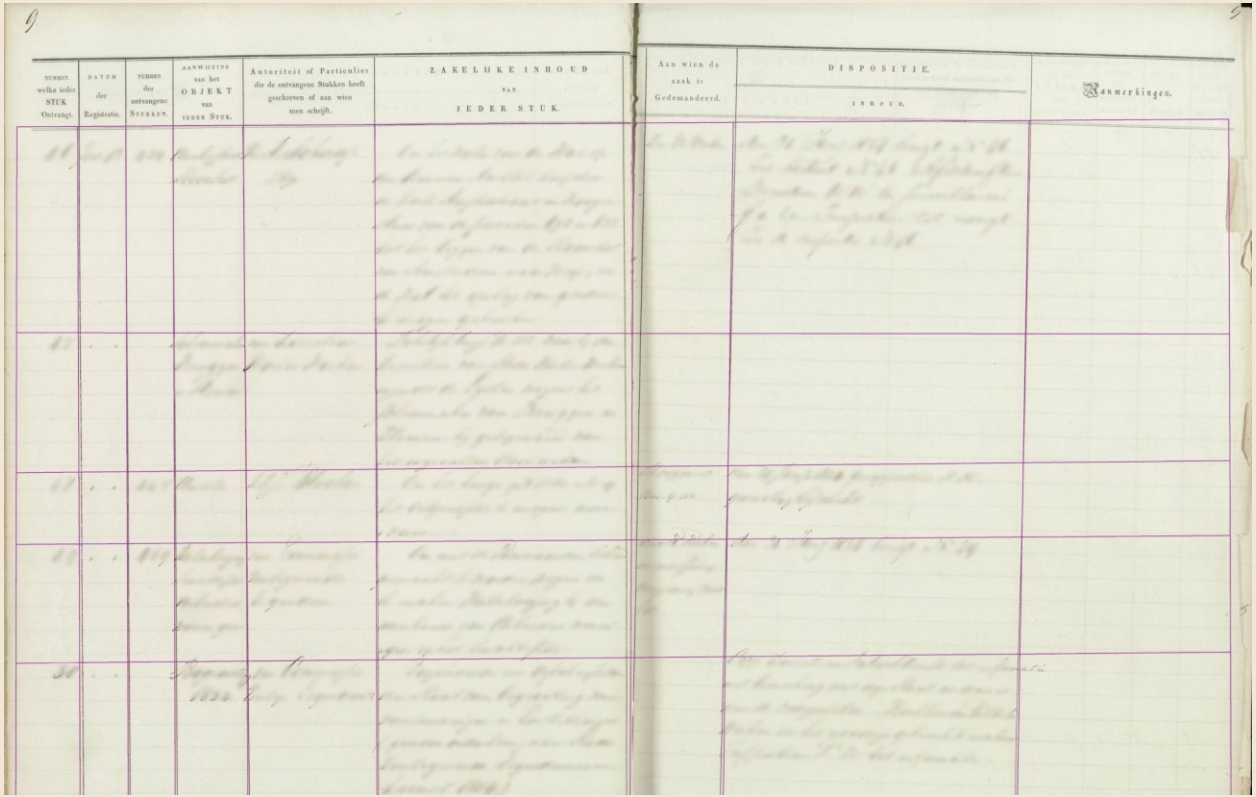
Riadky a stĺpce



Skosené tabuľky



Viacriadkové bunky

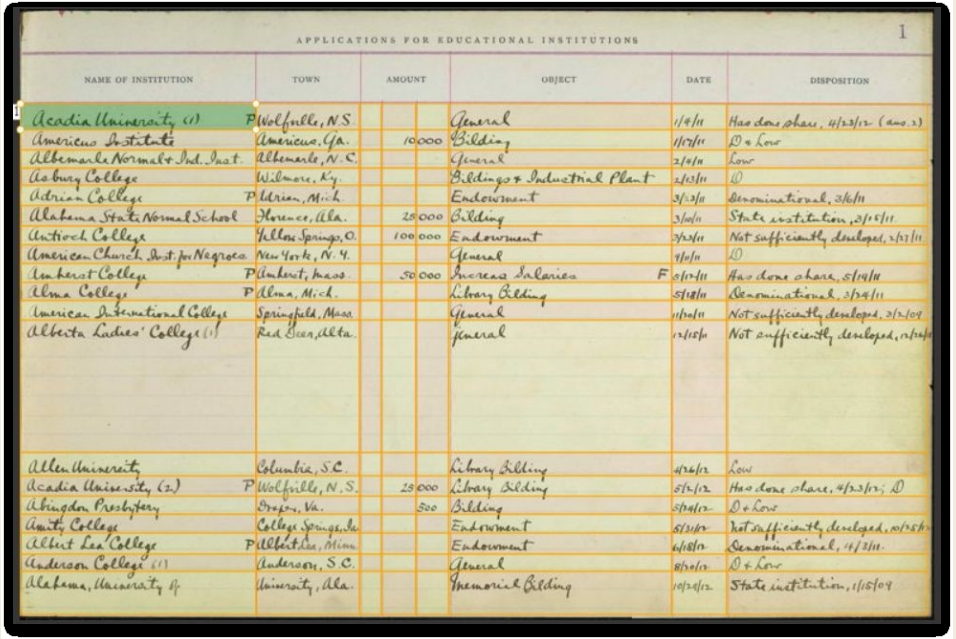


Modely pre tabuľky

Ground Truth tvorba v editore:

Tabuľka

- Stĺpce
- Riadky



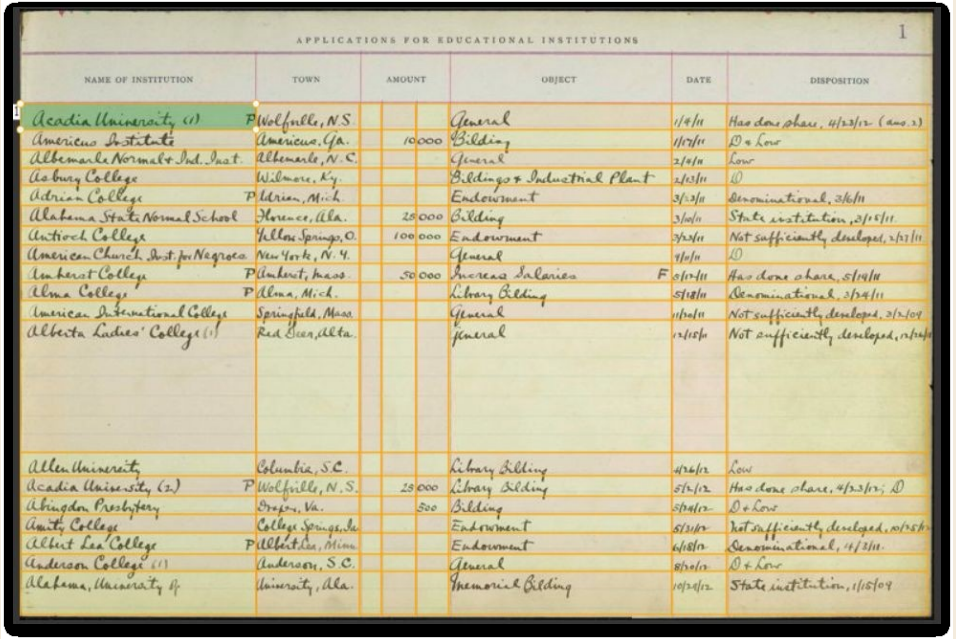
APPLICATIONS FOR EDUCATIONAL INSTITUTIONS 1

NAME OF INSTITUTION		TOWN	AMOUNT	OBJECT		DATE	DISPOSITION
Acadia University (1)	P	Wolfville, N.S.		General		1/9/11	Has done share, 4/23/12 (see 2)
Americus Institute		Americus, Ga.	10,000	Building		1/17/11	D + Loan
Albemarle Normal + Ind. Inst.		Albemarle, N.C.		General		2/9/11	Loan
Asbury College		Wilmore, Ky.		Buildings + Industrial Plant		2/13/11	D
Adrian College	P	Adrian, Mich.		Endowment		3/2/11	Denominational, 3/6/11
Alabama State Normal School		Florence, Ala.	25,000	Building		3/10/11	State institution, 3/15/11
Antioch College		Yellow Springs, O.	100,000	Endowment		3/24/11	Not sufficiently developed, 4/27/11
American Church Inst. for Negroes		New York, N.Y.		General		4/11/11	D
Amherst College	P	Amherst, Mass.	50,000	Increase Salaries	F	5/17/11	Has done share, 5/19/11
Alma College	P	Alma, Mich.		Library Building		5/18/11	Denominational, 5/24/11
American International College		Springfield, Mass.		General		11/26/11	Not sufficiently developed, 3/12/09
Alberta Ladies' College (1)		Red Deer, Alta.		general		12/10/11	Not sufficiently developed, 11/14/11
Allen University		Columbia, S.C.		Library Building		4/26/12	Loan
Acadia University (2)	P	Wolfville, N.S.	25,000	Library Building		5/1/12	Has done share, 4/23/12; D
Abingdon Presbytery		Drapo, Va.	500	Building		5/14/12	D + Loan
Amity College		College Springs, Ia.		Endowment		5/31/12	Not sufficiently developed, 10/25/11
Albert Lea College	P	Albert Lea, Minn.		Endowment		6/18/12	Denominational, 11/21/11
Anderson College (1)		Anderson, S.C.		General		8/10/12	D + Loan
Alabama, University of		University, Ala.		Memorial Building		10/12/12	State institution, 11/15/09

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



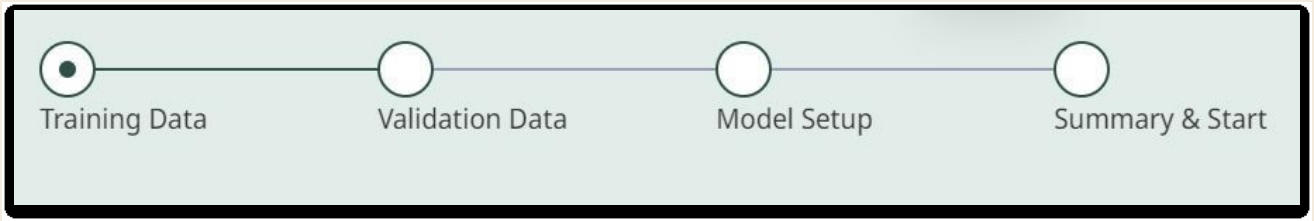
APPLICATIONS FOR EDUCATIONAL INSTITUTIONS 1

NAME OF INSTITUTION		TOWN	AMOUNT	OBJECT		DATE	DISPOSITION
Acadia University (1)	P	Wolfville, N.S.		General		1/9/11	Has done share, 4/23/12 (ans. 2)
Americus Institute		Americus, Ga.	10,000	Bilding		1/17/11	D + Low
Albemarle Normal + Ind. Inst.		Albemarle, N.C.		General		2/9/11	Low
Asbury College		Wilmore, Ky.		Bildings + Industrial Plant		2/13/11	D
Adrian College	P	Adrian, Mich.		Endowment		3/2/11	Denominational, 3/6/11
Alabama State Normal School		Florence, Ala.	25,000	Bilding		3/10/11	State institution, 3/15/11
Antioch College		Yellow Springs, O.	100,000	Endowment		3/23/11	Not sufficiently developed, 4/27/11
American Church Inst. for Negroes		New York, N.Y.		General		4/11/11	D
Amherst College	P	Amherst, Mass.	50,000	Increase Salaries	F	5/12/11	Has done share, 5/19/11
Alma College	P	Alma, Mich.		Library Bilding		5/18/11	Denominational, 5/24/11
American International College		Springfield, Mass.		General		11/10/11	Not sufficiently developed, 3/12/09
Alberta Ladies' College (1)		Red Deer, Alta.		general		12/10/11	Not sufficiently developed, 11/14/11
Allen University		Columbia, S.C.		Library Bilding		4/26/12	Low
Acadia University (2)	P	Wolfville, N.S.	25,000	Library Bilding		5/1/12	Has done share, 4/23/12; D
Abingdon Presbytery		Drapo, Va.	500	Bilding		5/14/12	D + Low
Amity College		College Springs, Ia.		Endowment		5/31/12	Not sufficiently developed, 10/25/11
Albert Lea College	P	Albert Lea, Minn.		Endowment		6/18/12	Denominational, 11/21/11
Anderson College (1)		Anderson, S.C.		General		8/10/12	D + Low
Alabama, University of		University, Ala.		Memorial Bilding		10/12/12	State institution, 11/15/09

Modely pre tabuľky

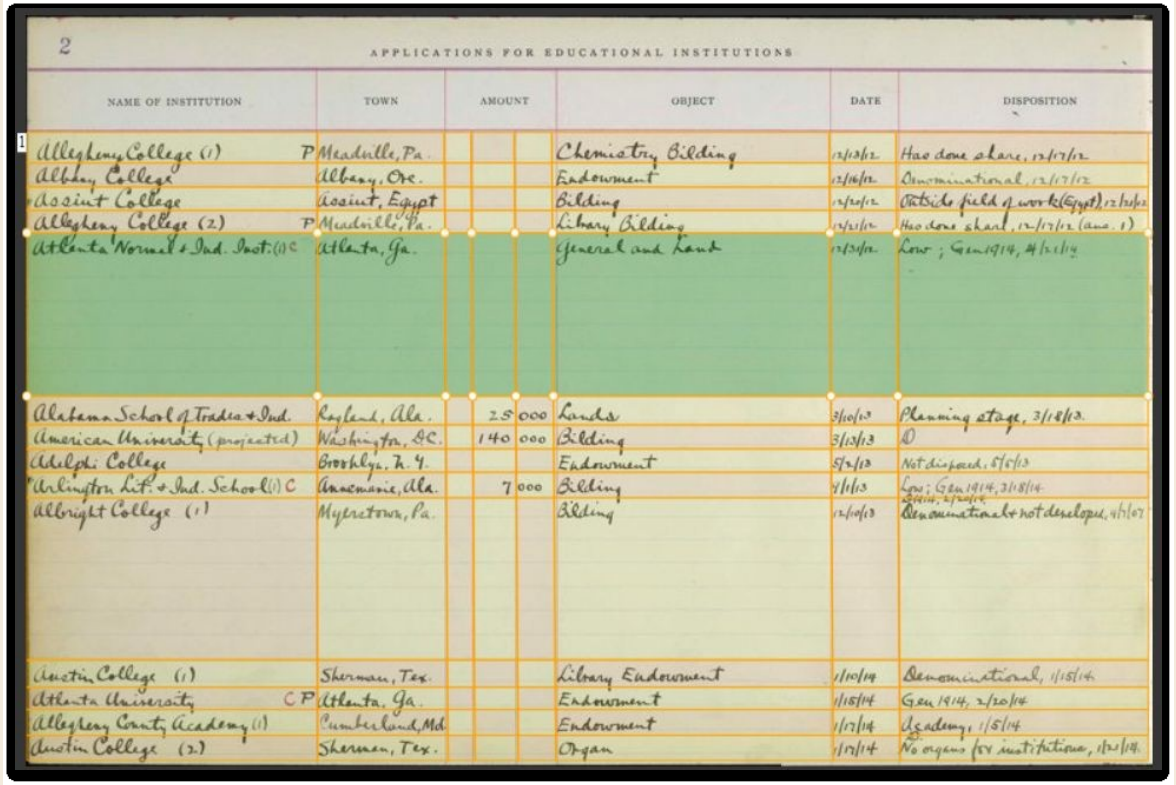
Tréning (beta.transkribus.eu):

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



Modely pre tabuľky

Ground Truth: 20 strán



2

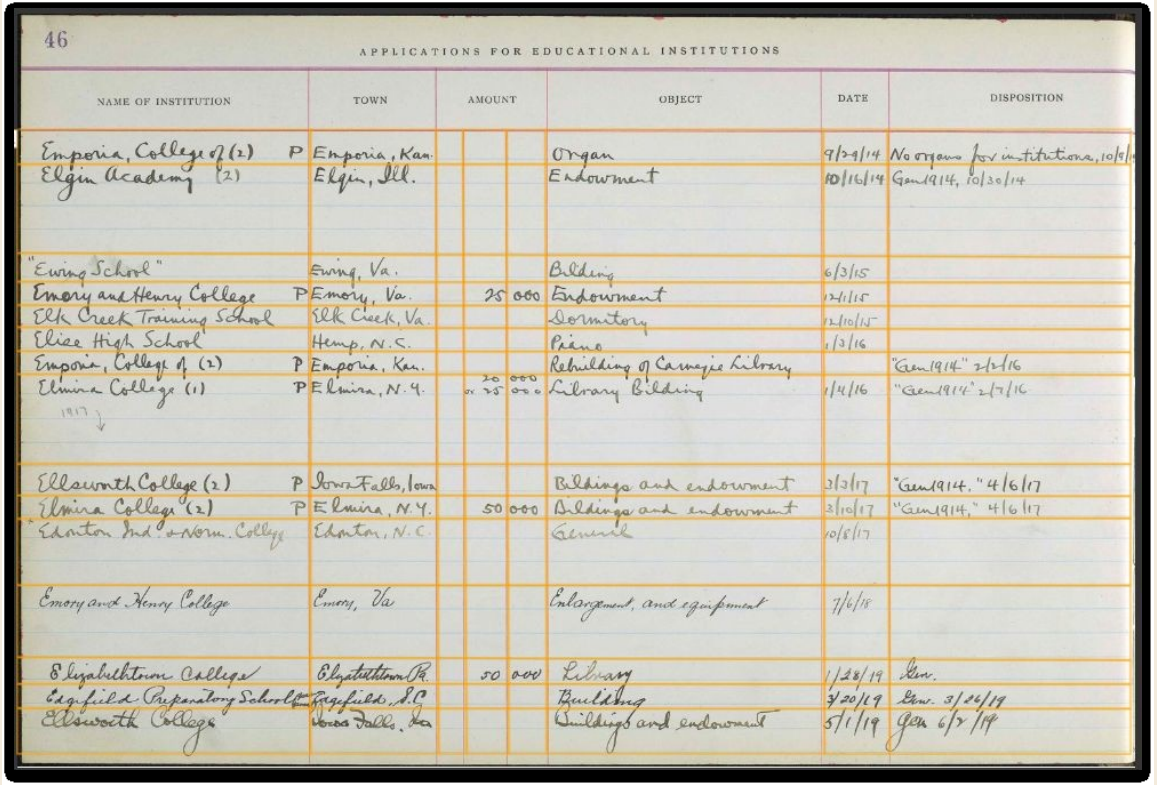
APPLICATIONS FOR EDUCATIONAL INSTITUTIONS

NAME OF INSTITUTION	TOWN	AMOUNT	OBJECT	DATE	DISPOSITION
Allegheny College (1) P	Meadville, Pa.		Chemistry Bilding	12/13/12	Has done share, 12/17/12
Albany College	Albany, Ore.		Endowment	12/16/12	Denominational, 12/17/12
Assiut College	Assiut, Egypt		Bilding	12/20/12	Outside field of work (Egypt) 12/28/12
Allegheny College (2) P	Meadville, Pa.		Library Bilding	12/21/12	Has done share, 12/17/12 (ans. 1)
Atlanta Normal + Ind. Inst. (1) C	Atlanta, Ga.		General and Land	12/31/12	Low; Gen 1914, 4/1/14
Alabama School of Trades + Ind.	Rayland, Ala.	25 000	Lands	3/10/13	Planning stage, 3/18/13.
American University (projected)	Washington, D.C.	140 000	Bilding	3/13/13	ID
Adelphi College	Brooklyn, N.Y.		Endowment	5/2/13	Not dis'pated, 5/6/13
Arlington Lit. + Ind. School (1) C	Anacomania, Ala.	7 000	Bilding	4/1/13	Low; Gen 1914, 31/8/14.
Albright College (1)	Myerstown, Pa.		Bilding	12/10/13	Denominational + not developed, 4/1/07
Austin College (1)	Sherman, Tex.		Library Endowment	1/10/14	Denominational, 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, 1/15/14
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)	P	Emporia, Kan.		Organ	9/29/14	No organs for institutions, 10/9/14
Elgin Academy (2)		Elgin, Ill.		Endowment	10/16/14	Gen1914, 10/30/14
"Ewing School"		Ewing, Va.		Bilding	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	
Elise High School		Hemp, N.C.		Piano	1/3/16	
Emporia, College of (2)	P	Emporia, Kan.		Rebuilding of Carnegie Library		"Gen1914" 2/2/16
Elmira College (1)	P	Elmira, N.Y.	20 000 or 25 000	Library Bilding	1/4/16	"Gen1914" 2/7/16
1917						
Ellsworth College (2)	P	Iowa Falls, Iowa		Bildings and endowment	2/3/17	"Gen1914." 4/6/17
Elmira College (2)	P	Elmira, N.Y.	50 000	Bildings and endowment	2/10/17	"Gen1914," 4/6/17
Elanton Ind'l & Norm. College		Elanton, 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.
Edgefield Preparatory School		Edgefields, S.C		Building	3/20/19	Gen. 3/26/19
Ellsworth College		Iowa Falls, Ia.		Buildings and endowment	5/1/19	Gen 6/7/19

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

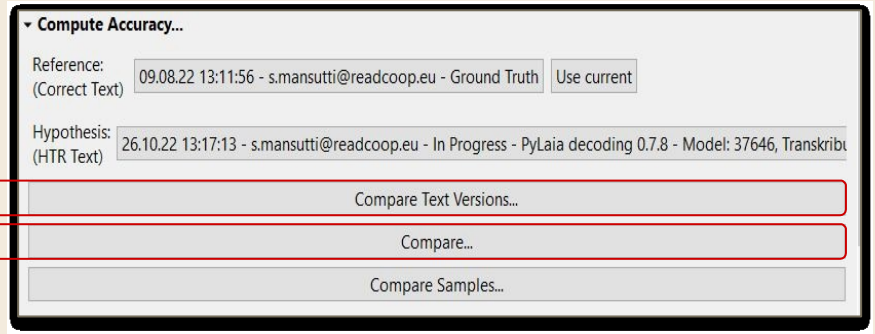
Transkribus

Výpočty presnosti transkripcie

Dve verzie tej istej stránky:

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

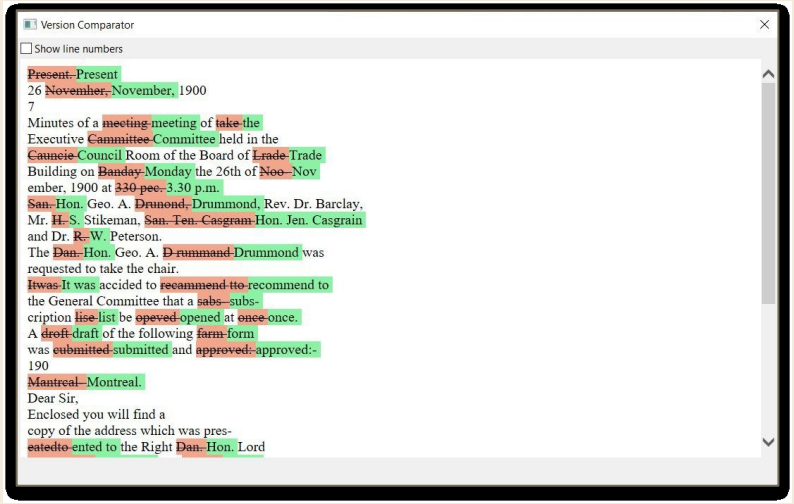
- **Porovnajte textové verzie (pozrite si rozdiely medzi dvoma vybratými verziami)**
- **Porovnať...(Compare)**
- **(porovnáva tieto dva prepisy**
- **a vypočítava chybovosť slov**
- **a chybovosť znakov)**



Výpočty presnosti transkripcie

Porovnať textové verzie

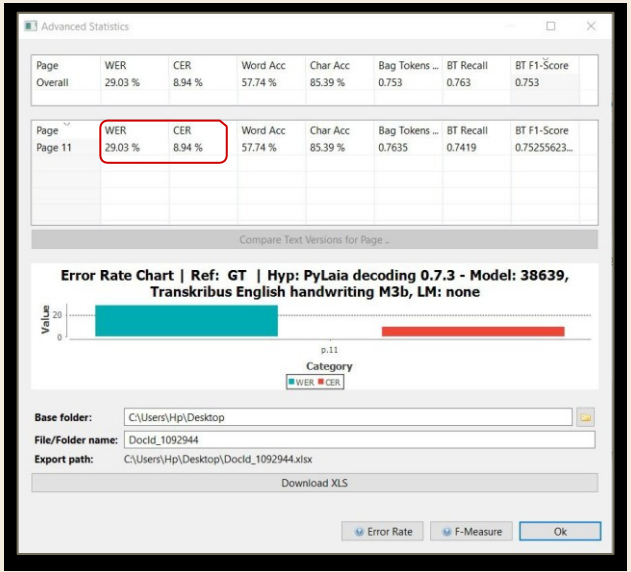
Ground Truth - model "Transkribus English handwriting M3b" bez jazykového modelu:



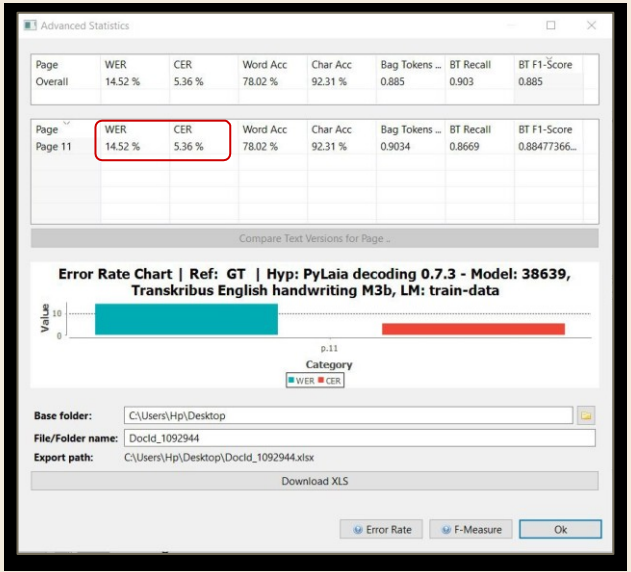
Ground Truth - “Transkribus anglický rukopis M3b" model s jazykovým modelom:

Výpočty presnosti transkripcie

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

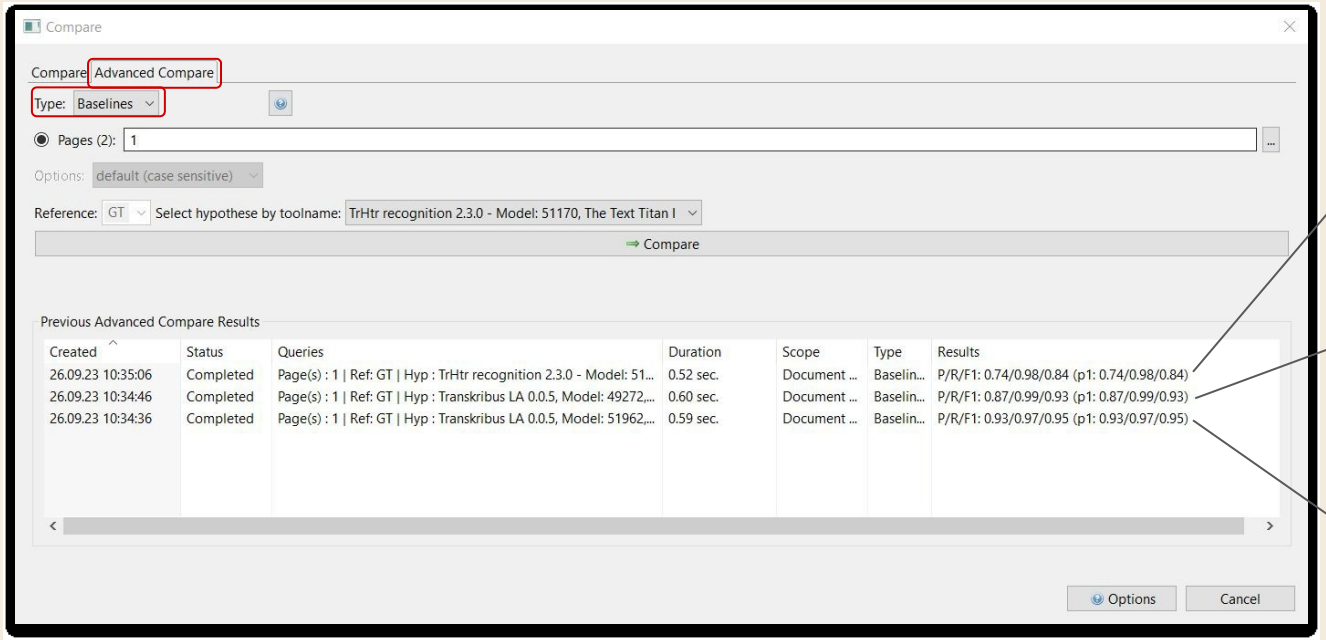


Ground Truth - “Transkribus anglický rukopis M3b" model s jazykovým modelom:



Výpočty presnosti transkripcie

Compare → Advanced Compare → Baselines

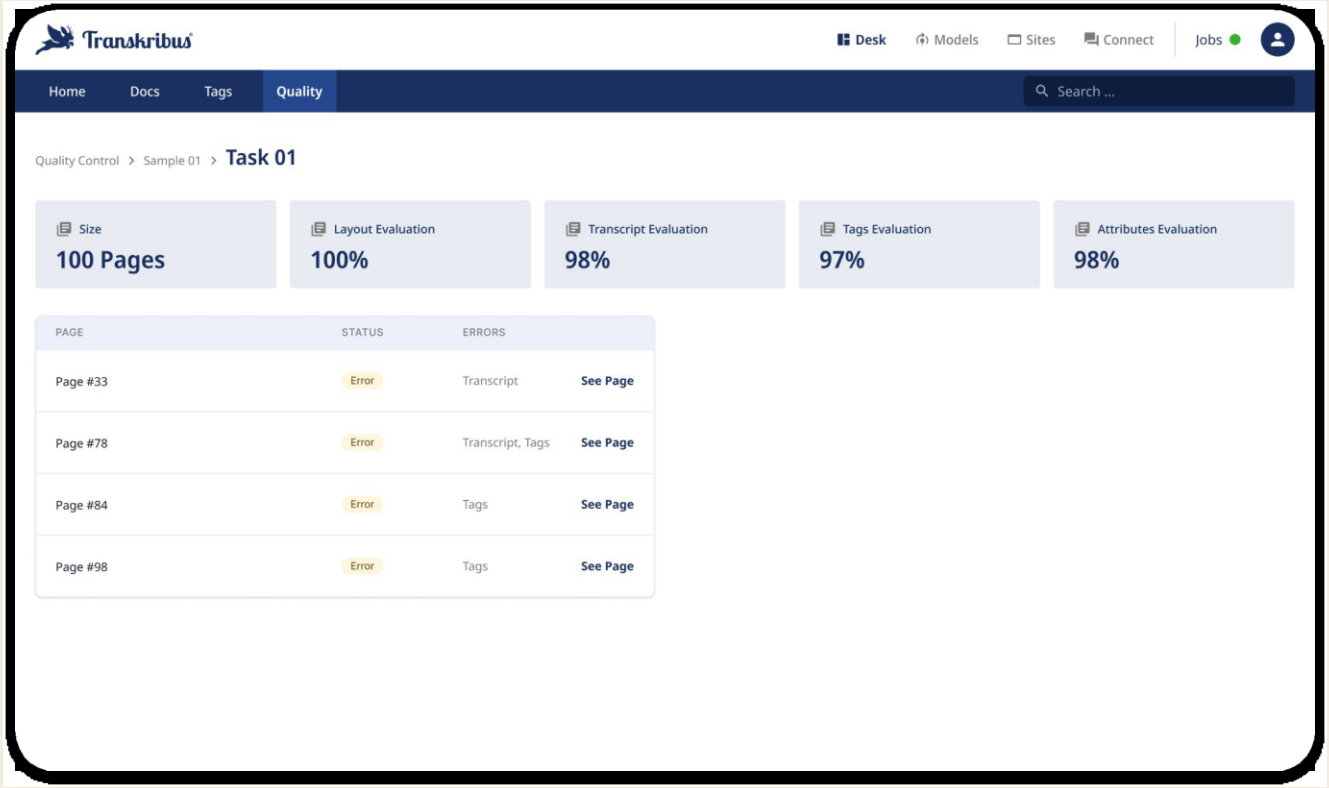


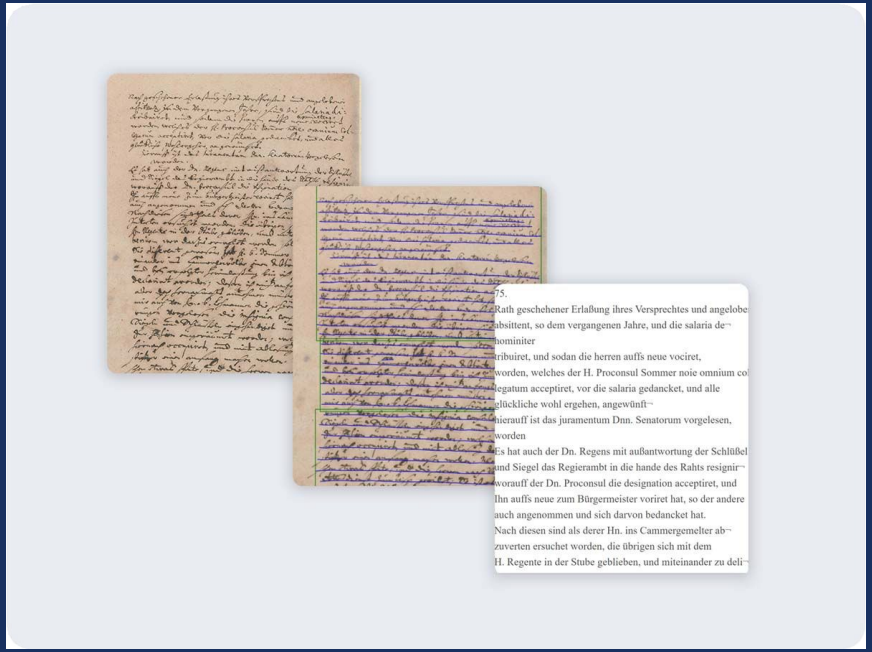
Predvolené rozloženie s rozpoznávaním textu

Základný model orientácie zmiešanej čiary

Základný model univerzálnych línií

Kontrola kvality





Publikačné modely v Transkribus

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

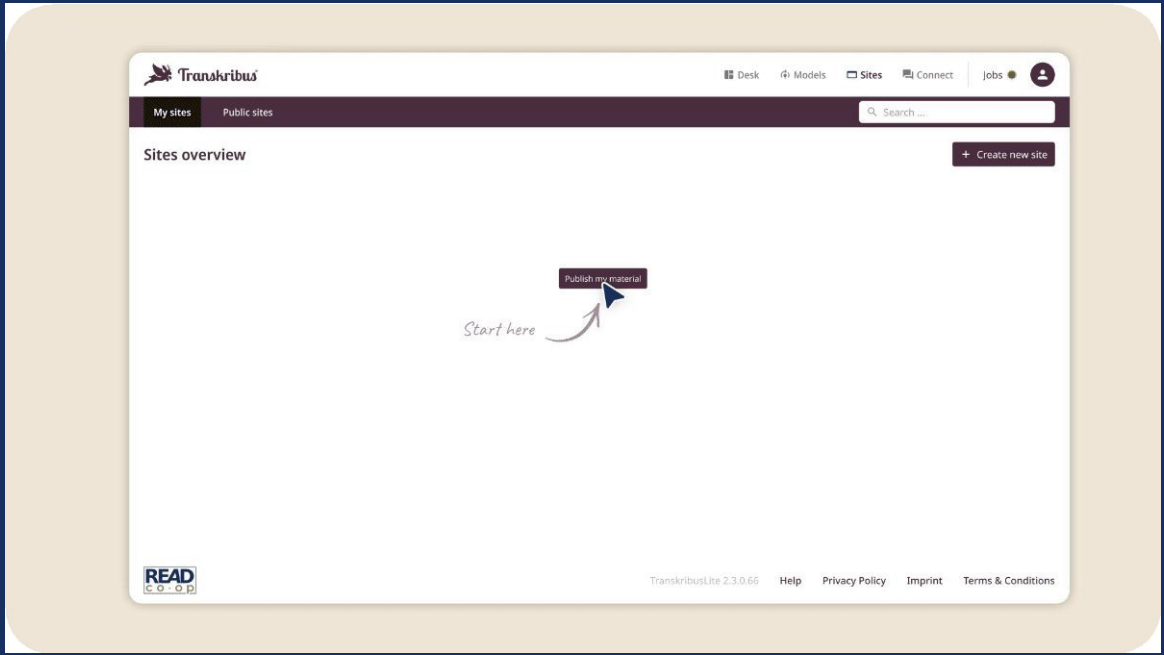
Publikačné modely

Ako publikovať model:
Kontaktujte nás prostredníctvominfo@readcoop.eu alebo prostredníctvomcontact form/help center aby ste nás informovali, že chcete zverejniť svoj model v rámci spoločnosti Transkribus

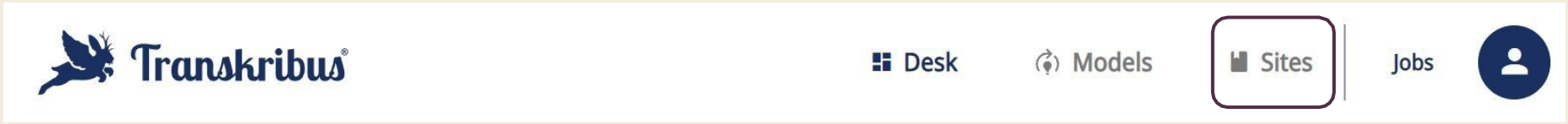
- Požiadavky: veľkosť tréningovej sady ~ 50 000 slov, CER 7%-5% alebo nižšia . Ak ide o model vyškolený na skript alebo jazyk, ktorý zatiaľ nemôžeme ponúknuť, tieto kritériá neplatia
- Poskytnúť stručný opis modelu, ktorý pomôže ostatným používateľom pochopiť použitý obsah školenia; Užitočné je aj pridanie 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

Plánované na rok 2024

Transkribus **Connect** je miesto, kde sa **exchange s**tane.

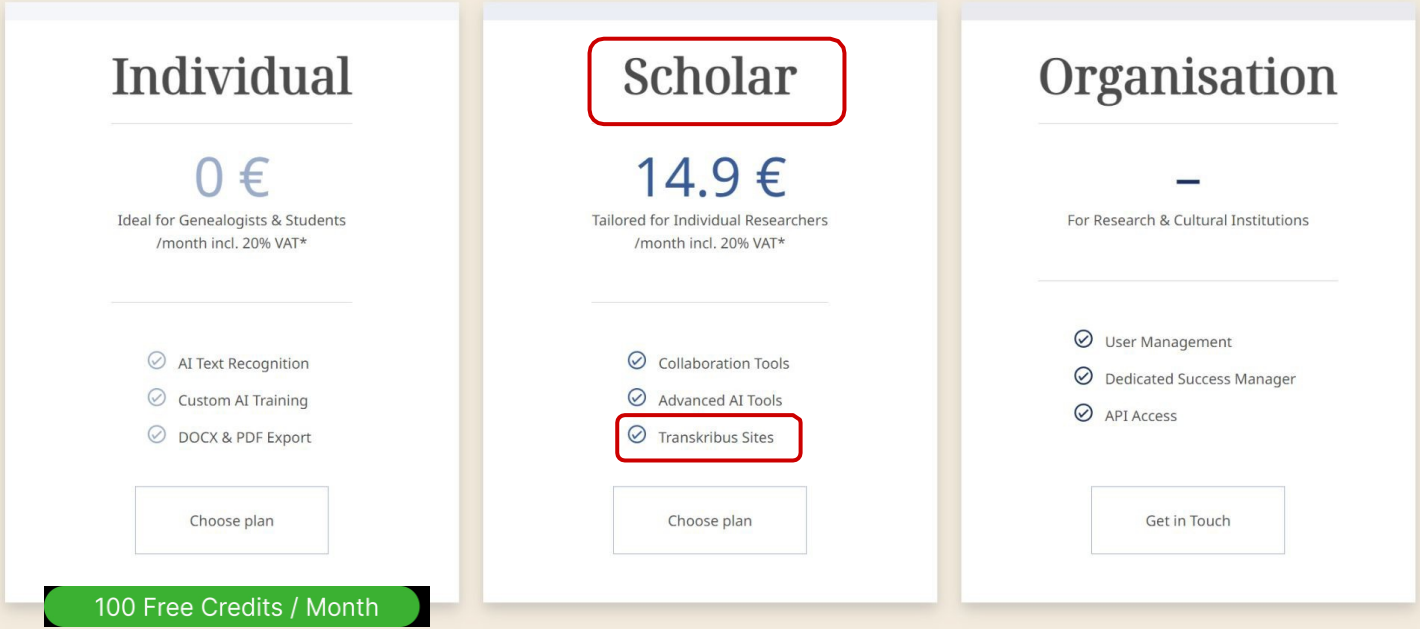


Transkribus stránky

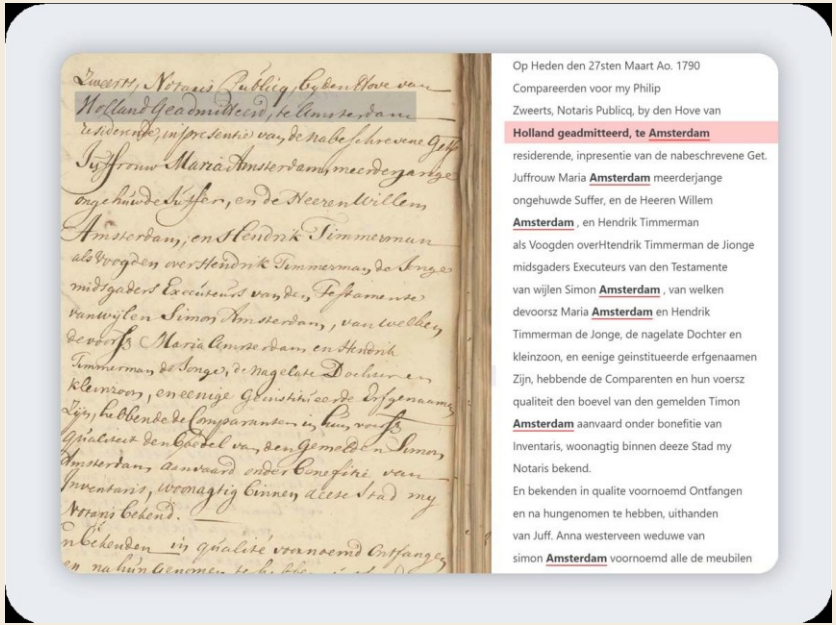


Transkribus **Connect** je miesto, kde sa **exchange s**tane.

Plány predplatného



Transkribus stránky - vlastnosti



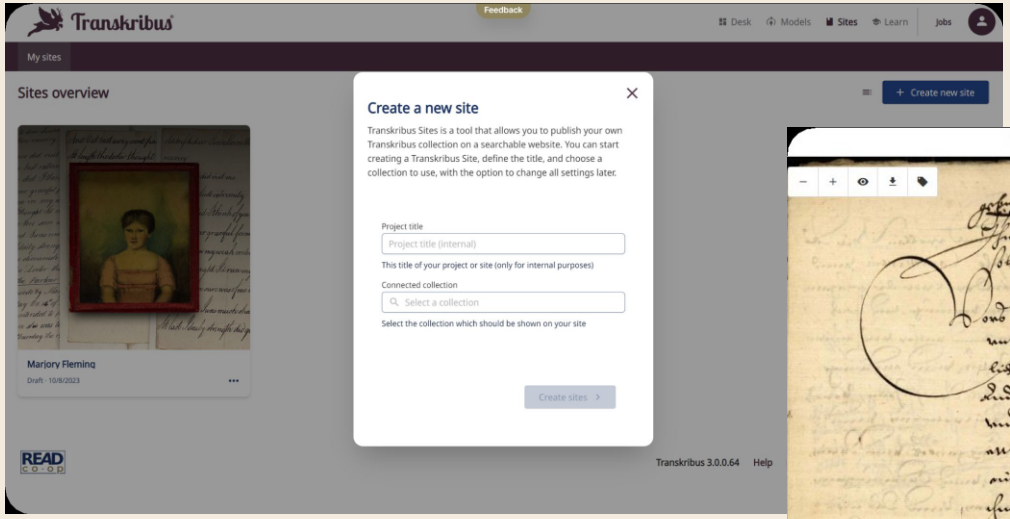
Op Heden den 27sten Maart Ao. 1790
Compareerden voor my Philip
Zweerts, Notaris Publicq, by den Hove van
Holland geadmitteerd, te Amsterdam
residerende, inpresentie van de nabeschrevene Get.
Juffrouw Maria **Amsterdam** meerderjange
ongehuwde Suffer, en de Heeren Willem
Amsterdam , en Hendrik Timmerman
als Voogden overHtendrik Timmerman de Jionge
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 Comparenten en hun voersz
qualiteit den boevel van den gemelden Timon
Amsterdam aanvaard onder bonefitie van
Inventaris, woonagtig binnen deeze 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

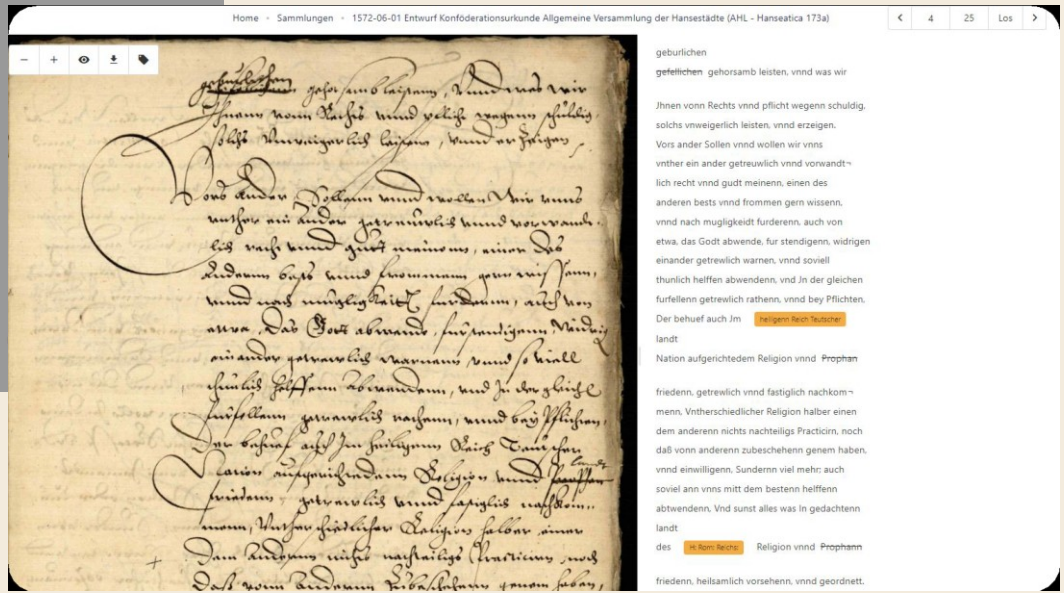
Jednoduché zdieľanie materiálu

Pohľad strana vedľa strany (obrázok-prepis)

Vylepšené možnosti vyhľadávania

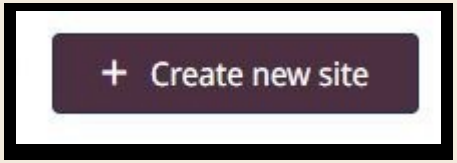
Transkribus stránky





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)

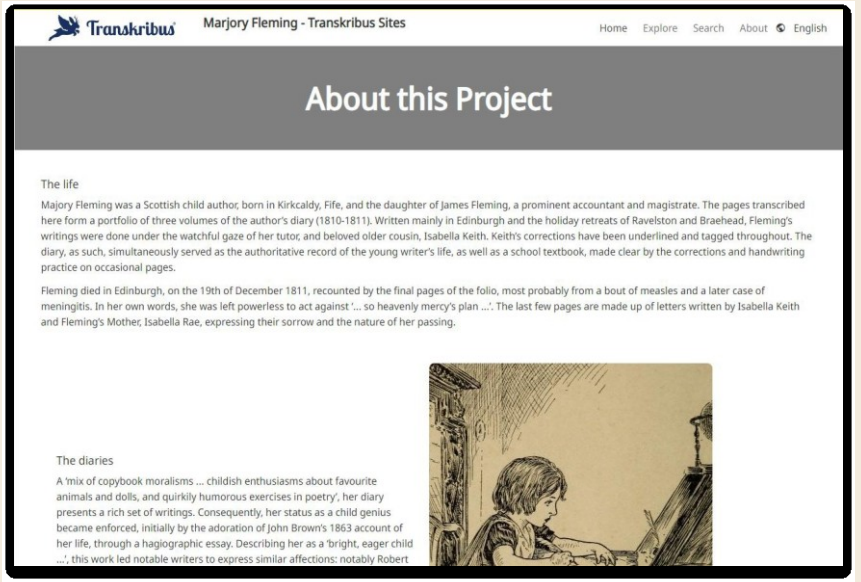
- Titul
- Stručný opis obsahu/stránky
- Obrázok pozadia domovskej stránky

Vaša prvá stránka Transkribus

O (About)

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

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



Vaša prvá stránka Transkribus

Preskúmať (Explore)

(Ako chcete nakonfigurovať stránku vyhľadávania):

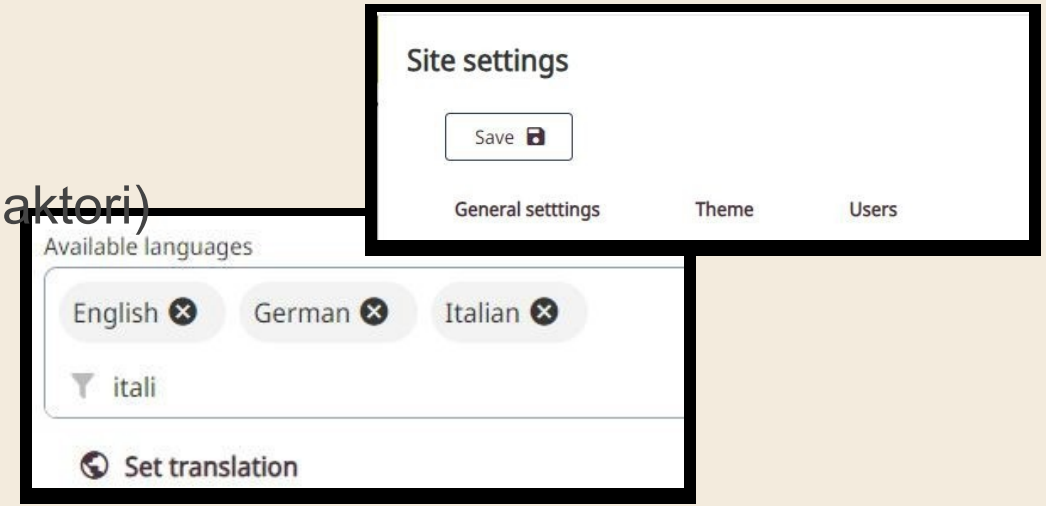
- Povolenie značiek prehľadávania
- Povolené značky (ak ste použili značky vo vašich dokumentoch Transkribus)
- Povoliť filtre a filter rokov (na základe metadát dokumentov Transkribus)

Vaša prvá stránka Transkribus

Read&Search - Demo (transkribus.eu)

Ďalšie nastavenia (Other settings):

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



Transkribus®

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

Transkribus®

Unlocking the past, together