

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. Digitálna 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
-

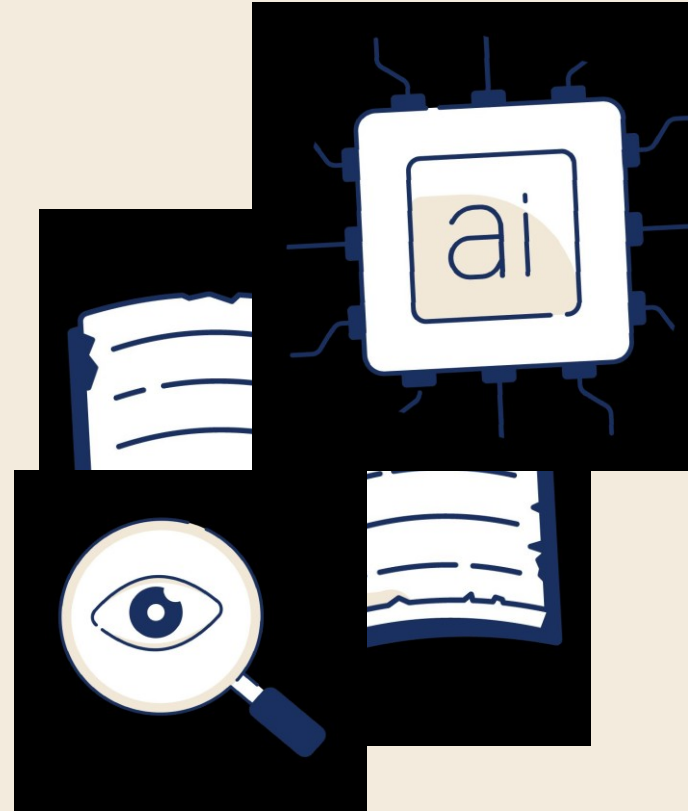




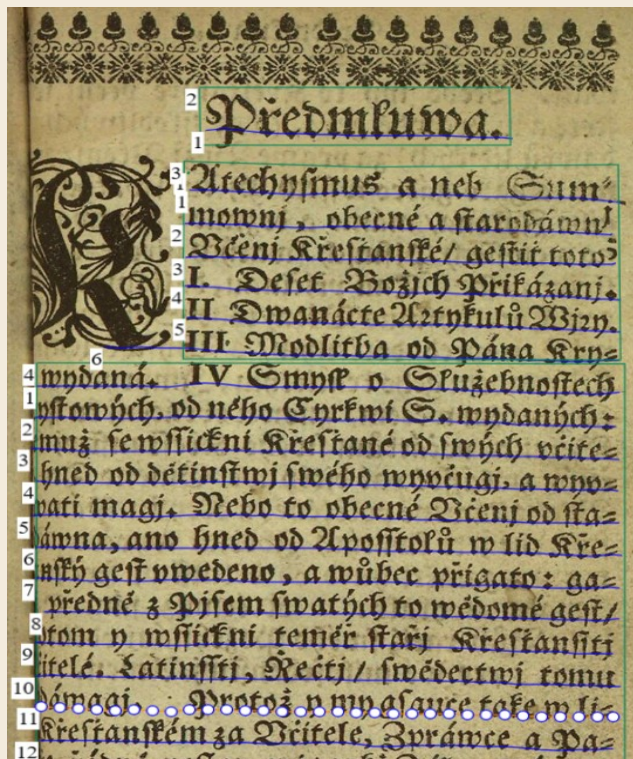
Úvod

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)



1 Předmluwa.

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

- 1 wydanä. IV. Smysl o. Skuzebnostech
- 2 ystowých. od něho Cyrkwí S. wydaných:
- 3 muzi se wssickni Křestane od swych: učite= | |
- 4 hned od dětinstwí swého wyučuj, a wyo=
- 5 wati magj. Nebo to obecné Vcenj od sta= |
- 6 dawna, ano hned od Aposstolů w lid Kre= |
- 7 nský gest wvedeno, a wübec prigato: ga= | |
- 8 předné z Písem swatých to wědomě gest /
- 9 ptom y wssickni teměr staří Křestansiti
- 10 litelé: Latinssti, Recti / swědectwi tomu
- 11 dáwaj. Protož y myjsauče take wli=

České dokumenty – práca študentov na Opavskej univerzite (Taufrová)



- 1 Čert a Prawda.
- 2 To gest:
- 3 welmi pěkně smyšlené, utěšené
- 4 Hystorie a Rozprávky
- 5 w několika stech,
- 6 kteréž se pro vyrazení mysli a pro
- 7 zasmání při dlouhé chvíli, y w každé
- 8 veselé společnosti, časem také y drobet
- 9 pro wybrausení rozumu dobře užije-
- 10 wati mohau.
- 11 Wydané podlé rukopisu
- 12 Hylarya, Jokosa, Astucya.
- 13 Gakozto II. Díl
- 14 k Zrcadlu Possetilostj.

Region 2

Region 3

- 1 Kraméryusowým nákladem.

Region 4

Region 5

- 1 W Praze, 1796.
- 2 kdostání w České Expedycy w Dominy-
- 3 kánské ulicy, u Hrabů w Nie. 373.

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

Arabské pohádky.

2 3 4

První Svazek.

7 8 9 10 11 12 13 14 15 16

I. Nuredyn, Peršský Princ.

**II. Abulfara Kupce podivni příběz
homé na cestách geho.**

20 21

Kraméřusovským nákladem.

22

W Praze, 1795.

23
24
25

† dostání w Čiasté novinářské Expedicích, w
Dominikánské ulici v Praze Kro. 373.

27 26

3^{re}

Liebt die mich besuch, Oja, ich bin bereit, was ich
 Alles gütlichwillig thun, was ich nur thun kann, als
 beifolgt, was ich zu thun an die Meinerer Namen die
 Gedächtnis der mich besuch, lassen können und so
 herzlich ganz?
 ferd, gütlichwillig und so ganz.

Specialis.

1^{re}

Wann? nicht wir oft
 hat die die Form, die mich besuch, und so ganz?
 die ich in die Form, die mich besuch, und so ganz?
 die ich in die Form, die mich besuch, und so ganz?

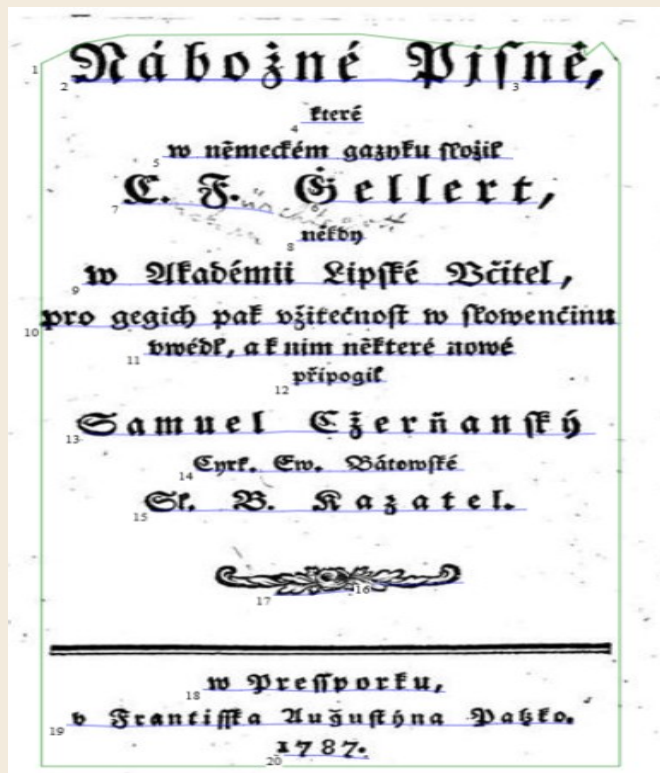
2^{re}

Was hat die die Form, die mich besuch, und so ganz?
 gelassen? nicht die mich besuch, und so ganz?
 welche die die Form, die mich besuch, und so ganz?
 welche die die Form, die mich besuch, und so ganz?

3^{re}

Ich bin bereit, was ich nur thun kann, als
 beifolgt, was ich zu thun an die Meinerer Namen die
 Gedächtnis der mich besuch, lassen können und so
 herzlich ganz?
 ferd, gütlichwillig und so ganz.

České dokumenty – práca študentov na Opavskej univerzite (Gajdošová)

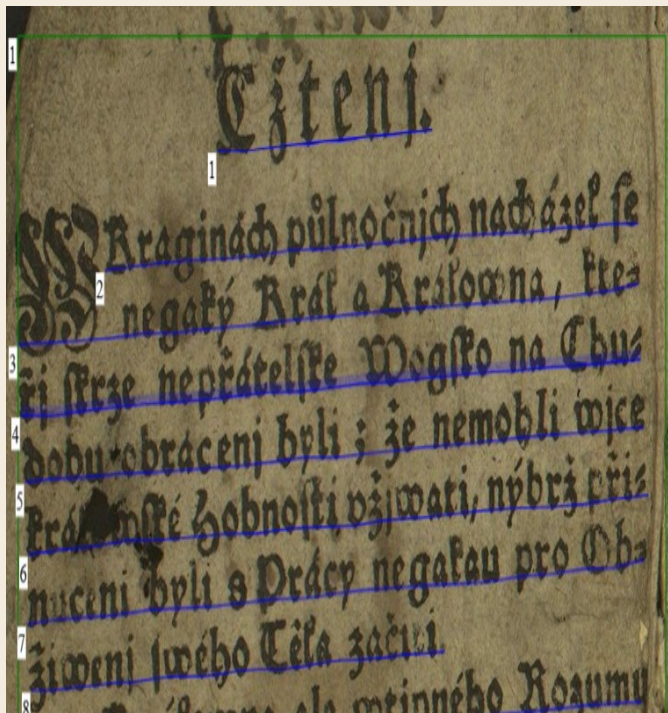


Region 1

Nábožné Písně

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

W Pressporku,
U Frantíška Augustýna Pazko.
1787

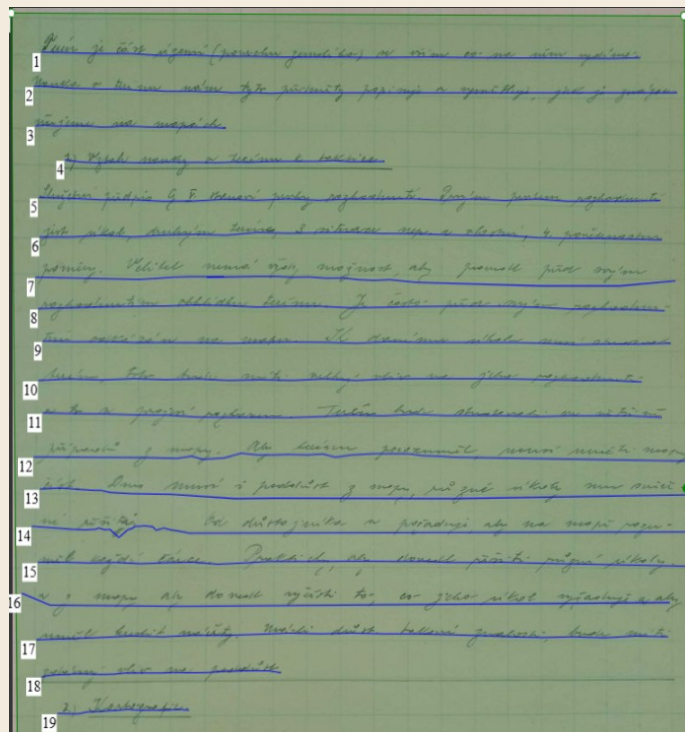


Region 1

Čtení.

W kraginách půlnočnjich nacházel se negaký Král a Králowna, kteří skrze nepřátelske Wogsko na Chudobu obráčení byli ; že nemohli wíce královské hodnosti užjwati, nýbrž přinuceni byli s Prácy negakau pro Obživění swého Těla začíti.

České dokumenty – práce studentů na Opavské univerzitě (Němec)



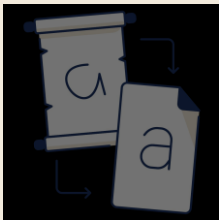
Region 1

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

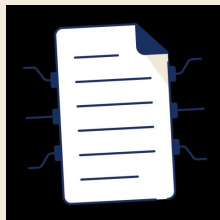
Začínáme: Prvé kroky v Transkribuse

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

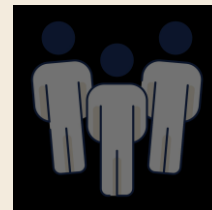
Čo Transkribus umožňuje?



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



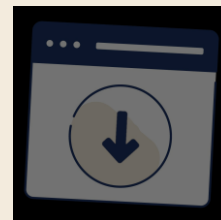
**Trénovanie modelov
umelej inteligencie**



Spolupráca



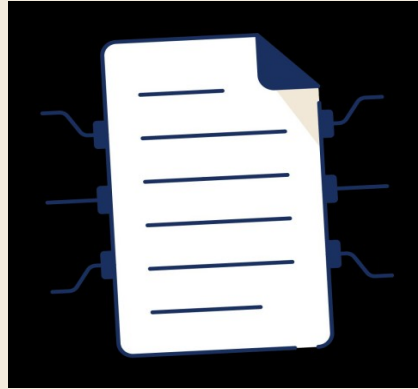
**Tagovanie štruktúry
a obsahu
dokumentov**



**Export dokumentov v
rôznych formátoch**

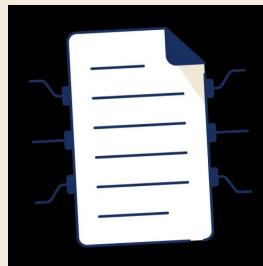
Trénovanie modelov umelej inteligencie

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

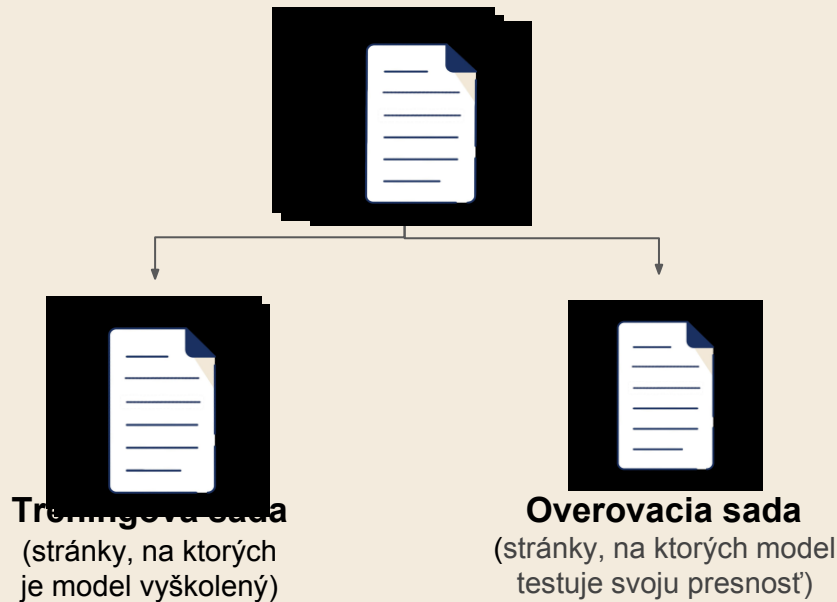
Tréningová sada (Training set)

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)

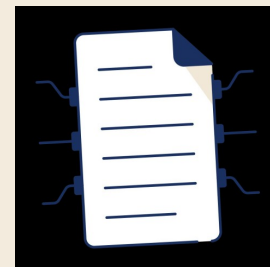


Tréning modelov

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



Trénovanie modelov AI



Modely trénovateľné s Transkribusom:

Text

Riadky

Bloky textu

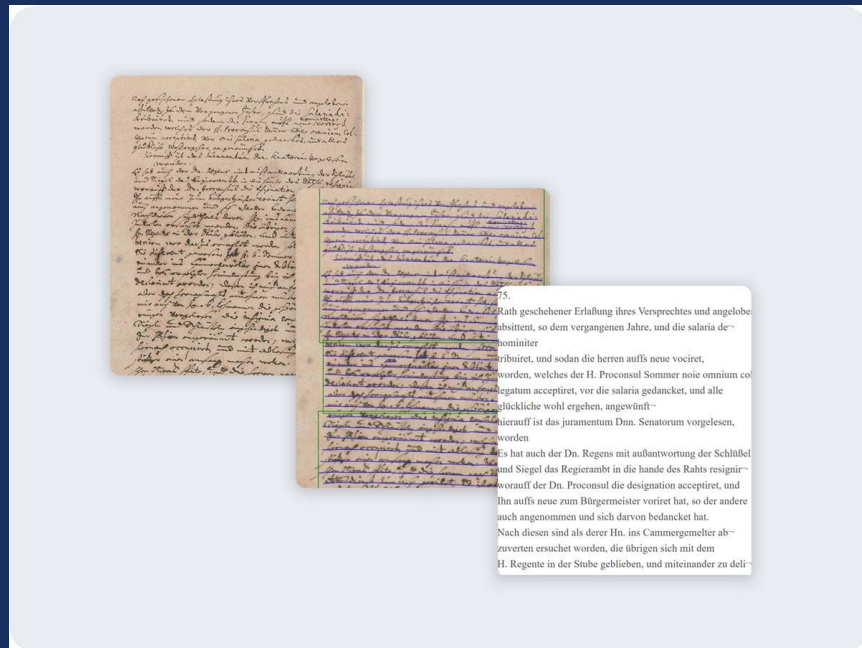
A screenshot of the Transkribus model training interface. It features a list of model types on the left and a '+ Train New Model' button on the right. Arrows point from the labels 'Text', 'Riadky', and 'Bloky textu' to the corresponding model types in the list.

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

+ Train New Model

Analýza rozloženia (segmentácia)

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



Kath. gescheneher Erlaubung ihres Versprechtes und angelobt
abstinent, so dem vergangenen Jahre, und die salaria de-
terminiret
abstinet, und sodan die heren auffz neue vociret,
worden, welches der H. Proconsul Sommer noie omnium co-
legatum acceptiret, vor die salaria gedancket, und alle
Blickliche wohl ergeben, angewünt
hierauff ist das juramentum Dni. Senatorum vorgelesen,
worden
Es hat auch der Dn. Regens mit aufantwortung der Schlüssel
und Siegel das Regieramt in die hande des Rechts resignir-
t worauff der Dn. Proconsul die designation acceptiret, und
Ihn auffz neue zum Bürgemeister vociret hat, so der andere
auch angenommen und sich darvon bedancket hat.
Nach diessen sind als derer Hn. ins Cammergemelter ab-
zuverten ersacht worden, die thbrig sich mit dem
H. Regente in der Stube geblieben, und miteinander zu deli-

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 (tlačný 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

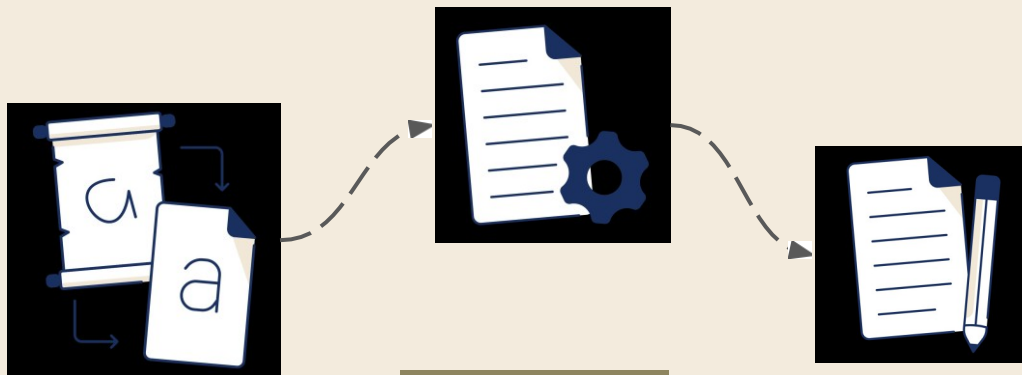
Uložte stránku ako GT "Základnú pravdu"!



Textové modely

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

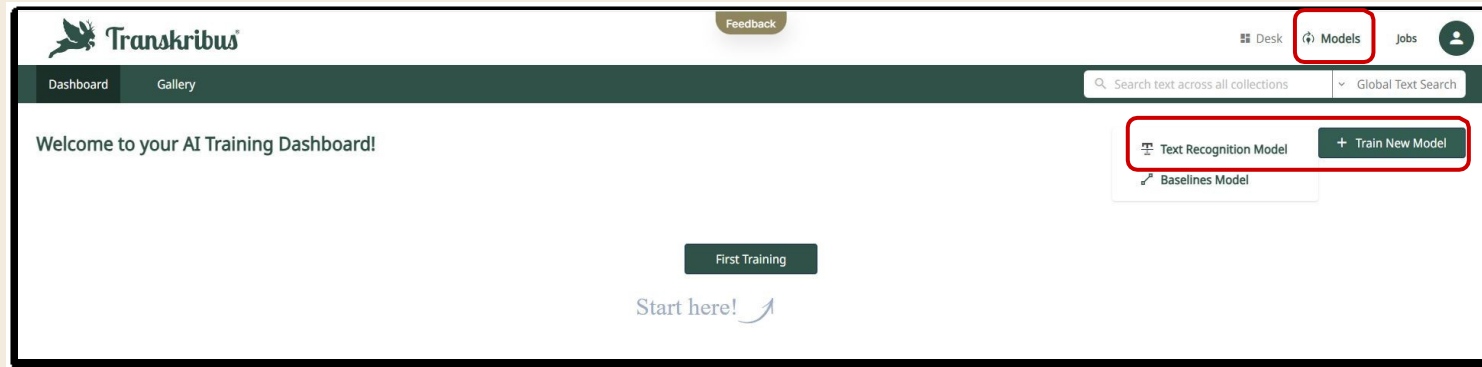
1. Vyberte stránky, ktoré chcete zahrnúť do Základnej pravdy (GT)
2. Spustenie rozpoznávania textu
3. Oprava automatických prepisov
4. Uložte stránku ako "Základnú pravdu,, (GT)



Textové modely

Po vytvorení prepisov (Základná pravda):

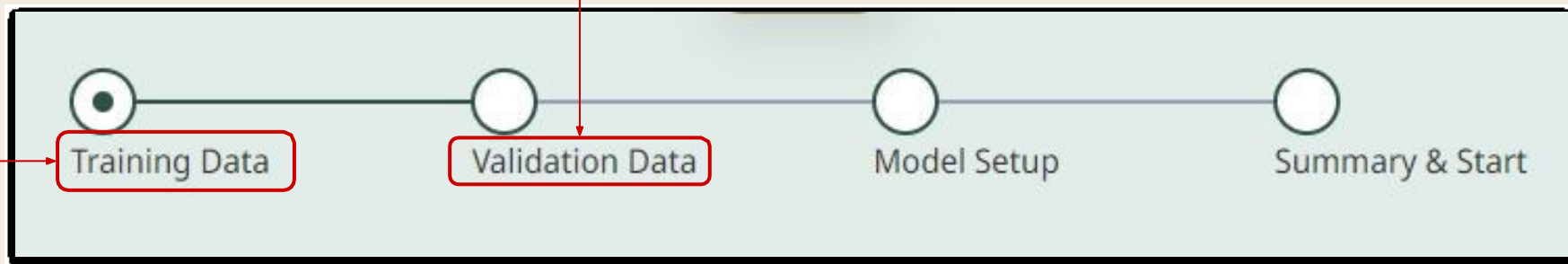
- prejdite do sekcie "Modely"
- kliknite na "Train New Model - Text Recognition Model"
- vyberte zbierku s prepismi (Základná pravda) Ground Truth



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

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

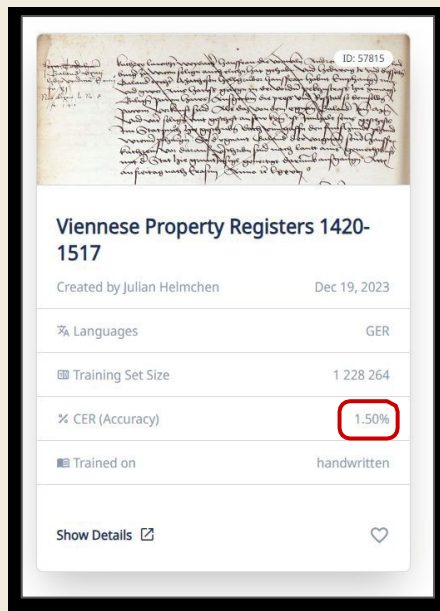
Rozšírené možnosti:

- **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

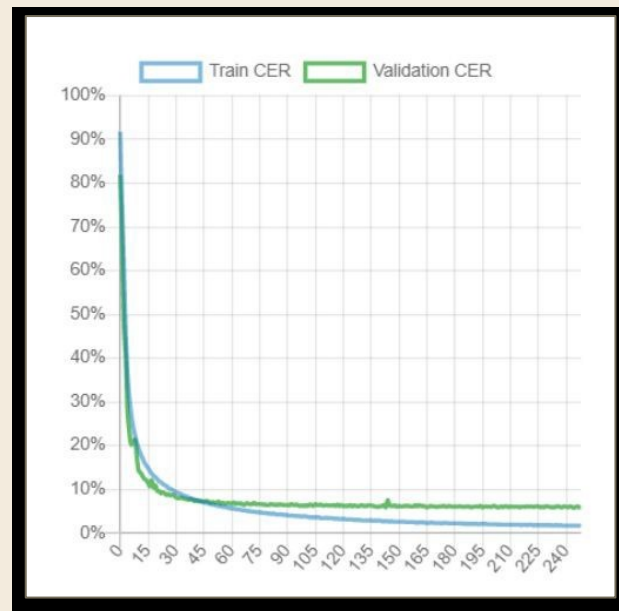
Po dokončení tréningu sa môžete pozrieť na podrobnosti modelu:

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 was 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

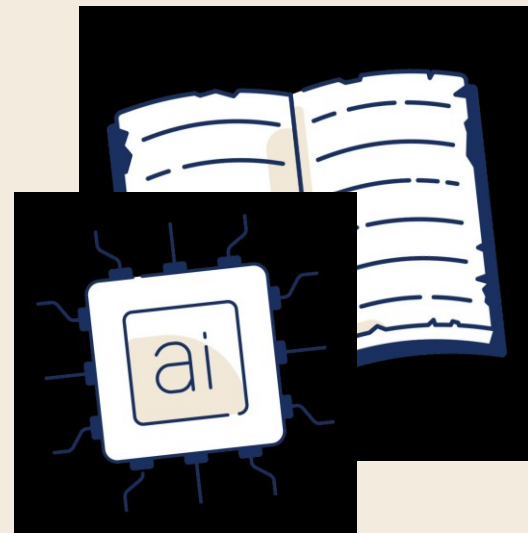


Textové modely

	CER (chybovosť znakov)	Tréningová sada
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

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

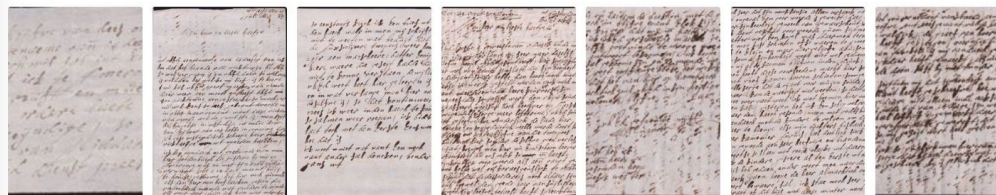
48329

Model description

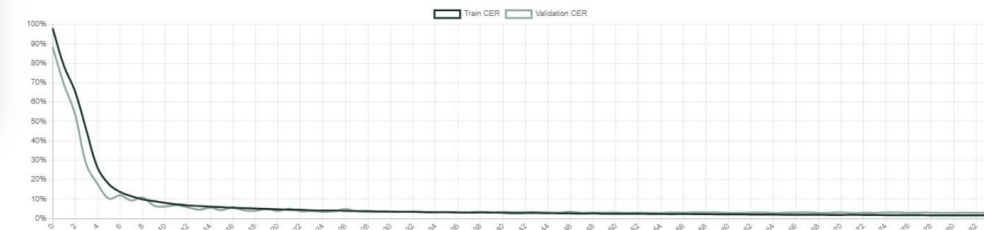
This is the first model created by the Utrecht Archives. It is based on a thousand letters Margaretha Turnor wrote to her husband during the late 17th century. She managed the castle of Amerongen, while her husband worked abroad as a diplomat for the Dutch Republic. Her letters provide an insight into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

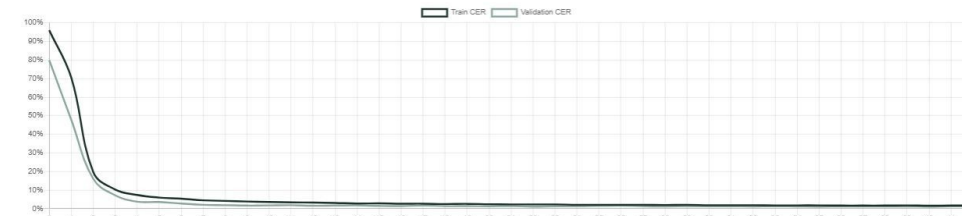
Model description

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.

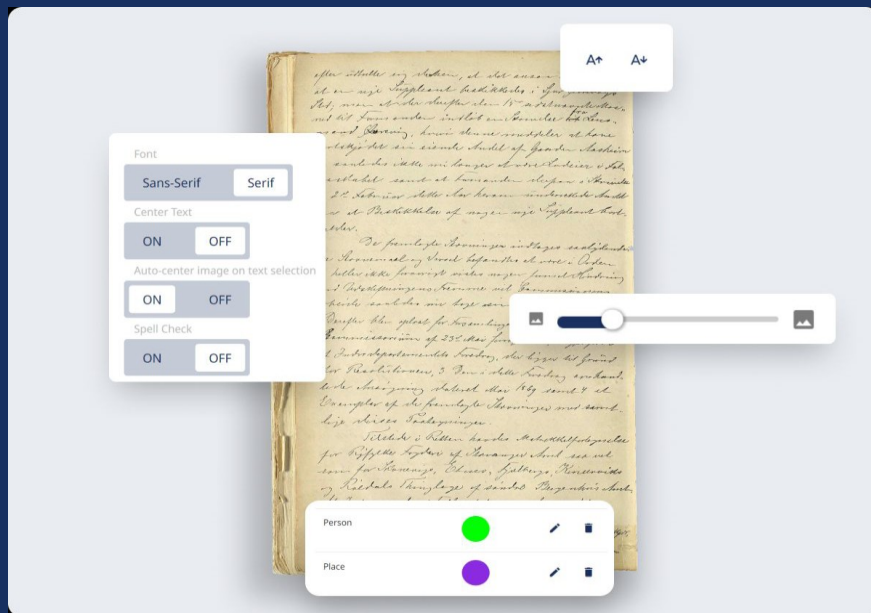
Training data



Training stats



Iteration	Train CER	Validation CER
0	95%	85%
1	85%	75%
2	25%	15%
3	15%	10%
4	12%	10%
5	11%	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):

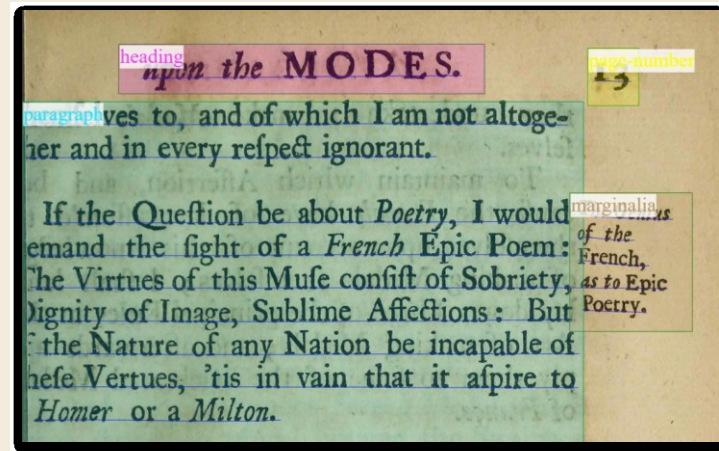
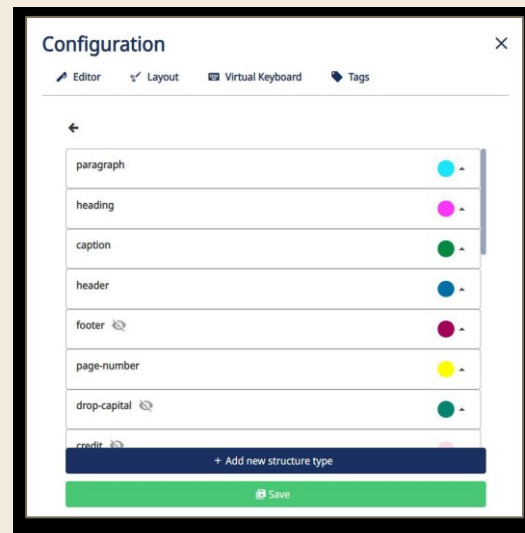
Slúžia na označenie prvkov štruktúry dokumentu

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 underlined 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, the word "place" is entered and underlined in purple. Further down, 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 skrátenu 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

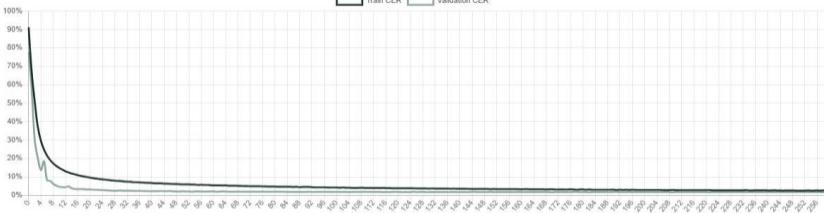
Seventh iteration of the collaborative UCL–University of Toronto model for processing medieval Latin manuscripts, particularly those containing a large quantity of abbreviated words. E-mail: criley@ucl.ac.uk.

Training data



View all >

Training stats



Epoch	Train CER	Validation CER
0	100%	100%
1	~50%	~50%
2	~10%	~10%
3	~5%	~5%
4	~3%	~3%
5	~2%	~2%
6	~1.7%	~1.7%
7	~1.7%	~1.7%
8	~1.7%	~1.7%
9	~1.7%	~1.7%
10	~1.7%	~1.7%
11	~1.7%	~1.7%
12	~1.7%	~1.7%
13	~1.7%	~1.7%
14	~1.7%	~1.7%
15	~1.7%	~1.7%
16	~1.7%	~1.7%
17	~1.7%	~1.7%
18	~1.7%	~1.7%
19	~1.7%	~1.7%
20	~1.7%	~1.7%
21	~1.7%	~1.7%
22	~1.7%	~1.7%
23	~1.7%	~1.7%
24	~1.7%	~1.7%
25	~1.7%	~1.7%
26	~1.7%	~1.7%
27	~1.7%	~1.7%
28	~1.7%	~1.7%
29	~1.7%	~1.7%
30	~1.7%	~1.7%
31	~1.7%	~1.7%
32	~1.7%	~1.7%
33	~1.7%	~1.7%
34	~1.7%	~1.7%
35	~1.7%	~1.7%
36	~1.7%	~1.7%
37	~1.7%	~1.7%
38	~1.7%	~1.7%
39	~1.7%	~1.7%
40	~1.7%	~1.7%
41	~1.7%	~1.7%
42	~1.7%	~1.7%
43	~1.7%	~1.7%
44	~1.7%	~1.7%
45	~1.7%	~1.7%
46	~1.7%	~1.7%
47	~1.7%	~1.7%
48	~1.7%	~1.7%
49	~1.7%	~1.7%
50	~1.7%	~1.7%
51	~1.7%	~1.7%
52	~1.7%	~1.7%
53	~1.7%	~1.7%
54	~1.7%	~1.7%
55	~1.7%	~1.7%
56	~1.7%	~1.7%
57	~1.7%	~1.7%
58	~1.7%	~1.7%
59	~1.7%	~1.7%
60	~1.7%	~1.7%
61	~1.7%	~1.7%
62	~1.7%	~1.7%
63	~1.7%	~1.7%
64	~1.7%	~1.7%
65	~1.7%	~1.7%
66	~1.7%	~1.7%
67	~1.7%	~1.7%
68	~1.7%	~1.7%
69	~1.7%	~1.7%
70	~1.7%	~1.7%
71	~1.7%	~1.7%
72	~1.7%	~1.7%
73	~1.7%	~1.7%
74	~1.7%	~1.7%
75	~1.7%	~1.7%
76	~1.7%	~1.7%
77	~1.7%	~1.7%
78	~1.7%	~1.7%
79	~1.7%	~1.7%
80	~1.7%	~1.7%
81	~1.7%	~1.7%
82	~1.7%	~1.7%
83	~1.7%	~1.7%
84	~1.7%	~1.7%
85	~1.7%	~1.7%
86	~1.7%	~1.7%
87	~1.7%	~1.7%
88	~1.7%	~1.7%
89	~1.7%	~1.7%
90	~1.7%	~1.7%
91	~1.7%	~1.7%
92	~1.7%	~1.7%
93	~1.7%	~1.7%
94	~1.7%	~1.7%
95	~1.7%	~1.7%
96	~1.7%	~1.7%
97	~1.7%	~1.7%
98	~1.7%	~1.7%
99	~1.7%	~1.7%
100	~1.7%	~1.7%

[Example](#)

Skratky

- Model trévaný 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)

خاتك افكارىنى تئور مقصدىله چقاردىمىز بو مجموعه قارشىسنده دهرىن برهمنوتيت دويوز و بولى
قىمىلى بروفىطىه نلقى ايدىوز . اكر . بووظىفه ايلهده ، مملكتك بك محتاج اولدىهي معارى دويمولرىنى
فعاليت دولرىلى . . اوباندىرمقده خدمت ايدىبيليرىنهك بختيارز .
« مکتب جمعی » ، ساغاد ، مطبعه وساتره اجر تئرىك بك بهالى اولدىهي بوزمانده عرفان توليد
اينمك ، هرکس ايجون فاندلى اولمق اوزره ساخه انتشاره آتلمق .

Tréningové modely pre RTL písmo

RTL skripty

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

2 verzie osmanko-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:

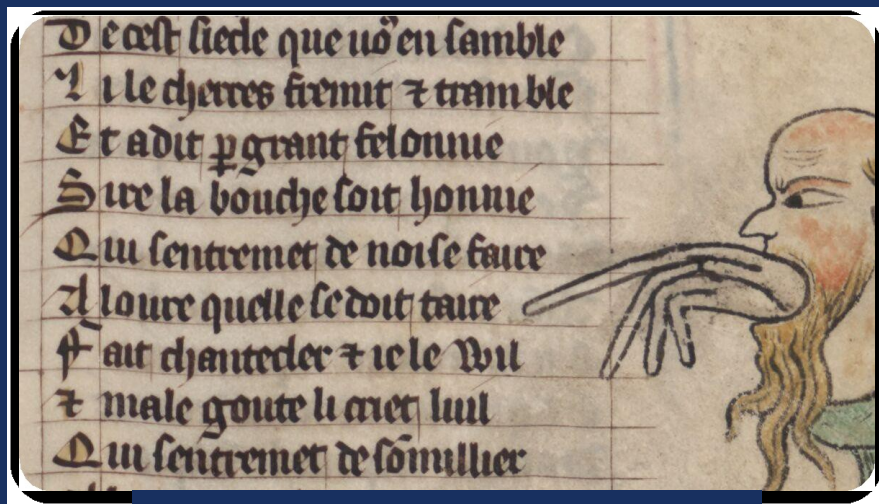
Manuálne spustenie segmentácie (rozpoznávania rozloženia) alebo označovanie rozloženia (oblasti textu + základné čiary) manuálne

- Prepis textu 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

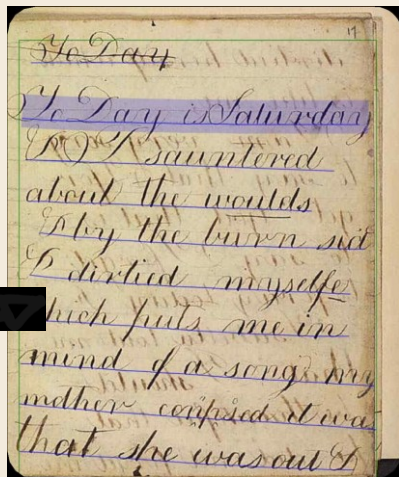


Paris, BnF, Fr. MS 12584 (13th century)

Rozpoznávanie rozloženia (Segmentácia)

Čo sa stane, keď sa stránka rozpozná?

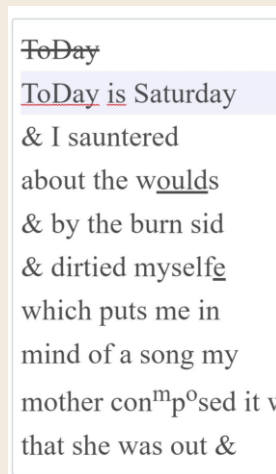
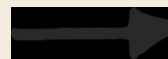
 Recognize



1. krok

Rozpoznávanie rozloženia

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



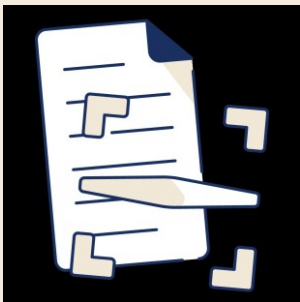
2. krok

Rozpoznávanie textu

Rozpoznávanie rozloženia (segmentácia)

1. krok

Rozpoznávanie rozloženia

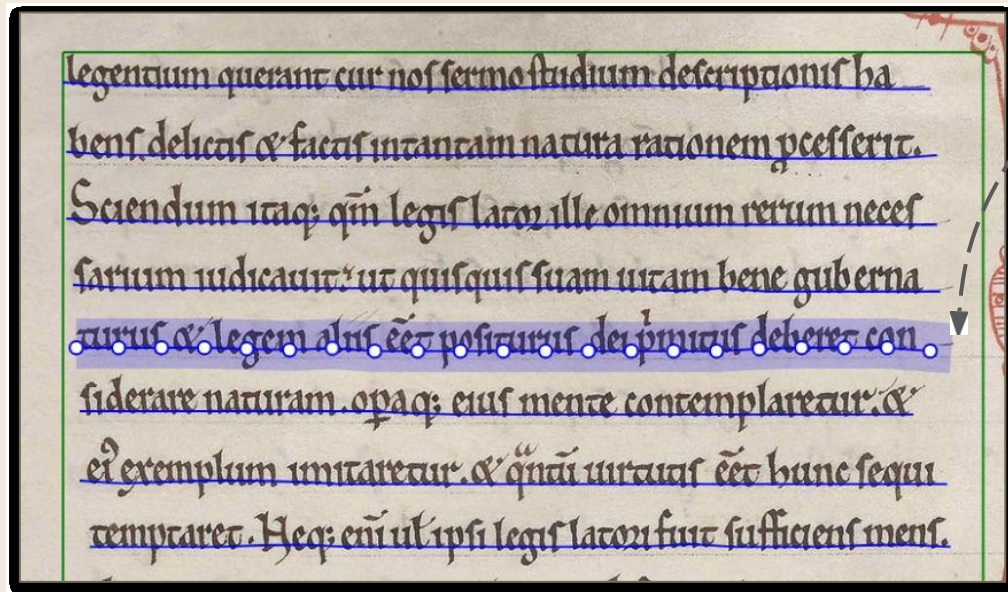


- Analýza rozloženia obrazu dokumentu
- Obrázok je potrebné rozdeliť na textové oblasti a základné čiary
- Základ pre rozpoznávanie a pre transkripciu (prepis)

Tri piliere rozloženia (segmentácie)

1) **Základná čiara** (Baseline):

Členená čiara prebiehajúca pozdĺž spodnej časti riadka rukou písaného textu



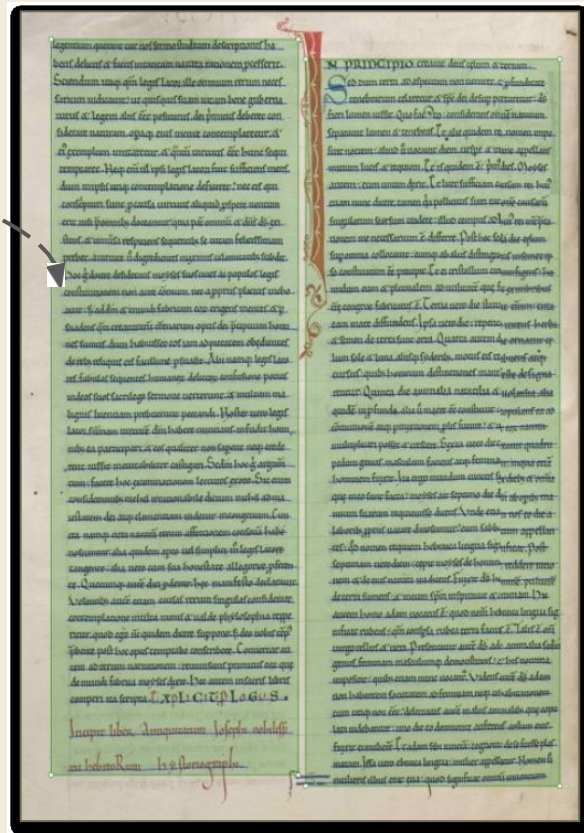
Tri piliere rozloženia (segmentácie)

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))

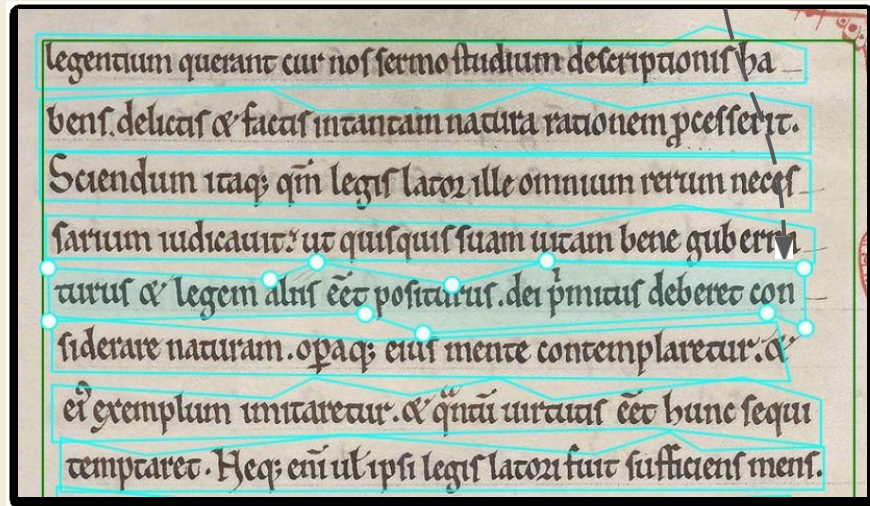
Bloky textu (Text region)



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)

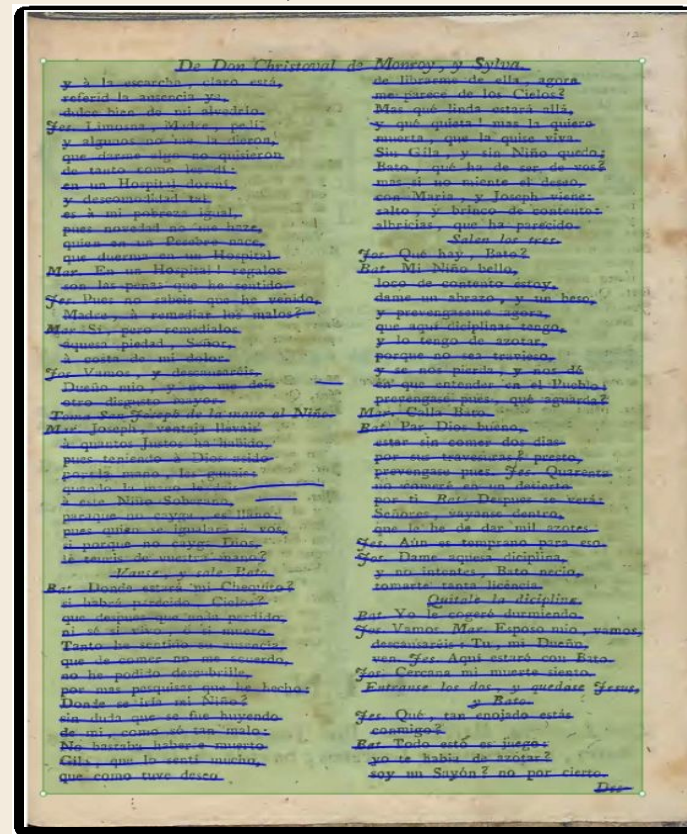


Rozpoznávanie rozloženia (segmentácia)

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 11 / Mr 1836

~~Black Mountain~~

go to town the advice
has been, though Dr Bent
has advised to turn out
out of doors if I
would not. People
are in with our friend
that I did not
go & I will go
have your my
near and more
than 20 times.

to me to have
to Adeney it be
I will you &
believe me.
your my
after. - A kind
Oscar Weston.

Nepresné základné čiary

Nepresné základné čiary

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

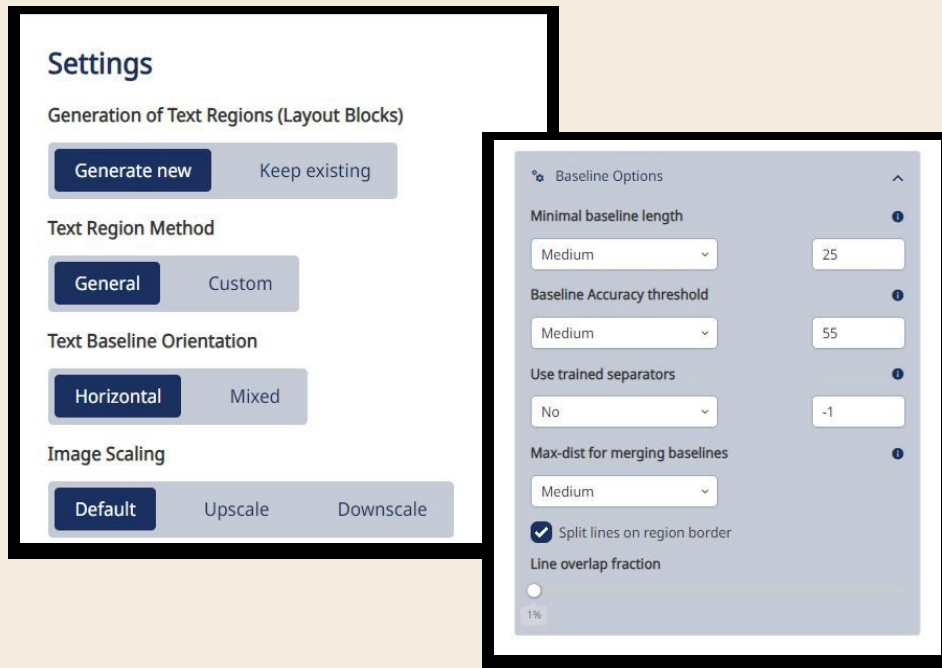
1) Použitie iného verejného modelu základných čiar (Baseline model)

:

- 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)



The image shows two overlapping panels from a software settings interface. The left panel, titled 'Settings', contains general configuration options. The right panel, titled 'Baseline Options', shows advanced settings for text baseline generation.

Settings

Generation of Text Regions (Layout Blocks)

Generate new Keep existing

Text Region Method

General Custom

Text Baseline Orientation

Horizontal Mixed

Image Scaling

Default Upscale Downscale

Baseline Options

Minimal baseline length: Medium (25)

Baseline Accuracy threshold: Medium (55)

Use trained separators: No (-1)

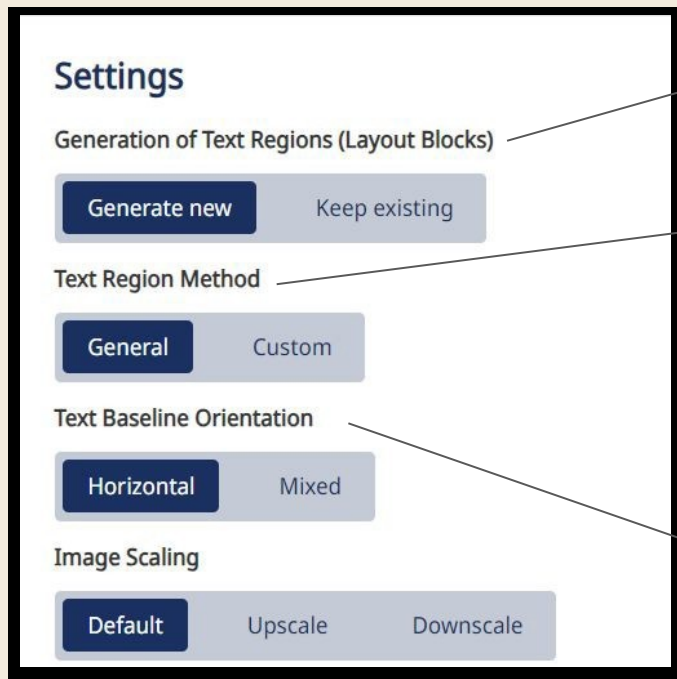
Max-dist for merging baselines: Medium

Split lines on region border

Line overlap fraction: 1%

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:

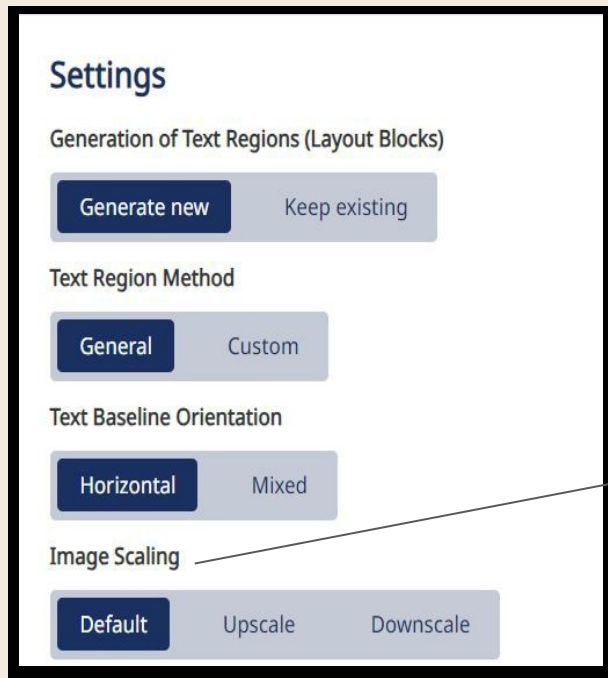
General (Všeobecné): zoskupí čiary zľava doprava

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)

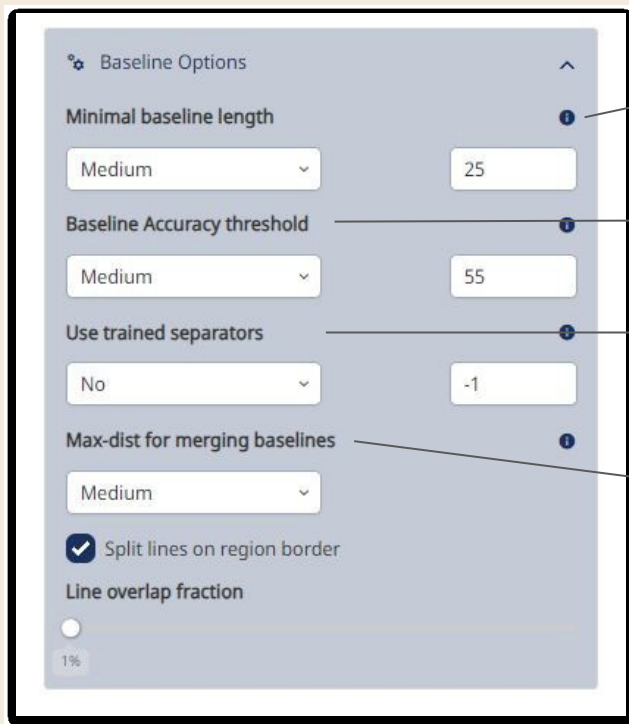


Škálovanie obrázka:

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 tréovaných 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%

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

Iba ak zachováte existujúce bloky textu:

Delené čiary na hranici regiónu:

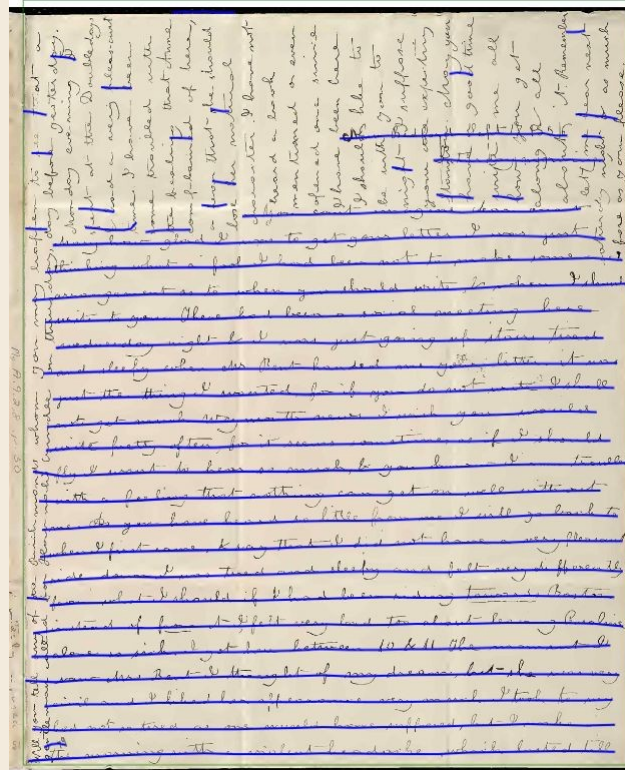
Aby čiary striktno dodržiavali hranicu regiónu.

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>		

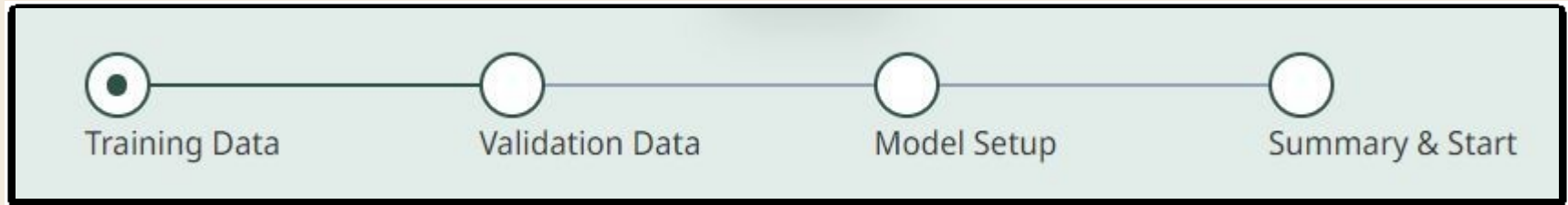
Tréning modelu základných čiar (Baseline Model)

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)

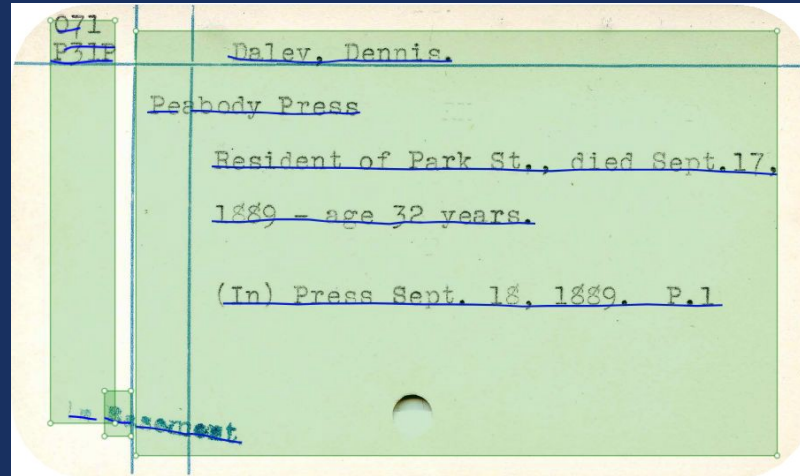
The screenshot displays the Text Recognition interface. At the top, there is a 'Text Recognition' header with a 'Layout' tab highlighted in a red box. Below this, a search bar contains 'NL-RISA_199_226' and a 'Start Recognition' button. On the right, 'Credits needed: 0.00' and 'Available: 0.00' are shown. The main area is divided into two sections: a list of models on the left and a detailed view of a selected model on the right.

Model List:

NAME	WORDS
[Redacted]	v3
[Redacted]	v2
[Redacted]	v1
Medieval manuscript with glosses	2 366
Notes and miscellaneous materiel	8 468

Model Details (v2):

- Private Model (ID: 4880)
- by s.mansutti@readcoop.eu
- 19/12/2022
- Languages
- Training Set Size
- CER (Accuracy): 5.19%
- Trained on: handwritten

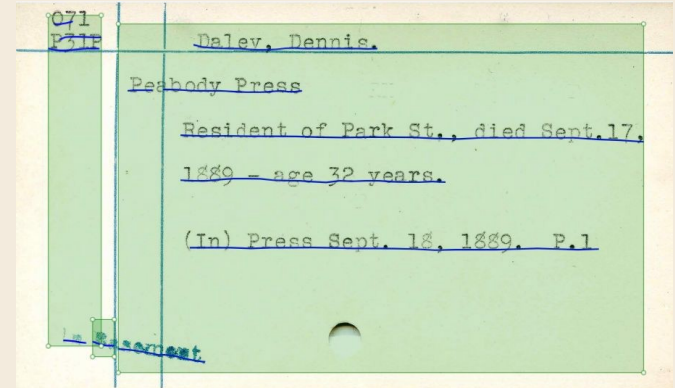
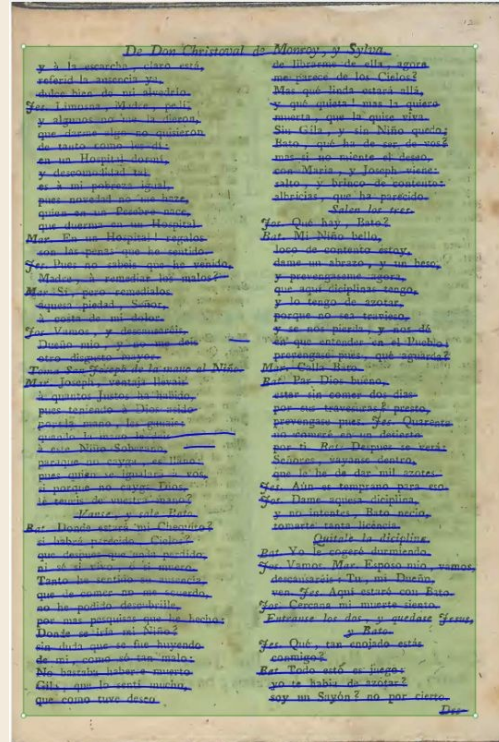


Nepresné bloky textu

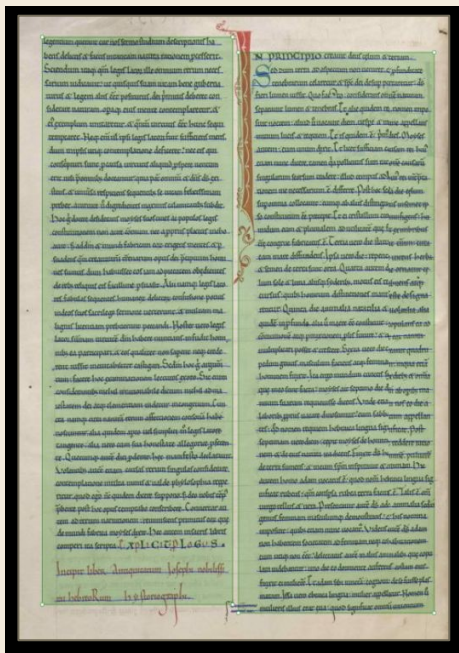
Nepresné bloky textu

Example 1

Example 2



Analýzy rozloženie/segmentácia (rozpoznanie textu)



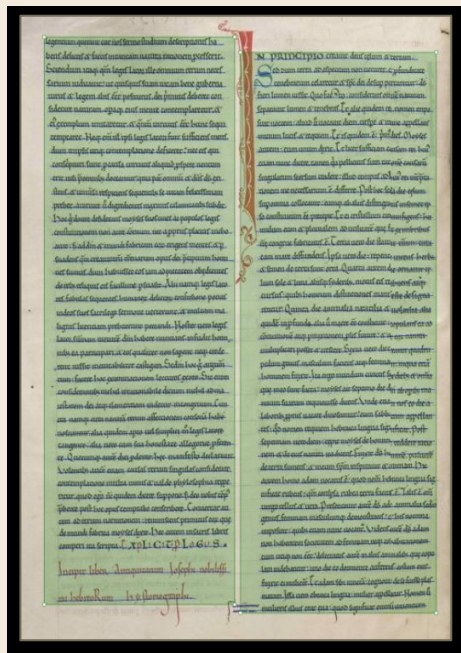
Bloky textu:

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)



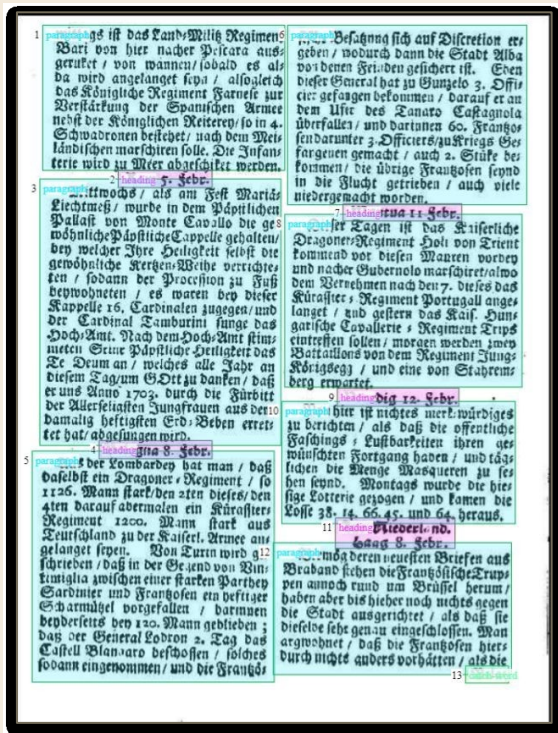
Bloky textu:

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 poľa (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>Lindler</i>		
seit 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. Kant. Lager					

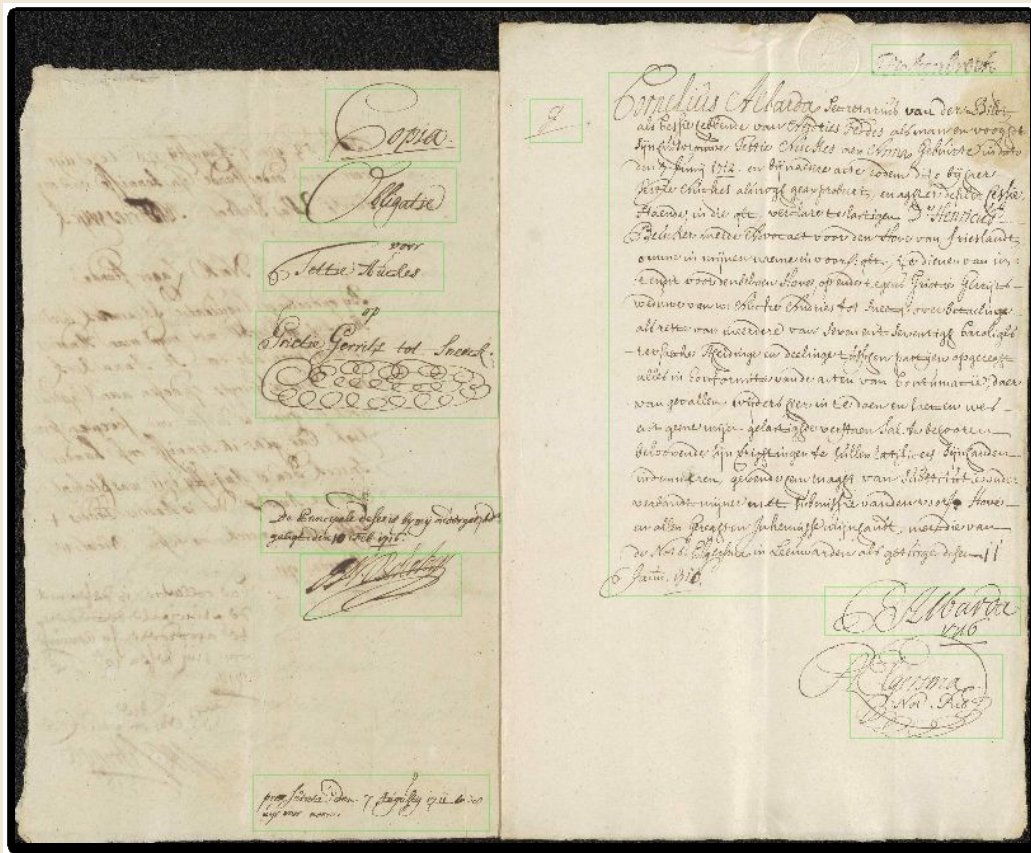
Modely poľa (Beta)

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		Linsler
	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% 1907

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: Beruf 100% 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%

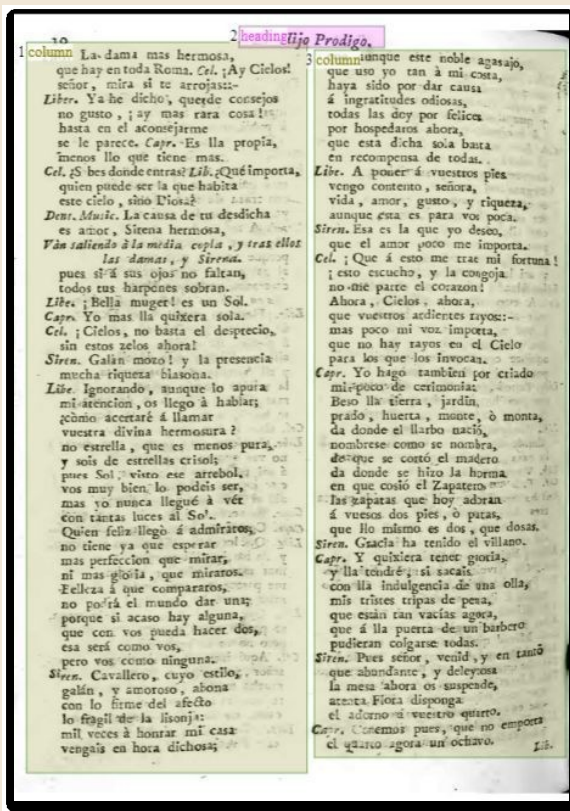
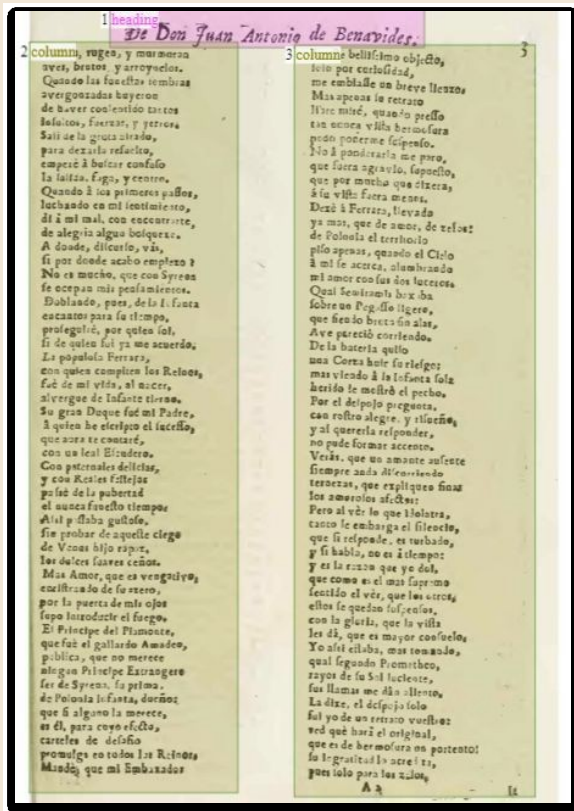
in verheiratet mit geboren in 100%

Wohnung: Wohnung 98%

Vordruck Nr. 304b (Wahlbild unvollst.) 9.48 380 000

Dr. 150-Nr. 945 Staatsdruckerei Berlin 2706

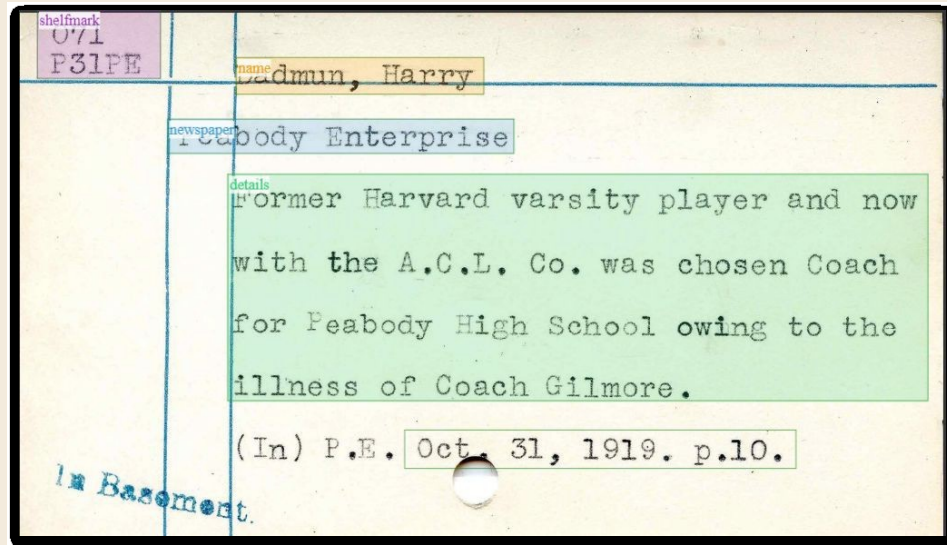
Viac stĺpcové rozloženie textu



Modely poľa (Beta)

Pripravte si cca 50 strán tréningových dát:

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



Modely poľa (Beta)



Desk

Models

Sites

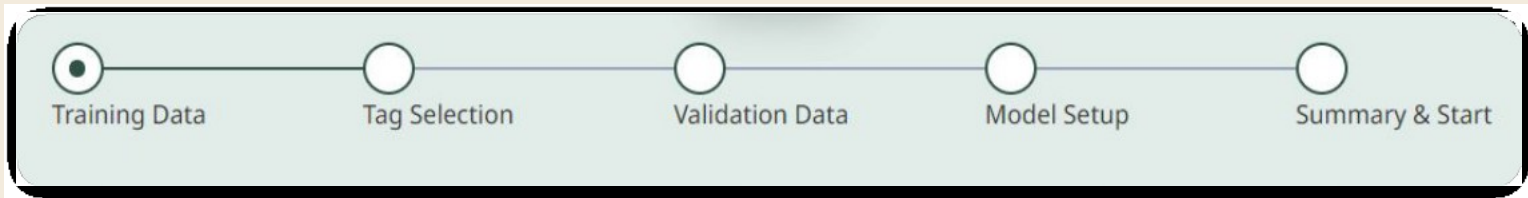
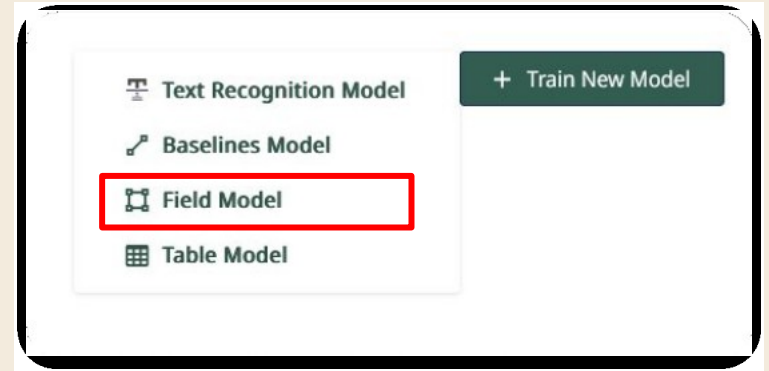
Jobs



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 poľí:**

- 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 vyškoli, → 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

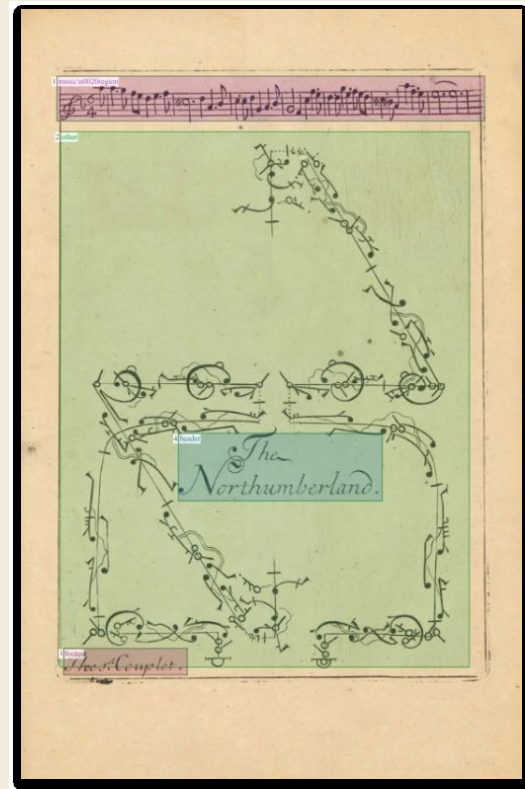
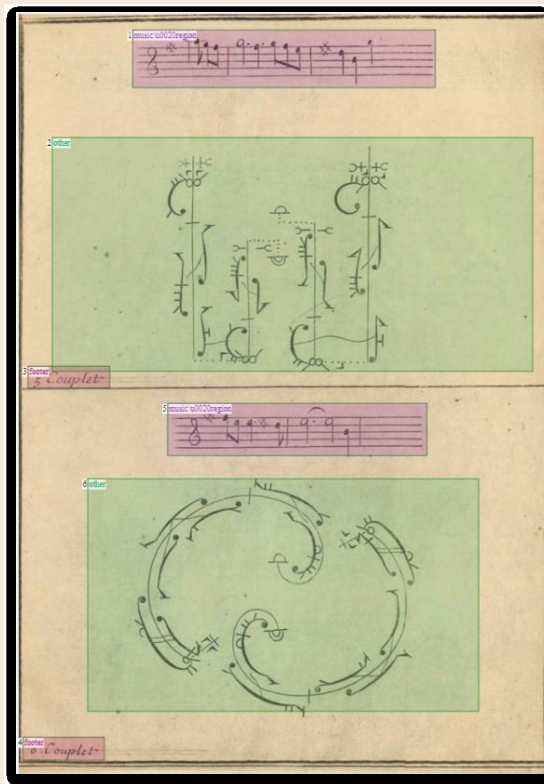
Príklady Modelov polí

Ground Truth:

30 strán

5 tagov

[Example](#)



Príklady Modelov polí

Example

1 **Header**
Anno 1746. (Num. 17.) 26. Februarii.
Wienerisches
DIARIUM.
 10 **Imprimt** Ihrer K&Misch: Kaiserl., auch zu Singarn, und W&heim K&ngl. Maj. Streybeten
In dem neuen Michaeler-Haus/ bey Joh. Peter v. Spelen.

4 **Imperator-singl**
 5 **beobacht** **Italien.**
Genna 29. Jan.
 6 **ergangen**
 7 **beobacht**
 8 **beobacht**
 9 **beobacht**
 10 **beobacht**
 11 **beobacht**
 12 **beobacht**
 13 **beobacht**

Ergangenen Samstag kamen abermalen zwey Catalonische Schiffe mit 3600 Spanischen Soldaten von Villa franca und Sonnens tags Abends eine unferige Feluke 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. fruhe in diesem Gegebenen ein Engländisches Krieges-Schiff mit vier ektigen Flaagen sich anken gesehen. Briefe von Barcello unter dem 15. dieses geben / das selbst nach und nach immerhin Drucken und Fahr- Züge antommen / weiche jetzerzeit gar bald nach hiesigen Gegenden befördert werden : es sollen auch würllich einige Cavallerie- Regimente zur Verstärkung der Spanischen Armee allschon nach unsrer Stränken im würllichen Anmarck seyn. Seit einigen Tagen hat man angefangen die von Savonia nebst vielen andern Krieges-Geräte hieher gekommenen Artillerie zu der Spanischen Armee nach der Kommanden abzuschicken. So seynd auch Dienstag 4. Sambdich / Schiffe / und 5. andere Neapolitanische Fahr-Zeuge in 11. Tagen von Gaetra mit Artillerie und andern Krieges-Vorrat hiet an-

zu kommen ; Auch seynd auf einer Französischer Tartane 150. Mann von dem Französischen Regiment Lothringen von Villa franca hier angelangt / von welchem Ort selbe zugleich mit 3. andern mit derley Truppen beladenen hiet her bestimmten Schiffen abgefeglet. Eine Kroenische Pinke / so in 12. Tagen von Marfilen 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. Stuck-Kugeln alhier an.

W&ch Gelegenheit da dieser Tag beyder Königl. Majestäten zu Vortice zu Mittag geseisset / wurde unter diesen dem Russischen Reichs / Wices Konsten Grafen von Woronzow nebst seiner Frauen Gemahlin der gantz hiesige König. Pallast / samt allen Kleinodien / wie auch denen Königlich den Vermählungs-Kutschen / und alle übrige kostbare Einrichtung gezeiget. Vergangene Woche ist eine hiesige wohl ausgerüstete Volleotte mit 15000. Mann zu Bezahlung unserer Truppen in der Lombardey von hier nach Genna abgedisct worden. Gestern Vortice

1 **beobacht**
 2 **beobacht**
 3 **beobacht**
 4 **beobacht**
 5 **beobacht**
 6 **beobacht**
 7 **beobacht**
 8 **beobacht**
 9 **beobacht**
 10 **beobacht**
 11 **beobacht**
 12 **beobacht**
 13 **beobacht**

ist das Land-Milch Regimente Bari von hier nach Picara ausgezuckert / von wannen / sobald es als da wird angelangt seyn / also gleich das Königl. Regiment Farneo zur Verstärkung der Spanischen Armee nebst der Königl. Reiteren / so in 4. Schwadronen beisset / nach dem Weislandischen marschiren solle. Die Infanterie wird zu Mer aberschicket werden.

Bestimmung sich auf Discretion ergeben / wodurch dann die Stadt Alba von denen Freuden gesehet ist. Eben diese General hat zu Gungelo 3. Officier gefangen bekommen / darauf er an dem Ufer des Tanato Casagnola überfallen / und darinnen 60. Frankosen darunter 3. Officiers / zu Kriegs Gefangen gemacht / auch 2. Stübe der Kommen / die übrige Frankosen seynd in die Flucht getrieben / auch viele niedergemachet worden.

am Fest Mariæ Liechtenst. / wurde in dem Päpstlichen Pallast von Monte Cavallo die gewöhnliche Päpstliche Capelle gehalten / bey welcher Ihre Heiligkeit selbst die gewöhnliche Kircheng-Weih verrichteten / sodann der Procecion zu Fuß beynaheten / es waren bey dieser Kapelle 16. Cardinale zugegen / und der Cardinal Tamburini fungte das Hoch-Alt. Nach dem Hoch-Alt stimmeten Seine Päpstliche Heiligkeit das Te Drum an / welches alle Tage an diesem Tagum G&D zu danken / doch er uns Anno 1702. durch die Fürbitte der Allerheiligsten Jungfrauen aus dem damalig bestigsten Erd-Wehen errettet hat / abgefungen wird.

der Kombardey hat man / 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 seyn. Von Turin wird geschrieben / daß in der Gegend von Buntimiglia zwischen einer starken Partbey Sardiner und Frankosen ein heftiger Scharmüel vorzuefallen / darinnen beyderseits der 120. Mann geblieben ; das der General Lodron 2. Tag das Castell Pianaro beschoosen / solches sodann eingenommen / und die Franck-

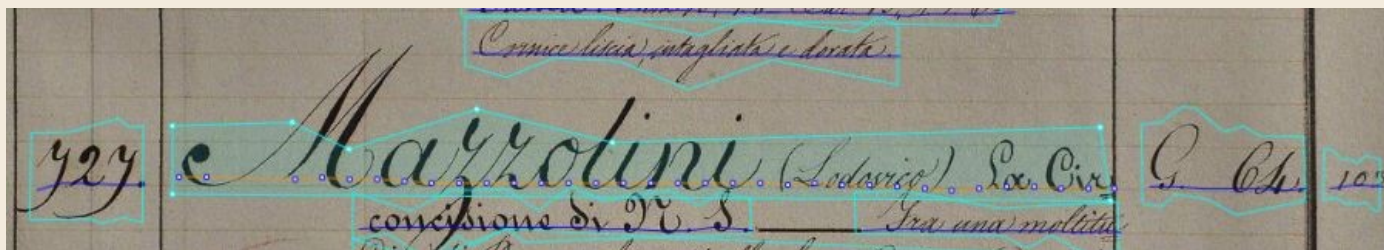


Nepresné polygóny (mnohouholníky)

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 (ans. 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

NAME OF INSTITUTION	TOWN	AMOUNT	OBJECT	DATE	DISPOSITION
Lutheran Ladies' Seminary	Reading, Minn.		Sanctuary	1/10/0	Not sufficiently developed, 2/10/07
Lombard College	P. Julesburg, 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, 1/10/0, 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	2/10/0	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

NAME OF INSTITUTION	TOWN	AMOUNT	OBJECT	DATE	DISPOSITION
Lutheran Ladies' Seminary	Reading, Minn.		Sanctuary	1/10/0	Not sufficiently developed, 2/10/07
Lombard College	P. Julesburg, 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, 1/10/0, 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	2/10/0	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	Yellow 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, Miss.		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:

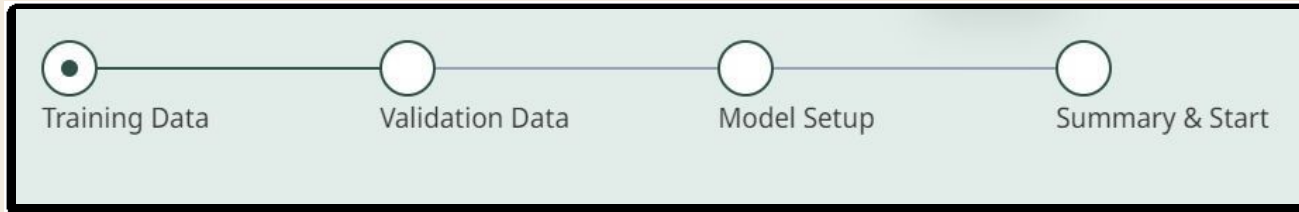
- **Jednoduché dabuľky:** 20 strán GT
- **Ťažké tabuľky:** 50 strán GT
- **mix rôznych tabuliek:** 50 až 100 strán GT v závislosti od počtu tabuliek

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, 5/14/11	
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	12/12/11	Not sufficiently developed, 1/1/11	
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, 10/15/11	
Albert Lea College	Pell City, Ala.		Endowment	4/18/12	Denominational, 11/2/11	
Anderson College (1)	Anderson, S.C.		General	8/15/12	D + Low	
Alabama University of	University, Ala.		Memorial Building	10/2/12	State institution, 1/15/109	

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 hand	12/21/12	Low; Gen 1914, 4/15/14
Alabama School of Trade & Ind.	Rayland, Ala.	25 000	lands	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)		P Emporia, Kan.			Organ	7/24/14	No organs for institutions, 10/6/14				
Elgin Academy (2)		Elgin, Ill.			Endowment	10/16/14	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)		P Elmira, N.Y.	20	000	Library Building	1/4/16	"Gen 1914" 2/7/16				
			or 25	000							
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

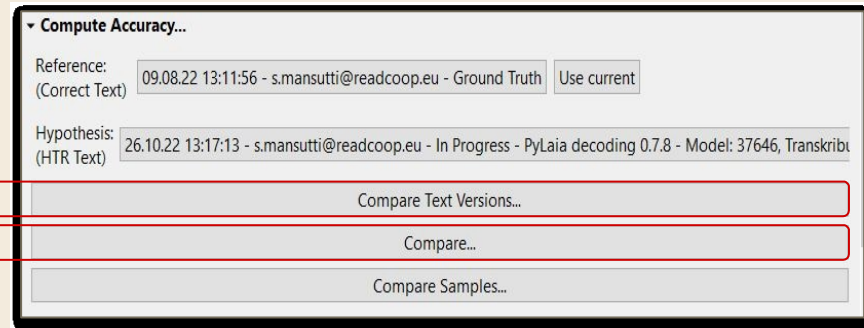


Výpočty presnosti transkripcie

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:

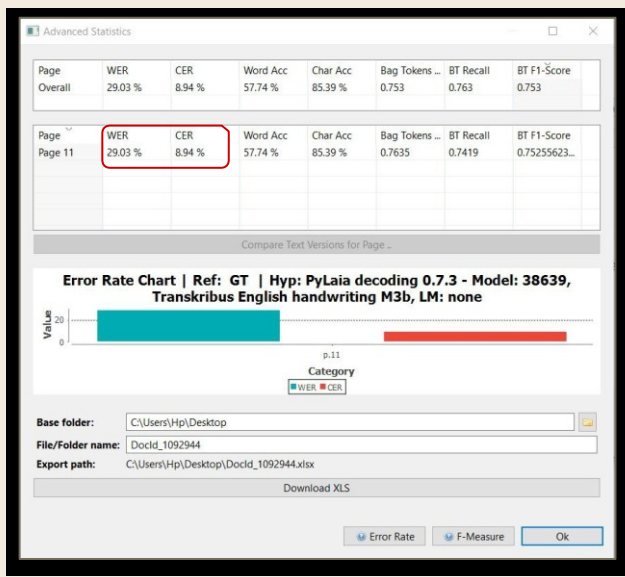


Ground Truth - "Transkribus anglický
rukopis M3b" model s jazykovým
modelom:

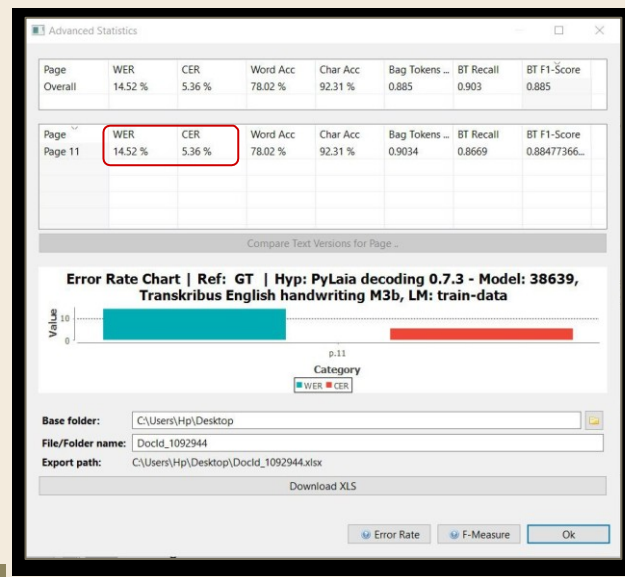


Výpočty presnosti transkripce

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
línii

Kontrola kvality

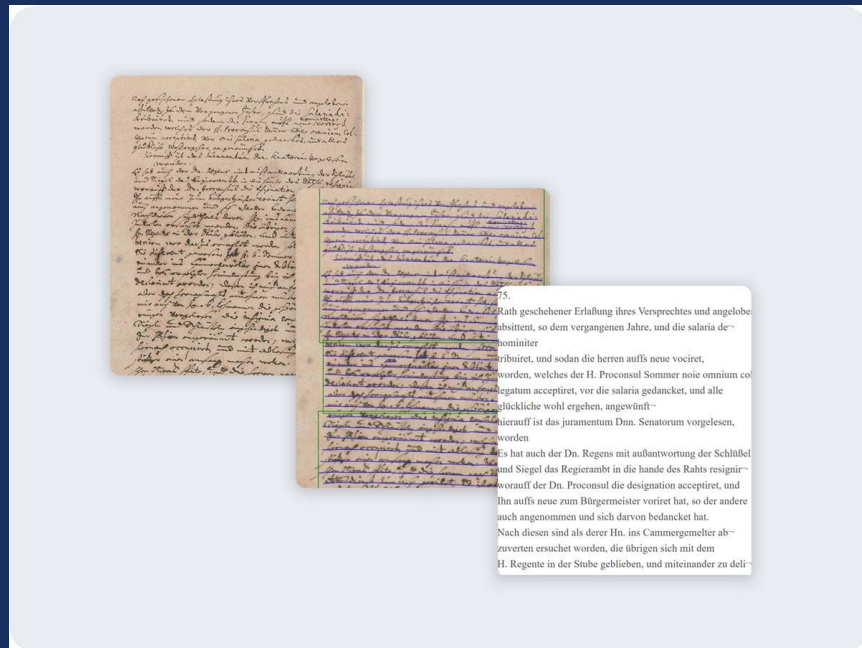
The screenshot displays the Transkribus Quality Control interface. At the top, the Transkribus logo is on the left, and navigation links for Desk, Models, Sites, Connect, and Jobs are on the right. Below the navigation bar, the breadcrumb path is "Quality Control > Sample 01 > Task 01".

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 prídanie 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

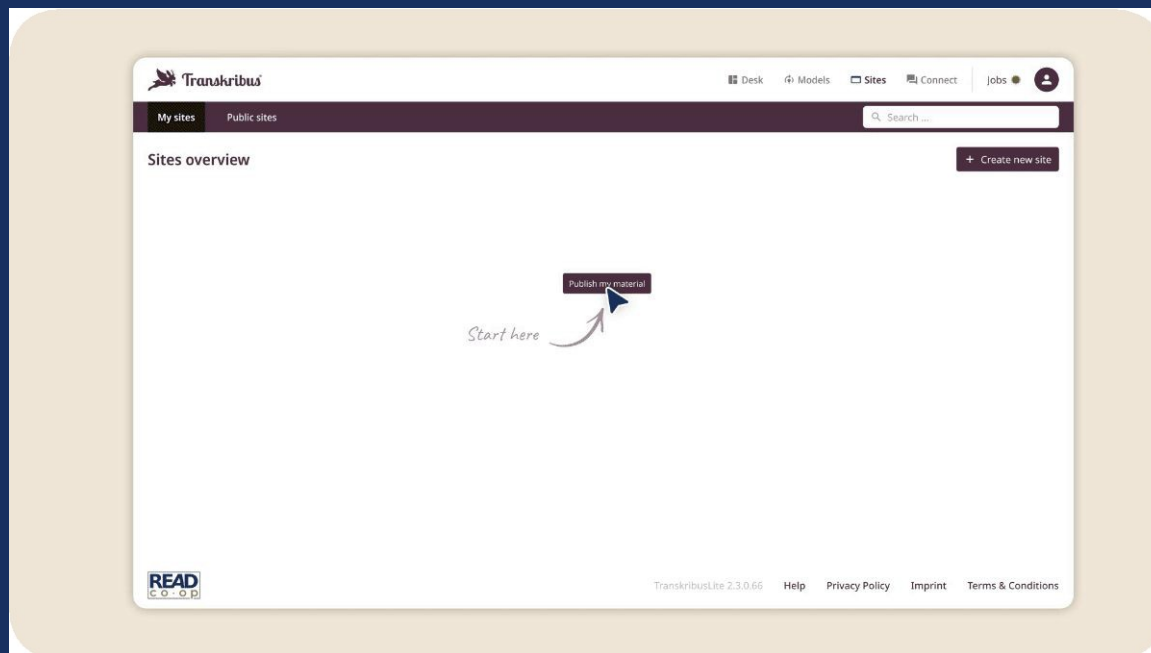
 **Connect**

Jobs 



Plánované na rok 2024

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



Transkribus stránky



 Desk

 Models

 Sites

Jobs



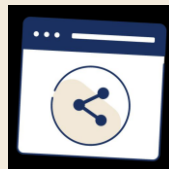
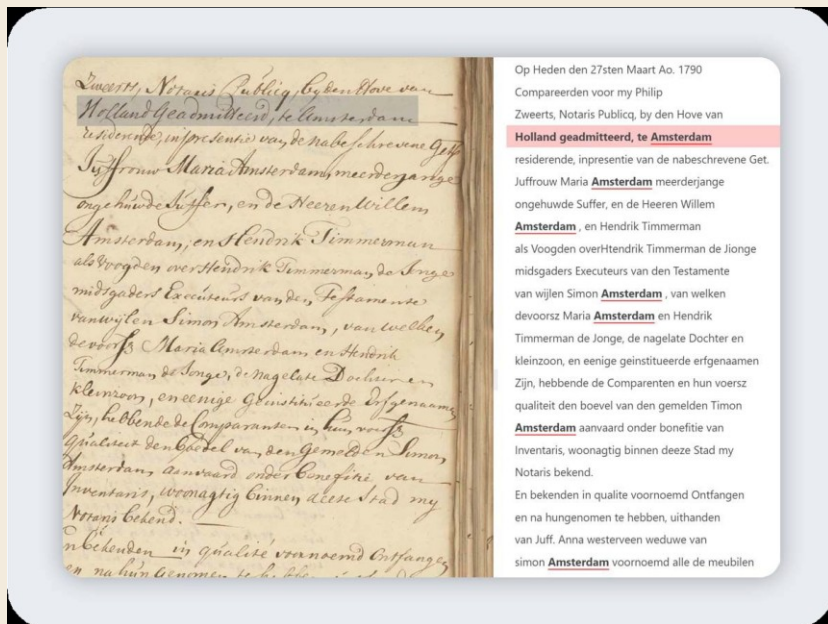
Transkribus **Connect** je miesto, kde sa **exchange** stane.

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



Jednoduché zdieľanie materiálu

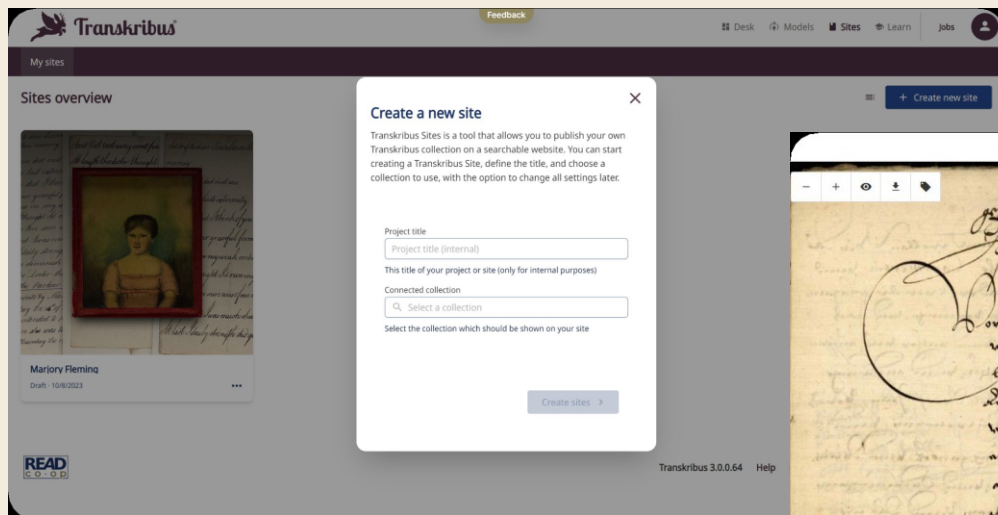


Pohľad strana vedľa strany (obrázok-prepis)



Vylepšené možnosti vyhľadávania

Transkribus stránky



The screenshot shows the Transkribus interface with a 'Create a new site' modal dialog open. The dialog has a close button (X) in the top right corner and a '+ Create new site' button in the top right corner of the background. The dialog 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.'

Project title

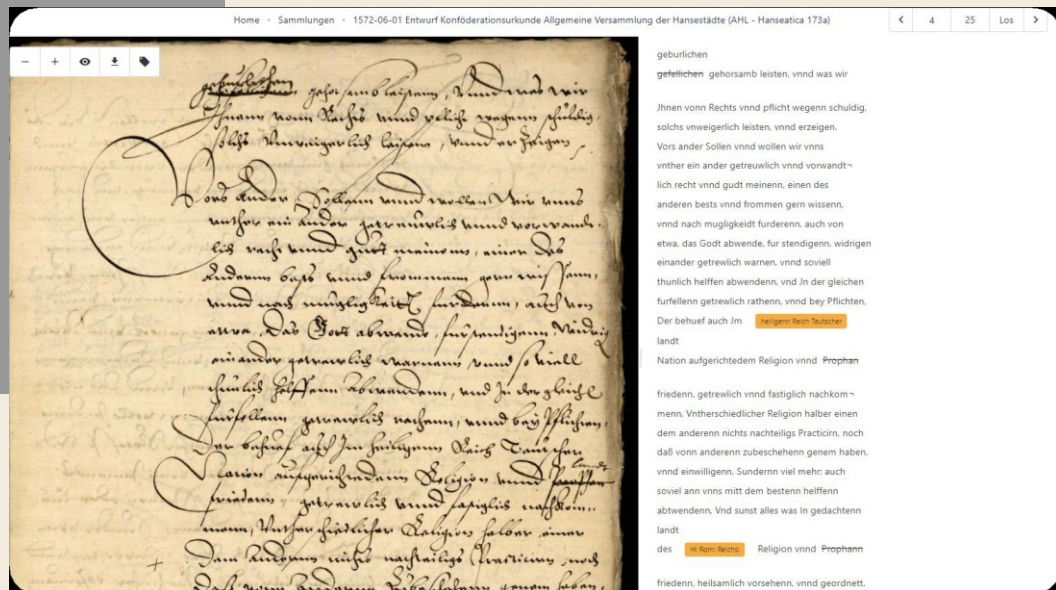
This title of your project or site (only for internal purposes)

Connected collection

Select the collection which should be shown on your site

[Create sites >](#)

Transkribus 3.0.0.64 Help

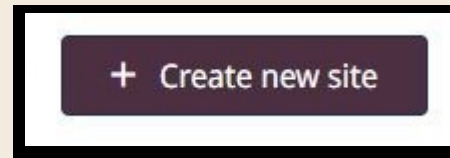


The screenshot shows a document viewer interface for a handwritten manuscript. The document is titled '1572-06-01 Entwurf Konföderationsurkunde Allgemeine Versammlung der Hansestädte (AHL - Hanseatica 173a)'. The viewer includes navigation controls (back, forward, search, zoom) and a page number '4' of '25' pages. The manuscript text is in German, written in a cursive script. The viewer also displays a 'geburlichen' (birth) record on the right side, which is transcribed from the manuscript. The text on the right reads: 'geburlichen gehorsamb leisten, vnnnd was wir solchs vnnweigerlich leisten, vnnnd erzeigen. Vnther ein ander getrewlich vnnnd vorwandtlich recht vnnnd gndt meinnen, einen des anderen bests vnnnd frommen gern wissenn, vnnnd nach muglichkeit fuderenn, auch von etwa, das Godt abwende, fur stendigen, widrigen einander getrewlich warnen, vnnnd soviel thunlich helffen abwendenn, vnnnd In der gleichen furtellenn getrewlich rathenn, vnnnd bey Pflichten. Der behuef auch im heiligen Reich Nutscher Nation aufgerichtetem Religion vnnnd Prophan friedenn, getrewlich vnnnd fastiglich nachkommenn, Vntherschiedlicher Religion halber einen dem anderen nichts nachtheiligs Practicirn, noch daß vonn anderen zubesehenn genem haben, vnnnd einwilligen, Sundern viel mehr: auch soviel ann vnns mitt dem bestenn helffen abwendenn, Vnnnd stund alles was In gedachtem landt des heiligen Reichs 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

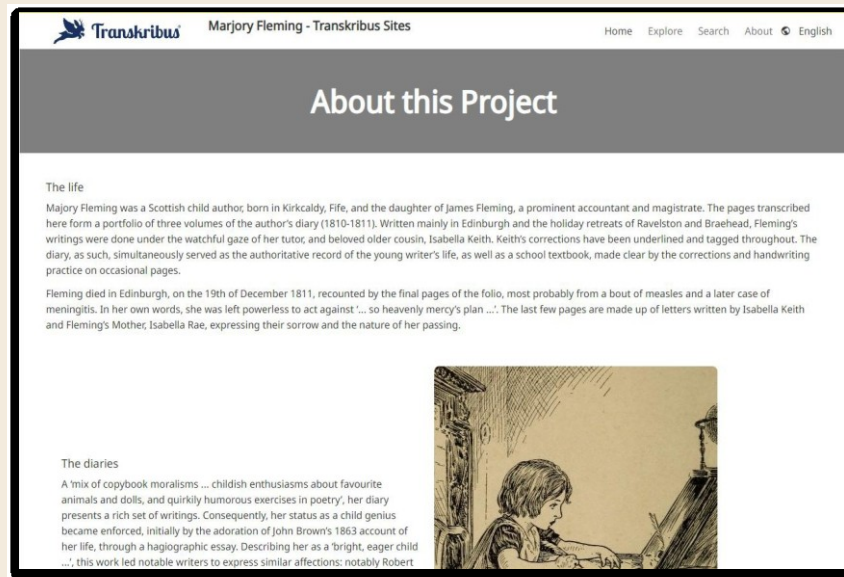


Vaša prvá stránka Transkribus

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 father James Fleming, and her education. The 'The diaries' section describes the content of her diaries, including moralisms, childish 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

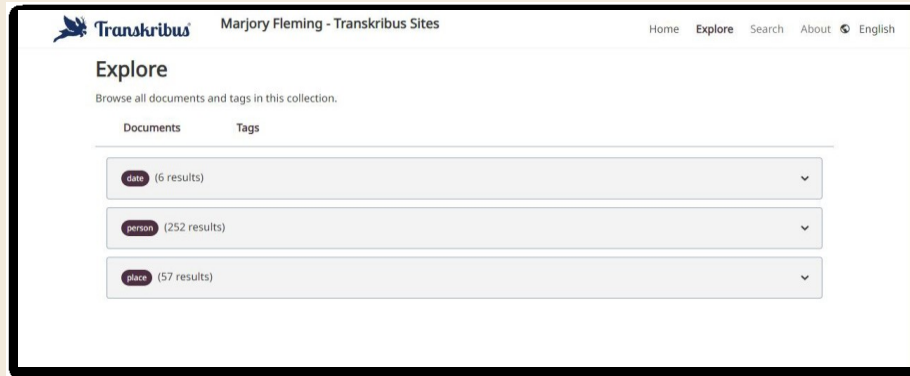


Vaša prvá stránka Transkribus

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)

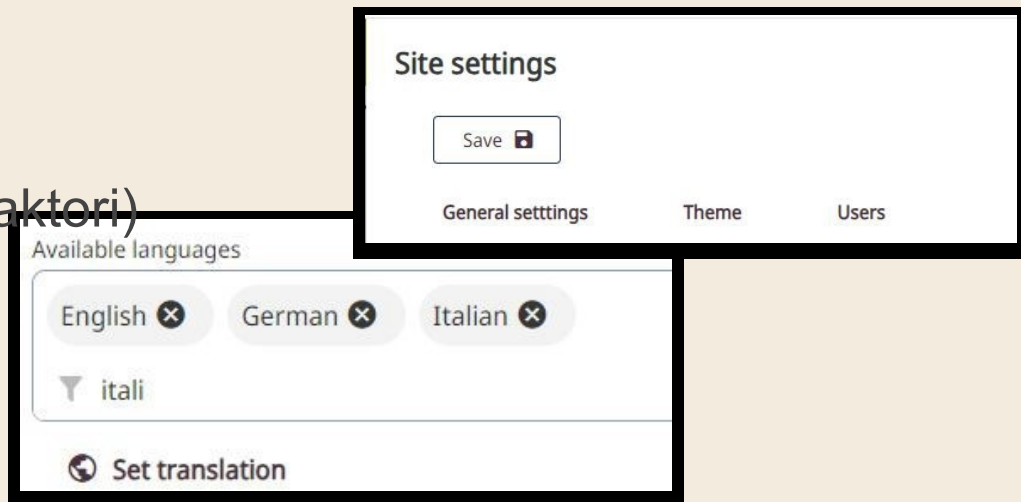


Vaša prvá stránka Transkribus

[Read&Search - Demo \(transkribus.eu\)](https://transkribus.eu)

Ďalšie nastavenia (Other settings):

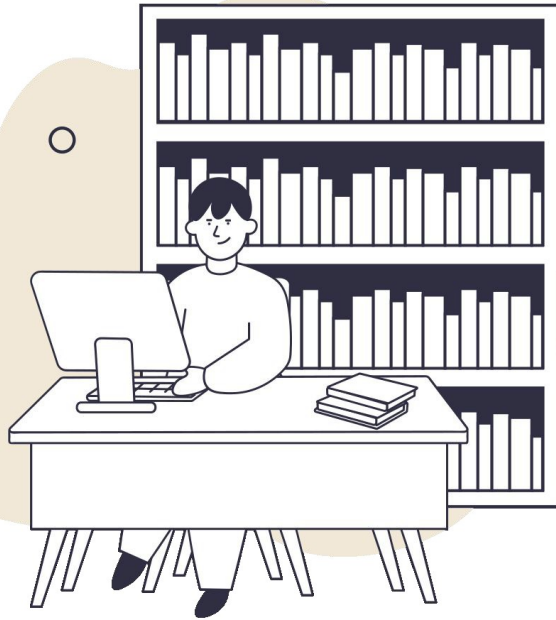
- Jazyky + možnosť úpravy prekladov
- Súkromie
- Motív (logo a farba)
- Používatelia (vlastník, redaktori)





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

