

Manuscript processing in software environment and using artificial intelligence

Les Futurs Fantastiques

2021.12.09.

Kata Ágnes SZÜCS

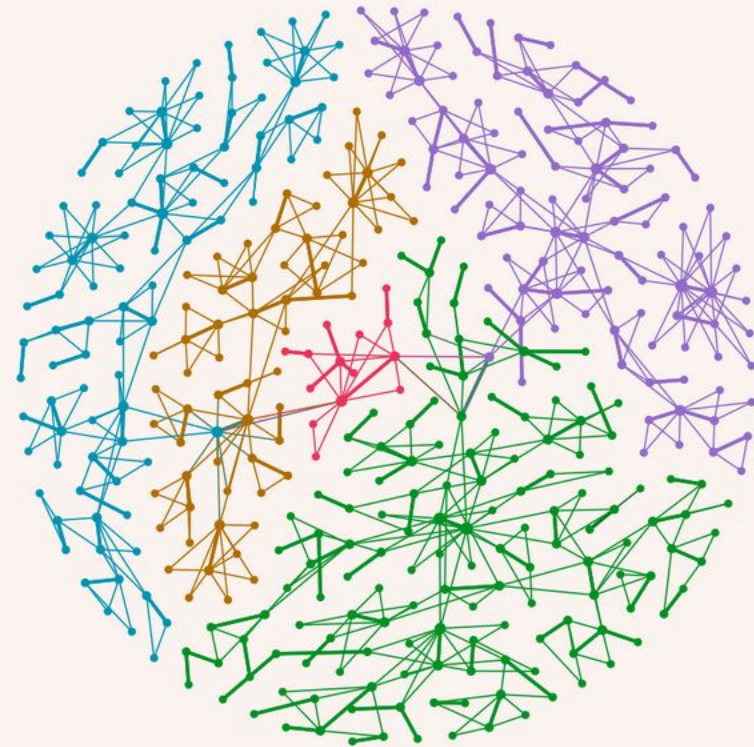
Eszter MIHÁLY – project manager

Elindultunk!

Digital Humanities Platform

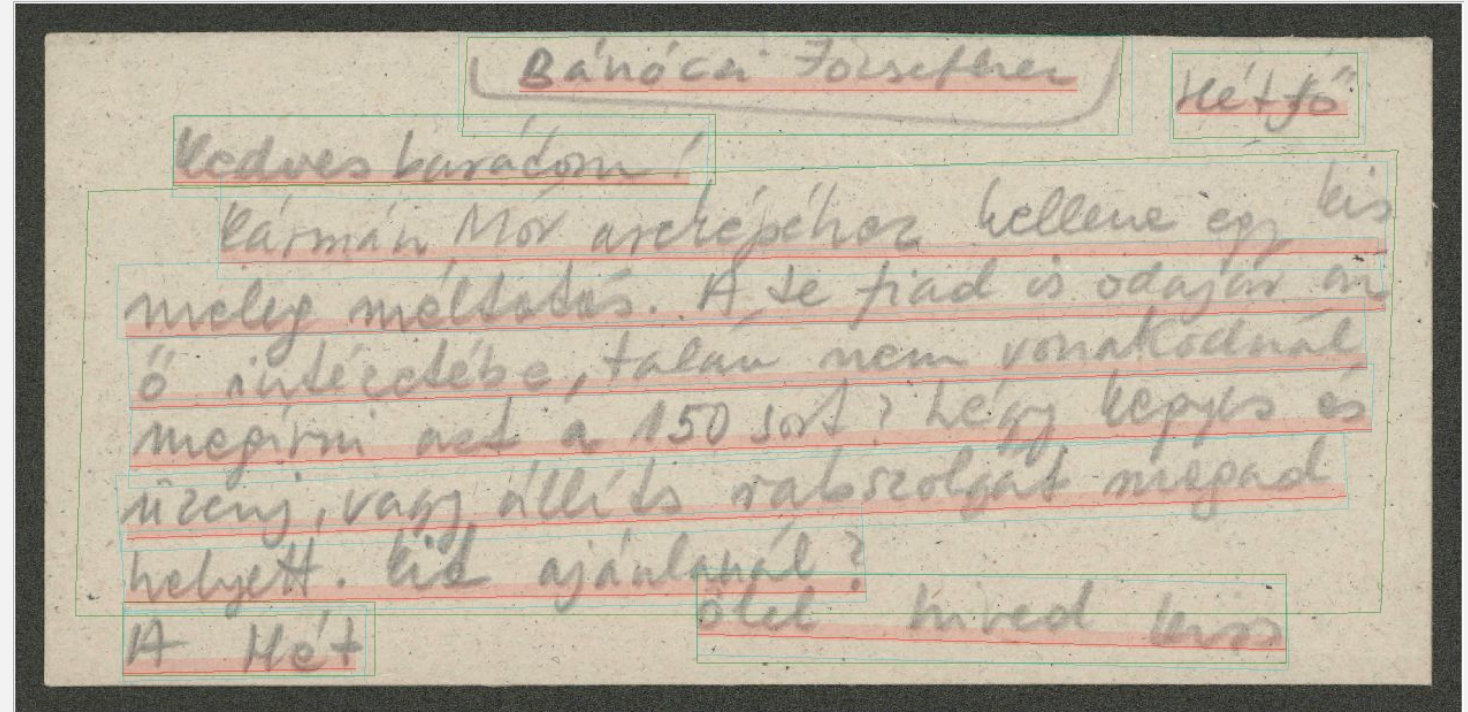
< A Digitális Bölcsészeti Platform célja digitális szövegkiadások és a szövegfeldolgozáson alapuló kreatív tartalmak közzététele.

Tudj meg többet a [dhupláról](#)!



Public collection's tasks concerning manuscripts

- describe, provide metadata, ingest
- archiving, **long-term preservation**
- make it accessible to the public



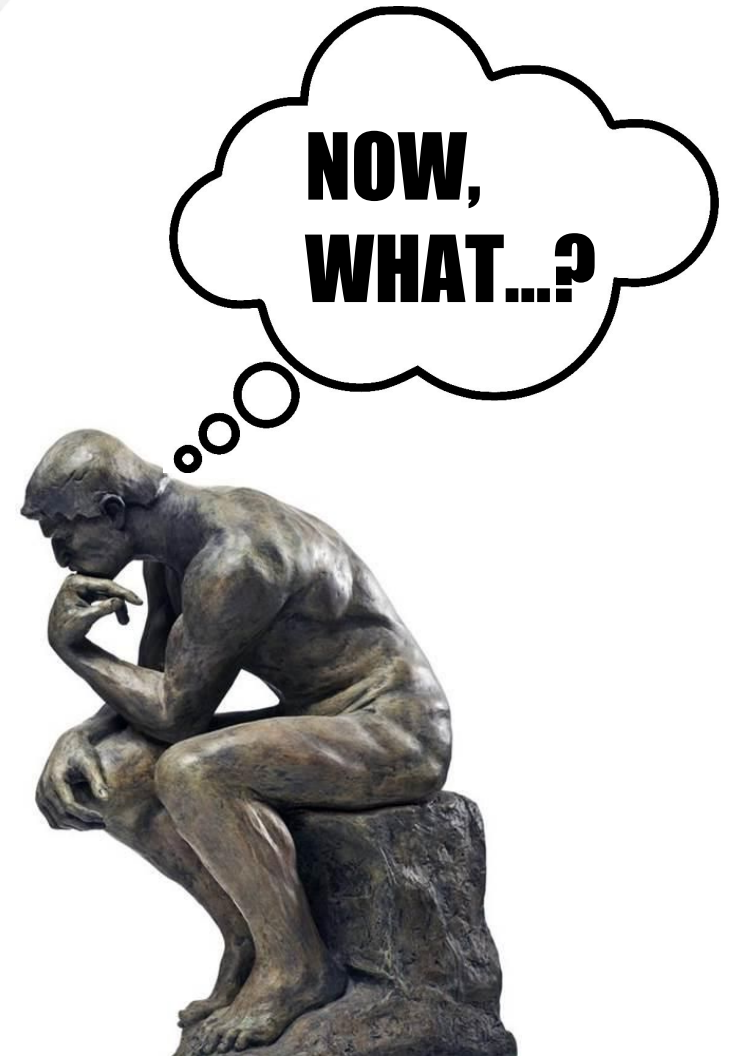
... is different in the digital age

New challenges, problems

- coordination of the digitisation
- integration of digital humanities tools
- reconciling the collection and the digital humanities approach
- providing human resources (e.g. transcription of text)

→ solutions: universal and individual

- A universal framework for public collections of manuscripts.
- Editorial environment



First steps



Common denominators

- Develop the digitizing rules
- Establishing naming conventions
- Creating records and statements of project resources
- Schedule
- Setting up a **content management environment**
- Selection of tools
- Assigning roles
- Workflow planning



In Progress

TR
L
BL
W
...
H
V
L
A
...
H
V
L
A
...

A HÉT
VIII., NÉPSZÍNHÁZ-UTCA 22.
TELEFON 61-38.

Budapest, 1909. sept. 28.

Kedves fiam Mór⁵icz Zsigmond!
Küldök egy csipetnyi novellát. A
Sallaiármúhoz felemeltem 5 forint. Ha
e kitére kifizérte⁶ volna, rögtön kiadnám
Ha meglátogatna, nagyon jó lenne!

1-1 A. Hét
2-1 VIII., Nép⁵színház-utca. 22.
3-1 Telefon. 61,38.
4-1 Budapest, 1909. sept. 28
5-1 Kedves fiam Mór⁵icz Zsigmond!
6-1 Küldök egy csipetnyi novellát. A

— +1 B I X² X² U S ...

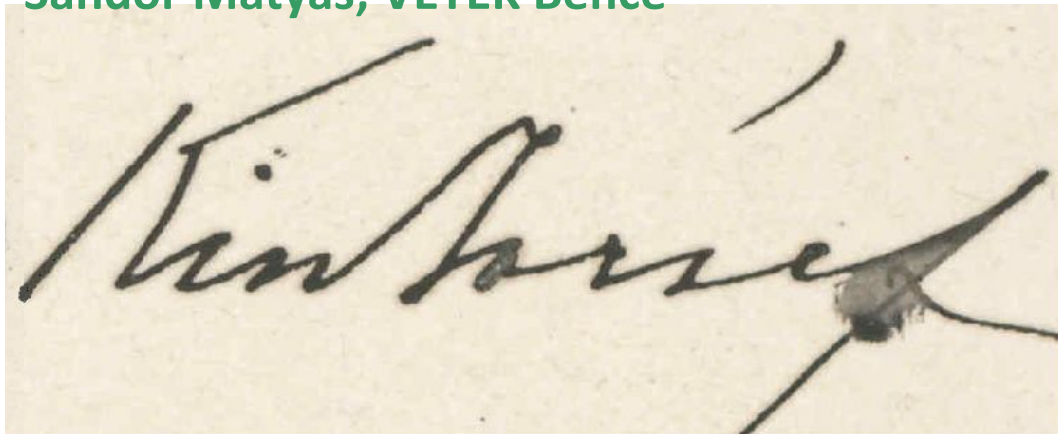
dtb DIGITÁLIS
BOLCSÉSZETI
KÖZPONT

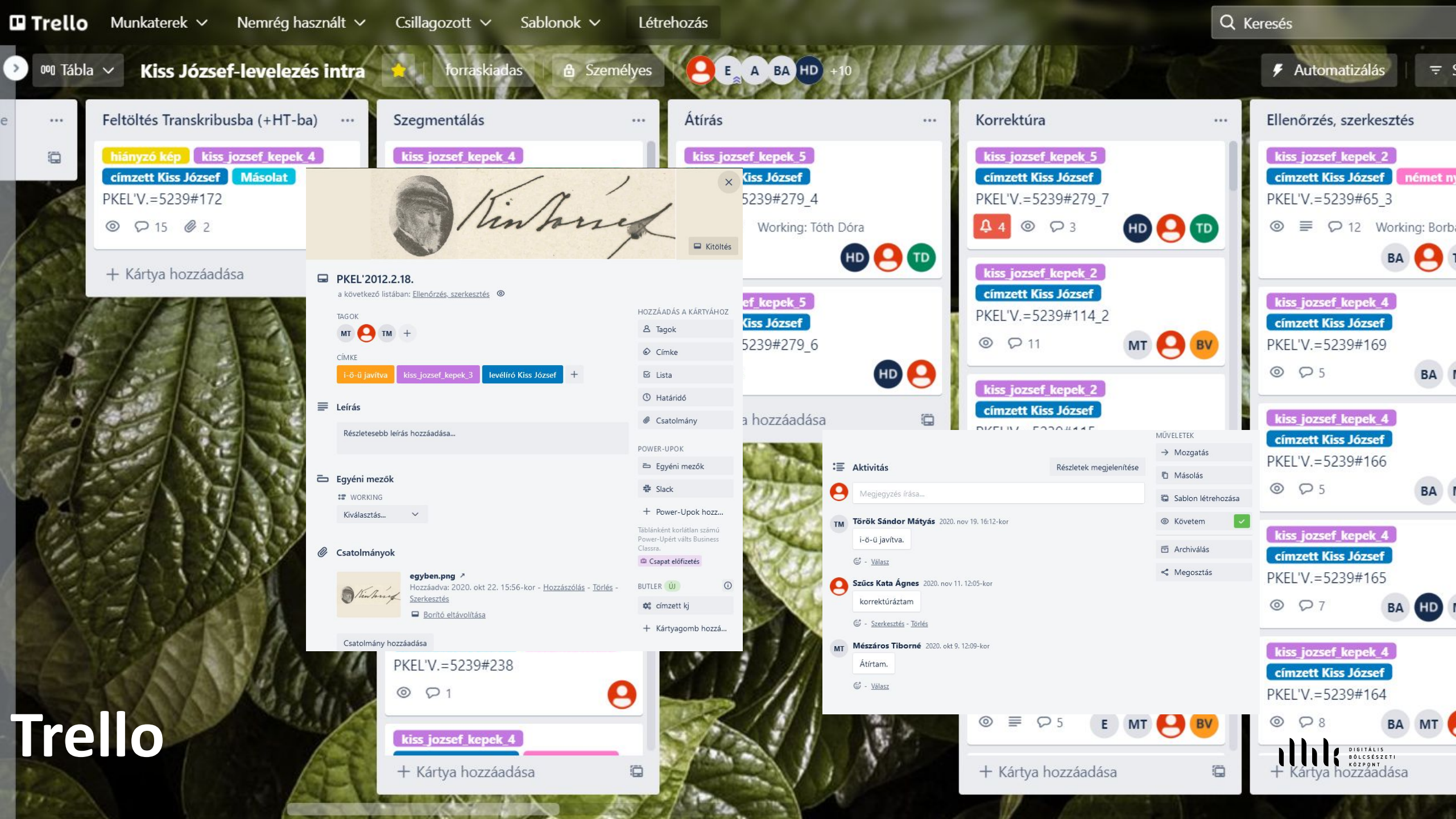
József Kiss' correspondence

Pilot project

MIHÁLY Eszter – project manager

BORBÁS Andrea; HORVÁTH Dániel; MÉSZÁROS Tiborné; SZŰCS Kata Ágnes; TÓTH Dóra; TÖRÖK Sándor Mátyás; VÉTEK Bence





Trello

Stats

The scale:

- 261 correspondent József Kiss
- 1422 addressee József Kiss
- ~1700 (1683) letters in total

State of coverage (approx.)

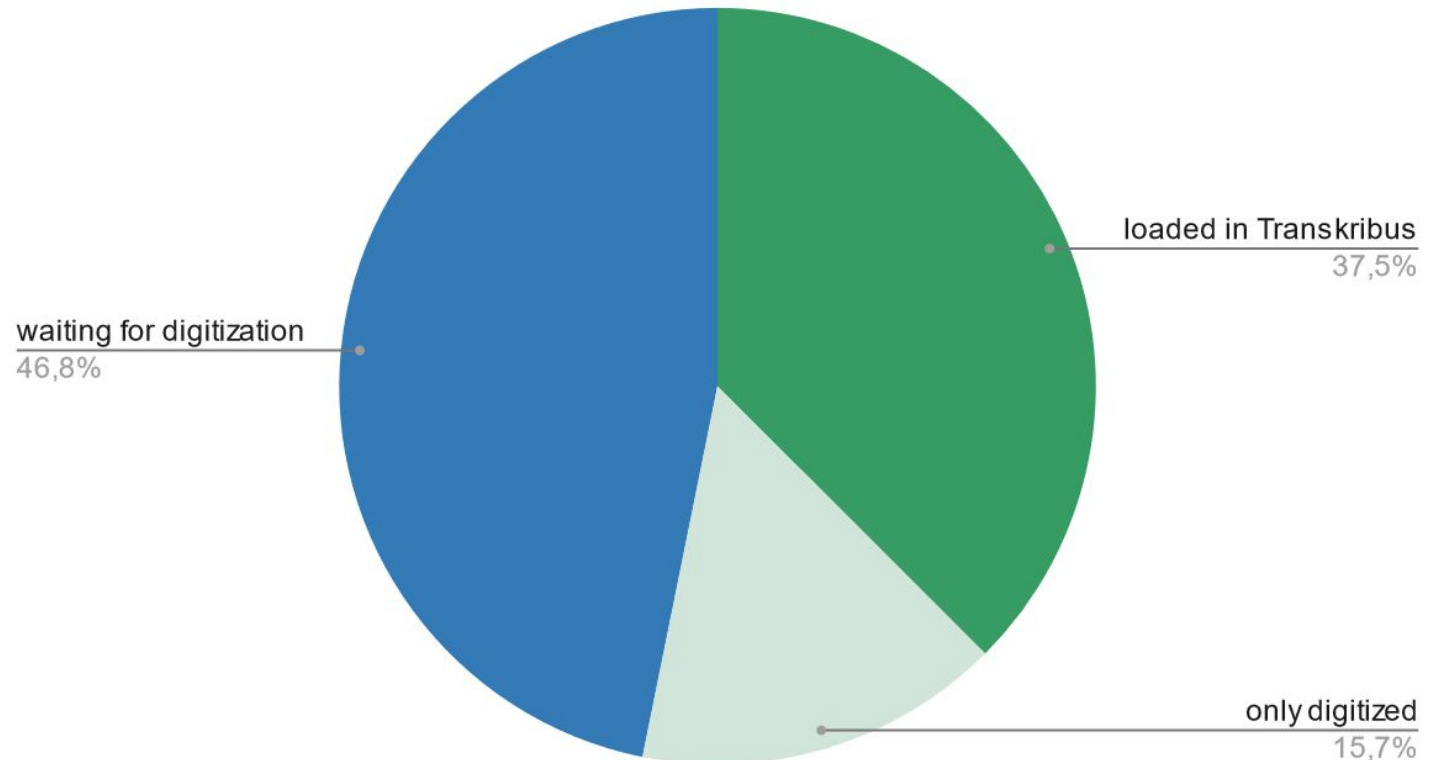
- 40% loaded in Transkribus
- 30% only digitized
- 30% waiting for digitization

HTR model

- first the handwriting of József Kiss
- currently experimenting with miscellaneous handwritings from the correspondents

The correspondence of József Kiss

The level of processing



Level of publication

1.0 OPAC

Double-layered pdf version with the transcribed text

- Full-text search in the catalogue
- Filter options



PETŐFI IRODALMI MÚZEUM

Gyűjtemények Névtér Díjak Adattárak Böngészés

Műtárgytípus

Levél (113)

Alkotó/Közreműködő

Kiss József (51)

Ady Endre (1)

Apponyi Albert (1)

Anyag

papír (4)

Keletkezés/Megjelenés
helye

Budapest (39)

H. n. (26)

Pest (8)

Levél címzettje

Kiss József (62)

Fincicky Mihály (8)

Ismeretlen (4)

Ábrányi Kornél
1907. dec. 16Ady Endre (1877-1919)
[1907. nov. 22.]Ágai Adolf (1836-1916)
É.n. szept. 14Apponyi Albert (1846-1933)
É.n. febr. 7Aradi Magántisztviselők Egyesülete
1912. febr. 10Arányi Zsigmond (1873)
1909. dec. 20?Baksay István (1820-1910)
1892. júl. 7Bálint Rezső (1885-1945)
1918. jan. 6Balogh József (1962)
1913. nov. 30Filter
options



PETŐFI LITERARY MUSEUM

[Back to results](#)

< 4/6 >

[Collections](#) [Namespace](#) [Awards](#) [Data store](#) [Browse](#)

HU EN

Labeled

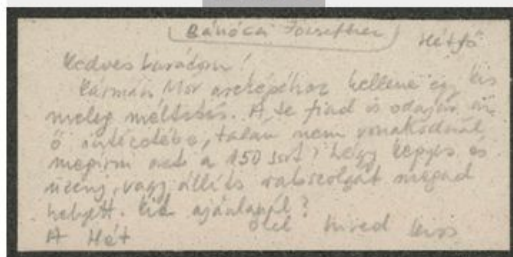
MARCXML

LIDO

<https://resolver.pim.hu/bib/PIM1225902>

Download

Megnyitás



Letöltés

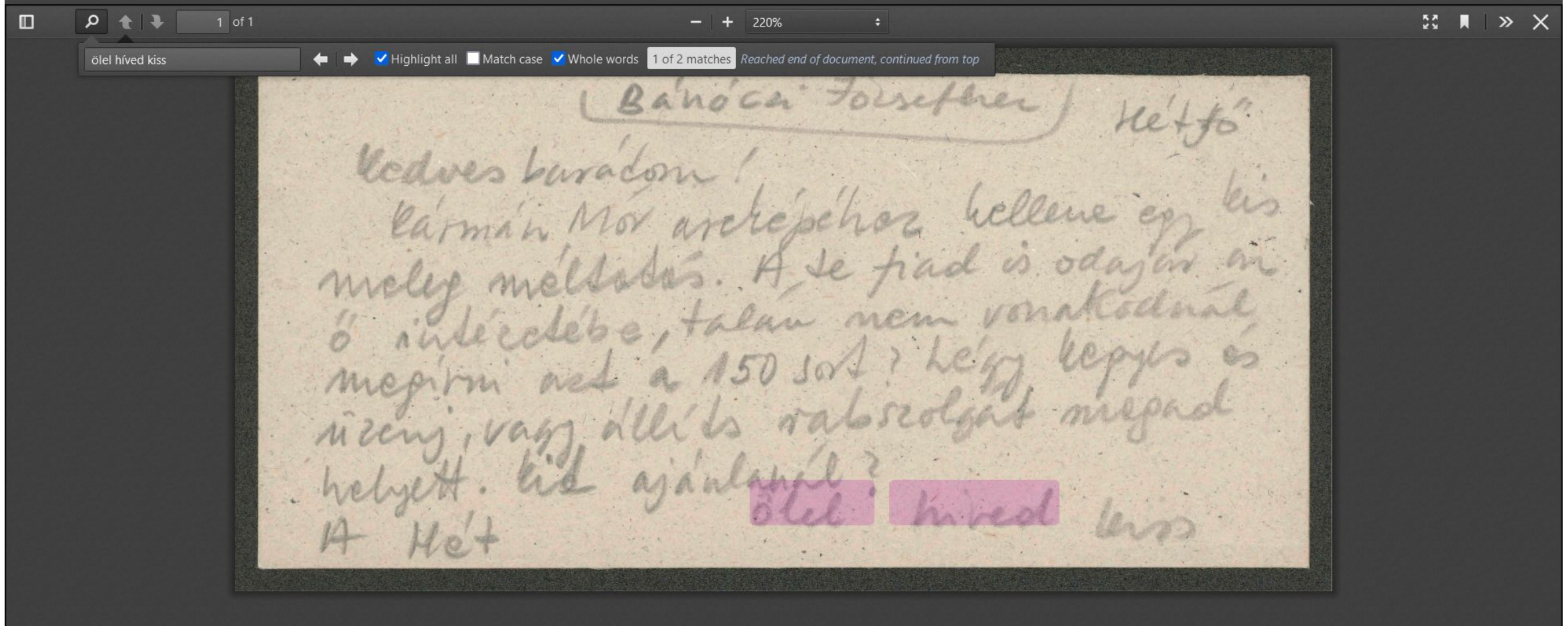
Writer: Kiss József (1843-1921)**Addressee:** Bányóczy József (1849-1926)**Description:** 1f**Annotation / Notes:** Az eredeti levél ceruzairásos másolata?**Language:** Hungarian**Document type:** letter

<< < 1 > >>

Museum ▾	Collection ▾	Legacy ▾	Call number ▾	Inventory number ▾
PIM	PIM - Kéziratanyag	Kiss József-hagyaték	V. 5239/1	V. 5239/1

<< < 1 > >>

Hits in whole text (3 hits) :1. PKEL'V=5239#1.pdf / 1. page nr : ... állíts rabszolgát magad helyett. Kit ajánlanál? **Ölel híved Kiss** A Hét ...



Double-layered PDF

Level of publication

2.0
dHUpIa

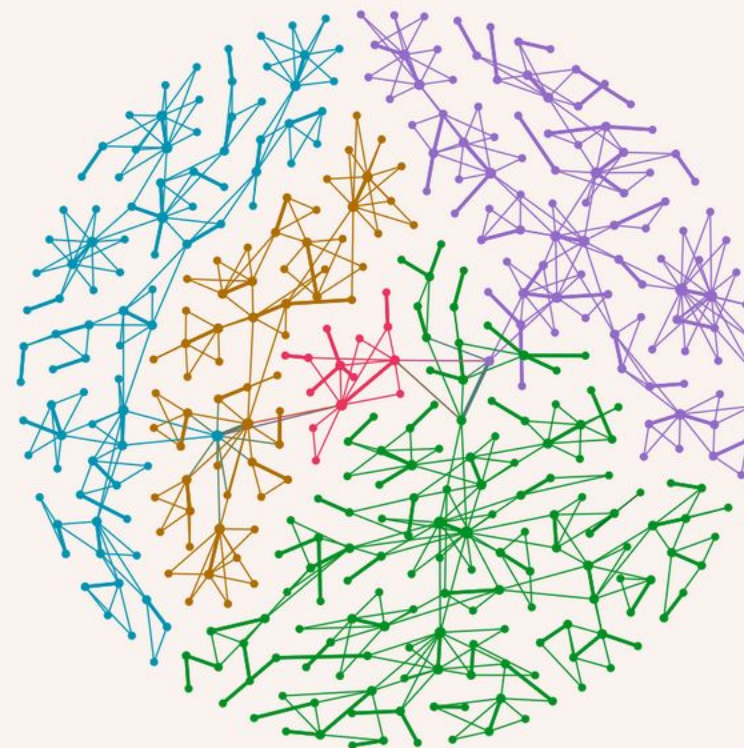
TEI-XML publication: text-image linking. Multiple view options.

Elindultunk!

Digital Humanities Platform

< A Digitális Bölcsészeti Platform célja digitális szövegkiadások és a szövegfeldolgozáson alapuló kreatív tartalmak közzététele.

Tudj meg többet a [dhupláról](#)!



Küldjön egy csipetnyi novellát, a
salláriumbot felemelem 5 frttal. Ha
e hétre liferálhatna, rögvest kiadnám.

A Hét

VIII., Népszínház-utca 22.

Telefon 61-38.

Budapest, 1909 sept 28

Kedves fiam Móricz Zsigmond!

Küldjön egy csipetnyi [rövidítés] A

salláriumot felemelem 5 frttal. Ha

e hétre liferálhatna, rögvest kiadnám.

Ha meglátogatna, nagyon jót tenne ve-

lem, mert még mindig beteg vagyok.

Collegialis szeretettel

híve Kiss

Text-image linking

Ngos

Moricz Zsigmond

urnak



VIII., Népszínház-utca 22.



A Hét

Kedves fiam Moricz Zsigmond!

Kedves fiam Móricz Zsigmond!

Küldjön egy csipetnyi novellát. A
salláriumot felemelem 5 frttal. Ha
e hétre liferálhatna, rögvest kiadnám.
Ha meglátogatna, nagyon jót tenne ve-
lem, mert még mindig beteg vagyok.

Collegialis szeretettel

híve Kiss

Multiple view options

VIII., Népszínház-utca 22.



A Hét

VIII., Népszínház-utca 22.

Telefon 61–38.

Budapest, 1909 sept 28

Kedves fiam Móricz Zsigmond!

Küldjön egy csipetnyi novellát. A
salláriumot felemelem 5 frttal. Ha
e hétre liferálhatna, rögvest kiadnám.
Ha meglátogatna, nagyon jól tene ve-
lem, mert még mindig beteg vagyok.

Collegialis szeretettel

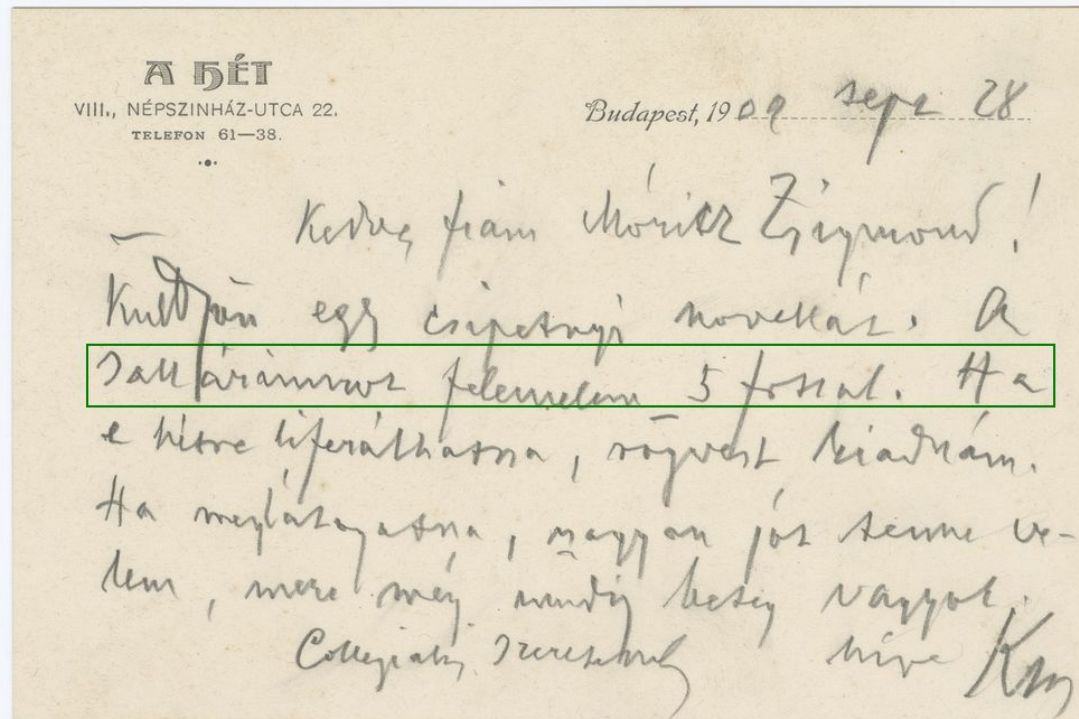
híve Kiss

🔍 Nagyít 🔍 Kicsinyít

← Előző → Következő

◀ Vissza

🖨 Teljes képe



Traditional representation

cím	típus	keletkezés
<u>Jékey Aladár – Kiss József (1890. márc. 11.)</u>	levél	1890-03-11
<u>Kiss József – Endrei Zalán (1902-04-29)</u>	levél	1902-04-29
<u>Balla Ignác – Kiss József (1902-07-28)</u>	levél	1902-07-28
<u>Kiss József – Bácskai Hírlap (1904-12-20)</u>	levél	1904-12-20
<u>Ady Endre – Kiss József (k.n.)</u>	levél	[1907-11-22]
<u>Ábrányi Kornél – Kiss József (1907-12-16.)</u>	levél	1907-12-16
<u>Kiss József – Ismeretlen (1908-02-02)</u>	levél	1908-02-02
<u>Kiss József – Tömörkény István (1908-02-05)</u>	levél	1908-02-05
<u>Kazay László – Kiss József (1908-09-25)</u>	levél	1908-09-25
<u>Kiss József – Móricz Zsigmond (1909-09-28)</u>	levél	1909-09-28
<u>Kiss József – Magyar Hitelbank Igazgatósága (1909-12-13)</u>	levél	1909-12-13
<u>Arányi Zsigmond – Kiss József (1909-12-30)</u>	levél	1909-12-30
<u>Antal Sándor – Kiss József (1908-12-29)</u>	levél	1911-12-11
<u>Kiss József – Vargha Gyula (1914-01-17)</u>	levél	1914-01-14

keresés a gyűjteményben

[szűrési feltételek törlése](#)

TÍPUS

levél (42)

SZERZŐ / LEVÉLÍRÓ

Kiss József (29)

n.a. (1)

Ady Endre (1)

Angyalffy Erzsébet (1)

Antal Sándor (1)

Apponyi Albert (1)

FELADÁS HELYE

n.a. (39)

Budapest (2)

[Budapest] (1)

CÍMZETT

Kiss József (13)

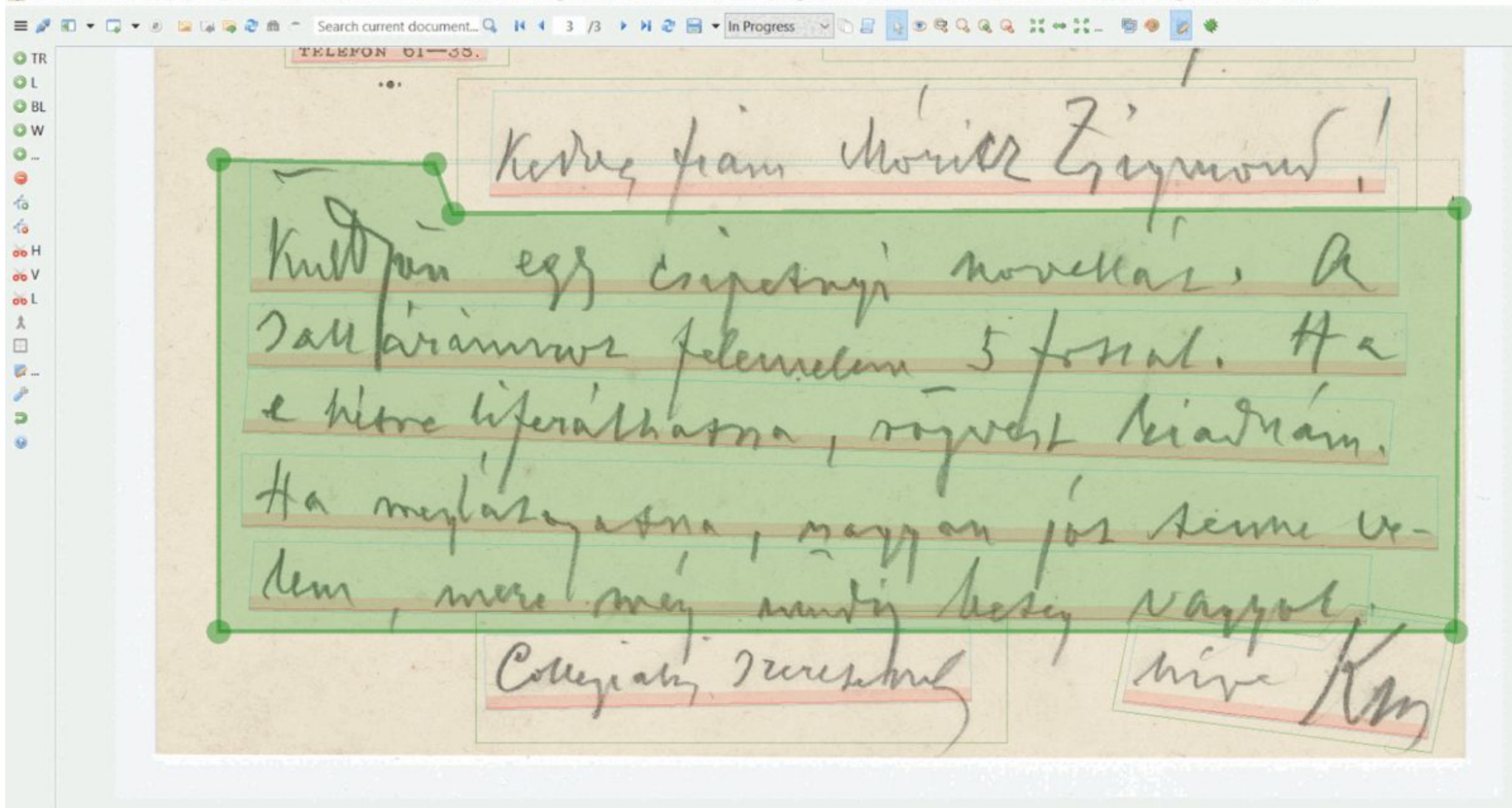
n.a. (10)

Filtering based on the metadata and annotation

The role of the HTR

- First Hungarian public handwriting recognition model
- Building increasingly general models
- Creation of dictionaries

house → "house"



Segmentation

Search current document... 3 / 3 In Progress

TR
L
BL
W
...
H
V
L
A
...
...

A HÉT
VIII., NÉPSZÍNHÁZ-UTCA 22.
TELEFON 61—38.

Budapest, 1909 sept 28

Kedves fiam Mórícz Zsigmond!

Küldjön egy csipetnyi novellát, a
salláriumot felemelem 5 frttal. Ha
e hétre liferálhatna, rögvst kiadnám.

1-1 A Hét
2-1 VIII., Népszínház-utca 22.
3-1 Telefon 61—38.
4-1 Budapest, 1909 sept 28
5-1 Kedves fiam Mórícz Zsigmond!
6-1 Küldjön egy csipetnyi novellát. A
6-2 salláriumot felemelem 5 frttal. Ha
6-3 e hétre liferálhatna, rögvst kiadnám.

Connected lines

Handwritten Text Recognition



Official How to's and videos are available on the Transkribus webpage

https://transkribus.eu/wiki/images/3/34/HowToTranscribe_Train_A_Model.pdf

What do you need?

- HTR model for a given language OR
- Transcribed manuscript document containing at least 5000-15000 words

What makes a model good?

If the corpus consists of texts

- written by one hand
- written roughly in the same period
- one type of medium or genre (e.g. diary, correspondence, account book, etc.)



Creating HTR model

Transkribus v1.13.1 (17_12_2020_14:02)

Server Overview Layout Metadata **Tools**

▼ Layout Analysis

Method: CITIab Advanced

● Current page

☒ Find Text Regions

☒ Find Lines in Text Regions

▼ Text Recognition

Method: HTR (CITIab HTR+ & PyLaTeX) Models...

▼ Compute Accuracy...

Reference: (Correct Text) Choose... Use current

Hypothesis: (HTR Text) Choose... Use current

▼ Other Tools

● Current page

Add Polygons to Baselines

HTR Training

Model Name: Language: CITIab HTR+ Nr. of Epochs: 50

Description: Base Model: Choose... Omit lines by tag: ☐ gap ☐ unclear Reset to defaults

Documents HTR Model Data

- > 569257 - PKEL'V.=5239#148 (4 pages)
- > 569254 - PKEL'V.=5239#145 (4 pages)
- > 569252 - PKEL'V.=5239#144_4 (2 pages)
- > 569070 - PKEL'V.=5239#144_3 (3 pages)
- > 569069 - PKEL'V.=5239#144_2 (1 pages)
- > 569065 - PKEL'V.=5239#144_1 (1 pages)
- > 557610 - PKEL'V.=5239#143 (3 pages)
- > 557278 - PKEL'V.=5239#142 (2 pages)
- > 557273 - PKEL'V.=5239#141 (1 pages)
- > 557263 - PKEL'V.=5239#139_2 (2 pages)
- > 555480 - PKEL'V.=5239#140 (3 pages)
- > 555472 - PKEL'V.=5239#139_1 (2 pages)
- > 555459 - PKEL'V.=5239#138 (2 pages)
- > 555458 - PKEL'V.=5239#137 (4 pages)

Filter

Overview

Transcript version: Latest transcript

Training Set

ID	Title	Pages
----	-------	-------

Remove selected entries from training set

Validation Set

ID	Title	Pages
----	-------	-------

Remove selected entries from validation set

Training Validation automatic selection of validation set ☐ 2% from train ☐ 5% from train ☐ 10% from train

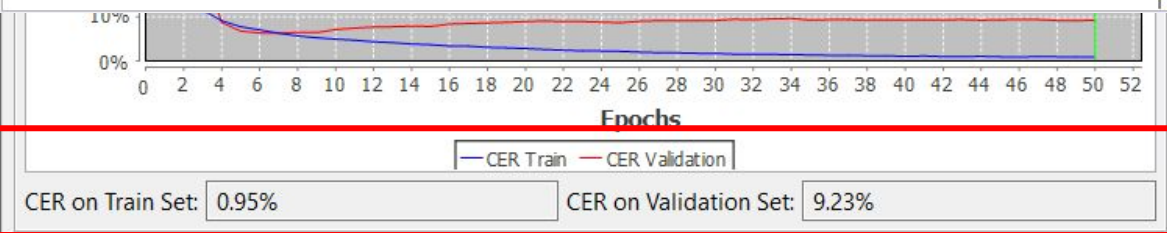
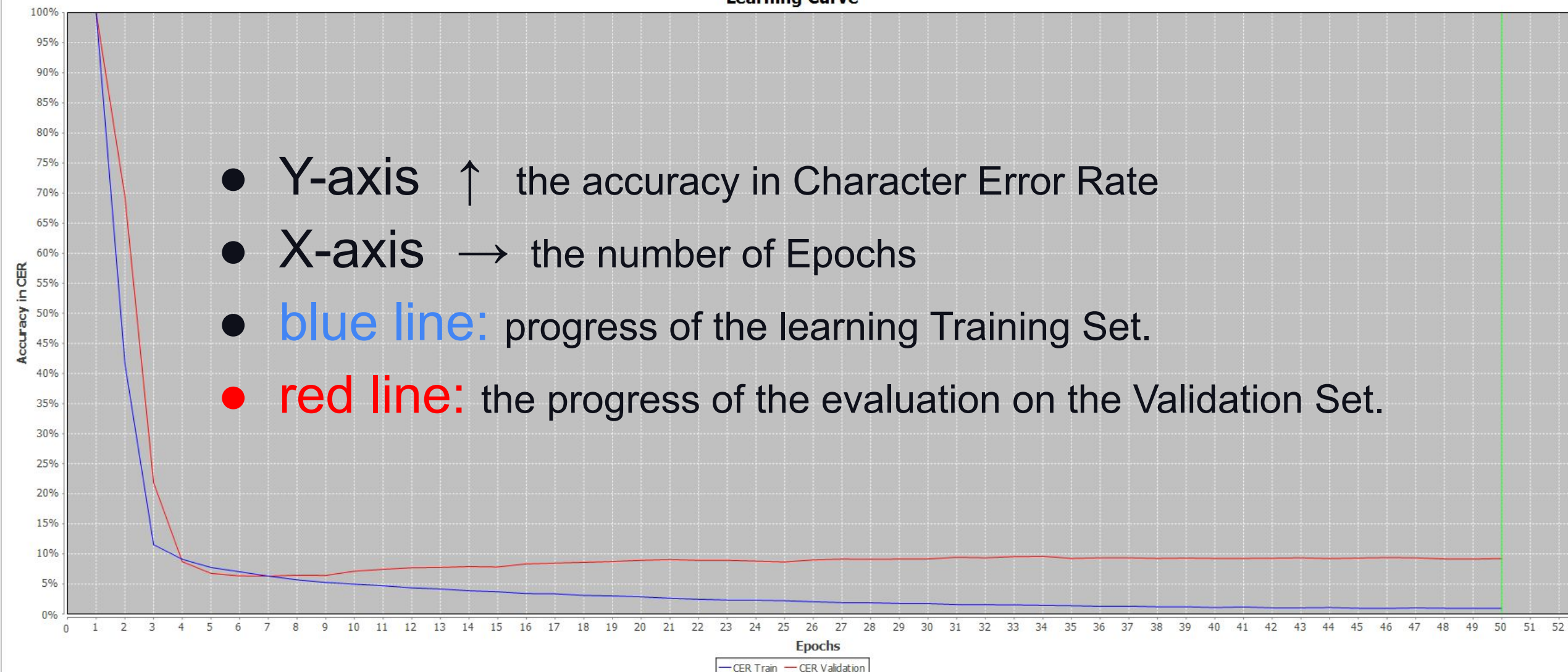
OK Cancel

Base Model

Trainig set (90%)

Validation set (10%)

Learning Curve

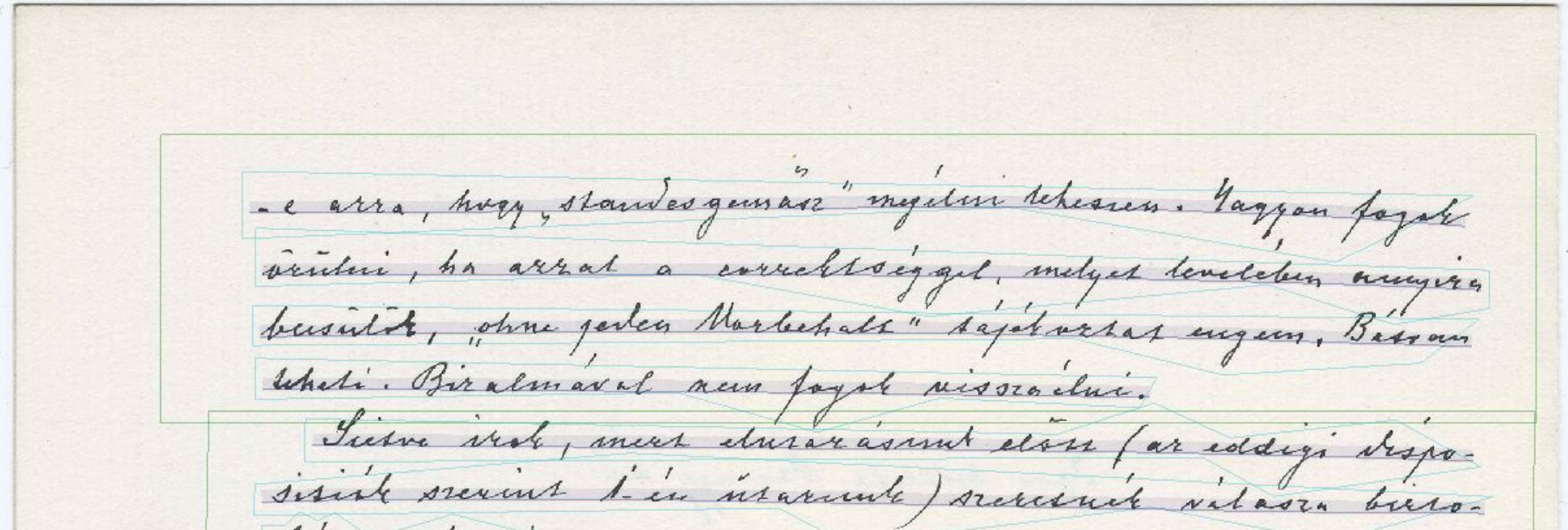


CER = Character Error Rate

Training models – the results on one handwriting

Model Name	Train Set	Validation set	Epoch	CER on Validation Set	CER on Train Set	Training Time
KEZ17_Kiss József kézírása_1	455	42	75	7,39%	4,79%	2h 47m
KEZ17_Kiss József kézírása_2	455	42	100	7,15%	3,81%	3h 34m
KEZ17_Kiss József kézírása_3	455	42	125	7,11%	2,99%	4h 17m
KEZ17_Kiss József kézírása_4	455	42	150	7,01%	2,58%	5h 14m
KEZ17_Kiss József kézírása_5	455	42	200	6,94%	2,13%	6h 53m

6,94%



Version Comparator

☒ Show line numbers

- 1-1 # ~~e~~-e arra, hogy ~~standesgemasz~~ ~~megelni~~ „standesgemász” megélni lehessen. Nagyon fogok
1-2 # örülni, ha azzal a correktséggel, melyet levelében annyira
1-3 # becsülök, ~~öline~~ „ohne jeden Vorbehalt” tájékoztat Vorbehalt” tájékoztat engem. ~~Bassán~~ ~~Bátran~~
1-4 # teheti. Bizalmával nem fogok ~~vissza elni~~ visszaélni.
2-1 # Sietve írok, mert elutazásunk előtt (az eddigi dispo-
2-2 # ~~sitiók~~ ~~sitiók~~ szerint 1-én utazunk) szeretnék válasza ~~biztó~~ ~~birto~~
2-3 # kában lenni.
3-1 # Önt még egyszer teljes ~~biztmunk~~ ~~bizalmunk~~ és rokonszenvünk-
3-2 # ~~ról~~ ~~biztósítva~~ ~~ról~~ ~~biztosítva~~ maradtam baráti ~~jóindulat~~ ~~jóindulattal~~
4-1 # ~~kész~~ ~~Kész~~ híve
5-1 # Kiss József

ment és rokonszenvünk-
arasi jóindulattal

Kiss hve

Improvement options

- increase the amount of data
- Use Base Model (for larger corpus)
- incorporation of dictionaries
- manually modifying the shape of polygons / preserving their original shape
- avoid over-learning and bias (do not mix the data from the training set and the validation set for different models, as this can lead to false-positive results in the percentage of the character error rate value.
- selection of more representative training and validation sets
- increase the number of epochs (may be time-consuming)
- use language model (automatically built from training data/custom dictionary)

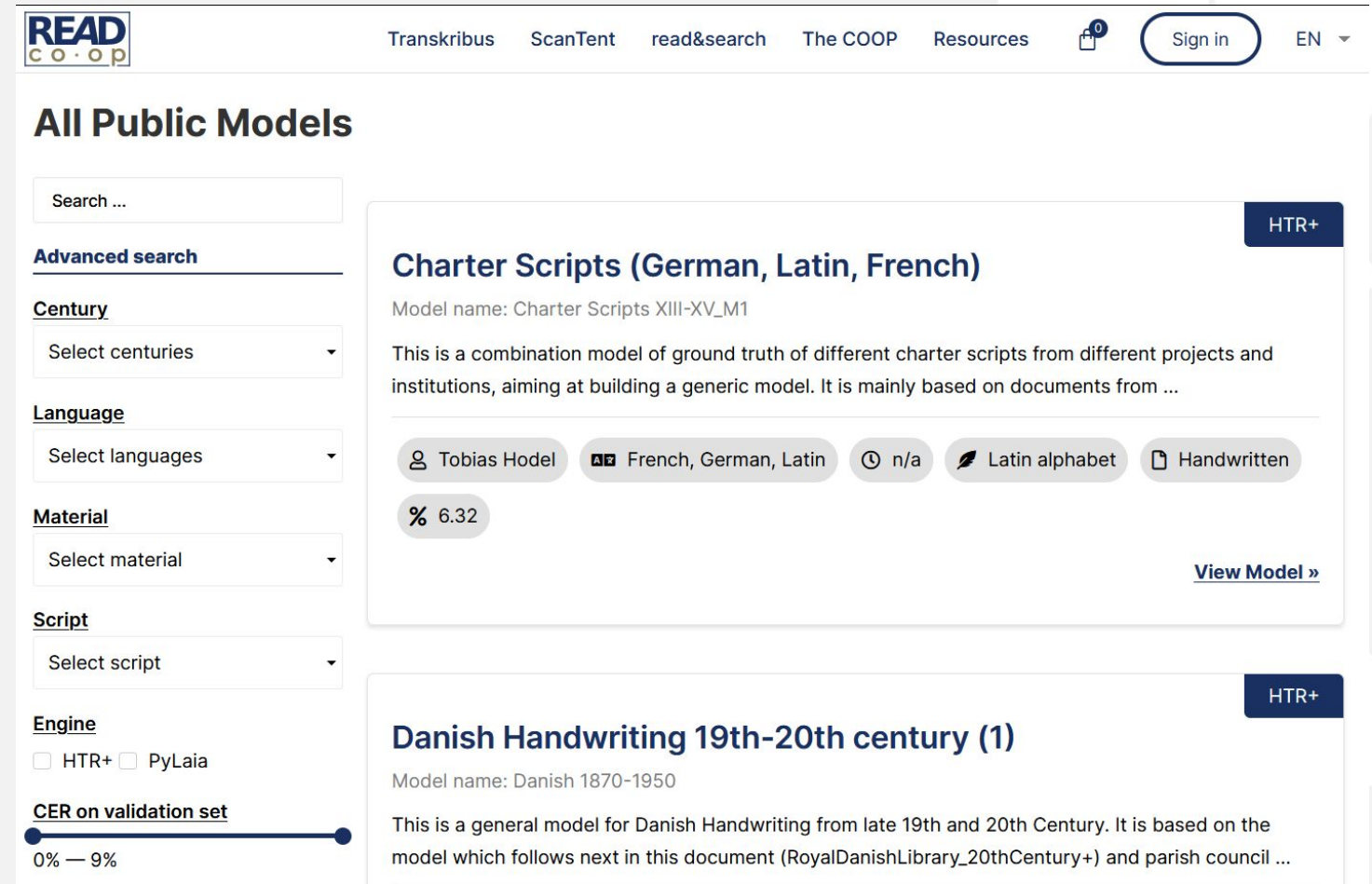


Training models – the results on mixed handwriting

Model name	Selection	Technology	Base model	Training Set (pages)	Validation set (pages)	Epoch	CER on Train Set	CER on Validation Set
Vegyes kézírás_9	aut 10% (2)	HTR+	NO	481	53	150	4.69%	12.38%
Vegyes kézírás_18	aut 10% (2)	HTR+	YES (6.94%)	481	53	150	8.28%	12.34%
Vegyes kézírás_13	manu (1)	HTR+	YES (6.94%)	484	54	150	4.06 %	10.11 %
Vegyes kézírás_15	manu (1)	HTR+	NO	484	54	150	5.1 %	10.92 %

Next steps

1. HTR-integration into workflow
2. Creating new models (e.g. Zsigmond Móricz correspondence)
3. Building a better and better general Hungarian model
4. Publishing models (in progress)



READ
COOP

Transkribus ScanTent read&search The COOP Resources Sign in EN

All Public Models

Search ...

Advanced search

Century
Select centuries

Language
Select languages

Material
Select material

Script
Select script

Engine
☐ HTR+ ☐ PyLaia

CER on validation set
0% — 9%

Charter Scripts (German, Latin, French)

Model name: Charter Scripts XIII-XV_M1

This is a combination model of ground truth of different charter scripts from different projects and institutions, aiming at building a generic model. It is mainly based on documents from ...

Tobias Hodel French, German, Latin n/a Latin alphabet Handwritten

% 6.32

[View Model »](#)

Danish Handwriting 19th-20th century (1)

Model name: Danish 1870-1950

This is a general model for Danish Handwriting from late 19th and 20th Century. It is based on the model which follows next in this document (RoyalDanishLibrary_20thCentury+) and parish council ...

