

Automatikus kézírás-felismertetés

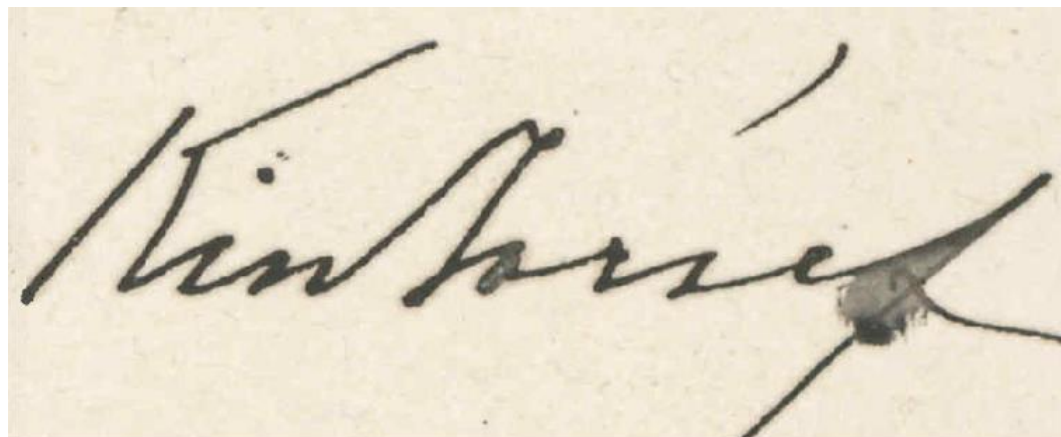
Transkribus és a HTR

Mihály Eszter – Szűcs Kata
Ágnes
2021. január 26.

Kiss József- projekt

Pilot projekt

Borbás Andrea; Horváth Dániel; Mészáros Tiborné;
Mihály Eszter; Szűcs Kata Ágnes; Tétényi Csaba; Tóth
Dóra; Török Sándor Mátyás; Vétek Bence



Trello

I. A Transkribus-ról röviden

Vázlat

- I. Mi az a Transkribus
- II. Mit tudunk az automatikus
kézírás-felismeretetésről
- III. Milyen eredményeket
értünk el eddig.

Server Overview Layout Metadata Tools

Logout szucs.kata.agnes@gmail.com

Document...

Find

Document Manager

User Manager

Versions

Jobs

Recent Documents...

User activity

Collections: kiss_jozsef (45634, Editor)

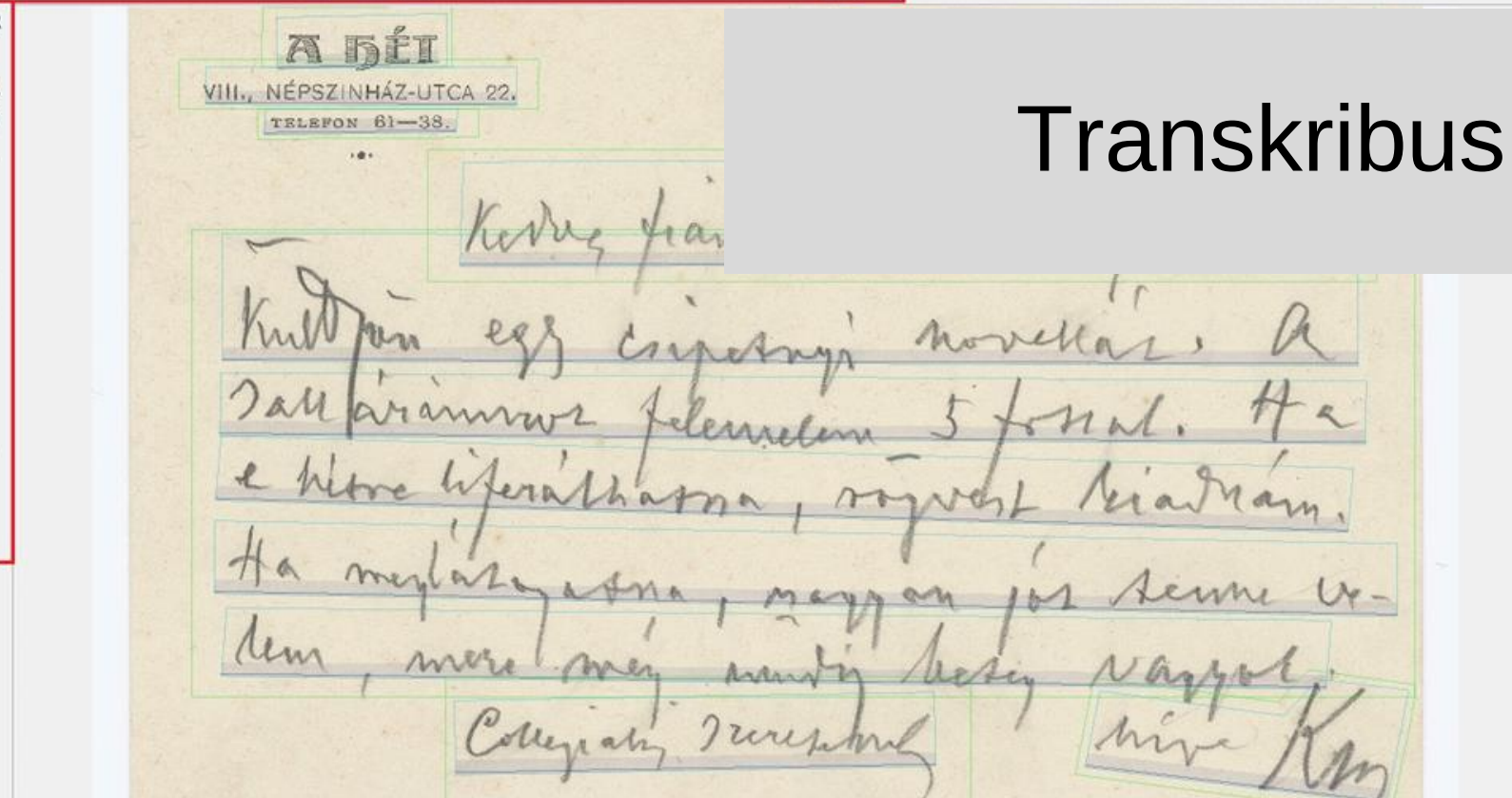
Col-ID

Documents HTR Model Data

1-100 / 215 1 3 Doc-ID

ID	Title	Pages	Uploader	Uploaded
569...	PKEL'V.=5239#148	4	szucs.kata.a...	Mon Jan 11...
569...	PKEL'V.=5239#145	4	szucs.kata.a...	Mon Jan 11...
569...	PKEL'V.=5239#144_4	2	szucs.kata.a...	Mon Jan 11...
569...	PKEL'V.=5239#144_3	3	szucs.kata.a...	Mon Jan 11...
569...	PKEL'V.=5239#144_2	1	szucs.kata.a...	Mon Jan 11...
569...	PKEL'V.=5239#144_1	1	szucs.kata.a...	Mon Jan 11...
557...	PKEL'V.=5239#143	3	esztter.miha...	Thu Dec 17...
557...	PKEL'V.=5239#142	2	esztter.miha...	Wed Dec 1...
557...	PKEL'V.=5239#141	1	esztter.miha...	Wed Dec 1...
557...	PKEL'V.=5239#139_2	2	esztter.miha...	Wed Dec 1...
555...	PKEL'V.=5239#140	3	esztter.miha...	Mon Dec 1...
555...	PKEL'V.=5239#139_1	2	esztter.miha...	Mon Dec 1...
555...	PKEL'V.=5239#138	2	esztter.miha...	Mon Dec 1...
555...	PKEL'V.=5239#137	4	esztter.miha...	Mon Dec 1...
554...	PKEL'V.=5239#136	3	esztter.miha...	Fri Dec 11 ...
554...	PKEL'V.=5239#135	1	esztter.miha...	Fri Dec 11 ...
554...	PKEL'V.=5239#134	3	esztter.miha...	Fri Dec 11 ...
554...	PKEL'V.=5239#133	3	esztter.miha...	Fri Dec 11 ...
554...	PKEL'V.=5239#131	1	esztter.miha...	Fri Dec 11 ...
554...	PKEL'V.=5239#130	4	esztter.miha...	Fri Dec 11 ...
545...	PKEL'V.=5239#117	4	esztter.miha...	Fri Nov 27 ...
543...	PKEL'V.=5239#111	36	esztter.miha...	Mon Nov 2 ...

100



Transkribus

- 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 Móríc Zsigmond!
- 6-1 Küldjön egy csipetnyi novellát. A
- 6-2 salláriumot felemelem 5 frttal. Ha
- 6-3 e hétre liferálhatna, rögvest kiadnám.



A DÉL

II., NÉPSZÍNHÁZ-UTCA 22.

TELEFON 361-38.

...

Budapest, 1909. szept. 28.

Kedves fiám Moricz Zsigmond!

Küldök egy csipetnyi novellát, a
Talláránkosz felemeltem 5 forint. Ha
e little liferálthasna, rögzvet kiadnam.

Ha meglátogatna, nagyon jó lenne ve-
lem, mert még mindig kesey vagyok.

Collegian, Zsigmond

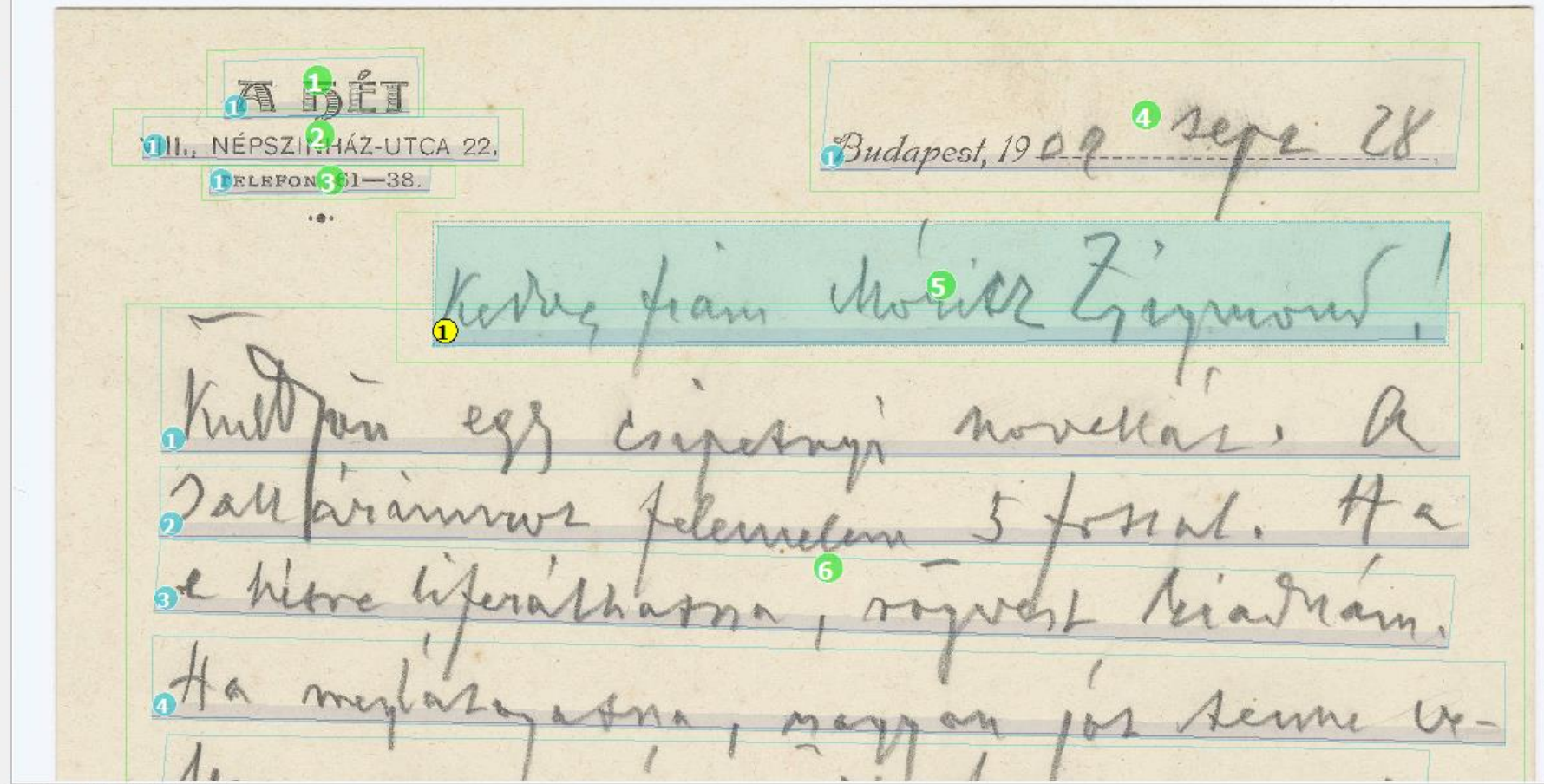
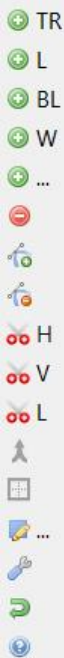
híve Km

Text
Region

Base
Line

Line

A sorok összekötötése



5-1 Telefon 01-38.

4-1 Budapest, 1909 sept 28

5-1 Kedves fiam Móricz Zsigmond!

6-1 Küldjön egy csipetnyi novellát. A

6-2 salláriumot felemelem 5 frttal. Ha

6-3 e hétre liferálhatna, rögvest kiadnám.

6-4 Ha meglátogatna, nagyon jót tenne ve-

6-5 lem, mert még mindig beteg vagyok.

7-1 Collegialis szeretettel



II. A kézírás- felismertetésről



Hivatalos How-to-k a Transkribus oldalán
<https://transkribus.eu/wiki/images/3/34/HowToTranscribeTrainAModel.pdf>

Mi kell hozzá?

- Adott nyelvhez kötődő HTR modell
- Min. 5000–15000 szót tartalmazó átírt kéziratos dokumentum

Nyelvi modell készítés

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:

Description:

CITIab HTR+

Nr. of Epochs: 50

Base Model:

Omit lines by tag: ☐ gap ☐ unclear

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
----	-------	-------

Validation Set

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

☒ Training

☒ Validation

automatic selection of validation set

☐ 2% from train

☐ 5% from train

☐ 10% from train

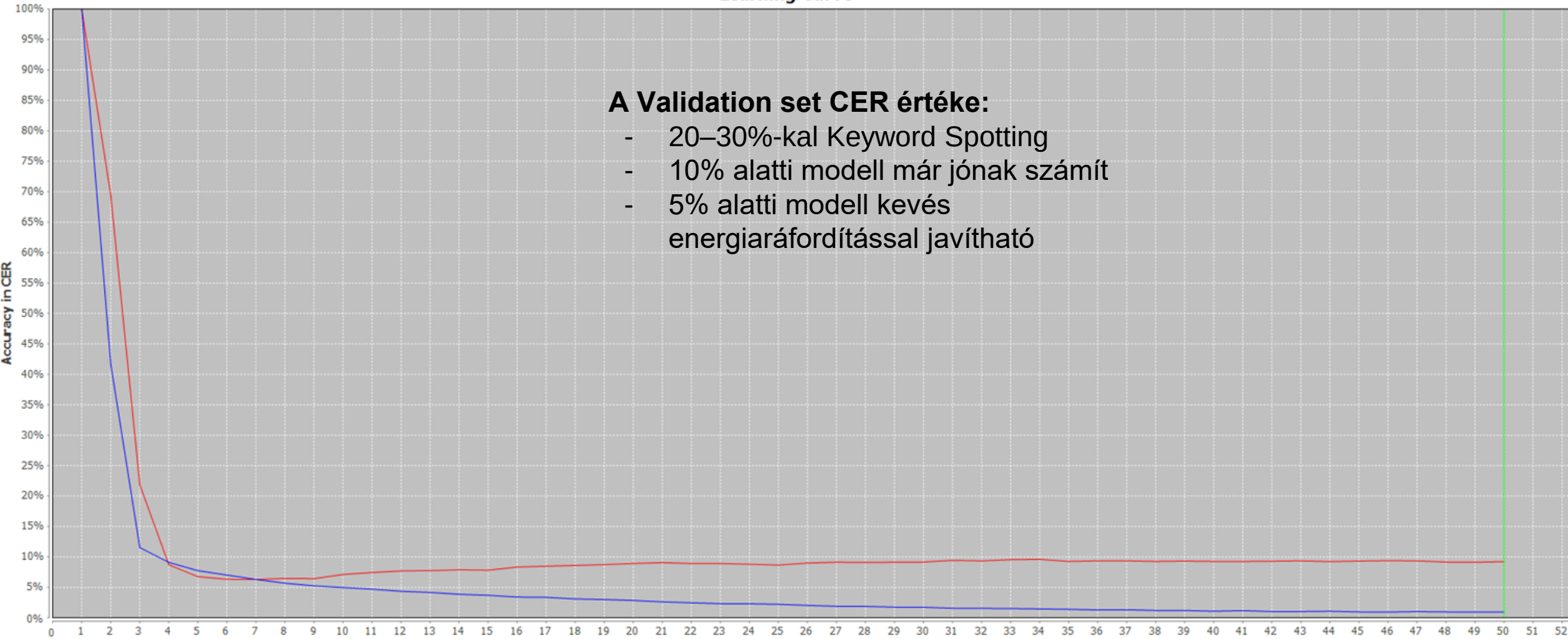
Trainig set
(90%)

Validation set
(10%)

Name:

Language:

Learning Curve



A Validation set CER értéke:

- 20–30%-kal Keyword Spotting
- 10% alatti modell már jónak számít
- 5% alatti modell kevés energiaráfordítással javítható

0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52

Epochs

CER Train

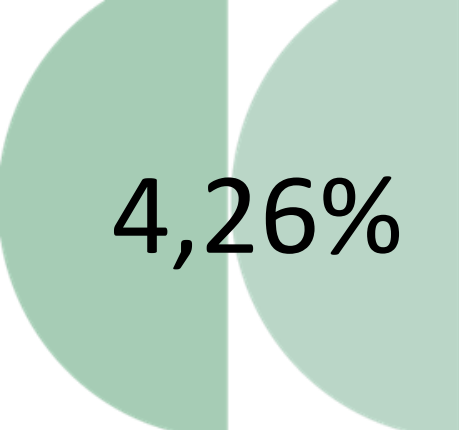
CER Validation

CER on Train Set: 0.95%

CER on Validation Set: 9.23%

Modellépítés – eredmények

	KJ_keze_osszes_1	KJ_keze_osszes_2	KJ_keze_osszes_4	KJ_keze_osszes_3	KJ_keze_osszes_5	KJ_keze_osszes_6
Eredmények						
CER on Train Set	3,05%	6,54%	2,58%	2,26%	1,77%	2,13%
CER on Validation Set	5,82%	10,70%	8,38%	4,26%	6,08%	6,59%
Nr. of Epochs	50	50	50	50	100	50
Összes fájl	366	366	369	369	369	366



4,26%



an Andra delusani 2 orang
Kolumnas actosini. Hk
in ingon naga 10 sort. 20
ryben.

Steli
hive

7-1 kérek határozom felehetet.

8-1 Öleli

9-1 hive

 Version Comparator

☒ Show line numbers

1-1 # A Hét

2-1 # VIII., Népszínház-ut Népszínház-utca 22.

3-1 # Telefon 61—38.

4-1 # Budapest, 1910 marcz ~~11~~ 11.

5-1 # Kedves barátom!-barátom.

6-1 # Már ~~legnap kerestetem~~ tegnap kerestettem Önt. Talán tudna délutáni 2 ~~óraig~~ óráig

6-2 # az Elektra muzsikájáról vagy 3/4 kolumnát alkotni. Ha

6-3 # A-a fő próban próbán volt, kormer jerezőb is Kramer Terézről is írjon vagy 10 sort. Jv Jó

6-4 # arczképünk van róla ebben a szerepben.

7-1 # ~~kérek határozom felehetet.~~ Kérek határozott feleletet.

8-1 # Öleli

9-1 # ~~hive~~-hive

10-1 # Kiss-Kiss József

III. Tapasztalatok

Mitől lesz “jó” egy modell?

- egy kéz által írt szövegek
- korban egymáshoz közel keletkezett szövegek
- egyféle forrástípus (pl. napló, levelezés, számadáskönyv, stb.)

Első fázis

homokozó, próba 2			
validation set	7	training set	61
CER	9,23%	CER	0,95%
0	átüt a tinta	2	átüt a tinta
0	boríték	1	boríték
3	ceruza	23	ceruza
0	címzés	1	címzés
0	firka	1	firka
0	gépirat	1	gépirat
0	nem KJ kézírás	0	nem KJ kézírás
0	kép. firka.	0	kép. firka.
1	német	1	német
0	nyomtatott kézírás	0	nyomtatott kézírás
3	toll	31	toll

Csak Kiss József kézírásával				
	próba 2 (homokozó)	Kiss József kézírása (kiss_jozsef_mas olat)	kj2	kj3
Train set				
pages	61	193	137	229
lines	1041	2855	2538	2946
words	5003	12971	12227	13066
Validation set				
pages	7	22	15	22
lines	132	313	223	340
words	560	1408	797	1446
Eredmények				
CER on Train Set	0.95%	1.99%	1.32%	1.89%
CER on Validation	9.23%	12.66%	17.29%	9.72%
Nr. of Epochs	50	100	100	100

kj3			
validation set	22	training set	229
CER	9,72%	CER	1,89%
0	átüt a tinta	2	átüt a tinta
3	boríték	32	boríték
4	ceruza	72	ceruza
1	címzés	15	címzés
0	firka	2	firka
0	gépirat	0	gépirat
0	nem KJ kézírás	0	nem KJ kézírás
0	kép. firka.	10	kép. firka.
0	német	1	német
0	nyomtatott kézírás	0	nyomtatott kézírás
14	toll	95	toll

Második fázis

- tollal/ceruzával/írógéppel/nyomtatottan írt levelek külön kezelésének megszüntetése

	KJ_keze_valogatas_1	KJ_keze_valogatas_2
Base Model	kj3	nincs
Train set	aut_ts_2	aut_ts_2
pages	302	302
lines	4170	4170
words	18941	18941
Validation set	aut_vs_2	aut_vs_2
pages	33	33
lines	558	558
words	2681	2681
Eredmények		
CER on Train Set	2,81%	6,11%
CER on Validation Set	6,13%	11,49%
Nr. of Epochs	50	50
Összes fájl	335	335

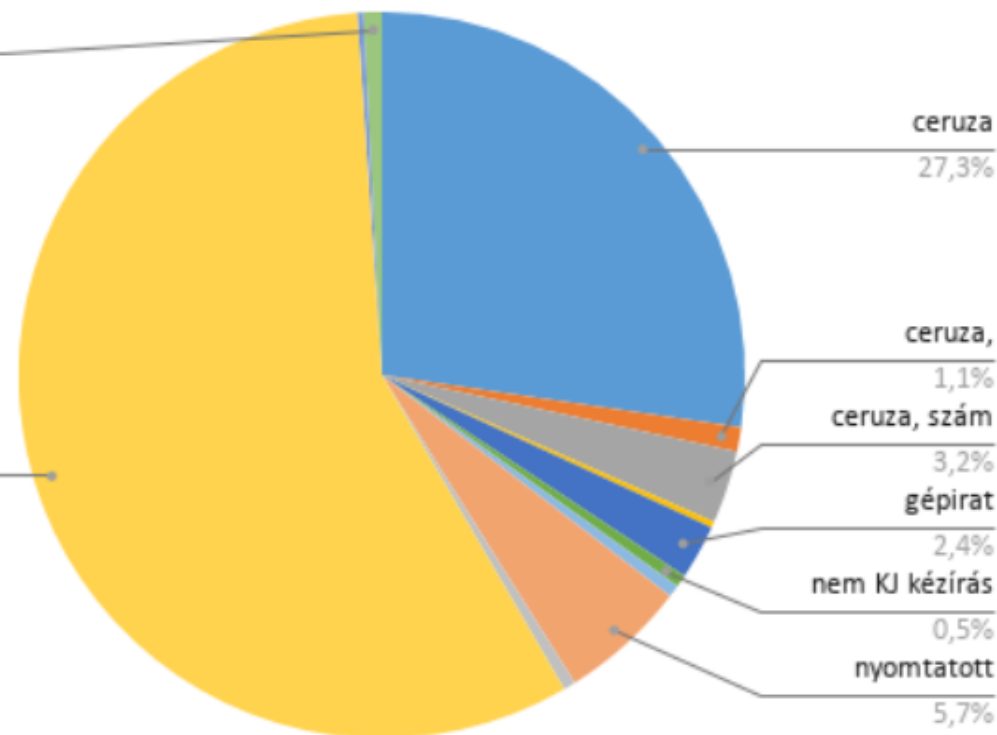
	KJ_keze_osszes_1	KJ_keze_osszes_2
Base Model	kj3	nincs
Train set	aut_ts_1	aut_ts_1
pages	330	330
lines	4399	4399
words	20039	20039
Validation set	aut_vs_1	aut_vs_1
pages	36	36
lines	459	459
words	2120	2120
Eredmények		
CER on Train Set	3,05%	6,54%
CER on Validation Set	5,82%	10,70%
Nr. of Epochs	50	50
Összes fájl	366	366

Előzetes felmérések

	összes	validation set
ceruza	101	8
ceruza, nyomtatott	4	1
ceruza, szám	12	2
francia	1	0
gépirat	9	2
nem KJ kézírás	2	1
német, toll	2	1
nyomtatott	21	4
nyomtatott kézírás	2	1
toll	212	16
toll, ceruza	1	0
toll, nyomtatott	3	1

toll, nyomtatott
0,8%

toll
57,3%



ze_osszes_6

ze_osszes_1

ut_ts_1

330

4399

20039

ut_vs_1

36

459

2120

2,13%

6,59%

50

366

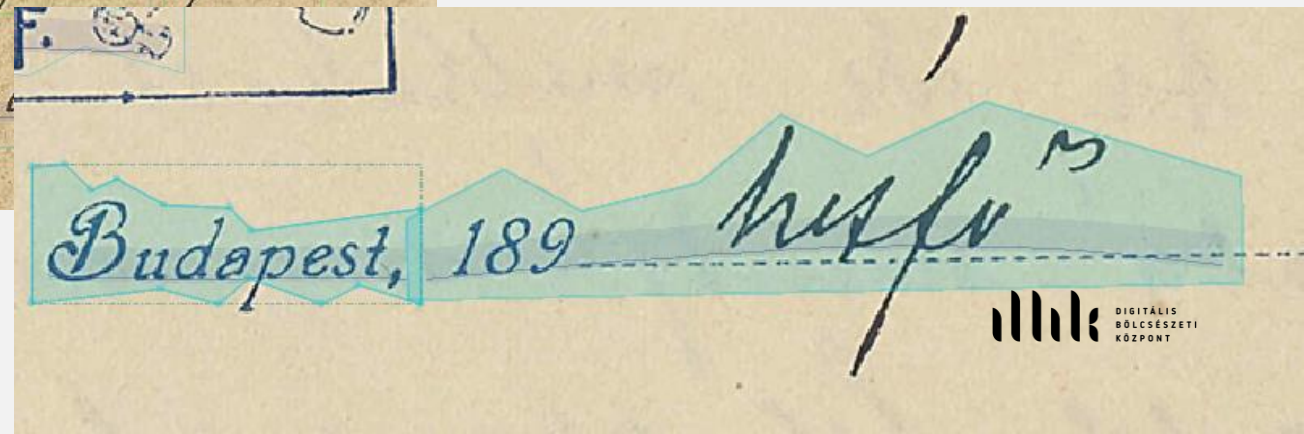
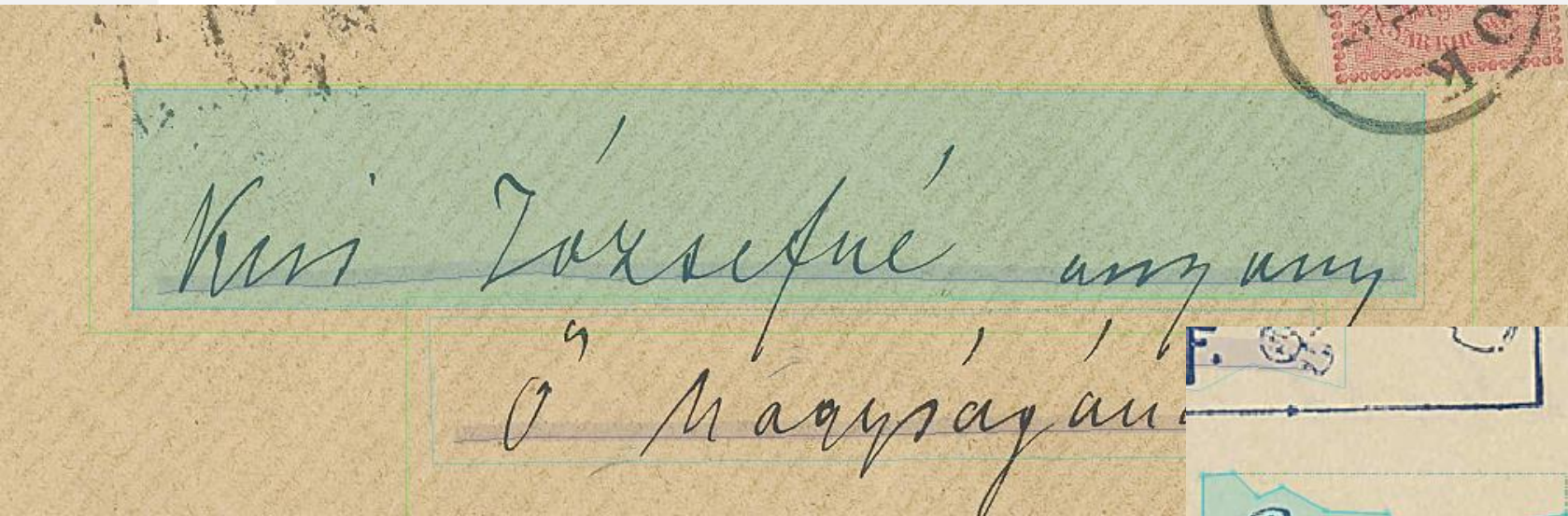


Értelmezési lehetőségek

- a felosztás módja: automatikus vagy manuális
- milyen alapmodell használatos milyen felosztással

Javítási lehetőségek

- adatmennyiség növelése
- Base Model használata
- szótárak beépítése
- polygon-ok alakjának manuális módosításával / eredeti alakjának megőrzésével
- a túltanulás elkerülése



„SIOFOK”

BALATON-FÜRDŐ RÉSZVÉNY-TÁRSASÁG.

Budapest, 1802. júl. 27.

Kedvesem! Megszoktem hazulról, mert egyre rosszabbul lettem.
Ma este Rólusra megyek. A mai napot Gíófokon és Balaton-
füreden töltöm. Ottónak tegnap 15 forintot küldtem.
Ne maradjatok a fekete házban. Ha nem jól érzeték
magatokat, különösen te, még mindig elmehetsz Bartfára,
vagy máshova. Milyen az ottam orvos? Vannak-e már
ismerőseid. Könyveket küldtem. Sányi arczképének nagyon
megörültem. Pá! Ölel és csakol híved

Zsófi

Leveles várat Rólusra. Minden nap újai palatka körü-
ltek. Pá!

AHÉT

MAJALY IRODALMI ÉS MŰVELÉSI KÖZLÖNY
KÖZMŰVELÉSI
SS JÓZSEF

TÁRSASÁGOK ÉS KIADÓ

HIVATAL

VII. ERZSÉBET-KÖRUT 6

Budapest, 185

húgli

Kedvesem!

Parkáncni és marad szeretni,
tehát én, annius levelemet most,
ma le nem mehetek, hanem csak
a his mégis egy jóval vasár-
nap. Vélhető, hogy jó egész nap.
Nem vagy. Vigyázz magadra,
nehogy valahányszor. 1. 7. 0

1-1 stofóe

2-1 Galaton-Pordó résztenz-Társaság.

3-1 Budapest, 1802 júl 27

4-1 Kedvesem! Megszoktem hazulról, mert egyre rosszabbul lettem.

4-2 Ma este Rólusra megyek. A mai napot Gíófokon és Balaton-

4-3 füreden töltöm. Ottónak tegnap 15 frtot küldtem.

5-1 Ne maradjatok a fekete házban. Ha nem jól érzeték

5-2 magatokat, különösen te, még mindig elmehetsz Bartfára,

5-3 vagy máshova. Milyen az ottam orvos. Vannak-e már

5-4 ismerőseid. Könyveket küldtem. Sányi arczképének nagyon

5-5 megörültem. Pá. Ölel és csakol híved

1-1

1-2 6 66a

1-3 s J

1-4 8

1-5 6

1-6 Ke Jst.

1-7 Kée KK. á

1-8 7á7

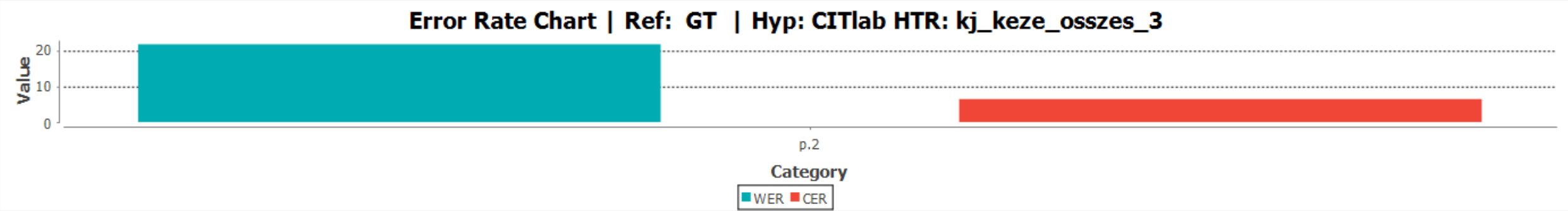
1-9 Kr á

1-10 Ti

Page	WER	CER	Word Acc	Char Acc	Bag Tokens ...	BT Recall	BT F1-Score	
Overall	21.82 %	6.52 %	65.15 %	88.36 %	0.798	0.806	0.798	

Page	WER	CER	Word Acc	Char Acc	Bag Tokens ...	BT Recall	BT F1-Score	
Page 2	21.82 %	6.52 %	65.15 %	88.36 %	0.8056	0.7909	0.79816513...	

Compare Text Versions for Page ..



Base folder:

File/Folder name:

Export path:

Download XLS

Kitekintés - lehetőségek



PETŐFI IRODALMI MÚZEUM

[Vissza a találati oldalhoz](#)[Gyűjtemények](#) [Névtér](#) [Díjak](#) [Adattárak](#) [Böngészés](#)

DIGIPHIL



HU EN

Címkés

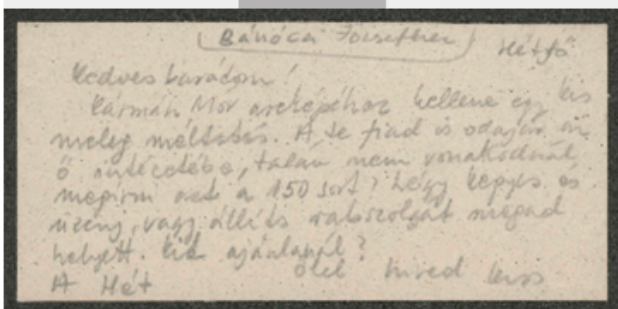
MARXML

LIDO

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

Letöltés

Megnyitás

**Levélíró:** Kiss József (1843-1921)**Címzett:** Bánóczy József (1849-1926)**Terjedelem:** 1 f**Megjegyzés:** Az eredeti levél ceruzairásos másolata?**Nyelv:** Magyar**Dokumentumtípus:** Levél

« < 1 > »

Múzeum ▾	Gyűjtemény ⇅	Hagyatéknev ▾	Raktári jelzet ⇅	Leltári szám ⇅
PIM	PIM - Kéziratanyag	Kiss József-hagyaték	V. 5239/1	V. 5239/1

« < 1 > »

Találatok a teljes szövegben (3 találat) :

1.oldal - állíts rabszolgát magad helyett. Kit ajánlanál? **Ölel híved Kiss** A Hét<https://resolver.pim.hu/bib/PIM1225902>

Kimeneti lehetőségek

- tei-xml
- pdf

Következő lépések

1. Letisztított, végleges modell létrehozása:

Kiss József által írt levelek egész korpuszán új modellt építeni a megszerzett tanulságok fényében -- későbbi alapmodell lesz egy vegyes kézírás-modellhez

1. Beépítés workflow-ba
2. Újabb modellek létrehozása (pl. Móricz-levelezés)
3. Modellek közzététele
4. Egyre jobb általános magyar nyelvű modell építése

Name	Language	Curator	Technology	Created	ID	nrOfWords
 Acta_17 HTR+	german, latin	alverman@uni-greifswald.de	CITlab HTR+	25.09.20	26284	594628
 Acta_17 PyLaia	german, latin	alverman@uni-greifswald.de	PyLaia	04.11.20	27337	594628
 Admiraliteit Zeeland 1606	Dutch	brndnbg@zeelandnet.nl	CITlab HTR+	18.09.19	16804	56765
 BM_Kortrijk_M1+	dutch	guenter.hackl@transkribus.eu	CITlab HTR+	04.06.19	14324	11915
 BnF_Newseye_M2+	french	guenter.hackl@transkribus.eu	CITlab HTR+	28.02.19	11203	434100
 Charter Scripts XIII-XV_M1	German, Latin, French	tobias.hodel@uzh.ch	CITlab HTR+	23.12.19	19872	938325
 Combined_Full_VKS_2	Church Slavonic	achim.rabus@slavistik.uni-freiburg.de	CITlab HTR+	14.04.19	12425	393079
 DAT 18. Jh M3b_Pylaia	German	anne-christine.guenther@tu-darmstadt.de	PyLaia	03.12.20	28419	102875
 Danish 1870-1950	Danish	ajam@aarhus.dk	CITlab HTR+	29.11.19	18968	348716
 Danish 1870-1950 v3.5	Danish	kpind@aarhus.dk	CITlab HTR+	06.06.20	24424	1603622
 Danish Fraktur SB 19th century v....	Danish	poul.steen@gmail.com	CITlab HTR+	20.02.20	21669	30983
 Devanagari mixed M1	Hindi	guenter	CITlab HTR	08.06.18	4017	25938
 Devanagari_nagara_M1	Hindi	merkel@ub.uni-heidelberg.de	CITlab HTR	12.11.18	7884	19243
 DiJeSt for Hebrew Script Languag...	heb, yid, lad, jrb	sinai.rusinek@mail.huji.ac.il	CITlab HTR+	15.04.20	23005	774192
 Dutch Margaretha Turnor 17th C...	Dutch	r.companje@hetutrechtsarchief.nl	CITlab HTR+	05.07.19	15303	36289
 Dutch Mountains (18th Century)	Dutch	jirsireinders1989@hotmail.com	CITlab HTR+	21.02.20	21683	1384893
 Dutch Notarial Model 18th Century	Dutch	jirsireinders1989@hotmail.com	CITlab HTR+	17.07.19	15708	622904
 Dutch manuscript poetry 1603-16...	Dutch	bj.m.caers@hum.leidenuniv.nl	CITlab HTR+	04.02.20	21187	51788
 Dutch_Gothic_Print	Dutch (16th, 17th, 18th century)	info@caromein.nl	CITlab HTR+	28.11.19	18944	51143
 Dutch_Gothic_Print_Pylaia	Dutch_Gothic	info@caromein.nl	PyLaia	16.09.20	26131	51143
 Dutch_Romantype_Print	Dutch	info@caromein.nl	CITlab HTR+	12.12.19	19423	88105
 Dutch_Romantype_Pylaia	Dutch_Roman_Print	info@caromein.nl	PyLaia	16.09.20	26138	88105
 Edelfelt M13+	swedish	maria.vainio-kurtakko@sls.fi	CITlab HTR+	31.10.19	18107	438817
 English Writing M1	English	Unknown	CITlab HTR	08.04.16	133	0
 English writing M2	English	guenter	CITlab HTR+	20.12.19	19809	1234795



Köszönjük a figyelmet!