Comparing ML OCR Engines on

Texts from 19th Century Written in the Romanian Transitional Script

Marc Frîncu, Marius E. Penteliuc, Simina Frîncu, Gheorghe Bran, and Manuela Zănescu

Comparing ML OCR Engines on

Texts from 19th Century Written in the Romanian Transitional Script

Talking Points

Context | Proposal | Experiments | Results | Conclusions

Context

Talking Points

Proposal

Experiments

Results

Conclusions

Latin ▶ Lorem ipsum dolor

RTS ▶ Лорем іпсᵾм дол

Cyrillic ▶ Лорем ипсум дол

4500+

RTS Documents from
9 queried Libraries

a lot more...

Documents from 3642 Libraries, Church and Museum Collections

IV.

Inвidiea.*)

Прin пiсма diаволѹлѹї а iнтрат моартеа
ꙟн лѹме шi чеї din партеа лѹї ꙟл кав-
тъ пре dъпсѹл. Ꙟцел. 2, 24.

Тот пъкатѹл din лъѹнтрѹ — din а iнiмеї токмеалъ,
Се ѹрzеще dѹпъ zica вѹнѹлѹї мънтѹiтор;
Ба ё фъкѹт ꙟн сiмцiмiнте маї наiнте de iвеалъ,
Маї наiнте de че фапта лар вedi ла черкътор. —
Кънd ънтеїї оаменї воїеа фъкъторѹлѹї кълкаръ,
Мърецiеа маї наiнте ї ꙟнтiнace кѹ пъкат.
Маї наiнте de че фѹрѹл ꙟшї ꙟнтinde неагра геаръ,
De пъкатѹл лъкомiеї кѹцетѹ'ї с'а шi стрiкат.
Маї наiнте de че фапта перѹшiнеї се плiнеще,
Сѹфлетѹл прin реоа понтъ тъпъ dѹс с'а шi сmiнтiт.
Маї наiнте de че Каiн ѹчiсесе тiрънеще,

Proposal

Talking Points

Context

Experiments

Results

Conclusions

Proposal

	Tesseract	Transkribus
Architecture	RNN with LSTM units	VGG with BLSTM units
Learning Type	Fine-tuning a Cyrillic Script model	Fine-tuning a Russian model
Learning Rate	0.0001	
Training Epochs	250, 50000	250

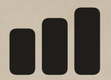
Experiments

Talking Points

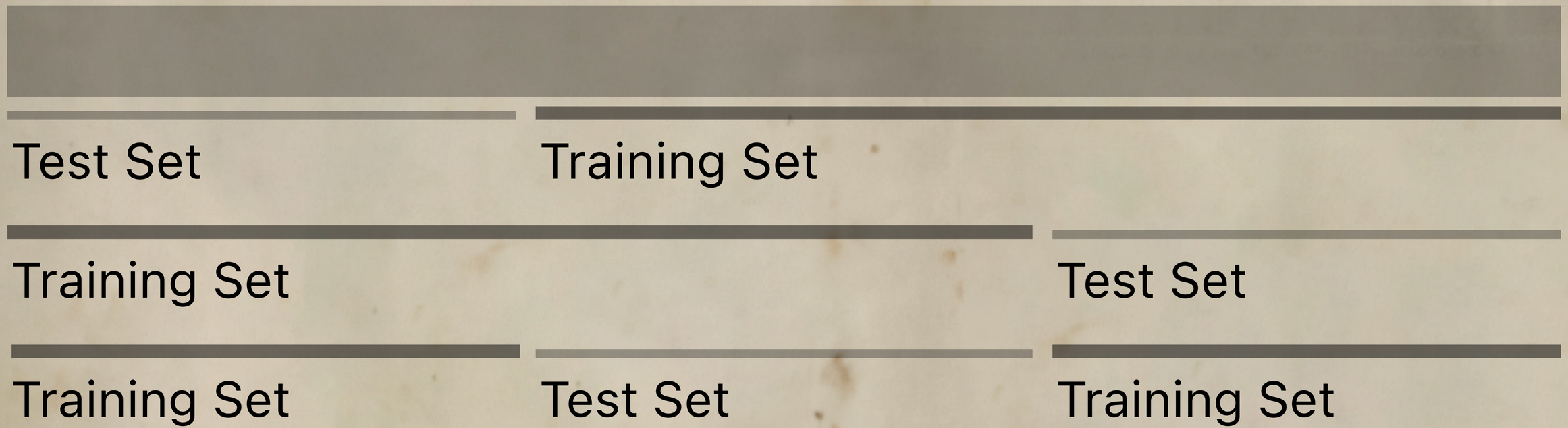
Context

Proposal

Results

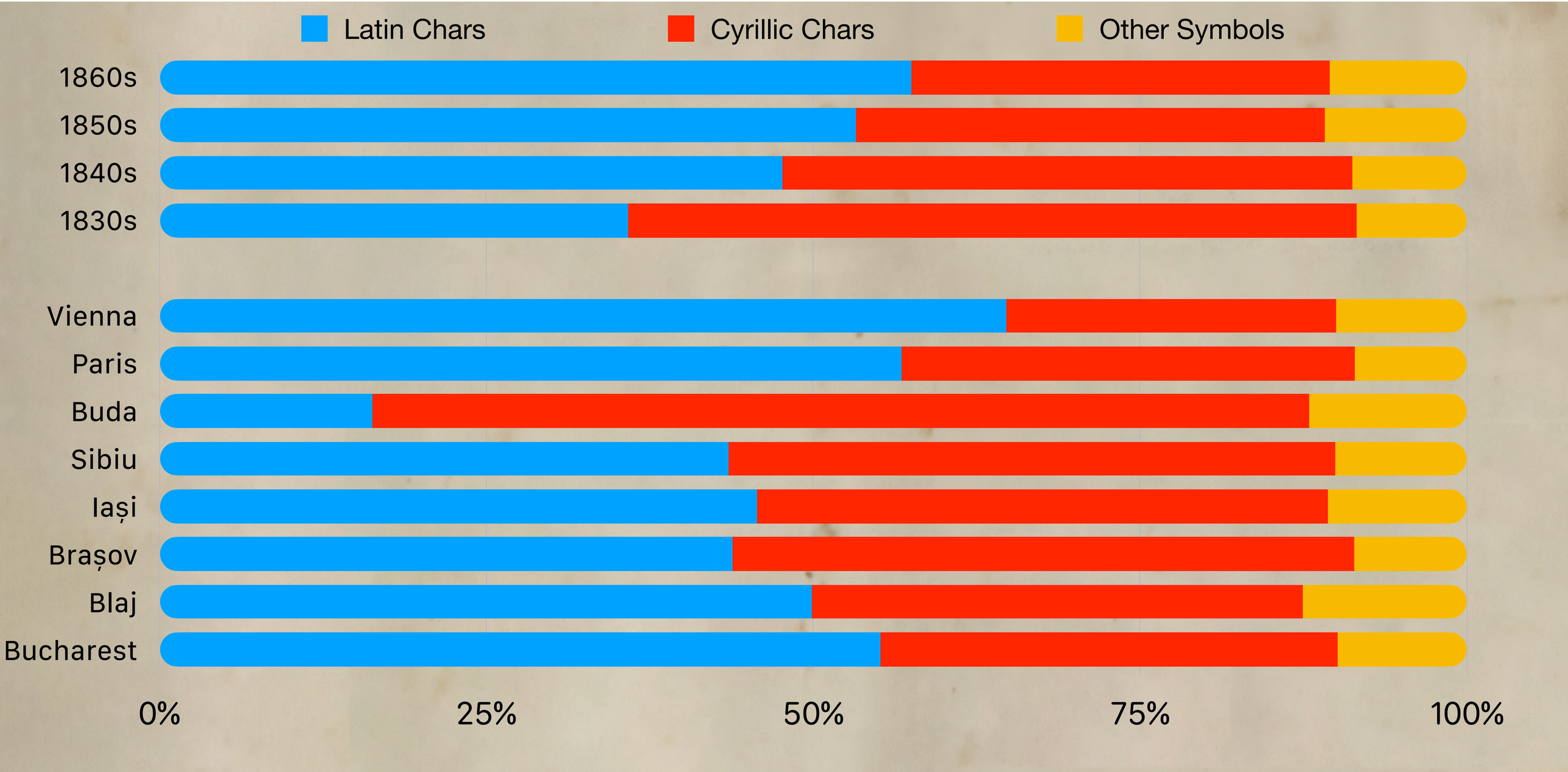


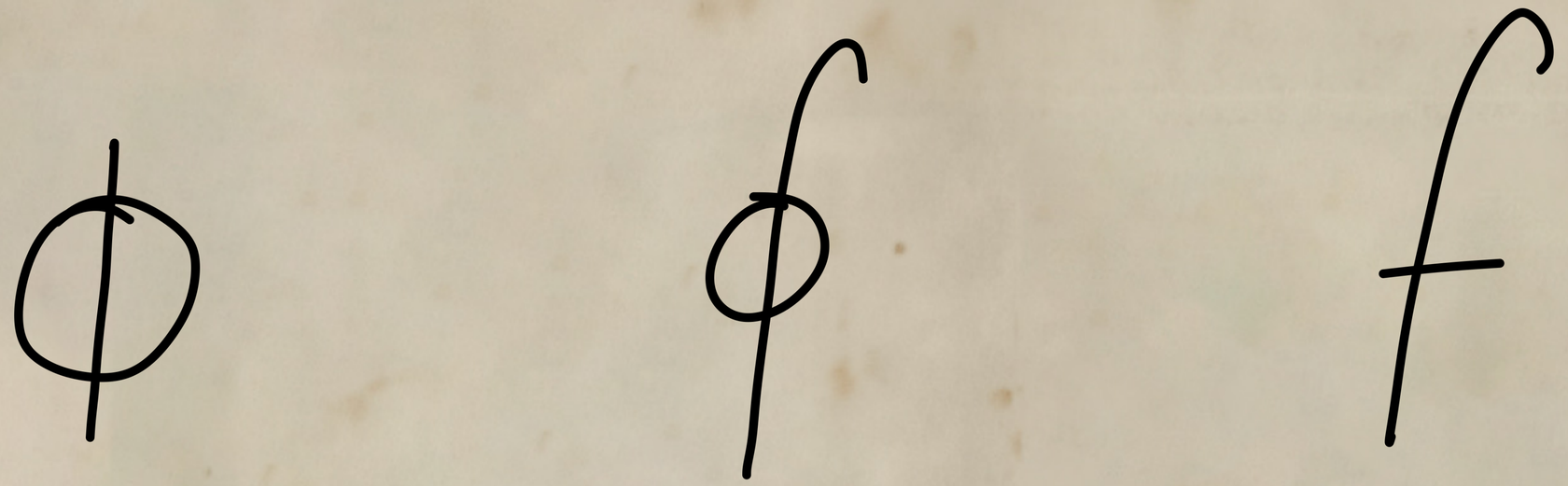
Conclusions


Test Set
Training Set
Training Set
Test Set
Training Set
Test Set
Training Set

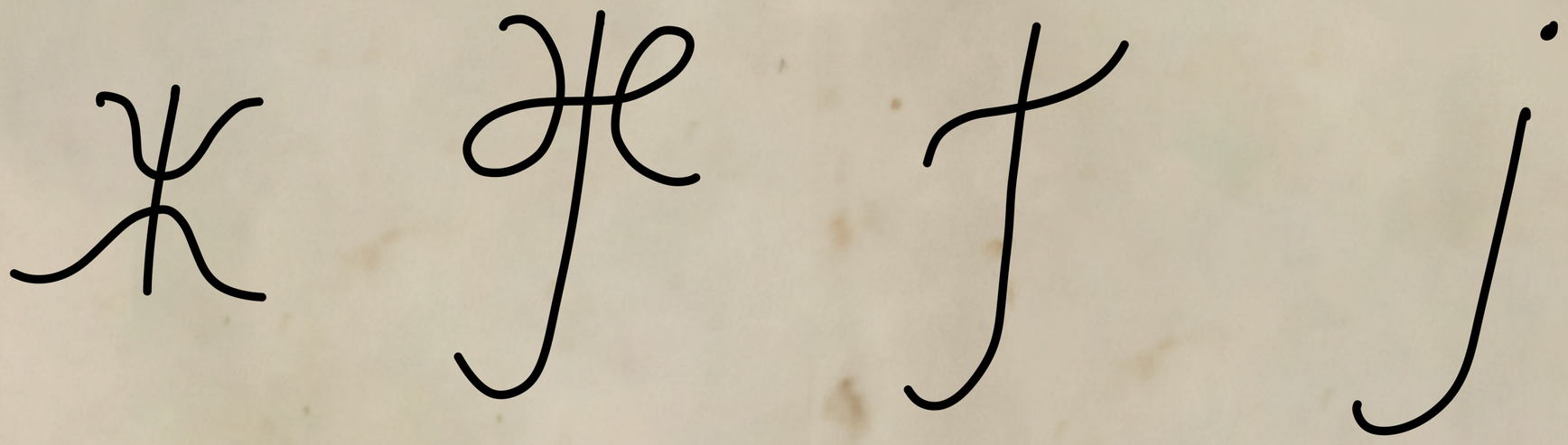
Average Dimension	300 × 900 → 2000 × 3000 pixels
Average Size	90kB → 10MB
	29 790 words
	153 286 characters

Experiments • Dataset

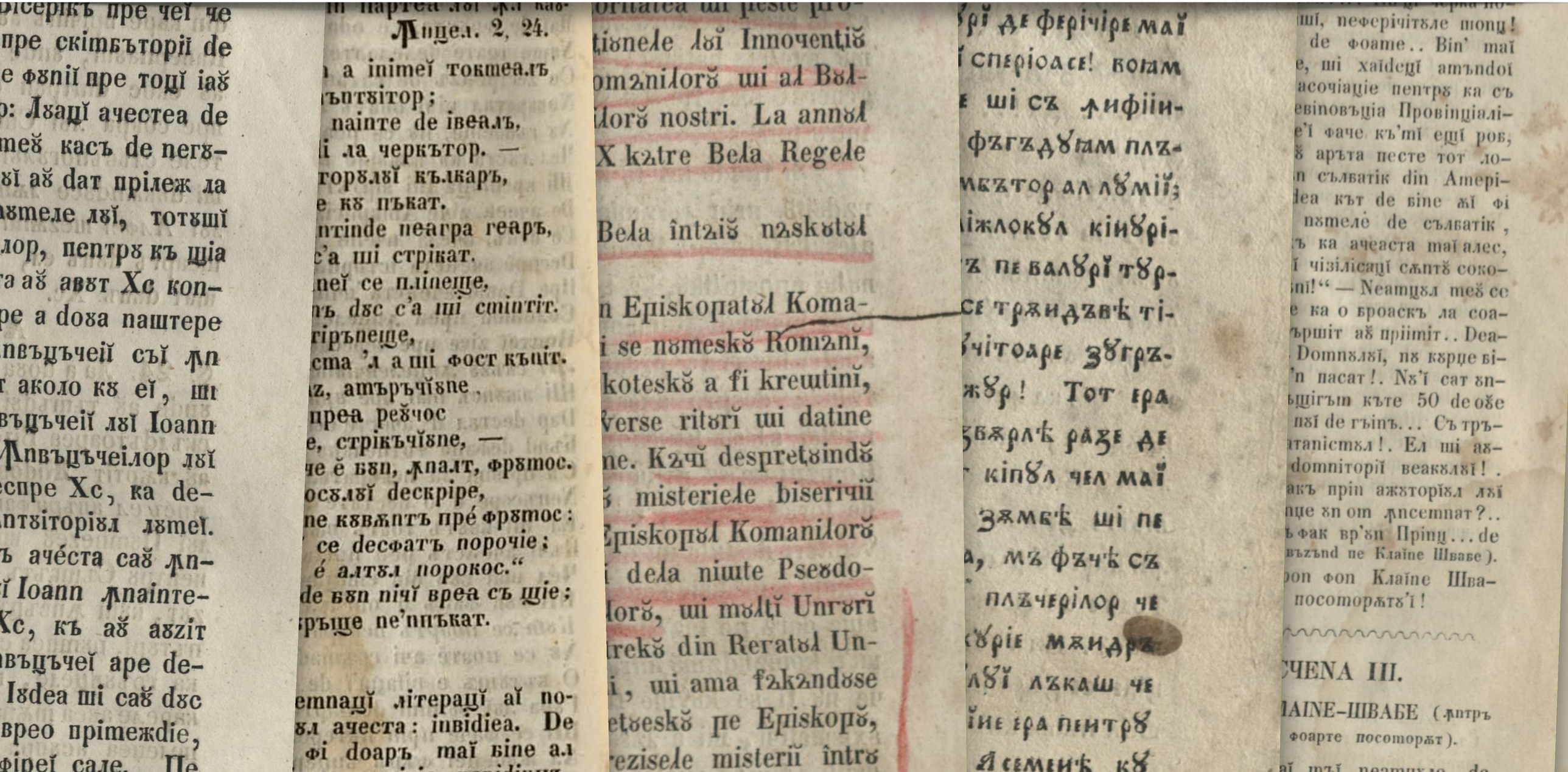




Chrillic F to Latin F



Chrillic J to Latin J



Лпцел. 2, 24.

n Episkopatŭl Koma-
i se numeskŭ Romanĭ,
koteskŭ a fi kreuținĭ,
Verse ritŭrĭ ui datine
ne. Kaĭ despretŭindŭ
misteriele biserichĭ
Episkopŭl Komanilorŭ
i dela niute Pseudo-
lorŭ, ui multĭ Unrŭrĭ
trekŭ din Reratŭl Un-
i, ui ama fakandŭse
etŭeskŭ pe Episkopŭ,
rezisele misterii întru

ui, neferichitŭle ponu!
de foame.. Bin' mai
e, ui haĭdeți amandoi
asochiație pentru ka sŭ
evinovățiia Provințiali-
e'i face kŭ'mi eți rob,
ŭ arăta peste tot lo-
n sălbatik din Ameri-
lea kŭt de bine mi fi
numele de sălbatik,
ŭ ka acheasta mai ales,
i chizilisați sfintŭ soko-
ni!" — Neatsŭl teŭ se
e ka o broaskŭ la soa-
ărșit aŭ priimit.. Dea-
Domnŭlui, nu karde bi-
'n pasat!. Nu'i sat un-
ășigăm kăte 50 de oăe
lui de găinŭ... Sŭ trŭ-
atanistŭl!. El ui aŭ-
domnitorii veakŭlui!.
akŭ prin ajutorĭul lui
nde un om apsemnat?..
ŭ fak vr'un Prinț... de
vzund ne Klaine Shvabe).
on fon Klaine Shva-
posotorăt'i!

CHENA III.

LAINE-ШВАБЕ (лптръ
фоарте посоторжт).

Scenarios	Groups	Tesseract	Transkribus
Region	8	✓	
Decade	4	✓	
Publisher	44	✓	
Transformations	1	✓	
OCR engine (control)	3	✓	✓

× 3 folds = 180 models

Results

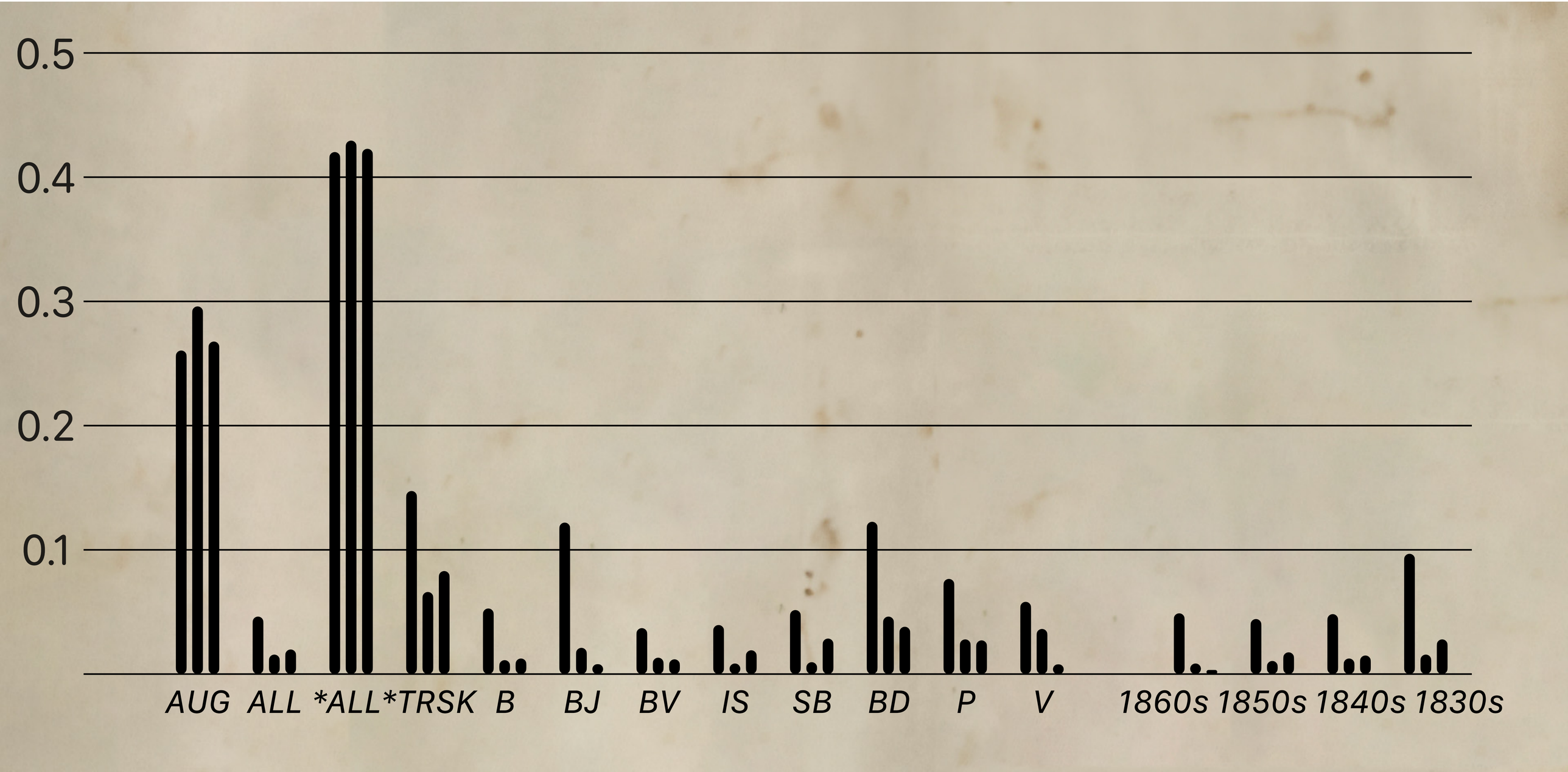
Talking Points

Context | Proposal | Experiments | Conclusions

Results

$$CER = \frac{S + D + I}{N}$$

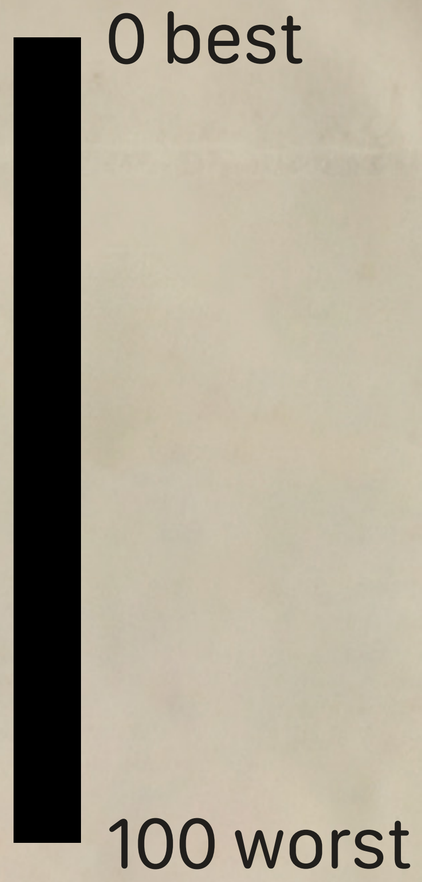
CER	Character Error Rate
S	Substitution
D	Deletion
I	Insertion
N	Number of Characters (baseline)


0.5
0.4
0.3
0.2
0.1
AUG
ALL
*ALL
*TRSK
B
BJ
BV
IS
SB
BD
P
V
1860s
1850s
1840s
1830s

Results • Performance

Results • Comparison

	1830s	*1840s*	*1850s*	*1860s*
1830s	0.10	0.13	0.18	0.33
1840s	0.11	0.05	0.09	0.17
1850s	0.15	0.08	0.03	0.10
1860s	0.15	0.08	0.03	0.05



Results • Comparison

	1830s	*1840s*	*1850s*	*1860s*
1830s	0.10	0.13	0.18	0.33
1840s	0.11	0.05	0.09	0.17
1850s	0.15	0.08	0.03	0.10
1860s	0.15	0.08	0.03	0.05

Results • Comparison

	1830s	*1840s*	*1850s*	*1860s*
1830s	0.10	0.13	0.18	0.33
1840s	0.11	0.05	0.09	0.17
1850s	0.15	0.08	0.03	0.10
1860s	0.15	0.08	0.03	0.05

	B	*BJ*	*BV*	*IS*	*SB*	*BD*	*P*	*V*
Bucharest	0.04	0.39	0.30	0.13	0.23	0.45	0.27	0.30
Blaj	0.20	0.08	0.50	0.40	0.50	0.39	0.56	0.52
Brașov	0.04	0.36	0.04	0.03	0.05	0.31	0.20	0.32
Iași	0.05	0.42	0.15	0.04	0.10	0.36	0.20	0.34
Sibiu	0.05	0.38	0.10	0.04	0.04	0.29	0.20	0.33
Buda(pest)	0.15	0.22	0.43	0.41	0.48	0.08	0.58	0.63
Paris	0.05	0.43	0.25	0.06	0.16	0.47	0.06	0.23
Viena	0.14	0.45	0.44	0.24	0.32	0.58	0.35	0.07

	B	*BJ*	*BV*	*IS*	*SB*	*BD*	*P*	*V*
Bucharest	0.04	0.39	0.30	0.13	0.23	0.45	0.27	0.30
Blaj	0.20	0.08	0.50	0.40	0.50	0.39	0.56	0.52
Brașov	0.04	0.36	0.04	0.03	0.05	0.31	0.20	0.32
Iași	0.05	0.42	0.15	0.04	0.10	0.36	0.20	0.34
Sibiu	0.05	0.38	0.10	0.04	0.04	0.29	0.20	0.33
Buda(pest)	0.15	0.22	0.43	0.41	0.48	0.08	0.58	0.63
Paris	0.05	0.43	0.25	0.06	0.16	0.47	0.06	0.23
Viena	0.14	0.45	0.44	0.24	0.32	0.58	0.35	0.07

	B	*BJ*	*BV*	*IS*	*SB*	*BD*	*P*	*V*
Bucharest	0.04	0.39	0.30	0.13	0.23	0.45	0.27	0.30
Blaj	0.20	0.08	0.50	0.40	0.50	0.39	0.56	0.52
Brașov	0.04	0.36	0.04	0.03	0.05	0.31	0.20	0.32
Iași	0.05	0.42	0.15	0.04	0.10	0.36	0.20	0.34
Sibiu	0.05	0.38	0.10	0.04	0.04	0.29	0.20	0.33
Buda(pest)	0.15	0.22	0.43	0.41	0.48	0.08	0.58	0.63
Paris	0.05	0.43	0.25	0.06	0.16	0.47	0.06	0.23
Viena	0.14	0.45	0.44	0.24	0.32	0.58	0.35	0.07

Results • Time

Conclusions

Talking Points

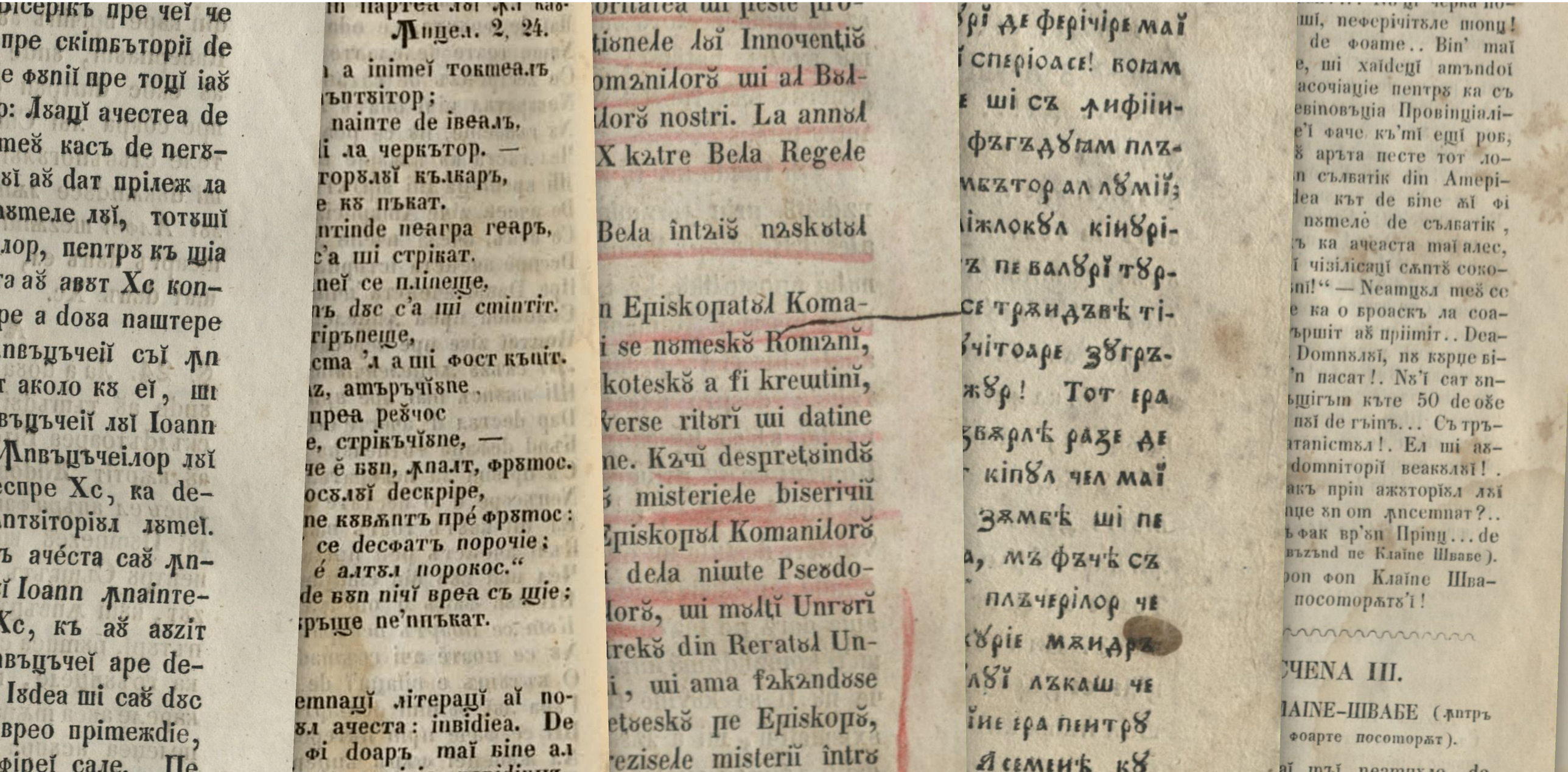
Context

Proposal

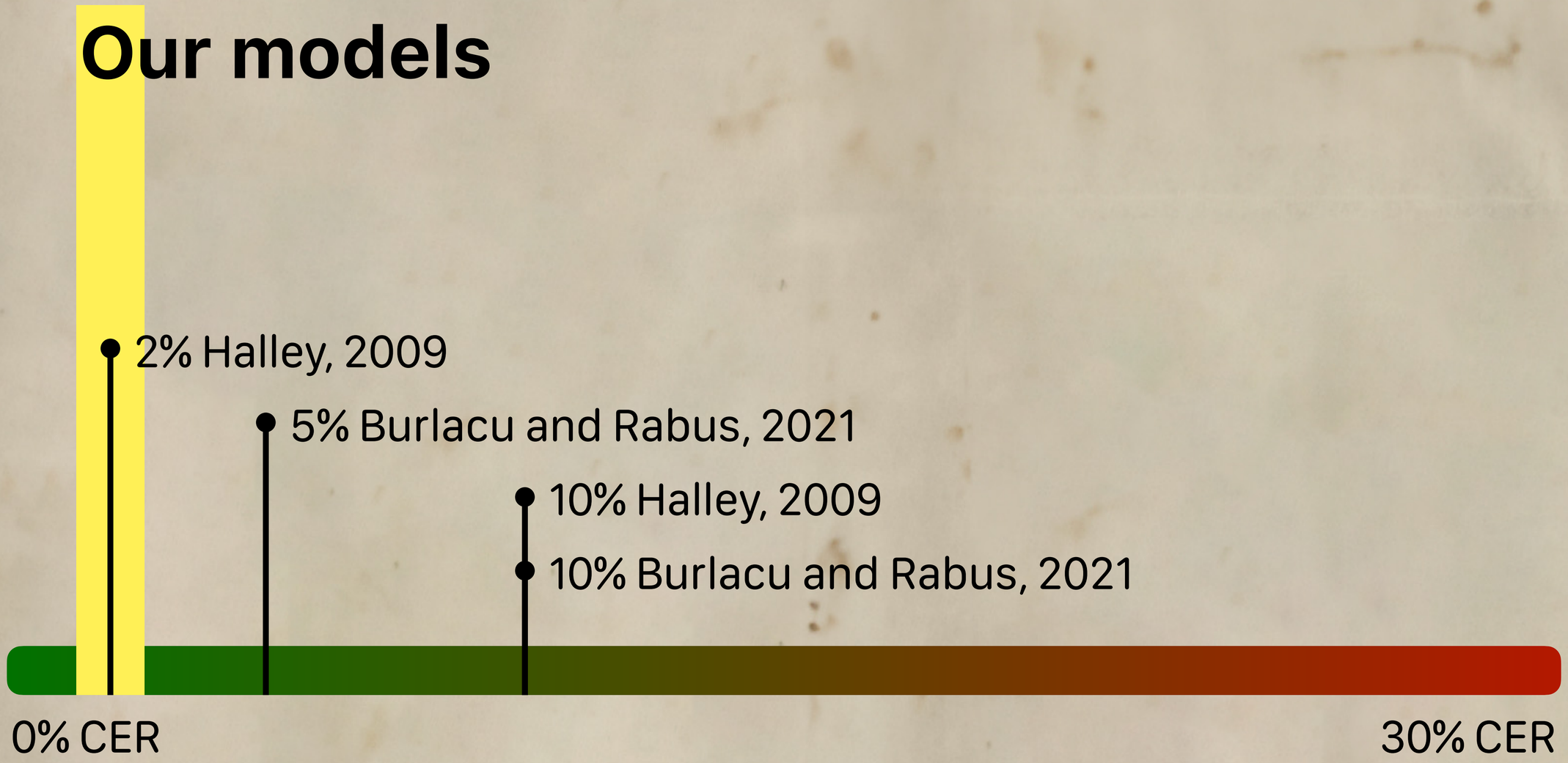
Experiments

Results

Conclusions



Our models


2% Halley, 2009
5% Burlacu and Rabus, 2021
10% Halley, 2009
10% Burlacu and Rabus, 2021
0% CER
30% CER

Conclusions

RTS to Latin

Transliteration Service

Playground | Box Editor | Words Pairing

Upload Documents

Image

Choose File | no file selected

Box File

Choose File | no file selected

Download Documents

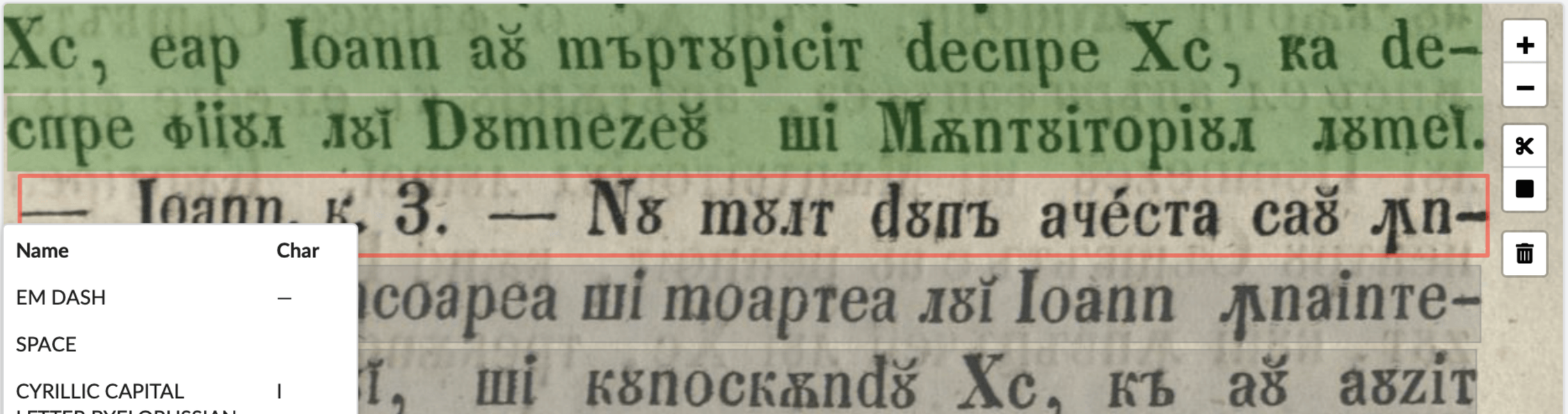
Box File

sampleImage.box

Ground Truth

sampleImage.gt.txt

¶ | Romanian Transitional Script custom



Conclusions

Search

MARIUS E. PENTELIUC AND 1 COLLABORATOR · UPDATED 6 MONTHS AGO

27 | New Notebook | Download (133 MB)

19th-Century Romanian Transitional Script

OCR Dataset for the Romanian Transitional Script of 19th Century

Data Card | Code (1) | Discussion (0) | Settings

About Dataset

Edit

This dataset consists of 147 pages of Romanian texts written in the **Romanian Transitional Script** (RTS). RTS is a mix of Latin and Cyrillic characters that were used in the 19th century in the Romanian provinces to facilitate the transition from the Romanian Cyrillic Script to the modern Latin Script. The images cover the period between 1833 and 1864. The selected texts cover a diverse range of literary genres, including poems, novels, dramas, stories, newspapers, and religious texts.

The dataset was obtained from the Central University Libraries (BCU) of Timișoara, Iași, and Cluj-Napoca through their free online platforms or by request. The scanned images are provided in JPEG and PNG formats, with dimensions ranging from approximately 300 by 900 pixels to 2000 by 3000 pixels. The file sizes vary between 70 KB and 10 MB.

To ensure **diversity**, the dataset includes images with *various fonts, styles, regions, publishers, and years*. It *covers all three main Romanian provinces' key publishing regions* (Bucharest - B, Iasi - IS, Brasov - BV, Sibiu - SB, Blaj - BJ) as well as *some located outside Romania* that printed texts in RTS (Vienna - V, Budapest - BD, Paris - P). It comprises 4353 lines of text, totaling 29,790 words and 153,286 characters. Among these characters, there are 59,819 Cyrillic characters, 25,837 Latin characters, 52,163 overlapping characters (identical symbols), and 15163 other characters (e.g., punctuation, digits). The images below summarize its content per publisher and decade. More statistics (including

Usability
10.00

License
Other (specified in descripti... Edit

Expected update frequency
Quarterly Edit

Tags
Image | Text | Linguistics | Research | Model Comparison

Comparing ML OCR Engines on

Texts from 19th Century Written in the Romanian Transitional Script

Comparing ML OCR Engines on

Texts from 19th Century Written in the Romanian Transitional Script

Q/A

Backup Slide

RTS	Romanian Transitional Script
RNN	Recurent Neural Network
VGG	Very Deep Convolutional Networks
B/LSTM	Bidirectional / Long Short-Term Memory
OCR Flavors	Hand-writing, printed text, and case-specific
Tagging	30 minutes using Custom Designed Box Editor
Dataset Span	147 pages between 1833 — 1864
Unique Chars	Not appearing in Unicode
CER Score	Excluding spaces due to inconsistent use

Backup Slide

Dataset

Project

Playground