Challenges and Solutions in

# Transliterating 19th-Century Romanian Texts from the Transitional to the Latin Script

#### Challenges and Solutions in

# Transliterating 19th-Century Romanian Texts from the Transitional to the Latin Script

#### **Talking Points**

Context

**Related Work** 

Challenges



Proposal

**Experiments** 

Conclusions







Latin >

• Lorem ipsum dolor

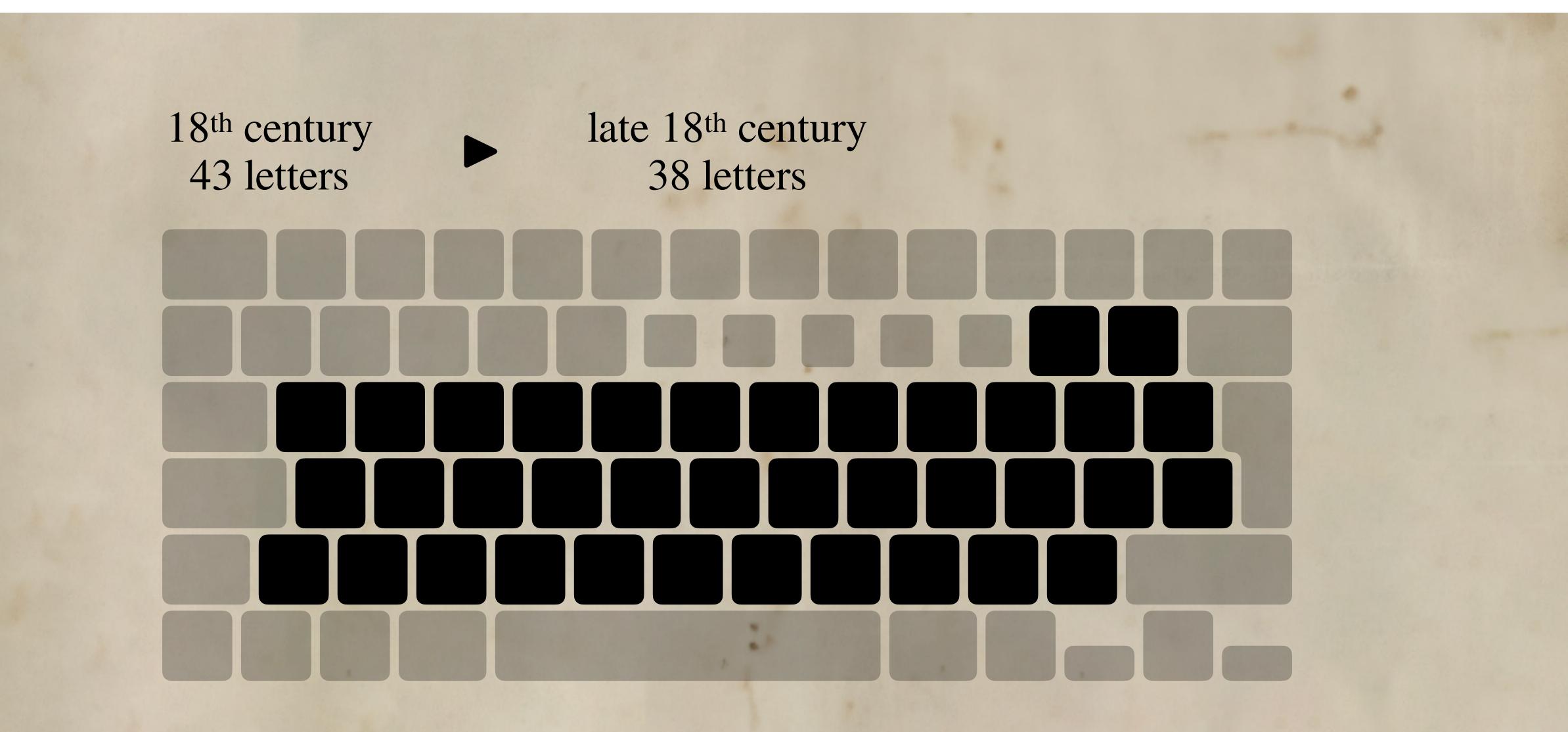
Cyrillic >

- Лорем ипсум доло

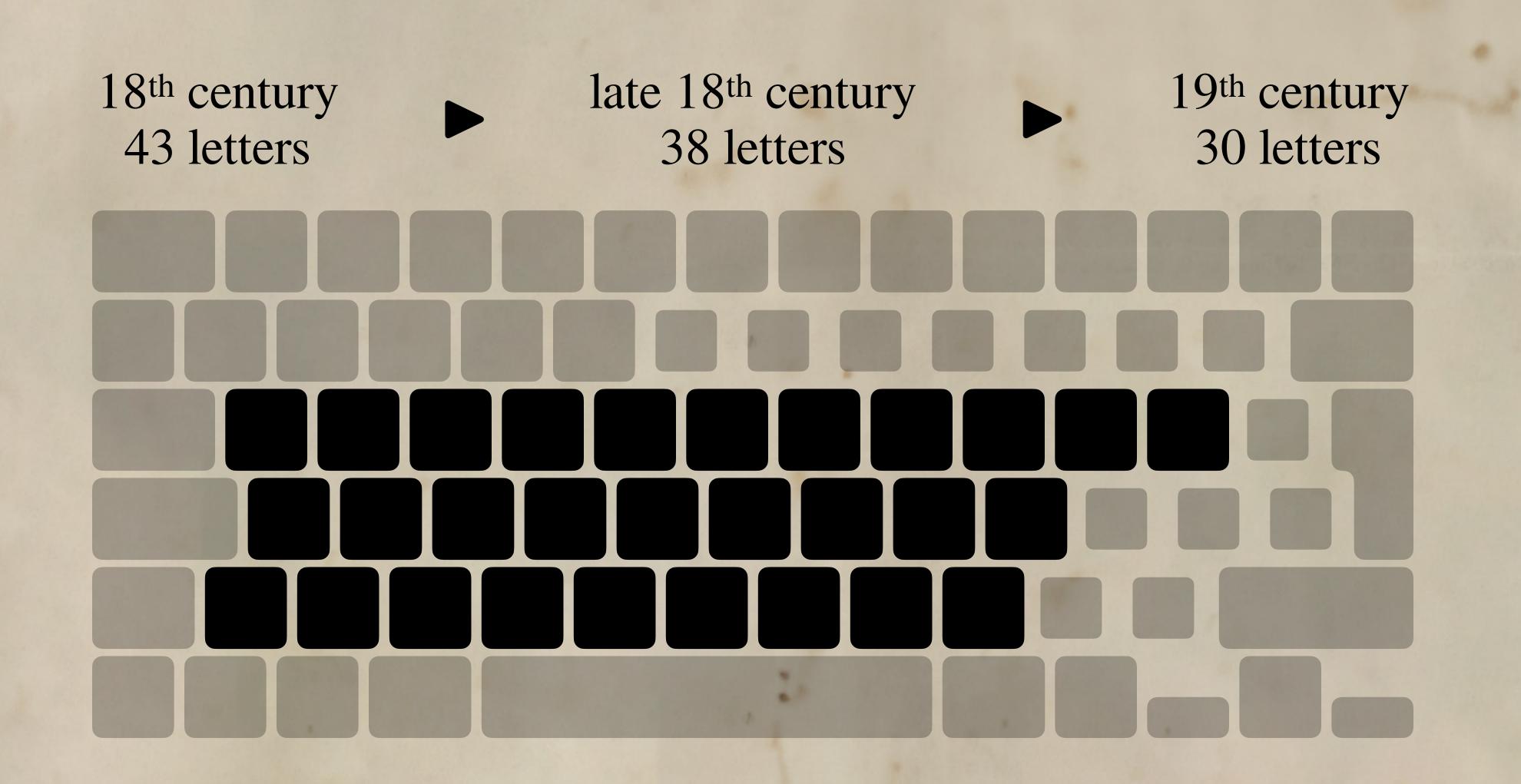




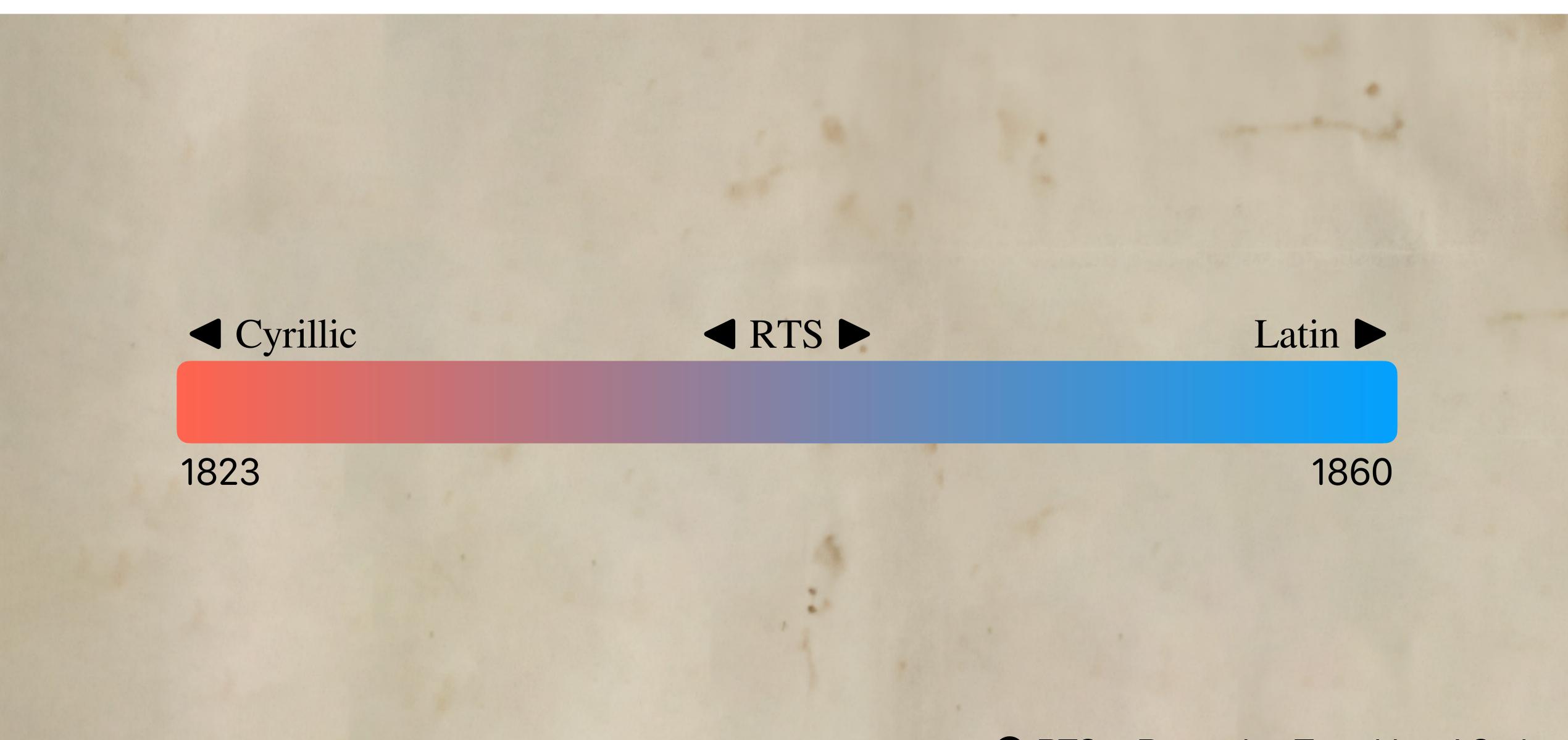












#### Аптыва Пасхъ.

Фiind апро́пе Пацціле Жідовілорь, сав світ Іісьс да Гервсаліт, ші ав гъсіт да Бісерікъ пре чет че віndea ої, пяї ші порятьї, ші пре скіть торії de ғапі шеzжnd, ші фъкжnd вічів de фвиії пре тоці іав скост din Бісерікъ, ші ат zic лор: Лваці ачестеа de аічеа, ші пу фачеці каса татълуі тей касъ de nersцъторіе. De ші фапта ачеста а ляї ай daт прілеж ла твлиї, ка съ креадъ длярв пвтеле лві, тотвші licsc их се дикреdea пре cine лор, пептря къ щіа пре тоці. Іоапп к. 2. Dвпъ ачеста ав авыт Xc копворыре татпікъ кв Nikodim decupe a dosa naштере ші тжитвіре, ші сай две кв дивъцъчей съй ди пътжитил Івдеїї, щі ай петрекит аколо ки еї, щі ав ботегат. — Маї de парте Апвъцъчей ляї Ioann ай вестіт ляї феспре дипиндіреа Ливъцъчеілор ляї Xc, eap Ioann ав търтврісіт decupe Xc, ка deспре фіївл ляї Двипегей ші Мжитвіторівл лящеї. — Іоапп. к. 3. — Ne meat dent avécta caë anтътплат пріпсоареа ші тоартеа ляї Ioann Anainteтергъторівляї, ші квпоскалов Хс, къ ав авгіт фарісеіт, кът къ ел таї твлиї дивъцъчеї аре deкжт loann: пептры ачееа ай льсат Isdea mi сай dыс ди Галілеа, ка съ ns віпъ ди врео прітежdie, къчі дикъ из вепісе вретеа жертоіреї сале. Пе калеа ачеста і сай длятьшилат де ай авит ворбіре ки msiepea Camapineanкъ ла пядял ляї Іаков, впde STYT do susana muiona ven mi da vong vonit Cama

- Interesting research topic.
- ► Multi-step study process.
- ► OCR fails to recognize RTS.
- ► New ML models need large volume of documents.
- ▶ Big Data problem.

► Same sound, same graphic

OAEIKMOT

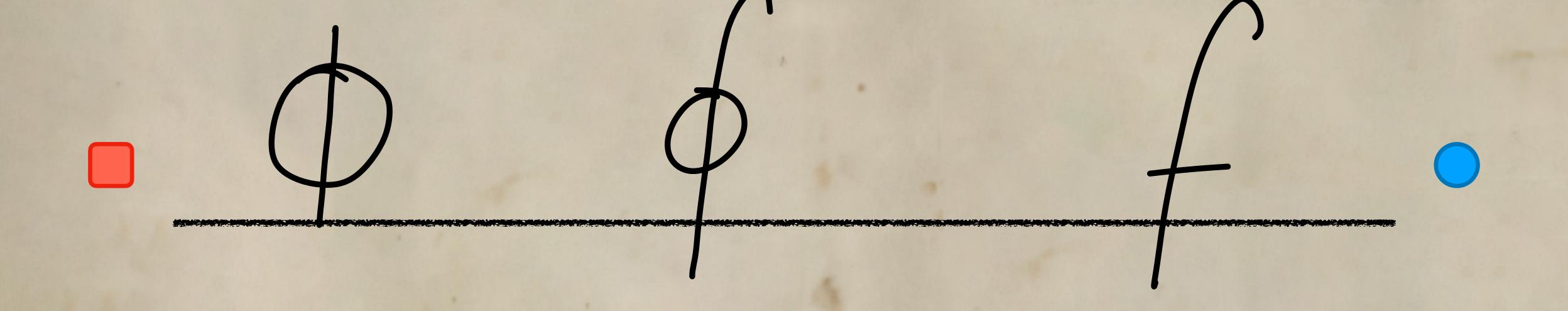
OAEIKMOT

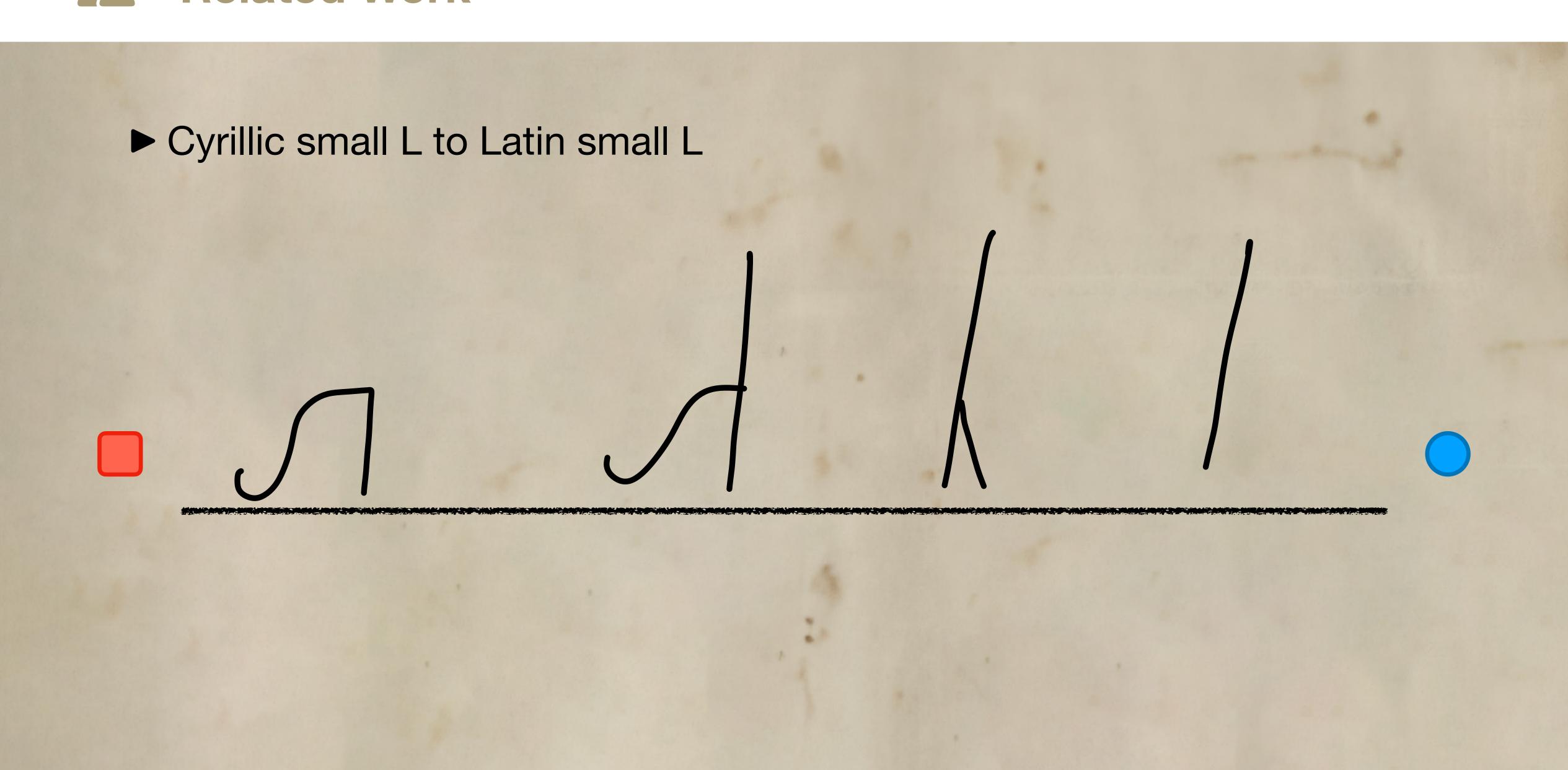
► Same sound, different graphic

OVSNRH

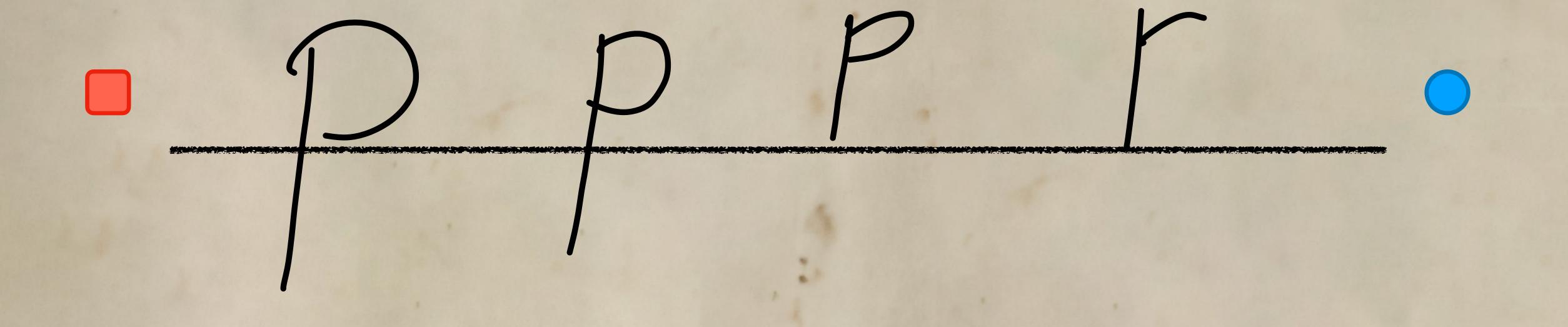
BCHPX

► Cyrillic small F to Latin small F

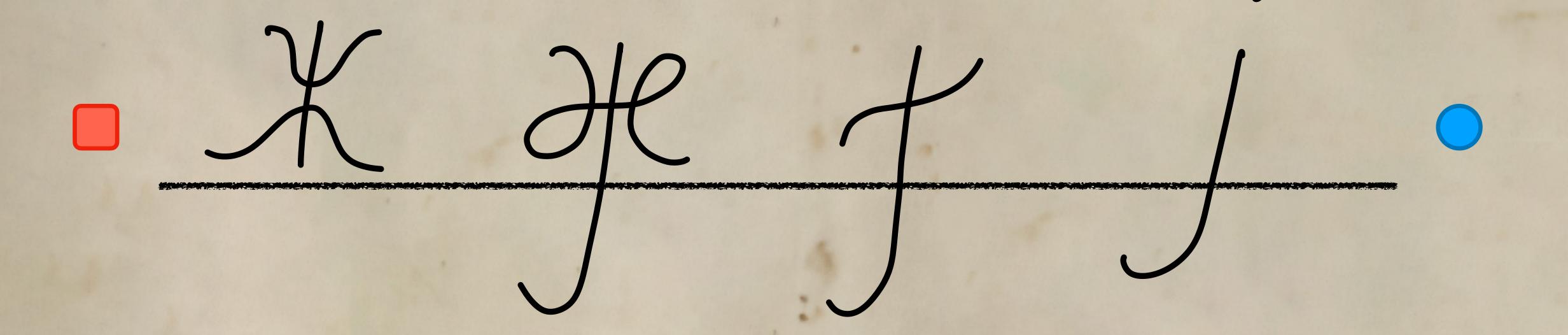


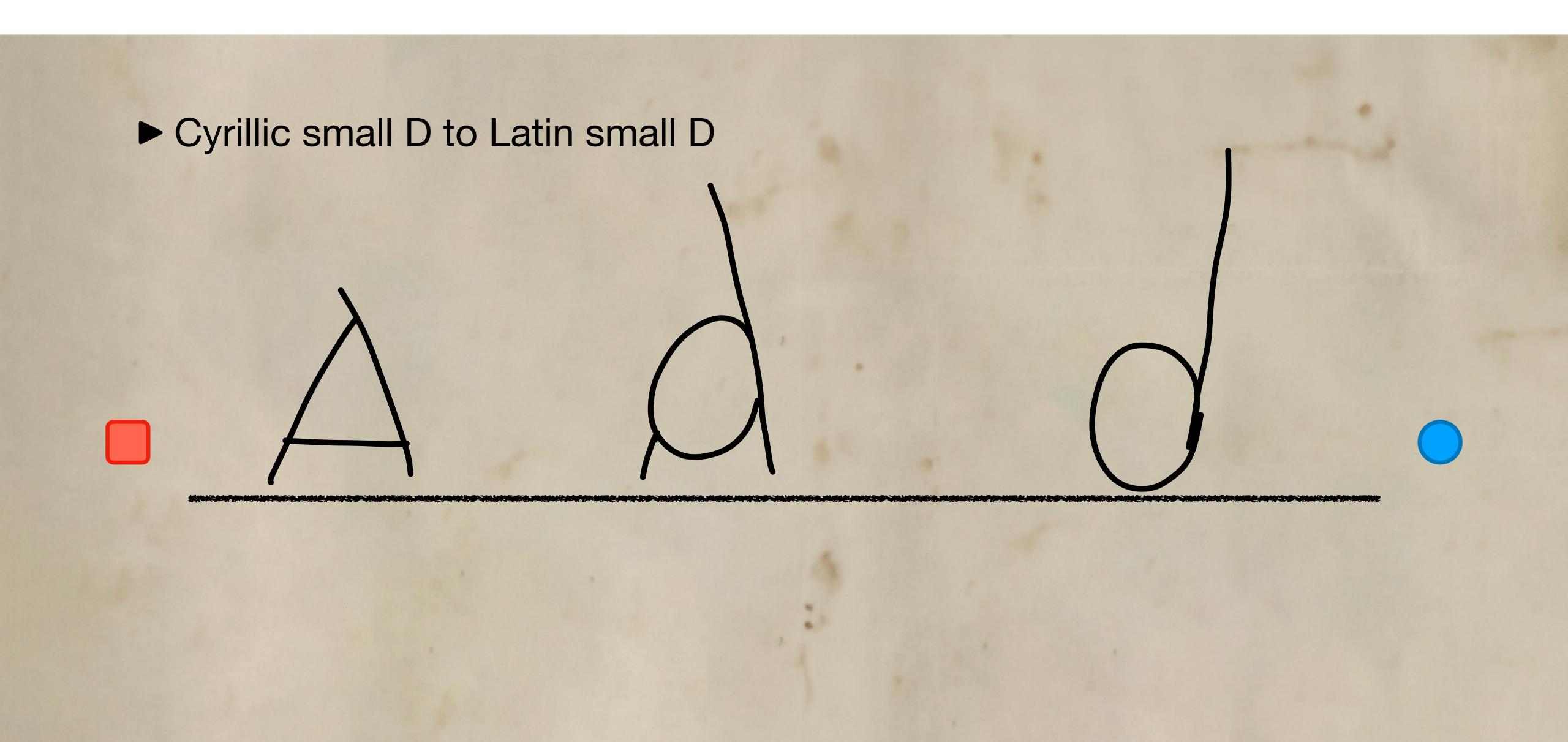


► Cyrillic small R to Latin small R



► Cyrillic small J to Latin small J





#### Related Work

Boian et al., 2014

Cojocaru et al., 2016

Demidova and Burteva, 2017

Burlacu and Rabus, 2021 **Dataset Size** 

Unavailable

Unavailable

Unavailable

30 900 words

Software

ABBY FineReader (proprietary)

ABBY FineReader (proprietary)

ABBY FineReader (proprietary)

Transkribus (proprietary)

Result

63–95.4% correct words

96% Accuracy

99% Accuracy

10% CER (lower is better)

Our Study

30 pages

Tesseract (open source)

## Challenges · Processing

Document
Scanning
Training Set
Labeling
Validation Set
Labeling
Testing Set
Labeling
ML Model Training

Effort	
Manual	
Manual	
Manual	
Manual	
Automated	

	Time
	Long
30	0 min / page
30	0 min / page
30	0 min / page
	- 2200 sec / raining fold

Space
100 KB - 10 MB / scan
3 - 10 KB / file
3 - 10 KB / file
3 - 10 KB / file
50 MB / checkpoint 9 MB / trained



#### **Challenges • OCR and Transliteration**

#### AGE-RELATED DAMAGE

- ► Thick binding, ripped stitching, or broken spines.
- ➤ Creases, folds, wrinkles, and undulation due to humidity changes.
- ► Moisture halos, ink discoloration, foxing, burns, tearing, grease stains, glue residue.
- ▶ Presence of post-printing elements.

#### OCR CHALLENGES

- ► Typesetting using various inks, typefaces, and fonts.
- ► Text visible from the verso.
- ➤ Single vs multi-column layouts, framed and/or manually underlined text.
- ► Glossing with marginal or interlinear notations.

# 1

### **Challenges • OCR and Transliteration**

отсерткъ пре чет че пре скіть торії de е филії пре тоці іай о: Лваці ачестеа de тей касъ de nersві ай дат прілеж ла втеле ляї, тотяші лор, пептря къ щіа а ав авыт ХС копре а dosa паштере пвъцъчей съй дл г аколо кв еї, ші въцъчей ляї Ioann Іпвъцъчеілор льї спре Xc, ка deптвіторівл летеї. в ачеста сав дпi Ioann Anainte-1860 Brilly let ape de-Isdea Sibiu врео пришежие, pipei care.

Лицел. 2, 24. а іпітеї токшеаль ъптвітор; nainte de івеаль, і ла черкътор. горуляї кълкаръ, е кв пъкат. nrinde nearpa геаръ, са ші стрікат. пет се плинеще, пъ две с'а щі стіптіт. ста 'л а ші фост къпіт. г, атъръчівпе, преа ревчос е, стрікъчівне, че е въп, жпалт, фрвтос. осълът дескріре, пе кввжитъ пре фрвтос: се десфатъ порочіе; е алтъл порокос." de въп пічт вреа съ щіе; **1847** 

Brașov

фі doapъ maĭ вine ал

tisnele loi Innogentis omanilors wi al Boldors nostri. La annal X katre Bela Regele Bela întziš naskutul n Episkopatal Komase nemeske Romani, koteskš a fi kreutini, verse riteri wi datine ne. Kzui despretsinds misteriele biseriqii dpiskopsl Komanilors dela niute Pserdolors, wi malti Unrari **1850** ull Vienna misterii întro

PI DE PEPITIPE MAI CHEPIOACE! BOIMM е ші съ лифіінφZΓZA8mm nAZ= MEZTOP AN NEMITS іжлоква кінврі-Z HE BANSPITSP-CE TPAHAZBE TI-PHITOAPE 38 FPZж8р! Тот гра SBRON'S PAGE AF KINSA HEA MAI 38ME'E WI DE а, ма фачь са ПАДЧЕРІЛОР ЧЕ 1835 Bucharest ACEMENIK K8

ші, пеферічіткле шопц! de poame.. Bin' mai e, mi xaideni amundoi асочіаціе пентря ка съ евіповъціа Провіпціаліей фаче къ'ті ещі роб, в аръта песте тот лоп сълватік din Amepidea кът de sine мі фі пътеле de сълбатік, ъ ка ачеаста таї алес, чізілісаці сжитё сокоmi! " — Neamusл meš се е ка о вроаскъ ла соаършіт аў пріітіт.. Dea-Domnsляї, па карце ві-'n пасат!. Ns'ї сат впьщігът къте 50 de обе път de гътпъ... Съ тръатапіствл!. Ел ші авdomniтopiĭ веаквлвĭ!. акъ прін ажеторієл леї пце вп от дисетнат?.. ь фак вр'яп Пріпц...de въхънд пе Клаїне Шваве). ооп фон Клаїне Швапосоторжтв'і! **1852** CHENA III.

lași

фоарте посоторыт).

#### Challenges · OCR and Transliteration

вindea ої, пві ші порвтыї, ші пре скітьторії de ғапі шеzжnd, ші фъкжnd вічів de фвий пре тоці іав скося din Бісерікъ, ші ая zic лор: Лваці ачестеа de аічеа, ші пу фачеці каса татълуї тей касъ de nersцъторіе. De ші фапта ачеста а люї ай daт прілеж ла твлиї, ка съ креадъ длярв пвтеле ляї, тотвші licsc их се дикреdea пре cine лор, пептря къ щіа пре тоці. Іоапп к. 2. Dвпъ ачеста ав аввт Хс копворые татпікъ кв Nikodim decnpe a dosa naштере ші тжптвіре, ші сав двс кв дпвъцъчей съй дп пътжитил Івдеїї, щі ай петрекит аколо ки еї, щі ав ботегат. — Маї de парте Ливъцъчей ляї Ioann ай вестіт ляї феспре дитивиціреа Ливъцъчеілор ляї Xc, eap Ioann ав търтврісіт decпре Xc, ка deспре фіївл леї Детпегей ші Мжптвіторівл летеї. — Ioann. к. 3. — Ne ment dent ачеста саё лптътплат пріпсоареа ші тоартеа ляї Ioann Anainteтергъторівляї, пі впосканій Хс, къ ай авгіт 1860 фарісеії, кут къ дивъцъчеї ape deкжт loann: пептря / Sibiu car Isdea mi cas dsc лп Галілеа, ка им вить ди врео прітежdie, къчі дикъ из вепісе вретеа жертфіреї сале.

іпцел. 2, 24. еї токшеалъ р; de iвeaль, кътор. еагра генръ, рікат. uneme, a mi cmintir. пі фост къпіт. qĭsne, lisne, кпалт, фрвтос. скріре, ъ пре фрвтос: тъ порочие: порокос." вреа съ щіе; пъкат. ARREST OF SA тераці аї поinsidiea. De mai sine a.1

si Innoventis rs mi al Boltri. La annol Bela Regele паtal Komaeskš Romani, a fi krewtini, sri wi datine despretsinds iele biseriqii I Komanilors imte Psesdomølti Unreri Rerated Unna fakandsse пе Episkops, misterii întro

PITIPE MAY CE! ROMM **У**ИФІІН-MAM TAZ-18р1 т8р-ZBK Ti-38rpz-TOT EPA PASE AE HEA MAI Mi UE 24 CZ INOP 4E RHAPE ALL 4E БЕ (мптръ ентр8 эторыт).

птвле шопц! ... Bin' maï ni amundoi ептръ ка съ Провінціаліті ещі ров, сте тот лоdin Amepisine al oi е сълбатік, та таї алес, сжитё сокопцвл тей се скъ ла соаlimir.. Deaпв кврџе ві-Na'i car ane 50 de oše ... Сътръ-Ел ші аввеакълъї!. квторівл лві incemnar?.. Пріпц...de taine III Base ). raine IIIBara'i! mmm

## Ţ.

neжdie,

e. He

#### **Challenges • OCR and Transliteration**

ть пре dжиски. Пидел. 2, 24. upe पर्धा पर **Бъторіі** de Тот платви din лъвитря — din a inimei токшенлъ е тоці іав Се вргеще двит гіса ввивляї штитвітор; честеа de Ба е фъквт и сіщцітінте таї паінте де івеаль, de ners-Маї nainте de че фанта лар веdі ла черкътор. — Кънd ъптей oan eni войеа фъкъторили кълкаръ, прілеж ла Мъреціеа тай пай те ї дитіпасе ку пъкат. і, тотуші Mai nainte de че фуръл дий дитинde nearpa геаръ, рв къ щіа De пъкатил лъкотіеї канетаї са ші стрікат. Xc кon-Маї паінте de че фанта первшінеї се плінеще, Сволетва пріп реоа почть тъпъ две с'а щі стіптіт. паштере Maĭ nainre de че Каin svicece тіръпеще, сы дп Пре Абел — ал ляї лъвитов піста 'л а ші фост къпіт. Піста е пріхапъ таре пії веках, атъръчівне, s ei, mi De вечіпа порочіре ші аплакс преа рекчос isi Ioann De стръіна реоъ старе, серъчіе, стрікъчівне, еілор деї Nimensi фър' пътаї сіеш вреа че з въп, дпалт, фрвтос. , ka de-Еак' ачі 'п пяціпе ворье а целосяляї дескріре, Dеспре ачеаста Съпт-Атвросів пе кавжить пре фрвтос: a smel. "Чел фър' гей де къштігаташі се сесфатъ порочіе; cas An-Dap целосвя ва съ кръпе къпd е алувя порокос." nainte-Чел фър' гез іввеще ревл, ші де въп пічі вреа съ щіе: eğ aszit Daр пелосыл пыта 1847 ыще ге'ппъкат. ape decay dec

<sup>\*)</sup> Піста. Еў — Brasov аці літераці ай пострі дптревинцаю ковынол ачеста: insidiea. De се веdе преа латіп, пу ар фі doapъ mai sine ал

si Innoventis	pivipe mai ce! Botam	ітвле mong! Bin' mai
ri. La annal Bela Ri Collegia		ептръ ка съ Провіпціалі- ті ещі ров,
No second supply the second supply to the second su		сте тот ло- din Amepi- sine жі фі е сълбатік,
patel Koma-	898 T89- 	та та та алес, сжите соко- пцел те се
a fi kreutini,	38rpz-	скъ ла соа- літіт Dea- пъ кърџе ві- Nъ'ї сат ъп-
despretsinds	PAZE AE	е 50 de обе Сътръ- Ел ші ав-
iele biseriuï l Komanilore	MI UE	веаквляї!. квторївл ляї псетпат? Пріпцde
mute Pseudo- multi Unruri	inop 4E	taine Шваве). taine Шваве). taine Шваве). taine Шваве).
n Reratsl Un- ma fakandsse	RHAPE ALL HE	I.
пе Еріsкоръ, misterii întro	к к8 к к8	БЕ (дитръ эторыт).



### **Challenges • OCR and Transliteration**

ipe vei ve цел. 2, 24. Бъторіі de токшеа.гъ е тоці іаў честеа de веаль, de nersьтор. ьлкаръ, прілеж ла і, тотвші агра геаръ, грв къ щіа ікат. Хс копmeille, mi cmintit. паштере сы дл фост къщт. risne, s ei, mi oc Barrel der isi Ioann isne, еілор льї палт, фрвтос. , ka deскріре, ъ пре фрвтос: a asmel. тъ порочие: cas Anпорокос." nainteвреа съ щіе; ппъкат. es aszit de greens en 3/ ape deітераці аї поcay dec : insidiea. De neжdie, mai sine a.1 e. He

Vintele de la Denzre. Neroțiațienele lei Innogenție III ks Ioannitis İmperatorish Andrews "Freroris katre Regele Bela intal Like Katalanda Katalan 1234 îndreptà Рара Гreгогів IX katro Unrariei srmatórea skrisóre. Regelsi Unrariei. "Prekeme ame înțelese, în Episkopatel Komanilors se afla nimte popli, kari se nemeskë Romani, mi ka tote ka ka namele se sokoteska a fi kremtini, însa sepë o kredința avandă diverse riteri mi datine faks fapte kontrare auestsi neme. Kaui despretsinds biserika Romana na priimeskä misteriele biseriqii dela venerabilal nostra frate, Episkopal Komanilora kare e Dieuesanzl lokzlvi, ui dela nimte Psezdo-Episkopi, karii tinë ritel Franilorë, mi melti Unreri 1850 din Rerats Unmi Germani, mi alti o rariei ka sa loksésk / Vienna ama fakandsse sně pople ke Romann, despreteské pe Episkopě, mi grijmeskë dela dinmii grezisele misterii între

PITIPE MAI птвле шопц! ... Bin' maï CE! ROIAM gi amundoi ептръ ка съ Провінціаліті ещі ров, сте тот лоdin Amepisine al oi е сълбатік, та таї алес, сжитё сокопцвл тев се AZBE TIскъ ла соаlimir.. Dea-38rpzпв кврџе ві-Na'i car an-TOT EPA e 50 de oše ... Сътръ-PASE AE Ел ші аввеаквляї!. TEN MAI квторівл льї incemnar?.. MI UE Пріпц...de 24 CZ taine III Base ). raine IIIBa-INOP 4E ra'i! mmm RHAPE CALL 4E БЕ (мптръ интр8 эторыт).



### **Challenges • OCR and Transliteration**

ipe vei ve Бъторіі de е тоці іаў честеа de de nersпрілеж ла ї, тотвші грв къ щіа Хс коппаштере і съї дп s ei, mi isi Ioann еілор льї , ka dea asmei. сав дпnainteape decay dec neжdie, mai sine a.1 e. He

цел. 2, 24. nnoventië токшеалъ ni al Bol-La annul івеаль, Regele ікат. ineille, il Komami cminrir. Romani, фост къпіт. kreutini, nsne, ui datine oc Fatoal der isne, pretsinds палт, фрвтос. biseriqii скріре, ъ пре фрвтос: manilors тъ порочие: Psevdoпорокос." вреа съ щіе; ti Unreri ппъкат. rated Unde excours as akandsse ітераці аї по-Episkops, : insidiea. De

erii întro

ДЕКЯТЕ ОРІ ЗІОВ АМ ФЕКВТ ПЛАНВРІ ДЕФЕРІЧІРЕ МАЇ 418 AATE AEKRT BICSPINE HENE MAT CHEPIOACE! BOMM СЕ Ампреви лекрері испотрівіте ші се лифіінцед пефінца. Тікилосвя! Амі фигидвим пли-DE LES DE LE TEATOS VEN CRIMEZTOP AN ASMIT; лор віеції, ші оферічіре неклінтіти пе валврі тврв врате. Ли че в ніверс амирітор се тряндив тіиереціле меле! кв че феце стрилвчітогре звгривим тоате лекреріле деп Ампрежер! Тот ера пентря міне фрямос; тот Амі адвярль раде де BECENIE; TOT MI CE AMPZHIMA CERT KINEN 4EN MAÏ амирітор; натвра литреги лмі зямью ші пе тоати зіла скімбяндвіші подолел, ми фин си віт ка сянт мврітор, Ан сянва пахчерілор че CEAZHU III 1835 ESKEPIE MAHAPA ME TAIMEAM AH MIXAOKEA MEDELLEASI AEKAW 48 CANTACIA AMISH T Bucharest INE EPA TIENTOS MIHE SH BILTOP COAPTE AFTIZPTAT. A CEMEH'S KY

чітвле шопц! e.. Bin' maï egi amundoi пентръ ка съ Провінціаліmi emi pos, сте тот лоdin Amepisine al oi е сълбатік, та таї алес, сжить сокотивл тей се скъ ла соаiimir.. Deaпв кврџе ві-Ns'i car sne 50 de oše ... Сътръ-Ел ші аввеакълъї!. івторівл лві incemnar?.. Пріпц...de атпе Шваве). aine IIIBa-81! mmm

БЕ (дитръ

торыт).



#### Challenges · OCR and Transliteration

ipe vei ve ITIPE MAY цел. 2, 24. nnoventis Бъторіі de E! BOIM токшен.гъ ii al Bolе тоці іаў лифіінчестеа de La annsi в веаль, de nersьтор. прілеж ла ь. каръ, і, тотуші агра геаръ, nasketel грв къ щіа ікат. PI 78p-Хс копineille. il Komami cmintit. ZBK Tiпаштере Romani, 38rpzсъй дл фост къщт. kreutini, s ei, mi nsne, OT EPA ui datine loc Larragh ısı Ioann IZE AE isne, pretsinds еілор льї спалт, фрвтос. EN MAI biseriqii , ka deскріре, ъ пре фрвтос: MI UE a asmei. manilor тъ порочіе: сав дп-4 6 CZ порокос." Psezdonainteвреа съ щіе; ti Unreri 100 4E пиъкат. eğ aszit rated Unидра BE SKROTE BUSIN ape deakandsse ітераці аї по-W 48 cay dec dniskopo, итр8 : insidiea. De neжdie, mai sine a.1 erii între e. He

костівт de връжітор. Църапії спъріеці фвгв ип кръчтъ стрігьиd:

Еака фракъ тъ! сака фракъ.

10Р - (кътръ църапі) № Фоціці оатепі въпі, къ доаръ пъ'с Скараоцкі!. Къпд вреа дракъ съ тъпчеаскъ оатепії, ел се фаче цъбекътор сав Ісправнік ... Ав фацітв!.. с'ав фас!... Аша фага тоці чеї че ть дитьлиескв. — Еать-ть'с дисфжршіт ші шарлатап ... De патра ляні de ziлe аквт, алерг din сат ли сат ші daš репрегентації акроватіче.. Ев Іорга Оътіап!... Ші че фракт ера съ фак? — Къпф m'o гъсіт тошви тев ди влестетатвл чел de чеасорпік, ші къпd m'o dar пе тъпа котісарі лор ера съ певкпееск дисъ m'am ствичіт воїпічеще din гіареле че тъ пріпсесерь de гът; тіат луат кълкъіле de-а втере ші m'am dsc. m'am dsc, m'am dsc пъп' че m'am възът департе де капіталіе, ші де Іцік, ші de Adџie, ші maĭ алес de стахіа чеа de Гахіцъ. Атъпче ат литрат дитр'о кръчить ка съ тъ odixneck!. Къпд дл кръчтъ, чесъ въд?... пе хер фон Клаїне Шва-Ee!. sn neamy че къпос Cadarspa ші каре се dsче

че.. De odaтъ жті тръспеше

Мъ апропії de neamusa meš шії

zік: "Мъї шопихле!. decкideні вре-

кап о idee de челе корп

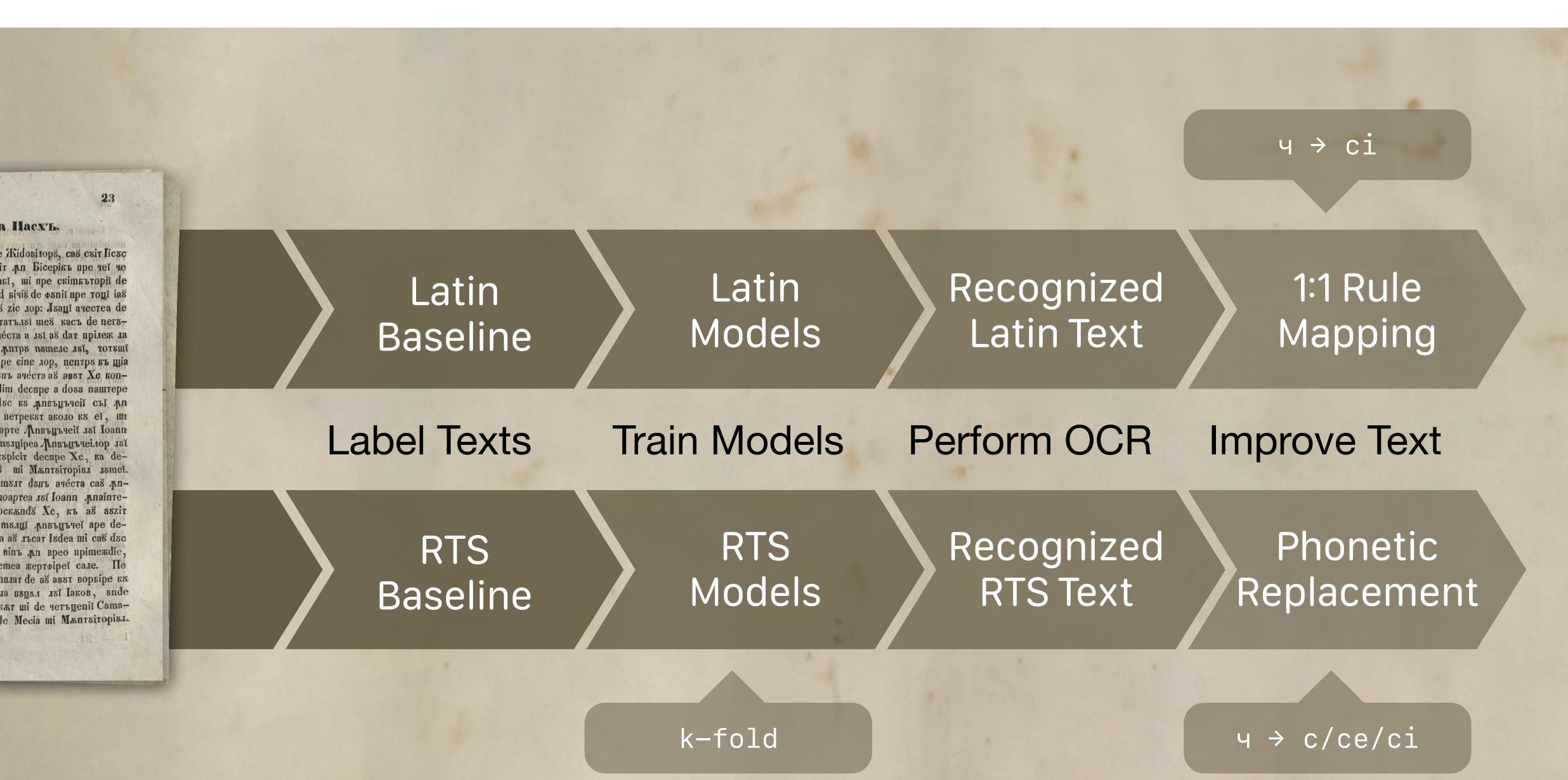
рокъл ди Гаші, пеферічітьле шопц! къ-мі mspi de фоате.. Bin' mai віне къ mine, ші хаїdegї amъndoї съ фачет о асочіаціе пентря ка съ спеквлът певіповъціа Провінціалілор.. Тъ тей фаче къті ещі ров, ші ей те-оїй аръта песте тот локъл ка пе ъп сълватік din Атеріка.. Жі віdea кът de віne жі фі пріїтіт свв пътеле de сълбатік, дитр'о епохъ ка ачеаста таї алес, snde oamenii чізілісаці сжить сокотіці de певяпі! " — Neamusa meš ce віта ла тіпе ка о вроаскъ ла соаре, ші дпсфършіт аў пріітіт.. Deaтъпчі, славъ Domnълъї, пъ кърџе віпе лаптеле 'п пасат!. № ї сат впde съ пъ къщігът къте 50 de обе ші къте 10 пъї de гъіпъ... Съ тръеаскъ шарлатапіствл!. Ел ші аврыл сжить domnitopii веакылы!. Чine щie daкъ прin ажьторівл льі пъ воїв ацівние ви от дисетнат?.. Поате съ тъ фак вр'яп Пріпц...de Cadarspa. (възънд пе Клаїне Швабе). Еака ші бароп фоп Клаїпе Шваве!.. Бре!. посоторътв'і!

CHENA III.

my avalory

ка съ dee peпреzenтації ціmnастіа SIPF8, КЛАІНЕ-ШВАБЕ (Аптръ піп фреапта фоарте посоторыт).

10Pr - Че аї тъї пеатикле, de



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

$$BCER = \sum_{i=1}^{W} \left( \frac{S_i + D_i + I_i}{N_i} \right) / W$$

CER Character Error Rate

BCER Bag of Characters Error Rate

S Substitution

D Deletion

I Insertion

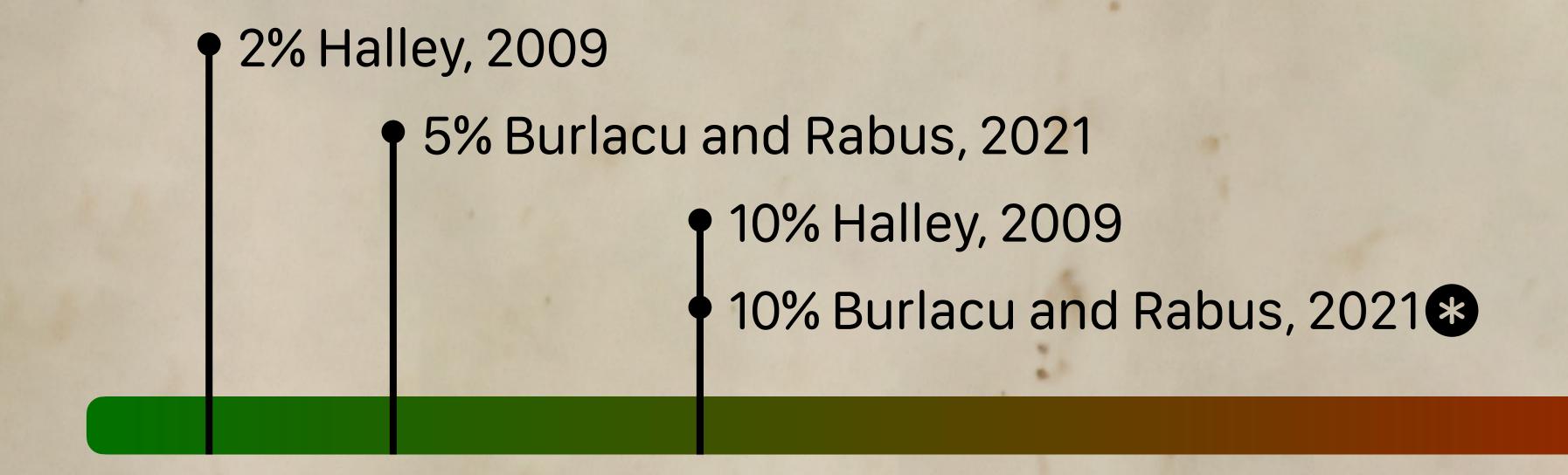
N Number of Characters (baseline)

W Number of Words

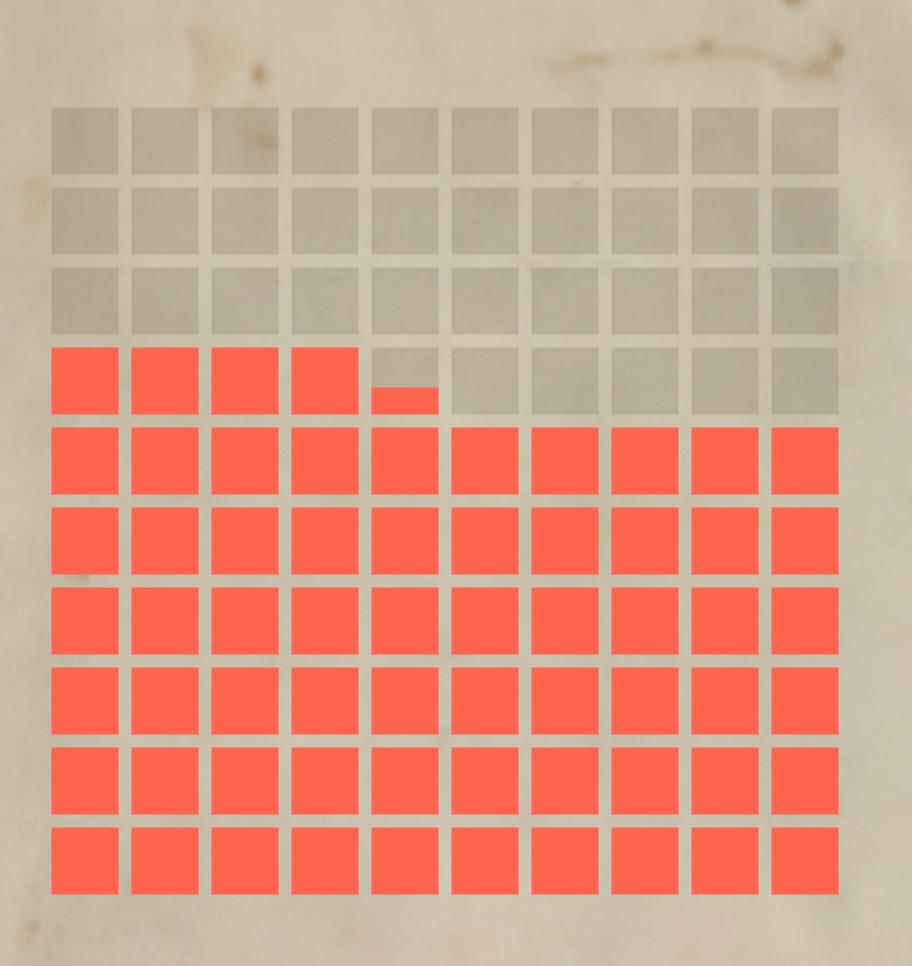
## Experiments

0% CER

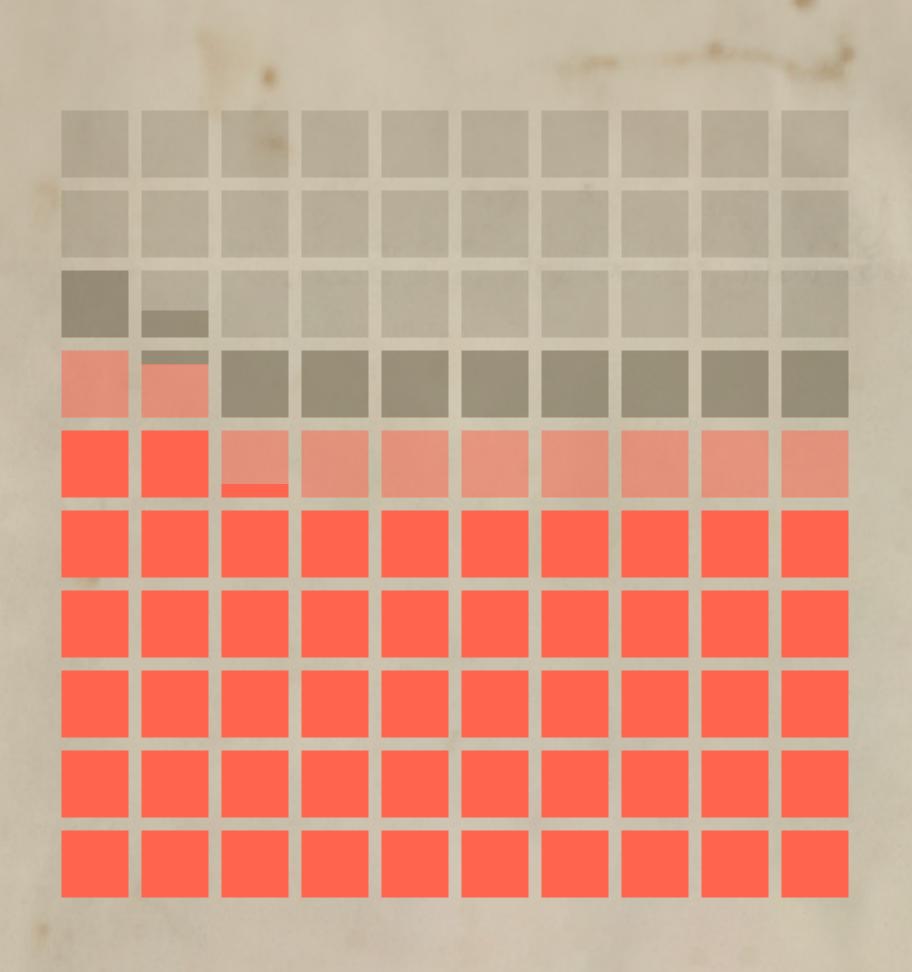
$$CER = \frac{S + D + I}{N}$$



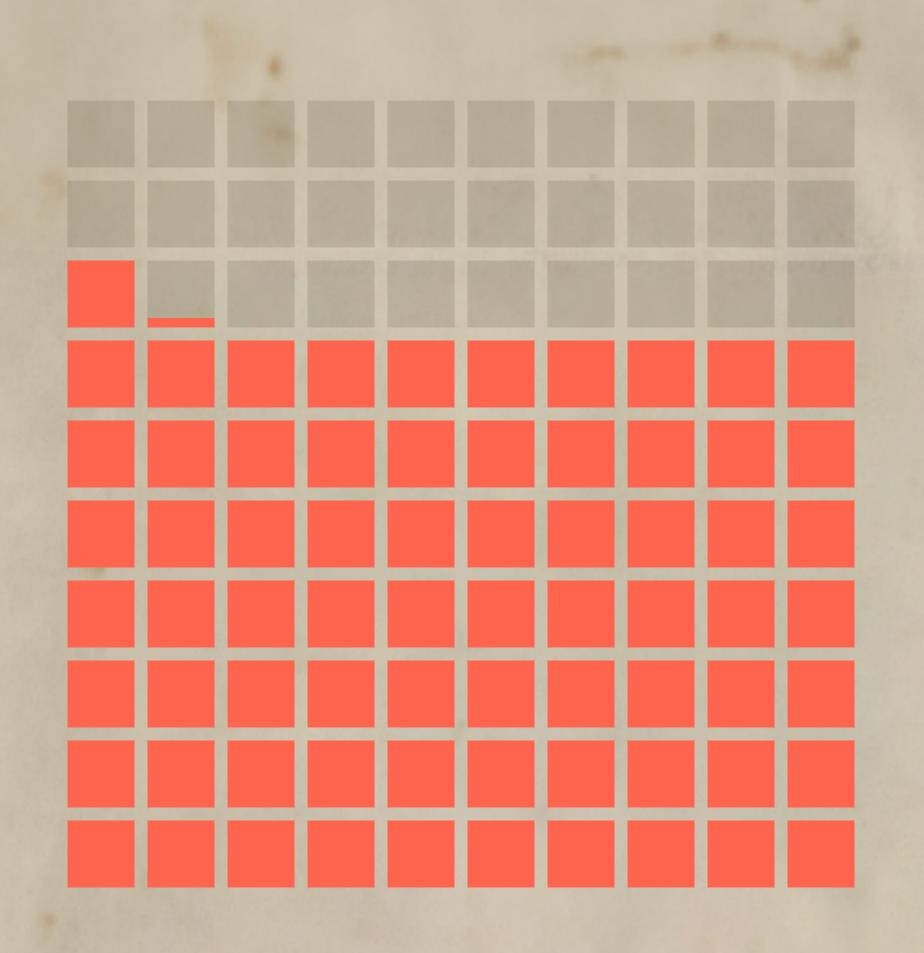
- ► 30 pages
- ► 24148 characters (64.4% Cyrillic)



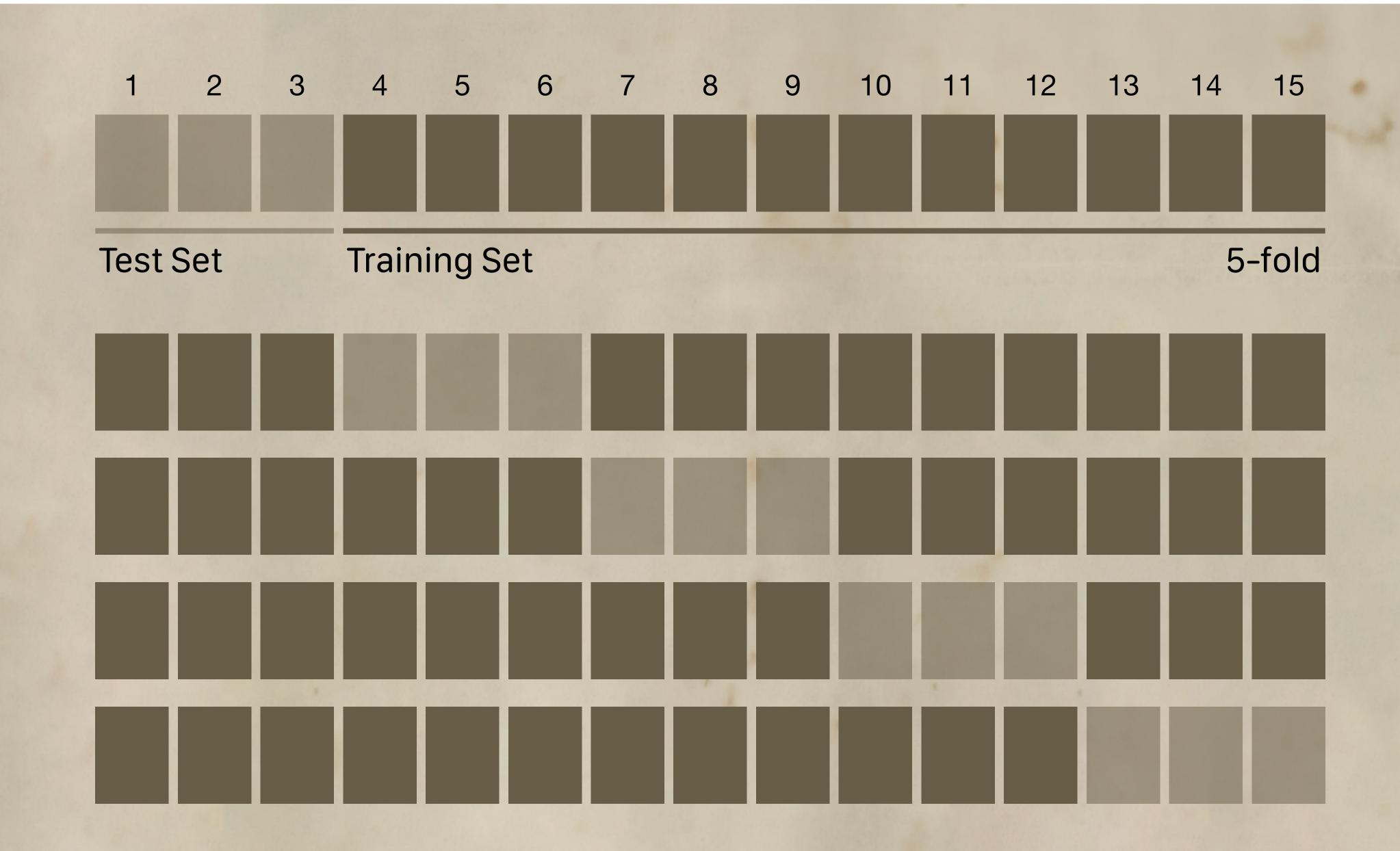
- ► 30 pages
- ► 24148 characters (64.4% Cyrillic)
- ► Page average
- ► 61.8% ± 9.6% Cyrillic

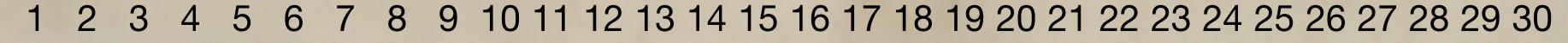


- ► 30 pages
- ► 24148 characters (64.4% Cyrillic)
- Page average
- ► 61.8% ± 9.6% Cyrillic
- ► Test page
- ► 745 characters (71.14% Cyrillic)









3-Fold Test Set

**◀** Best results

10-Fold Test Set

29-Fold Test Set

Е віпе іввіділор, съ пв свферіт даръ таї твлт житре пої пре dsxsл чел рей ал люшії, чела че вреа пвтаї а се трвфі ші а се търі! Съ къйтьт а дитокті фачереа ші пезвіпца, лекрареа ші сеферіпца поастръ аша, ка пріп ачелеа сіптър Dътпегев съ се лавде, ші пятеле ляї съ се преа търеаскъ! Ші ка съ віецвіт неквртат дл втілінцъ, аша воіт поі а авеа пе'пчетат лиаттеа окілор пъкътошіеа ші пітічіеа поастръ, ші пічі къпd а zesiта търітыл есетплы, каре Мънтыторыл чел втіліт 'л а dar поаъ. De твлте орі воіт а пе adъче aminre de въвіптеле че вісеріка 'nainre de вътева гіле пё а стрігат: "Квиеть о отвле, къ цержпъ ші чепвшъ ещї, ші іаръ цержиъ ші чепвить те веї фаче." Ка вашешил чел втіліг воіт de тилте орї таї алес ди тіппвл с. аціви (пост) пліні de къіпцъ а пе вате пентвл гръind: "Dsmnezesae фі Andspar mie пъкътосвавї!" Дл таїна покънцеї воіт пої a ne

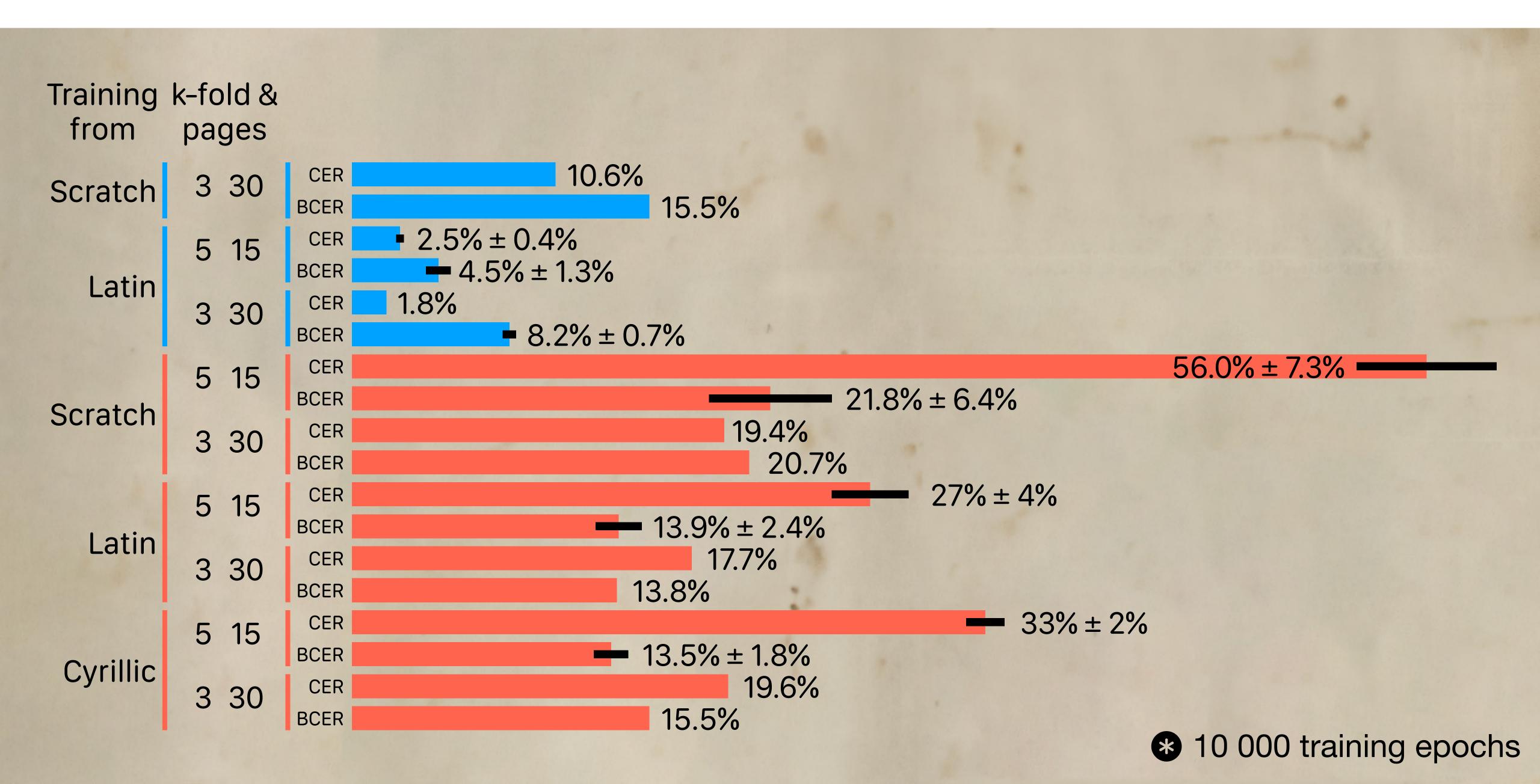
- 12 -

втъ на вапъ плъчере да фантеле на капсва inimel втіліте ле фаче. Ел ле челе та паціп натипьпітоаре— е съпт пліпе de теріт ла Damnezes. нъ на атът та търіт се ва дпив на кът та адъпн с'а дпріп латеа ачелста щі с'а дппітічіт. втіліт, щі 'п ачелсть щі 'п чел

р, съ на свферім даръ таї твят ы чел рей ал лишії, чела че вреа и а се търі! Съ къйтът а дитокті а, лекрарен ші сеферіпца поастръ t cinrep Demnezeš съ се лавde, ші реа търеаскъ! Ші ка съ віецзіт цъ, ama воіт nol a авеа ne'nчетат чътошіеа ші nimiчiea noacтръ, щі гърітви есешпив, каре Мънтвіторви поаъ. De твлте орг воин а пе інтеле че вісеріка 'паінте de къте-"Квиетъ о отвле, къ цермпъ іарь цержив піі чепвить те веї чел втіліт воіт de твлте орї г. аџїви (пост) пліні de къінцъ ьind: "Demnezesae фі Andspar п таїна покъінцеї воіт пої а не ей втіліте фъръ ретрацере щі сфжринт кв впіліть іпіть а пе a dparocreĭ, вршънd a ziче ка сжит вредия, на съ дити скит

#### 0

#### Experiments · Results





#### Experiments · Results

#### RECOGNITION FAILURE TYPOLOGY

- ► Graphic similarity between letters, accented letters, or numbers resembling them visually.
- Lack of previous training. E.g. Greek symbols present in the dataset.
- ► Errorsin transliterating certain double consonants like II (double L).

#### Conclusions

- ➤ We addressed transliteration of Romanian texts from the 19<sup>th</sup> century.
- ➤ We proposed a Tesseract-based solution on two targets: Latin and RTS.
- Latin initial results are good, but phonetically interpreting the text is challenging.
- ► RTS results indicate the need for a richer dataset.
- ► Further work considers these aspects.
- ▶ Dataset is available on Kaggle as: "19th-Century Romanian Transitional Script"

Challenges and Solutions in

# Transliterating 19th-Century Romanian Texts from the Transitional to the Latin Script

Thank you