

Challenges and Solutions in

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

Marc Frîncu, Simina Frîncu and Marius E. Penteliuc

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



Context

18th century
43 letters



* not a standard keyboard layout



Context

18th century
43 letters



late 18th century
38 letters



* not a standard keyboard layout



Context

18th century
43 letters



late 18th century
38 letters



19th century
30 letters



* not a standard keyboard layout



Context



* RTS = Romanian Transitional Script



Дѣла Пасхѣ

Фіинд апропе Пашіле Жідовіторѣ, саѣ сѣит Іисус
дн Іерусалим, ші аѣ гѣсѣт дн Бісерікѣ пре чеї че
віндеа ої, пѣї ші порѣмѣї, ші пре скітѣторіи де
ѣанї шежнд, ші фѣкжнд вічіѣ де фѣнії пре тоѣ іаѣ
скосѣ дн Бісерікѣ, ші аѣ зіс лор: Лѣаѣ ачестеа де
аїчеа, ші пѣ фачеѣї каса татѣлї меѣ касѣ де пегѣ-
ѣторіе. Де ші фапта ачѣста а лѣї аѣ дат прілеж ла
мѣлїї, ка сѣ крѣадѣ днтрѣ пѣмеле лѣї, тотѣшї
Іисус пѣ се днкрѣдеа пре сїне лор, пѣнтрѣ кѣ шїа
пре тоѣ. Іоанн к. 2. Дѣпѣ ачѣста аѣ авѣт Хс кон-
ворѣре таїнікѣ кѣ Нікодїм деспре а доѣа паштере
ші мѣпѣре, ші саѣ дѣс кѣ днвѣѣѣей сѣї дн
пѣмѣпѣл Іѣдеї, ші аѣ пѣтрекѣт аколо кѣ еї, ші
аѣ вотеѣат. — Маї де парте днвѣѣѣей лѣї Іоанн
аѣ вестїт лѣї деспре днмѣлїреа днвѣѣѣейлор лѣї
Хс, еар Іоанн аѣ мѣртѣрісїт деспре Хс, ка де-
спре фїїл лѣї Дѣмнеѣѣ ші Мѣпѣїторїл лѣмѣї.
— Іоанн. к. 3. — Нѣ мѣлїт дѣпѣ ачѣста саѣ дн-
тѣмѣлат прїсоареа ші тоартеа лѣї Іоанн днпайте-
мерѣторїлї, ші кѣпоскжндѣ Хс, кѣ аѣ азїт
фарїсеїї, кѣм кѣ ел маї мѣлїї днвѣѣѣей аре де-
кѣт Іоанн: пѣнтрѣ ачѣеа аѣ лѣсат Іѣдеа ші саѣ дѣс
дн Галїлеа, ка сѣ пѣ віпѣ дн врео прїмеждіе,
кѣчї днкѣ пѣ венїсе времеа жѣртѣїреї сале. Пе
каїеа ачѣста і саѣ днтѣмѣлат де аѣ авѣт ворѣре кѣ
мѣїреа Сатарїеанкѣ ла пѣѣл лѣї Іаков, ѣнде
атѣт де ачѣстѣ мѣїреа кѣл ші де днѣторїї Сата-

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



Related Work

► Same sound, same graphic

• A E I K M O T

■ A E I K M O T



Related Work

► Same sound, different graphic

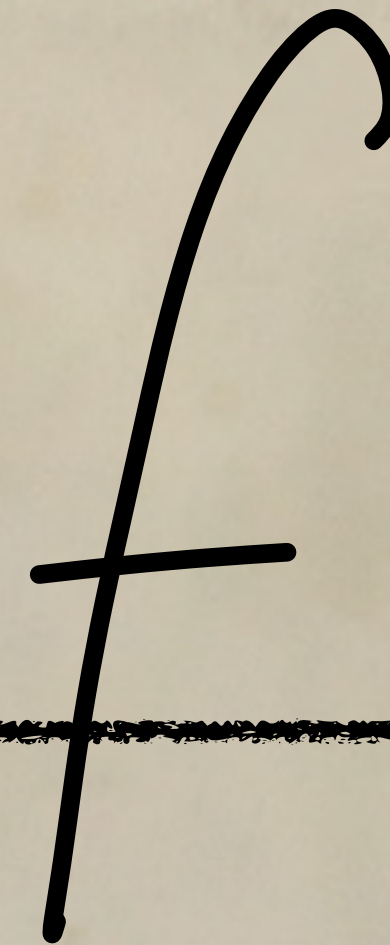
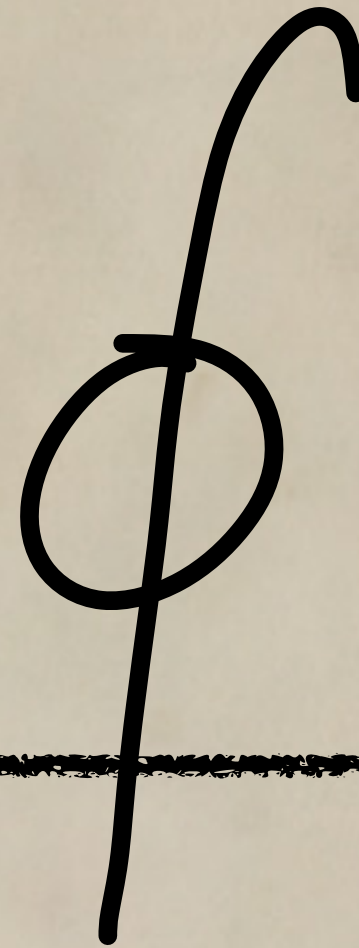
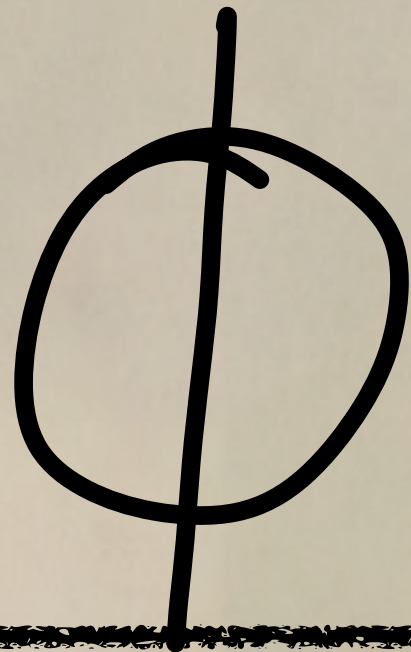
● V S N R H

■ B C H P X



Related Work

- Cyrillic small F to Latin small F





Related Work

► Cyrillic small L to Latin small L



л

Л

л

l





Related Work

- Cyrillic small R to Latin small R

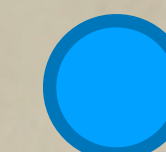


Р

р

P

r





Related Work

- Cyrillic small J to Latin small J



ж

je

j

j





Related Work

- Cyrillic small D to Latin small D



А

д

d





Related Work

	Dataset Size	Software	Result
Boian et al., 2014	Unavailable	ABBY FineReader (proprietary)	63–95.4% correct words
Cojocaru et al., 2016	Unavailable	ABBY FineReader (proprietary)	96% Accuracy
Demidova and Burteva, 2017	Unavailable	ABBY FineReader (proprietary)	99% Accuracy
Burlacu and Rabus, 2021	30 900 words	Transkribus (proprietary)	10% CER (lower is better)
Our Study	30 pages	Tesseract (open source)	



Challenges • Processing

	Effort	Time	Space
Document Scanning	Manual	Long	100 KB – 10 MB / scan
Training Set Labeling	Manual	30 min / page	3 – 10 KB / file
Validation Set Labeling	Manual	30 min / page	3 – 10 KB / file
Testing Set Labeling	Manual	30 min / page	3 – 10 KB / file
ML Model Training	Automated	17 – 2200 sec / training fold	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.

! Challenges • OCR and Transliteration

Бисерикъ пре чеї че
пре скітѣторіі де
е фхніі пре тоці іаѣ
р: Лхаці ачестеа де
меѣ касъ де перѣ-
хі аѣ дат прілеж ла
хмеле лѣї, тотѣші
лор, пентрѣ къ шіа
а аѣ авѣт Хс коп-
ре а доѣа паштере
пвѣцѣчей сѣї лп
аколо кѣ еї, ші
вѣцѣчей лѣї Іоанн
Лпвѣцѣчелор лѣї
спре Хс, ка де-
пѣторіѣл лѣмеї.
ѣ ачѣста саѣ лп-
ї Іоанн лпайнте-
Хс, кѣ
вѣцѣчей аре де-
Іѣдеа
врео прімеждіе,
фіреї сале. Пе

1860

! Sibiu

Андел. 2, 24.
а інімеї токмеалѣ
ѣптѣтор;
пайнте де івеалѣ,
і ла черкѣтор. —
торѣлѣ кѣларѣ,
е кѣ пѣкат.
птінде неагра геарѣ,
сѣа ші стрікат.
неї се плінеце,
лѣ дѣс сѣа ші смінтіт.
гірѣнеце,
ста ѣа ші фост кѣніт.
лѣ, амѣрѣчїне,
преа реѣчос
е, стрікѣчїне, —
не ѣ вѣн, лпалт, фрѣмос.
осѣлѣї дескрїре,
не кѣвѣнтѣ пре фрѣмос:
се десфатѣ порочїе;
ѣ алѣлѣ порокас.“
де вѣн пічї вреа сѣ шїе;
рѣце не пѣкат.

1847

! Braşov

Onitatea lui Iesie pro-
tionele lui Innocentiis
omninoſorſi ui ad Bal-
lorſi nostri. La annul
X katre Bela Regele
Bela intziſ naskotul
n Episkopatul Roma-
ni se nămeskă Romani,
koteskă a fi kreutini,
verse rituri ui datine
ne. Kăci despretindă
misteriele bisericii
Episkopat Komaniilorſi
de la niute Pseudo-
lorſi, ui multſi Ungari
trekă din
i, ui amă făkandose
etăeskoſe pe Episkopoſi,
rezisele misterii întră

1850

! Vienna

ри де ферічїре маї
сперіоасе! котам
е ші сѣ лїфїїн-
фѣгѣдѣтам пѣз-
мѣѣтор ал лѣмії;
іжлокѣл кінѣрі-
ѣ пѣвалѣрітѣр-
се трѣндѣѣтї-
ѣчїтоаре зѣгрѣ-
жѣр! Тот іра
зѣѣрлѣ рѣѣе де
кінѣл чѣл маї
зѣмѣѣ ші пѣ
а, мѣ фѣчѣ сѣ
пѣзѣерілор чѣ
ѣрїе
лѣї лѣкѣлѣ чѣ
іне іра пінтрѣ
лѣсеменѣ кѣ

1835

! Bucharest

ші, неферічїтѣле шонѣ!
де фоаме... Bin' mai
e, ші хайдеѣ амѣндої
асочїаѣе пентрѣ ка сѣ
евїновѣѣїа Провїнѣїалї-
еї фаче кѣмі еѣї рѣѣ,
ѣ арѣта песте тот лѣ-
п сѣлѣатїк дїн Амерї-
їеа кѣт де вїне жї фї
пѣмелѣ де сѣлѣатїк,
ѣ ка ачѣаста маїалес,
ї чїзїлісѣїї сѣпѣтѣ соко-
їні!“ — Neamul meſe
е ка о вѣоаскѣ ла соа-
ѣршїт аѣ прїїмїт... Dea-
Domnălă, пѣ кѣрѣ вї-
п пасат!. Năї сат ѣп-
ѣшїгѣм кѣте 50 де оѣе
пѣї де гѣпѣ... Сѣ трѣ-
атанїсѣмѣл!. Ел шї аѣ-
domnїторїї веакѣлѣї!.
акѣ прїп ажѣторїѣл лѣї
пѣ ѣп ом лпсѣмнат?...
ѣ фак вѣрѣп Прїпѣ... де
вѣзѣнд не Клаїне Шѣѣѣ).
роп фон Клаїне Шѣѣ-
посоторѣтѣї!
CHENA III.
LAINE-Шѣѣѣ
фоарте посоторѣтѣ).

1852

! Iaşi



Challenges • OCR and Transliteration

... vindea oi, năi și porumbi, și pre scîmbătorii de
cani șezând, și făcând biric de fani pre totii iaș
skocș din Biserikă, și aș zic lor: Lăsați aчестea de
aіchea, și nă faceți каса татълăи меѣ касъ de negъ-
цъторіе. De și факта ачеста а лăи аș dat прілеж ла
мăлăи, ка съ креадъ лăнтръ нăмеле лăи, тотъші
lăscă nă се лăкредеа пре сіне лор, пентрă къ шіа
пре тоцї. Ioann к. 2. Дъпъ ачеста аș авѣт Хс коп-
ворбіре таінікъ къ Nikodim деспре а доха наштере
ші мăнтăіре, ші саș дăс къ лăвѣцъчей сѣи лă
пъмăнтăл lădeи, ші аș петрекѣт аколо къ еї, ші
аș ботезат. — Маї de пәрте Лăвѣцъчей лăи Ioann
аș вестіт лăи деспре лăмăлăіреа Лăвѣцъчеілор лăи
Хс, еар Ioann аș мъртърісіт деспре Хс, ка de-
спре фііхл лăи Dămnezeș și Мăнтăіторіхл лăмеі.
— Ioann. к. 3. — Nă мăлт дъпъ ачеста саș лă-
тъмлат пріпсоареа ші тоартеа лăи Ioann лăаінте-
мергъторіхлăи, ші кѣпоскѣдъ Хс, къ аș аșзіт
фарісеіи, кѣм къ лăмăлăи лăвѣцъчей are de-
кѣт Ioann: пентрă лăмăлăи лăвѣцъчей аș гѣат lădeа ші саș дăс
лă Галілеа, ка съ нă вѣнъ лă вreo прімеждіе,
къчї лăкъ nă venice времеа жертфїреї сале. Пе

... еа лăи лăи лăи
... индел. 2, 24.
... еї токмеалъ
... р;
... de івеалъ,
... кѣтор. —
... кѣ
... ат.
... еагра гѣаръ,
... рікат.
... іпеліе,
... а ші сміптіт.
... ші фост кѣніт.
... чїне
... нос
... іне, —
... лăалт, фрѣмос.
... скріре,
... ѣ пре фрѣмос:
... тѣ порочіе;
... порокоc.“
... і вреа съ шіе;
... нѣкат.

... și preste pro-
... lăi Innocențîș
... cș și al Băl-
... tri. La anul
... Bela Regele
... opatșl Koma-
... eskș Romani,
... a fi kreutini,
... rī ui datine
... despretșindș
... iele bisericii
... l Komaniilorș
... iute Pseudo-
... mălti Ungri
... Regatșl Un-
... na făcândșse
... pe Episcopș,
... misterii întrș

... річїре маї
... се! котам
... лифїи-
... там плз-
... кїндрі-
... рїтѣр-
... азвѣ ті-
... зѣгрз-
... Тот іра
... разе де
... чел маї
... ші пе
... зчѣ сѣ
... ілор че
... кїдръ
... лаш че
... нїтрѣ
... ѣ къ

... ітѣле монц!
... Bin' maї
... dї amăndoi
... ентръ ка съ
... Провінціалі-
... ті едї ров,
... сте тот ло-
... din Ameri-
... bine și фї
... е сѣлватік,
... та маї алес,
... сѣптѣ соко-
... нцл меѣ се
... скъ ла соа-
... ііміт... Dea-
... нă кърде ві-
... Năi сат ѣн-
... e 50 de оѣе
... Съ трѣ-
... Ел ші аѣ-
... веакѣлăи!
... кѣторіхл лăи
... мсепнат?..
... Prinț... de
... іаіне Швабе).
... іаіне Шва-
... гѣї!
... I.
... АБЕ (лăнтръ
... торѣт).



1860



Sibiu

Text from verso



Challenges • OCR and Transliteration

пре чеі че
вѣторіі де
е тоді іаѣ
честеа де
де негѣ-
прілеж ла
і, тотѣші
грѣ къ шіа
Хс коп-
паштере
сѣі лп
ѣі, ші
іаі Іоанн
еілор лѣі
ка де-
л лѣмеі.
саѣ лп-
лпайнте-
аѣ аззіт
аре де-
саѣ дѣс
междіе,
е. Пе

лп лѣме ші чеі дп партеа лѣі лп каѣ-
тѣ пре дѣлсѣл. Лпдѣл. 2, 24.
Тот пѣкатѣл дп лѣхнтрѣ — дп а іпмеі токшеалѣ
Се ѣрзеде дѣпѣ зіса вѣнѣлѣі тѣнтѣтор;
Ба ѣ фѣкѣт лп сімпімінте маі пайнте де івеалѣ,
Маі пайнте де че фѣпта лар веіі ла черкѣтор. —
Кѣнд ѣптеіі оапепі воіеа фѣкѣторѣлѣі кѣларѣ,
Мѣреіеа маі пайнте і лптіпѣсе кѣ пѣкат.
Маі пайнте де че фѣрѣл лпш лптінде неагра геарѣ,
Де пѣкатѣл лѣкоміеі кѣпетѣі сѣ ші стрікат.
Маі пайнте де че фѣпта перѣшпнеі се плінеіе,
Сѣфлетѣл прпп реоа поітѣ тѣпѣ дѣс сѣ ші смптіт.
Маі пайнте де че Каіп ѣісесе тірѣнеіе,
Пре Авел — ал лѣі лѣхнтрѣ пісма 'л а ші фост кѣпйт.
Пісма ѣ пріханѣ таре ші пѣказ, атѣрѣчїне,
Де вечїна порочїре ші аплаѣс преа реѣчос
Де стрѣіпа реоѣ старе, серѣчїе стрікѣчїне, —
Нїменѣі фѣр' пѣтаі сіеш вреа че ѣ вѣп, лпалт, фѣрѣтос.
Еак' ачі 'п пѣіне ворѣе а целосѣлѣі дескрїре,
Деспре ачеаста Сѣпт-Амѣросїѣ не кѣвѣптѣ пре фѣрѣтос:
„Чел фѣр' зеѣ де кѣштїгаташї се десфатѣ порочїе;
Дар целосѣл ва сѣ кѣрѣе кѣнд ѣ алѣл порокас.“
Чел фѣр' зеѣ іѣвеіе реѣл, ші де вѣп пічі вреа сѣ шїе;
Дар целосѣл пѣтаі реѣл реѣше не пѣкат.



1847



Braşov

*) Пісма. Еѣ лпдѣл літерѣці аї по-
стрі лптрѣвїпдѣлѣ кѣвѣптѣл ачестѣ: іпвідїеа. Де
се веде преа латїп, пѣ ар фї доарѣ маі вїне ал

лп рѣсте рѣ-
ї Іппочентїѣ
ѣ шї ал Вѣл-
рі. La annѣл
Bela R...
їѣіш пѣскѣтѣл
патѣл Кѣма-
скѣ Ромѣні,
а фї кѣрѣтїні,
їрі шї датїне
деспрѣсїндѣ
їеле бїсерїчїї
л Кѣманїлѣрѣ
їїште Псеѣдо-
мѣлті Унгѣрі
л Регѣтѣл Ун-
ма фѣкѣндѣсе
ре Епїскѣпѣ,
мїстерїї іптрѣ

річіре маі
се! вѣтам
лїфїїн-
лпм пѣлѣ-
їїш пѣскѣтѣл
кїндѣрі-
їрі тѣр-
азвѣ ті-
зѣгрѣ-
Тот іра
раѣе де
чел маі
шї пѣ
зчѣ сѣ
їлор че
кїндѣрѣ
лш че
їптрѣ
ѣ кѣ

їтѣле шѣпѣ!
... Вїн' маі
її амѣндѣі
їптрѣ ка сѣ
Прѣвїпціалї-
мї еічі рѣв,
сте тот лѣ-
дп Амѣрі-
вїне жї фї
е сѣлѣватїк,
та маі алес,
сѣптѣ соко-
пѣл теѣ се
скѣ ла соа-
їїмїт... Деа-
пѣ кѣрѣе вї-
Нѣі сѣт ѣп-
е 50 де оѣе
... Сѣ трѣ-
Ел шї аз-
веакѣлѣі!
кѣторїѣл лѣі
лпсѣмнат?...
Прїпѣ... де
їїне Шѣѣе).
їїне Шѣѣ-
гѣі!
І.
АБЕ (лптрѣ
їптрѣт).

Stains



Challenges • OCR and Transliteration

Markings Ruptures



1850



Vienna



Challenges • OCR and Transliteration

Different Fonts



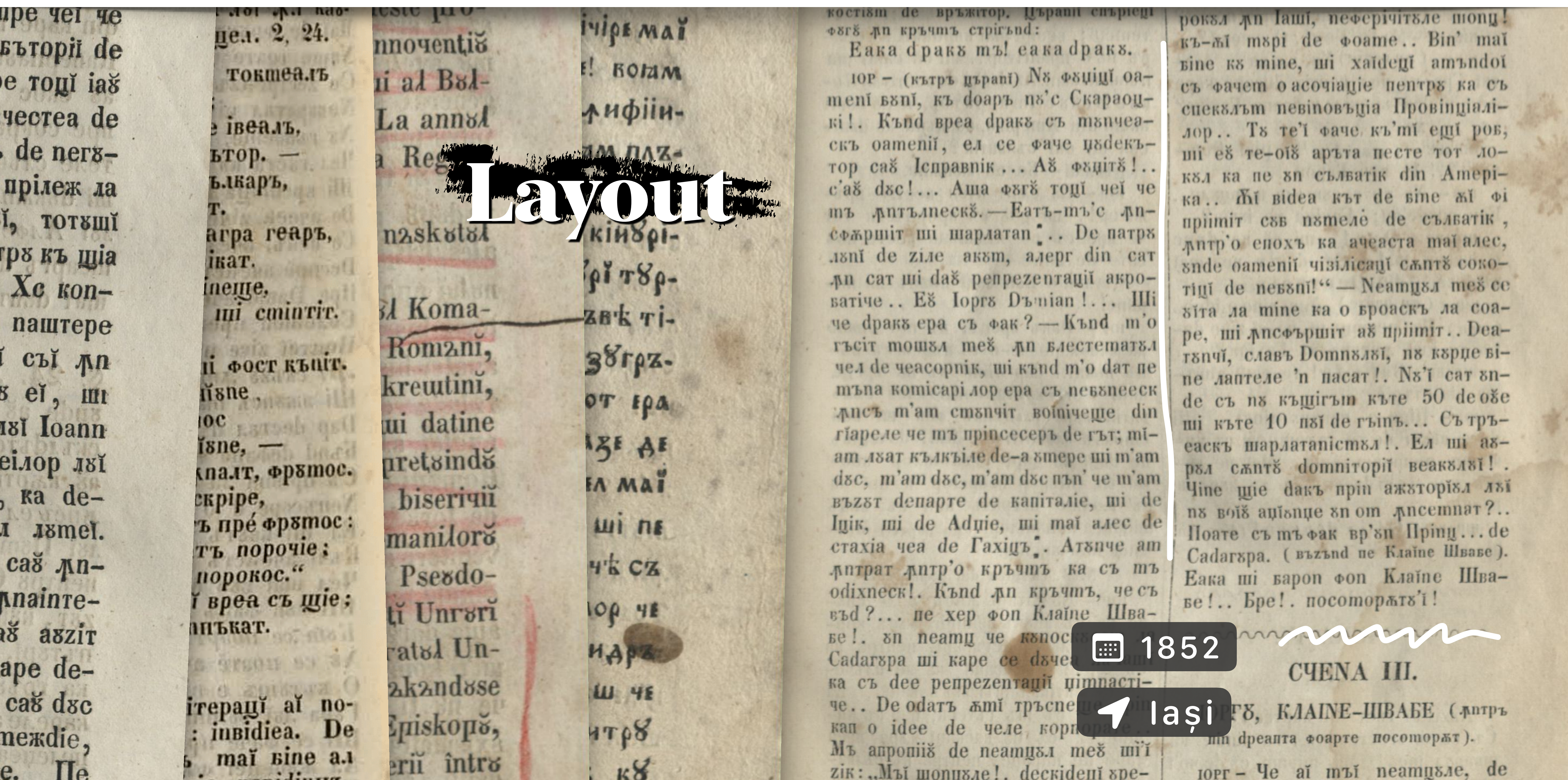
1835



Bucharest

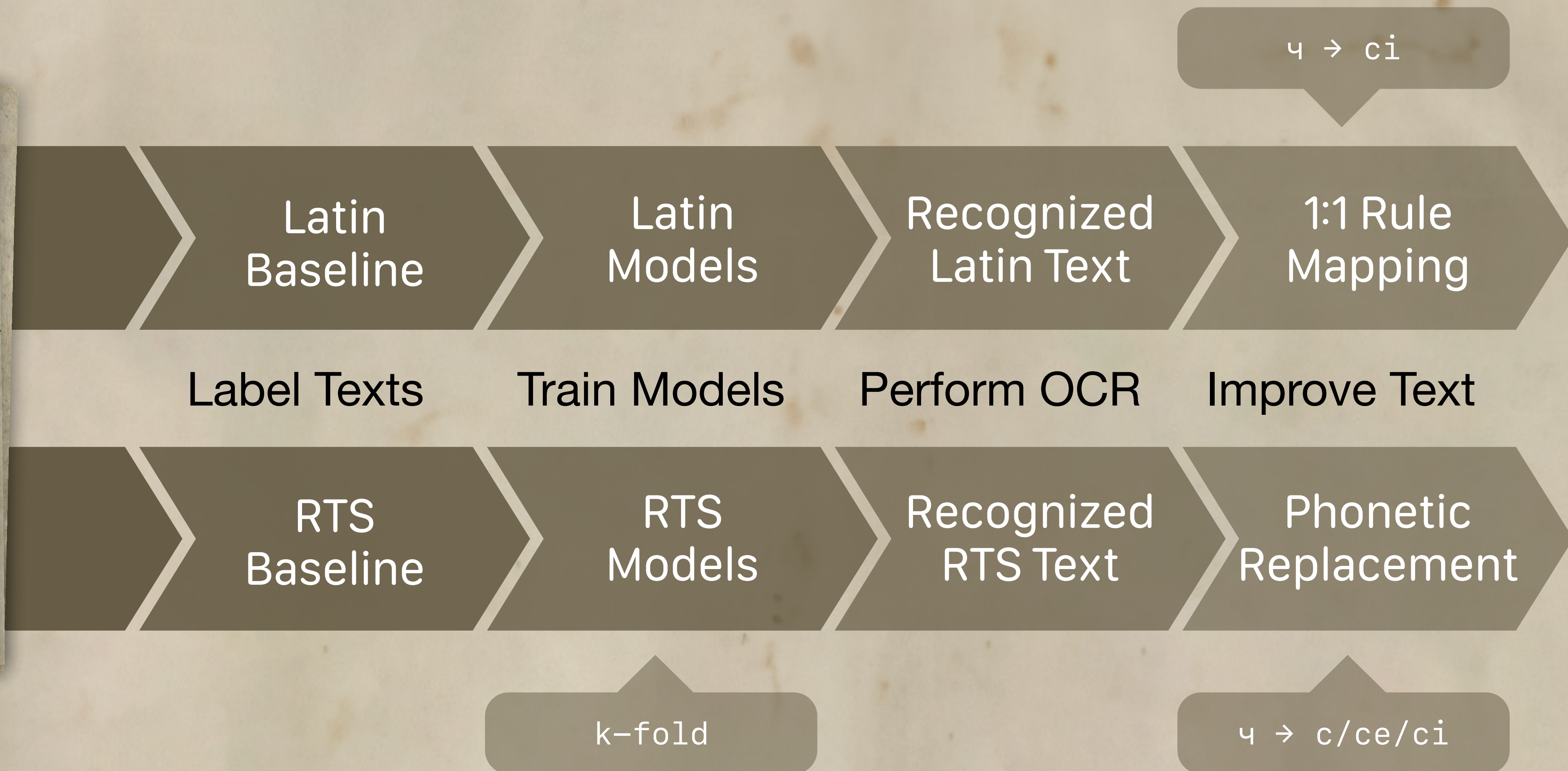
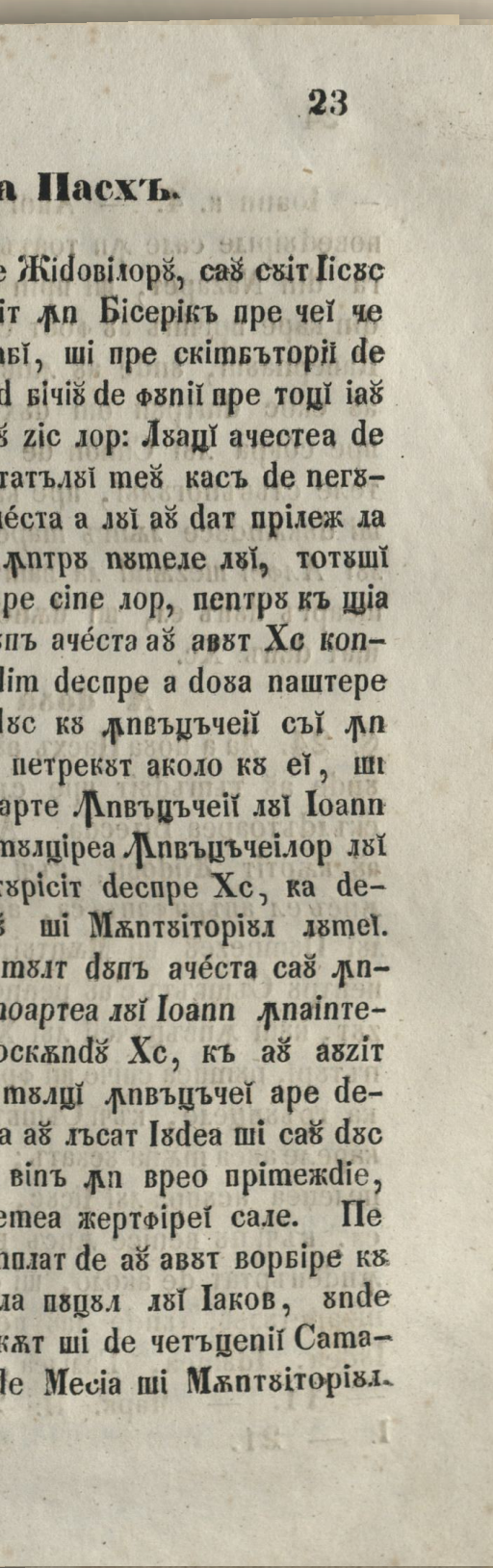


Challenges • OCR and Transliteration





Proposal





Experiments

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

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

CER Character Error Rate

BCER Bag of Characters Error Rate

S Substitution

D Deletion

I Insertion

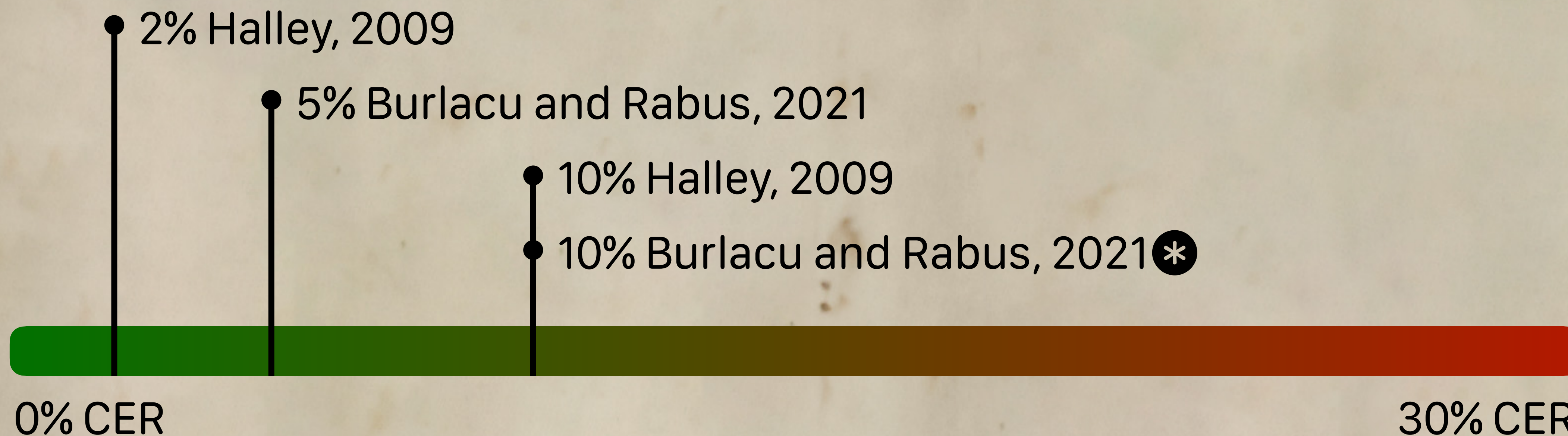
N Number of Characters (baseline)

W Number of Words



Experiments

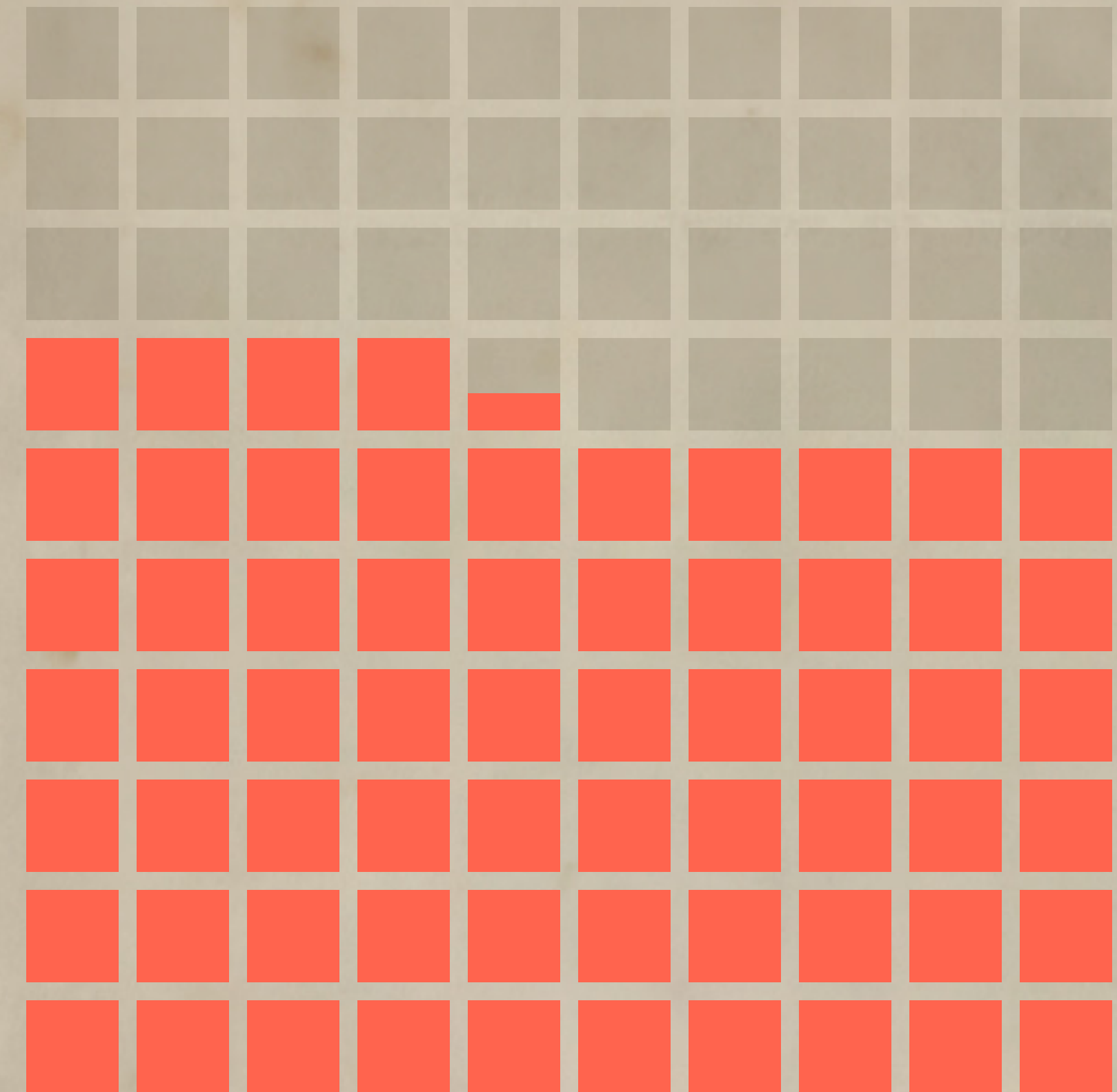
$$CER = \frac{S + D + I}{N}$$





Experiments • Setup

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





Experiments • Setup

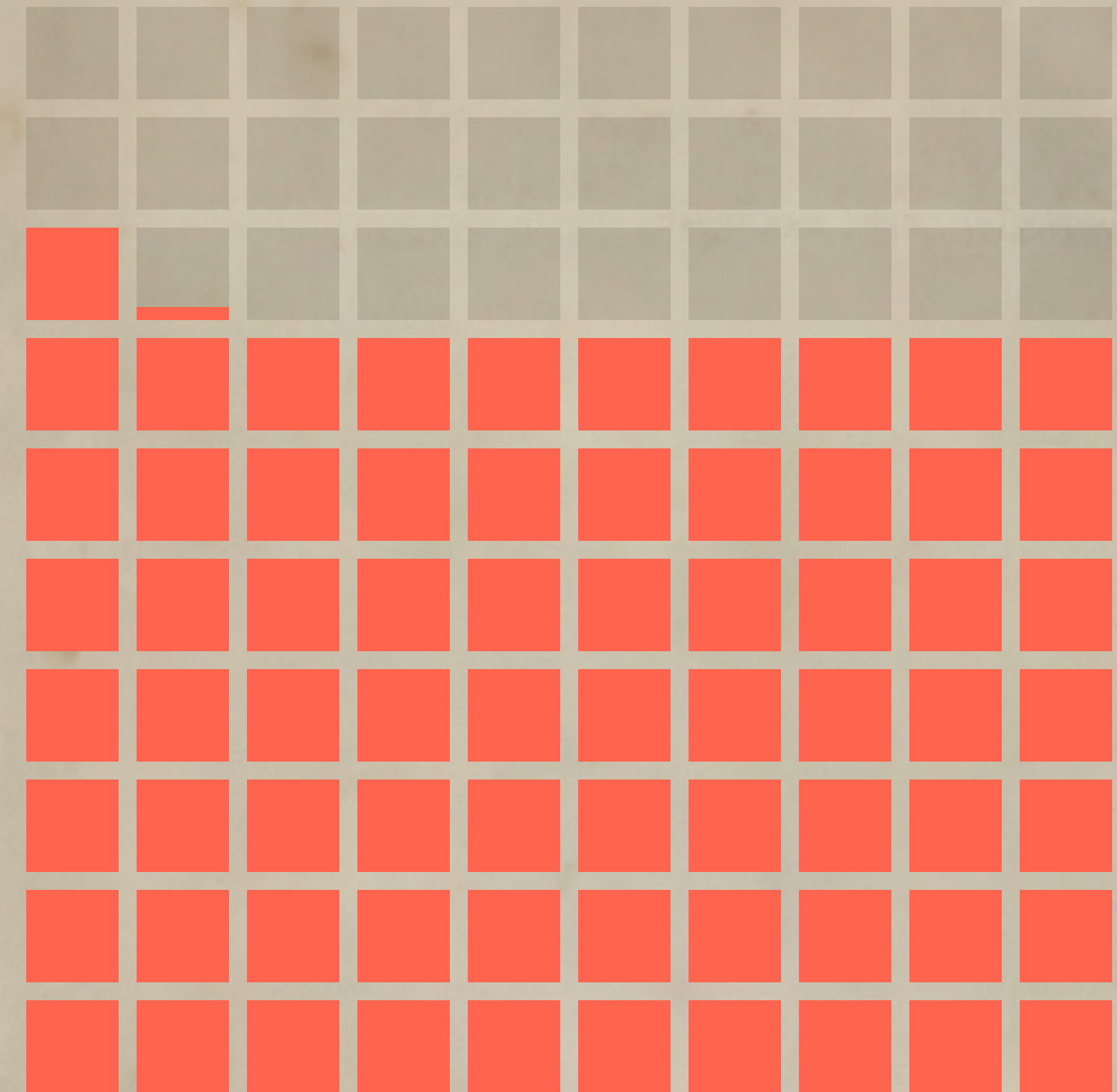
- ▶ 30 pages
- ▶ 24148 characters (64.4% Cyrillic)
- ▶ Page average
- ▶ $61.8\% \pm 9.6\%$ Cyrillic





Experiments • Setup

- ▶ 30 pages
- ▶ 24148 characters (64.4% Cyrillic)
- ▶ Page average
- ▶ 61.8% \pm 9.6% Cyrillic
- ▶ Test page
- ▶ 745 characters (71.14% Cyrillic)





Experiments • Setup





Experiments • Setup



3-Fold Test Set

◀ Best results

10-Fold Test Set

29-Fold Test Set



Experiments • Setup

— 12 —

ние, къ татъ се зѣтъ къ възъ плъчере ла фантеле
лѣ, че ел дн аскъска инимей змилѣ ле фаче. Ел
ние, къ шѣ лѣкрѣрѣ челе маѣ пѣдѣн къмпѣнѣтоаре —
дн змилѣдѣ фѣкѣте сѣнт плѣне де мерѣт ла Dѣmpеzeѣ.
Ел нѣе дн адевр, къ кѣ атѣт маѣ тѣрѣт се ва дн
пѣдѣ одѣнѣарѣ, кѣ кѣт маѣ адѣнк сѣ днѣѣосѣт прѣн
змилѣдѣ инимей дн лѣмеа ачеаста шѣ сѣ днѣнѣмѣчѣт.
Шѣ аша вѣне е челеѣ змилѣт, шѣ 'н ачеастѣ шѣ 'н чеа
алаатѣ лѣме.

Е вѣне ѣвѣдѣлор, сѣ пѣ сѣферѣм дѣрѣ маѣ тѣлѣт
дѣнѣре поѣ пре дѣхѣл челе реѣ ал лѣмѣѣ, челе че вреа
лѣмаѣ а се тѣрѣфѣ шѣ а се тѣрѣ! Сѣ кѣѣтѣм а дѣнокѣмѣ
фачереа шѣ пѣзѣнѣдѣ, лѣкрѣреа шѣ сѣферѣнѣдѣ поастрѣ
ауѣа, ка прѣн ачелеа сѣнѣр Dѣmpеzeѣ сѣ се лѣѣде, шѣ
лѣмеле лѣѣ сѣ се преа тѣреаскѣ! Шѣ ка сѣ вѣедѣѣм
пѣкрѣмѣт дн змилѣдѣ, аша воѣм поѣ а авеа пѣнѣчетѣт
дѣнѣнѣтеа оѣлор пѣкѣтошѣеа шѣ нѣмѣчѣеа поастрѣ, шѣ
нѣчѣ кѣнд а зѣѣта тѣрѣтѣл есеплѣ, каре Мѣнтѣѣторѣл
челе змилѣт 'л а дѣт поаѣ. De тѣлѣте орѣ воѣм а пѣ
адѣче амѣнѣте де кѣвѣнтеле че вѣсерѣка 'нѣнѣте де кѣте-
ва зѣле пѣ а стрѣгат: „Кѣѣегѣ о омѣле, къ дѣрѣжѣнѣ
шѣ чепѣжѣнѣ еѣѣ, шѣ ѣарѣ дѣрѣжѣнѣ шѣ чепѣжѣнѣ те веѣ
фаче.“ Ка вѣмѣнѣл челе змилѣт воѣм де тѣлѣте орѣ —
маѣ алес дн тѣмпѣл с. аѣѣѣн (пост) плѣнѣ де кѣнѣдѣ
а пѣ вѣте пѣнтѣл грѣнѣд: „Dѣmpеzeѣле фѣ дѣндѣрат
мѣе пѣкѣтосѣлѣѣ!“ Дн таѣна покѣнѣдѣѣ воѣм поѣ а пѣ

— 12 —

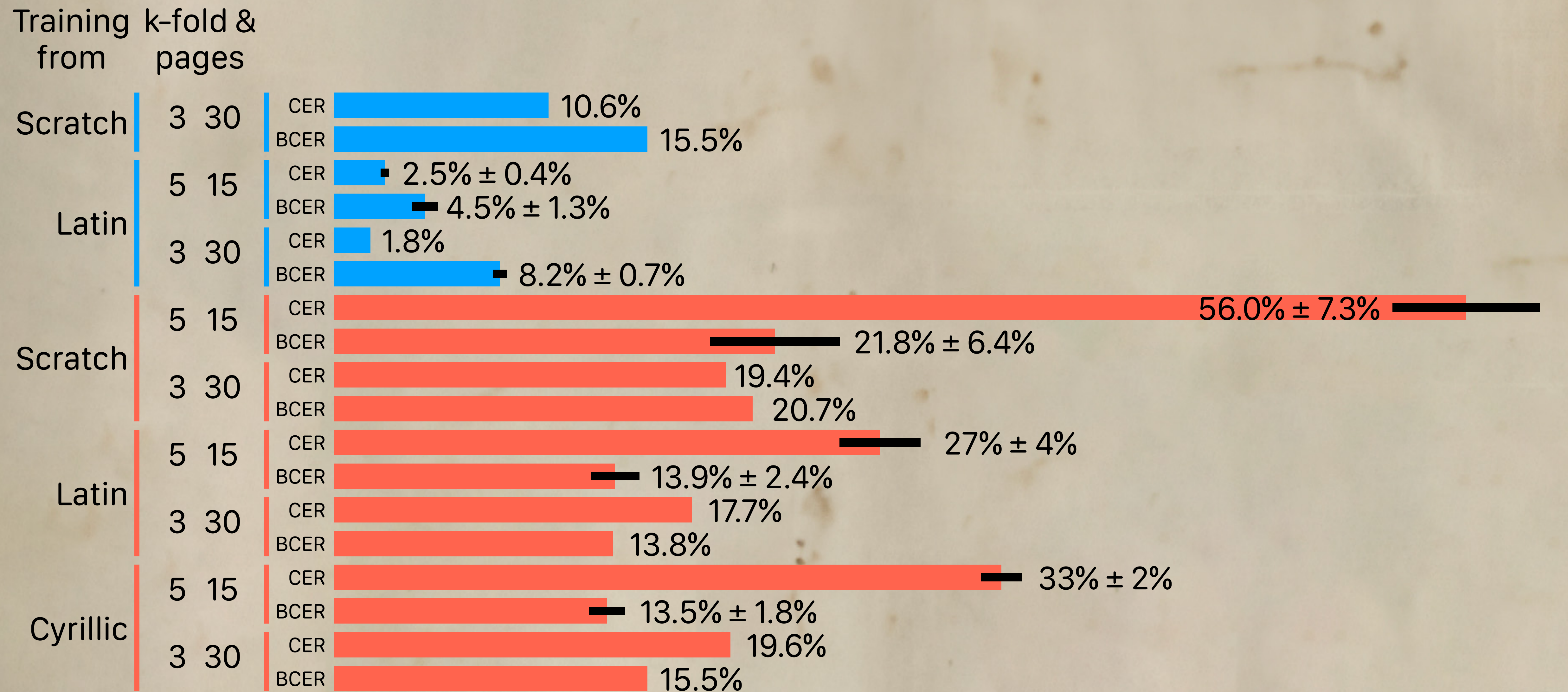
зѣтъ къ възъ плъчере ла фантеле
кѣска инимей змилѣте ле фаче. Ел
ле челе маѣ пѣдѣн къмпѣнѣтоаре —
е сѣнт плѣне де мерѣт ла Dѣmpеzeѣ.
къ кѣ атѣт маѣ тѣрѣт се ва дн
кѣт маѣ адѣнк сѣ днѣѣосѣт прѣн
лѣмеа ачеаста шѣ сѣ днѣнѣмѣчѣт.
ѣ змилѣт, шѣ 'н ачеастѣ шѣ 'н чеа

р, сѣ пѣ сѣферѣм дѣрѣ маѣ тѣлѣт
л челе реѣ ал лѣмѣѣ, челе че вреа
шѣ а се тѣрѣ! Сѣ кѣѣтѣм а дѣнокѣмѣ
а, лѣкрѣреа шѣ сѣферѣнѣдѣ поастрѣ
сѣнѣр Dѣmpеzeѣ сѣ се лѣѣде, шѣ
преа тѣреаскѣ! Шѣ ка сѣ вѣедѣѣм
дѣ, аша воѣм поѣ а авеа пѣнѣчетѣт
кѣтошѣеа шѣ нѣмѣчѣеа поастрѣ, шѣ
тѣрѣтѣл есеплѣ, каре Мѣнтѣѣторѣл
поаѣ. De тѣлѣте орѣ воѣм а пѣ
нтеле че вѣсерѣка 'нѣнѣте де кѣте-

„Кѣѣегѣ о омѣле, къ дѣрѣжѣнѣ
ѣарѣ дѣрѣжѣнѣ шѣ чепѣжѣнѣ те веѣ
челе змилѣт воѣм де тѣлѣте орѣ —
с. аѣѣѣн (пост) плѣнѣ де кѣнѣдѣ
ѣнѣд: „Dѣmpеzeѣле фѣ дѣндѣрат
л таѣна покѣнѣдѣѣ воѣм поѣ а пѣ
нѣѣ змилѣте фѣрѣ рѣтраѣере шѣ
сѣѣрѣнѣт къ змилѣтѣ инѣмѣ а пѣ
а драгостѣѣ, зѣтѣнд а зѣче ка
сѣнт вѣедѣѣлѣ, ка сѣ дѣнѣтѣ сѣѣт



Experiments • Results



* 10 000 training epochs



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.
- ▶ Errors in transliterating certain double consonants like ll (double L).



Conclusions

- ▶ We addressed transliteration of Romanian texts from the 19th 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