Bringing the Old Writings Closer to Us: Deep Learning in Deciphering Cyrillic Romanian

Eduard Coman[1], Andreea Dascălu[1], Claudiu Marinescu[1], Petru Rebeja[2], Anca Vasilescu[1],
Nicolae Cleju[3], Gabriela Haja[4], Dan Cristea[2,5]

[1] Department of Mathematics and Computer Science, “Transilvania” University of Brașov

[2] Doctoral School of Computer Science, “Alexandru Ioan Cuza” University of Iași

[3] Department of Electronics and Telecommunications Engineering, “Gheorghe Asachi” Technical University of Iași

[4] “Alexandru Philippide” Institute of Philology, Iași branch of the Romanian Academy

[5] Institute for Computer Science, Iași branch of the Romanian Academy

eduard.coman@student.unitbv.ro, andreea-a.dascalu@student.unitbv.ro, claudiu.marinescu@student.unitbv.ro,
petru.rebeja@gmail.com, vasilex@unitbv.ro, nikcleju@gmail.com, gabihaja@gmail.com, dan.cristea@acadiasi.ro

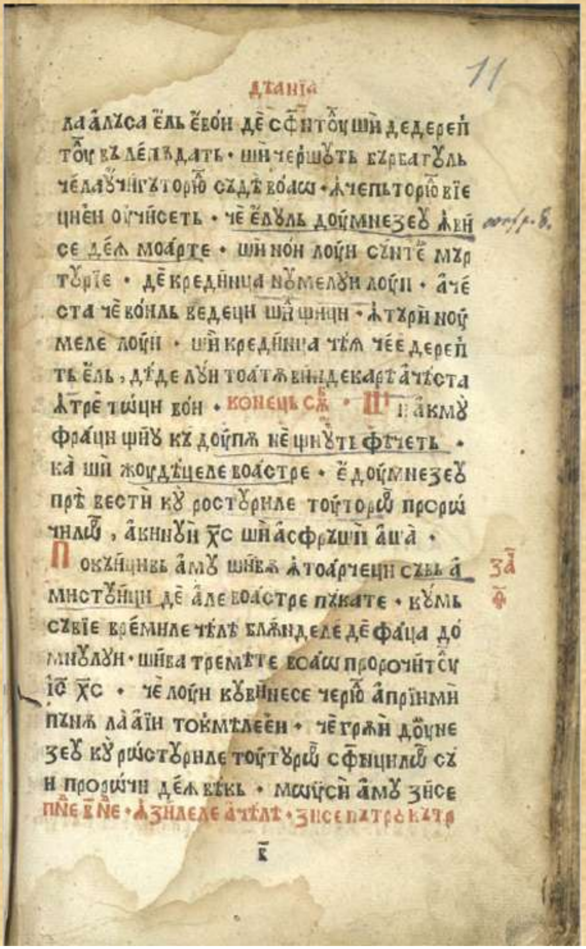
Motivation

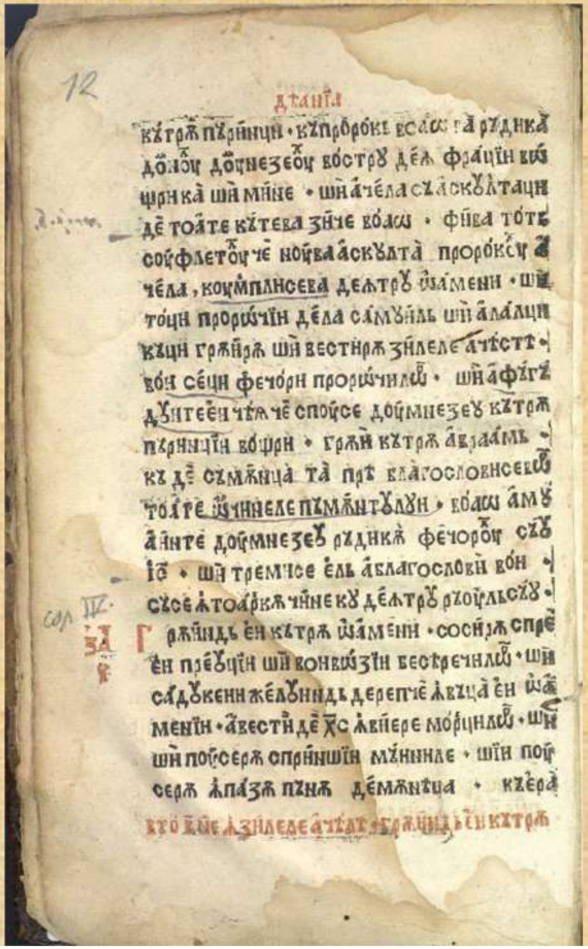
- Cyrillic alphabet: used for writing Romanian, on the territory of today's Romania between the 16th and the middle of 19th century
  - 1862: prince Alexandru Ioan Cuza, the ruler of united Moldova and Wallachia, imposes Latin as the official alphabet
- Only a small part of the thousands of Cyrillic Romanian documents inventoried in libraries and official archives, in Romania and abroad [1, 2], have been published in critical editions
- Researchers, students, editing houses lack a technology able to interpret them in the Latin script

[1] Bianu I., Hodoş N., and Simionescu D. (1903-1944). Bibliografia românească veche. 1508–1830. Tom. I–V, Ediţiunea Academiei Române, Bucureşti, 2490 p.

[2] Cândea V. (2011, 2012, 2014). Mărturii românești peste hotare, Editura Biblioteca Bucureștilor, Editura Academiei Române, Editura Muzeului Literaturii

Examples of documents: print, good quality pages

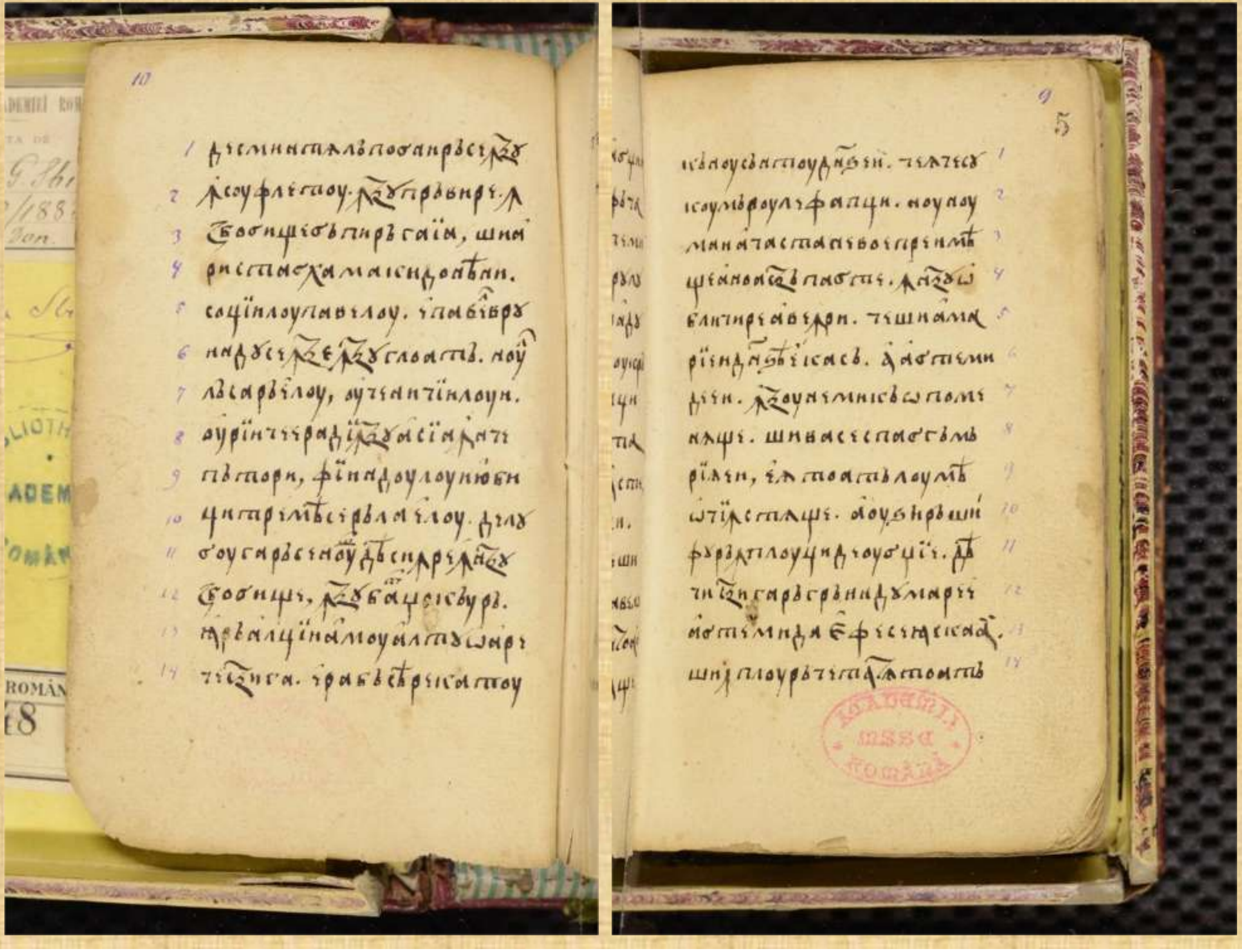




Apostol românesc, Brașov, 1563, printed by Coresi

Facsimiles: Gabriela Dumitrescu, Romanian Academy Library (RAL)

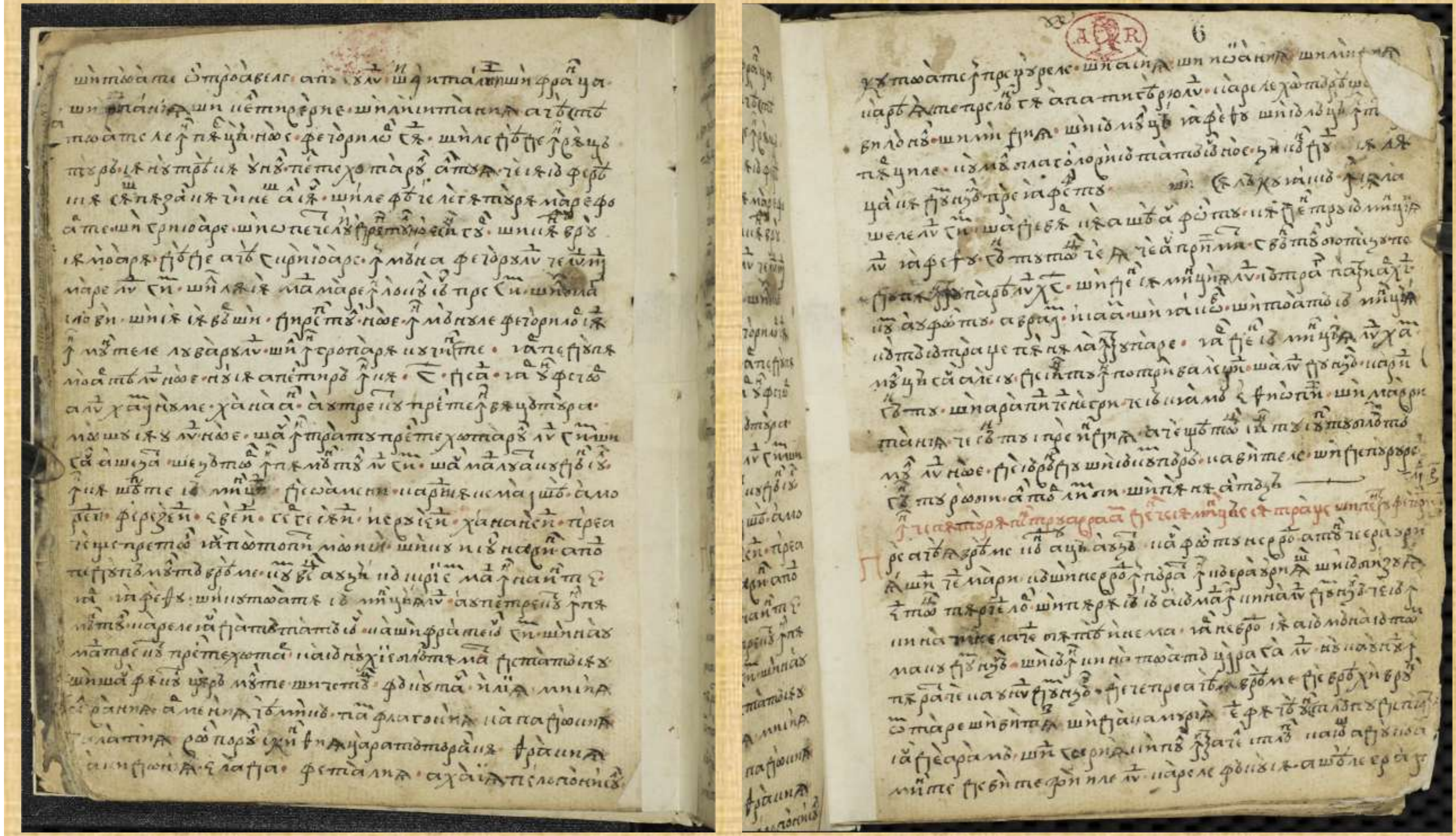
Examples: uncials, clean pages



Codicele Voronețean (The Voroneț Codex), a copy of a translation from Slavic, containing fragments of the New Testament

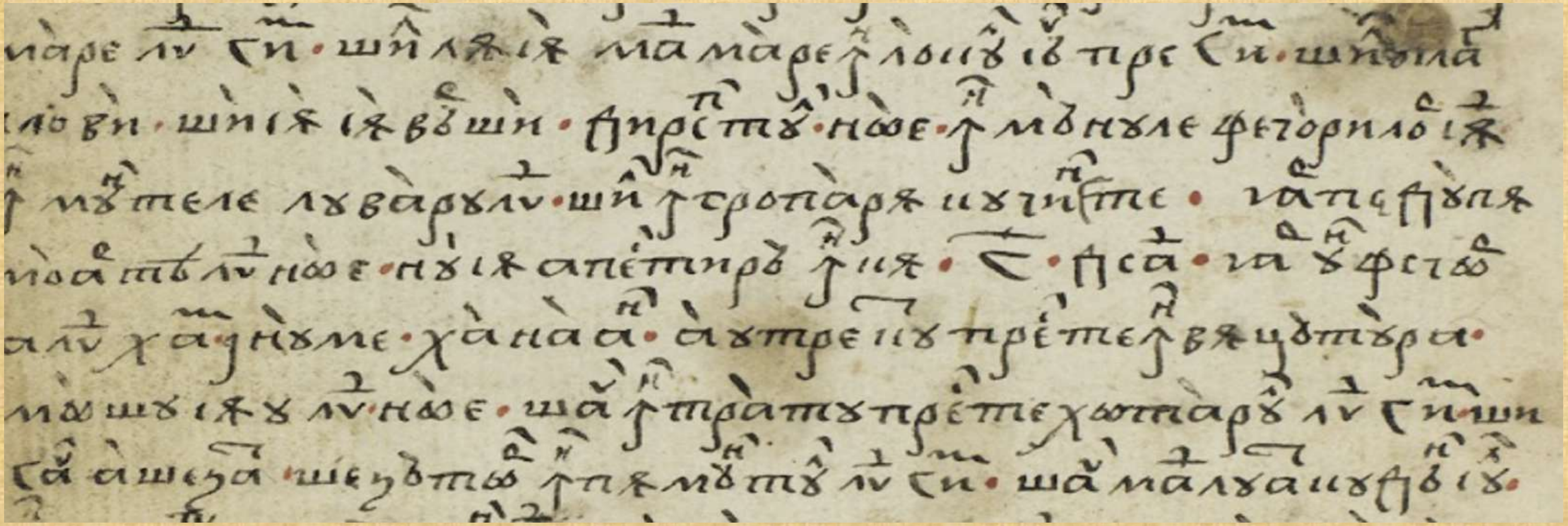
Facsimiles: Gabriela Dumitrescu, RAL

Example: Hronograf (sec. XVII), translation attributed to Nicolae Milescu Spătarul



Uncials in inclined lines, interlinear writing (characters and sequences over-posed)

A detail



нàре лꙋ си · шн лѫ їꙋ ма мàре ꙟлоциꙋ їꙋ пре си · шнꙋила
лоꙋвн · шни їꙋ їꙋ вꙋшн · Ꙟнрестꙋ ноѥ · ꙟ мꙋнчле фечорилоꙋ їꙋ
ꙟ мꙋтеле лꙋ варꙋлꙋ · шн ꙟ сротпàрѫ нꙋ ꙟнсте · гапиеꙟꙋлѫ
ноатѣ лꙋ ноѥ · нꙋ їꙋ апетнрѣ ꙟ нѫ · с · ꙟса · га ꙋ фечоꙋ
алꙋ хаамꙋнє · хàнаа · аꙋтреиꙋ претеле ꙟ вѫцꙋтꙋра ·
ноꙋ шꙋ їꙋ ꙋ лꙋ ноѥ · шаꙋ ꙟтрàтꙋ претнѥ хотарꙋ лꙋ си · шн
са ашеза · шеꙃꙋтоꙋ ꙟпѫмꙋтꙋ лꙋ си · шаꙋ малꙋанꙋꙟꙋїꙋ ·

Under development…

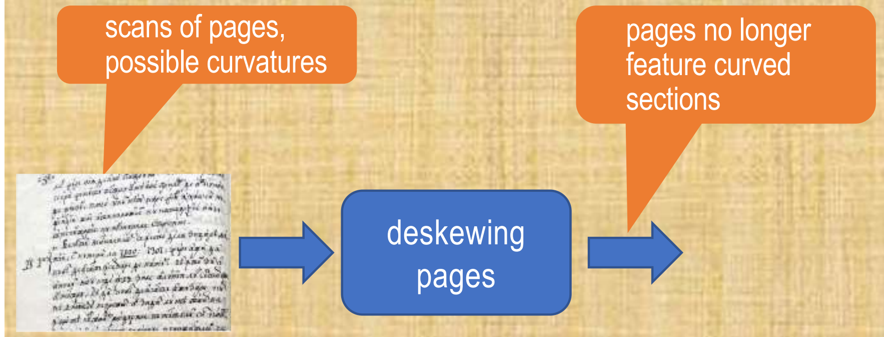
… a technology for responsive browsing of old Romanian documents by deciphering the Cyrillic writings into the Latin alphabet

- a large collection of scanned Cyrillic Romanian documents has been acquired
- metadata and manual annotation of graphical objects have been added in pages
- deep learning models to detect rows and characters in pages have been implemented and tested

Still to be done:

- deskewing curvatures of lines
- linearisation of writing, including interlinear
- labelling individual Cyrillic characters with Latin symbols
- assembling characters and sequences into words of the Old Romanian

A bird-eye view of the technology

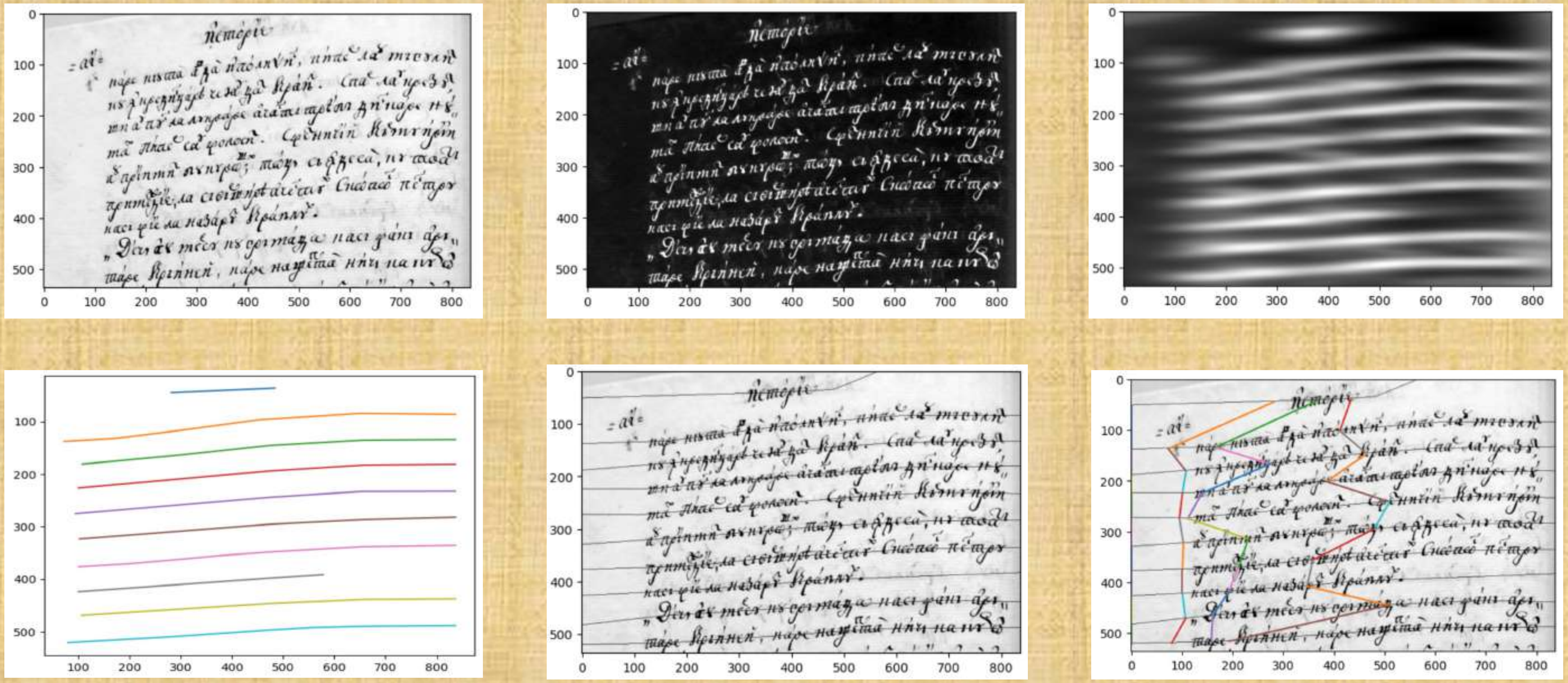


scans of pages, possible curvatures

deskewing pages

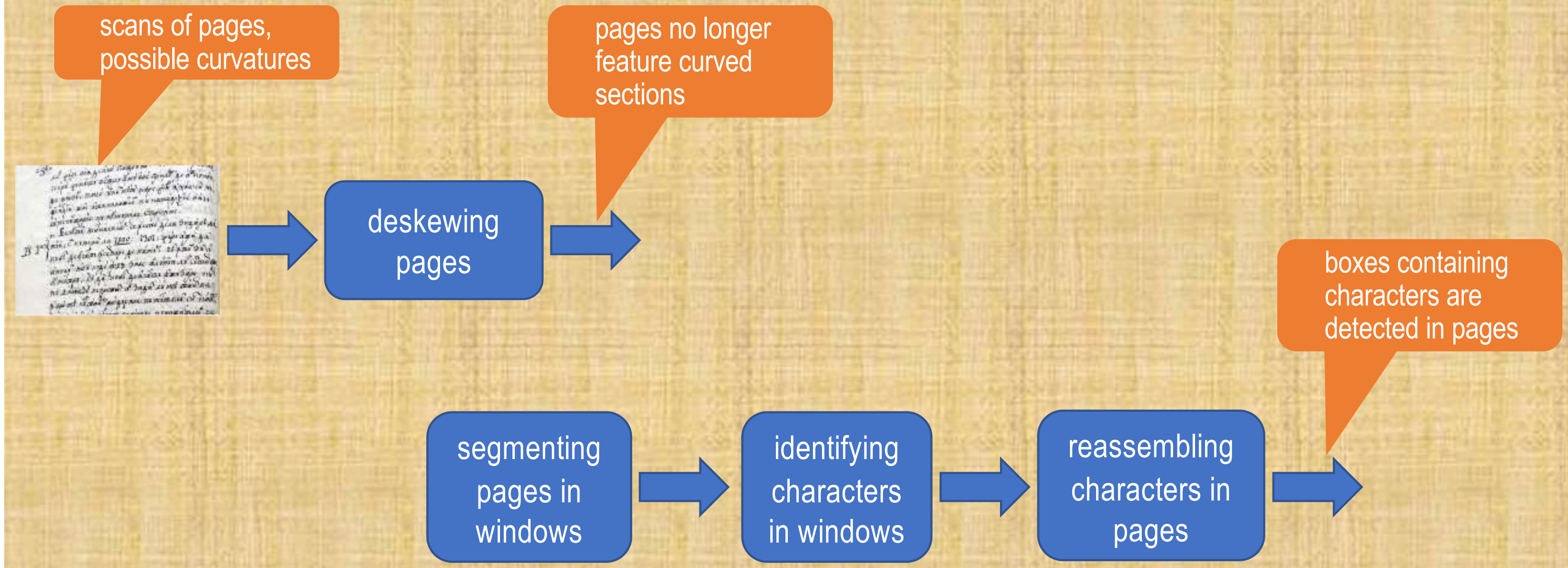
pages no longer feature curved sections

Deskewing pages



Work in progress

A bird-eye view of the technology



scans of pages, possible curvatures

deskewing pages

pages no longer feature curved sections

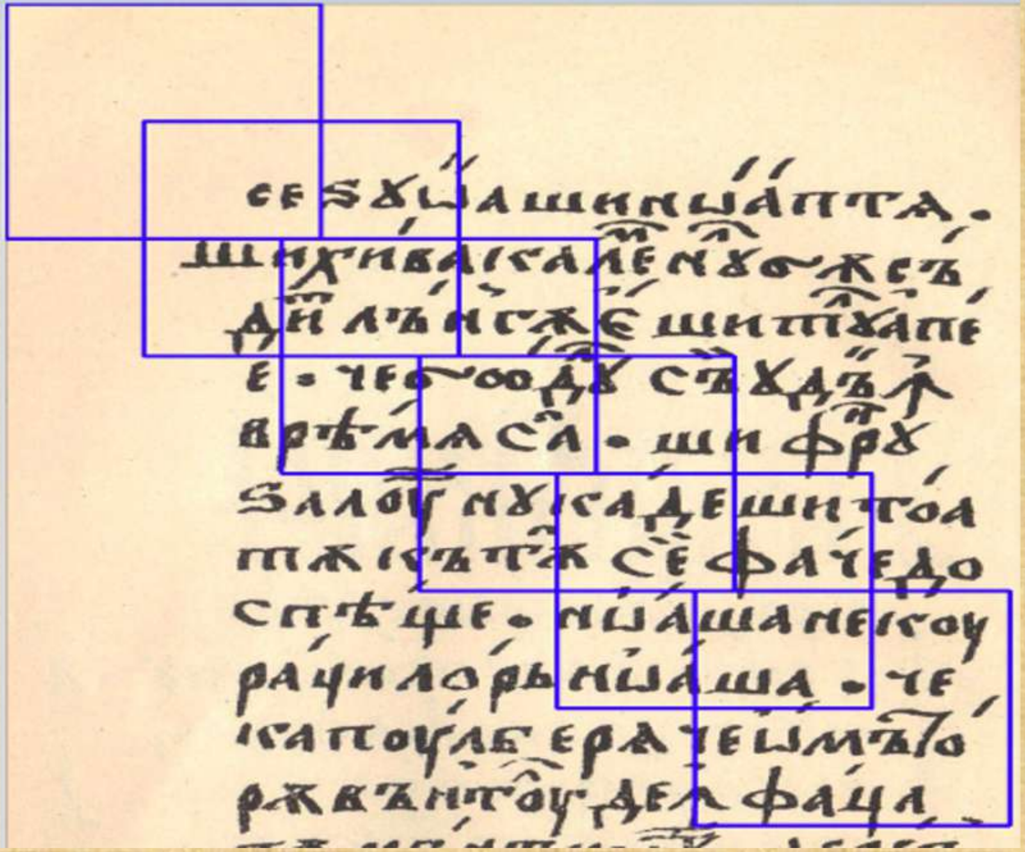
segmenting pages in windows

identifying characters in windows

reassembling characters in pages

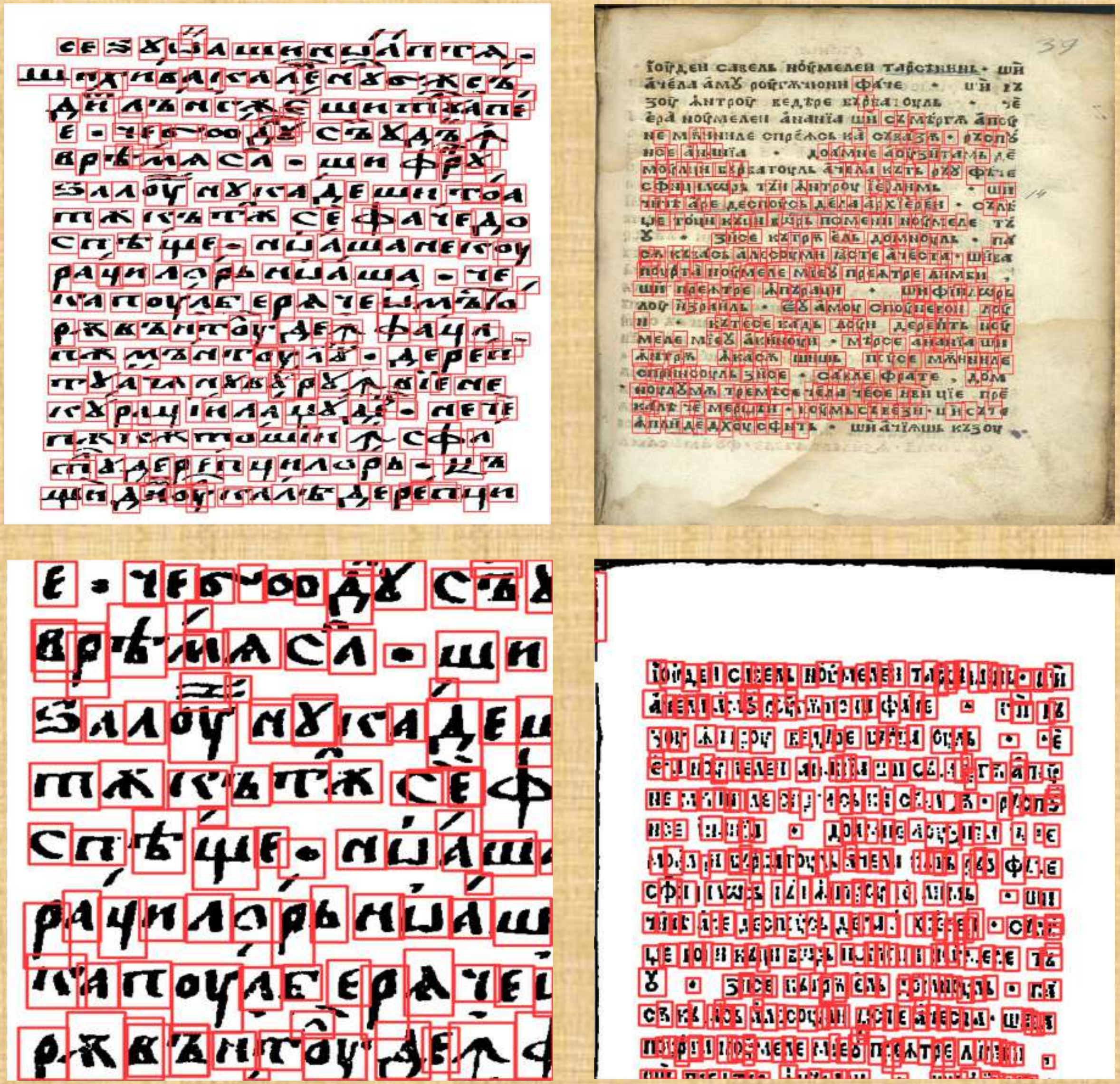
boxes containing characters are detected in pages

The upper part of a page in the process of window segmentation



се ѕꙋѡашиниѡаптѧ.
шихривалсалениꙋѡжесъ
дин лъникѫє шинипиꙋхапе
е. чео одꙋ съхадъ
врѣмаса. ши фрꙋ
ѕалоу нꙋкадешитоа
пѫ іоъ тѫ се фачедо
спѣще. ниѡаша некоу
рацилорь ниѡаша. че
капоуле ера чеѡмъто
рѫвъниоу деа фаца

Characters detection with a Yolo.v5 implementation: best results in a window segmentation



in page segmentation

in window segmentation

From (Cristea *et al.*, 2022)

Characters detection

Experiments

- Data from 13 Cyrillic Romanian books (prints and uncials) covering centuries XVIth to XVIIIth
- Training set in WS: 1000 windows of size 500 by 500 pixels, including 140,324 annotated boxes of characters.

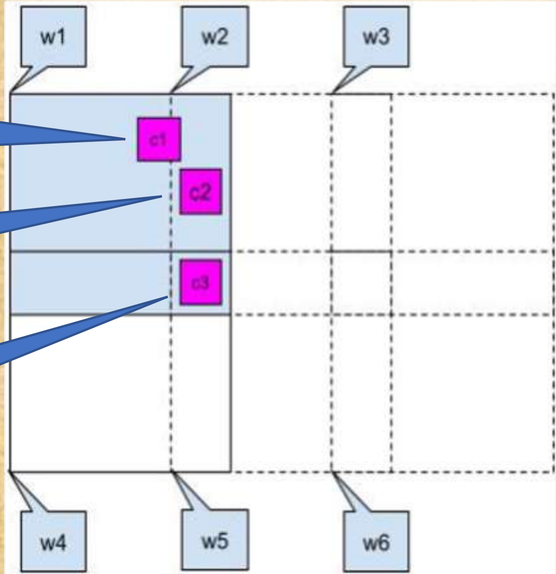
Validation

- Validation set: 500 windows of size 500 by 500 pixels, including 48,359 annotated boxes of characters.
- Average Precision at an Intersection over Union (IoU) threshold of 0.5 => 0.99
- Baseline: a YOLOv5 detection of characters on the original unsegmented pages and a conventional training procedure => 0.82

From: Cristea, D., Rebeja, P., Pădurariu, C. (2022). Applying YOLO.v5 Learning in Detecting Old Cyrillic-Romanian Characters, submitted to the 17th ConsILR Conference, Chișinău, Nov.

Characters in windows are assembled back in pages

Positions of 6 windows in the original page during detection

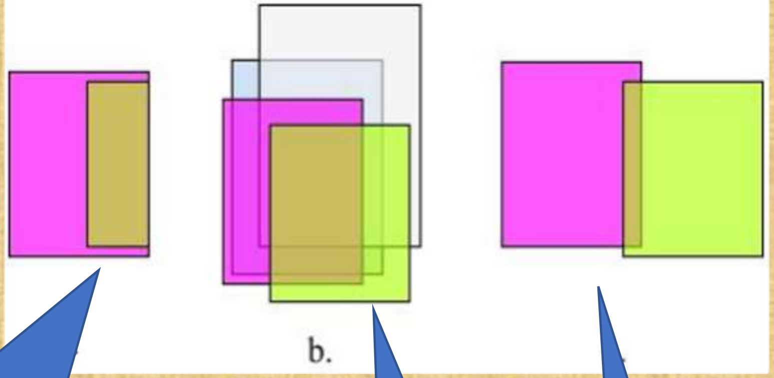


c1 appears once un-fragmented (in w1) and once fragmented (in w2)

c2 appears twice (in w1 and w2)

c3 appears 4 times (in w1, w2, w4 and w5

Possible relative positions of instances of character frames



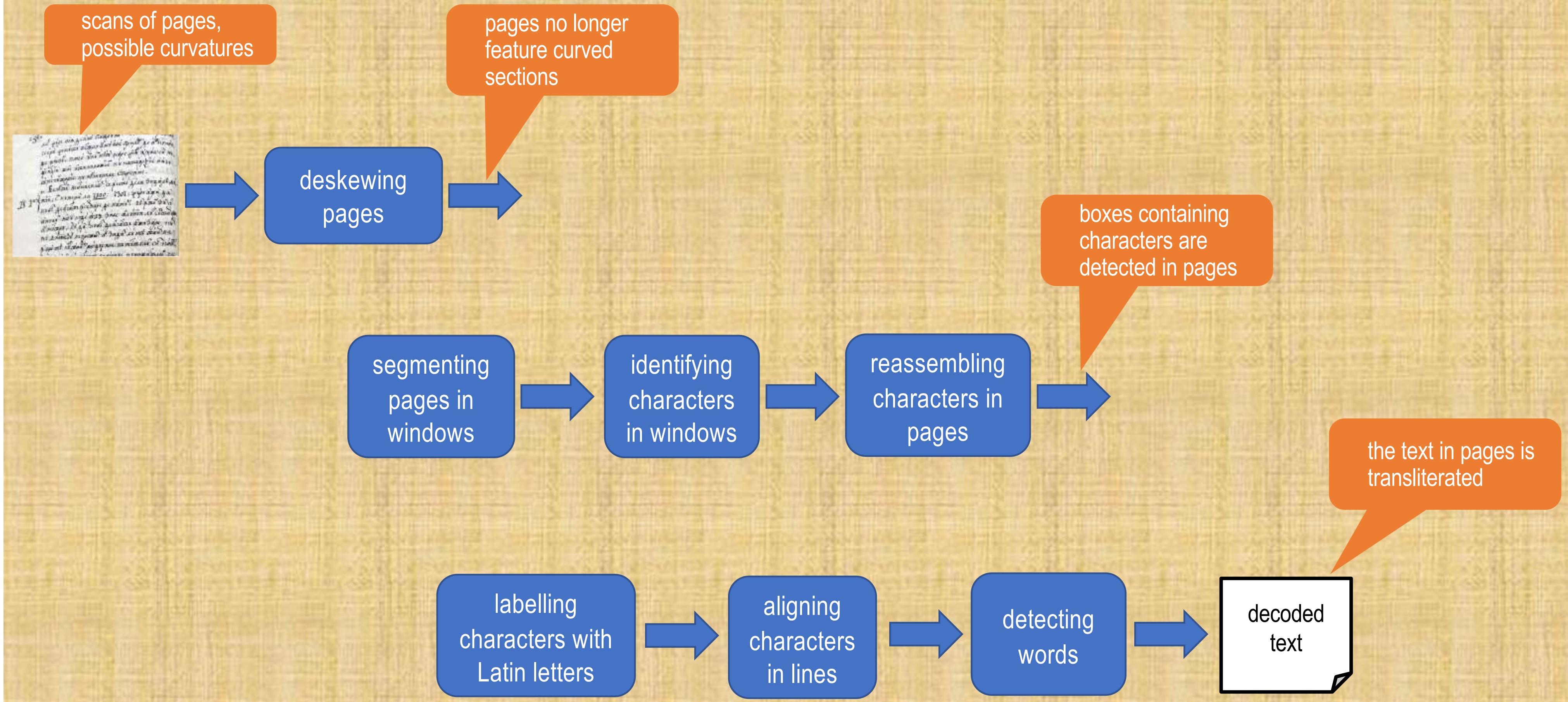
The same character is detected un-fragmented in one window and fragmented in another window

One character produces more frame instances

The frames of two characters slightly intersect (both collected in the same window).

Work in progress

A bird-eye view of the technology



scans of pages, possible curvatures

deskewing pages

pages no longer feature curved sections

segmenting pages in windows

identifying characters in windows

reassembling characters in pages

boxes containing characters are detected in pages

labelling characters with Latin letters

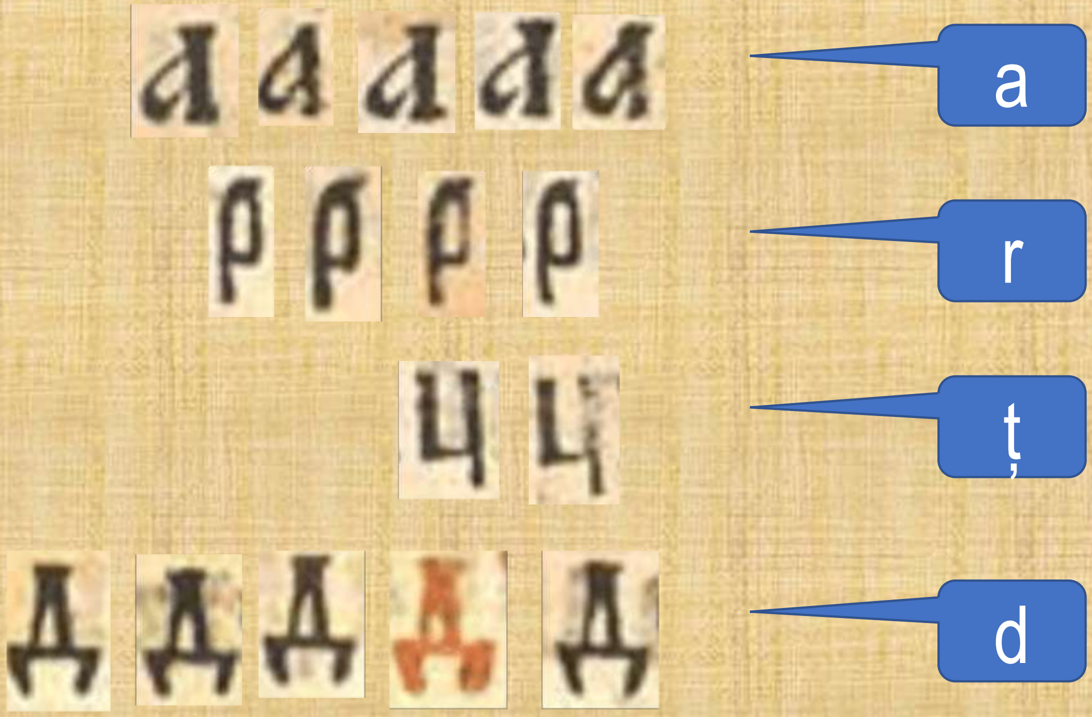
aligning characters in lines

detecting words

decoded text

the text in pages is transliterated

Labelling characters is a Character Recognition task

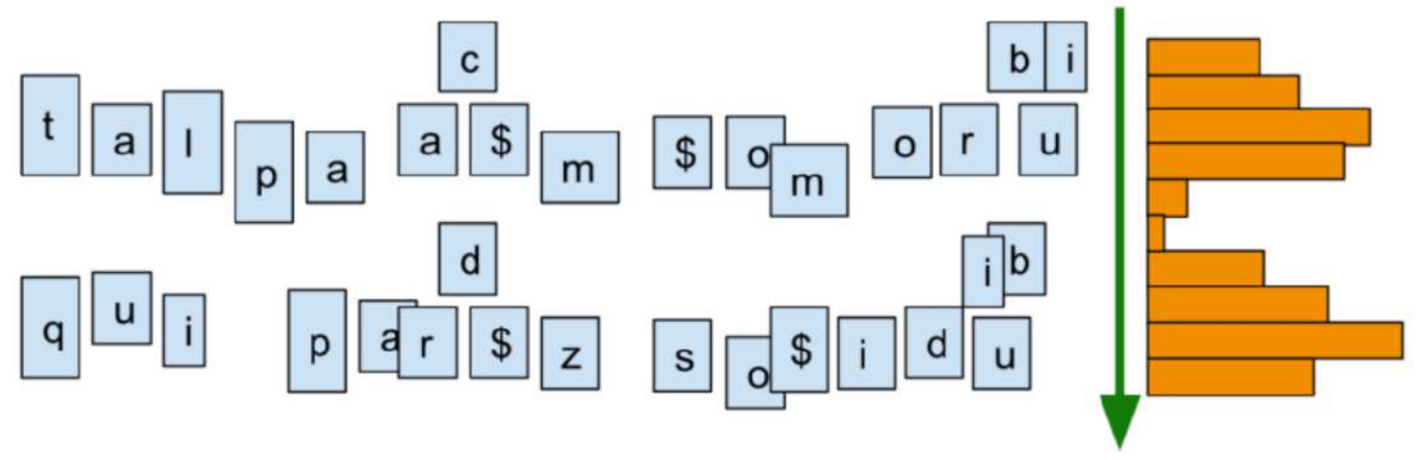


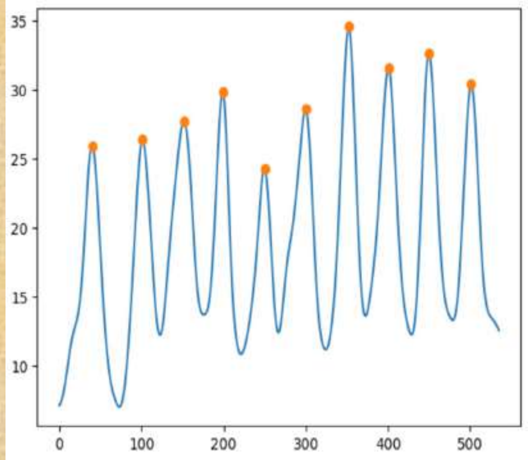
- Till now, an F1-measure of 0.84 with a classical linear classifier
  - including more convolutional layers (with descending no. of filters), wedged with max pooling, flattening, dense and dropout layers

Work in progress

Characters are recognised, but they are not aligned in lines

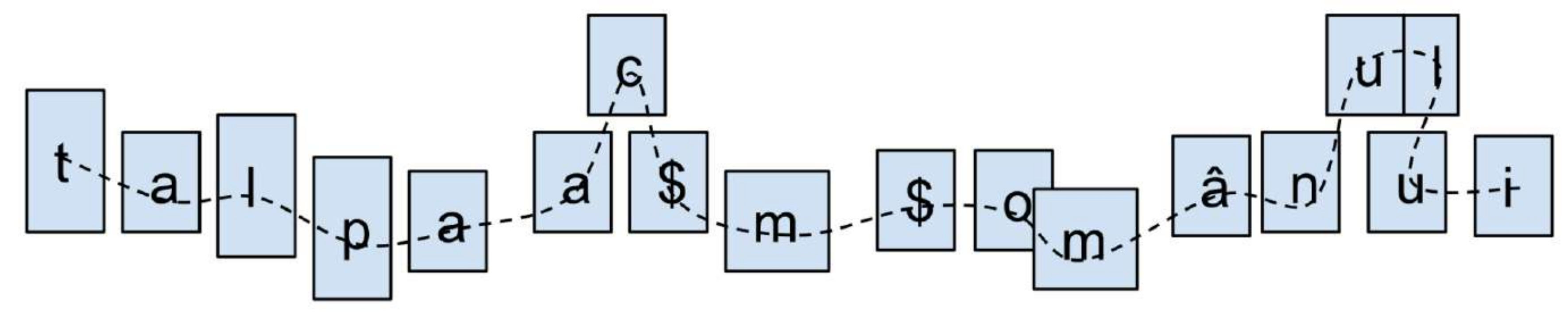
- Vertical position of character boxes help to detect the position of lines
- Compute the peaks on the histogram of pixels



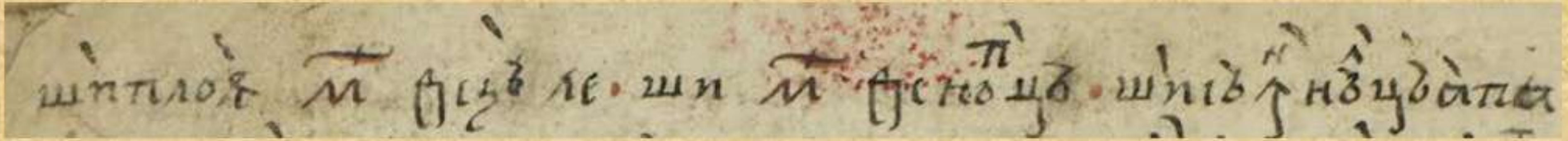


Appreciate the position of interlinear characters on the lines detected in the histogram

- Horisontal position of character boxes help to negociate the position of characters in lines



Guessing missing white spaces and reconstructing words…



si▫ploă 60 de▫zî⊠le și 60 de▫no↑p↓ți. si▫să î↑n↓nă↑l↓ță▫apa

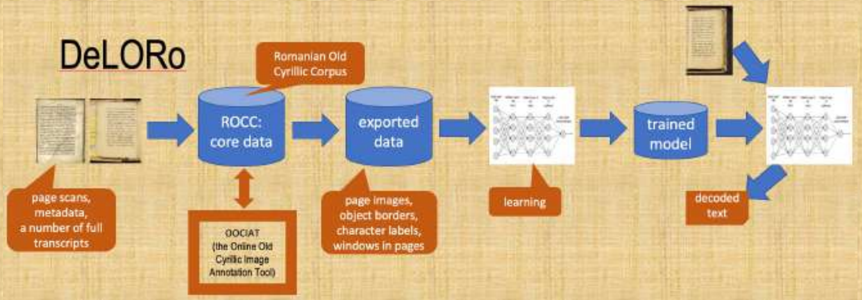
Hronograf, pag. 2 verso, sec. XVII

Work in progress

Conclusions

- We described a pipeline of modules aimed to configure an AI rooted technology for curation of old Romanian writings
- Our approach is a follow up of the DeLORo project, opening a door for gaining a deeper inside into old Romanian documents, understanding what's in there, and thus recuperating the Cyrillic content in the Latin script
- Among many possible applications: reverse search from text into image…
- Further developments:
    - Extending the character recognition module to decipher characters of the ”transition alphabet”, used in Romanian provinces in the middle of the XIXth
    - Building POS-taggers and lemmatisers for old Romanian, with particularisations for diachronicity and synchronicity
        - “How can I find occurrences of the word <<haină>> in the 12 novels from the XIXth century included in the collection of *Astra Data Mining* of the journal Transilvania Sibiu?”

Acknowledgements


DeLORo
Romanian Old Cyrillic Corpus
ROCC: core data
exported data
trained model
page scans, metadata, a number of full transcripts
OOCIAT (the Online Old Cyrillic Image Annotation Tool)
page images, object borders, character labels, windows in pages
learning
decoded text

- Project DeLORo PN-III-P2-2.1-PED-2019-3952, no. 400PED: “Deep Learning for Old Romanian” (2020-2022).
- The members in the project from the Institute of Computer Science, the Institute of Philology - Iași branch of the Romanian Academy, and from the Faculty of Mathematics and Computer Science of the University of Bucharest.
- Romanian Academy Library, Bucharest – for offering scans of old books and metadata.
- Assoc. prof. Roxana Vieru and her students from the Master in Paleolinguistics, Faculty of Letters from “Alexandru Ioan Cuza” University of Iași, which contributed with hundreds of hours of manual annotation to acquire data used to train the technology.

Thank you!